McGRAW-HILL

ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

2 ANS-BIN

www.MHEST.com

A

Anseriformes — Azurite

Anseriformes

An order of birds comprising two families, the screamers (Anhimidae) of South America and the worldwide waterfowl (Anatidae). They are closely related to the Galliformes (fowls), with the screamers being a rather intermediate group. The waterfowl and the fowls are closely related within the Aves and are often placed within a distinct superorder, the Galloanserae. The large, flightless diatrymids (giant ground birds) of the early Tertiary appear to be specialized offshoots of the Anseriformes. *See* GALLIFORMES; GASTORNITHIFORMES.

Classification. The order Anseriformes is divided into the suborder Anhimae, containing the single family Anhimidae (screamers; 3 species), and the suborder Anseres, including only the family Anatidae (ducks, geese, and swans; 147 species). The waterfowl (Anatidae) are further subdivided into seven subfamilies: the primitive Anseranatinae (magpie goose of Australia; this genus is sometimes placed in a separate family), Dendrocygninae (tropical tree ducks), Anserinae (swans and geese), Tadorninae (shelducks of the Old World and South America), Anatinae (true ducks), Merginae (mainly Northern Hemisphere sea ducks and mergansers), and Oxyurinae (stiff-tailed ducks).

Screamers. The South American screamers (*Anhima*) are turkey-sized, fowllike aquatic birds with a short, heavy, chickenlike beak. The legs are of medium length and heavy, with four toes having only a basal web. They fly slowly but soar well. Screamers live in marshes, walk on mats of floating vegetation, and sometimes swim. They feed on vegetable matter, are gregarious, and nest in solitary pairs. The nest is a pile of rushes in reed beds. The clutch is three to six eggs. Both parents incubate and care for the chicks, which are downy and leave the nest right after hatching.

Waterfowl. The waterfowl vary in size from the tiny pygmy geese (*Nettapus*) to large swans (*Cygnus*). They occur worldwide in fresh and marine (coastal) waters. All species have strong legs and feet with webbed toes and a flattened bill with comblike lamellae or teeth. The plumage is waterproof and varies from pure white to multihued to all black; females usually have a brown, cryptic plumage. The tongue is large and fleshy and serves in filter-feeding. Waterfowl feed on both plants and animals, obtained by filtering (many true ducks), grazing (swans and geese), and diving (sea ducks, mergansers, stiff-tailed ducks, and some true ducks). Most diving ducks feed on mollusks and water plants. In mergansers (*Mergus*), the lamellae of the narrowed bill are modified into toothlike structures suitable for catching fish. All species swim well, and most are strong fliers. A few, such as the Hawaiian goose (*Branta sandvicensis*), are almost completely terrestrial. Two of the three species of the South American steamer ducks (*Tachyeres*) are flightless, as were several larger fossil species.

Most species are gregarious except at breeding; the Australian black swan (*Cygnus atratus*) is an exception as it breeds in large colonies. Waterfowl are monogamous with a strong pair bond (some mate for life) and elaborate courtship. Courtship and pair formation is usually completed on the wintering ground, with the pair returning to the birth area of the female. The nest is usually solitary and can be placed on the ground, in burrows, in holes in trees, on cliffs, and in marshes; it is usually lined with down. The clutch is from 2 to 16 eggs, incubated by both sexes or in most species by the female alone. The young leave the nest right after hatching and are cared for by both sexes (for example, swans and geese) or by the female alone (most ducks). The males molt into an eclipse plumage similar to the female plumage and then into the breeding plumage. All flight feathers are molted simultaneously, leaving the bird flightless for several weeks.

Many high-latitude species migrate, wintering as large mixed flocks in more southern fresh waters or along the marine coast. Steller's eider (*Polysticta stelleri*) and spectacled eiders (*Somateria fischeri*) winter in large flocks in open waters within the frozen Bering Sea.

The magpie goose (*Anseranas semipalmata*) of Australia is primitive among the anatids, and is a good intermediate form between the screamers and the advanced waterfowl. Its feet are only partly webbed, and it molts its flight feathers gradually instead of simultaneously. However, it possesses the typical broadened, somewhat-flattened bill, with a thin covering of skin found in the waterfowl.

Fossil record. The screamers (Anhimidae; 3 species) do not have a fossil record. Waterfowl are first found as fossils in the early Oligocene, but not commonly until the Miocene, which is surprising because the heavy bones of aquatic anatids would be expected to fossilize readily if these birds were present. Apparently, the radiation of the anatids into the modern subfamilies and genera only took place in the Miocene, as fossils known from the early Miocene are difficult to place into modern taxa. The much older long-legged and wading *Presbyornis* (Eocene) appears to be a primitive anseriform and may be a connecting link between the Charadriiformes and the Anseriformes. *See* CHARADRIIFORMES.

Economic importance. The waterfowl are of immense economic importance. The mallard (*Anas platyrhynchos;* the common domesticated duck), muscovy duck (*Cairina mosebata*), gray-lag goose (*Anser anser;* the common domesticated goose), and swan goose (*Anser cygnoides*) have been domesticated since ancient times for their flesh, eggs, and feathers. A special breed of domesticated duck is used in southeastern Asia for ridding the rice fields of animal pests. Species of waterfowl, especially the mute swan (*Cygnus olor*), the black swan (*Cygnus atratus*), and the mandarin duck (*Aix galericulata*), are commonly kept as ornamental birds on ponds in parks and estates. The down of the common eider (*Somateria mollissima*) is collected in great quantities for quilting. Waterfowl of many species are among the most popular game birds. Some species of waterfowl have become a pest in urban areas, notably the Canada goose (*Branta canadensis*) in the eastern United States, the black swan in New Zealand, and the mallard, which has been introduced into areas in the Southern Hemisphere where it has become common and has threatened some other members of the genus *Anas* by interbreeding and competition.

Conservation and extinction. Many species of waterfowl have declined in numbers from overshooting and in more recent years from loss of habitat, both breeding areas and wintering grounds. Several species are now extinct, one example being the Labrador duck (*Camptorhynchus labradorius*) of the east coast of North America. Recovery and spread of the trumpeter swan (*Cygnus buccinator*) is an outstanding example of saving a species of waterfowl close to extinction (see **illustration**). The Hawaiian goose (*Branta sandvicensis*) has been reintroduced successfully into its native islands from stocks bred in captivity. Most species of waterfowl are carefully protected and managed under strict conservation laws.

Trumpeter swan (*Cygnus buccinator*). (*Photo by Gerald and Buff Corsi; © 2000 California Academy of Sciences*)

Waterfowl are of intense interest to ornithologists, aviculturists, zoo keepers, conservationists, and game managers. They are one of the most popular groups of birds kept by zoos and aviculturists, with most species breeding readily in captivity. *See* AVES.

Walter J. Bock

Bibliography. J. Delacour et al., *The Waterfowl of the World*, 4 vols., 1954–1964; E. A. Johnsgard, *Ducks, Geese, and Swans of the World*, 1978; J. Kear, *Ducks, Geese, and Swans*, Oxford University Press, 2005.

Antarctic Circle

The Antarctic Circle is an imaginary line that delimits the northern boundary of Antarctica. It is a distinctive parallel of latitude at approximately 66°30′ south. Thus it is located about 4590 mi (7345 km) south of the Equator and about 1630 mi (2620 km) north of the south geographic pole.

All of Earth's surface south of the Antarctic Circle experiences one or more days when the Sun remains above the horizon for at least 24 h. The Sun is at its most southerly position on or about December 21 (slightly variable from year to year). This date is known as the summer solstice in the Southern Hemisphere and as the winter solstice in the Northern Hemisphere. At this time, because Earth is tilted on its axis, the circle of illumination reaches 23.50° to the far side of the South Pole and stops short 23.50° to the near side of the North Pole.

The longest period of continuous sunshine at the Antarctic Circle is 24 h, and the highest altitude of the noon Sun is 47° above the horizon at the time of the summer solstice. The long days preceding and following the solstice allow a season of about 5 months of almost continuous daylight.

Six months after the summer solstice, the winter solstice (Southern Hemisphere terminology) occurs on or about June 21 (slightly variable from year to year). On this date the Sun remains below the horizon for 24 h everywhere south of the Antarctic Circle; thus the circle of illumination reaches 23.50° to the far side of the North Pole and stops short 23.50° to the near side of the South Pole. *See* ARCTIC OCEAN.

Tom L. McKnight

Bibliography. A. H. Strahler and A. Strahler, *Introducing Physical Geography*, 4th ed., 2005; H. Veregin (ed.), *Rand McNally Goode's World Atlas*, 21st ed., 2004.

Antarctic Ocean

The Antarctic Ocean, sometimes called the Southern Ocean, is the watery belt surrounding Antarctica. It includes the great polar embayments of the Weddell Sea and Ross Sea, and the deep circumpolar belt of ocean between 50 and 60°S and the southern fringes of the warmer oceans to the north. Its northern boundary is often taken as 30°S (**Fig. 1**). The Antarctic is a cold ocean, covered by sea ice during the winter from Antarctica's coast northward to approximately 60°S.

The remoteness of the Antarctic Ocean severely hampers the ability to observe its full character. The sparse data collected and the more recent addition of data obtained from satellite-borne sensors have led to an appreciation of the unique role that this ocean plays in the Earth's ocean and climate. Between 50 and 60°S there is the greatest of all ocean currents, the Antarctic Circumpolar Current sweeping seawater from west to east, blending waters of the Pacific, Atlantic, and Indian oceans. Observed within this current is the sinking of cool (approximately 4°C; 39.2°F), low-salinity waters to depths of near 1 km (0.6 mi), which then spreads along the base of the warm upper ocean waters or thermocline of more hospitable ocean environments. The cold polar atmosphere spreading northward from Antarctica removes great amount of heat from the ocean, heat which is carried to the sea surface from ocean depths, brought into the Antarctic Ocean from warmer parts of the ocean. At some sites along the margins of Antarctica, there is rapid descent of cold (near the freezing point of seawater, −1.9°C; 28.6°F) dense water, within thin convective plumes. This water reaches the sea floor, where it spreads northward, chilling the lower 2 km (1.2 mi) of the global ocean, even well north of the Equator.

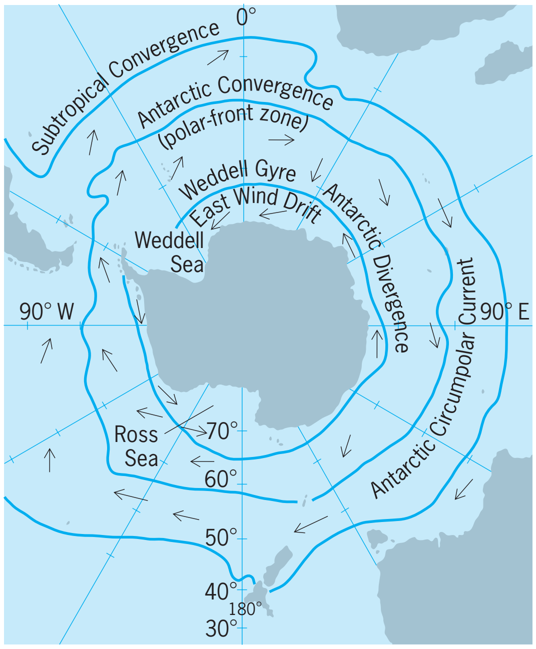


Fig. 1. Direction of the surface circulation and major surface boundaries of the Antarctic Ocean.

Oceanographic investigations. There has been much exploration of the Antarctic Ocean since the sixteenth century. Curiosity, exploration for rich unclaimed lands, and the need for new ship routes were the impetuses. There was the search for Terra Australis Incognita, the proposed great southern continent. Sightings of many of the larger islands in the southwestern Pacific Ocean were first considered to be this long-sought land, but circumnavigation proved that these were islands and, with the exception of a few of the smaller ones, not economically worth exploiting. In the 1770s, under the command of Captain James Cook, the first scientific voyages were carried out in the Antarctic Ocean and added much to the knowledge of the Southern Hemisphere's geography and climate. From the size of the icebergs observed and a latitudinal variation of the pack ice field, Cook surmised a frozen continent not symmetrical with the geographic South Pole. The nineteenth century witnessed further explorations of the Antarctic Ocean and the delineation of the coastlines of Antarctica. This series of explorations was sparked mainly by the growing seal and whaling industries.

In the twentieth century, further scientific studies have been carried out by the German ship *Deutschland* in the Weddell Sea and by the circumpolar expeditions of the English vessels *Discovery II* and *William Scoresby*. The United States has contributed much through the Navy Deep Freeze operations and the 1962–1972 circumpolar study of the Antarctic Ocean by the National Science Foundation–sponsored ship, USNS *Eltanin*. In the period 1975–1980, an intensive study of the Drake Passage was the focus of a program called International Southern Ocean Studies (ISOS). In October-November 1981, the first modern expedition well into the seasonal sea ice cover was carried out along the Greenwich Meridian in a joint U.S.-Soviet oceanographic program.

Currents. The major flow is the Antarctic Circumpolar Current, or West Wind Drift (Fig. 1). Along the Antarctic coast is the westward-flowing East Wind Drift. The strongest currents are in the vicinity of the polar front zone and restricted passages such as the Drake Passage, and over deep breaks in the meridionally oriented submarine ridge systems (**Fig. 2**).

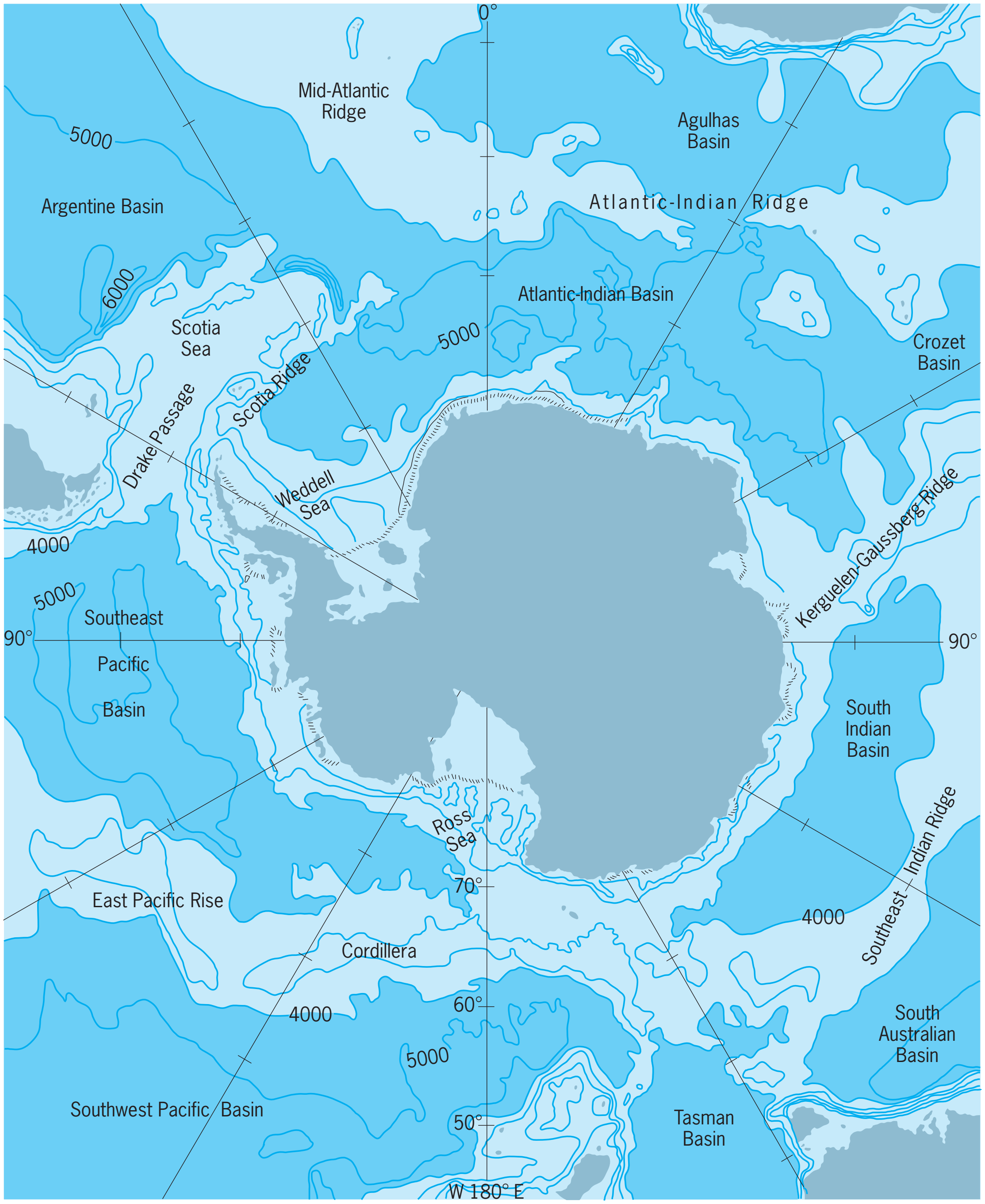

0°
Mid-Atlantic Ridge
Agulhas Basin
5000
Argentine Basin
Atlantic-Indian Ridge
Atlantic-Indian Basin
6000
Scotia Sea
5000
Crozet Basin
Scotia Ridge
Drake Passage
Weddell Sea
Kerguelen-Gaussberg Ridge
4000
5000
Southeast Pacific Basin
90°
90°
South Indian Basin
Ross Sea
Southeast Indian Ridge
70°
East Pacific Rise
4000
Cordillera
4000
60°
South Australian Basin
5000
Southwest Pacific Basin
50°
Tasman Basin
W 180° E

Fig. 2. Bottom topography of the Antarctic Ocean. The depths shown are in meters. 1 m = 3.3 ft.

International Southern Ocean Studies. The structure and volume transport of the Antarctic Circumpolar Current are central to the understanding of the general circulation of the Antarctic Ocean. Estimates of the total volume transport through the Drake Passage based on observations prior to 1970 varied from zero to 2.4×10^8 m^3/s (8.5×10^9 ft^3/s). To resolve this controversy and to study the dynamics and structure of the Antarctic Circumpolar Current, the International Southern Ocean Studies program was organized. Major emphasis was placed on monitoring the volume transport, determining the thermohaline and chemical properties of the waters passing through the Drake Passage, and investigating the nature and magnitude of meridional exchanges of heat and salt. As part of the 6-year program, 65 current meter moorings were maintained, and more than 600 hydrographic stations were taken during some 15 oceanographic cruises.

Transport through Drake Passage. The volume transport of an oceanic current can be decomposed into a baroclinic transport and a barotropic transport. The baroclinic transport arises from gradients in the horizontal and vertical density or mass field. The barotropic transport is due to a uniform pressure field and gives rise to a velocity which does not vary with depth.

The density of seawater is a function of pressure, temperature, and salinity and can be obtained from routine hydrographic measurements. Thus, from a series of hydrographic stations it is possible to determine the mass field and calculate the baroclinic transport. To determine the barotropic velocity and the total velocity profile, it is necessary to obtain the total velocity at least at one depth (with, for example, a current meter). By comparing this velocity with the baroclinic velocity, it is possible to determine the barotropic component and the total transport. The baroclinic transport through the Drake Passage between the free surface and the depth at 3000 decibars or 30 megapascals (approximately 3 km or 1.8 mi) is relatively constant (9.5×10^7 m^3/s or 3.4×10^9 ft^3/s $\pm 15\%$).

Though the magnitude of the current is small, there is little attenuation of flow with depth, which results in the great volume transport of the Antarctic Circumpolar Current. ISOS data show the Circumpolar Current carries about 1.25×10^8 m^3/s (4.4×10^9 ft^3/s) of ocean water through the Drake Passage. About 1×10^8 m^3/s is due to the baroclinic pressure field, the rest being barotropic. Variations from the mean of 8×10^7 m^3/s occur, closely correlated to variations in the wind field, which drives the ocean circulation.

Antarctic Circumpolar Current. Constant-density surfaces in the Drake Passage do not rise uniformly toward the south, but shoal in a series of distinct zones of large horizontal density gradients. These regions are vertically aligned and appear as a series of steplike structures, approximately 50 km (30 mi) wide (**Fig. 3**). Associated with these regions of large horizontal density gradient are high-velocity cores where the Antarctic Circumpolar Current attains surface velocities of 40 cm/s (16 in./s).

In the upper waters, these horizontal density discontinuities are associated with water-mass boundaries and in the deep waters with steplike rises of the core of Circumpolar Deep Waters. The polar front zone, the transition region between Antarctic and Subantarctic Surface Waters, is confined to the region between the two central density steps and is approximately 200 km (120 mi) wide. Small-scale mixing processes cause the Antarctic and Subantarctic

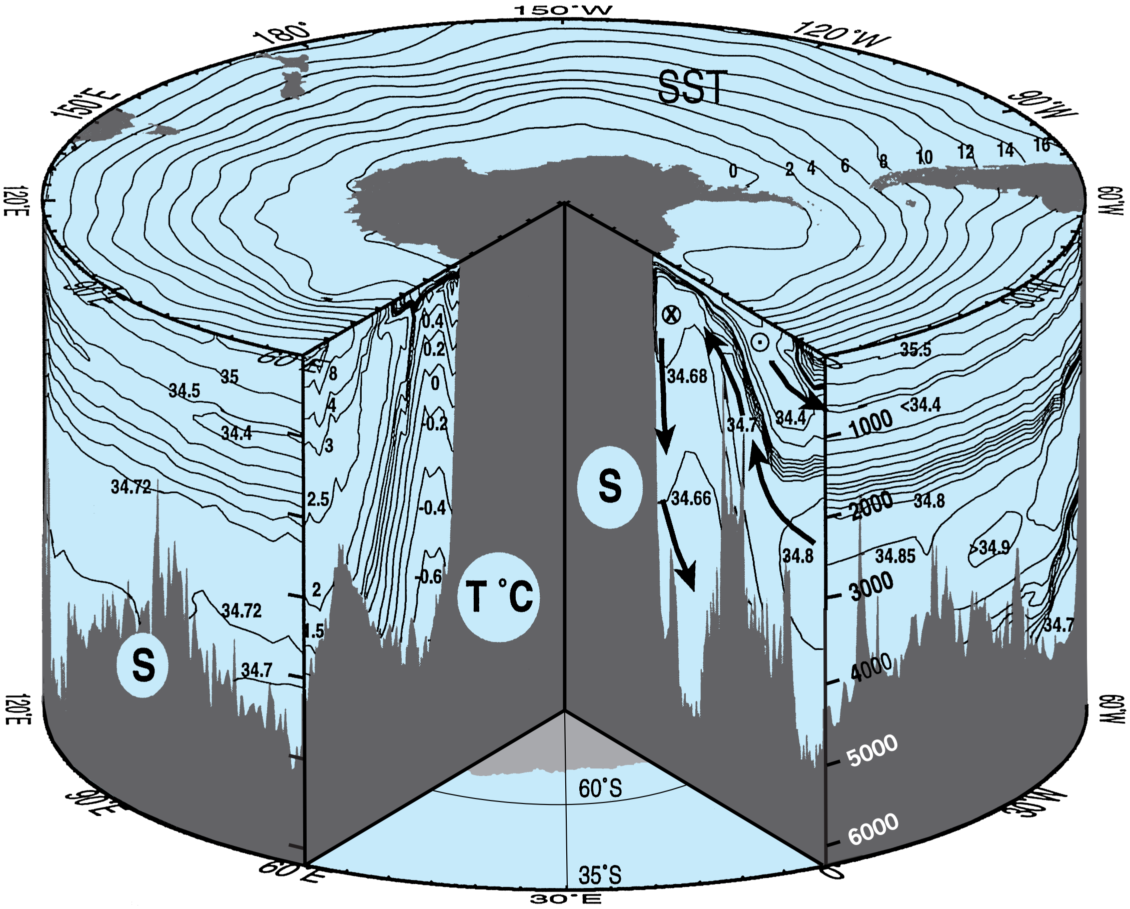


Fig. 3. Antarctic Ocean temperature and salinity. Isotherms of the annual average sea surface temperature (SST in °C) are shown on the plane of the sea surface. Winter SST adjacent to Antarctica is at the freezing point of seawater, −1.9°C. Sea surface temperature increases with distance from Antarctica. The core of the eastward flowing Antarctic Circumpolar Current and associated polar front occurs near the 4°C isotherm. The right-hand plane of the slice shows the salinity (S) values as a function of latitude (the 35°S and 60°S latitudinal circles are shown along the floor of the figure) and depth (shown in thousands, in meters), values in approximately parts per thousand of dissolved sea salt ($^0/_{00}$). These data are derived from the oceanographic observations along the Greenwich Meridian shown at the floor of the figure. (*After A. Gordon, The Southern Ocean, J. Mar. Educ., 15*(*3*)*, 1999*)

Surface Waters to interleave with vertical scales of order 100 m (330 ft) and less. However, these processes are relatively ineffective at transporting heat and salt across the zone.

Meanders, rings, and meridional heat and salt fluxes. The meridional transfer of heat and salt across the polar front zone is apparently accomplished by large-scale processes associated with instabilities of the Antarctic Circumpolar Current. Although the transport of the Antarctic Circumpolar Current is large and primarily toward the east through the Drake Passage, there are large disturbances called waves or meanders which give rise to north-south velocities. The meanders involve the entire water column, and have 250-km (160-mi) wavelengths and eastward phase velocities of 10–15 cm/s (4–6 in./s). These meanders are attributed to baroclinic and barotropic instabilities. Because the meanders transport colder waters northward and warm waters southward, they decrease the potential energy (baroclinic instability) associated with the inclined density surfaces and give rise to a poleward heat flux. The calculated poleward heat flux for the Drake Passage is larger than the required circumpolar averaged heat flux needed to balance heat lost by the Antarctic Ocean to the atmosphere.

Occasionally, the waves or meanders in the Antarctic Circumpolar Current become unstable, growing in time until the meander pinches off, a ring is formed, and the current and front are reestablished with a primarily zonal orientation. (Rings are commonly observed near other major oceanic current systems, such as the Gulf Stream and Kuroshio.) When a northward meander pinches off, colder Antarctic waters become enclosed by Subantarctic waters, giving rise to a cold-core clockwise-rotating cyclonic ring. Similarly, a southward meander can pinch off to form a warm-core ring (**Fig. 4**). In the Drake Passage the existence of such rings is well documented, having been photographed from satellites, studied from ships, and sensed with current meters. In other regions of the Antarctic Circumpolar Current, the existence of rings has also been inferred from hydrographic and current meter measurements.

The rings observed in the Antarctic Circumpolar

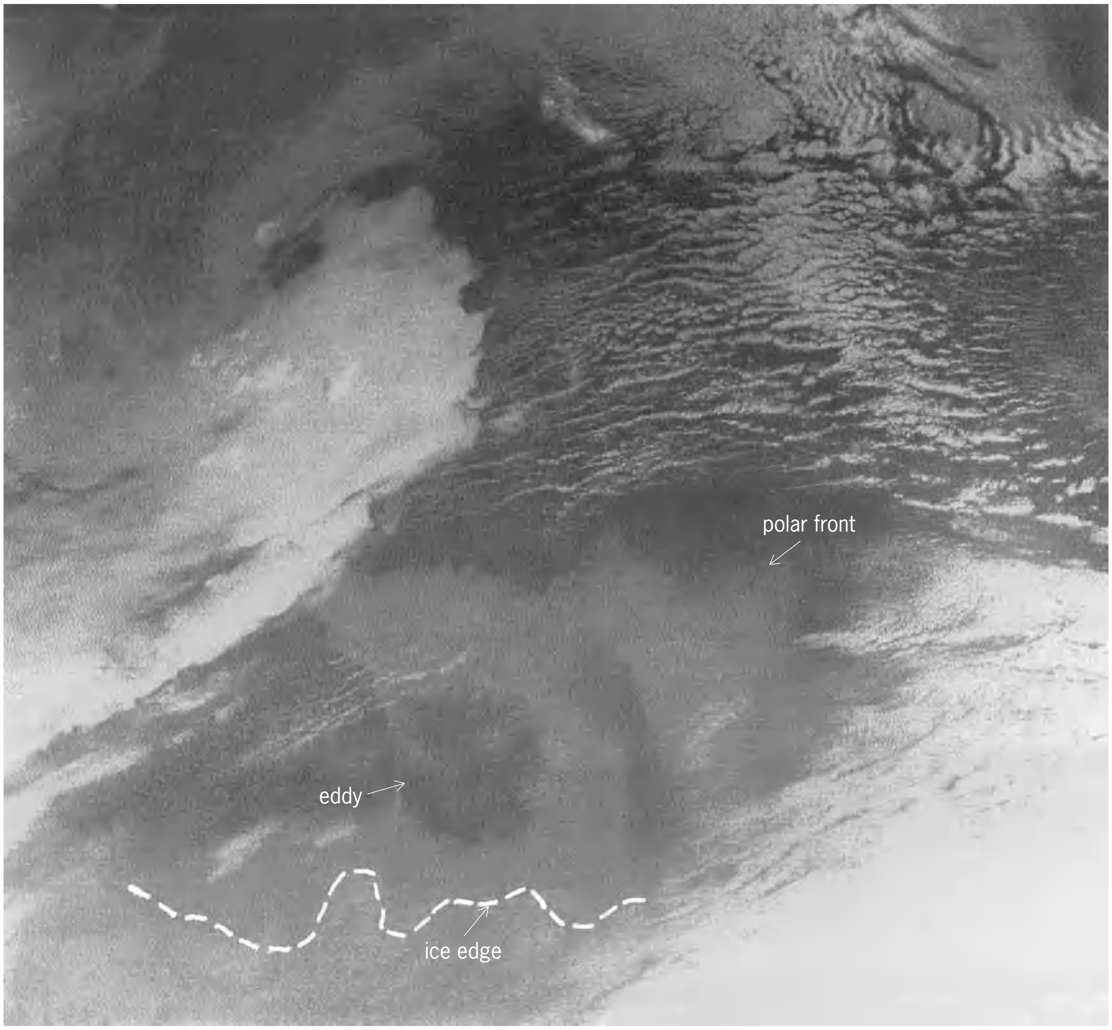


Fig. 4. Thermal infrared image from a polar orbiting satellite over the Drake Passage. The polar front, a warm–core ring (eddy), and the edge of the pack ice are indicated. (*From R. Legeckis, Oceanic polar front in the Drake Passage: Satellite observations during 1976, Deep–Sea Res., 34:701-704, 1977*)

Current are typically 3 km (1.8 mi) deep and 80–120 km (50–74 mi) in diameter, involving a volume of water equal to about one-tenth the water in all the world's fresh-water lakes. The average temperature difference between the waters inside and outside a cold-core ring is 0.25°C (0.45°F). Thus the northward ejection of a cold-core ring represents a significant heat transfer. The formation of 200 such rings per year in the Antarctic Circumpolar Current would account for one-half of the heat lost by the ocean to the atmosphere south of the polar front.

Between the Antarctic Circumpolar Current and Antarctica, the baroclinic velocity is very small. Characteristic velocities of less than 5 cm/s (2 in./s) are typical. However, the deep-reaching nature of these currents yields significant volume transports. A dominant circulation feature is the clockwise-flowing Weddell Gyre, which transfers (7.6×10^7 m^3/s; 2.7×10^9 ft^3/s) water between the Antarctic Circumpolar Current and the coastal East Wind Drift. The northward-flowing limb of the Weddell Gyre, just east of the Antarctic Peninsula, meets the Pacific water carried into the Atlantic via Drake Passage, forming a strong frontal zone over the Scotia Ridge (Fig. 2), called the Weddell-Scotia Confluence.

A smaller clockwise-flowing gyre exists within the Ross Sea, and possibly east of the Kerguelen-Gaussberg Ridge (Fig. 2). These circulation cells, with the Weddell Gyre, exchange water between the circumpolar belt and the Antarctic coastal region. *See* OCEAN CIRCULATION.

The nonzonal character of the flow is associated with irregularities in the bottom topography. On approaching a ridge, the current accelerates and turns northward. It becomes more diffuse and turns southward when approaching a basin.

Water masses. The seawater comprising the Antarctic Ocean can be conveniently divided into several water masses. The criteria used for this division are based on the temperature and salinity (T/S relation) of the water. Each water mass will fall in a different region of the diagram (**Fig. 5**). In schematic form, Fig. 3 shows the positions of these water masses and the average meridional circulation of the Antarctic Ocean. Within each of the water masses is an associated core layer. This layer contains the most undiluted water characterizing the water mass. It is observed as an extreme (maximum or minimum) in temperature, salinity, or dissolved-oxygen concentration.

As observed in the T/S diagram, the coldest, freshest water is the Antarctic Surface Water. The cooling results from the loss of heat to the atmosphere, and the low salinity is due to the excess of precipitation over evaporation. The heat and salt needed to maintain this water at a constant temperature and salinity over a period of years are supplied by the wind-induced upwelling at a rate of 4×10^7 m^3/s (1.4×10^9 ft^3/s) of the southward-flowing deep water discussed above. This water mass is called the Circumpolar Deep Water.

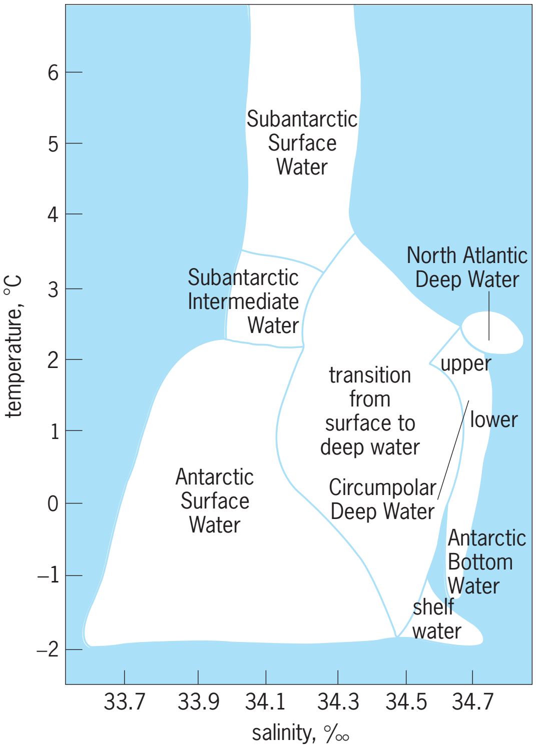


Fig. 5. Generalized diagram of the several water masses of the Antarctic Ocean; division of the water masses is based on the temperature–salinity relationship. $^\circ F = (^\circ C \times 1.8) + 32^\circ$.

The Antarctic Surface Water is bounded to the north by thick, homogeneous Subantarctic Surface Water, which is warmer and slightly more saline than the Antarctic Surface Water. A polar front zone separates these water masses (Fig. 1).

The polar front zone is characterized by a large surface-temperature gradient. Its position varies in a series of meanders or waves within a 3–4° latitude band. It is also called the Antarctic Convergence, because there is occasionally evidence of a convergence of the two surface waters. During convergence, the mixture of both surface waters sinks and flows northward as Subantarctic Intermediate Water. However, later work suggests that this does not always occur; at times, signs of divergences are found, or signs of neither convergence nor divergence but rather a simple meeting of two water masses. The direct contribution of winter-cooled Subantarctic Surface Water to Subantarctic Intermediate Water, with a significant component from the Antarctic Surface Water, may also occur.

In contrast to the Arctic, the Antarctic Ocean is not landlocked, but then it is not exactly fully open to the lower warmer latitudes, as the Antarctic Circumpolar Current surrounding Antarctica effectively blocks the passage of warmer surface waters from lower latitudes into high southern latitudes. Antarctica's glacial state of the last tens of millions of years is related to the establishment of the deep circumpolar belt and the development of the Antarctic Circumpolar Current. Even at depth the Antarctic Circumpolar Current tends to hem in southern polar waters, but there are a few deep cracks in the ridge system

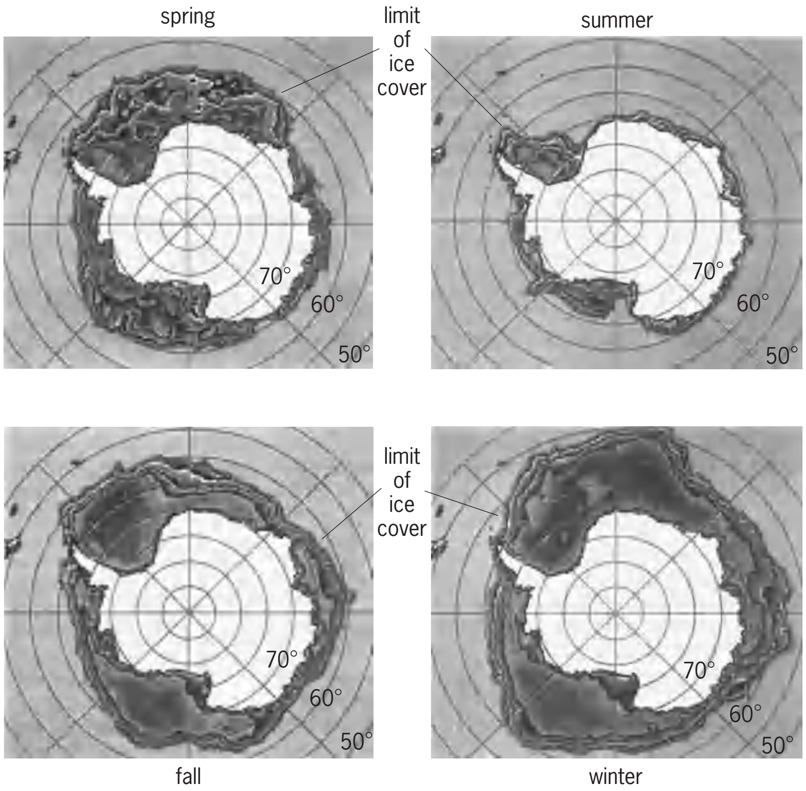


Fig. 6. Ice cover limits in the Antarctic Ocean at various seasons.

that surrounds Antarctica, allowing for meridional exchange of deep and bottom water.

Along its 21,000 km (13,020 mi) path the Antarctic Circumpolar Current transports 130×10^6 m^3 (4.6×10^9 ft^3) of seawater each second toward the east, making it the strongest of the ocean's currents. The polar front within the center of the Antarctic Circumpolar Current separates the very cold polar surface waters to the south from the somewhat warmer waters to the north. The Antarctic Circumpolar Current path and associated polar front position are aligned with the position of the maximum westerly winds, but guided to a significant degree by bottom topography, following along the northern flank of a series submarine ridges. The Antarctic Circumpolar Current and polar front reach south of 60°S, near New Zealand and near 48°S in the South Atlantic. As the ocean surface temperature pattern responds to the circulation pattern and the atmosphere responds to the sea surface temperature, there is the surprising result that the bottom topography influences the position of the maximum westerly winds. *See* SEAWATER.

Deep relatively saline water (>34.7‰) spreads poleward and upwells toward the sea surface (Fig. 3). It is balanced by northward flow of lower-salinity waters (<34.4‰) near 1 km (0.6 mi) [Antarctic Intermediate Water] and by sinking of slightly lower-salinity water along the continental slope of Antarctica. This process (saltier water in, fresher water out) removes the slight excess of regional precipitation from the Southern Ocean. The temperature is based on data collected at the same points as used for the salinity section. Shallowing of isotherms is evident as the deep water rises up toward the sea surface. There it is cooled and sinks, flooding the bottom layers with water of less than 0°C (32°F). This cold bottom water spreads well into the global ocean (Antarctic Bottom Water). The low salinity (Antarctic Intermediate Water) is shown as less than 34.45 band near 1 km (0.6 mi). More saline deep water spreading southward is seen near the 4 km (2.5 mi) depth.

At many sites along the continental margin of Antarctica, notably within the Weddell Sea (the extreme southern Atlantic Ocean), deep-reaching convective plumes of very cold surface water descend over the continental slope into the deep ocean. This Antarctic Bottom Water, high in dissolved oxygen, ventilates the bottom-layer adjacent ocean and eventually cools the lower 2 km (1.2 mi) of the global ocean. The transport of bottom water of less than −0.7°C (30.7°F) emanating from the Weddell Sea is estimated to vary from around 2 to 4×10^6 m^3/s (92 to 184×10^6 ft^3/s). Perhaps an equal amount descends at other sites around Antarctica.

Sea ice. Naturally the extreme cold of the polar regions causes an extensive ice field to form over the southern regions of the Antarctic Ocean. The extent of the ice is seasonal (**Fig. 6**) in that during the October-to-March period the area decreases, and it increases during the remaining months. The seasonal difference in the volume of sea ice is estimated as 2.3×10^{19} grams (8.1×10^{17} oz). Satellite photographs reveal that the sea ice field is not uniform, but has many large polynyas. The sea ice plays an important role in the heat balance since it reflects much more solar radiation (and therefore heat) into space than would be the case for a water surface. The polynyas would therefore be of special interest in radiation and heat-balance studies. In addition to the ice formed at sea, the ice calving at the coast of Antarctica introduces icebergs into the ocean at a rate of approximately 1×10^{18} g/year (3.5×10^{12} oz/year). *See* HEAT BALANCE, TERRESTRIAL ATMOSPHERIC; ICEBERG; SEA ICE.

In winter, a thin veneer of frozen seawater or sea ice stretches from Antarctica northward, reaching half the distance to the Antarctic Circumpolar Current. Antarctic Ocean sea ice is an important part of the global climate system, and changes in its behavior have important feedbacks to the global climate. Scientists recognize that accurate predictions of future climate requires that the ocean and atmosphere forces that govern Antarctic Ocean sea ice be understood, and that they be properly simulated in numerical global climate models.

South of the Antarctic Circumpolar Current, there is a significant transfer of the warm deep water into the surface layer of the ocean. This heat limits the thickness of the winter sea ice cover to between 0.5 and 1 m (1.6 and 3.3 ft), in sharp contrast to Arctic where upwelling of relatively warm water is small, and the sea ice is greater than 1 m (3.3 ft) thick. Upwelling of the warm deep water contributes on average 40 W/m^2 into the atmosphere during the winter months, attaining values well above 100 W/m^2 during storm periods. As the heat is transferred to the atmosphere, cold surface water is produced. It is this water that sinks along the Antarctic margins.

Glacial (fresh-water) ice and the ocean meet along the shores of Antarctica. This occurs not only at the northern face of the ice sheet but also at hundreds of meters depth along the bases of floating ice shelves. Ocean-glacial ice interaction is believed to be a major factor in controlling Antarctica's glacial ice mass balance and stability.

During the austral winters of 1974 to 1976, near the Greenwich Meridian and 66°S in the vicinity of a seamount called Maud Rise, the ice displayed strange behavior, which has not been repeated since. A large region normally covered by sea ice in winter remained ice-free, though it was surrounded by sea ice. This remarkable climate anomaly is referred to as the Weddell Polynya.

Normally the waters over the deep regions of the Weddell Sea remain stratified during the winter, with the thick relatively warm (>0°C; 32°F) saline deep water, separated from cold (near freezing, −1.85°C; 28.7°F), fresher, thin surface layer, only slightly less dense than that of the deep water. Upwelling of the warm deep water provides heat to the surface layer and atmosphere. As long as there is enough fresh water, including that stored within the sea ice, within the surface layer to dilute and make more buoyant the cooled deep water, this configuration remains in place. The immense storage of heat within deep water is sufficient to easily remove the winter sea ice cover if it were allowed to more freely and reach the sea surface. Should upwelling of deep water somehow overwhelm the surface-layer fresh-water storage, the system would abruptly flip into a convective state. This is exactly what happened during the Weddell Polynya event. Ocean cooling to nearly 3 km (1.8 mi) was observed, indicative of an average heat flux into the winter atmosphere over the polynya of 135 W/m^2.

Sediments. The composition of the sediments are influenced by two factors: the high biological productivity in the nutrient-rich euphotic zone, and the input of glacial debris carried seaward by icebergs. The siliceous ooze (diatoms and radiolaria) is found roughly between the polar front zone and the limit of the pack ice. To the north, the siliceous ooze is replaced by calcareous ooze and red clay in deeper basins. Within the pack ice field, sediments are of glacial marine type (**Fig. 7**).

The differences in concentration of diatom and radiolaria species with depth in the ooze indicate past minor climatic changes of the Antarctic region. A major transition in sediment types is found to have occurred 1.6 × 10^6 years ago when the sediments were red clay, indicating lower productivity. This suggests that the present Antarctic circulation has existed for at least the past 1.6 × 10^6 years. *See* MARINE SEDIMENTS.

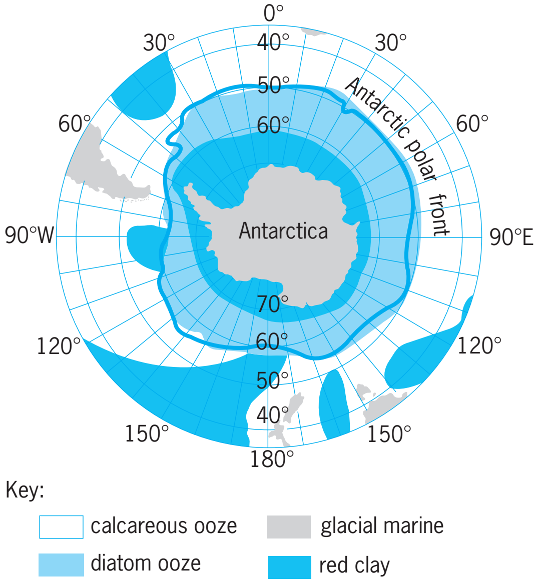


Fig. 7. Generalized distribution of the four sediment types which are found around Antarctica. (*After J. D. Hays, Quaternary sediments of the Antarctic Ocean, Prog. Oceanogr., 4:117–131, 1967*)

Biology. Antarctic Ocean marine mammal populations of whales and seals were the targets of much exploitation in the early twentieth century. Many species were nearly depleted. The marine mammals are sustained by high productivity of plankton within the Antarctic Ocean, particularly associated with frontal zones and the diverse and complex ecosystem supported at the northern fringes of the seasonal sea ice. The estimated initial stock of 884,000 metric tons (975,000 tons) of baleen whales has diminished to 307,000 metric tons (338,000 tons), mostly among the larger (more commercial) whales.

There has been renewed interest in the living marine resources of the Antarctic Ocean which were once "whale food," particularly in the inch-long shrimplike Antarctic krill, *Euphausia superba*. Krill form large schools or swarms with densities of up to 60,000 individuals per cubic meter (1700 individuals per cubic foot) of ocean, extending from ten to hundreds of meters with perhaps giant schools of over 1 km (0.6 mi) in horizontal dimension. Significant krill populations exist near the boundaries of the Weddell Gyre, north of the Ross Sea, and over the continental margins of Antarctica at a number of sites. Estimates of the total annual sustainable yield of krill (9 × 10^6 metric tons) is nearly double the total world fish and shellfish catch.

Many nations have exploited this source of protein in the Antarctic Ocean. Because of the remoteness of the Antarctic Ocean and the seasonal sea ice cover, the full life cycle, feeding habits, and dynamics of the swarming activity of krill are not known; management of this living resource requires such knowledge. An international program called Biological Investigations of Marine Antarctic Systems and Stocks (BIOMASS) was developed in the early 1980s to investigate krill and the population dynamics, food chain, and behavior of a number of elements in Antarctic Ocean living resources, notably whales, seals, birds, fish, and squid. *See* MARINE FISHERIES.

Arnold L. Gordon

Bibliography. D. T. Georgi and J. M. Toole, The Antarctic Circumpolar Current and the oceanic heat and freshwater budgets, *J. Mar. Res.*, 40(supp.):183–917, 1982; A. L. Gordon, Polar oceanography, *Rev. Geophys. Space Phys.*, 12(5):1124–1131, 1983; A. L. Gordon, Weddell Deep Water variability, *J. Mar. Res.*, 40(supp.):199–217, 1982; A. L. Gordon and T. N. Baker, *Southern Ocean Atlas: Objective Contouring and Grid Point Data Set*, 1982; A. L. Gordon and E. M. Molinelli, *Southern Ocean Atlas: Thermohaline-Chemical Distributions and the Atlas Data Set*, 1982; W. D. Nowlin, Jr., and M. Clifford, The kinematic and thermohaline zonation of the Antarctic Circumpolar Current at Drake Passage, *J. Mar. Res.*, 40(supp.):481–507, 1982; D. Walton, *Antarctic Science*, 1987.

Antarctica

Antarctica is the coldest, windiest, and driest continent. The lowest temperature ever measured on Earth, −89.2°C (−128.5°F), was recorded at the Russian Antarctic station of Vostok in July 1983. Katabatic (cold, gravitational) winds with velocities up to 50 km/h (30 mi/h) sweep down to the coast and occasionally turn into blizzards with 150 km/h (nearly 100 mi/h) wind velocities. Antarctica's interior is a cold desert averaging annually only a few centimeters of water-equivalent precipitation, while the coastal areas average 30 cm (12 in.).

Antarctica's area is about 14 million square kilometers (5.4 million square miles), which is larger than the contiguous 48 United States and Mexico together (**Fig. 1**). Antarctica ranks fifth in size among the continents, and is followed by Europe and Australia. About 98% of it is buried under a thick ice sheet, which in places is 4 km (13,000 ft) thick, making it the highest continent, with an average elevation of over 2 km (6500 ft).

Most of Antarctica is covered by ice, and some mountains rise more than 3 km (almost 10,000 ft) above the ice sheet. The largest range is the Transantarctic Mountains (**Fig. 2**) separating East from West Antarctica. The highest peak is Mount Vinson, 5140 m (16,850 ft), in the Ellsworth Mountains. Other ranges, such as the Gamburtsev Mountains in East Antarctica, are completely buried, but isolated peaks called nunataks frequently thrust through the ice around the coast.

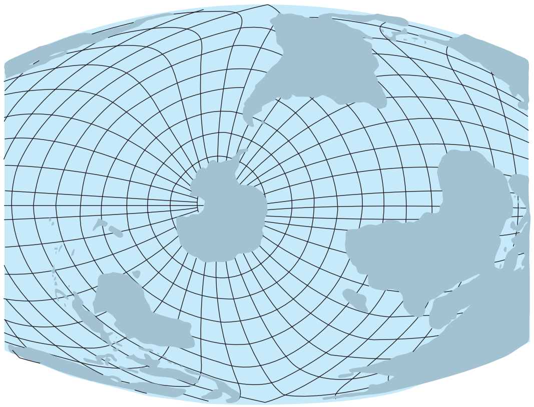


Fig. 1. Equal-area projection showing Antarctica's size in relation to that of the nearest continents.

The Antarctic continent is surrounded by a large belt of sea ice. At its maximum annual extent in September, sea ice covers an area of over 18 million square kilometers (7 million square miles); however, it melts during the summer season to a minimum extent of about 2–3 million square kilometers (1 million square miles). The ice reaches to 52°S latitude in some areas and has an average thickness of about 1.5 m (5 ft). In summer, it can generally be found only close to the coast and in the Weddell Sea.

The waters of the Southern Ocean are considered Antarctic waters south of the Antarctic Convergence. This convergence marks a sharp boundary between cold Antarctic water and warmer water to the north, and it lies roughly between 50 and 60°S latitude. Inside or close to the Antarctic Convergence lie the subantarctic islands. South of the Antarctic Convergence there are numerous icebergs; however, icebergs have also been sighted as far north as 45°S latitude. *See* ANTARCTIC OCEAN; ICEBERG.

Discovery. It is unclear who first sighted the Antarctic continent. The English navigator James Cook crossed the Antarctic Circle for the first time in his two great voyages between 1772 and 1776 and circumnavigated Antarctica, but he failed to sight the continent. Thaddeus von Bellingshausen, sent by the Russian czar to be the first to circumnavigate Antarctica since Cook, sighted land at 70°S latitude, 2°W longitude, in 1820 but did not clearly recognize the ice field he saw as land. A few days later, the Englishman Edward Bransfield sighted the Antarctic Peninsula, but his log books are now lost, making it difficult to substantiate any claims of discovery. The American sealer Nathanial Palmer sighted land almost a year earlier in the same region that Bransfield saw, but his log entries are vague.

Ice sheet. The Antarctic ice sheet is the largest remnant of previous ice age glaciations. It has probably been in place for the last 20 million years and perhaps up to 50 million years. It is the largest reservoir of fresh water on Earth, with a volume of about 25 million cubic kilometers (6 million cubic miles). If melted, Antarctic ice would raise the level of the world's oceans by about 60 m (200 ft). The ice load has depressed the continent by about 600 m (2000 ft) on average, and if it were removed, Antarctica would appear as an archipelago of scattered, mountainous islands. Glaciers flow out from this ice sheet and feed into floating ice shelves along 30% of the Antarctic coastline. The two biggest ice shelves are the Ross and Filchner-Ronne. These shelves may calve off numerous large tabular icebergs, with thicknesses of several hundred meters, towering as high as 70–80 m (250 ft) above the sea surface. The largest berg ever sighted in Antarctic waters was 335 km (208 mi) long and 97 km (60 mi) wide, about the size of Belgium. *See* GLACIOLOGY.

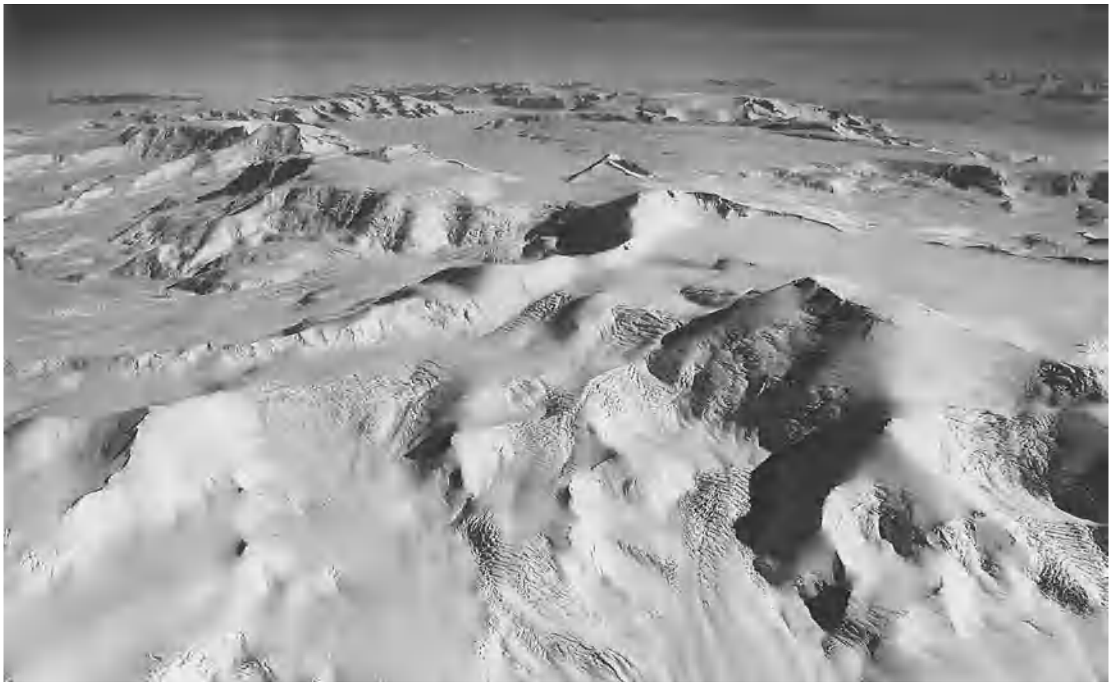


Fig. 2. Aerial view of the Transarctic Mountains. (*U.S. Navy photograph*)

Life on land. Year-round life on land in Antarctica is sparse and primitive. North of the Antarctic Peninsula, a complete cover of vegetation, including moss carpets and only two species of native vascular plants, may occur in some places. For the rest of Antarctica, only lichen, patches of algae in melting snow, and occasional microorganisms occur. In summer, however, numerous migrating birds nest and breed in rocks and cliffs on the continental margins, to disappear north again at the beginning of winter. South of the Antarctic Convergence, 43 species of flying birds breed annually. They include petrels, skuas, terns, cormorants, and gulls. Several species of land birds occur on the subantarctic islands. The largest and best-known of the Antarctic petrels are the albatrosses, which breed in tussock grass on islands north of the pack ice. With a wing span of 3 m (10 ft), they roam freely over the westerly wind belt of the Southern Ocean. *See* PROCELLARIIFORMES.

Ocean life. In contrast to the generally sparse and primitive lifeforms on the Antarctic continent, the Southern Ocean teems with life. Nutrient- and oxygen-rich, it is one of the most productive oceans and supports a rich and diverse food web. The base of this food web is phytoplankton (tiny sea plants), which support a large crop of zooplankton (small sea animals), about half of which are the 3–6-cm-long (1–2-in.) shrimplike krill. Krill are the primary food of baleen whales and crabeater seals, and they are also eaten in substantial quantities by penguins, flying birds, and fish. *See* PHYTOPLANKTON.

Seven species of whales frequent Antarctic waters, including the blue, fin, humpback, and right whales, all of which are protected by international convention, and the sei, sperm, and minke whales. There are also six species of seals: crabeater, Weddell, elephant, leopard, Ross, and fur seals. Some species were severely diminished in numbers by sealing but are now recovering. There are also 21 species of penguins scattered throughout the Southern Hemisphere. The most species and the largest colonies of penguins, some of them containing several hundred thousand birds, occur on the subantarctic islands. Only two species, Emperor and Adélie penguins, are found on the Antarctic continent. *See* CETACEA; SPHENISCIFORMES.

Aurora. Modern science in Antarctica encompasses many areas of research. A fascinating phenomenon studied in the polar regions is the aurora. Auroras (northern or southern lights) are caused by subatomic particles (electrons and protons) from the Sun colliding with molecules in the upper atmosphere and producing light. The typical form of an aurora is curtain- and draperylike, moving and rippling rapidly across the sky. Bands, arcs, rays, and spirals occur as auroras form. The color of the aurora is determined by the atoms and molecules present, violet for nitrogen and red and green for atomic oxygen, and the intensity is determined by the energy of the subatomic particles that strike them. Auroras interfere with communication systems, may cause power blackouts, and can perhaps even affect the climate. *See* AURORA.

Ozone hole. A discovery was made in 1983 that ozone (O_3), a gas that occurs in small quantities in the atmosphere, has been steadily decreasing over Antarctica. The reduction is greatest in October, when the polar upper atmosphere is coldest. Temporary ozone losses as high as 50% have been recorded, forming a hole of low ozone concentrations over Antarctica that is as large as the United States. In recent years, ozone concentrations have been lower, and the size of the hole has been larger and the hole has lasted longer. The explanations for this ozone hole are complex but heavily implicate chlorofluorocarbons (CFCs), the industrial materials used as coolants in refrigerators and air conditioners and for making plastic foams. Ozone, although present in very small quantities, effectively absorbs the ultraviolet radiation in sunlight that causes sunburn and skin cancer. Ultraviolet radiation is also reported to cause cataracts and to weaken the immune system.

Ice core records. Some of the most spectacular scientific results from Antarctica in recent years have come from deep ice cores, particularly from the Russian station Vostok (**Fig. 3**), where ice core records date back about 400,000 years. The method of reconstructing past climates is through analyses of isotope ratios, since the ratio of oxygen-18 to oxygen-16 in the ice depends on the air temperature at the time when the ice was deposited as snow on the surface. Another interesting result comes from the analyses of the trace gases contained in the tiny air bubbles in the ice. As the snow turns into ice, the air is trapped and can give an indication of the chemical composition of the atmosphere at that time. Pre-industrial revolution levels of carbon dioxide (CO_2), for example, have been found to be much lower (275 parts per million) than at present (360 ppm). *See* CLIMATE HISTORY; PALEOCLIMATOLOGY.

Gondwanaland. Up until about 140 million years ago, East Antarctica formed the central block of a giant supercontinent called Gondwanaland. When this continent broke up, the continents of Antarctica, Australia, Africa, and South America, as well as the Indian subcontinent, drifted to their present positions. In reconstructing Gondwanaland, the edges of the continental shelves are matched, and rocks and fossils on either side are compared. There are many examples supporting the idea that these continents were once joined together. One crucial line of evidence arises from the discovery of about 280-million-year-old tillites—consolidated ground moraines from a former widespread and broadly synchronous glaciation of the southern continents. *See* CONTINENTS, EVOLUTION OF.

International Geophysical Year. International scientific cooperation flourished during the International Geophysical Year (IGY) which took place from June 1957 to December 1958. It was the first major worldwide scientific effort that involved Antarctica. About 50 stations were maintained by 12 nations in Antarctica, with the United States establishing a station at the South Pole and the Soviet Union establishing

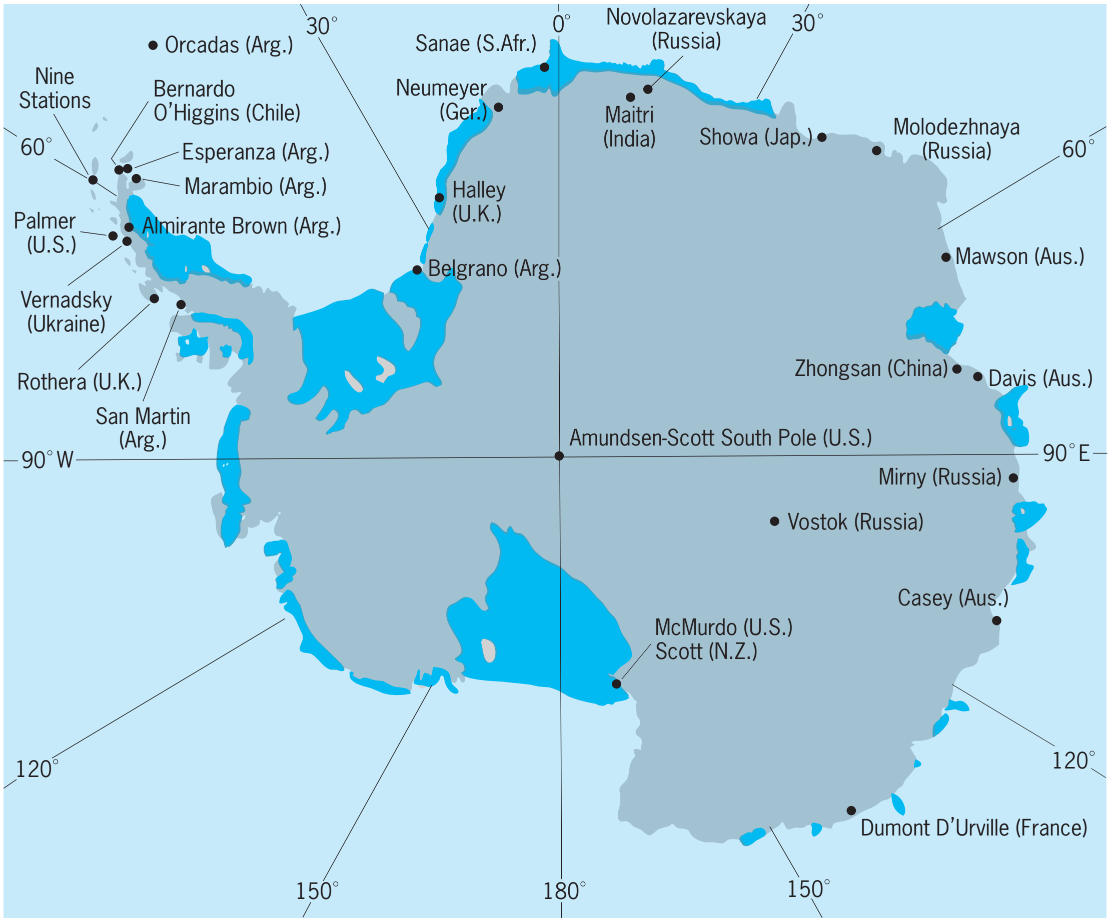

30°
0°
30°
60°
60°
90° W
90° E
120°
120°
150°
180°
150°
Orcadas (Arg.)
Sanae (S.Afr.)
Novolazarevskaya (Russia)
Nine Stations
Bernardo O'Higgins (Chile)
Neumeyer (Ger.)
Maitri (India)
Showa (Jap.)
Molodezhnaya (Russia)
Esperanza (Arg.)
Marambio (Arg.)
Halley (U.K.)
Palmer (U.S.)
Almirante Brown (Arg.)
Belgrano (Arg.)
Mawson (Aus.)
Vernadsky (Ukraine)
Rothera (U.K.)
San Martin (Arg.)
Zhongsan (China)
Davis (Aus.)
Amundsen-Scott South Pole (U.S.)
Mirny (Russia)
Vostok (Russia)
Casey (Aus.)
McMurdo (U.S.)
Scott (N.Z.)
Dumont D'Urville (France)

Fig. 3. Stations occupied during winter 2002. Gray areas show the continental ice sheet; color areas are ice shelves and glacier tongues. There are nine stations on King George Island: Jubany (Argentina), Commandante Ferraz (Brazil), Capitan Arturo Prat (Chile), Presidente Eduardo Frei (Chile), Escudero (Chile), Great Wall (China), King Sejong (Korea), Arctowski (Poland), and Bellinghausen (Russia).

one near the geomagnetic pole. Extensive traverses were conducted with large oversnow vehicles, and the Commonwealth Transantarctic Expedition completed the first crossing of Antarctica. The IGY activities in Antarctica made a significant contribution to the study of weather and climate, the upper atmosphere, and the antarctic ice sheet and underlying bedrock. The cooperation between scientists of all the participating nations produced a lasting impact.

Antarctic Treaty. The United States convened a conference in 1958 to discuss the future of Antarctica, and a treaty was signed on December 1, 1959, by the nations that had been active in Antarctica during the IGY: Argentina, Australia, Belgium, Chile, France, Japan, New Zealand, Norway, South Africa, the Soviet Union, the United Kingdom, and the United States. The treaty temporarily freezes all territorial claims and sets aside Antarctica for peaceful purposes only. In the meantime, 32 other nations have joined the initial 12 in signing the treaty. The overriding achievement of the Antarctic Treaty has been to bring peace and political stability to a major world region. Consultative meetings have established standards for environmental protection, sharing of research results, tourism, and protection of special areas. The Antarctic Treaty has an indefinite duration and continues to provide the framework for cooperation and harmony in Antarctica. Gunter Weller

Bibliography. D. J. Drewry, R. M. Laws, and J. A. Pyle (eds.), *Antarctica and Environmental Change*, Royal Society, Clarendon Press, Oxford, 1993; R. Fifield, *International Research in the Antarctic*, Oxford University Press, New York, 1987; D. W. H. Walton (ed.), *Antarctic Science*, Cambridge University Press, New York, 1987.

Antares

α Scorpii, a cool supergiant star of spectral type M1Ib, whose red color stands out in the midsummer sky. With an effective temperature of approximately 6000°F (3600 K), Antares resembles Betelgeuse, the brightest of the red supergiants, and would fill the solar system beyond the orbit of Mars if it replaced the Sun. Antares has a distance of 185 parsecs (600 light-years) from the Sun, and its angular diameter of about 0.045 arc-second has been measured by interferometric and lunar occultation methods. Red supergiants of this type originate as stars with mass at least 20 times that of the Sun. Such stars quickly evolve through successive stages of thermonuclear fusion of heavier and heavier elements in their cores with lighter elements undergoing fusion in concentric shells surrounding the hot core. Eventually, the core builds up with iron, the most tightly bound of all atomic nuclei. The fusion of iron nuclei absorbs rather than releases energy, and the supergiant star implodes in a type II supernova explosion. This fate is likely for Antares in less than a million years. Prior to this dramatic event, Antares will have shed up to 50% of its mass through a stellar wind of material blown away from the star into the surrounding interstellar medium. *See* BETELGEUSE; SCORPIUS; SPECTRAL TYPE; STELLAR EVOLUTION; SUPERGIANT STAR; SUPERNOVA.

Antares is a member of an association of young and primarily hot stars, and is gravitationally bound in a binary star system with a hot blue star of spectral type B3V. The two stars orbit each other with a period of about 900 years, from which their masses can be determined to be about 15 and 7 times that of the Sun for the red and blue components respectively. The interaction of the strong wind of matter from the supergiant with the radiation from the less massive, hot companion produces an unusual nebulosity surrounding the hot star, contributing emission lines to that star's spectrum. Antares is also a source of radio-wavelength emission arising separately from each component of the binary. *See* BINARY STAR; NEBULA; STAR. Harold A. McAlister

Bibliography. E. E. Bloemhof and R. M. Danen, Direct measurement of the inner radius of the dust shell around the cool supergiant star α Scorpii, *Astrophys. J.*, 440: L93–L96, 1995; A. Frankoi, D. Morrison, and S. C. Wolff, *Voyages Through the Universe*, 3d ed., Brooks/Cole, 2004; J. B. Kaler, *The Hundred Greatest Stars*, Copernicus Books, 2002; J. M. Pasachoff and A. Filippenko, *Astronomy in the New Millennium*, 3d ed., Brooks/Cole, 2007; G. L. Verschuur, Journey into the Galaxy, *Astronomy*, 21(1):32–39, January 1993.

Anteater

A name associated with several animals in four different orders of living mammals (see **table**). They are so named because they are insectivorous, having a diet of ants and termites which they detect mainly by smell. These animals provide a good example of the evolutionary concept of convergence, in which similar adaptations are made by organisms of different groups. The animal most frequently associated with this name is the giant anteater (*Myrmecophaga tridactyla*), a ground-dwelling member of the family Myrmecophagidae in the order Xenarthra (see **illustration**). Three arboreal species—the northern tamandua (*Tamandua mexicana*), southern tamandua (*T. tetradactyla*), and the silky, or pygmy, anteater (*Cyclopes didactylus*)—are also classified in this family. They are found in Central and South America from Mexico to northern Argentina. Although they share membership in Xenarthra with sloths and armadillos, anteaters are the only toothless members of the order. Tamanduas and silky anteaters do possess either distinct tooth sockets or vestiges, which indicate that the teeth have been lost in the course of phylogenetic development. All members of the order Xenarthra differ from all other mammals by having xenarthrous vertebrae, lumbar vertebrae that have additional articulations (xenarthrales) to provide reinforcement for digging.

Morphology and diet. Anteaters range from 155 mm, or 6 in. (silky anteater) to about 1200 mm or 4 ft (giant anteater) in head-plus-body length. The giant anteater has a dense, coarse, stiff gray coat with

Classification of animals commonly referred to as anteaters

Scientific name	Common name	Range
Order Monotremata		
Family Tachyglossidae		
Tachyglossus aculeatus	Short-nosed echidna or spiny anteater	Australia, Tasmania, New Guinea
Zaglossus bruijni	Long-nosed echidna	New Guinea
Order Dasyuromorphia		
Family Myrmecobidae		
Myrmecobus fasciatus	Numbat or banded anteater	Australia, New South Wales
Order Pholidota		
Family Manidae		
Manis pentadactyla	Chinese pangolin	Southeast Asia
M. crassicaudata	Indian pangolin	Southeast Asia
M. javanica	Malayan pangolin	Southeast Asia
M. tricuspis	Tree pangolin	Senegal, Kenya, Zambia
M. gigantea	Scaly anteater or giant pangolin	Senegal, Uganda, Angola
M. temmincki	Cape pangolin	Chad, Sudan, South Africa
M. tetradactyla	Long-tailed tree pangolin	Senegal, Uganda, Angola
Order Xenarthra		
Family Myrmecophagidae		
Myrmecophaga tridactyla	Giant anteater	Central and South America
Tamandua mexicana	Northern tamandua or lesser anteater	Mexico to Peru
T. tetradactyla	Southern tamandua or lesser anteater	South America
Cyclopes didactylus	Silky anteater	Central and South America

black and white shoulder stripes. Adults weigh 22–39 kg (48–86 lb), although animals in zoos may weigh as much as 60 kg (132 lb). The dense, short fur of tamanduas is tan to dark brown with patches of black or reddish brown on the trunk. Completely black or dark brown individuals are found in some regions of South America. They weigh 3.2–7 kg (7–15.5 lb). The silky anteater has soft, silky gray to yellowish-orange fur with a darker middorsal stripe. It weighs 375–410 g (13.2–14.4 oz).

Anteaters have elongated, tapered snouts and tubular mouths. The long, narrow, rounded tongues are longer than their heads. The tongue is covered in minute, posteriorly directed spines (filiform papillae) and is wet with saliva when the animal is feeding. It is not sticky as is sometimes reported. The retractor muscle of the tongue is highly efficient, with motion pictures showing 160 tongue strokes a minute by a giant anteater at a termite mound. Approximately 35,000 ants and/or termites are consumed daily. Their keen sense of smell allows anteaters to be selective in choosing which anthills or termite mounds are suitable for clawing open for food. Neither anthills nor termite mounds are destroyed during feeding, as was once believed. Rather, the anteater takes only a portion of the population, thus conserving its source of food. The stomach does not secrete hydrochloric acid as in most mammals, but depends on the formic acid of the ingested ants to assist with digestion.

Giant anteater (*Myrmecophaga tridactyla*). (*Photo © 2001 by John White*)

Tamanduas may occasionally eat small fruits, while silky anteaters supplement their diet with bees and wasps. The ears are short and rounded and the eyes are small. The muscular, prehensile tail and long claws (up to 400 mm or 16 in.) of the arboreal tamanduas and silky anteater are used to grip branches. The strong foreclaws in all four species are also used to open ant and termite nests and as defensive weapons. When walking, only the outer edge of the hand, with claws retracted, makes contact with the ground and only half the sole of the hind foot makes contact. The four toes of the hind foot of the silky anteater are joined by a callused sole whose surface broadens on the inside to make an elastic rounded cushion for grasping. The genus name *Cyclopes* (Greek *kyklos* = "circle"; Latin *pes* = "foot") alludes to this feature. When agitated, posterior glands of tamanduas produce a musky secretion that has a very disagreeable odor.

Ecology and distribution. Giant anteaters are diurnal and prefer grassland, swamps, and lowland tropical forest. They swim well, even crossing wide rivers. A study of giant anteaters in Brazil showed a population density of 1.2 individuals per square kilometer. Each giant anteater lives a solitary existence during most of the year. Contact between the sexes occurs only at the time of estrus. Adults maintain their own territories, which may overlap considerably. They sleep in a shady trench or depression on the ground. Giant anteaters have the lowest recorded body temperature for a placental mammal, 32.7°C (90.9°F). The species is classified as "vulnerable" by the International Conservation Union (IUCN) and is on Appendix 2 of the CITES.

Tamanduas are also diurnal and inhabit savannas, thorn scrub, and wet and dry forests, whereas silky anteaters inhabit tropical forests. The silky anteater is the most arboreal of the four species and will not voluntarily descend to the ground. This squirrel-sized, nocturnal species spends the daylight hours usually curled up in a ball in the fork or hollow of a kapok tree.

Reproduction and development. Giant anteaters mate in the fall and give birth to a single precocial young in the spring. Gestation ranges from about 150 to 190 days. A newborn weighs approximately 1.2 kg (2.5 lb). Immediately following its birth, the baby will climb onto its mother's back and be carried wherever she goes. Mothers may carry their young until they are half-grown, at which time they are almost as big as the mother. The young nurse for about six months and may remain with their mother for as long as two years. Little is known about reproduction in tamanduas and silky anteaters. Tamanduas usally give birth to a single young in the spring following a gestation period of 130–190 days. The young is carried on the mother's back and remains with its mother for about one year. It is known that both silky anteater parents care for their young by regurgitating crushed insects to feed the baby. *See* MONOTREMATA; PHOLIDOTA.

Donald W. Linzey

Bibliography. D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Antelope

The name given to a group of African and Asian hoofed ruminants of the order Artiodactyla. The 91 species in 31 genera are classified in five subfamilies (Bovinae, Cephalophinae, Hippotraginae, Antilopinae, and Caprinae) in the family Bovidae. This group includes the bongos and kudus (*Tragelaphus*), elands (*Taurotragus*), duikers (*Cephalophus* and *Sylvicapra*), waterbucks (*Kobus*), reedbucks (*Redunca*), roan and sable antelopes (*Hippotragus*), oryx (*Oryx*), wildebeests or gnus (*Connochaetes*), impalas (*Aepyceros*), blackbucks (*Antilope*), gerenuks (*Litocranios*), gazelles (*Gazella*), and springbucks (*Antidorcas*). One genus (*Pseudonovibos*) was described only on the basis of a series of 14 horns from Cambodia and Viet Nam; the living animal has never been seen by scientists.

Eland bull (*Taurotragus derbianus*).

Morphology. Antelopes vary in size from the royal antelope (*Neotragus pygmaeus*), which has a head and body length of about 500 mm (19–20 in.), a shoulder height of 250–305 mm (10–12 in.), and a weight of 2–3 kg (4–6.5 lb), to the giant eland (*Taurotragus derbianus*; see **illustration**) which may have a head and body length of 1800–3450 mm (70–134 in.), a shoulder height of 1000–1800 mm (39–70 in.), and a weight of 400–1000 kg (875–2200 lb). The royal antelope is the second smallest hoofed mammal in the world, after some individuals of Asiatic chevrotains or mouse deer (*Tragulus javanicus*) that inhabit southeast Asia. *See* CHEVROTAIN.

All adult male antelopes and most adult females possess unbranched horns composed of a bony core attached to the frontal bones of the skull and covered by a hard sheath of horny material. With the exception of the four-horned antelope (*Tetracerus*) of India and Nepal, all individuals possess a single pair of horns. The length and configuration of the horns varies considerably among the various species of antelopes. The dental formula is I 0/3, C 0/1, PM 3/3, M 3/3 × 2 for a total of 32 teeth. The lower incisors generally project forward and serve as cutting teeth, while the upper lip and tongue serve the function of upper incisors. These mammals feed by twisting grass, stems, and leaves around the tongue and cutting the vegetation off with the lower incisors. All ruminants have a four-compartment stomach and chew the cud. When food is initially swallowed, it goes into the largest of the four compartments (rumen) where it is acted upon by bacteria. It is later regurgitated while the animal is resting, chewed, and swallowed a second time for thorough digestion. The ulna and fibula are reduced, and the main foot bones are fused into a cannon bone. On each foot the third and fourth toes are well developed, whereas the second and fifth toes are reduced or absent. *See* DENTITION.

Reproduction and development. Reproduction may be seasonable in some species, whereas in others mating may occur throughout the year. Gestation periods range from 168 to 277 days. There is normally a single offspring, but twins are frequent in some species of gazelles. Sexual maturity is attained between 6 (dwarf antelope) and 48 (eland) months of age. Longevity may be over 23 years (eland).

Ecology. Antelopes are grazers and browsers. Depending on the species, antelopes may inhabit grasslands, brushy areas, wooded savannahs, open forests, and, in the case of the klipspringer (*Oreotragus*) and beira (*Dorcatragus*), even rocky hills, mountains, and cliffs. Most are gregarious, although some (*Oreotragus, Ourebia, Raphicerus, Neotragus, Madoqua*, and *Dorcatragus*) are usually solitary and territorial.

Excessive hunting, excessive overgrazing by domestic livestock, agricultural development, and other habitat modifications have adversely affected most populations of antelopes. The International Conservation Union (IUCN) has designated some subspecies and/or populations of most genera as critically endangered, endangered, threatened, conservation dependent, or vulnerable. *See* ARTIODACTYLA; MAMMALIA.

Donald W. Linzey

Bibliography. D. Macdonald, *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Antenna (electromagnetism)

The device that couples the transmitter or receiver network of a radio system to space. Radio waves are used to transmit signals from a source through space. The information is received at a destination which in some cases, such as radar, can be located at the transmitting source. Thus, antennas are used for both transmission and reception. *See* RADAR.

Size and configuration. To be highly efficient, an antenna must have dimensions that are comparable with the wavelength of the radiation of interest. At long wavelengths such as the part of the spectrum used in broadcasting (a frequency of 1 MHz corresponds to a free-space wavelength λ of 300 m), the requirement on size poses severe structural problems, and it is consequently necessary to use structures that are portions of a wavelength in size (such as 0.1 λ or 0.25 λ). Such antennas can be described as being little more than quasielectrostatic probes protruding from the Earth's surface. The characteristics of antenna properties can be expressed in terms of the ratio of the dimensions of the antenna to the wavelength of operation. Therefore, in describing antennas, the dimensions are given in wavelengths.

In order to control the spread of the energy, it is possible to combine antennas into arrays. As the wavelength gets shorter, it is possible to increase the size of the antenna relative to the wavelength; proportionately larger arrays are also possible, and techniques that are familiar in acoustics and optics can be employed (**Fig. 1**). For example, horns can be constructed with apertures that are large compared with the wavelength. The horn can be designed to make a gradual transition from the transmission line, usually in this case a single-conductor waveguide, to free space. The result is broadband impedance characteristics as well as directivity in the distribution of energy in space. Another technique is to use an elemental antenna such as a horn or dipole together with a reflector or lens. The elemental antenna is essentially a point source, and the elementary design problem is the optical one of taking the rays from a point source and converting them into a beam of parallel rays. Thus a radio searchlight is constructed by using a paraboloidal reflector or a lens. A very large scale structure of this basic form used as a receiving antenna (together with suitably designed receivers) serves as a radio telescope. Antennas used for communicating with space vehicles or satellites are generally large (compared to wavelength) structures as well. *See* RADIO TELESCOPE; SPACE COMMUNICATIONS; TRANSMISSION LINES; WAVEGUIDE.

Types of problems. Antenna problems are frequently formulated in one of the following forms. (1) Given a specific radiating system, it is required to find its directional characteristics, polarization of fields, gain, input impedance, and other important properties which need to be known in order that the communication system may be designed. (2) Given a directional characteristic, for example, a certain beam-width, polarization, and beam orientation, it is required to find or select a radiating system that will produce it. Certain other constraints such as impedance behavior and physical size will have to be met. This antenna problem may be classed as a synthesis problem.

Elemental dipole antenna. Radiation properties of many antennas may be obtained by assuming that the current distribution on the antenna consists of a superposition of current elements of appropriate magnitude and, where necessary, of appropriate phase. With knowledge of the fields produced by a current

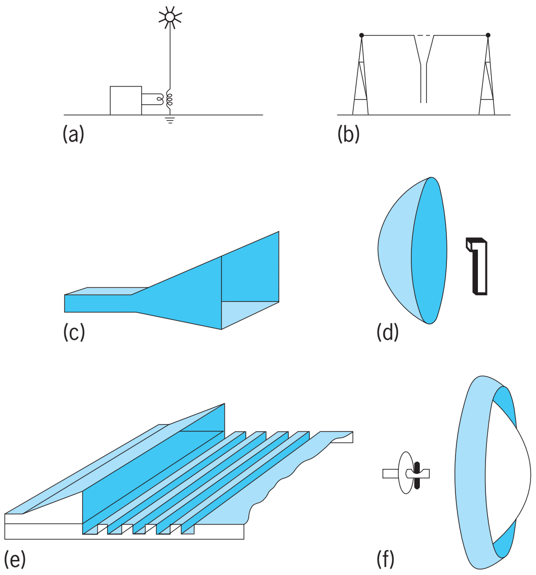


Fig. 1. Various types of antennas. (*a*) Top-loaded vertical mast; (*b*) center-fed horizontal antenna; (*c*) horn radiator; (*d*) paraboloidal reflector with a horn feed; (*e*) corrugated-surface wave system for end-fire radiation; (*f*) zoned dielectric lens with a dipole-disk feed. *(After D. J. Angelakos and T. E. Everhart, Microwave Communications, Krieger, 1983)*

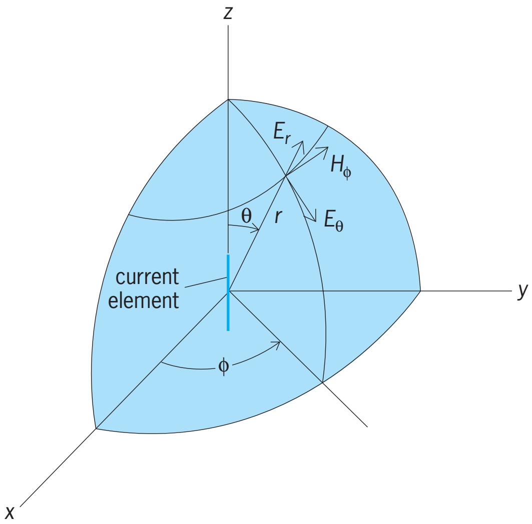


Fig. 2. Spherical coordinate system for an electric dipole.

element, by superposition it is then possible to obtain the composite fields of the antenna. Care must be taken to add the fields with proper consideration for their vector and phasor properties. An equivalent current source may also be defined for aperture-type antennas. For these reasons and because additional characteristics of antennas may be illustrated by the simple current element, the elemental dipole is a good starting point in the study of antennas.

For a differential current element of cosinusoidal time variation oriented along the z axis, the electric (E) and magnetic (H) field components, in a spherical coordinate system (**Fig. 2**), are given by Eqs. (1)–(3), where the differential current element

$$E_r = 60\beta^2 Idz\left[\frac{\cos(\omega t-\beta r)}{(\beta r)^2} + \frac{\sin(\omega t-\beta r)}{(\beta r)^3}\right]\cos\theta \tag{1}$$

$$E_\theta = 30\beta^2 Idz\left[\frac{-\sin(\omega t-\beta r)}{\beta r} + \frac{\cos(\omega t-\beta r)}{(\beta r)^2} + \frac{\sin(\omega t-\beta r)}{(\beta r)^3}\right]\sin\theta \tag{2}$$

$$H_\phi = \frac{\beta^2}{4\pi} Idz\left[\frac{-\sin(\omega t-\beta r)}{\beta r} + \frac{\cos(\omega t-\beta r)}{(\beta r)^2}\right]\sin\theta \tag{3}$$

is a current $I\cos(\omega t)$ in a very short piece of wire of length dz, r is the radial distance to the observation point, θ and ϕ are the angular coordinates, $\beta = 2\pi/\lambda$, $\omega = 2\pi f$, and λ and f are, respectively, the wavelength and frequency of the radio wave. *See* COORDINATE SYSTEMS; ELECTROMAGNETIC RADIATION; ELECTROMAGNETIC WAVE TRANSMISSION.

The product $\beta r = 2\pi r/\lambda$ represents the distance in terms of wavelengths. Hence for large values of r/λ ($\beta r \gg 1$) only the $1/(\beta r)$ terms are retained. The far-field components of significance are then given by Eqs. (4) and (5). The first part of the field expressions shows that the magnitude of the intensity depends on the ratio of the differential current element to the wavelength. The second part, in square brackets, is the variation of field intensity with angle and distance from the element. It represents the radiation pattern of the elemental antenna. The last term contains the wave phenomenon, a radio wave traveling from the current element outward in a radial direction (**Fig. 3**).

$$E_\theta = \frac{-60\pi Idz}{\lambda}\left[\frac{\sin\theta}{r}\right]\sin(\omega t-\beta r) \tag{4}$$

$$H_\phi = -\frac{Idz}{2\lambda}\left[\frac{\sin\theta}{r}\right]\sin(\omega t-\beta r) = \frac{E_\theta}{120\pi} \tag{5}$$

For a small magnetic dipole (loop), the far-field components are given by Eqs. (6) and (7), where dA is the differential area of a loop of current I. The radiation pattern of the magnetic dipole is the same as that

$$E_\phi = 120\pi^2\frac{IdA}{\lambda^2}\left[\frac{\sin\theta}{r}\right]\cos(\omega t-\beta r) \tag{6}$$

$$H_\theta = -\pi IdA\left[\frac{\sin\theta}{r}\right]\cos(\omega t-\beta r) = \frac{-E_\phi}{120\pi} \tag{7}$$

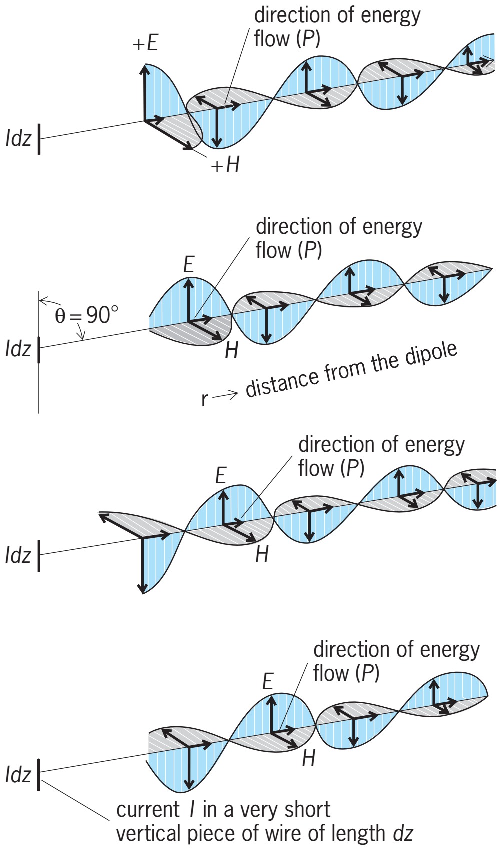

direction of energy flow (P)
+E
Idz
+H
direction of energy flow (P)
E
θ = 90°
Idz
H
r → distance from the dipole
direction of energy flow (P)
E
Idz
H
direction of energy flow (P)
E
Idz
H
current I in a very short vertical piece of wire of length dz

Fig. 3. Radiation field of an elemental dipole at various instants in a cycle, for a position in the plane that perpendicularly bisects the dipole, and at a radical distance *r* that is large compared to the wavelength.

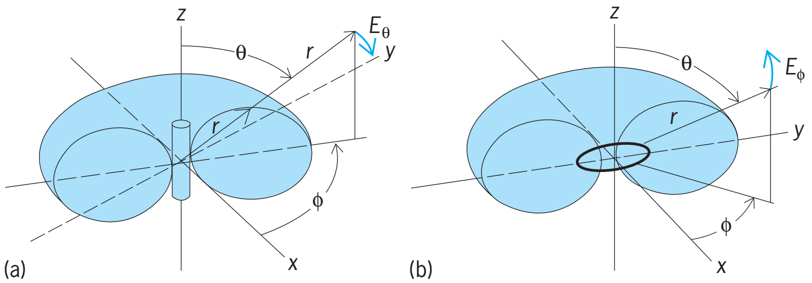


Fig. 4. Far-field patterns. (*a*) Small electric dipole. (*b*) Small magnetic dipole.

of the electric dipole, except that the electric field and magnetic field polarizations are interchanged.

Pattern shape. A small electric or magnetic dipole radiates no energy along its axis, the contour of constant energy being a toroid (**Fig. 4**). The most basic requirements of an antenna usually involve this contour in space, called the radiation pattern. The purpose of a transmitting antenna is to direct power into a specified region, whereas the purpose of a receiving antenna is to accept signals from a specified direction. In the case of a vehicle, such as an automobile with a car radio, the receiving antenna needs a nondirectional pattern so that it can accept signals from variously located stations, and from any one station, as the automobile moves. The antenna of a broadcast station may be directional; for example, a station in a coastal city would have an antenna that concentrated most of the power over the populated land. The antenna for transmission to or from a communication satellite should have a narrow radiation pattern directed toward the satellite for efficient operation, preferably radiating essentially zero power in other directions to avoid interference. Each special application of radio waves has its unique requirements for radiation patterns. *See* DIRECTIVITY; RADIO BROADCASTING.

For highly directional beam-type patterns, the concept of directivity is useful in the description of pattern shape. A simple approximation is based on beam area B at the half-power beam width of the pattern, that is, at the angles to the sides of the central beam axis where the radiated power is half the power radiated along the beam axis. Beam area b is defined as the solid angle through which all power would be radiated if the power were constant over the complete solid angle and at maximum level. For these hypothetical conditions, directivity D is defined as the surface of the sphere (4π radians) divided by beam area B in radians, as shown in Eq. (8). If the

$$D = \frac{4\pi}{B} \tag{8}$$

beam is described in terms of its two widths, θ and ϕ (in degrees), then $D = 41{,}253/(\theta\phi)$. These approximations neglect power radiated from the secondary lobes and power that is lost in the antenna conductors. Keeping the secondary lobes small is important to reduce interference in communications systems having adjacent cofrequency transmitters.

Total radiated power P_r from an antenna can be computed by integrating the energy flow outward through an imaginary sphere surrounding the antenna system. For a hypothetical spherical radiation pattern surface which has a radius of r, the radiated power P_r is given by Eq. (9), where E_s is the field

$$P_r = \frac{4\pi r^2 E_s^2}{120\pi} \tag{9}$$

strength of the radiated electric field at the surface of the imaginary sphere, and 120π is the characteristic resistance of free space. If the radiated power is 1 kW, the field strength E_s at the surface of the imaginary sphere 1 mi (1.609 km) in radius is 107.6 mV/m. This uniform spherical radiator has no directivity and hence no gain with a lossless antenna system. It is referred to as an isotropic radiator and is used as the standard for comparison of the gain of directional antennas. This antenna gain can be defined as the ratio of maximum field strength from the antenna of interest relative to the field strength that would be radiated from a lossless isotropic antenna fed with the same input power. In relation to directivity D, gain G can be expressed as in Eq. (10), where η is antenna efficiency.

$$G = \eta D \tag{10}$$

Efficiency. Antenna efficiency η is defined as the radiated power divided by the input power to the antenna. Input per P_i supplies antenna system losses P_l and the radiated power P_r, as shown by Eq. (11).

$$\eta = \frac{P_r}{P_l + P_r} \tag{11}$$

Losses in an antenna system take place in the resistance of the transmission line to the antenna, the coupling network between line and antenna, antenna conductors, insulators, and the ground system. For electrically small antennas, even small-loss resistances can significantly reduce the efficiency. In the design of transmitting antennas, the heating effect of losses must also be considered together with losses due to corona. *See* CORONA DISCHARGE.

Polarization. The plane of the electric field depends on the direction in which the current flows on the antenna (Fig. 3). The electric field is in a plane orthogonal to the axis of a magnetic dipole (Fig. 4). This dependence of the plane of the radiated electromagnetic wave on the orientation and type of antenna is termed polarization. A receiving antenna requires the same polarization as the wave that it is to intercept. By combining fields from electric and magnetic dipoles that have a common center, the radiated field can be elliptically polarized; by control of the contribution from each dipole, any ellipticity from plane polarization to circular polarization can be produced. Some antennas of the conical horn type lend themselves to symmetrical excitation of wave guides. Circular polarization can be easily detected with appropriate feed arrangements such as probes placed orthogonal to each other within the waveguide. The orientation and sense of the circular polarization components can be ascertained and the accompanying signals separated. Similar cross-polarized (simultaneous vertical and horizontal) signals of the same or different frequencies can be detected. Thus, for a single transmitting or receiving station it is possible to almost double the useful capability. Such cross-polarization systems are used extensively in both terrestrial and satellite microwave facilities. *See* POLARIZATION OF WAVES.

Impedance. The input impedance Z_{in} of an antenna is the ratio of the voltage to current at the terminals connecting the transmission line and transmitter or receiver to the antenna. The impedance can be real for an antenna tuned at one frequency but generally would have a reactive part at another frequency. The impedance can be determined through theoretical and other methods, some of which involve numerical computations. For simple antennas, approximate input or self-impedance can be computed from a knowledge of radiation patterns and approximate current distribution. For more complicated radiating structures, input impedance is usually measured experimentally. *See* ELECTRICAL IMPEDANCE.

The concept of antenna impedance may be illustrated by the example of the simplest wire conductor antenna. The current distribution and charge distribution on the wire antenna can be approximated to a fair degree of accuracy by the distributions on an open-ended transmission line for which the current is zero at its end and the charge accumulation is a maximum. For a quarter-wavelength section of line, at the input, the current is a maximum and the charge is a minimum. Since the electric field is proportional to the charge and represents the voltage, the ratio of voltage to current represents a low impedance. As the wires are separated into a V antenna, the antenna impedance still remains low since the current and voltage at the input have not changed. When the configuration is fully expanded into the half-wave antenna, the input impedance is low but not negligible since the radiation of energy must be accounted for. The power carried by the emitted radiation is I^2R_r, where R_r is called the radiation resistance. For this half-wave antenna case, R_r has a value of about 73 ohms with a small reactive part. For the electrically small antenna (elemental dipole), the radiation resistance is given by Eq. (12). For $dz/\lambda = 0.05$,

$$R_r = 80\pi^2(dz/\lambda)^2 \qquad (12)$$

R_r is about 2 ohms. The input impedance for this small antenna has a very large capacitive reactance which when tuned out with a corresponding inductor yields a resistive input of 2 ohms plus a resistance due to the imperfectly conducting antenna and inductor wires. This is not a very efficient antenna: its efficiency is low, and its low input resistance is difficult to match to the transmission line.

As a component of a receiving circuit, the antenna may be represented by Thévenin's or Norton's equivalence theorems. In the former the representation is a simple series circuit containing impedance, and also containing voltage induced from the incoming wave. Useful power is calculated in terms of the resulting current in the resistance of the load circuit. *See* NETWORK THEORY; THÉVENIN'S THEOREM (ELECTRIC NETWORKS).

If a current element of length dz is placed in a field E parallel to it, the root-mean-square voltage induced is E_s. If the element is tuned, the available power P that it captures is shown by Eq. (13). The average power per square meter P_{av} in the plane wave is shown in Eq. (14), which indicates that the effective area of the elemental dipole is $3\lambda^2/(8\pi)$.

$$P = \frac{E^2\lambda^2}{320\pi^2} \qquad (13)$$

$$P_{\text{av}} = \frac{E^2}{120\pi} \qquad (14)$$

If the reactive component of antenna impedance is appreciable, techniques similar to those used in other transmission circuits are required to obtain maximum power; that is, the impedance of the connected circuit must be conjugate to that of the antenna. Reflections caused by a mismatched antenna can cause trouble at a distant point in receiving circuitry. The solution may be to obtain a better match or to employ a nonreciprocal device which discriminates against reflected waves. *See* IMPEDANCE MATCHING.

When a driven antenna is in proximity to others, the radiation pattern of the antenna can be altered so as to yield the desired fields. The presence of the other elements can modify the input impedance of the antenna, called driving-point impedance. The combination of the elements and driven antenna constitute an array, as discussed below.

Because of the interaction between the elements of an array, a distinction is made between a driven element to which a transmission line directly connects and a parasitic element that couples into the array only through its mutual electromagnetic coupling to the driven elements. Any conductor, such as a guy wire, in the vicinity of a primary or driven antenna element may act as a parasitic element and affect antenna impedance and radiation pattern (**Fig. 5***a*). The driving-point impedance and the radiation pattern are controlled by the tuning and spacing of parasitic elements relative to driven elements. If a parasitic element reduces radiation in its direction from a driven element, it is called a reflector; if it enhances radiation in its direction, it is called a director. *See* YAGI-UDA ANTENNA.

Array antennas. An array of antennas is an arrangement of several individual antennas so spaced and phased that their individual contributions add in the preferred direction and cancel in other directions. One practical objective is to increase the signal-to-noise ratio in the desired direction. Another objective may be to protect the service area of other radio stations, such as broadcast stations. *See* SIGNAL-TO-NOISE RATIO.

The simplest array consists of two antennas. Using phasor notation, the azimuth pattern is given by

-S

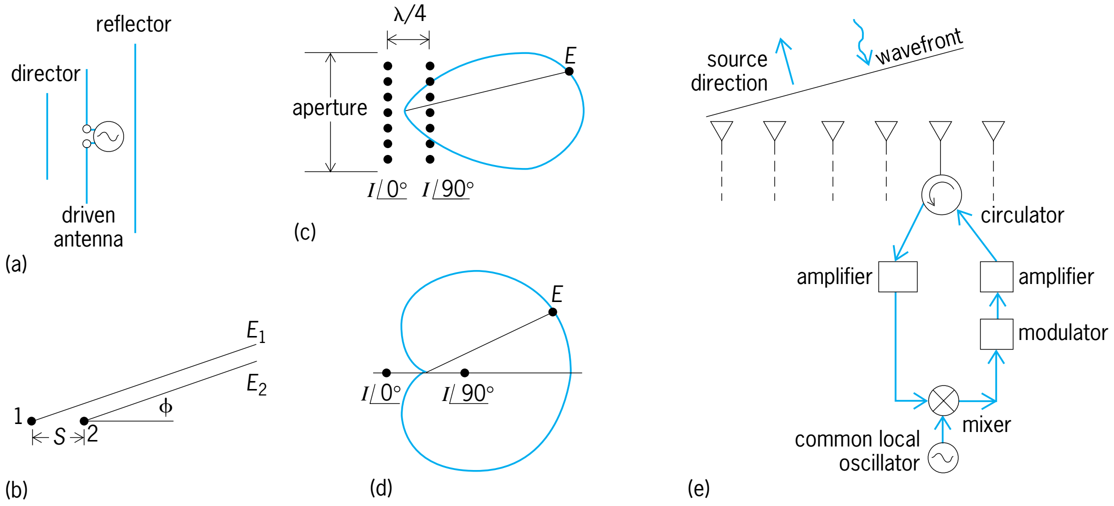


Fig. 5. Antenna arrays. (*a*) Simple three-element antenna. (*b*) Two-antenna array. (*c*) Broadside box array. (*d*) Unidirectional (cardioid pattern) array. (*e*) Retrodictive array. Each element has the same electronics circuit. (*Part e after R. E. Collin, Antennas and Radiowave Propagation, McGraw-Hill 1985*)

Eq. (15), where E is field strength for unit distance at

$$E = E_1 \,\underline{/0^\circ} + E_2 \,\underline{/2\pi(S/\lambda)\cos\phi + \psi_2} \tag{15}$$

azimuth angle ϕ, E_1 and E_2 are nondirectional values of the field strength of the individual antennas, S is the spacing from antenna 1 to antenna 2, and ψ_2 is the electrical phasing of antenna 2 with respect to antenna 1 (Fig. 5*b*). *See* ALTERNATING-CURRENT CIRCUIT THEORY.

The four variables that the simple two-antenna array provides make possible a wide variety of radiation patterns, from nearly uniform radiation in azimuth to a concentration of most of the energy into one hemisphere, or from energy in two or more equal lobes to radiation into symmetrical but unequal lobes.

In a directional antenna a feeder system is required for the proper division of power and control of the phase of the radiated fields. The feeder system is also called upon to match impedances between the transmitter of transmission line and the antenna.

For further control over the radiation pattern a preferred arrangement is the broadside box array. In this array, antennas are placed in a line perpendicular to the bidirectional beam. Individual antenna currents are identical in magnitude and phase. The array can be made unidirectional by placing an identical array 90° to the rear and holding its phase at 90° (Fig. 5*c*). The directivity of such a box array increases with the length or aperture of the array.

Another popular arrangement is the in-line array, in which the current in each antenna is equal but the phase varies progressively to give an end-fire unidirectional pattern (Fig. 5*d*). The end-fire array, for a given number of elements, does not have as much gain as the broadside box array.

If antennas are stacked vertically in line to form a collinear array with the currents of each antenna in phase, a symmetrical pattern results in the azimuth plane. This type of array is used for commercial broadcasting to increase the nondirectional ground coverage for a given transmitter power.

One useful antenna system consists of many, usually identical, antenna elements that are connected through separate networks capable of introducing signal phase changes so that a progressive phase shift along the array results in a corresponding shift in direction of the radiation beam. Of course, to obtain the broadside array property mentioned above, the phase-shift difference is maintained at zero or a multiple of 360°. For the end-fire array, maximum radiation is desired along one end of the array and a minimum, or zero, in the back radiation. The progressive phase shift ζ (a linearly increasing phase delay) depends on the uniform spacing d of the elements. For an array with N elements, $\zeta = \pi/N$ and $d = \lambda/(2N)$, where λ is the wavelength.

Further use of array concepts has enabled improvements in communications. By introducing a network for each antenna element (Fig. 5*e*), it is possible to receive a signal from a source direction and to return a signal in the direction of the source. The returned signal can be modulated or amplified or have its frequency changed. Such an array is called a retrodirective array. Basically, the array seeks out the incoming signal and returns one of useful characteristics, such as that which is needed for the communication between a moving vehicle and a stationary or slowly moving source.

The incoming pilot signal excites the individual elements with a progressive phase delay; the element closest to the source receives the incoming wavefront in advance of the other array elements, with the element farthest from the source experiencing the largest phase delay. In order that a transmitted signal (with information) be sent in the direction of the source, it is necessary that the phase delays be converted into phase advances. That is to say, the element farthest from the source must have the largest phase advance, and the other elements of the array must have progressively lesser phase advances. The overall array thus transmits toward the source direction.

The phase reversal (from a phase delay to phase advance) is accomplished at each array element through the use of a circulator (separating the incoming signal from the outgoing), a common local oscillator of twice the frequency of communication, a mixer, a separation of the resulting frequencies, and a transmission of the responding signal with information sent to the source.

Antenna arrays can also be designed to adapt their radiation or receiving patterns through the use of proper signal processing of the energy received by each element of the array. It is possible to enhance the signal received from a desired direction or to reject an unwanted signal or interference from another direction. The signal processing involves phase comparison and phase-locked loops to obtain the desired results. Such antenna arrays are called adaptive arrays. *See* PHASE-LOCKED LOOPS.

It is possible to use more than two antennas in an array to achieve a desired pattern shape. If a symmetrical pattern is satisfactory, the antennas can be in line. If the arrays are not in line, the protection directions can be specified by the nulls of the individual pairs of antennas used in the array. When necessary, the number of pairs of antennas can be increased to provide further control of the pattern shape.

Bandwidth. The bandwidth of an antenna may be limited by pattern shape, polarization characteristics, and impedance performance. Bandwidth is critically dependent on the value of Q; hence the larger the amount of stored reactive energy relative to radiated resistive energy, the less will be the bandwidth. *See* Q (ELECTRICITY).

For efficient operation, a low standing-wave ratio at the antenna input terminals is necessary over the operating frequency range. Usually the input impedance will vary in both resistance and reactance over the desired frequency interval. It is, therefore, of interest to determine the bandwidth that can be achieved for a given standing-wave ratio. For a particular antenna the bandwidth that can be obtained by compensating networks has a theoretical limit; therefore, if a wider band is required, it becomes necessary to select antenna types with the inherent characteristic of wider bandwidth, such as the frequency-independent antennas discussed below.

Electrically small antennas. Antennas whose mechanical dimensions are short compared to their operating wavelengths are usually characterized by low radiation resistance and large reactance. This combination results in a high Q and consequently a narrow bandwidth. Current distribution on a short conductor is sinusoidal with zero current at the free end, but because the conductor is so short electrically, typically less than 30° of a sine wave, current distribution will be essentially linear. By end loading to give a constant current distribution, the radiation resistance is increased four times, thus greatly improving the efficiency but not noticeably altering the pattern.

An end-fed monopole antenna is common at very low frequencies for long-range communication, commercial broadcasting, and mobile use. The small loop antenna is used extensively for direction finding and navigation. Where height is a limiting factor at higher frequencies, the monopole height can be reduced by forming the conductor into a helical whip or by top loading.

Slot antenna. Where no height is permitted, a slot is used. Just as the magnetic loop is related to the electric dipole, so a slot in a conductive surface is related to a conductive wire in space. The conductive surface is usually the outside surface of a waveguide within which the radio energy travels. A slot in the waveguide allows the energy to radiate. Typically the slot is narrow and a half-wavelength long. The configuration of the electric field radiated from the slot is the same as the magnetic field from a wire of like dimensions; thus one is the dual of the other. The slot may be fed by a transmission line connected across its narrow dimension, by a resonant cavity behind it, or (as mentioned above) from a waveguide. Because it is flush with a metallic surface, a slot antenna is advantageous in aircraft.

Broadband antenna. The impedances of very thin wire antennas such as half-wave antenna dipoles and quarter-wave monopoles are highly sensitive to frequency, although the radiation pattern of these antennas is insensitive to frequencies off resonance. The impedance can have large capacitive reactances for frequencies lower than resonance and inductive reactances for frequencies above resonance. This variation can be minimized by choosing dipoles that are thick, by minimizing reflections at the feed points by using conical conductor, or by combinations of these choices.

Nonresonant antennas. Long-wire antennas, or traveling-wave antennas, are usually one or more wavelengths long and are untuned or nonresonant.

Horizontal single-conductor type. The radiation pattern of a long conductor in free space depends upon its length in wavelengths. A $0.5\text{-}\lambda$ conductor will radiate broadside, but as it is made longer the pattern splits, the major lobe coming closer to, but never reaching, the direction of the conductor. The pattern is a figure of revolution of its own cross section around the conductor as an axis.

When a horizontal unterminated line with standing electric current waves of uniform amplitude is placed above perfect ground, it has a radiation pattern symmetrical about the center with its major lobes closest to the line in both directions. In the practical case with constant phase velocity and exponential variation in amplitude, the pattern lobes will be unsymmetrical with larger lobes toward the open end. The total number of lobes is the same as the number of $0.5\text{-}\lambda$ lengths of the conductor. If the far end of the line is terminated, the antenna radiates only one major lobe in the direction the wave travels down the line to the termination.

Arrays of long-conductor type. Because a single long-conductor terminated antenna has rather high side lobes and a major lobe at an angle to the axis of the conductor, two long terminated conductors can

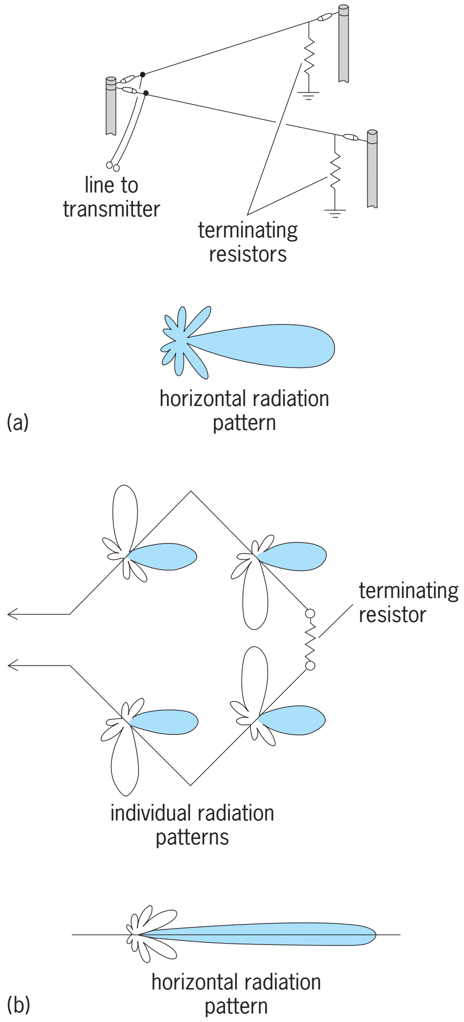

line to transmitter
terminating resistors
horizontal radiation pattern
(a)
terminating resistor
individual radiation patterns
horizontal radiation pattern
(b)

Fig. 6. **Arrays of long-conductor antennas. (*a*) Horizontal terminated V. (*b*) Horizontal terminated rhombic.**

be combined into a horizontal V array (**Fig. 6*a***). The angle between the V array is determined by the length of the conductors and is usually chosen so that the major lobes will add in phase.

The commonest type of long-conductor array is the rhombic antenna (Fig. 6*b*). Major lobes of four legs add in phase to form the resultant major lobe, while major lobes at right angles tend to cancel, leaving only smaller side lobes. Because of its simplicity, low cost, and wide bandwidth, the rhombic array is widely used for both transmission and reception where the side lobes can be tolerated and the real estate dimensions permit.

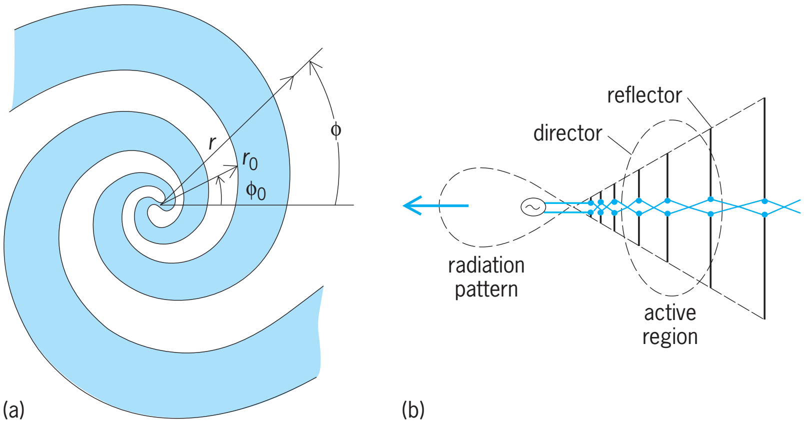


Fig. 7. **Frequency-independent antennas. (*a*) Equiangular spiral (*after D. J. Angelakos and T. E. Everhart, Microwave Communications, Krieger, 1983*). (*b*) Log-periodic structure.**

Dielectric type. A dielectric material, formed into either a solid rod or a hollow cylinder, is a waveguide. The wavelength of a wave entirely inside a large, solid dielectric rod is less than the free-space wavelength; rather, the wavelength is this free-space length divided by the square root of the dielectric constant of the rod. If the rod diameter is large compared to the wavelength inside it, most of the energy in the wave travels inside the dielectric. Even if the rod diameter is somewhat less than the wavelength inside the dielectric, the wave travels along the rod, with some of the energy inside the dielectric and some outside. However, if the diameter of the rod is reduced below a half-wavelength, the velocity of the wave increases to that of free space and the wave continues beyond the end of the rod into free space. The major lobe is then in the direction of the rod. *See* DIELECTRIC MATERIALS.

Frequency-independent antennas. Antennas that consist of lengths of wires are sensitive to changes in frequency as the wire lengths become equal to a half-wavelength, a full wavelength, and so forth. An infinitely long biconical antenna would have characteristics independent of frequency since its dimension is specified by angle only, but, of course, if such an antenna is truncated its length is also specified, making it frequency-dependent.

There are two principal approaches to constructing frequency-independent antennas. The first is to shape the antenna so that it can be specified entirely by angles; hence when dimensions are expressed in wavelengths, they are the same at every frequency. Planar and conical equiangular spiral antennas adhere to this principle. The energy radiates as the wave progresses along an equiangular spiral antenna, so that beyond a dimension of circumference of the spiral equal to the wavelength the antenna structure can be terminated. This criterion determines the lowest frequency of practical application. The preciseness of the input portion of the spiral determines the highest frequency of operation. If r_1 is the radial distance to the spiral corresponding to a wavelength λ_1, such that Eq. (16) is valid, then at a new

$$r_1 = r_0 e^{a(\phi_1 - \phi_0)} \tag{16}$$

wavelength λ_2 this radial distance should be given by Eq. (17) if the antenna characteristics are to be

$$r_2 = (\lambda_2/\lambda_1) r_1 \tag{17}$$

frequency-independent. This yields Eq. (18). Here a,

$$r_2 = r_0 e^{a(\phi_2 - \phi_0)} \tag{18}$$

r_0, and ϕ_0 represent the spiral parameters, and ϕ_1 and ϕ_2 are the angles at which the radial distance is r_1 and r_2 respectively (**Fig. 7*a***).

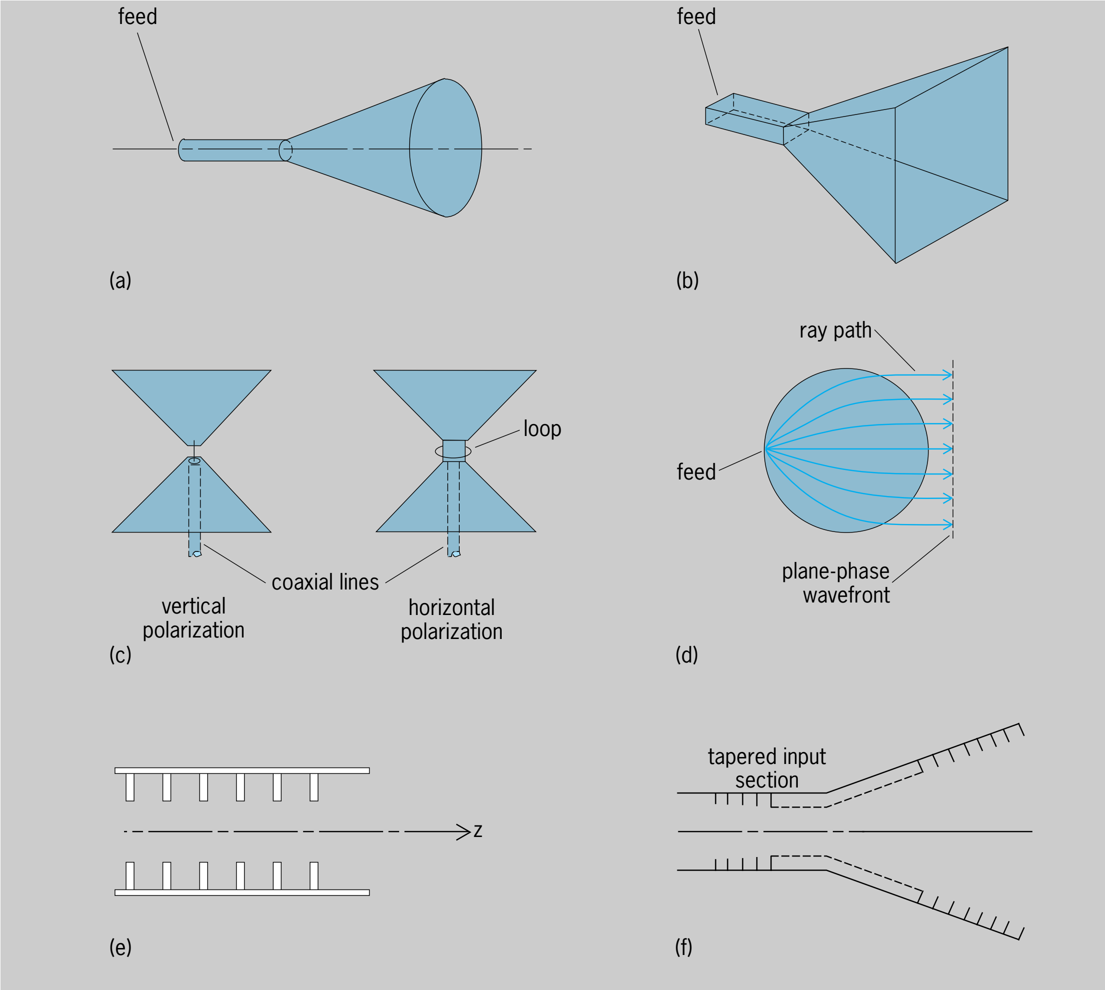


Fig. 8. Direct-aperture antennas. (*a*) Conical horn. (*b*) Pyramidal horn. (*c*) Biconical horns, electrically fed for vertical polarization or magnetically fed for horizontal polarization. (*d*) Luneberg lens with dielectric sphere between feed and plane-phase wavefront. (*e*) Corrugated circular waveguide. (*f*) Corrugated conical horn. (*Parts e and f after R. E. Collin, Antennas and Radiowave Propagation, McGraw-Hill, 1985*)

The second approach depends upon complementary shapes. According to this principle, which is used in constructing log-periodic antennas, before the structure shape changes very much, when measured in wavelengths, the structure repeats itself. This periodicity in the exponent of the structural dimensions yields a class of antennas which is frequency-independent in the sense that as the frequency of the input signal is changed, another set of elements comes into resonance, with the same electrical dimensions for each frequency, within the limits set by the fineness of the element spacing (Fig. 7*b*).

By combining periodicity and angle concepts, antenna structures of very large bandwidths become feasible.

Direct-aperture type. When they are to be used at short wavelengths, antennas can be built as horns, mirrors, or lenses. Such antennas use conductors and dielectrics as surfaces or solids in contrast to the antennas described thus far, in which the conductors were used basically as discrete lines. *See* MICROWAVE OPTICS.

Horn radiators. The directivity of a horn increases with the size of the mouth of the horn; however, tolerances must be held close if the variations in magnitude and phase distribution are to be low enough to maintain high gain over a broad band. Conical and pyramidal horns are commonly used for high-gain beams (**Fig. 8***a* and *b*). Biconical horns (Fig. 8*c*) have an omnidirectional pattern with either vertical or horizontal polarization.

Luneberg lens. If in a dielectric sphere the index of refraction varies with distance from the center of the sphere, a plane wave falling on the surface of one hemisphere is focused at the center of the opposite hemisphere. Such an arrangement is called a Luneberg lens (Fig. 8*d*). By reciprocity, energy is fed into the lens at the focal point for transmitting; energy is removed from the lens at this point for receiving. Only the feed point need be moved around the lens to steer the pencil beam.

For high frequencies a horizontal conducting grid structure serves as a Luneberg lens antenna. Because it will focus rays from any azimuth direction, several high-frequency beams can be generated in different azimuth directions at the same time.

Corrugated conical horns. A very special antenna system that is used either as a feed of a reflector type or as a direct radiator is the corrugated conical horn antenna (Fig. 8*e* and *f*). For the ordinary metallic-wall waveguide and horn, the metal walls present boundary conditions for the electric field that are different from those for the magnetic field. By replacing the metal walls by pseudowall-nonwall in the corrugated structure, the boundary conditions become impartial to the type of field: electric or magnetic. The hybrid wall appears as both a short circuit and an open circuit. The latter is obtained by depressions of length comparable to a quarter wavelength at some middle frequency. This horn has a high degree of circular symmetry in its radiation and receiving patterns.

Reflector type. By using reflectors it is possible to achieve high gain, modify patterns, and eliminate backward radiation. A low-gain dipole, a slot, or a horn, called the primary aperture, radiates toward a larger reflector called the secondary aperture. The large reflector further shapes the radiated wave to produce the desired pattern. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

Plane sheet reflectors. One common plane reflector antenna consists of a dipole parallel to a flat conducting sheet (**Fig. 9***a*). The perfectly conducting screen creates an image of the physical structure (mirror image) but with the currents of the image flowing in the opposite direction. The forward pattern is the same as that of a two-element dipole array. If the screen is large enough, it prevents all back radiation. Usually the screen is limited in size, and some radiation exists to the rear. Gain depends on spacing and can exceed 6 dB above a free-space dipole.

Corner reflectors. If two flat conducting sheets intersect along the y axis at an angle α (Fig. 9*b*), an effective directional antenna results when a dipole is placed a distance S from the corner. Gain depends on corner angle α, dipole spacing S, and antenna losses; gain can exceed 12 dB for $\alpha = 45^\circ$. *See* CORNER REFLECTOR ANTENNA.

Parabolic reflectors. If an isotropic source is placed at the focus F of a parabola, the radiated wave is reflected from the parabolic surface as a plane wave at the aperture plane (Fig. 9*c*). The physical size of the parabola determines the size of the aperture plane.

If the parabolic reflector is a surface of translation (Fig. 9*d*), an in-phase line source serves as the primary aperture, such as the two dipoles in this illustration. Beam width in the vertical, z direction is controlled by the parabola aperture, while the azimuth beam width, in the y direction, is controlled by the length of the cylindrical parabola.

If the parabolic reflector is a surface of rotation, it converts a spherical wave from an isotropic source at its focus to a uniform plane wave at its aperture. The presence of the primary antenna in the path of the reflected wave has two disadvantages. (1) The reflected wave modifies the input impedance of the primary radiator. (2) The primary radiator obstructs the reflected wave over the central portion of the aperture (Fig. 9*e*).

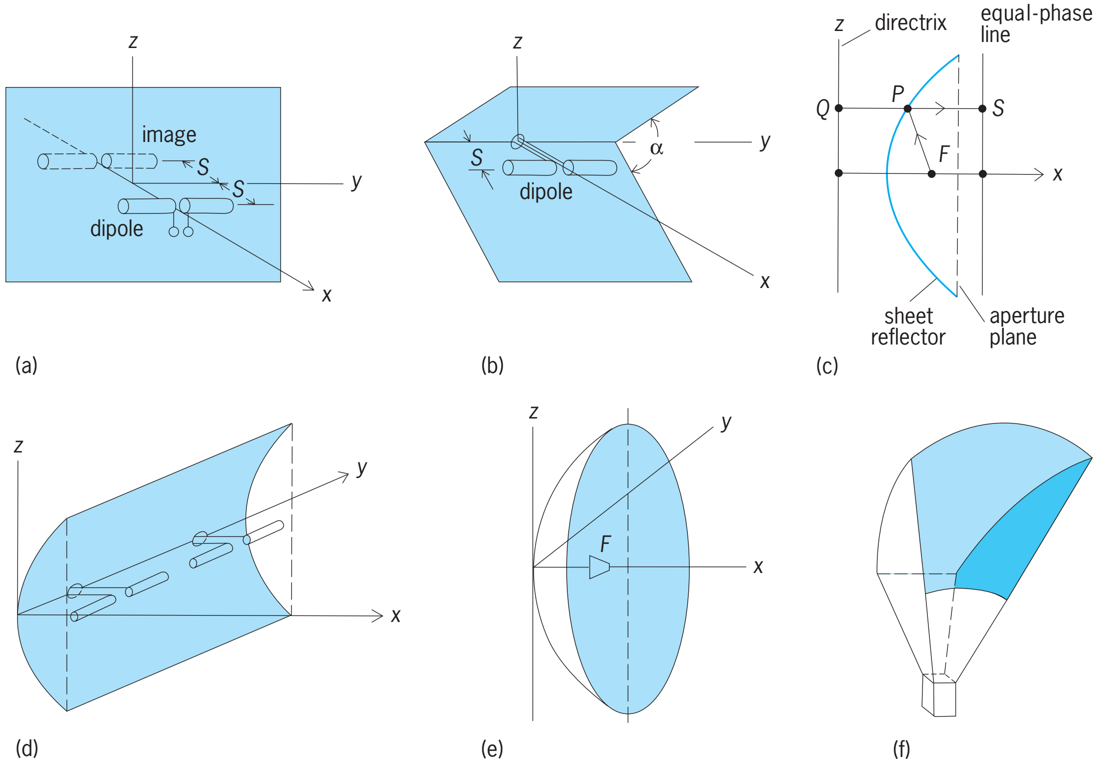


Fig. 9. Reflector antennas. (*a*) Plane sheet reflector. (*b*) Corner reflector. (*c*) Parabolic reflector, where $FP = QP$ and $FP + PS$ is a constant for all points P on the reflector. (*d*) Cylindrical parabolic reflector. (*e*) Paraboloidal reflector. (*f*) Horn paraboloid reflector.

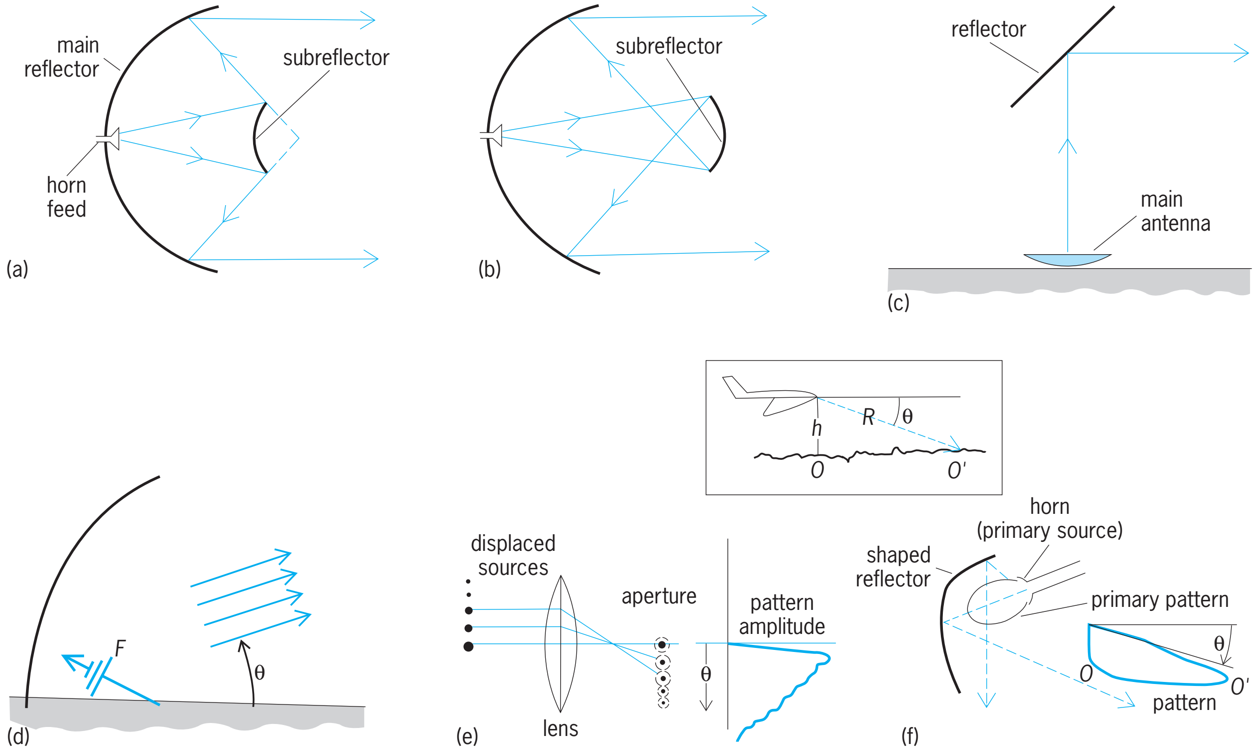


Fig. 10. Two-element systems. (*a*) Cassegrain. (*b*) Gregorian. (*c*) Periscope. (*d*) Spherical steerable system. Feed is located at focus *F*. θ = elevation angle. (*e*) Distributed sources. (*f*) Shaped reflector. Inset to parts *e* and *f* shows geometry of aircraft radar application. Power emitted in the direction of *O′* must be greater than that emitted from *O* to compensate for weaker signals reflected from *O′* because range *R* to *O′* is greater than distance *h* to *O*. (*Part e and f after D. J. Angelakos and T. E. Everhart, Microwave Communications, Krieger, 1983*)

To avoid both of these disadvantages, only a portion of the paraboloid can be used; the focus, where the primary aperture must be placed, is then to one side of the secondary aperture. For example, a primary-aperture horn placed at one side can illuminate the parabolic reflector (Fig. 9*f*). This type of antenna is commonly used in point-to-point microwave systems, because it is broadband and has a very low noise level. Because of its high front-to-back ratio, the horn paraboloid reflector is used at the ground station for satellite communication.

Two-reflector antennas. A beam can be formed in a limited space by a two-reflector system. The commonest two-reflector antenna, the Cassegrain system, consists of a large paraboloidal reflector. It is illuminated by a hyperbolic reflector, which in turn is illuminated by the primary feed (**Fig. 10***a*). The Gregorian system (Fig. 10*b*) is similar, except that an elliptical reflector replaces the hyperbolic reflector and consequently must be placed farther out than the focal point of the paraboloid. This system is less compact than the Cassegrain system, and less frequently used. A variation of the Gregorian involves a shaping of the primary reflector to further modify the pattern; it is called the Dragonian.

In microwave relay systems, one common practice is to place the paraboloid at ground level (Fig. 10*c*). The feed system is readily accessible, and long lengths of transmission line are not required. A flat sheet at the top of a tall tower redirects the beam over obstacles to the next antenna.

By using a portion of a hemisphere (Fig. 10*d*) illuminated by a movable feed, the spherical reflecting surface causes the wave to be reflected off the ground to form a beam at a specified elevation angle. With this arrangement it is possible to steer the beam both in elevation and in azimuth. This system is used in the Arecibo instrument, a 1000-ft (300-m) radio telescope antenna.

A pattern of desired shape can alternatively be obtained by displacing the primary source (Fig. 10*e*) or by shaping the reflector (Fig. 10*f*). An example of the application of such patterns is an aircraft radar system.

Low-profile antennas. A series of antennas are useful in situations which require a low profile. The slot antennas mentioned above constitute a large portion of this group. In essence, replacing a wire (metal) by a slot (space), which is a complement of the wire, yields radiation characteristics that are basically the same as those of the wire antenna except that the electric and magnetic fields are interchanged. The impedance Z_d of the driven antenna, in the simplest case of a dipole antenna, is related to that of the equivalent driven slot antenna Z_s by Eq. (19). A

$$Z_s Z_d = \frac{(120\pi)^2}{4} \qquad (19)$$

folded dipole antenna has an impedance $Z_{fd} = 4Z_d$ that is related to the impedance Z_{fs} of the equivalent slot antenna by Eq. (20).

$$Z_{fs}Z_{fd} = \frac{(120\pi)^2}{4} = Z_{fs}(4Z_d) \qquad (20)$$

Because flush-mounted antennas present a low profile and consequently low wind resistance, slot-type antennas have had considerable use in aircraft, space-launching rockets, missiles, and satellites. They have good radiation properties and are capable of being energized so as to take advantage of all the properties of arrays, such as scanning, being adaptive, and being retrodirective. These characteristics are obtained without physical motion of the antenna structures. Huge slot antenna arrays are commonly found on superstructures of aircraft carriers and other naval ships, and slot antennas are designed as integral parts of the structure of aircraft, such as the tail or wing.

Patch antenna. This antenna consists of a thin metallic film which is attached to a dielectric substrate mounted on a metallic base. Depending on its use, the patch can be of different shapes and can be driven in various fashions. Driven at one end, the radiated electric field at this end has a polarization that is in phase with the radiated electric field at the farther end of the patch antenna.

MMIC antennas. Planar antennas are designed as integral parts of monolithic microwave integrated circuits (MMICs). Coupling between can be effected through the use of planar (flush-mounted) antennas fabricated directly on the microelectronics chips (integrated circuits). This arrangement eliminates the need for coaxial lines, which at these microwave frequencies exhibit considerable losses. As is the case with other planar antennas, it is possible to design circuitry so as to obtain many, if not all, the properties of arrays mentioned above. The elements of these arrays can take on the form of slot antennas or patch antennas (of course with suitable modification for use on the MMICs). *See* MICROWAVE.

Diogenes J. Angelakos

Bibliography. American Radio Relay League, *ARRL Antenna Book*, 18th ed., 1997; D. J. Angelakos and T. E. Everhart, *Microwave Communications*, 1968, reprint 1983; C. A. Balanis, *Antenna Theory: Analysis and Design* 2d ed., 1996; R. E. Collin, *Antennas and Radiowave Propagation*, 1985; R. C. Johnson, *Antenna Engineering Handbook*, 3d ed., 1993; J. D. Kraus, *Antennas*, 2d ed., 1988; W. L. Pritchard et al., *Satellite Communication Systems Engineering*, 2d ed., 1992; W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design* 2d ed., 1997.

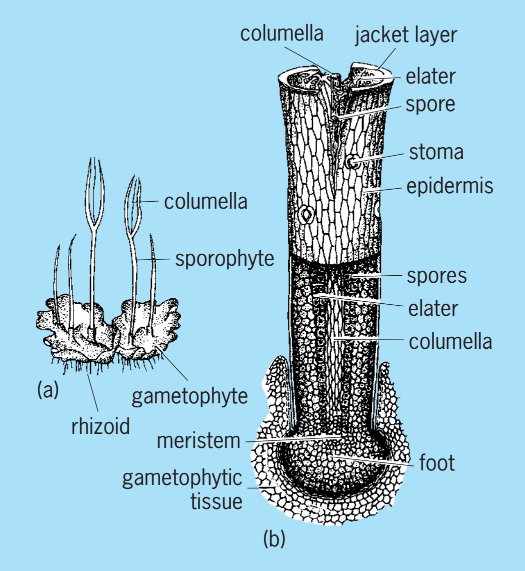
Anthocerotopsida

A small class of the plant division Bryophyta, made up of plants commonly called hornworts. There is a single order with six genera.

Gametophyte structure. The gametophytes are flat thalli, often forming rosettes of uniform thickness (see **illus.**) or, in *Dendroceros*, with a thickened midrib and crisped unistratose wings. Ventral scales are absent. Rhizoids are unicellular, with thin, smooth walls. The thallus of undifferentiated tissue is sometimes provided with cavities containing mucilage (and often occupied by *Nostoc* colonies) and ventral pores leading to them. The cells are thin-walled and lack oil bodies; at least those at the surface have one, two, or more large chloroplasts with a central pyrenoid. Slime papillae (or mucilage hairs) are lacking. The stalked antheridia develop from internal cells and occur singly or in groups in cavities beneath the upper surface of the thallus and become exposed by the rupture of overlying tissue. The archegonia, also endogenous, are embedded in the dorsal tissue of the thallus. Paraphyses are lacking. The sporophyte is surrounded at the base by a tubular involucre (or completely surrounded in *Notothylas*). A calyptra is lacking.

Sporophyte structure. The sporophyte consists of a massive foot and an erect, long-cylindric, green capsule (see illus.), though in *Notothylas* the capsule is horizontal and spindle-shaped. The capsule, indeterminate in growth owing to a basal meristem, dehisces from the apex into two valves which are usually twisted when dry. The wall consists of severa l cell layers in a more or less solid tissue having one, two, or more chloroplasts per cell. Stomata with two guard cells are usually present. The spore sac, derived from the amphithecium, surrounds and overarches the slender columella (lacking in *Notothylas*). Spore maturation proceeds from the apex downward. The tetrahedral spores are mingled with pseudelaters of one to several cells, with or without spiral bands. A protonema is lacking. The haploid chromosome number is 5 or 6. *See* REPRODUCTION (PLANT).

The elongate capsule dehiscing by two valves and its indeterminate growth from a basal meristematic



Hornwort *Anthoceros*. (*a*) Thalloid gametophyte with long hornlike sporophytes. (*b*) Elongated sporophyte with basal foot in gametophytic tissue. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

tissue are unique, as are the large chloroplasts with a central pyrenoid and the endogenous origin of sex organs. (The chloroplasts are sometimes single in the gametophyte and paired in the sporophyte.) The spore mother cells undergo meiosis directly, but the elater mother cells usually undergo several mitotic divisions before differentiation as pseudelaters of an unreduced chromosome number. As a result, the pseudelaters may be more numerous than the spores. (In the Hepaticopsida, by contrast, the spore mother cells undergo meiosis directly or undergo several mitotic divisions before meiosis, whereas the elater mother cells mature directly into diploid elaters.) *See* ANDREAEOPSIDA; BRYOPHYTA; BRYOPSIDA; HEPATICOPSIDA; SPHAGNOPSIDA. Howard Crum

Bibliography. M. Fulford, Contemporary thought in plant morphology: Hepaticae and Anthocerotae, *Phytomorph*, 14:103–119, 1964; G. G. Häel de Menéndez, Estudio de las Anthocerotales y Marchantiales de la Argentina, *Opera Lilloana*, 7:1–297, 1962; J. Proskauer, Studies on the morphology of *Anthoceros*, pt. I., *Ann. Bot*., n.s., 12:237–265, 1948, pt. II, 12:427–439, 1948; J. Proskauer, Studies on Anthocerotales, pt. III: The genera *Anthoceros* and *Phaeoceros, Bull. Torrey Bot. Club*, 78:331–349, 1951, pt. IV, 80:65–75, 1953; J. Proskauer, Studies on Anthocerotales, pt. V: The genera Anthoceros and Phaeoceros, *Phytomorph*, 7:113–135, 1957, pt VI: On spiral thickenings in the columella and its bearing on phylogeny, *Phytomorph*, 10:1–19, 1960.

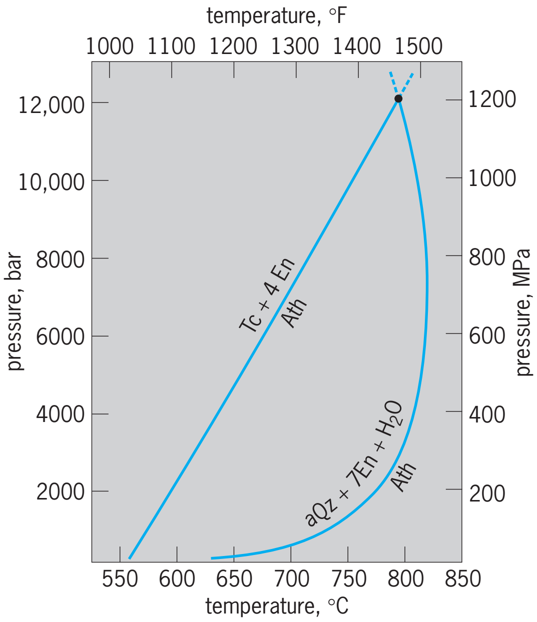
Anthophyllite

A magnesium-rich orthorhombic amphibole with perfect {210} cleavage and a color which varies from white to various shades of green and brown. It is a comparatively rare metamorphic mineral which occurs as slender prismatic needles, in fibrous masses, and sometimes in asbestiform masses. Anthophyllite may occur together with calcite, magnesite, dolomite, quartz, tremolite, talc, or enstatite in metacarbonate rocks; with plagioclase, quartz, orthopyroxene, garnet, staurolite, chlorite, or spinel in cordierite-anthophyllite rocks; and with quartz and hematite in metamorphosed iron formations and with talc, olivine, chlorite, or spinel in metamorphosed ultrabasic rocks. Anthophyllite is distinguished from other amphiboles by optical examination or by x-ray diffraction, and from other minerals by its two cleavage directions at approximately 126° and 54°.

Anthophyllite has the general formula below, with

$$(\mathrm{Mg, Fe^{2+}})_{7-x}\mathrm{Al}_x(\mathrm{Al}_x\mathrm{Si}_{8-x}\mathrm{O}_{22})(\mathrm{OH, F, Cl})_2$$

$x < 1.0$. For aluminum-poor varieties, up to about 40% of the Mg may be replaced by Fe^{2+}; higher iron contents result in the formation of the monoclinic amphibole cummingtonite. Increasing the aluminum content in anthophyllite beyond $x = 1.0$ results in the formation of the orthohombic amphibole gedrite; aluminous anthrophyllite can accommodate more Fe^{2+} than Al-poor varieties.



Stability field of anthophyllite in pressure-temperature space calculated with the program GEO-CALC. Broken lines extend from the invariant point at which the equilibrium terminates. Ath = anthophyllite, En = enstatite, Fo = forsterite, *a*Qz = quartz (*a* is variable coefficient), Tc = talc.

The stability field of anthophyllite in pressure–temperature space (see **illus.**) is bounded at low temperatures by the equilibrium 4 enstatite + talc = anthophyllite and at high temperatures by the equilibrium anthophyllite = 7 enstatite + quartz + H_2O. *See* AMPHIBOLE; BIOPYRIBOLE; CUMMINGTONITE.

J. V. Chernosky

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 2B: *Double-Chain Silicates*, 2d ed., 1997.

Anthozoa

A class of phylum Cnidaria (formerly known as Coelenterata) whose members are marine, live exclusively as polyps, and occur as solitary individuals or in clonal or colonial groups. Most adult anthozoans are attached to a firm object or the sea bottom but some burrow into soft sediments, and rare ones are planktonic. As is typical of cnidarians, sexual reproduction commonly gives rise to a planktonic larva, the planula, which metamorphoses to a polyp, usually when it settles onto the substratum. *See* CNIDARIA.

Defining characteristics. Anthozoans are distinguished from members of other classes of Cnidaria by the absence of a medusoid phase in the life cycle as well as by certain characteristic features of the polyp: (1) hollow tentacles, into which the coelenteron (gastrovascular cavity) extends; (2) some to many cells in the mesoglea; (3) gametes of endodermal origin; (4) an actinopharynx (also called a stomodeum), a throatlike tube that extends from

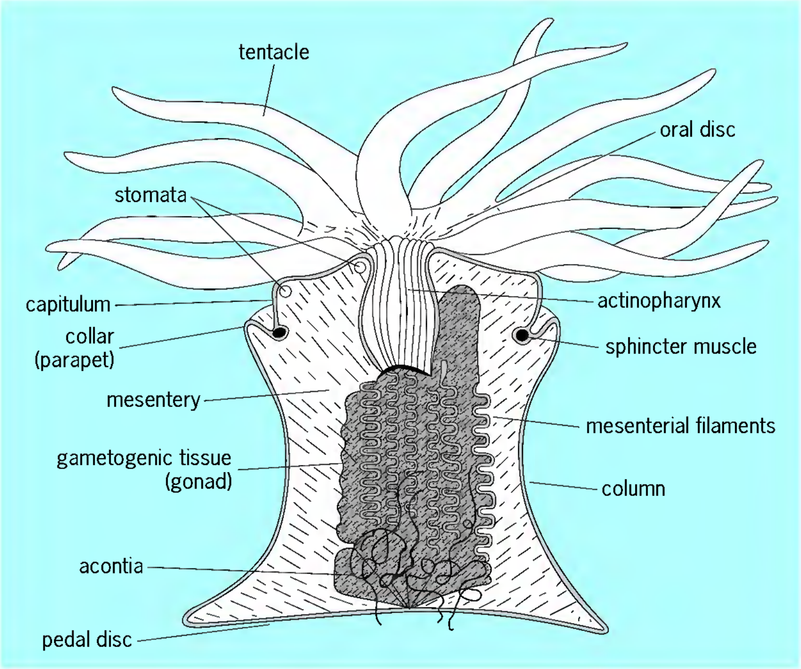


Fig. 1. Longitudinal section through a sea anemone. (*Reprinted with permission from Interactive Glossary of Actinarian (Sea Anemone) Anatomical Terms*, http://darwin.nhm.ku.edu/inverts/InterGlossary/basicintro.htm)

the mouth partly into the coelenteron (**Fig. 1**); and (5) mesenteries, which are radially arrayed sheets of tissue extending from the body wall that partition the coelenteron into wedge-shaped compartments. Some or all mesenteries, termed "complete," extend from the body wall to the actinopharynx, whereas "incomplete" ones do not reach the actinopharynx (**Fig. 2**). A polyp of the subclass Octocorallia has eight mesenteries and eight tentacles; each tentacle has small side branches, pinnately arrayed. A polyp of the subclass Hexacorallia typically has roughly a multiple of six tentacles and mesenteries, although some have octamerous, decamerous, or, rarely, irregular symmetry.

Running the length of the actinopharynx, which is more or less ovoid in cross section, are one or more grooves called siphonoglyphs Fig. 2, which are smoother and have longer cilia than the rest of the actinopharynx. Octocorals and some hexacorals have one siphonoglyph; many hexacorals have two (or rarely more), located at diametrically opposite ends of the actinopharynx. The arrangement of retractor muscles on mesenteries that connect to the siphonoglyphs may differ from that of the other mesenteries. The presence of siphonoglyph(s) and these anatomically distinctive mesenteries provide a bilaterality to animals that belong to a group commonly considered radially symmetrical. Thus anthozoans are described as biradial, with elements of both radial and bilateral symmetry.

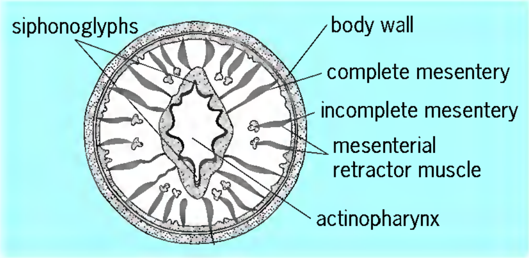


Fig. 2. Cross section (at the level of the actinopharynx) through the column of a sea anemone. (*Adapted from L. H. Hyman, The Invertebrates: Protozoa Through Ctenophora, vol. 1, McGraw-Hill, 1940*)

Classification. Class Anthozoa includes soft, stony, and black corals, sea pens, sea fans and whips, and sea anemones sensu lato (that is, in the broad sense—including all its subordinate taxa). The current classification of Anthozoa is listed below. Two groups of corals that lived in the Paleozoic Era, Rugosa and Tabulata, may be included in Anthozoa; but without knowing their internal anatomy, their placement and how modern anthozoans are related to them are uncertain.

Class Anthozoa
  Subclass Octocorallia (Alcyonaria)
    Order: Alcyonacea [soft corals]
      Helioporacea [blue corals]
      Pennatulacea [sea pens]
  Subclass Hexacorallia (Zoantharia)
    Order: Actiniaria [sea anemones in the strict sense]
      Antipatharia [black corals]
      Ceriantharia [tube anemones]
      Corallimorpharia [corallike anemones]
      Scleractinia [stony corals]
      Zoanthidea [mat anemones]

The three groups of octocorals share so many features that there is no doubt they constitute a clade which, it has been proposed, merits class status. Some classification schemes put Ceriantharia in its own subclass because it differs so much in morphology from other hexacorals; although molecular sequence data support the distinctness of this group, whether it belongs with hexacorals is still debated. Previously, ceriantids were thought to be most closely allied to antipatharians, based on a resemblance between antipatharian polyps and cerianthid larvae; but differences between the two groups are so great that this relationship is no longer considered tenable. Actiniarians, scleractinians, and corallimorpharians share many morphological features, including details of mesentery and tentacle arrangement. The position of the Zoanthidea is obscure; these animals have some features reminiscent of rugosans, but because nothing is known about the soft tissue morphology of the latter, and the former do not form the skeletons, which are all that is known of rugosans, it is impossible to determine if this resemblance is phylogenetically significant. *See* PHYLOGENY.

Morphology. As is typical of cnidarian polyps, at the distal end of the cylindrical column is the oral disc (Fig. 1). The oval or slitlike mouth at its center, which opens into the actinopharynx, is surrounded by the tentacles, each of which arises from the space between two mesenteries, so its cavity is part of the coelenteron. The eight tentacles of an octocoral are arrayed in a single cycle, as are the six in most antipatharians; those of a cerianthid are arrayed in two cycles, one immediately around the mouth and one at the margin of the oral disc; in other hexacorals, the tentacles may cover all or only part of the oral disc, and be arrayed radially or in cycles. In colonial anthozoans, the proximal (basal) end of the body is connected to other polyps or to a common mass of tissue from which other polyps also arise. In solitary anthozoans, the base may form a pedal disc (Fig. 1), by which the animal attaches to something solid, the base may be more or less pointed or bulbous and used for burrowing into and anchoring in soft sediments.

The coelenteron of an anthozoan is filled with seawater, which, in combination with the animal's musculature, serves as a hydrostatic skeleton. Water is driven into the coelenteron largely through the siphonoglyphs (Fig. 2) by the action of the cilia that distinguish these structures. The column has ectodermal longitudinal and endodermal circular muscles by means of which the animal shortens and narrows, respectively; the tentacle muscles are similarly arrayed. Each mesentery typically has longitudinal muscles along one side and transverse ones along the other.

The skeleton of a cnidarian, if present, functions only to support and protect the animal—it does not serve movement. Actiniarians, ceriantharians, and corallimorpharians do not form skeletons; zoanthideans of many species incorporate debris, such as sand grains and sponge spicules, into the mesoglea of the column wall, which provides a sort of adventitious skeleton. Mineralic skeletons, produced by all octocorals and scleractinians, are made of calcium carbonate in the crystal form aragonite. The scleractinian calcareous skeleton consists of a cup around each polyp, with radiating septa that parallel the mesenteries; in a colony, cups of adjacent polyps may share walls, or the cups may be separate.

Most octocorals form sclerites (sometimes termed spicules) that are commonly shaped like needles, discs, or rods and are embedded in the mesoglea (gelatinous layer between the ectoderm and endoderm) of living tissue. The sclerites of a helioporacean colony are fused into a solid skeleton that appears similar to scleractinian corals except that it is blue; the sclerites of the organ-pipe coral (order Alcyonacea) are fused into long red parallel tubes linked by transverse platforms. The color of these skeletons is imparted by iron salts, which scleractinian skeletons lack and so are white. A colony of antipatharians and some alcyonaceans, such as the sea fans and sea whips, is supported by an internal organic skeleton that may be calcified to some degree; the colony of some pennatulaceans has an internal stiffening rod of organic material.

Physiology. All cnidarians are carnivorous, and most capture large prey with the tentacles, which are richly supplied with stinging cells called nematocysts and, in hexacorals, with spirocysts; in dimorphic (or trimorphic) species of octocorals, polyps of only one type feed. The tentacles deliver the prey to the mouth and it is swallowed (through the actinopharynx). Small planktonic prey enter the coelenteron, with water pumped in to support the hydrostatic skeleton. Filaments borne along the free edges of many or all mesenteries, which consist largely of gland cells and nematocysts, enwrap prey items to effect digestion. Undigested food is egested through the single body opening, the mouth. Many shallow-water anthozoans, particularly tropical ones, contain zooxanthellae, microscopic photosynthetic organisms that have a symbiotic relationship with marine organisms of several kinds, and live within endodermal cells of anthozoans. Some of the carbon fixed by photosynthesis of the zooxanthellae is leaked to the anthozoan host; some of the oxygen generated in the photosynthetic process is presumably also available to the host. Reciprocally, nitrogenous wastes and carbon dioxide from the animal may be used by the photosymbionts. Like all cnidarians, anthozoans have no specialized structures for respiration and excretion, which are effected through the entire body surface.

Although sedentary, some anthozoans (especially actiniarians and pennatulaceans) are capable of limited locomotion as adults. The nerve net of an anthozoan has no aggregations of nerve cells or specialized nervous structures. Otherwise, however, the nerves function essentially as they do in other animals. Members of a colony may be poorly or well integrated neurologically. Some pennatulaceans and zoanthids are bioluminescent. *See* BIOLUMINESCENCE.

Reproduction and development. Anthozoans of all species can undergo sexual reproduction (although not all polyps of a colony may be able to do so). Some species of anthozoans are gonochoric (have separate sexes) whereas some are hermaphroditic; hermaphroditism can be serial (in which the animal changes from one sex to the other during its lifetime) or simultaneous (in which the animal produces eggs and sperm at the same time), and some hermaprodites are self-fertile. Gametes are formed in endodermal cells of the mesenteries and sink into the mesenterial mesoglea to mature; when ripe, the gametes break through the mesoglea and endoderm surrounding them, to be freed into the coelenteron. Typically they are then released (through the actinopharynx and mouth) into the surrounding water, where fertilization and larval development occur. The zygote develops into either a ciliated swimming larva, the planula, or directly into a young polyp. In some species, only sperm are released into the water, whence they are taken into a female and the eggs are fertilized; some or all of larval development occurs in the female's coelenteron or on the surface of her body.

Many, but not all, anthozoan species can undergo asexual reproduction by a variety of processes, including budding (a term that is not synonymous with asexual reproduction), fission, and laceration. The products of asexual reproduction are genetically identical to one another, thus constituting a clone;

if the clonemates remain physically attached to one another, the multipolyp result is termed a colony. Asexual reproduction requires regenerative powers to reconstitute the missing body parts; members of species that do not undergo asexual reproduction typically have poorer regenerative powers than those that do, and some anthozoans, while able to heal, cannot regenerate at all. *See* REGENERATION (BIOLOGY).

Ecology. Anthozoans occur in all oceans of the world at all depths, including trenches. They can be ecologically very important, and they dominate the substratum in some habitats. Since most anthozoans require full or nearly complete salinity, and are intolerant of sedimentation and desiccation, they are poorly represented in estuaries and turbid environments, and high in the intertidal zone. Many groups are most diverse in the shallow tropics, but all groups (including scleractinian corals) occur in cold and deep water. [Deep-water scleractinians, however, do not live deeper than 5–6 km (16,404 ft–19,685 ft), probably due to the decrease in concentration of calcium carbonate in seawater with depth.]

Because species of coral that build shallow-water reefs are adversely affected by low temperature, coral reefs are confined to tropical and subtropical regions. Nearly all species of these reef-forming corals contain zooxanthellae, so tropical coral reefs are also confined places that can be reached by sufficient sunlight for photosynthesis [thus not at depths much greater than about 100 m (330 ft) or in caves]. Deep-water coral reefs differ from shallow tropical reefs: they are not latitudinally confined; the diversity of species is much lower; octocorals are relatively more important in forming them; and the scleractinians have less robust skeletons than their tropical reef-forming counterparts (which seems to be related to their lack of zooxanthellae). *See* REEF.

Daphne G. Fautin

Fossil record and evolution. Fossil anthozoans are known from the late Precambrian (during the Ediacaran geologic period, about 600 million years ago) and from rocks of every later geologic period. Fossilized soft polyps are extremely rare, but several groups of anthozoans with a mineralic external skeleton were well preserved and have good (and even excellent) records. The Paleozoic Rugosa and Tabulata are common in rocks of Middle Ordovician to Permian age, and the Mesozoic-Cenozoic Scleractinia are common from the Middle Triassic to the present. Rugosans and tabulates were important reef builders in the Middle Silurian to early Late Devonian, and scleractinians were important reef builders from the Jurassic through the Cretaceous, and again from the Miocene to the present. Fossil Octocorallia are represented by impressions from the Ediacaran period, and although sclerites are recognized sporadically through time, the fossil record is not good enough to trace the evolutionary pattern within this group or its relationship to the Hexacorallia. *See* EDIACARAN BIOTA; GEOLOGIC TIME SCALE; RUGOSA; TABULATA.

The relationship among the classes of Cnidaria had been uncertain until the discovery that the mitochondrial deoxyribonucleic acid (DNA) of anthozoans is circular, as it is in all non-cnidarians, whereas that of cubozoans, hydrozoans, scyphozoans, and staurozoans is linear. Thus the non-anthozoans, now commonly referred to as the Medusozoa, form a clade that is derived (evolutionarily advanced), in this feature relative to Anthozoa, so Anthozoa is the sister group to all other cnidarians. *See* ANIMAL EVOLUTION; CUBOZOA; HYDROZOA.

William A. Oliver, Jr., Daphne G. Fautin

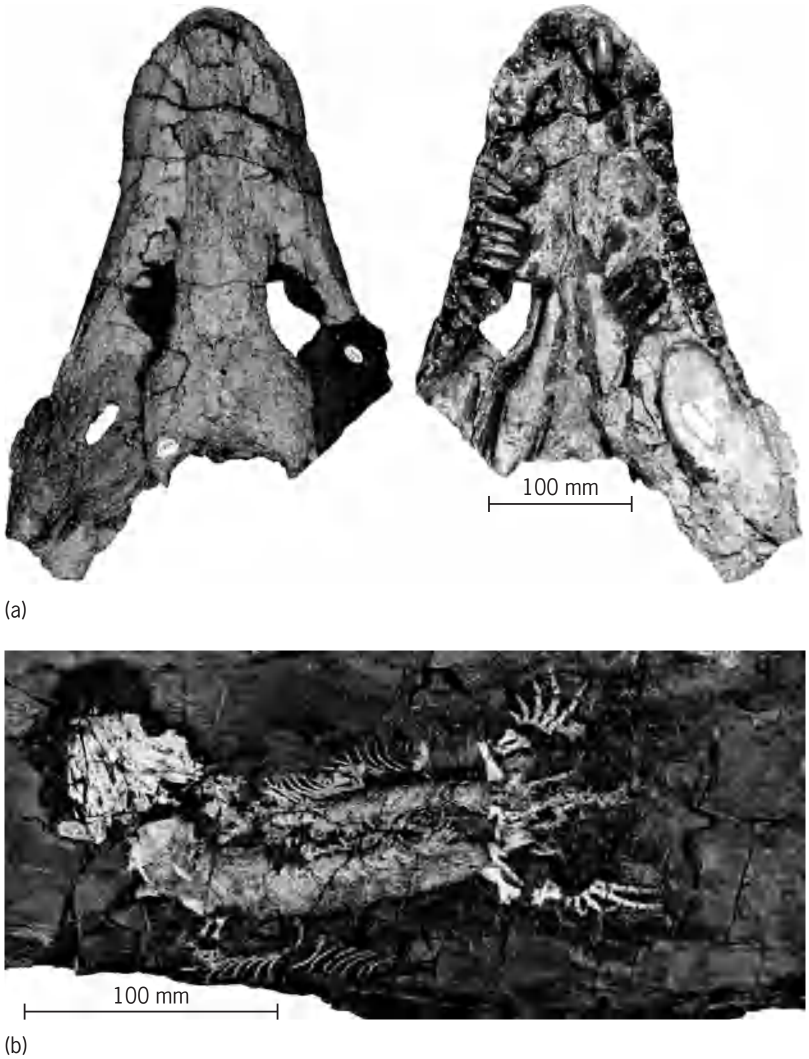
Bibliography. D. F. Dunn, Cnidaria, pp. 669–706 in *Synopsis and Classification of Living Organisms*, vol. 1, S. P. Parker (editor-in-chief), McGraw-Hill, New York, 1982; L. H. Hyman, *The Invertebrates: Protozoa Through Ctenophora*, vol. 1, McGraw-Hill, New York, 1940; J. W. Wells and D. Hill, Anthozoa—general features, pp. F161–F165 in *Treatise on Invertebrate Paleontology*, vol. F: *Coelenterata*, Geological Society of America and University of Kansas Press, Lawrence.

Anthracosauria

An order of Paleozoic tetrapods that arose in the Early Carboniferous (Mississippian) (345 million years ago) and lasted until the Early Triassic (about 245 mya). The term anthracosaurs (meaning "coal reptiles") encompasses a range of genera that may not all be close relatives; rather, they form a stem group (members of the lineage leading to amniotes, but not amniotes themselves) possibly related to amniotes (reptiles, birds, and mammals). They share a number of features, such as the pattern of skull roof bones and vertebral construction, that distinguish them from temnospondyls, a larger group of Paleozoic tetrapods that are more closely related to modern amphibians. Anthracosaurs also share a number of features that may be primitive for tetrapods, including a "skull table" unit that is only loosely attached to the cheek, and a palate in which bones almost meet along the midline.

Anthracosaurs include both terrestrial and aquatic forms, with a subgroup, the embolomeres, being large, long-bodied, crocodilelike fish eaters. Embolomeres include genera such as *Pholiderpeton* and *Anthracosaurus* (**illus.** *a*) from the Carboniferous of Great Britain and *Archeria* from the Early Permian (about 290 mya) of the United States. Embolomeres are characterized by having two ring- or horseshoe-shaped parts forming the central portion (centrum of their vertebrae), and a prong or "horn" on the tabular bone, at the back of the skull (illus. *a*).

The earliest anthracosaurs, *Silvanerpeton* (illus. *b*) and *Eldeceeon*, are from Scotland and are Viséan in age (about 340 million years old). *Eldeceeon* was terrestrial, but *Silvanerpeton* was probably partially aquatic. Another mid-Carboniferous (about 320 mya) form, *Eoherpeton*, was also probably terrestrial, though it is known only from a few incomplete specimens. *Gephyrostegus*, from the Late Carboniferous (about 300 mya) of eastern Europe, is a terrestrial form which at one time was thought to be a close ancestor of reptiles, though its position among the anthracosaurs is



Anthracosaur features. (*a*) *Anthracosaurus russelli*, one of the largest and fiercest anthracosaurs, from Great Britain. The left view shows the skull roof, with the central "skull table" bearing a short tabular horn; the right view shows the palate with enormous teeth. (*b*) *Silvanerpeton miripedes*, one of the earliest anthracosaurs, from Scotland. The head is flattened, and the body has been "kippered (slit open)," with the ribs separated from the vertebral column. The dermal gastralia (scalelike structures) formed a protective armour (seen well here) down the middle of the animal.

disputed. The most recent anthracosaurs (last in their line) have been found in Russia, where some embolomere-like forms have reverted to terrestrial habits. *See* AMNIOTA; AMPHIBIA; TEMNOSPONDYLI; TETRAPODA. Jennifer A. Clack

Bibliography. M. J. Benton, *Vertebrate Palaeontology*, 3d ed., Blackwell, Oxford, 2005; R. L. Carroll, *Vertebrate Paleontology and Evolution*, W. H. Freeman, New York, 1987; J. A. Clack, *Gaining Ground: The Origin and Early Evolution of Tetrapods*, Indiana University Press, Bloomington, 2002; T. R. Smithson, Anthracosaurs, in *Amphibian Biology*, vol. 4, pp. 1053–1063, ed. by H. Heatwole and R. L. Carroll, Surrey Beatty, Chipping Norton, NSW, 2000.

Anthrax

An acute, infectious worldwide zoonotic disease caused by the spore-forming bacterium *Bacillus anthracis* to which most animals, especially grazing herbivores, are susceptible. In natural conditions, human infections (predominantly cutaneous) usually result from contact with infected animals or contaminated animal products, such as hides or wool. The awareness of the use of *B. anthracis* as a bioterrorist weapon has considerably increased. Anthrax is endemic as a zoonosis in many areas of Africa, Asia, and the Americas, where spores can lie dormant in the soil for many years and commonly affect grazing animals such as sheep, cattle, and goats. *See* ZOONOSES.

Bacteriology. *Bacillus anthracis*, the causative agent of anthrax, derives its name from the Greek word for coal, *anthraxis*, because of the black skin lesions it causes. *Bacillus anthracis* is the only obligate pathogen within the genus *Bacillus* and is a large gram-positive, aerobic, spore-bearing bacillus, 1–1.5 × 3–10 micrometers in size. It is nonmotile and nonhemolytic (does not destroy red blood cells) on horse or sheep's blood agar, grows at 37°C, and forms typical white colonies, which are characteristically tacky on teasing with a loop. The oval spores

(0.75 × 1 μm) resist drying, cold, heat, and disinfectants and can survive in soil for decades.

Spores do not form in host tissues, unless the infected body fluids are exposed to air that contains a lower level of carbon dioxide. Provided no antibiotics have been prescribed, *B. anthracis* is readily isolated from clinical specimens such as skin or blood cultures. In environmental samples, because of the presence of other nonpathogenic *Bacillus* spp., a selective medium such as polymixin lysozyme ethylenediaminetetraacetic acid (EDTA) thallus acetate agar is recommended for isolating specimens.

R. C. Spencer

Pathogenesis. *Bacillus anthracis* possesses three virulence factors: lethal toxin, edema toxin, and a poly-D-glutamic acid capsule. Lethal toxin is composed of two proteins, lethal factor (LF, 82 kilodaltons, a metalloprotease) and protective antigen (PA, 63 kDA), complexed at a ratio of seven protective antigens per lethal factor. The protective antigen is produced by the anthrax bacillus at a molecular weight of 83 kDA, but must be cleaved by either serum or target cell surface proteases to 63 kDa before it complexes with lethal factor to form lethal toxin. The edema toxin is composed of edema factor (EF, 89 kDa, adenylate cyclase) and protective antigen, and it is believed to complex in a manner similar to that seen for lethal toxin. Protective antigen plays a central role in that it is required for transport of lethal factor and edema factor into host target cells. Macrophages appear to be the primary host target cells for lethal toxin, whereas neutrophils appear to be the target cells for edema toxin in addition to other cells involved in edema formation. The third virulence factor is the capsule, which inhibits phagocytosis through its negatively charged poly-D-glutamic acid composition.

All three toxin components are encoded by a plasmid, pXO1, whereas the genes required for capsule synthesis (cap B, cap C, and cap A) are encoded by the smaller pXO2 plasmid. The pXO2 plasmid also encodes a depolymerase enzyme, which cleaves the capsule to form small glutamic acid peptides that also appear to play a role in virulence. Strains lacking either or both plasmids are avirulent, such as the veterinary vaccine Sterne strain, which lacks the pXO2 plasmid.

John W. Ezzell

Infection in humans occurs after introduction of the spore through a break in the skin (cutaneous anthrax) or entry through mucosa (gastrointestinal anthrax). After ingestion by macrophages at the site of entry, germination of the vegetative form then occurs, followed by extracellular multiplication, together with the production of capsule and toxins. In inhalation anthrax, spores of 1–2 μm in diameter are inhaled and deposited in the alveolar spaces from where they are transported by the lymphatics to the mediastinal lymph nodes, where they cause mediastinal lymphadenopathy and hemorrhage (that is, enlarged mediastinal lymph glands and bleeding within them). Vegetative bacteria then spread through blood and lymph where, if unchecked, they cause septicemia (blood poisoning). In turn, large amounts of exotoxins are produced, which are responsible for the overall symptoms and death.

Clinical features. The symptoms of anthrax infection in humans are dependent on the route through which the infection is contracted:

Cutaneous anthrax. Cutaneous anthrax is by far the most common form, accounting for over 95% of naturally occurring cases. It is caused by direct contact with spores, especially via open wounds. Following an incubation period of 1–3 days, localized itching precedes the appearance of a papular lesion that turns vesicular; 2–6 days later the lesion develops a black eschar (dry crust), giving anthrax its name.

Inhalation/pulmonary anthrax. Inhalation/pulmonary anthrax follows infection via the lungs, where the spores can easily cross into the blood at the level of the alveoli. Initially, infection resembles the common cold and is characterized by nonspecific symptoms including fever, nonproductive (dry) cough, malaise, and fatigue. Chest x-ray at this stage often shows mediastinal widening (due to mediastinal lymphadenopathy) and pleural effusions (abnormal accumulation of fluid in the pleural space, the area between the membranes lining the lungs and the chest cavity). Two to five days after initial symptoms there is an abrupt onset of marked pyrexia (fever) and severe respiratory depression often accompanied by dyspnea (labored breathing), cyanosis (bluish tinge due to lack of oxygen), and stridor (peculiar, harsh, vibrating sound produced during respiration). Untreated, patients will go into shock and die within 24–36 h.

Gastrointestinal anthrax. Gastrointestinal anthrax is rare and usually follows consumption of contaminated meat. It is characterized by acute inflammation of the gastrointestinal tract. Initial signs of nausea, anorexia, vomiting, and pyrexia are followed by abdominal pain, hematemesis (vomiting of blood), and severe diarrhea, which is often bloody. Bacteremia (presence of bacteria in the blood) may develop after 2–3 days, which if untreated is usually fatal.

Person-to-person transmission. Person-to-person spread of anthrax is very rare, and contacts need not be treated unless they have been exposed to the same source of infection. Although direct exposure to exudates (excretions) from cutaneous anthrax lesions may result in secondary cutaneous infection, there have been no known cases of person-to-person transmission of inhalation disease.

Treatment. Antibiotic therapy is the recommended treatment for anthrax. Route of administration and type of medication vary depending on the type of disease, whether diagnosis has been confirmed, and organism sensitivity.

Inhalation and gastrointestinal anthrax. When the diagnosis of inhalation or gastrointestinal anthrax is suspected but not confirmed, it may be necessary to start empirical treatment to protect against the possibility of anthrax. However, in these circumstances, it will also be necessary to treat concurrently for other causes of acute respiratory illness. Recommended initial treatment for inhalational and gastrointestinal anthrax is with intravenous ciprofloxacin (with a change to oral delivery when appropriate) for 60 days. Treatment can be changed to penicillin if the organism is found to be sensitive.

Cutaneous anthrax. Treatment for cutaneous anthrax should be initiated with oral ciprofloxacin 500 mg twice daily for 7 days. This can be changed to penicillin if the organism is found to be sensitive. If there is suspicion of deliberate release, treatment may need to be continued for up to 60 days to protect against inhalation anthrax, which may have been acquired concurrently.

Prophylactic treatment. In the event of a known exposure to anthrax spores, antibiotic prophylaxis should be initiated as soon as possible with oral ciprofloxacin (a fluoroquinolone) or with doxycycline (a tetracycline) if fluoroquinolones are unavailable or contraindicated. Prophylaxis should continue until *B. anthracis* exposure has been excluded. If exposure is confirmed, prophylaxis should continue for 60 days. During this period, no special precautions are required for exposed persons; however, they should receive an anthrax information sheet and be instructed to seek medical attention immediately in the event of any suspicious symptoms.

Pediatric use of fluoroquinolones and tetracyclines can be associated with adverse effects that must be weighed against the risk of developing a serious disease. If *B. anthracis* exposure is confirmed, the organism must be tested for penicillin susceptibility. If susceptible, exposed children may be treated with oral amoxycillin. *See* ANTIBIOTIC.

Vaccination. Vaccination is the most cost-effective form of mass protection. Although the first anthrax animal vaccine was developed by Pasteur in 1881, human vaccines did not emerge until the middle of the twentieth century.

Animal vaccines. Pasteur's initial vaccine had problems with declining potency and variability in immunogenicity (ability to induce an immune response), which led to the search for a more stable and effective vaccine. Sterne developed an attenuated spore vaccine, based on an avirulent nonencapsulated strain derived from a subculture from a case of bovine anthrax. This has proved to be extremely safe and effective, requiring little modifications. Although extremely effective, repeated vaccinations are required for long-term protection, because a single dose provides immunity for only about one year.

Human vaccines. Vaccination of humans with live spores, either by scarification or subcutaneous injection, was developed in the former Soviet Union and China. In Russia a strain analogous in its derivation to the Sterne strain is used.

The vaccine developed in the United States is a filtrate of a bovine strain of *B. anthracis* V770-NPI-R. It was licensed in 1972 for administration to those in at-risk occupations. It is not licensed for use in children or pregnant women, and at present is not available for civilian use. The vaccine is licensed to be given in a six-dose series over an 18-month period. The current vaccine from the United Kingdom consists of a filtrate of an aerobic supernatant from the nonencapsulated Sterne strain of *B. anthracis*. The UK vaccine was introduced for workers in at-risk occupations in 1965, and licensed for general human use in 1979. At present, on empirical grounds, boosters are administered 6 months after the initial series of three doses (at 0, 3, and 6 weeks), and annually thereafter.

Future vaccines. Considerable effort has been expended in the development of next-generation vaccines to meet current licensing criteria, but with improved safety and efficacy performance characteristics. The favorites are subunit vaccines that contain whole-length (83,000 kDa) recombinant protective antigen. *See* VACCINATION.

R. C. Spencer

Bibliography. L. Baillie and T. D. Read, *Bacillus anthracis*: A bug with attitude, *Curr. Opin. Microbiol.*, 4:78–81, 2001; T. C. Dixon et al., Anthrax, *N. Engl. J. Med.*, 341:815–826, 1999; J. Guilleman, *Anthrax: The Investigation of a Deadly Outbreak*, University of California Press, Berkeley, 1999; R. C. Spencer, *Bacillus anthracis*, *J. Clin. Pathol.*, 56:182–187, 2003; M. N. Swartz, Recognition and management of anthrax—an update, *N. Engl. J. Med.*, 345:1607–1610, 2001; P. C. B. Turnbull, Definitive identification of *Bacillus anthracis*—a review, *J. Appl. Microbiol.*, 2:237–240, 1999.

Anthropology

The observation, measurement, and explanation of human variability in time and space. This includes both biological variability and the study of cultural, or learned, behavior among contemporary human societies. These studies are closely allied with the fields of archeology and linguistics. Studies range from rigorously scientific approaches, such as research into the physiology, demography, and ecology of hunter-gatherers, to more humanistic research on topics such as symbolism and ritual behavior. *See* ARCHEOLOGY; PHYSICAL ANTHROPOLOGY.

Anthropology lacks a unified theory comparable to neo-Darwinian evolution in the biological sciences and is characterized, instead, by a wide variety of subfields that analyze and integrate studies of human behavior in different ways. Social-cultural anthropology examines the various ways in which learned techniques, values, and beliefs are transmitted from one generation to the next and acted upon in different situations. Most studies stress the historical development and internal structure and workings of particular cultural traditions, and anthropologists have amassed detailed bodies of documentation on different human societies, ranging from small, nomadic hunter-gatherer groups like the Eskimo or Australian aboriginals to complex, sedentary societies like India, Iran, or Japan. Such ethnographic studies are carried out within a relativistic framework which assumes that one cannot evaluate the behavior and attitudes of one society by applying the standards of another. This attitude has been a useful corrective to earlier attempts to judge non-Western behavior by Western standards. Given this view, social-cultural anthropologists strive, by various means, to achieve a reliable understanding of each cultural system in its own terms, with as little distortion as possible due to ethnocentrism.

Significant, too, within social-cultural anthropology are cross-cultural studies that seek to identify essential structural or behavioral properties of human society. While anthropologists like Leslie White have proposed general laws of cultural evolution, others like Claude Lévi-Strauss have argued for universal principles of human cognition by examining how people in different societies classify natural and social phenomena by means of logical systems based upon the principle of binary opposition. Modern scholars have sought to identify universal patterns of symbolic behavior and belief, and there are other social-cultural anthropologists actively testing these kinds of propositions in particular cases.

Increasingly, too, social-cultural anthropologists have applied their training and skills to issues of contemporary importance such as economic development in third world countries, public policies affecting ethnic minorities, and changes arising from contact between different societies (especially Western and non-Western ones). Sometimes referred to as applied anthropology, such studies are often made in situations where conflicting social values or expectations may arise. A cogent example is a series of studies into the social consequences arising from technical and economic changes due to the so-called Green Revolution, involving the introduction of new and more productive strains of rice into traditional rice-growing societies in southern Asia. This nonacademic area of social-cultural anthropology, where application is emphasized over basic research, assumed a larger role in social planning and policymaking of all kinds, ranging from the actions of individual corporations (such as Western oil companies operating in traditional Moslem societies in the Middle East) to national governments.

Cultural linguistics is closely allied with both the goals and methods of social-cultural anthropology, especially with respect to the way in which linguists strive for a reliable understanding of how each different language works according to its own sound system (phonology) and grammatical structure. Linguists have produced a growing body of valuable sources, including dictionaries, grammars, and texts of particular languages and comparative analyses, that enable them to infer historical relationships between different languages that share common ancestry. Such controlled, empirical studies, free of ethnocentric bias, also provide finely tuned views of verbal and behavioral interaction in the context of particular situations, such as non-Western artistic performances or rituals. *See* PSYCHOLINGUISTICS.

There has also been a developing tendency in anthropology toward integration of different subfields. For example, ethnoscience is a subject in which anthropologists apply approaches derived from linguistics to understand the grammatical structure and manipulation of cognitive perceptions by people in different societies of such things as color, weather, and biotic environment. Studies range from traditional Philippine farmers' plant taxonomies and agricultural practices to long-distance sailing and navigation by Micronesian islanders in outrigger canoes. Another growing subfield is ethnoarcheology, in which observations of material behavior (especially discard) in contemporary societies are used to interpret the archeological remains of prehistoric cultures. Modern anthropology is characterized by its breadth and diversity of approaches to the study of variability in human behavior. Richard A. Gould

Bibliography. M. Douglas, *Natural Symbols,* 1972; M. Harris, *The Rise of Anthropological Theory,* 1968; D. Levinson and Martin J. Malone, *Toward Explaining Human Culture: Critical Review of the Findings of Worldwide Cross-Cultural Research,* 1980; C. Lévi-Strauss, *The Savage Mind,* 1967; E. F. Moran, *Human Adaptability,* 1979; J. E. Pfeiffer, *The Emergence of Man,* 1978; E. B. Ross (ed.), *Beyond the Myths of Culture: Essays in Cultural Materialism,* 1980.

Anthropometry

The systematic quantitative representation of the human body. Anthropometric techniques are used to measure the absolute and relative variability in size and shape of the human body.

Instrumentation. Depending on the objective, anthropometric techniques involve instrumentation that may include weighing scale, anthropometer, skinfold calipers, body volume tanks, and bioelectrical impedance analyzers that are used for measurements of body size and body composition. Similarly, radiographic instruments and x-ray scanners such as dual-energy-ray absorption meters and ultrasound densitometers are used for quantifying cortical bone density, bone mass, subcutaneous fat density, and lean body mass.

Quality control. Anthropometry follows a rigorous set of guidelines that include standardization of the measurement techniques, uniform landmarks, and establishing conditions of the measurements. Despite its simplicity, it requires a great deal of quality control to ensure minimal error of measurements. Such an objective is usually obtained through meticulous training in the technique that results in replicable measurements.

Applications. In biological anthropology and human paleontology, anthropometry is the technique of choice for quantifying variability and relationship of fossils and extant populations. Anthropometric measurements of the head, face, and long bones are also used in analyzing fossil taxa (using measurements from radiographs). The mathematical approaches are used in describing the size and proportions of various fossil hominids. For example, through measurements of long bones it is known that the earliest human ancestors such as *Australopithecus afarensis* were about 3.5–5 ft (1.05–1.5 m) tall. In the same vein, it has been established that *Homo erectus* and Neandertals were as tall as modern-day humans, contrary to previous assumptions. *See* ANIMAL GROWTH; BIOMETRICS.

Anthropometry is the most universally applicable, inexpensive, and noninvasive method available to assess the nutritional history throughout life. It has been used to assess and predict the health, nutritional history, welfare, performance, and survival of nonindustrialized and industrialized societies. For example, since fat is the main form of energy storage, and body muscle is composed largely of protein, anthropometric measurements of body composition provide indirect estimates of energy and protein reserves of the body. These reserves are sensitive to changes in dietary intake. Reserves can be depleted during chronic malnutrition, resulting in muscle wasting, while during overnutrition reserves can grow, resulting in obesity. *See* NUTRITION.

To achieve such objectives, anthropologists and population biologists have developed various references that can be used as base lines for expressing the absolute and relative deviation from the average. Techniques of data analysis include the expression of individual values in the form of Z scores (the individual value minus the reference mean for the age and sex, divided by the corresponding standard deviation). This approach permits the investigator to express the measurements in terms of Z score units from the mean. Another approach involves expressing individual values in the form of percentiles placement. For this purpose, the investigator needs to compare the individual value to the percentile ranges given in the anthropometric standard. Thus, an individual measurement may be expressed as being either close to the 50th percentile or above or below the 95th or 5th percentile.

Anthropometry is also essential to the field of forensics, specifically forensics anthropology, which is concerned with the relationship between medicine and the law. Forensic anthropologists (so-called bone detectives) make extensive use of anthropometry in human identification, whether for isolated cadavers, commingled remains, victims of mass disasters, or genocide victims. As applied to metacarpal and phalangeal (reterning to bones of the hand) patterns, anthropometry helps in the diagnosis of congenital malformation syndromes (dysmorphogenesis states), even in fetuses. Anthropometric measurements of the head and face are also extensively employed in orthodontic diagnosis, in treatment planning, and following orthodontic treatment. Measurements made from cephalometric radiographs also serve in the identification of syndromes. Further extensions of cephalometry measurement of the living human head), in three dimensions (cartesian anthropometry), are used in sculpting head forms not just for equipment design but for use in the plastic and reconstructive surgery of accident victims. Forensic anthropologists are involved in the identification of human remains using techniques that include direct skeletal measurements, allowing the identification of the age at death, sex, and cause of death, and deduction of the life history of the deceased. For example, identification of healed bone fractures, age, and sex of Neandertals has permitted anthropologists to infer the life-span and hunting techniques used. Analysis of the skeletal remains revealed that these fossil humans did not live beyond the age of 50 years and practiced manual (hand-to-hand) hunting of their game, resulting in fractures similar to those suffered by rodeo cowboys. *See* NEANDERTALS.

The use of anthropometry for the design of clothing, equipment of all sorts, and interiors is relatively new in the history of humankind. For example, through anthropometric techniques to establish human dimensions, gas masks, oxygen masks, dust masks, and respirators as well as military helmets were designed. Anthropometry has enabled the production of clothes that accurately match the various sizes of the population. The dimensional standards used for sports equipment and military applications have been designed to match the distribution of size in the population as whole. Hence, anthropometry plays an essential role in the armed forces, by providing information about the necessary dimensions for military equipment of all sorts to accommodate individuals in the military service. Similarly, the equipment used in space demands knowledge of human dimensions. Anthropometry has evolved from the manual to automated approach, whereby measurements are entered directly into the computer for data processing.

Anthropometry plays an essential role in all areas of human endeavor concerned with the relative and absolute quantification of the human body. It serves a profound function in all aspects of science and human activity. *See* ANTHROPOLOGY; PHYSICAL ANTHROPOLOGY. A. Roberto Frisancho

Bibliography. W. M. Bass, *Human Osteology: A Laboratory and Field Manual of the Human Skeleton*, 4th ed., 1995; B. Bogin, *Pattern of Human Growth*, 1988; C. Bouchard, R. Malina, and L. Perusse, *Genetics of Fitness and Physical Performance*, 1997; P. B. Eveleth and J. M. Tanner, *Worldwide Variation in Human Growth*, 1976; A. R. Frisancho, *Anthropometric Standards for the Assessment of Growth and Nutritional Status*, 1990; S. Garn, Physical anthropology: Methods, *Med. Phys.*, 3:25–30, 1960; T. Lohman, A. Roche, and R. Martorell (eds.), *Anthropometric Standardization Reference Manual*, 1988; F. Spencer (ed.), *History of Physical Anthropology*, vol. 1, 1997.

Antibiotic

The original definition of an antibiotic was a chemical substance that is produced by a microorganism and, in dilute solutions, can inhibit the growth of, and even destroy, other microorganisms. This definition has been expanded to include similar inhibitory substances that are produced by plants, marine organisms, and total- or semi-synthetic procedures.

History. The first to apply antibiotic therapy, over 2500 years ago, were the Chinese. They were aware of the therapeutic properties of moldy soybean curd applied to infections, and used this material as standard treatment. L. Pasteur and J. F. Joubert recognized the clinical potential of antibiosis in 1877. They observed that large quantities of anthrax bacilli could be given to an animal without adverse effects, provided that ordinary bacteria were given at the same time. The modern era of antimicrobial therapy was launched in the mid-1930s with the discovery of prontosil; sulfamethoxazole, a derivative of this synthetic compound, is still in clinical use (see **table**). Penicillin, discovered by A. Fleming in 1928, was not developed until the early 1940s because of the low potency and high instability of the crude material. The potency and therapeutic potential of penicillin was recognized by H. Florey and E. Chain. World War II provided the impetus for the Anglo-American research effort that led to the large-scale production of penicillin and its development as a systemic therapeutic agent. In the mid-1940s, cephalosporin C was discovered by G. Brotzu and streptomycin by S. Waksman and his associates. The latter discovery was the result of a well-planned, systematic, and intensive search for antibiotics in soil actinomycetes and fungi. Subsequent observations on the mode of action of bacitracin, penicillin, polymyxin, gentamycin, chloramphenicol, tetracycline, and amphotericin revealed the importance and therapeutic potential of antibiotics and added momentum to the search for new ones. Since the 1940s, thousands of antibiotics have been isolated and identified; some have been found to be of value in the treatment of infectious disease. They differ markedly in physicochemical and pharmacological properties, antimicrobial spectra, and mechanisms of action.

Biosynthesis. Penicillin is produced by strains of the fungus *Penicillium notatum* and *P. chrysogenum*. Most of the other antibiotics in clinical use are produced by actinomycetes, particularly streptomycetes (natural antibiotics) [see table]. Other antibiotics, such as nalidixic acid and clotrimazole, are produced by chemical synthesis (synthetic antibiotics). Based on structure, the major antibiotic classes are the β-lactams (penicillins and cephalosporins), aminoglycosides, macrolides, tetracyclines, quinolones, rifamycins, polyenes, azoles, glycopeptides, and polypeptides.

Manufacture. The key step in the production of natural antibiotics is a fermentation process. Strains of microorganisms, selected by elaborate screening procedures from randomly isolated pure cultures, are inoculated into sterile nutrient medium in large vats and incubated for varying periods of time. Both cell growth and antibiotic production, which are influenced by culture conditions such as composition of the medium, temperature, aeration, and pH, are monitored during this period.

Different strains of a single microbial species may differ greatly in the amounts of antibiotics they produce. Strain selection is thus the most powerful tool in effecting major improvements in antibiotic yield. In addition, variations in culturing conditions often markedly affect the amount of antibiotic that is produced by a given strain. *See* BACTERIAL PHYSIOLOGY AND METABOLISM; FERMENTATION.

The production of an antibiotic in quantity by fermentation and its subsequent recovery from the culture medium usually require 5–8 days. The process can be shortened sometimes by establishing the optimal conditions for growth and antibiotic production. Paper, thin-layer, and high-pressure liquid chromatography, techniques designed to separate minute quantities of different substances, allow rapid detection of the antibiotic compounds. Fluorescence, ultraviolet, and infrared spectroscopy are often used to monitor antibiotic distribution in various fractions, obtained by such procedures as solvent extraction and column or high-pressure liquid chromatography. *See* CHROMATOGRAPHY; SPECTROSCOPY.

Chemical modifications of antibiotics produced by fermentation processes have led to semisynthetic ones with improved antimicrobial activity or pharmacological properties. For example, chemically modified penicillins, such as methicillin, are effective against bacteria resistant to the parent compound, penicillin G. Other penicillins, such as ampicillin, are active against a wider variety of microorganisms than the parent compound. Still others are better absorbed by the body or have greater chemical stability, more favorable tissue distribution and excretion, or reduced toxicity.

In the United States, the Federal Pure Food and Drug Act of 1906, as amended in 1938 and 1962, requires that all drugs for human use must be approved by the Food and Drug Administration (FDA) for purity, efficacy, and safety. Before an antibiotic is tested in humans, extensive microbiological, pharmacological, and toxicological studies are needed to support claims for efficacy and safety. The data are compiled and submitted to the FDA as part of an application for an Investigational New Drug. If the FDA standards are met, the antibiotic is approved for human clinical trials. The trials study the safety and pharmacokinetic profile of the new drug in healthy volunteers (phase I trials) and its efficacy in patients with different infectious diseases (phase II and III trials). Data from these trials are then submitted to the FDA to support a New Drug Application. If approved by the FDA, the antibiotic can be marketed for human use. Other countries have their own regulatory agencies, with most European nations bound by the European Economic Community's rules and regulations. *See* PHARMACEUTICALS TESTING.

Antimicrobial activity. In general, microorganisms are divided into bacteria, fungi, viruses, and protozoa. All four groups can cause infectious diseases in animals and humans, though the majority of infections are caused by bacteria. Most antibiotics are active against bacteria. Although for the proper treatment of serious infections cultures and antibiotic sensitivities are required, antibiotic therapy is often empiric, with etiology being inferred from the clinical features of a disease.

Bacteria. Bacteria are divided by a staining reaction into the gram positive and the gram negative; each

-s

group comprises a wide variety of different species. Gram-positive bacteria may be subdivided by another staining reaction into acid-fast and non-acid-fast bacteria. *See* STAIN (MICROBIOLOGY).

Staphylococci, pneumococci, and streptococci are the more common gram-positive organisms, while enterobacteria, *Pseudomonas*, and *Hemophilus* are the most common gram negative. Tubercle bacilli are the most important acid-fast organisms. Certain antibiotics, such as erythromycin and vancomycin, are effective only against gram-positive bacteria. Others, such as cephalosporins, quinolones, tetracyclines, and chloramphenicol, are effective against both gram-positive and gram-negative bacteria and are referred to as broad-spectrum antibiotics. *See* BACTERIA; MEDICAL BACTERIOLOGY.

Fungi. Pathogenic fungi may be divided on the basis of their pathogenicity into true pathogens, such as *Histoplasma*, and opportunistic pathogens, such as *Candida*, *Aspergillus*, and *Cryptococcus*. The opportunistic occur mainly in debilitated and immunocompromised patients. Although hundreds of antibiotics active against fungi have been isolated, clinically useful ones are amphotericin B, nystatin, griseofulvin and the azole antifungals. As broad-spectrum antifungals, amphotericin is active against systemic infections and nystatin against local infections. Griseofulvin is active against fungi that cause skin diseases. Azoles, which constitute perhaps the largest group of antifungals, are broad-spectrum and synthetic. *See* FUNGI; MEDICAL MYCOLOGY; OPPORTUNISTIC INFECTIONS.

Viruses. Viruses may be divided on the basis of nucleic acid content (RNA and DNA), morphology, and composition of the virus shell (enveloped and nonenveloped viruses). With some viruses that cause mild infections, such as the common-cold viruses (rhinoviruses), treatment is symptomatic. With others, such as the polio, smallpox (now eradicated), and hepatitis B viruses, the only way to prevent disease is by vaccination. With still other viruses (see table), antibiotics, mostly synthetic, are the appropriate treatment. Clinically useful antibiotics are ribavirin, acyclovir, and zidovudine, which are active against, respectively, respiratory, herpes, and human immunodeficiency viruses. *See* ANIMAL VIRUS; VACCINATION.

Protozoa. Protozoa may be divided, on the basis of the site of infection, into intestinal (*Entamoeba histolytica, Giardia lamblia*), urogenital (*Trichomonas vaginalis*), blood (*Plasmodium falciparum*), and tissue (*Pneumocystis carinii*, which was reclassified as a fungus). Protozoan diseases such as malaria, trypanosomiasis, and amebiasis are particularly common in the tropics, in populations living under poor housing and sanitary conditions. In the developed countries, *P. carinii* is the most important opportunistic pathogen, being associated almost exclusively with acquired immune deficiency syndrome (AIDS). Antibiotics active against protozoa include metronidazole, trimethoprim-sulfamethoxazole, and quinine. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); MEDICAL PARASITOLOGY; PROTOZOA.

Antitumor activity. The pioneering observation in 1952 on the antitumor activity of actinomycin sparked an intensive search for antitumor antibiotics in plants and microorganisms. Among the antibiotics used clinically against certain forms of cancer are daunorubicin, doxorubicin, mitomycin C, and bleomycin. Although antitumor antibiotics are significantly toxic to the patient, the lack of nontoxic therapeutic alternatives has required their continued use. *See* CANCER (MEDICINE).

Mechanism of action. Antibiotics active against bacteria are bacteriostatic or bacteriocidal; that is, they either inhibit growth of susceptible organisms or destroy them. Bactericidal activity may be necessary for the eradication of infections where the host immune system is impaired. On the basis of their mechanism of action, antibiotics are classified as (1) those that affect bacterial cell-wall biosynthesis, causing loss of viability and often cell lysis (penicillins and cephalosporins, bacitracin, cycloserine, vancomycin); (2) those that act directly on the cell membrane, affecting its barrier function and leading to leakage of intracellular components (polymyxin); (3) those that interfere with protein biosynthesis (chloramphenicol, tetracyclines, erythromycin, spectinomycin, streptomycin, gentamycin); (4) those that affect nucleic acid biosynthesis (rifampicin, novobiocin, quinolones); and (5) those that block specific steps in intermediary metabolism (sulfonamides, trimethoprim). Some antibiotics, such as sulfonamides, penicillins, cephalosporins, and quinolones, are specific inhibitors of bacterial enzymes. *See* ENZYME; SULFONAMIDE.

Antibiotics active against fungi are fungistatic (ketoconazole, fluconazole) or fungicidal (amphotericin, nystatin). Their mechanisms of action include (1) interaction with the cell membrane, leading to leakage of cytoplasmic components (amphotericin, nystatin); (2) interference with the synthesis of membrane components (ketoconazole, fluconazole); (3) interference with nucleic acid synthesis (5-fluorocytosine); and (4) interference with microtubule assembly (griseofulvin). *See* FUNGISTAT AND FUNGICIDE.

Pharmacological properties. For an antibiotic to be effective, it must first reach the target site of action on or in the microbial cell. It must also reach the body site at which the infective microorganism resides in sufficient concentration, and remain there long enough to exert its effect. The concentration in the body must remain below that which is toxic to the human cells. The effectiveness of an antibiotic also depends on the severity of the infection and the immune system of the body, being significantly reduced when the immune system is impaired. Complete killing or lysis of the microorganism may be required to achieve a successful outcome. *See* IMMUNITY.

Antibiotics may be given by injection, orally, or topically. When given orally, they must be absorbed

-S

Spectrum of activity and clinical utility of antibiotics and other antimicrobial agents

Infecting microorganism:			*Staphylococcus aureus*	*Staphylococcus epidermidis*	*Streptococcus pneumoniae*	*Streptococcus pyogenes*	*Streptococcus viridans*	*Enterococcus faecalis*	*Corynebacterium diphtheriae*	*Clostridium tetani*	*Clostridium difficile*	*Bacillus anthracis*	*Mycobacterium tuberculosis*	*Escherichia coli*	*Klebsiella pneumoniae*	*Salmonella* species	*Shigella* species
Antibiotic			Disease produced*														
Generic name	Trade name	Produced by	Bacteremia Endocarditis Osteomyelitis Pneumonia Skin infections Toxic shock syndrome	Bacteremia Endocarditis Urinary tract infections	Bronchitis Meningitis Otitis media Pneumonia	Pharyngeal infections Rheumatic fever Scarlet fever Septic sore throat Tonsillitis	Bacteremia Endocarditis Wound infections	Bacteremia Endocarditis Urinary tract infections	Diphtheria	Tetanus	Antibiotic-associated colitis	Anthrax	Tuberculosis	Abdominal infections Pneumonia Traveler's diarrhea Urinary tract infections	Pneumonia Rhinitis	Bacteremia Gastroenteritis Typhoid fever	Dysentery
Acyclovir	Zovirax	Chemical synthesis															
Amantadine	Symmetrel	Chemical synthesis															
Amikacin	Amikin	*Streptomyces* sp. and chemical modification	⊕	⊕				⊕					+	⊕	⊕	+	+
p-Aminosalicylic acid	PAS	Chemical synthesis											⊕				
Amoxicillin	Amoxil, Polymox, Trimox, etc	*Penicillium chrysogenum* and chemical modification	+	+	⊕	⊕	⊕	⊕				+		⊕		⊕	
Amphotericin B	Fungizone	*Streptococcus nodosus*															
Ampicillin	Polycillin, Principen, etc.	*Penicillium chrysogenum* and chemical modification	+	+	⊕	⊕	⊕	⊕		+	+	+		⊕	+	⊕	⊕
Azidothymidine	AZT, Retrovir	Chemical synthesis															
Azithromycin	Zithromax	*Streptomyces* sp. and chemical modification	⊕	⊕	⊕	⊕	⊕	+	⊕	+	+	⊕					
Azlocillin	Azlin, Securopen	*Penicillium chrysogenum* and chemical modification	+	+	⊕	⊕	⊕	⊕		+	+	+		⊕	+	⊕	⊕
Aztreonam	Azactam	Chemical synthesis												⊕	⊕	+	+
Bacitracin	Bacitracin	*Bacillus subtilis* *Bacillus licheniformis*	+	+	+	+	+	+	+	+	+	+					
Cefactor	Ceclor	*Cephalosporium* sp. and chemical modification	⊕	⊕	⊕	⊕			+					⊕	⊕	+	+
Cefadroxil	Duricef	*Cephalosporium* sp. and chemical modification															⊕
Cefamandole	Mandol	*Cephalosporium* sp. and chemical modification															⊕
Cefoperazone	Cefobid	*Cephalosporium* sp. and chemical modification	⊕	⊕	⊕	⊕	⊕	⊕	+							+	
Cefotaxime	Claforan	*Cephalosporium* sp. and chemical modification	⊕	⊕	⊕	⊕			+					⊕	⊕	+	+
Cefoxitin	Mefoxin	*Streptomyces* sp. and chemical modification	⊕	⊕	⊕	⊕	⊕		+					⊕	⊕	+	+
Ceftazidime	Fortaz, Tazicef, Tazidime	*Cephalosporium* sp. and chemical modification	⊕	+	⊕	+								⊕	⊕	+	+
Ceftriaxone	Rocephin	*Cephalosporium* sp. and chemical modification	⊕	⊕	⊕	⊕	+		+					⊕	⊕	⊕	⊕
Cephalexin	Keflex	*Cephalosporium* sp. and chemical modification	⊕	⊕	⊕	⊕	+		+					+	+	+	+
Cephalothin	Keflin	*Cephalosporium* sp. and chemical modification	⊕	⊕	⊕	⊕	⊕		+	+				⊕	⊕	⊕	⊕
Chloramphenicol	Chloromycetin	*Streptomyces venezuelae*	+	+	⊕	⊕	+	⊕	+	⊕	⊕	+		+	⊕	⊕	⊕
Ciprofloxacin	Cipro	Chemical synthesis	⊕	⊕	⊕	⊕	+	⊕	+	+	+			⊕	⊕	+	⊕
Clarithromycin	Biaxin	*Streptomyces* sp. and chemical modification	⊕	⊕	⊕	⊕	⊕	+	⊕	+	+	⊕					
Clindamycin	Cleocin	*Streptomyces lincolnensis* and chemical modification	⊕	⊕	⊕	⊕	+		+	⊕	⊕						
Clotrimazole	Lotrimin, Mycosporin, etc.	Chemical synthesis															
Cyclophosphamide	Cytoxan, Neosar	Chemical synthesis															
Cycloserine	Seromycin	*Streptomyces orchidaceus*											⊕				
Dideoxyadenosine		Chemical synthesis															⊕
Dideoxycytidine	Hivid	Chemical synthesis															
Dideoxyinosine		Chemical synthesis															
Doxycycline	Vibramycin	*Streptomyces* sp. and chemical modification	+	+	+	+	+	+		+	⊕	⊕	+	⊕	⊕	⊕	⊕
Econazole	Spectazole	Chemical synthesis															
Erythromycin	Erythrocin, Ilotycin, etc.	*Streptomyces erythreus*	⊕	⊕	⊕	⊕	⊕	+	⊕	+	+	⊕					
Ethambutol	Myambutol	Chemical synthesis											⊕				
Fluconazole		Chemical synthesis															

*+ signifies test-tube activity. ⊕ signifies clinical usefulness.

																																					Disease produced*	
				⊕									+	⊕	+	+	+	+	+	+	+			+		+											Brucellosis	*Brucella* species
				+									⊕	⊕	+		⊕	⊕	⊕	⊕	⊕			⊕		⊕	⊕			⊕		⊕		⊕			Urinary tract infections	*Proteus* species
				+									⊕	+			+	⊕		⊕	⊕					⊕								⊕			Bacteremia Pneumonia Urinary tract infections Wound and burn infect.	*Pseudomonas* species
				⊕									+	⊕												+								+			Plague	*Pasteurella* species
		+		⊕								+	⊕	⊕	⊕	+	⊕	⊕	⊕	⊕	⊕			⊕	+	⊕	⊕	+		⊕		⊕		+			Bacteremia Bronchitis Meningitis Otitis media	*Haemophilus influenzae*
		⊕		⊕								⊕																⊕									Whooping cough	*Bordetella pertussis*
		⊕		⊕								⊕	+	⊕	+	+	⊕	+	⊕	⊕	+			+	+	⊕	⊕	⊕		⊕		⊕					Gonorrhea Meningitis Otitis media	*Neisseria* species
				⊕										⊕																							Cholera	*Vibrio* species
		⊕		⊕										+	+	+	⊕		+	⊕	+			+				⊕									Syphilis, yaws	*Treponema* species
																	⊕			⊕																	Lyme disease	*Borellia* species
				+							⊕		+	⊕			+	⊕	⊕	⊕	⊕																Abdominal infections Bacteremia Mixed wound infections	*Bacteroides* species
		⊕		⊕								⊕																⊕									Legionnaire's disease (pneumonia)	*Legionella pneumophila*
		⊕		⊕								⊕		⊕														⊕									Lymphogranuloma venereum Nongonococcal urethritis Pneumonia Psittacosis Trachoma	*Chlamydia* species
		+	⊕	⊕								+		⊕														+									Rickettsial pox Spotted fever Typhus	*Rickettsia* species
+			⊕							⊕																					⊕						Candidiasis (topical systemic)	*Candida* species
+			+							+																					⊕						Aspergillosis	*Aspergillus* species
+			+							+																					⊕						Histoplasmosis	*Histoplasma capsulatum*
⊕										+																					⊕						Meningitis Systemic infections	*Cryptococcus neoformans*
																																					Pneumonia	*Pneumocystis carinii*
																																					Dermatophytosis	*Trichophyton* and *Epidermophyton* spp.
																																					Amebiasis	*Entamoeba histolytica*
																																					Trichomoniasis	*Trichomonas vaginalis*
									⊕																												Mononucleosis Burkitt's lymphoma	Epstein-Barr virus
																																				⊕	Cold sores Encephalitis Genital herpes Keratoconjunctivitis	Herpes simplex virus
																																				⊕	Chicken pox Herpes zoster	Varicella zoster virus
																																					Hepatitis	Hepatitis (A, B, C) virus
																																					Congenital defects Hepatitis Systemic infection	Cytomegalovirus
																																					Genital warts	Papilloma virus
																																			⊕		Influenza	Influenza virus
																																					Bronchitis Pneumonia	Respiratory syncytial virus
					⊕	⊕	⊕															⊕	⊕						⊕								AIDS AIDS-related complex	Human immunodeficiency virus

Spectrum of activity and clinical utility of antibiotics and other antimicrobial agents (cont.)

Infecting microorganism:			*Staphylococcus aureus*	*Staphylococcus epidermidis*	*Streptococcus pneumoniae*	*Streptococcus pyogenes*	*Streptococcus viridans*	*Enterococcus faecalis*	*Corynebacterium diphtheriae*	*Clostridium tetani*	*Clostridium difficile*	*Bacillus anthracis*	*Mycobacterium tuberculosis*	*Escherichia coli*	*Klebsiella pneumoniae*	*Salmonella* species	*Shigella* species
Antibiotic			Disease produced*														
Generic name	Trade name	Produced by	Bacteremia Endocarditis Osteomyelitis Pneumonia Skin infections Toxic shock syndrome	Bacteremia Endocarditis Urinary tract infections	Bronchitis Meningitis Otitis media Pneumonia	Pharyngeal infections Rheumatic fever Scarlet fever Septic sore throat Tonsillitis	Bacteremia Endocarditis Wound infections	Bacteremia Endocarditis Urinary tract infections	Diphtheria	Tetanus	Antibiotic-associated colitis	Anthrax	Tuberculosis	Abdominal infections Pneumonia Traveler's diarrhea Urinary tract infections	Pneumonia Rhinitis	Bacteremia Gastroenteritis Typhoid fever	Dysentery
5-Fluorocytosine	Ancobon	Chemical synthesis															
Fusidic acid	Fucidine	*Fucidium coccineum*	⊕	⊕	+	+	+										
Gentamycin	Garamycin	*Micromonospora purpurea*	⊕	⊕				⊕					+	⊕	⊕	+	+
Griseofulvin	Fulvicin, Grifulvin	*Penicillium griseofulvum*															
Imipenem		*Streptomyces cattleya* and chemical modification	⊕	⊕	⊕	⊕	⊕	⊕		⊕	⊕			⊕	⊕	⊕	⊕
Isoniazid	Cotinazin, Rimifon, Nydrazid	Chemical synthesis											⊕				
Itraconazole	Sporanox	Chemical synthesis															
Kanamycin	Kantrex	*Streptomyces kanamyceticus*	+	+										⊕	⊕	+	+
Ketoconazole	Nizoral	Chemical synthesis															
Methicillin	Azapen, Staphcillin	*Penicillium chrysogenum* and chemical modification	⊕	⊕	+	+										+	
Metronidazole	Flagyl, Protostat	Chemical synthesis								⊕	⊕						
Mezlocillin	Baypen, Mezlin	*Penicillium chrysogenum* and chemical modification	+	+	⊕	⊕	⊕	⊕		+	+	+		⊕	+	⊕	⊕
Miconazole	Monistat	Chemical synthesis															
Nafcillin	Nafcil, Unipen	*Penicillium chrysogenum* and chemical modification	⊕	⊕	+	+	+										
Netilmicin	Netromycin	*Micromonospora inyoesis* and chemical modification	+	+									+	⊕	⊕	+	+
Nitrofurantoin	Furadantin	Chemical synthesis	+	+				⊕	+					⊕	⊕		
Norfloxacin	Noroxin	Chemical synthesis	⊕	⊕				+						⊕	⊕	+	+
Novobiocin	Albamycin	*Streptomyces niveus* *Streptomyces spheroides*	+	+	+	+		+									
Nystatin	Mycostatin	*Streptomyces noursei*															
Oxacillin	Bactocill, Prostephen	*Penicillium chrysogenum* and chemical modification	⊕	⊕	+	+	+										
Penicillin G		*Penicillium notatum* *Penicillium chrysogenum*	+	+	⊕	⊕	⊕	+	⊕	⊕	⊕	⊕					
Piperacillin	Isipen, Pipracil	*Penicillium chrysogenum* and chemical modification	+	+	⊕	⊕	⊕	⊕		+	+	+		⊕	+	⊕	⊕
Polymyxin B	Aerosporin	*Bacillus polymyxa*												⊕	⊕	+	+
Ribavirin	Virazole	Chemical synthesis															
Rifampicin	Rifadin, Rifamate, Rimactane	*Nocardia mediterranea* and chemical modification	⊕	⊕	+	+	+	+	+	+	+		⊕				
Rimantadine		Chemical synthesis															
Spectinomycin	Trobicin	*Streptomyces spectabilis*															
Streptomycin		*Streptomyces* sp.	+	+			+	+	+			+	⊕	+	+	+	+
Sulfamethoxazole	Gantanol	Chemical synthesis	⊕	⊕	+	+								⊕	⊕	⊕	⊕
Teicoplanin		*Actinoplanes teichomyceticus*	⊕	⊕	+	+		⊕									
Tetracycline	Achromycin, Sumycin, etc.	*Streptomyces* sp.	+	+	+	+	+	+		+	⊕	⊕	+	⊕	⊕	⊕	⊕
Ticarcillin	Aerugipen, Monapen, Ticar	*Penicillium chrysogenum* and chemical modification	+		⊕	⊕	⊕	⊕		+	+	+		⊕	⊕	⊕	⊕
Tobramycin	Nebcin, Tobracin	*Streptomyces tenebrarius*	⊕	⊕				⊕					+			+	+
Trimethoprim	Proloprim, Trimpex	Chemical synthesis	⊕	⊕	⊕	⊕	⊕	+						⊕	⊕		
Vancomycin	Vancocin	*Streptomyces orientalis*	⊕	⊕	⊕	⊕	⊕	⊕		⊕	⊕	⊕					
Vidarabine	Ara A	Chemical synthesis															
Zidovudine	Retrovir, AZT	Chemical synthesis															

*+ signifies test-tube activity. ⊕ signifies clinical usefulness.

																																					Disease produced*	
						⊕		+	⊕					+						+		+							+			+					Brucellosis	*Brucella* species
			⊕	⊕	⊕	+		⊕	+					+	⊕					⊕	⊕	⊕			+				+			⊕		⊕			Urinary tract infections	*Proteus* species
				⊕	⊕	+			+					⊕						⊕		⊕							⊕			⊕		⊕			Bacteremia Pneumonia Urinary tract infections Wound and burn infect.	*Pseudomonas* species
			+	+		⊕		+	⊕					+								+							+			⊕		+			Plague	*Pasteurella* species
			⊕	+	⊕	⊕		⊕	+					+	⊕							+			⊕							⊕		+			Bacteremia Bronchitis Meningitis Otitis media	*Haemophilus influenzae*
						⊕																															Whooping cough	*Bordetella pertussis*
			+		⊕	⊕		+	+	⊕		+		+	⊕	⊕				+		+			⊕							⊕					Gonorrhea Meningitis Otitis media	*Neisseria* species
						⊕		+	+					+																		+					Cholera	*Vibrio* species
						⊕			+					+																		+					Syphilis, yaws	*Treponema* species
																⊕																					Lyme disease	*Borellia* species
						+								+												⊕						⊕					Abdominal infections Bacteremia Mixed wound infections	*Bacteroides* species
			⊕			⊕		⊕				⊕																									Legionnaire's disease (pneumonia)	*Legionella pneumophila*
						⊕		⊕				⊕																									Lymphogranuloma venereum Nongonococcal urethritis Pneumonia Psittacosis Trachoma	*Chlamydia* species
						⊕																															Rickettsial pox Spotted fever Typhus	*Rickettsia* species
																		⊕						⊕				⊕		⊕						⊕	Candidiasis (topical systemic)	*Candida* species
																		+						+				+		+						+	Aspergillosis	*Aspergillus* species
																		+						+				⊕		⊕						+	Histoplasmosis	*Histoplasma capsulatum*
																		+						+				+		+						⊕	Meningitis Systemic infections	*Cryptococcus neoformans*
			⊕					⊕																													Pneumonia	*Pneumocystis carinii*
																																		⊕			Dermatophytosis	*Trichophyton* and *Epidermophyton* spp.
																										⊕											Amebiasis	*Entamoeba histolytica*
																										⊕											Trichomoniasis	*Trichomonas vaginalis*
																																					Mononucleosis Burkitt's lymphoma	Epstein-Barr virus
	⊕																																				Cold sores Encephalitis Genital herpes Keratoconjunctivitis	Herpes simplex virus
																																					Chicken pox Herpes zoster	Varicella zoster virus
																																					Hepatitis	Hepatitis (A, B, C virus)
																																					Congenital defects Hepatitis Systemic infection	Cytomegalovirus
																																					Genital warts	Papilloma virus
											⊕																										Influenza	Influenza virus
													⊕																								Bronchitis Pneumonia	Respiratory syncytial virus
⊕							⊕	⊕												⊕	⊕						⊕										AIDS AIDS-related complex	Human immuno-deficiency virus

into the body and transported by the blood and extracellular fluids to the site of the infecting organisms. When they are administered topically, such absorption is rarely possible, and the antibiotics then exert their effect only against those organisms present at the site of application.

Microbial resistance. The therapeutic value of every antibiotic class is gradually eroded by the microbial resistance that invariably follows broad clinical use. For example, resistance to penicillins was recognized soon after the introduction of this class, and use eventually shifted to cephalosporins. Early resistance to macrolides (erythromycin) and chloramphenicol severely limited their use. The emergence of antibiotic resistance is simply natural selection in action. The driving force is antibiotic pressure. In turn, antibiotic resistance is the driving force for the discovery of new antibiotics.

Some bacteria are naturally resistant to certain antibiotics (inherent resistance). For example, enterobacteria are naturally resistant to erythromycin and fusidic acid by virtue of their outer membrane. Clinical resistance is commonly due to the emergence of resistant organisms following antibiotic treatment (acquired resistance). This emergence, in turn, is due to selection of resistant mutants of the infective species (endogenous resistance) or, usually, to transfer of resistance genes from other, naturally resistant species (exogenous resistance). A major challenge in antimicrobial chemotherapy is the horizontal spread of resistance genes and resistant strains, mostly in the hospital but also in the community. The consequences are increased patient morbidity and mortality, reduced drug options, and more expensive and toxic antibiotics.

Detection. Antibiotic resistance is initially detected by susceptibility testing, which provides the resistance phenotype of a given pathogen and has practical implications for patient treament. Rapid detection of resistance and pathogen identification are critical for the rational use of antibiotics and implementation of infection control measures. In the absence of such information, treament is empiric, usually involving broad-spectrum agents, which exacerbates resistance development. Inadequate infection control measures encourage dissemination of resistant strains. The molecular features of resistance are assessed by combinations of biochemical techniques (such as function assays) and molecular biology techniques such as DNA restriction analysis, DNA probes, and polymerase chain reaction (PCR).

Genetics. The genetics of antibiotic resistance are best understood in bacteria. Resistance can result from gene mutations, transfer of R-plasmids between strains or species, and movement of genetic elements (transposons, integrons, gene cassettes) between plasmids or chromosomes in the same organism or to a new organism. For example, R-plasmids were responsible for the spread of sulfonamide resistance to *Shigella* in the 1950s, and penicillin resistance from enterobacteria to *Neisseria* and *Haemophilus* in the 1970s. Plasmids are extrachromosomal circular pieces of DNA that replicate independently of chromosomal DNA and may be present in many copies in each cell. Transposons are smaller pieces of DNA that occasionally move from one site on the chromosome to another site on the chromosome or plasmid. They can replicate only as part of a chromosome or plasmid. Integrons may be part of a plasmid or transposon, while gene cassettes may be part of an integron or may exist free in a circularized (but nonreplicating) form. Gene cassettes and integrons can contain one or more resistance genes. Some resistance genes may have originated from antibiotic-producing organisms or other fortuitously resistant species, or may have evolved from genes coding for normal, mechanistically related cell functions. *See* BACTERIAL GENETICS; PLASMID; TRANSPOSONS.

Mechanisms. The development of resistance varies with the microorganism and the antibiotic. The mechanisms of resistance may be classified as follows: (1) decreased uptake, as with the aminoglycosides, or increased efflux, as with tetracyclines; (2) increased destruction of the drug, as with penicillins, aminoglycosides, and chloramphenicol; (3) altered target site, as with sulfonamides, trimethoprim, streptomycin, erythromycin, and rifampicin; and (4) increased concentration of a metabolite antagonizing the drug action, as with sulfonamides. *See* DRUG RESISTANCE.

Impact on disease. The effects of antibiotics on disease in humans, animals, and plants are discussed below.

Humans. Since not all infectious diseases are reportable to health officials within a community, it is difficult to determine the true incidence of many such illnesses. It is estimated, however, that the average duration of many infectious diseases and the severity of certain others have decreased significantly since the introduction of antibiotic therapy. The dramatic drop in mortality rates for such dreaded diseases as meningitis, tuberculosis, and septicemia offers striking evidence of the effectiveness of these agents. Bacterial pneumonia, bacterial endocarditis, typhoid fever, and certain sexually transmitted diseases are also amenable to treatment with antibiotics. So are infections that often follow viral or neoplastic diseases, even though the original illness may not respond to antibiotic therapy.

Credit for better health, longer life, and diminished mortality rates must be given to a variety of factors. Improved sanitation and housing, immunization programs, and better nutrition have complemented advances in antibiotic therapy to achieve substantial control of infectious diseases. Thus, the vast majority of deaths in developed countries today result from degenerative diseases rather than infections. However, worldwide, infectious diseases remain the leading cause of death. *See* EPIDEMIOLOGY.

Animals. Antibiotics in small amounts are widely used as feed supplements to stimulate growth of livestock and poultry. They probably act by inhibiting organisms responsible for low-grade infections and by reducing intestinal epithelial inflammation. Many experts believe that this use of antibiotics contributes to the emergence of antibiotic-resistant bacteria that could eventually pose a public health problem. They advocate that those antibiotics used in human therapy should not be used for animal growth promotion. Such concerns may eventually affect the practice of promoting animal growth with antibiotics. *See* ANIMAL FEEDS.

In cattle, sheep, and swine, antibiotics are effective against such economically important diseases as bacterial diarrhea, pneumonia, leptospirosis, foot rot, mastitis, and infections of the reproductive and urinary tracts. The use of antibiotics in dogs and cats closely resembles their use in human medical practice. Administration of the antibiotic may be by oral, parenteral, or topical means.

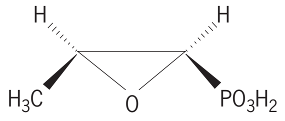
In fish farms, antibiotics are effective against such economically important diseases as bacterial kidney disease in salmon. They are usually added to the food or applied to the fish by bathing. The incidence of infections in fish, and animals in general, may be reduced by the use of disease-resistant stock, better hygiene, and better diet. *See* AQUACULTURE.

Plants. Although effective against many microorganisms causing disease in plants, antibiotics are not widely used to control crop and plant diseases. Some of the limiting factors are instability of the antibiotic under field conditions, the possibility of harmful residues, and expense. Nevertheless, antibiotic control of some crop pathogens is being practiced, as is true of the rice blast in Japan, for example. *See* PLANT PATHOLOGY. Nafsika H. Georgorapadakou

Chemistry of major groups. Most antibiotics are produced by fermentation of microorganisms or semisynthetic approaches that involve chemical derivatization of the naturally occurring antibiotics. Some antibiotics have been isolated from terrestrial plants and marine organisms. Over 13,000 fermentation-derived antibiotics have been reported. Many thousands of synthetic derivatives have been prepared from these compounds in efforts to improve their biological properties. However, relatively few of the natural or semisynthetic compounds have reached commercial status since most of these compounds are not active in an animal model, are too toxic, or have very limited efficacy.

Microorganisms that produce antibiotics include actinomyces, algae, bacteria, and fungi. During the fermentation of pure cultures of these organisms, the antibiotics are frequently produced as complex mixtures of many closely related components. It is not uncommon to have complexes of 10 or more components; however, there are usually one or two major components, with the others present at very low levels. At one time the very minor components were usually not detected, but modern analytical instrumentation such as high-performance liquid chromatography and liquid chromatography/mass spectrometry has greatly facilitated the identification of these compounds. On a number of occasions, more than one apparently unrelated antibiotic complex is produced by the same microorganism.

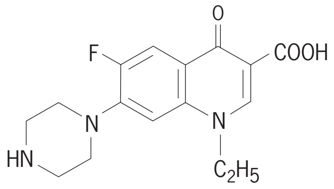
Generally, the chemical structures of these compounds are complex; however, some are very simple with only a few carbon atoms. Total synthesis of even the relatively simple antibiotics can be challenging because of the multiplicity of sterogenic centers and functionality that these compounds frequently possess. The smallest clinically used antibiotic is fosfomycin (**1**), which contains only three carbon atoms. This compound is produced by fermentation of strains of *Streptomyces*. Most of the fermentation-derived antibiotics have molecular weights in the range of 300–800.



(**1**)

The major antibacterial antibiotics in terms of structure class are (in order of commercial importance) cephalosporins, penicillins, quinolones (synthetic), macrolides, aminoglycosides, and tetracyclines. Compared to antifungal, antiviral, antituberculin, and anticancer antibiotics, the antibacterial antibiotics are the most widely used. All of these antibiotics are fermentation products or are derived from fermentation products, except the quinolones, which are synthetic.

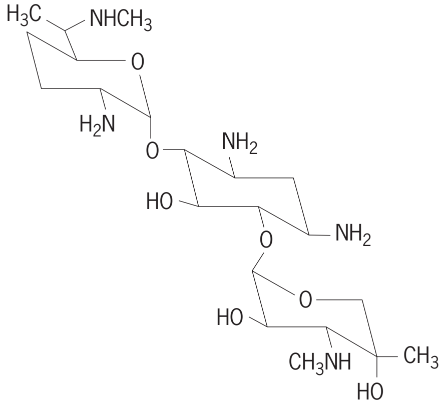
The first clinically used compound in the quinolone family was nalidixic acid, which is orally effective and has activity against gram-negative bacteria. Chemical modifications of this basic structure have led to compounds such as norfloxacin (**2**), which has an improved antibacterial spectrum and significantly less susceptibility to resistance. By the original definition, these compounds are not antibiotics since they have not been produced by microorganisms, but current usage includes them.



(**2**)

There have been many systems to classify the fermentation-derived antibiotics based on structure, mechanism of action, biosynthesis, producing organism, or other features. Although chemical classification is usually the most informative, it is complicated by the complex nature and diversity of chemical structures of these compounds. The following classification of some of the major groups is based on chemical structure, is generally consistent with the more widely used chemical schemes, and covers most of the commercially important antibiotics.

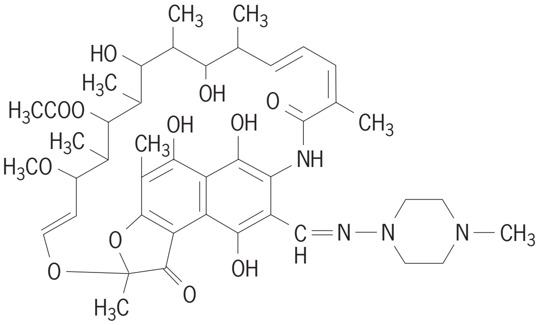
Aminoglycosides. This group of antibiotics is characterized by compounds containing amino sugars and deoxystreptamine or streptamine. Representative members are gentamicin [structure (**3**) is gentamycin



(3)

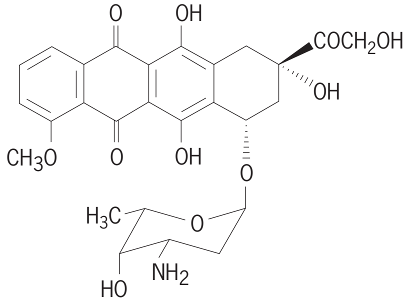
C_1] and neomycin; they are used primarily for gram-negative bacterial infections treated by injection and for topical applications, respectively. These antibiotics destroy bacteria by interfering with their protein biosynthesis. All of the members of this family of antibiotics are extremely water-soluble, polybasic compounds, reflecting the highly polar units that make up these molecules. *See* AMINO SUGAR; POLAR MOLECULE.

Ansamycins. Antibiotics in this family are large-ring structures that contain benzenoid or naphthalenoid aromatic units in which nonadjacent positions are bridged by an aliphatic chain to form the macrocyclic ring. An amide bond is always found at one of the aliphatic-aromatic junctions. Rifampin (**4**), a semisynthetically derived compound, is a member of this family. It kills bacteria by inhibiting their ribonucleic acid (RNA) polymerase. Clinically, it is used against infections by gram-positive bacteria and mycobacteria. *See* MACROCYCLIC COMPOUND.



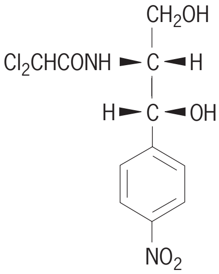
(4)

Anthracyclines. Antibiotics in this family are characterized by having a tetrahydrotetracenequinone chromophore linked to one or more sugars by glycosidic bonds. Doxorubicin (**5**), also known as adriamycin, is a member of this family and is one of the most important clinical antitumor agents. Anthracyclines are antibiotics that appear to destroy bacteria by intercalating with their deoxyribonucleic acid (DNA). Although most anthracyclines are very active against gram-positive bacteria, they are quite toxic to mammals and are not used as antibacterial antibiotics. The clinically used antibacterial antibiotics have much greater safety margins. Doxorubicin, however, has selective activity against certain types of tumors and, with carefully controlled dosing to minimize toxic effects against normal cells, this antibiotic has been very effective for cancer chemotherapy.



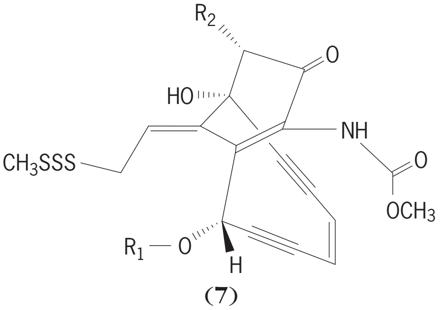
(5)

Chloramphenicol. Many naturally occurring antibiotics belong to families that are relatively large, with 20 or more closely related structures. In contrast, there are only a few compounds that are closely related to chloramphenicol (**6**). The structure of chloramphenicol is relatively simple: it is a nitrobenzene derivative of dichloroacetic acid. This antibiotic has a broad spectrum of antibacterial activity and kills bacteria by the inhibition of protein biosynthesis. Although chloramphenicol is very effective clinically, its use has been limited because of the potential toxic side effects. *See* NITROAROMATIC COMPOUND.



(6)

Enediynes. This is a unique family of antitumor antibiotics in which all members contain a cyclic array of two acetylenic units and a double bond. Calicheamicin and esperamicin are representative members of this family. The aglycone portion [structure (**7**), where R1 and R2 represent sugars and aromatic groups] of these antibiotics has been called a

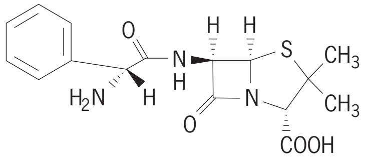


(7)

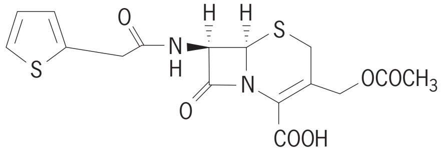
warhead, since it can be triggered to generate diradicals which destructively interact with cellular DNA to cause single- and double-strand cleavage and subsequent cell death. This damaging effect to DNA is probably responsible for the antibacterial and antitumor activities of these compounds. They are the most potent antitumor agents that have been discovered and are approximately 1000 times more potent than adriamycin (**5**).

Glycopeptides. Antibiotics within this family are relatively large molecules with molecular weights frequently in the range of 1400–2000. They all contain peptide units, with a number of aromatic amino acids that have apparently formed cyclic structures by phenolic oxidative coupling. Attached to this aglycone by glycosidic linkages are usually two or more sugar moieties. Vancomycin and avoparcin are glycopeptide antibiotics that are used commercially for medical and agricultural applications. These compounds inhibit bacterial cell-wall biosynthesis by interaction of the antibiotic aglycones with bacterial cell-wall peptides terminating in acyl-D-ala-D-ala. The term acyl-D-ala-D-ala represents the peptides in the bacterial cell wall that end in acyl-D-alanyl-D-alanine. The acyl term indicates linkage to the peptide through the carboxyl group of the bacterial peptide. Peptides ending in this D-ala-D-ala sequence bind to glycopeptide antibiotics. Vancomycin is effective clinically against gram-positive bacteria, especially staphylococci that are resistant to other antibiotics.

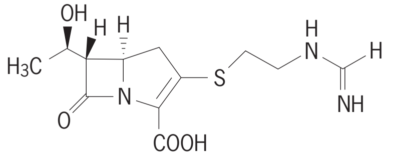
β-Lactams. For chemical classification the β-lactam antibiotics are within the general class of peptide antibiotics. All antibiotics in this family contain a β-lactam ring, and within the family are several substructure groups. The more important groups include the penicillins, cephalosporins, carbapenems, and monobactams as represented by ampicillin (**8**), cephalothin (**9**), imipenem (**10**), and aztreonam (**11**), respectively. Ampicillin and cephalothin are produced semisynthetically, whereas imipenem and aztreonam are derived from total syntheses. Although the latter two compounds are synthetic, the key compounds leading to their discovery were fermentation-derived antibiotics. The β-lactam family is commercially the most important class of antibiotics used for bacterial infections. The β-lactams are relatively nontoxic, since they selectively inhibit bacterial cell-wall biosynthesis. The development of resistance by bacteria, frequently from acquisition of new β-lactamases that inactivate the antibiotics, has spurred the synthesis of thousands of new β-lactam derivatives in efforts to find compounds with improved biological properties. Another important group of β-lactams comprises the β-lactamase inhibitors such as clavulanic acid and tazobactam. These compounds have poor antibiotic activity but are excellent inhibitors of the enzymes that inactivate commercially important β-lactam antibiotics. Consequently, they are used in combination with some of the commercially important β-lactam antibiotics to improve the spectrum of activity. *See* DRUG RESISTANCE; LACTAM.



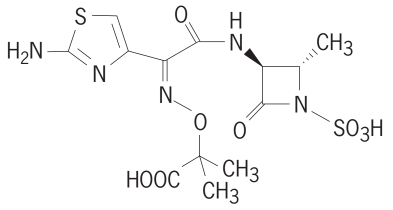
(8)



(9)

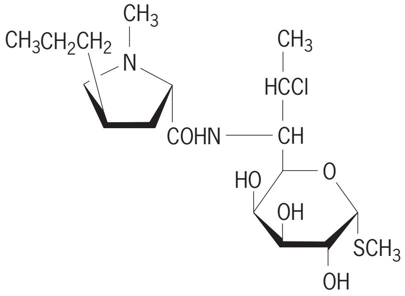


(10)



(11)

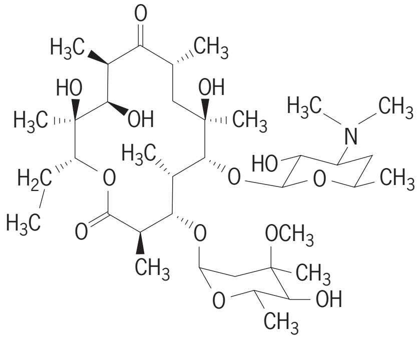
Lincomycin/celesticetin. Antibiotics within this small family are carbohydrate-type compounds containing a thiosugar (a sugar that has one or more of its oxygen atoms replaced with sulfur) linked by an amide bond to a substituted proline unit. Clindamycin (**12**) is clinically one of the most useful compounds in this family and is derived by chemical modification of lincomycin. Antibiotics in this family are effective against gram-positive aerobic and anaerobic bacteria. They destroy bacteria by inhibiting their protein biosynthesis.



(12)

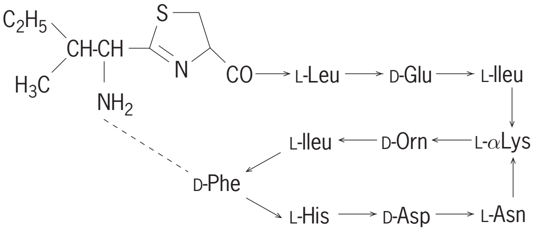
Macrolides. Macrolides constitute a broad category of antibiotics that are chemically defined as macrocyclic lactones. Within this broad category are a number of subfamilies. The two more prominent subfamilies are the antibacterial antibiotics related to erythromycin and the antifungal polyene antibiotics related to amphotericin B. The latter group are known as polyenes. *See* LACTONE.

The general class of antibacterial macrolides related to erythromycin have macrocyclic lactone rings that are 12, 14, or 16 membered. Erythromycin (**13**) has a 14-membered ring and is usually referred to as a basic macrolide, since an amino sugar is attached to its macrocyclic lactone through a glycosidic linkage and gives basic properties to the overall molecule. Other related macrolides such as neutramycin have neutral sugars. Both the neutral and basic macrolides of this type have activity against gram-positive bacteria and inhibit bacterial protein synthesis. Basic macrolides are important commercial antibiotics used for medical (for example, erythromycin) and agricultural (for example, tylosin) applications. Erythromycin is the drug of choice for treatment of legionnaire's disease. *See* LEGIONNAIRES' DISEASE.

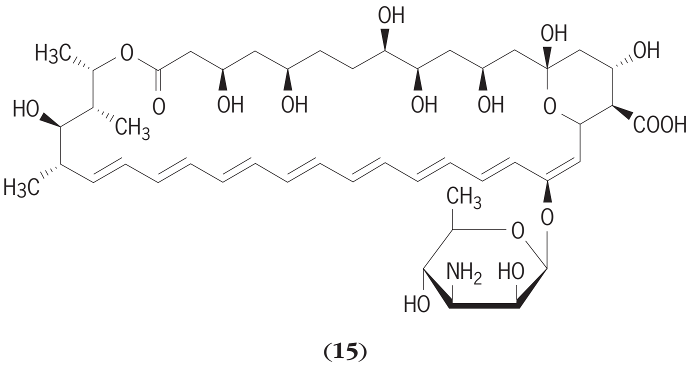


(**13**)

Nucleosides. This antibiotic family is relatively large, with over 200 known structures. The nucleoside antibiotics are analogs of the essential purine and pyrimidine nucleosides that are subunits of RNA and DNA. The differences between the essential nucleosides and the antibiotics are usually in the chirality and type of sugar, substitution on the sugar, the substitution pattern of the base and its oxidation state, and the type of bonding between the sugar and base units. The similarity in chemical structure to numerous critical metabolic products has resulted in nucleoside antibiotics having a number of modes of action as shown by the following examples: (1) puromycin inhibits bacterial protein synthesis by acting as an analog of aminoacyl-tRNA, which takes part in ribosomal peptide bond formation; and (2) the polyoxins interfere with fungal cell-wall assembly by inhibition of chitin synthetase. Many nucleoside antibiotics show antitumor or antiviral activity. Ara-A is used topically for treatment of herpes viral infections. *See* NUCLEOTIDE.



(**14**)



(**15**)

Peptides. Structurally the peptide antibiotics represent a very large and extremely diverse class. Examples of the families within this classification are the linear oligopeptides, cyclic oligopeptides, diketopiperazines, depsipeptides (subunits linked by amide and ester bonds), and β-lactams. The peptide antibiotics contain unusual amino acids residues such as β- or γ-amino acids, amino acids with the D-configuration, or amino acids totally unrelated to the common amino acids. Peptide antibiotics have been used for medical and agricultural applications. Bacitracin [structure (**14**) is bacitracin A] is used as a topical antibiotic in human therapy and as a growth promoter in animal feeds. Like many of the peptide antibiotics, it has a cyclic structure, contains several common amino acids in the D-configuration, and has an unusual amino acid containing a dihydrothiazole unit. *See* AMINO ACIDS.

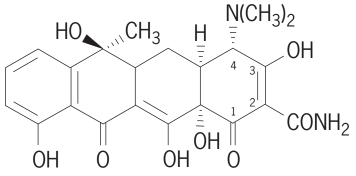
Polyenes. The antifungal polyenes are macrolides with 20- to 44-membered rings containing three to seven conjugated double bonds. They are classified as trienes, tetraenes, pentaenes, hexaenes, or heptaenes, depending on the length of the polyene chromophore. Many of the polyenes contain a glycosidically linked amino sugar, mycosamine or perosamine. Some of the heptaenes have an aromatic moiety, *p*-aminoactophenone or its *N*-methyl analog, on a side chain to the carbon involved in lactone formation. *Streptomyces* are the organisms usually found to produce these compounds. The polyenes usually have broad-spectrum antifungal activity, which results from interaction of these compounds with membrane sterols to change the fungal membrane permeability. Although the polyenes have excellent antifungal activity, only amphotericin B (**15**) is used parenterally for life-threatening antifungal infections such as candidiasis, aspergillosis, histoplasmosis, and coccidioidomycosis since the other polyenes are too toxic. Amphotericin B also has a small safety margin, but it is used because there are no better antifungal antibiotics for these applications.

Polyethers. Most of the polyether antibiotics are characterized by a linear series of tetrahydrofuran and tetrahydropyran moieties that frequently involve spiroketal systems (structures that possess a fusion of two rings, usually five or six membered). These compounds terminate in a carboxylic acid group.

Some polyether antibiotics have a sugar unit linked through a glycosidic bond to one of the tetrahydrofuran or tetrahydropyran rings. This sugar is almost always 4-*O*-methylamicetose. The large majority of these compounds are produced by *Streptomyces* species and to some extent the rarer actinomyces.

Polyethers selectively bind and transport certain cations through cellular membranes. Because of this unusual property, these compounds are referred to as ionophores. Each antibiotic has its own ion specificity. These antibiotics are active against gram-positive bacteria, mycobacteria, fungi, yeast, and certain protozoa. Their ion transport capabilities also makes them quite toxic for parenteral use in mammals. However, some of these compounds are commercially very useful as anticoccidial agents that are administered orally in feed to poultry. Monensin and maduramicin are very effective for this application. Maduramicin is one of the few polyether antibiotics containing a sugar different from 4-*O*-methylamicetose. *See* ION TRANSPORT; IONOPHORE.

Tetracyclines. The naturally occurring tetracyclines are a relatively small family of about 25 antibiotics that are produced mainly by *Streptomyces* species. These antibiotics contain a polyhydronaphthacene nucleus. None of the tetracyclines have a quinone functionality, which is common to numerous other tetracyclic antibiotics such as doxorubicin (**5**). Instead, the tetracyclines are characterized by having an array of hydroxyl and keto groups on one side of the molecule and a relatively fixed substitution pattern on the A-ring, which contains a carboxamide group ($CONH_2$) in the 2-position, as represented by tetracycline (**16**). The mechanism of action for most tetracyclinelike antibiotics is inhibition of bacterial protein synthesis. Although this family of naturally occurring antibiotics is relatively small, a number of the compounds have become commercial products for both medical and agricultural applications. The tetracyclines have a broad antibacterial spectrum, with activity against both gram-positive and gram-negative bacteria; and they are effective against *Rickettsia, Mycoplasma*, and *Spirochaetes*. They are orally active and exhibit relatively low toxicity. Chemical modifications have resulted in compounds such as minocycline with improved biological properties. *See* ANTIMICROBIAL AGENTS; CHEMOTHERAPY.



(**16**)

Donald B. Borders

Bibliography. B. Austin, *Marine Microbiology*, 1988; L. E. Bryan, *Microbial Resistance to Drugs*, 1988; B. W. Bycroft (ed.), *Dictionary of Antibiotics and Related Substances*, 1988; J. E. Conte and S. L. Barriere, *Manual of Antibiotics and Infectious Diseases*, 1988; W. Crueger and A. Crueger, *Biotechnology: A Textbook of Industrial Microbiology*, 2d ed., 1989; E. F. Gale et al., *The Molecular Basis of Antibiotic Action*, 2d ed., 1981; A. Kucers, N. Mck. Bennit, and R. J. Kemp, *The Use of Antibiotics*, 4th ed., 1987; B. Pratt and R. Fekety, *The Antimicrobial Drugs*, 1986; R. Reiner, *Antibiotics: An Introduction*, 1982; H. Umezawa (ed.), *Index of Antibiotics from Actinomycetes*, vol. 1, 1967, vol. 2, 1978; M. J. Weinstein and G. H. Wagman (eds.), *Antibiotics: Isolation, Separation and Purification*, 1978; M. Woodbine, *Antibiotics and Antibiosis in Agriculture*, 1977.

Antibody

A member of the immunoglobulin class of proteins that is produced and secreted by the B cells (or B lymphocytes) of the immune system in response to antigens (foreign substances). Antibodies are found in the tissue fluids and mucous membranes and are essential for protection against and recovery from infection. *See* ANTIGEN; IMMUNITY; IMMUNOGLOBULIN.

Historical perspective. In the 1890s, E. von Behring and S. Kitasato showed for the first time that a toxin-neutralizing activity appeared in the blood plasma of patients or animals that had survived tetanus or diphtheria. Furthermore, plasma from these animals could be transferred to other animals, protecting them. Although the chemical nature of the protective agent, antibody, was unknown, use of antitoxic plasma or serum to treat these and other infections soon became widespread. (Plasma is the liquid, acellular part of blood; serum is the liquid that remains after blood has clotted.) In fact, this was the only effective treatment for bacterial infections until the advent of antibiotics.

In a variety of tests, antibodies were shown to agglutinate bacteria or, in fresh serum, to lyse them; to precipitate in combination with antigen; to cause the skin wheals typical of allergy; and to make bacteria more readily ingested (opsonized) by phagocytic cells. Fierce debates raged about whether each of these activities was the property of different classes of antibody, or simply reflected the nature of the test. Once the structure of antibodies was unraveled, it became clear that several different classes existed.

The earliest experiments showed that antibodies were specific for the antigens that elicited them. They were heterogeneous (varying in charge and size) and migrated at different rates in an electrophoretic field, although generally in the gamma globulin region. Some antibodies were much larger than others, so they were named macroglobulin. When structural studies showed that all antibodies were closely related, the term "immunoglobulin" was adopted, and gamma globulin became immunoglobulin G (IgG), macroglobulin became IgM, and so on. *See* ELECTROPHORESIS; GLOBULIN.

Structure. The five classes of antibodies (IgM, IgD, IgG, IgA, and IgE) are structurally similar, consisting of a basic unit of two heavy (H) and two light (L) polypeptide chains. The H chains are linked to

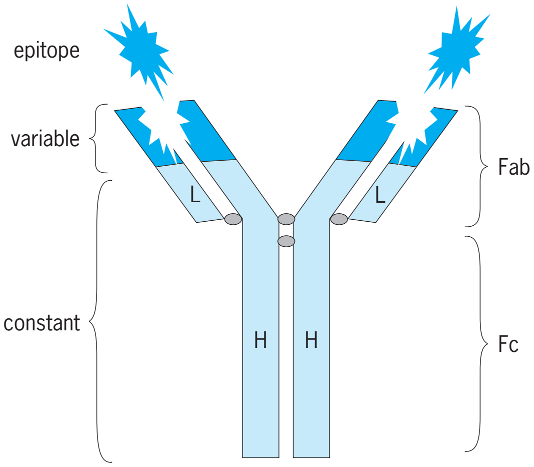


Fig. 1. Typical antibody molecule. There are two H and two L chains, held together by disulfide bonds, and the N-terminal variable domains are shown in dark color. The molecule is bilaterally symmetrical so two identical epitopes are bound. The antigen-binding regions (Fab) and the constant region (Fc) are indicated.

each other by one or more disulfide bonds, and each L chain is similarly linked to an H chain. In any one antibody molecule, the H and L chains are the same, so the molecule is said to have axial symmetry (**Fig. 1**).

X-ray crystallography and molecular sequence analyses show that each chain is composed of immunoglobulin folds, that is, two or more compact domains containing beta-pleated sheets that are held together by disulfide bonds. They probably all derive from an ancient ancestral domain and have a molecular weight of about 10,000 daltons. L chains have two domains, and H chains have four or five. When many antibodies were sequenced, it was found that the N- (or amino-) terminal domains of both H and L chains vary from antibody to antibody, while the other domains are more or less constant within a single class (**Fig. 2**). The variable domains (V_L and V_H) are where antigenic determinants (or epitopes) bind, based on complementarity of surface charge and hydrophobicity. (In this regard, antigen-antibody binding is similar to that between enzymes and substrates.) Between the first and second constant domains of H chains lies a flexible "hinge" peptide region (Fig. 2) that is readily cleaved by enzymes to produce two antigen-binding branches, called Fab, and a common fragment, known as Fc. The Fab regions confer specificity, whereas the Fc is responsible for the function of the particular class of antibody.

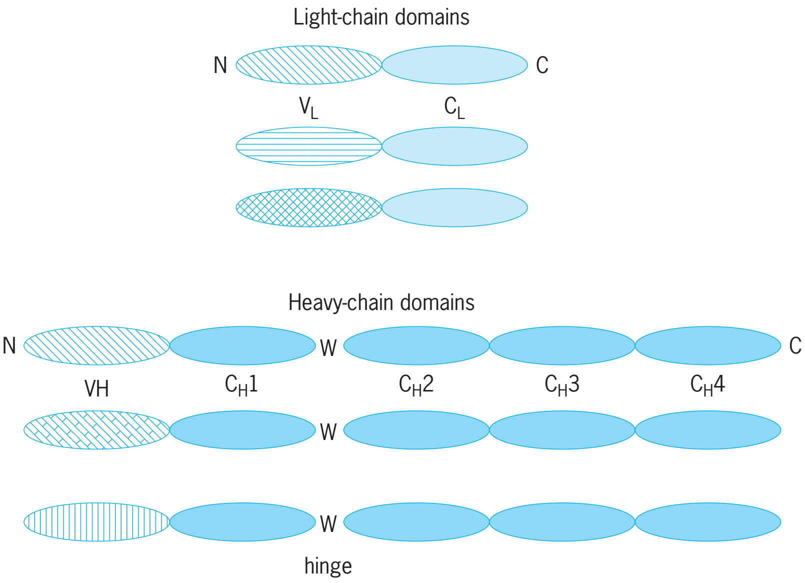


Fig. 2. Domain structure of three typical light and heavy chains shows their similarities and differences. The amino- and carboxy-terminals of the chains are marked N and C, respectively. The flexible hinge region is also shown. At the N terminals the first domain has a variable sequence that differs from antibody to antibody; it is to these domains that antigens bind. The C terminal domains are constant within any one class of immunoglobulin.

Because antibodies are bilaterally symmetrical, each can bind two epitopes (that is, they are divalent). Natural antigens usually have many epitopes, so they can be linked together by antibodies to form a large antigen-antibody complex, or immune complex; this is how, for example, bacteria are agglutinated by antibodies. Complexes between soluble antigens and antibodies become insoluble, which explains the phenomenon of precipitation. Immune complexes that form in the blood may get trapped in the basement membranes that underlie capillaries, resulting in harmful inflammation.

Antibody classes. Each of the five classes, or isotypes, of immunoglobulins probably arose over evolutionary time by gene duplication. This is often the case when a gene product confers a strong survival advantage. (With two copies, one copy of the gene can mutate without risk to the survival of the individual, and so evolve into a new useful gene; then there is even more pressure to duplicate, as every new variant may confer further survival advantages.) Each of the classes has unique structural and biological properties, while retaining overall similarity and function in immune responses. The specific properties of each class depend on the structure of the constant domains of its H chain; they all share the same pool of L chains. The very same L chains, and V_H domains, can be found in each of the five classes. In fact, a B cell that is making, for example, IgM against a certain epitope may be found a few days later to be making IgG or IgE of exactly the same specificity.

IgD. IgD consists of two L chains and two delta (δ) H chains. Although IgD is present in the blood, its major role is in the activation of B lymphocytes. Before exposure to antigen, a mature B cell expresses surface IgD (sIgD) and surface IgM (sIgM) of the same specificity. Antigen binding by both sIgD and sIgM activates the B cell.

IgM. IgM is the largest of the antibody molecules. Secreted IgM is composed of five basic units [composed of two L chains and two mu (μ) H chains] connected by disulfide binds to a J (joining) chain. Although it can bind 10 epitopes, in practice it binds only one or two. IgM is present in the blood (at a concentration of 100 mg/dL), but its large size hinders its movement from the blood to the tissue spaces. In addition, IgM is not capable of opsonization, as phagocytic cells do not have receptors for its Fc end.

Its main function in immune responses is to activate complement, a group of interacting proteins that are important mediators of inflammation and can attract and assist white blood cells in capturing bacteria and directly lyse bacterial cells. *See* COMPLEMENT.

IgG. IgG is a typical 2L–2H basic unit, with gamma (γ) H chains. It is the most abundant immunoglobulin class in blood (1000 mg/dL). It activates complement, but it is much less efficient than IgM. Since its Fc end fits receptors on most phagocytic white blood cells, it is capable of opsonization. IgG is the only isotype that can cross the placenta from human mother to fetus, providing vital antenatal and neonatal protection. However, if a mother is negative for the red blood cell antigen Rh(D) and her fetus is positive for the antigen, red cells from the fetus can cross the placenta and immunize the mother. The mother's IgG antibodies then cross the placenta in the other direction and destroy the fetal red cells, causing a condition known as hemolytic disease of the newborn. It is prevented by administering IgG against Rh(D) to the mother at 28 weeks' gestation and again at delivery. *See* RH INCOMPATIBILITY.

IgA. IgA is present in secretory fluids such as saliva, bronchial fluid, and milk. It is found in high concentrations in the colostrum (the first milk produced by the mother after giving birth) and continues to protect a baby's oral, nasal, and pulmonary mucosa as long as it nurses. IgA is produced by B cells located under the mucosa of secretory epithelia and released as a dimer of two basic units [of two L chains and two alpha (α) H chains] joined at their Fc ends by a J chain. The antibody is bound by receptors on the basal aspect of the mucosal cells and transported to the luminal side, where it is released into the secretions along with the receptor, known as secretory component. The secretory component seems to protect IgA from digestion by enzymes secreted by humans and many microorganisms.

IgE. IgE, a basic unit molecule with two L and two epsilon (ϵ) H chains. IgE antibodies are present in blood in very small amounts (micrograms per deciliter), but they are potent and can bind strongly to the membrane of tissue mast cells and blood basophils (the major effector components in an immediate allergic reaction).When antigen cross-links adjacent molecules of IgE, it causes mast cells and basophils to release their contents of many inflammatory mediators, mainly histamine. Thus, IgE is the antibody responsible for allergies and asthma. The mast cells triggered by IgE attract specialized phagocytes called eosinophils, which contain defense molecules that allow them, uniquely, to kill parasites. *See* ALLERGY.

Antibody production. Antibodies are produced by B lymphocytes and are the product of gene recombination events that lead to a unique V_L and V_H domain in each cell. Initially, this gene cassette (the set of gene segments encoding the H and L variable regions) is further recombined with the genes encoding the constant domains of the H chains of IgM and IgD. A primary ribonucleic acid (RNA) transcript is then made, and it is processed to messenger RNA (mRNA) for the mature H chains of IgM and IgD. These then combine with L chains as they are synthesized. Thus, the mRNAs that are translated into the H and L polypeptide chains are produced by various recombinations of gene segments that encode the variable and constant domains.

The resulting immunoglobulin is inserted into the plasma membrane, where it serves the B lymphocyte as an antigen receptor. When the cognate antigen binds to its corresponding sIgD and sIgM on the B cell and other environmental conditions are met, the B cell matures into an antibody-secreting plasma cell, which can release a million antibodies per minute. These secreted antibodies are identical to the ones that served as B-cell receptors.

IgM and IgD are always synthesized first and inserted into the B-cell membrane as receptors. If the B cell never encounters antigen, this is all that happens. If it does meet antigen, it will begin to secrete IgM (IgD is only there as a receptor). If so instructed by helper T cells, B cells will switch to making IgG, IgA, or IgE by recombining the V domain of their H chain gene with the corresponding new constant (or C) domain gene. Thus, the class of the secreted antibody changes, but the specificity for antigen does not. *See* CELLULAR IMMUNOLOGY; GENE; IMMUNOLOGICAL ONTOGENY.

Induction. After exposure to a new antigen, for example a vaccine, sensitive tests may detect IgM antibody in the blood in 2 or 3 days. For most antigens, the IgM concentration will peak in a week or two, and then decline as the IgM is replaced by IgG, the levels of which will decline in weeks to months. A second (booster) injection of antigen induces an IgG response which is quicker, rises more steeply, and is sustained longer. *See* VACCINATION.

Polyclonal and monoclonal antibodies. Each B cell (and each antibody) is specific for a single epitope, but most proteins have several epitopes (which can be 10 to 20 amino acids long), and organisms such as bacteria have hundreds. When many different B-cell clones (identical B cells) produce different antibodies against various antigenic determinants, or epitopes, the antibodies and the response are said to be polyclonal. A monoclonal antibody is produced by only one of these B-cell clones; it is not usually seen in nature but is produced by isolation of single B cells and their progeny. Newer methods are available to produce monoclonal antibodies, which are valuable in research, diagnosis, and therapy. *See* ANTIGEN-ANTIBODY REACTION; MONOCLONAL ANTIBODIES.

J. John Cohen

Bibliography. D. R. Davies and H. Metzger, *Structural Basis of Antibody Function*, in *Annual Review of Immunology*, 1:87–115, 1983; P. Holliger and P. J. Hudson, Engineered antibody fragments and the rise of single domains, *Nat. Biotech.*, 23:1126–1136, 2005; C. A. Janeway et al., *Immunobiology*, 6th ed., Garland Publishing, 2004; R. A. Manz et al., Maintenance of serum antibody levels, *Annu. Rev. Immunol.*, 23:367–386, 2005; W. E. Paul (ed.), *Fundamental Immunology*, 5th ed., Lippincott Williams & Wilkins, 2003.

-s

Anticline

A fold in layered rocks in which the strata are inclined down and away from the axes. The simplest anticlines (see **illus.**) are symmetrical, but in more highly deformed regions they may be asymmetrical, overturned, or recumbent. Most anticlines are elongate with axes that plunge toward the extremities of the fold, but some have no distinct trend; the latter are called domes. Generally, the stratigraphically older rocks are found toward the center of curvature of an anticline, but in more complex structures these simple relations need not hold. Under such circumstances, it is sometimes convenient to recognize two types of anticlines. Stratigraphic anticlines are those folds, regardless of their observed forms, that are inferred from stratigraphic information to have been anticlines originally. Structural anticlines are those that have forms of anticlines, regardless of their original form. *See* FOLD AND FOLD SYSTEMS; SYNCLINE.

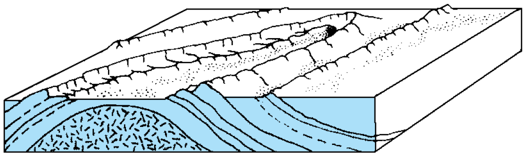


Diagram relating anticlinal structural to topography.

Philip H. Osberg

Antiferromagnetism

A property possessed by some metals, alloys, and salts of transition elements in which the atomic magnetic moments, at sufficiently low temperatures, form an ordered array which alternates or spirals so as to give no net total moment in zero applied magnetic field. **Figure 1** shows the simple antiparallel arrangement of manganese moments at temperatures below 72 K in the unit cell of manganese fluoride (MnF_2). The most direct way of detecting such arrangements is by means of neutron diffraction. *See* NEUTRON DIFFRACTION.

Néel temperature. This is the transition temperature (L. Néel, 1932) below which the spontaneous antiparallel magnetic ordering takes place. A plot of the magnetic susceptibility of a typical antiferromagnetic powder sample versus temperature is shown in **Fig. 2**. Below the Néel point, which is characterized by the sharp kink in the susceptibility, the spontaneous ordering opposes the normal tendency of the magnetic moments to align parallel to the applied field. Above the Néel point, the substance is paramagnetic, and the susceptibility χ obeys the Curie-Weiss law, as in Eq. (1), with a negative param-

$$\chi = C/(T + \theta) \qquad (1)$$

agnetic Curie-Weiss temperature $-\theta$. The Néel temperature is similar to the Curie temperature in ferromagnetism. *See* CURIE TEMPERATURE; CURIE-WEISS LAW; MAGNETIC SUSCEPTIBILITY.

The cooperative transition that characterizes antiferromagnetism is thought to result from an interaction energy U of the form given in Eq. (2), where

$$U = -2\Sigma J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j \qquad (2)$$

$\mathbf{S}_i$ and $\mathbf{S}_j$ are the spin angular momentum vectors associated with the magnetic moments of neighbor atoms i and j, and J_{ij} is an interaction constant which probably arises from the superexchange coupling discussed later, although formally Eq. (2) is identical to the Heisenberg exchange energy. If all J_{ij} are positive, the lowest energy is achieved with all $\mathbf{S}_i$ and $\mathbf{S}_j$ parallel, that is, coupled ferromagnetically. Negative J_{ij} between nearest-neighbor pairs (i, j) may lead to simple antiparallel arrays, as in Fig. 1; if the distant neighbors also have sizable negative J_{ij}, a spiral array may have lowest total energy. *See* FERROMAGNETISM; HELIMAGNETISM.

A simple lattice like MnF_2 (Fig. 1) can be divided into sublattice 1, containing all corner atoms, and sublattice 2, containing all body-centered atoms. Nearest-neighbor interactions connect atoms on different sublattices. On the average, the interaction may be replaced by a single antiparallel coupling between the total magnetizations M_1 and M_2 of the two sublattices. Each sublattice acts as if it

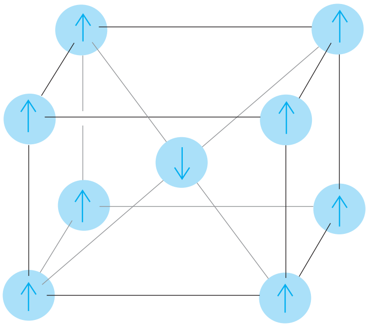


Fig. 1. Antiferromagnetism in manganese fluoride. Only manganese atoms are shown.

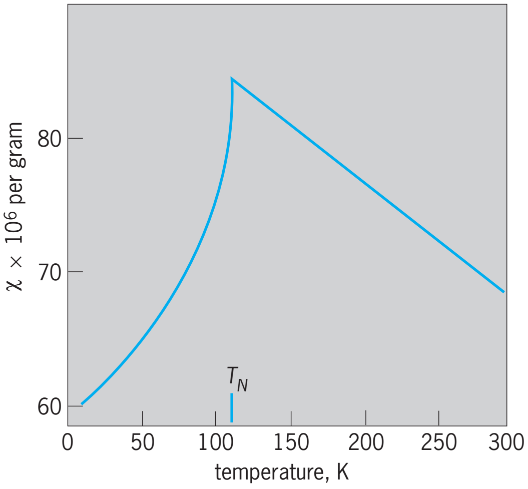


Fig. 2. Magnetic susceptibility of powdered manganese oxide. T_N = Néel temperature. (*After H. Bizette, C. F. Squire, and B. Tsai, 1938*)

were in a large internal magnetic field (Weiss field) proportional to the negative magnetization of the other sublattice. This elementary approach was first given by Néel in 1932 and is analogous to the Weiss molecular field theory of ferromagnetism. More exact treatments of Eq. (2) have been made, but the basic features of antiferromagnetism appear in this simple model.

At high temperatures, the sublattice magnetizations obey the Curie law, as in Eqs. (3) and (4), where

$$M_1 = (C'/T)(H_0 - \lambda M_2) \qquad (3)$$

$$M_2 = (C'/T)(H_0 - \lambda M_1) \qquad (4)$$

C' is the Curie constant for a sublattice, H_0 is an applied external field, and $-\lambda$ is the proportionality constant of the internal Weiss field. From Eqs. (3) and (4), Eq. (5) derived, is derived, which fits the Curie-Weiss law, Eq. (1), with the values shown in Eqs. (6).

$$\begin{aligned} \chi &= (M_1 + M_2)/H_0 \\ &= 2(C'/T)[1 + \lambda(C'/T)] \qquad (5) \end{aligned}$$

$$C = 2C' \qquad \theta = C'\lambda \qquad (6)$$

The condition that M_1 and M_2 can have finite values in the absence of H_0 (condition of spontaneous sublattice magnetization) is that the determinant of the coefficients of M_1 and M_2 in Eqs. (3) and (4) vanishes, which is satisfied at a temperature given by Eq. (7). This is the Néel temperature. Equations (3) and (4) hold only for $T \geq T_N$; the Curie law takes a complicated form for $T < T_N$. In this latter region, the sublattice magnetization varies with temperature essentially in the same manner as does the magnetization of ferromagnetism.

$$T_N = C'\lambda \qquad (7)$$

The preceding theory predicts $\theta/T_N = 1$; the experimental values (see **table**) range from 0.7 to 5. P. W. Anderson ascribed this disagreement to the oversimplified two-sublattice model. Anderson's multisublattice theory not only accounts for $\theta/T_N > 1$, but also predicts a variety of magnetic ordering arrangements, many of which have been confirmed by neutron diffraction. For example, the arrangement in MnO is shown in **Fig. 3**; the magnetic moments are all parallel in alternating planes.

Some representative antiferromagnets

Substance	Crystal type	Néel temp. (T_N), K	Curie-Weiss temp. (θ), K
MnF_2	Rutile	67	80
MnO	NaCl	122	610
FeO	NaCl	198	507
$KMnF_3$	Perovskite	88	158
$CuCl_2 \cdot 2H_2O$	Orthorhombic	4.3	4.5
CrSb	NiAs	723	550
Cr_2O_3	Al_2O_3	307	485
$ZnFe_2O_4$	Spinel	9	
EuTe	NaCl	7.8	6
MnTe	Hexagonal close-packed	403	690
La_2CuO_4	Orthorhombic	250	
$YBa_2Cu_3O_6$	Orthorhombic	500	

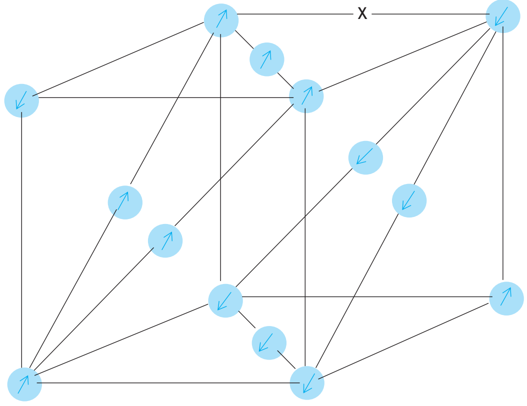


Fig. 3. Antiferromagnetism in manganese oxide. Only manganese atoms are shown.

Superexchange. This is an effective coupling between magnetic spins which is indirectly routed via nonmagnetic atoms in salts and probably via conduction electrons in metals. Consider the oxygen atom at the position labeled X in Fig. 3. The three dumbbell-shaped electronic wave functions of the oxygen will each overlap a pair of manganese atoms (**Fig. 4**). Along any one of these dumbbells, the ground state is $Mn^{2+}O^{2-}Mn^{2+}$, and the overlap mixes in the excited states $Mn^{+}O^{-}Mn^{2+}$ and $Mn^{2+}O^{-}Mn^{+}$, in which an electron "hops" from oxygen to manganese. The electron hops more easily if its magnetic moment is antiparallel to the manganese magnetic moment. Detailed consideration shows that there is an indirect tendency, from this mechanism, for the magnetic moments of the two manganese ions to be anti-parallel; this can be expressed by an energy of the form $-J_{ij}\,\mathbf{S}_i \cdot \mathbf{S}_j$, with negative J_{ij}. This coupling aligns the moments of second-neighbor manganese ions in an antiparallel array, as in Fig. 3. First neighbors are coupled by "right-angled" superexchange from π-like bonding. This is probably comparable to second-neighbor coupling in MnO but does not affect ordering, primarily because it is geometrically impossible for all first neighbors to be antiparallel to one another.

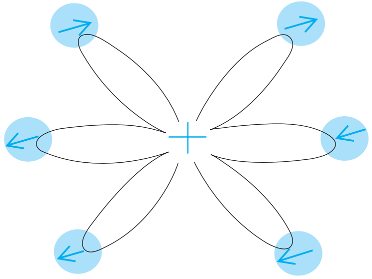


Fig. 4. Superexchange mechanism, in which coupling between magnetic spins of manganese atoms is indirectly routed through oxygen atom in center.

In metals the conduction electrons may play the "hopping" role ascribed above to O^{2-} electrons, or the antiferromagnetism may be related to periodic magnetic order in the electron energy bands.

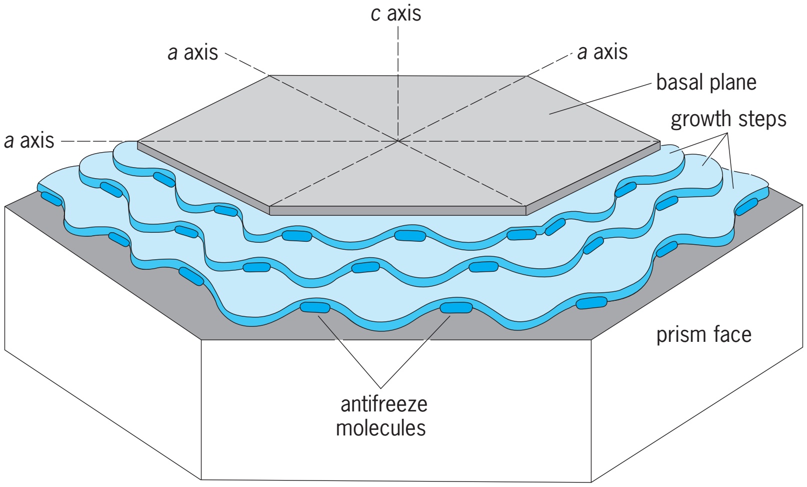
Magnetic anisotropy. The magnetic moments are known to have preferred directions; these are shown in Figs. 1 and 3. In MnO, it is not known exactly in which direction the moments point, except that it is some direction in the (111) planes. Anisotropic effects come from magnetic dipole forces (predominant in MnF_2 and in MnO) and also from spin-orbit coupling combined with superexchange. Some nearly antiparallel arrays, such as Fe_2O_3, show a slight bending (canting) and exhibit weak ferromagnetism. The anisotropy affects the susceptibility of powder samples and is of extreme importance in antiferromagnetic resonance. *See* MAGNETIC RESONANCE; MAGNETISM. E. Abrahams; E. Keffer; M. Brian Maple

Bibliography. N. Ashcroft and N. D. Mermin, *Solid State Physics*, 1976; D. Craik, *Magnetism*, 1995; D. C. Jiles, *Introduction to Magnetism and Magnetic Materials*, 1991; C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996.

Antifreeze (biology)

Glycoprotein or protein molecules synthesized by polar and north temperate fishes to enable them to survive in freezing seawater. Similar antifreezes are found in some insects, but relatively little is known about their structure and function.

Freezing-point depression. In a marine fish, the amount of salt and other small molecules in the blood depresses its freezing point to about 30°F (−0.8°C). In the winter, the polar oceans and the nearshore water of north temperate oceans are at the freezing point of seawater, 28.6°F (−1.9°C). In the absence of ice, many fishes survive by supercooling, a thermodynamic state of equilibrium in which a solution (the body fluids of the fish in this case) can be in liquid state, in the absence of ice nuclei, at a temperature lower than the equilibrium freezing point. However, polar waters are often laden with ice that can enter the fish by ingestion of seawater. Propagation of ice in the body fluids or tissues of the fish always leads to freezing damage and death. To avoid freezing, many fishes have evolved biological antifreezes that further lower the freezing point of their body fluids to 28°F (−2.2°C), which is 0.6°F (0.3°C) below the freezing point of seawater.



Model of adsorption inhibition as a mechanism of noncolligative lowering of the freezing point of water.

Freezing-point depression of water by a solute is one of the colligative properties and is solely a function of the number of dissolved particles. A 1-molal solution of sodium chloride depresses the freezing point of water by about 6.5°F (3.6°C). The freezing-point depression by biological antifreezes is 200 to 300 times greater than that predicted by their molal concentration, and is therefore a noncolligative process. However, antifreezes show the expected colligative effect on the melting point. This separation of freezing and melting points, sometimes referred to as thermal hysteresis, is a unique property of antifreezes not found in any other class of molecules.

Types. Two types of antifreeze have been isolated; they are either peptides or glycopeptides, the latter containing carbohydrate moieties. Antifreeze glycopeptides were discovered in Antarctic fishes belonging to the family Nototheniidae. They are a series of at least eight different-sized polymers, antifreeze glycopeptides 1 to 8, composed of different numbers of the basic repeating tripeptide unit, alanyl-alanyl-threonine with the disaccharide galactose-*N*-acetylgalactosamine linked to the threonine residue by an α-glycosidic linkage. The molecular masses range from 2600 to 34,000 daltons. The smaller antifreeze glycopeptides (6, 7, and 8) differ slightly from the larger ones in that a few alanines are replaced by prolines. Antifreeze glycopeptides of identical or very similar structure have subsequently been isolated from several northern cods belonging to the family Gadidae. *See* PEPTIDE.

Antifreeze peptides are relatively smaller molecules, mostly of several thousand daltons, and unlike the structurally conserved antifreeze glycopeptides, they have diverse composition and structure. They have been isolated from a number of northern fishes and two Antarctic species. Antifreeze peptides of the winter flounder and Alaskan plaice (flat fishes) and of several sculpins (cottids) have high alanine content. Their primary structures consist of stretches of nonpolar alanines separated by small clusters of polar residues, and their secondary structure is largely an alpha helix. Antifreeze peptides of sea raven, another cottid, however, are rich in cystine and possess beta structure. A third class of antifreeze peptides has been isolated from members of the family Zoarcidae (eel pouts), including two Antarctic, one Arctic, and one Atlantic ocean species. Unlike all the other antifreezes, this type has no biased distribution of amino acids and appears to lack defined secondary structure.

Mechanism of effect. Adsorption of antifreeze molecules to the surface of ice crystals, leading to inhibition of the crystals' growth, has been proposed to explain the noncolligative depression of the freezing point. Such adsorption has been demonstrated, but precisely how it is achieved is not entirely clear. For the antifreeze glycopeptides and the winter flounder antifreeze peptide, a lattice match between polar side chains and oxygen atoms in the ice crystal allowing hydrogen bonding may account for adsorption. The helical secondary structure of flounder antifreeze peptide places the side chains of the polar residues on one side of the helix, and those of the nonpolar alanine on the other. The spacing between many of the polar side chains is 0.45 nanometer and matches that between the oxygen atoms in the ice crystal parallel to the *a* axes. In the antifreeze glycopeptides, many of the hydroxyl groups of the disaccharide moieties are also spaced 0.45 nm apart.

An ideal ice crystal is a hexagon characterized by a *c* axis normal to three *a* axes (see **illus**.). The preferred direction of growth is along the *a* axes, presumably by water molecules joining the basal plane at the growth steps. Antifreeze molecules are thought to adsorb along the growth steps and force the ice to grow in between them, leading to the formation of many small, highly curved growth fronts. These curved fronts result in a large surface area-to-volume ratio, which raises the surface free energy. In order for freezing to continue, energy must be removed by further reducing the temperature; in other words, the freezing point is lowered. When the temperature is lowered below the freezing point of an antifreeze solution in the presence of a small seed ice crystal, ice growth occurs extremely rapidly along the *c* axis, the nonpreferred axis of growth, in the form of ice needles or spicules. Antifreeze molecules therefore not only inhibit ice crystal growth in between melting point and freezing point but also drastically change the crystal habit of ice growth. *See* COLD HARDINESS (PLANT); CRYPTOBIOSIS.

Arthur L. DeVries

Bibliography. A. L. DeVries, Biological antifreeze agents in cold-water fishes, *Comp. Biochem. Physiol.*, 73A:627–640, 1982; A. L. Devries, Role of glycopeptides and peptides in inhibition of crystallization of water in polar fishes, *Phil. Trans. Roy. Soc. London*, B304:575–588, 1984; A. L. DeVries, The role of antifreeze glycopeptides and peptides in the freezing avoidance of Antarctic fishes, *Comp. Biochem. Physiol.*, 90B:611–621, 1988; J. T. Eastman and A. L. DeVries, Antarctic fishes, *Sci. Amer.*, 254(11):106–114, 1986; J. A. Raymond and A. L. DeVries, Adsorption inhibition as a mechanism of freezing resistance in polar fishes, *Proc. Nat. Acad. Sci. USA*, 74:2589–2593, 1977; J. D. Schrag et al., Primary and secondary structure of antifreeze peptides from Arctic and Antarctic zoarcid fishes, *Biochim. Biophys. Acta*, 915:357–370, 1987.

Antifreeze mixture

A chemical substance that, when added to a liquid such as water, reduces the freezing point of the mixture. Antifreezes are used in a wide variety of applications, the most common being automotive cooling systems. Antifreeze liquids are also used in refrigeration systems (as a secondary coolant), heating and air-conditioning systems, ice skating rinks, and solar energy units, and as deicing agents for runways and aircraft. *See* ENGINE COOLING; REFRIGERATION.

Properties of a desirable antifreeze include the ability to depress the freezing point of the liquid (typically water), good solubility, high boiling point (to provide boil-over protection), chemical compatibility with materials of construction, good heat transfer properties, appropriate viscosity, and low cost. Freezing-point depression is a colligative property and follows Raoult's law, that is, the reduction in freezing point is proportional to the number of molecules of solute per unit of solvent. For example, 46 grams (1 mole) of methanol will depress the freezing point of 1000 grams of water to the same extent as will 342 grams (1 mole) of sucrose (1.86°C or 3.35°F). *See* SOLUTION.

Chemicals that have been used as antifreezes include glycols, glycerol, brines (such as calcium chloride), and alcohols. Ethylene glycol is the most common antifreeze used in automotive cooling systems because of the outstanding freezing-point depression effect, boil-over protection, heat transfer characteristics, high flash point, and low vapor pressure. Propylene glycol, diethylene glycol, and methanol have also been used to a limited extent. Propylene glycol has a lower oral toxicity to mammals than does ethylene glycol, but both glycols are readily biodegraded and are essentially nontoxic to aquatic life. *See* ETHYLENE GLYCOL.

Commercial automotive antifreezes contain corrosion inhibitors to protect the various types of metals in the cooling system. Different formulations are available for different materials of construction and type of service, but typical inhibitors for light-duty automotive use (passenger cars) include phosphates, borates, molybdates, and nitrates for steel and iron protection and thiazoles for copper and brass protection. To protect against corrosion of heat-rejecting aluminum surfaces, special silicate-based corrosion inhibitor packages are used. Antifoam agents and dyes are also frequently added. Heavy-duty applications, such as in diesel trucks, require a low-silicate formulation (because aluminum use in the engine is low, and potential silicate gel formation is avoided) and use supplemental cooling additives to protect against cavitation corrosion of the cylinder liners. Long-life coolants are specially formulated with carboxylic acid salt inhibitors designed to be nondepleting with use. Some locations around the world require phosphate-free inhibitor packages and use carboxylic acid salt, benzoate, and nitrate inhibitor packages. *See* INHIBITOR (CHEMISTRY).

Antifreeze may become unsuitable for use because of depletion of inhibitors, the presence of corrosion products or corrosive ions, or degradation of the glycol. At that point the fluid is replaced and the old fluid is discarded. Although the glycols are readily biodegraded, used antifreeze may pick up heavy metals from the cooling system (such as lead from the solder) so recycling of used coolant may be preferred for environmental and economic reasons. Recycling can be done a number of ways, but distillation of used antifreeze to remove water and reclaim the glycol is the most common.

Glycol antifreeze solutions are often used in aircraft deicing. These contain additional components for corrosion protection and wetting. Aircraft anti-icing fluids also contain a polymeric thickening agent to increase the fluid viscosity, which allows the fluid to adhere to the aircraft surface and provide protection against freezing for a limited period of time. These thickeners impart non-newtonian behavior so that at high shear, such as when the aircraft is taking off, the fluid viscosity is reduced and the fluid flows off the aircraft. Deicing compounds are also used on airport runways to break the ice-surface bond. Chemicals used in this application include glycols, urea, and potassium acetate. *See* NON-NEWTONIAN FLUID. Kathleen F. George

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed. vol. 3, 1992; I. L. Rozenfeld, *Corrosion Inhibitors*, 1981; Society of Automotive Engineers, *Engine Coolant, Cooling Systems, Materials and Components*, SP-960, 1993.

Antifriction bearing

A machine element that permits free motion between moving and fixed parts. Antifrictional bearings are essential to mechanized equipment; they hold or guide moving machine parts and minimize friction and wear. Friction wastefully consumes energy, and wear changes dimensions until a machine becomes useless.

Simple bearings. In its simplest form a bearing consists of a cylindrical shaft, called a journal, and a mating hole, serving as the bearing proper. Ancient bearings were made of such materials as wood, stone, leather, or bone, and later of metal. It soon became apparent for this type of bearing that a lubricant would reduce both friction and wear and prolong the useful life of the bearing. Convenient lubricants were those of animal, vegetable, or marine origin such as mutton tallow, lard, goose grease, fish oils, castor oils, and cottonseed oil. Egyptian chariot wheels show evidence of the use of mutton tallow for bearing lubrication.

The use of mineral oils dates back principally to the discovery of Drake's well at Titusville, Pennsylvania, in 1859. Petroleum oils and greases are now generally used for lubricants, sometimes containing soap and solid lubricants such as graphite or molybdenum disulfide, talc, and similar substances.

Materials. The greatest single advance in the development of improved bearing materials took place in 1839, when I. Babbitt obtained a United States patent for a bearing metal with a special alloy. This alloy, largely tin, contained small amounts of antimony, copper, and lead. This and similar materials have made excellent bearings. They have a silvery appearance and are generally described as white metals or as Babbitt metals. For many decades they have served as the measure of excellence against which other bearing materials have been compared. *See* ALLOY.

Wooden bearings are still used, however, for limited applications in light-duty machinery and are frequently made of hard maple which has been impregnated with a neutral oil. Wooden bearings made of lignum vitae, the hardest and densest of all woods, are still used. Lignum vitae is native only to the Caribbean area. This wood has a density of approximately 80 lb/ft^3 (1.3 g/cm^3) and has a resin content of some 30% by volume; thus it is remarkably self-lubricating. The grain is closely interwoven, giving the material high resistance to wear and compression and making it difficult to split. Applications are found in chemical processing and food industries where lignum vitae wood can successfully resist the destructive action of mild acids, alkalies, oils, bleaching compounds, liquid phosphorus, and many food, drug, and cosmetic compounds.

About 1930 a number of significant developments began to occur in the field of bearing metals. Some of the most successful heavy-duty bearing metals are now made of several distinct compositions combined in one bearing. This approach is based on the widely accepted theory of friction, which is that the best possible bearing material would be one which is fairly hard and resistant but which has an overlay of a soft metal that is easily deformed. **Figure 1** shows a number of combinations of bearing materials made up as indicated of combinations of hard and soft materials. **Figure 2** shows actual bearings in which graphite, carbon, plastic, and rubber have been incorporated into a number of designs illustrating some of the material combinations that are presently available.

Rubber has proved to be a surprisingly good bearing material, especially under circumstances in which abrasives may be present in the lubricant. Rubber bearings have found wide application in the stern-tube bearings of ships, on dredge cutter heads, on a number of centrifugal pumps, and for shafting bearings on deep-well pumps. The rubber used is a tough resilient compound similar in texture to that in an automobile tire. These bearings are especially effective with water lubrication, which serves as both coolant and lubricant.

Cast iron is one of the oldest bearing materials. Iron bearings were used in ancient India and China. With the advent of more complicated machinery during the industrial revolution, cast iron became a popular bearing material. It is still used where the duty is relatively light.

Porous metal bearings are frequently used when plain metal bearings are impractical because of lack of space or inaccessibility for lubrication. These bearings have voids of 16–36% of the volume of the bearing. These voids are filled with a lubricant by a vacuum technique. During operation they supply a limited amount of lubricant to the sliding surface between the journal and the bearing. In general, these

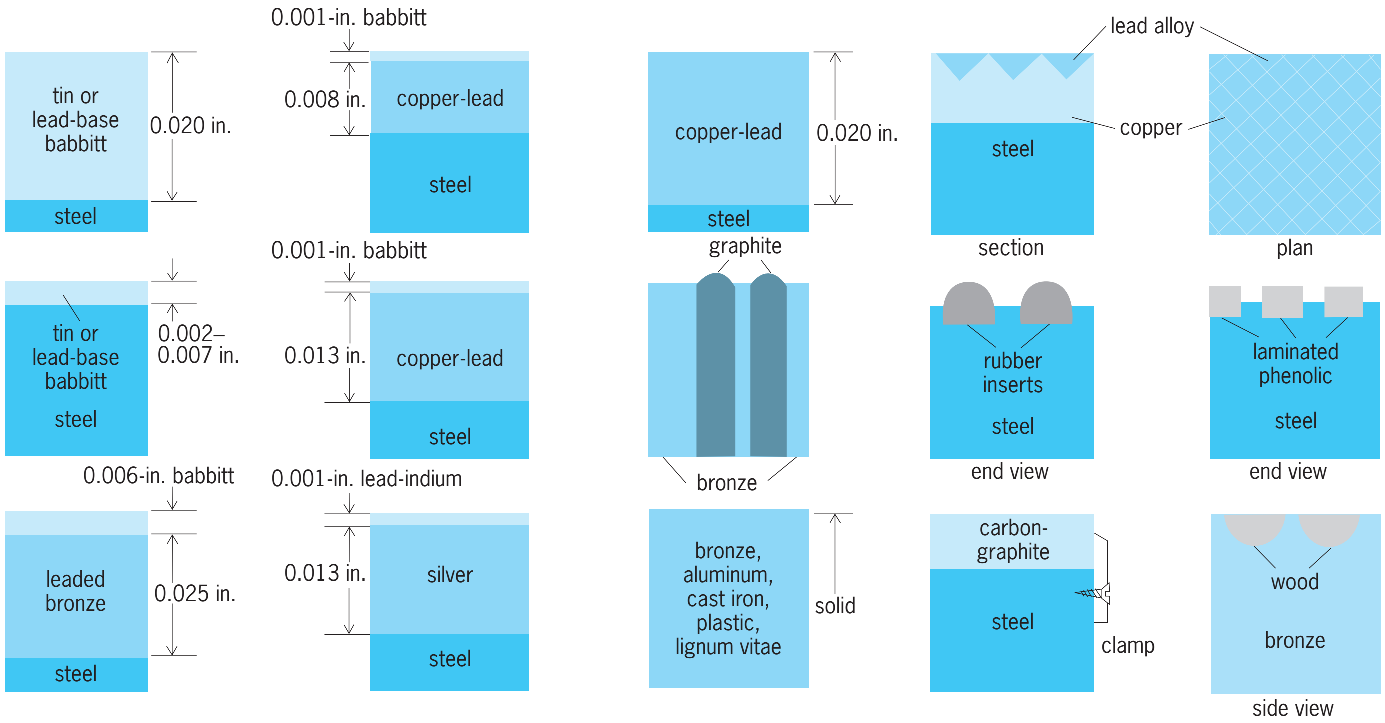

0.001-in. babbitt
lead alloy
tin or lead-base babbitt
0.020 in.
0.008 in.
copper-lead
steel
copper-lead
0.020 in.
copper
steel
steel
steel
section
plan
0.001-in. babbitt
graphite
tin or lead-base babbitt
0.002–0.007 in.
0.013 in.
copper-lead
rubber inserts
laminated phenolic
steel
steel
steel
steel
0.006-in. babbitt
0.001-in. lead-indium
bronze
end view
end view
leaded bronze
0.025 in.
0.013 in.
silver
bronze, aluminum, cast iron, plastic, lignum vitae
solid
carbon-graphite
steel
clamp
wood
bronze
steel
steel
side view

Fig. 1. **Schematic of a number of combinations of bearing materials made up of hard and soft materials. 1 in. = 2.54 cm. (*After J. J. O'Connor, ed., Power's Handbook on Bearings and Lubrication, McGraw-Hill, 1951*)**

bearings are satisfactory for light loads and moderate speeds.

In some areas recent research has shown that, surprisingly enough, very hard materials when rubbed together provide satisfactory bearing characteristics for unusual applications. Materials such as Stellite, Carboloy, Colmonoy, Hastelloy, and Alundum are used. Because of their hardness, these bearings must be extremely smooth and the geometry must be precise for there is little possibility that these materials will conform to misalignment through the process of wear.

Lubricants. Petroleum oils and greases have been fortified by chemical additives so that they are effective in reducing wear of highly stressed machine elements such as gears and cams. The additives include lead naphthenate, chlorine, sulfur, phosphorus, or similar materials. In general, compounds containing these elements are used as additives to form—through reaction with the metal surfaces—chlorides, sulfides, and phosphides which have relatively low shear strength and protect the surface from wear and abrasion.

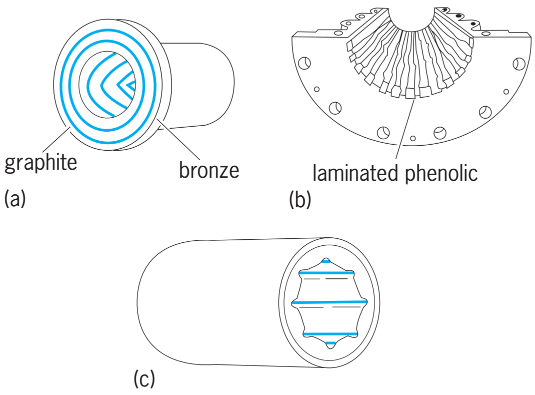

graphite
bronze
(a)
laminated phenolic
(b)
(c)

Fig. 2. **Bearings with (*a*) graphite, (*b*) wood, plastic, and nylon (*after J. J. O'Connor, ed., Power's Handbook on Bearings and Lubrication, McGraw-Hill, 1951*); (*c*) rubber (*Lucian Q. Moffitt, Inc.*).**

The method of supplying the lubricant and the quantity of lubricant which is fed to the bearing by the supplying device will often be the greatest factor in establishing performance characteristics of the bearing. For example, if no lubricant is present, the journal and bearing will rub against each other in the dry state. Both friction and wear will be relatively high. The coefficient of friction of a steel shaft rubbing in a bronze bearing, for example, may be about 0.3 for the dry state. If lubricant is present even in small quantities, the surfaces become contaminated by this material whether it is an oil or a fat, and depending upon its chemical composition the coefficient of friction may be reduced to about 0.1. Now if an abundance of lubricant is fed to the bearing so that there is an excess flowing out of the bearing, it is possible to develop a self-generating pressure film in the clearance space as indicated in **Fig. 3**. These pressures can be sufficient to sustain a considerable load and to keep the rubbing surfaces of the bearing separated. This is the type of bearing that is found on the crank shaft of a typical automobile engine. Unit pressures on these bearings reach and exceed 2500 psi (1.7×10^7 pascals) and with little attention will last almost indefinitely, provided that the oil is kept clean and free from abrasive particles

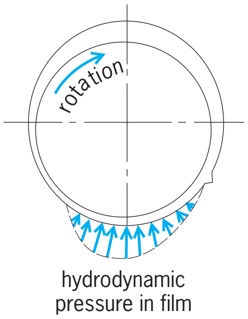


Fig. 3. Hydrodynamic fluid-film pressures in a journal bearing. (*After W. Staniar, ed., Plant Engineering Handbook, 2d ed., McGraw-Hill, 1959*)

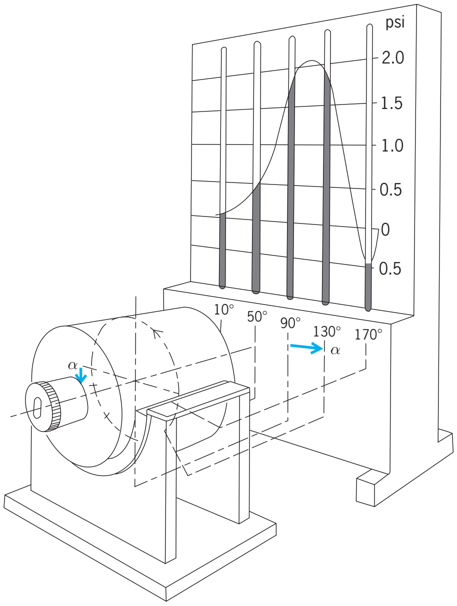


Fig. 4. Schematic of a test device for determining pressure in a journal bearing. 1 psi = 6.9 kPa. (*After H. Drescher, Air lubrication bearings, Eng. Dig., 15:103–107, 1954*)

and that the bearing materials themselves do not deteriorate from fatigue, erosion, or corrosion.

Figure 4 shows a schematic of a simple test device which indicates the pressure developed in the converging clearance space of the journal bearing. The position shown where the center of the journal is eccentric to the center of the bearing is that position which the journal naturally assumes when loaded. If the supply of lubricant is insufficient to fill the clearance space completely or if the load and speed of operation are not favorable to the generation of a complete fluid film, the film will be incomplete and there will be areas within the bearing which do not have the benefit of a fluid film to keep the rubbing surfaces apart. These areas will be only lightly contaminated by lubricant.

The types of oiling devices that usually result in insufficient feed to generate a complete fluid film are, for example, oil cans, drop-feed oilers, waste-packed bearings, and wick and felt feeders.

Oiling schemes that provide an abundance of lubrication are oil rings, bath lubrication, and forced-feed circulating supply systems. The coefficient of friction for a bearing with a complete fluid film may be as low as 0.001. **Figure 5** shows typical devices which do not usually supply a lubricant in sufficient quantity to permit the generation of a complete fluid film. **Figure 6** shows devices in which the flow rate is generally sufficient to permit a fluid film to form. Figure 6*a*, *b*, and *d* shows some typical forms of fluid-film journal bearings. The **table** shows current design practice for a number of bearings in terms of mean bearing pressure. This is the pressure applied to the bearing by the external load and is based on the projected area of the bearing. *See* ENGINE LUBRICATION; JEWEL BEARING; LUBRICANT.

Fluid-film hydrodynamic types. If the bearing surfaces can be kept separated, the lubricant no longer needs an oiliness agent, such as the fats, oils, and greases described above. As a consequence, many extreme applications are presently found in which fluid-film bearings operate with lubricants consisting of water, highly corrosive acids, molten metals, gasoline, steam, liquid refrigerants, mercury, gases, and so on. The self-generation of pressure in such a bearing takes place no matter what lubricant is used, but the maximum pressure that is generated depends upon the viscosity of the lubricant. Thus, for example, the maximum load-carrying capacity of a gas-lubricated bearing is much lower than that of a liquid-lubricated bearing. The ratio of capacities is in direct proportion to the viscosity.

Current practice in mean bearing pressure

Type of bearing	Permissible pressure, psi of projected area*
Diesel engines, main bearings	800–1500
Crankpin	1000–2000
Wristpin	1800–2000
Electric motor bearings	100–200
Marine diesel engines, main bearings	400–600
Crankpin	1000–1400
Marine line-shaft bearings	25–35
Steam engines, main bearings	150–500
Crankpin	800–1500
Crosshead pin	1000–1800
Flywheel bearings	200–250
Marine steam engine, main bearings	275–500
Crankpin	400–600
Steam turbines and reduction gears	100–220
Automotive gasoline engines, main bearings	500–1000
Crankpin	1500–2500
Air compressors, main bearings	120–240
Crankpin	240–400
Crosshead pin	400–800
Aircraft engine crankpin	700–2000
Centrifugal pumps	80–100
Generators, low or medium speed	90–140
Roll-neck bearings	1500–2500
Locomotive crankpins	1500–1900
Railway-car axle bearings	300–350
Miscellaneous ordinary bearings	80–150
Light line shaft	15–25
Heavy line shaft	100–150

*1 psi = 6.9 kPa.

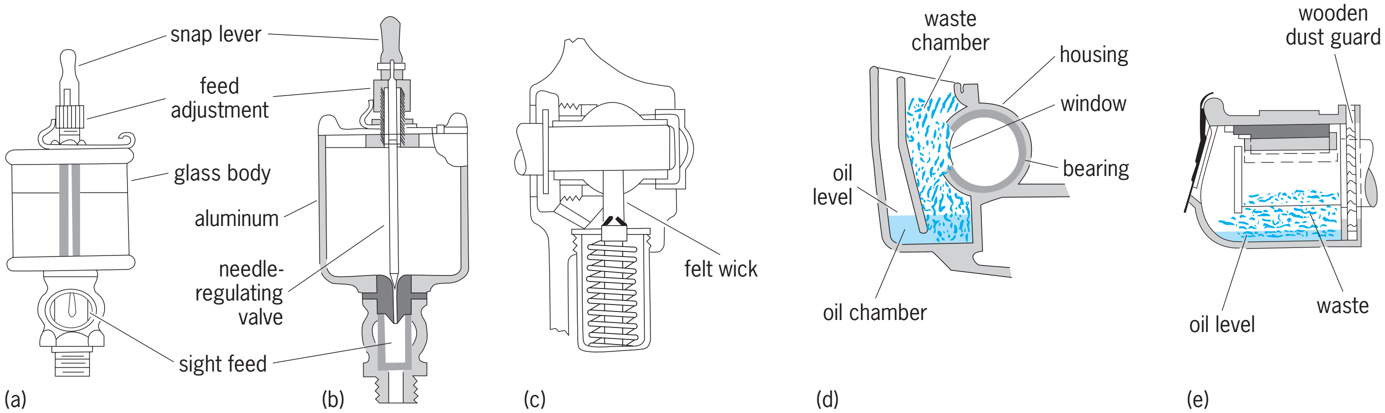


Fig. 5. Schematic showing some of the typical lubrication devices. (*a, b*) Drop-feed oilers. (*c*) Felt-wick oiler for small electric motors. (*d*) Waste-packed armature bearing. (*e*) Waste-packed railroad bearing. (*After W. Staniar, ed., Plant Engineering Handbook, 2d ed., McGraw-Hill, 1959*)

Considerable research has been directed toward the operation of machinery at extremes of temperature. On the low end of the scale this may mean −400°F (−240°C). On the upper end of the scale expectation is that some devices may be called upon to function at 2000–3000°F (1100–1600°C). Gas is the only presently known lubricant that could possibly be used for such extremes of temperature. Because the viscosity of gas is so low, the friction generated in the bearing is correspondingly of a very low order. Thus gas-lubricated machines can be operated at extremely high speeds because there is no serious problem in keeping the bearings cool. A rotor system has been operated on gas-lubricated bearings up to 433,000 revolutions per minute.

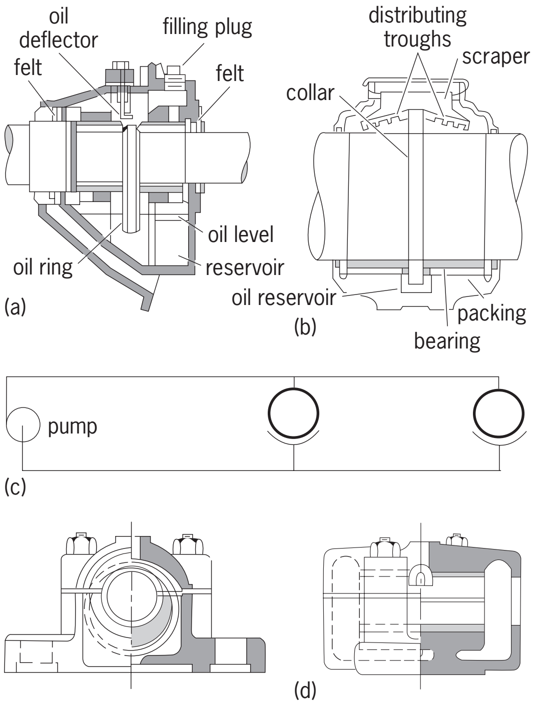


Fig. 6. Lubrication devices. (*a*) Ring-oiled motor bearings. (*b*) Collar-oiled bearing (*after W. Staniar, ed., Plant Engineering Handbook, 2d ed., McGraw-Hill, 1959*). (*c*) Circulating system for oiling bearings (*after J. J. O'Connor, ed., Power's Handbook on Bearings and Lubrication, McGraw-Hill, 1951*). (*d*) Rigid ring-oiling pillow block (*after T. Baumeister, ed., Standard Handbook for Mechanical Engineers, 8th ed., McGraw-Hill, 1978*).

The self-generating pressure principle is applied equally as well to thrust bearings as it is to journal bearings. One of the first commercial applications of the self-generating type of bearing was in the tilting-pad type of thrust bearing. One tilting pad is shown schematically in **Fig. 7*a***. There is certainly no question regarding the great value of the tilting-pad thrust bearing. It excels in low friction and in reliability. A model of a typical commercial thrust bearing is shown in Fig. 7*b*. The thrust bearing is made up of many tilting pads located in a circular position. One of the largest is on a hydraulic turbine at the Grand Coulee Dam. There, a bearing 96 in. (244 cm) in diameter carries a load of 2.15×10^6 lb (976,000 kg) with a coefficient of friction of about 0.0009. Large marine thrust bearings are of this type and transfer the entire thrust of the propeller to the hull of the ship.

Fluid-film hydrostatic types. Sleeve bearings of the self-generating pressure type, after being brought up to speed, operate with a high degree of efficiency and reliability. However, when the rotational speed of the journal is too low to maintain a complete fluid film, or when starting, stopping, or reversing, the oil film is ruptured, friction increases, and wear of the bearing accelerates. This condition can be eliminated by introducing high-pressure oil to the area between the bottom of the journal and the bearing itself, as shown schematically in **Fig. 8**. If the pressure and quantity of flow are in the correct proportions, the shaft will be raised and supported by an oil film whether it is rotating or not. Friction drag may drop to one-tenth of its original value or even less, and in certain kinds of heavy rotational equipment in which available torque is low, this may mean the difference between starting and not starting. This type of lubrication is called hydrostatic lubrication and, as applied to a journal bearing in the manner indicated, it is called an oil lift. Synchronous condensers need the oil lift when the unit is of large size. Rolling-mill and foil-mill bearings may be equipped with an oil lift

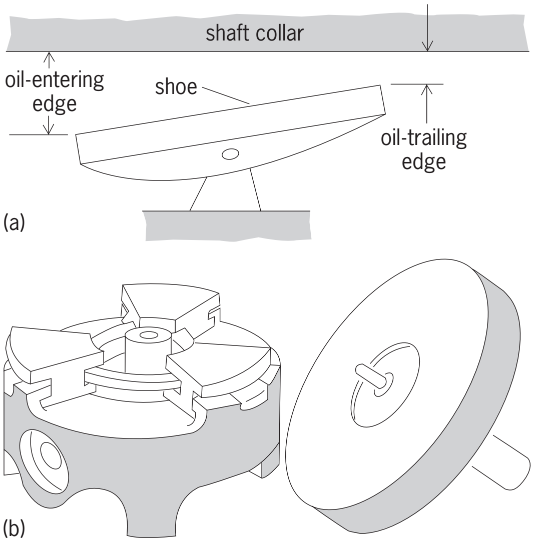


Fig. 7. Tilting-pad-type bearing. (*a*) Schematic of tilting pad (*after W. Staniar, ed., Plant Engineering Handbook, 2d ed., McGraw-Hill, 1959*). (*b*) Thrust bearing (*after D. D. Fuller, Theory and Practice of Lubrication for Engineers, Wiley, 1956*).

to reduce starting friction when the mills are under load. Occasionally, hydrostatic lifts are used continuously on bearings that are too severely overloaded to maintain a normal hydrodynamic or self-pressurizing oil film.

Hydrostatic lubrication in the form of a step bearing has been used on various machines to carry thrust. Such lubrication can carry thrust whether the shaft is rotating or not and can maintain complete separation of the bearing surfaces. **Figure 9** shows a schematic representation of such a step-thrust bearing. High-speed machines such as ultracentrifuges have used this principle by employing air as the lubricant at speeds upward of 100,000 rpm.

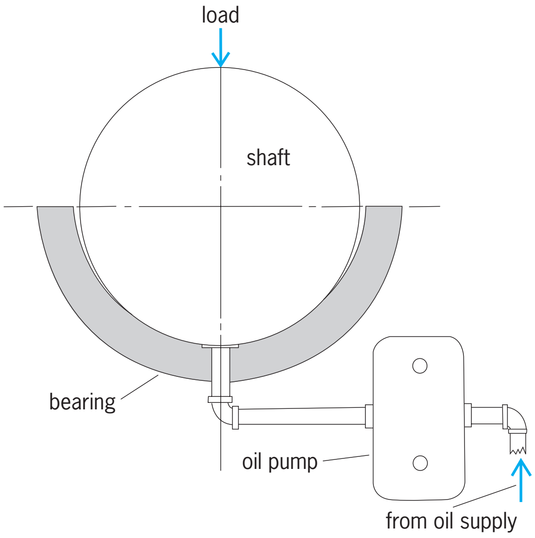


Fig. 8. Fluid-film hydrostatic bearing. Hydrostatic oil lift can reduce starting friction drag to less than one-tenth of usual starting drag. (*After W. Staniar, ed., Plant Engineering Handbook, 2d ed., McGraw-Hill, 1959*)

Large structures have been floated successfully on hydrostatic-type bearings. For example, the Hale 200-in. (5-m) telescope on Palomar Mountain weighs about 1×10^6 lb (500,000 kg); yet the coefficient of friction for the entire supporting system, because of the hydrostatic-type bearing, is less than 0.000004. The power required is extremely small and a $^{1}/_{12}$-hp clock motor rotates the telescope while observations are being made. Hydrostatic bearings are currently being applied to large radio telescopes and radar antennas, some of which must sustain forces of 5×10^6 lb (2×10^6 kg) or more, considering wind loads as well as dead weight. One such unit constructed at Green Bank, West Virginia, by the Associated Universities has a parabolic disk or antenna 140 ft (42 m) in diameter.

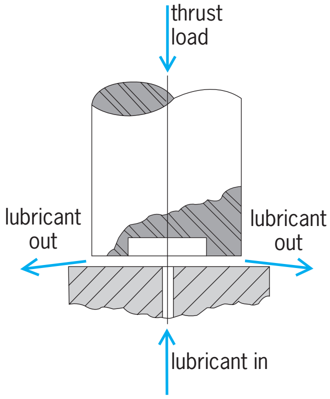


Fig. 9. Step bearing. (*After W. Staniar, ed., Plant Engineering Handbook, 2d ed., McGraw-Hill, 1959*)

Rolling-element types. Everyday experiences demonstrate that rolling resistance is much less than sliding resistance. The wheelbarrow, the two-wheeled baggage truck, and similar devices are striking examples of the reduction in friction by the use of the wheel. Heavy crates and similar objects are easily moved by introducing rollers under the leading edge of the load while the crate or object is pushed along. Egyptian engineers building the pyramids transported huge blocks of stone from the quarry to the building site by means of rollers. This principle is used in the rolling-element bearing which has found wide use.

The first major application of these bearings was to the bicycle, the use of which reached its peak just before the year 1900. In the development of the automobile, ball and roller bearings were found to be ideal for many applications, and today they are widely used in almost every kind of machinery.

Structure. These bearings are characterized by balls or cylinders confined between outer and inner rings. The balls or rollers are usually spaced uniformly by a cage or separator. The rolling elements are the most important because they transmit the loads from the moving parts of the machine to the stationary supports. Balls are uniformly spherical, but the rollers may be straight cylinders, or they may be barrel- or cone-shaped or of other forms, depending upon

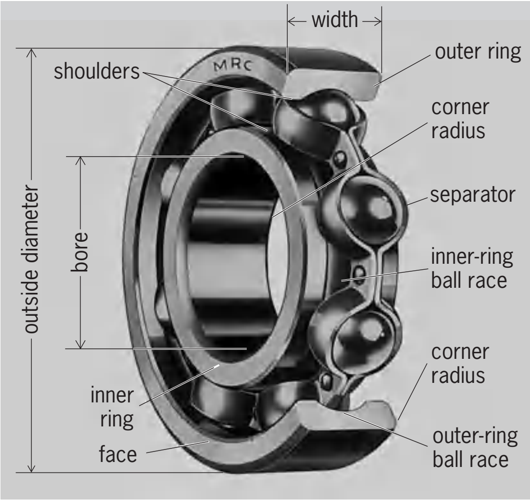


Fig. 10. Deep-groove ball bearing. (*Marlin-Rockwell*)

the purpose of the design. The rings, called the races, supply smooth, hard, accurate surfaces for the balls or rollers to roll on. Some types of ball and roller bearings are made without separators. In other types there is only the inner or outer ring, and the rollers operate directly upon a suitably hardened and ground shaft or housing. **Figure 10** shows a typical deep-grooved ball bearing, with the parts that are generally used.

These bearings may be classified by function into three groups: radial, thrust, and angular-contact bearings. Radial bearings are designed principally to carry a load in a direction perpendicular to the axis of rotation. However, some radial bearings, such as the deep-grooved bearings shown in Fig. 10, are also capable of carrying a thrust load, that is, a load parallel to the axis of rotation and tending to push the shaft in the axial direction. Some bearings, however, are designed to carry only thrust loads. Angular-contact bearings are especially designed and manufactured to carry heavy thrust loads and also radial loads.

Life. A unique feature of rolling-element bearings is that their useful life is not determined by wear but by fatigue of the operating surfaces under the repeated stresses of normal use. Fatigue failure, which occurs as a progressive flaking or pitting of the surfaces of the races and rolling elements, is accepted as the basic reason for the termination of the useful life of such a bearing.

Because the load on a bearing determines the severity of the stress in the surfaces of the races and the rolling elements, it follows that if the load is increased, the life of the bearing will be decreased, and conversely if the load is decreased, the life of the bearing will be increased. This is usually expressed by the relationship that the life of the bearing is inversely proportional to the load cubed. Thus, doubling the load will reduce the life of the bearing by a factor of 8.

The life of a bearing under a given load will therefore be a certain number of revolutions. If this number of revolutions is used up at a relatively high rate, the life of the bearing will be correspondingly short. If the total number of revolutions is used up at a low rate, the life of the bearing is correspondingly longer; that is, the life is inversely proportional to the speed. Life expectancy is a prediction based on results obtained from tests of a large number of identical bearings under identical loads and speeds. Individual bearings may deviate from this figure on a statistical basis, but manufacturers have followed the law of averages in establishing their ratings. For example, some manufacturers specify load and speed ratings for their bearings based on 3000 h of operation. The bearing manufacturer who uses 3000 h as a life-expectancy figure assumes that at least 90% of all bearings will last 3000 h under the specified conditions of load and speed. Based on statistical averages, however, this means that 10% of the bearings will fail before reaching the 3000-h life expectancy, 50% of the bearings will attain five times the design life, and a few bearings may reach 20–30 times the design life of 3000 h.

Characteristics. The various types of rolling-contact bearings can be identified in terms of their broad general characteristics (**Fig. 11**). A separable or a magneto-type bearing is useful where disassembly is frequent (Fig. 11*a*). The outer race is pressed firmly in the removable housing; the inner race may be pressed against a shoulder on the shaft. A deep-grooved bearing with a filling slot (Fig. 11*b*) allows more balls to be used in the bearing than are shown in Fig. 10 and will thus carry heavier radial loads. Because of the filling slot, however, it should be used only for light thrust loads. If the thrust is in one direction, the bearing should be mounted with the slot away from the thrust direction. The double-row radial bearing with deep grooves handles heavier radial and thrust loads than a single-roll bearing of the same dimensions. Internal self-aligning double-roll bearings (Fig. 11*d*) may be used for heavy radial loads where self-alignment is required. The self-aligning feature should not be used to correct poor design or assembly because excessive misalignment will harm the bearing. Thrust loads should be light because a thrust load will be sustained by only one row of balls. Figure 11*e* shows an external self-aligning bearing which requires a larger outside diameter, but has the advantage of being able to carry thrust in either direction, as well as providing a self-aligning feature. Angular contact bearings (Fig. 11*f*) provide for a maximum thrust and modest radial loads. They may be mounted back to back as duplex bearings and carry thrust in either direction. To minimize axial movement of such a bearing and the shaft that it restrains, these bearings may be preloaded to take up any possible slack or clearance. A ball bushing is shown in Fig. 11*g*. This is used for translational motion or motion of a shaft in its axial direction. Any such motion should be at a very low speed and with light radial loads. Roller bearings with short straight rollers are shown in Fig. 11*h*. They permit free movement in the axial direction when the guide lips are on either the outer or inner race. Needle bearings (Fig. 11*i*)

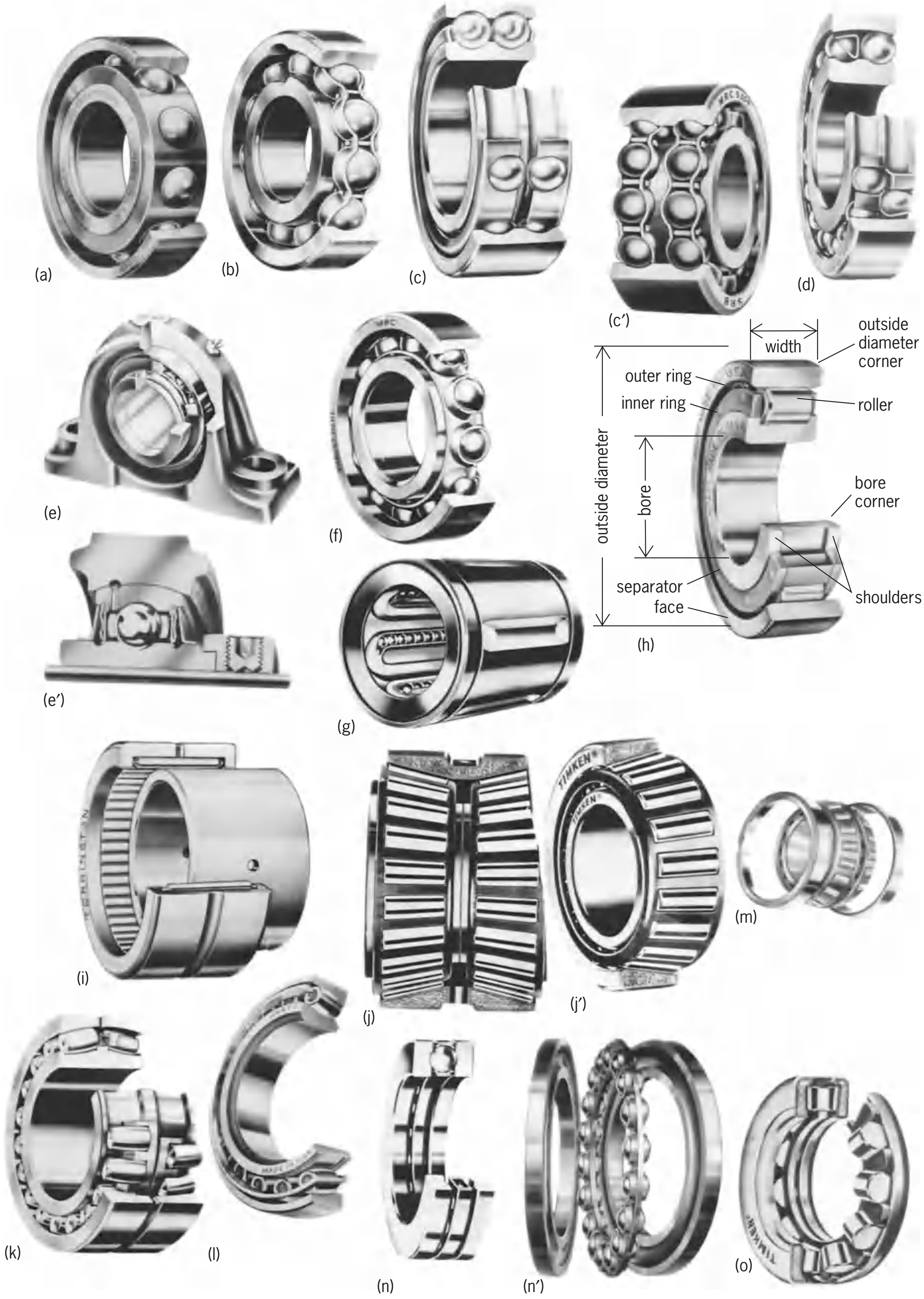

(a)
(b)
(c)
(c′)
(d)
(e)
(e′)
(f)
(g)
width
outside diameter corner
outer ring
inner ring
roller
outside diameter
bore
bore corner
separator
face
shoulders
(h)
(i)
(j)
(j′)
(m)
TIMKEN
(k)
(l)
(n)
(n′)
(o)

Fig. 11. Views of the various ball and roller bearings. (*a*) Separable angle-contact ball bearing (*New Departure*). (*b*) Deep-groove ball bearing with filling slot (*New Departure*). (*c*) Double-row deep-groove ball bearings (*SKF; Marlin-Rockwell*). (*d*) Internal self-aligning bearing (*SKF*). (*e*) Deep-groove ball bearing with spherical outer ring. (*f*) Standard angular-contact ball bearing (*Marlin-Rockwell*). (*g*) Ball bushing that permits unlimited travel, linear motion (*Thomson Industries*). (*h*) Roller bearing showing nomenclature (*Hyatt*). (*i*) Needle roller bearing (*Torrington*). (*j*) Tapered roller bearings, two-row and single-row (*Timken*). (*k*) Radial spherical roller bearings (*Torrington*). (*l*) Barrel roller bearing (*Hyatt*). (*m*) Concavex roller bearing (*Shafer*). (*n*) Single-direction ball-bearing thrust (*SKF; Marlin-Rockwell*). (*o*) Cylindrical roller thrust bearing (*Timken*).

have rollers whose length is at least four times the diameter. They may be furnished with a retainer but have their maximum load capacity when there is no retainer and the clearance space is completely filled with rollers. They are with or without an inner race and are most useful where space must be saved. If the shaft is used as an inner race, it should be hardened.

The tapered roller bearings shown in Fig. 11*j* permit heavy radial and thrust loads. These may be procured as double tapered rollers with two inner races and one outer or two outer races and one inner. The race cones intersect at the axis of the bearing. Figure 11*k* shows a self-aligning bearing with two rows of short barrel rollers. The spherical outer race is one piece, or it may be in two parts for preloading. The thrust is taken on the spherical surfaces of the center flange of an inner race. Figure 11*l* shows a barrel roller with spherical outer race. Figure 11*m* is a self-aligning bearing with hourglass rollers. This bearing is also built as a two-row bearing with a one-piece inner race.

The ball thrust bearing shown in Fig. 11*n* is used only for carrying thrust loads acting in the direction of the axis of the shaft. It is used for low-speed applications, while other bearings must support the radial load. Straight roller thrust bearings are made of a series of short rollers to minimize twisting or skewing. The tapered roller thrust bearing shown in Fig. 11*o* eliminates the twisting that may take place with the straight rollers but cause a thrust load between the ends of the rollers and the shoulder on the race.

Mountings. Many types of mountings are available for rolling-element bearings, and their selection will depend upon the variety of service conditions encountered. In the preferred method of mounting, the inner and outer races are held securely in place to prevent creeping or spinning of the races in the housing or on the shaft. In rigid mountings, provision must be made for expansion either by building slip clearances into one mount or by using a straight roller bearing on the free end of the shaft. A bearing can be mounted on a shaft with a snap ring on the outer race which locates the bearing against a shoulder. Bearings may also be mounted on shafts by means of an eccentric ring cut on the side of the extended inner race. A set screw can be used to hold the bearings in place. *See* FRICTION; WEAR.

Dudley D. Fuller.

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; J. E. Shigley and C. R. Mischke, *Bearings and Lubricaton*: *A Mechanical Designers' Handbook*, 1990.

Antigen

Any substance that causes the immune system to produce specific antibodies or T cells against it. An antigen may be a foreign substance from the environment (such as chemicals, bacterial or viral proteins, or pollen) or formed within the host's own body. Reactions of antigens with antibodies or T cells serve as a defense against microorganisms and other foreign bodies, but can be detrimental if the immune response is mounted against the "self," as in autoimmune disorders. Antigen-antibody complexes are used in laboratory tests for detecting the presence of either antigen or antibody to determine a subject's previous exposure to pathogens. *See* ANTIBODY; ANTIGEN-ANTIBODY REACTION.

Classes of antigens. Antigens can be classified into three distinct groups: immunogens, allergens, and tolerogens.

Immunogens. An immunogen is any substance that provokes an immune response. It is usually a protein or polysaccharide, and its ability to stimulate the immune response depends on many factors, including its foreignness, size, chemical composition, and conformation. Peptide sequences with high segmental mobility are frequently antigenic. Because an immunogen is rather large, it usually contains a large number of antigenic determinants, or epitopes (sites recognized by antibodies). Immunizing an animal with a protein results in the formation of numerous antibody molecules with different specificities, the number of different antibodies depending on the number of antigenic determinants. Since the introduction of technology for producing monoclonal antibodies, it has become possible to produce in the laboratory antibodies specific to virtually any antigen. *See* MONOCLONAL ANTIBODIES.

Immunogens may be introduced into an animal by ingestion, inhalation, contact with skin, or by injection into the bloodstream, skin, peritoneum, or other tissue.

Allergens. Allergens are substances that cause an allergic reaction by preferentially driving the immune system to an immunoglobulin E (IgE) response. Allergens can be ingested, inhaled, injected, or come into contact with skin. Allergens include dust mites, dander, peanuts, shellfish, and many other common environmental agents. *See* ALLERGY; IMMUNOGLOBULIN.

Tolerogens. Tolerogens are antigens that cause a specific immune non-responsiveness due to their molecular form. They are thought to provide only partial signals to the T or B cells, which respond by shutting down immune functions. Tolerogens are usually "self," that is, produced by the host's own body, which explains why, in general, we do not have an immune response to proteins normally made in the body. When this system does not work it is termed autoimmunity. *See* AUTOIMMUNITY; CELLULAR IMMUNITY.

Antibody response. Since antigens are complex molecules, the antibody response to them may also be complex. A mixture of antibodies may be produced, each corresponding to only a portion (epitope) of the original antigen surface (antigenic mosaic). If the antigenic entity is not a single molecular species but, as in the case of whole microorganisms, a mixture of separate antigens, the complexity is multiplied accordingly. A small part of an antigen may stimulate an antibody capable of recognizing not only that antigen but related antigens; this is known as a cross-reaction. The most famous example of a cross-reaction came from Edward Jenner's discovery that milkmaids infected with cowpox were protected from smallpox.

Sources of antigens. Bacteria, viruses, protozoans, and other microorganisms are important sources of antigens. These may be proteins or polysaccharides derived from the outer surfaces of the cell (capsular antigens), from the cell interior (somatic or O antigens), or from the flagella (flagellar or H antigens). Other antigens are either secreted by the cell or are released during cell death and disruption; these include many enzymes and toxins, of which diphtheria, tetanus, and botulinum toxins are important examples. The presence of antibody to one of these constituent antigens in human or animal sera is presumptive evidence of past or present contact with specific microorganisms, and this finds application in clinical diagnosis and epidemiological surveys. *See* BOTULISM; DIPHTHERIA; TETANUS; TOXIN.

Use in vaccines. In medicine, the ability of antigen to stimulate the production of antibodies or T cells is the basis of vaccines. In traditional antigen-based vaccines, the "target" antigens of a microorganism are produced and deliberately administered to induce an antibody response. There are four types of antigen-based vaccines: live, purified, recombinant, and synthetic. Live vaccines are made from viruses that have been attenuated (made nonpathogenic) by long culture in cells or animals until the dangerous genes have mutated. They are among our most effective vaccines. Purified antigen vaccines are those that are composed of molecules purified directly from a pathogen. Because these vaccines use a limited number of molecules from the parent pathogen, there is no danger of pathogen replication. Recombinant antigen vaccines are created by insertion of deoxyribonucleic acid (DNA) encoding an antigenic protein into bacteria, yeast, or viruses that infect cultured mammalian cells, or by direct transfection of the DNA into the mammalian cells. The organisms or cells will then produce the protein, which can be further purified for use as a vaccine. Synthetic antigen vaccines are those that use peptide antigens synthesized by automated machines. While the technology exists, this method of making vaccines is not yet in use. *See* VACCINATION.

Vaccines have been made from protein antigens derived from cancer cells. These aim to stimulate the patient's immune system to attack the cancer. Scientists have recently worked out the genetic codes of many of these cancer cell proteins, so they can make them in the laboratory in large quantities.

Margaret J. Polley; Zoë Cohen

Bibliography. J. Chinen and W. Shearer, Basic and clinical immunology, *J. Allergy Clin. Immunol.*, 116(2):411–418, 2005; T. Doan, R. Melzold, and C. Waltenbaugh, *Concise Medical Immunology*, Lippincott Williams & Wilkins, 2005; C. A. Janeway, Jr. et al., *Immunobiology*, Garland Publishing, 6th ed., 2004; W. E. Paul (ed.), *Fundamental Immunology*, Lippincott Williams & Wilkins, 5th ed., 2003; M. Tosi, Innate immune responses to infection, *J. Allergy Clin. Immunol.*, 116(2):241–249, 2005.

Antigen-antibody reaction

A reaction that occurs when an antigen combines with a corresponding antibody to produce an immune complex. A substance that induces the immune system to form a corresponding antibody is called an immunogen. All immunogens are also antigens because they react with corresponding antibodies; however, an antigen may not be able to induce the formation of an antibody and therefore may not be an immunogen. For instance, lipids and all low-molecular-weight substances are not immunogenic. However, many such substances, termed haptens, can be attached to immunogens, called carriers, and the complex then acts as a new immunogen capable of eliciting antibody to the attached hapten. *See* ANTIBODY; ANTIGEN.

Antibody-antigen specificity. An antibody, which belongs to a group of proteins in blood plasma called immunoglobulins, consists of two functionally and structurally distinct portions. The major portion is structurally identical for all antibodies of its immunoglobulin class and takes no part in the reaction with antigen. The other part varies markedly between antibodies and is responsible for binding antigen; in fact, its variability confers on it specificity, that is, the ability to react with one specific antigen among many. *See* IMMUNOGLOBULIN.

A molecule of antibody has two identical binding sites for one antigen or more, depending on its class. Each site is quite small and can bind only a comparably small portion of the surface of the antigen, which is termed an epitope. The specificity of an antibody for an antigen depends entirely upon the possession of the appropriate epitope by an antigen. The binding site on the antibody and the epitope on the antigen are complementary regions on the surface of the respective molecules which interlock in the antigen-antibody reaction. If a similar epitope occurs on a molecule other than the original immunogen, the antibody will bind to that molecule in a cross-reaction. The intensity with which an antibody binds to the antigen depends on the exactitude of the fit between the respective binding site and epitope, as well as some inherent characteristics of the reacting molecules and factors in the environment. The epitope must be continuous spatially, but not structurally: in other words, if the molecule of the antigen consists of several chains, like several pieces of string that have been dropped on top of one another, then an epitope may be formed by adjacent regions on two different chains, as well as by adjacent regions on the same chain. If the epitope is now modified either chemically (for example, by altering the hapten) or physically (for example, by causing the chains to separate), then its fit in the binding site will be altered or abolished, and the antigen will react with the antibody either less strongly or not at all.

Nature of immune complex. The immune complex formed in the reaction consists of closely apposed, but still discrete, molecules of antigen and antibody. Therefore, the immune complex can dissociate into the original molecules. The proportion of the dissociated, individual molecules of antigen and antibody to those of the immune complex clearly depends on the intensity of the binding. These proportions can be measured in a standardized procedure, so that the concentration of antigen [Ag], antibody [Ab], and the immune complex [AgAb] becomes known. A fraction is then calculated and called either the dissociation constant ($K_d = [\mathrm{Ag}] \times [\mathrm{Ab}]/[\mathrm{AgAb}]$) or the association constant ($K_a = [\mathrm{AgAb}]/[\mathrm{Ag}] \times [\mathrm{Ab}]$). The magnitude of either of these constants can be used subsequently to assess the intensity of the antigen-antibody reaction, for example, to select the most strongly binding of several antibodies for use in immunoassay; to compare the binding of an antibody to an immunogen with that to the corresponding hapten; or to detect modification in the epitope. *See* IMMUNOASSAY.

Only one epitope of its kind generally occurs on each molecule of antigen, other than that which consists of multiple, identical units, though many epitopes of different configuration are possible. Particles, however, either natural ones such as cells or suitably treated artificial ones made of, for example, latex or glass, typically carry multiple identical epitopes, as well as nonidentical ones, because their surfaces contain many molecules of the same antigen. An antibody molecule, bearing two or more binding sites, can become attached to a corresponding number of antigen molecules. If the antibody is specific for only one epitope, as is the case with monoclonal antibody, then it can link with one molecule of antigen at each binding site, but the antigen cannot link with yet another antibody, because it has no more identical epitopes. In a solution of single molecules of antigen, therefore, monoclonal antibody can form at most complexes no larger than those composed of one antibody molecule and the number of antigen molecules corresponding to that of the number of binding sites. By contrast, polyclonal antibody to an antigen consists of antibodies to a variety of epitopes on that antigen, and it can, therefore, link together a very large number of molecules. *See* MONOCLONAL ANTIBODIES.

Immune complexes comprising many molecules eventually reach sufficient size to scatter light, at which point they can be detected by nephelometry or turbidimetry; if their growth continues, they become visible as precipitates, which can also be assayed by such methods as immunodiffusion. Since particles typically carry many molecules of antigen, they can be, in principle, aggregated by monoclonal antibody as well as polyclonal antibody, though the latter is much more effective; and the reaction can be detected by inspection, for example the agglutination of bacteria or red cells, though much more sensitive methods, such as particle counting, are now available. Antigen-antibody reactions can also be detected at very low concentration of reactants through special techniques such as immunofluorescence and radioimmunoassay. *See* IMMUNOASSAY; IMMUNOFLUORESCENCE; IMMUNONEPHELOMETRY; RADIOIMMUNOASSAY.

The reaction between antigen and antibody is followed by a structural change in the remainder of the antibody molecule. The change results in the appearance of previously hidden regions of the molecule. Antibodies to these hidden regions are known as rheumatoid factors owing to their frequently high concentration in the blood of individuals suffering from rheumatoid arthritis. Some of these hidden regions have specific functions, however, such as binding complement. Fixation of complement by immune complexes has been used to detect and measure antigen-antibody reactions. *See* COMPLEMENT.

Factors affecting reaction. The antigen-antibody reaction is affected by various factors, including pH, temperature, and concentration, as well as the type of other dissolved substances. Acidic or markedly alkaline solutions, excessively high or low salt concentration, and the presence of chaotropic substances all prevent the formation of immune complexes or cause the dissociation of preformed ones. Substances that reduce charge on particles, allowing them to aggregate more easily, or reduce free water in solution, for example, polymeric hydrophilic materials such as poly(ethylene glycol), assist the formation of immune complexes.

Uses. The chief use of antigen-antibody reactions has been in the determination of blood groups for transfusion, serological ascertainment of exposure to infectious agents, and development of immunoassays for the quantification of various substances. *See* BLOOD GROUPS; IMMUNOLOGY; SEROLOGY.

Alexander Baumgarten

Bibliography. B. Cinader and R. G. Miller (eds.), *Progress in Immunology VI*, 1986; E. L. Cooper, *General Immunology*, 1982; E. A. Kabat, *Structural Concepts in Immunology and Immunochemistry*, 2d ed., 1976; N. R. Rose, H. Friedman, and J. L. Fahey, *Manual of Clinical Laboratory Immunology*, 3d ed., 1986.

Antihistamine

A type of drug that inhibits the combination of histamine with histamine receptors. These drugs are termed either H-1 or H-2 receptor antagonists depending on which type of histamine receptor is involved. H-1 receptor antagonists are used largely for treating allergies, and H-2 receptor antagonists are used to treat peptic ulcer disease and related conditions.

H-1 receptor antagonists. H-1 receptor antagonists have been available for many years and include agents such as diphenhydramine, chlorpheniramine, doxylamine, cyclizine, tripelennamine, promethazine, and terfenadine. Their primary therapeutic use is to antagonize the effects of histamine that is released from cells by antigen-antibody reactions; they can thus inhibit histamine-induced effects, such as bronchoconstriction, skin reactions, for example, wheals and itching, and nasal inflammation. These drugs, therefore, are quite effective in reducing allergy signs and symptoms, especially if they are administered before contact with the relevant antigen (for example, ragweed or other pollens); however they are not effective in treating asthma. *See* ALLERGY; ASTHMA.

Other uses of H-1 receptor antagonists may involve actions having little or nothing to do with their ability to antagonize histamine. For example, some (diphenhydramine, cyclizine, meclizine) are effective as anti-motion-sickness drugs, and some (diphenhydramine, doxylamine, pyrilamine) are important components of over-the-counter sleep aids. Although preparations that are marketed as being useful in treating the common cold often contain H-1 receptor antagonists, these drugs have no specific actions against the cold virus.

H-1 receptor antagonists have low toxicity. The chief adverse effect is sedation. However, this effect varies widely, both among the drugs and from individual to individual; in young children excitement may be seen. Another common set of effects caused by many of these drugs, including dry mouth, blurred vision, and urinary retention, can be ascribed to their anticholinergic actions. Agents such as terfenadine and loratadine have weaker sedative and anticholinergic actions than do diphenhydramine and doxylamine. Overdoses of H-1 receptor antagonists may be associated with excitement or depression, and although there is no pharmacologic antidote for these drugs, good supportive care should be adequate in managing cases of poisoning. *See* SEDATIVE.

H-2 receptor antagonists. H-2 receptor antagonists are much newer, but they rapidly became some of the most frequently prescribed drugs. Histamine stimulates gastric acid secretion by combining with H-2 receptors. By preventing this combination, H-2 antagonists can reduce acid secretion in the stomach, an effect that makes these drugs useful in managing various conditions.

The medical treatment of peptic ulcer disease, a common disorder in the Western world, involves giving antacids and other medication to reduce gastric acidity. The effectiveness of H-2 antagonists in reducing gastric acid has made these drugs a primary treatment for peptic ulcer. In patients with active ulcers, these drugs reduce symptoms and promote healing of the ulcers. They also prevent recurrence of ulcers. All of the H-2 antagonists (cimetidine, ranitidine, famotidine, nizatidine) appear equally effective in reducing gastric acidity. *See* ULCER.

Other conditions in which H-2 antagonists are used to lower gastric acidity include reflux esophagitis, stress ulcers, and hypersecretory states such as the Zollinger-Ellison syndrome, in which tumor cells secrete large amounts of the hormone gastrin, which stimulates gastric acid secretion. In these conditions, administration of H-2 antagonists reduces symptoms and promotes healing.

The toxicity of H-2 antagonists is quite low, and adverse effects are reported by only 1–2% of patients. The most common side effects are gastrointestinal upsets, including nausea, vomiting, and diarrhea.

One effect associated only with cimetidine is the ability to inhibit the metabolism of other drugs. Because actions of drugs taken concurrently with cimetidine may thus be enhanced or prolonged, dosages of those other drugs may have to be reduced in order to prevent toxicity. *See* HISTAMINE. Alan Burkhalter

Bibliography. B. Ulnas et al. (eds.), *Histamine and Histamine Antagonists*, 1991.

Antimatter

Matter which is made up of antiparticles. At the most fundamental level every type of elementary particle has its anti-counterpart, its antiparticle. The existence of antiparticles was implied by the relativistic wave equation derived in 1928 by P. A. M. Dirac in his successful attempt to reconcile quantum mechanics and special relativity. The antiparticle of the electron (the positron) was first observed in cosmic rays by C. D. Anderson in 1932, while that of the proton (the antiproton) was produced in the laboratory and observed by E. Segré, O. Chamberlain, and their colleagues in 1955. *See* ELECTRON; ELEMENTARY PARTICLE; POSITRON; PROTON; QUANTUM MECHANICS; RELATIVITY.

Properties of antiparticles. The mass, intrinsic angular momentum (spin), and lifetime (in the case of unstable particles) of antiparticles and their particles are equal, while their electromagnetic properties, that is, charge and magnetic moment, are equal in magnitude but opposite in sign. Some neutrally charged particles such as the photon and π° meson are their own antiparticles. Certain other abstract properties such as baryon number (protons and neutrons are baryons and have baryon number +1) and lepton number (electrons and muons are leptons and have lepton number +1) are reversed in sign between particles and antiparticles. *See* ANGULAR MOMENTUM; BARYON; LEPTON.

The quantum-mechanical operation of turning particles into their corresponding antiparticles is termed charge conjugation (*C*), that of reversing the handedness of particles is parity conjugation (*P*), and that of reversing the direction of time is time reversal (*T*). A fundamental theorem, the *CPT* theorem, states that correct theories of particle physics must be invariant under the simultaneous operation of *C*, *P*, and *T*. Simply put, the description of physics in a universe of antiparticles with opposite handedness where time runs backward must be the same as the description of the universe. One consequence of the *CPT* theorem is that the above-mentioned properties of antiparticles (mass, intrinsic angular momentum, lifetime, and the magnitudes of charge and magnetic moment) must be identical to those properties of the corresponding particles. This has been experimentally verified to a high precision in many instances; for example, the magnitude of the magnetic moments of positrons and electrons are equal to within 1 part in 10^{12}, the charge-to-mass ratios of antiprotons and protons are the same to within 1 part in 10^9, and the masses of the K^0 meson and its antiparticle are equal to 1 part in 10^{18}. *See* CPT THEOREM; PARITY (QUANTUM MECHANICS); SYMMETRY LAWS (PHYSICS); TIME REVERSAL INVARIANCE.

It is presently thought that leptons (electrons, muons, tau leptons, and their associated neutrinos) are fundamental particles. Baryons and mesons are composite particles, being made of yet more fundamental particles, quarks. As with all elementary particles, for every type of quark there is a corresponding antiquark. Baryons are composed of three quarks, and mesons of quark-antiquark pairs. Antibaryons are then made up of three antiquarks, while antimesons are made of the conjugate antiquarkquark pair of the associated meson. *See* BARYON; MESON; NEUTRINO; QUARKS.

Annihilation and production. When a particle and its antiparticle are brought together, they can annihilate into electromagnetic energy or other particles and their antiparticles in such a way that all memory of the nature of the initial particle and antiparticle is lost. Only the total energy (including the rest-mass energy, $E = mc^2$, where m is the rest mass and c is the speed of light) and total angular momentum remain. The annihilation of an electron and positron releases about 10^6 electronvolts (1 MeV) of electromagnetic energy (1.6×10^{-13} J). This can be compared with the energy of 4 eV released in a typical chemical reaction, or 200 MeV of a typical nuclear fission reaction. Annihilation of a proton and antiproton can release 1880 MeV of electromagnetic energy. *See* ELECTRONVOLT; JOULE'S LAW.

In the reverse process, antiparticles can be produced in particle collisions with matter if the colliding particles possess sufficient energy to create the required mass. For example, a photon with sufficient energy which interacts with a nucleus can produce an electron-positron pair. The creation of antibaryons or antileptons, for example, in collisions of protons with matter at high-energy accelerators, requires enough energy to create not only the antiparticle but also a corresponding particle such that total baryon or lepton number and charge can be conserved in the process. *See* ELECTRON-POSITRON PAIR PRODUCTION.

Since mesons do not possess baryon or lepton number, only charge, energy, and angular momentum need be conserved in their production. Thus, a process such as a collision of a proton with a proton can produce a single neutral pi meson, for example, $p = p \rightarrow p = p = \pi^0$. Other quantum numbers, such as strangeness and charm, must be conserved if production of mesons possessing these quantum numbers is to proceed through strong or electromagnetic interactions. In these cases a particle with the negative values of the particular quantum number must also be produced. Such a process is termed associated production. *See* CHARM; QUANTUM NUMBERS; STRANGE PARTICLES.

Matter-antimatter oscillations. Isolated neutral particles, notably K^0 and B^0 mesons, can spontaneously transform into their antiparticles via the weak interaction. These quantum-mechanical phenomena are termed K–$\bar{K}$ or B–$\bar{B}$ mixing, respectively. Mixing can lead to particle-antiparticle oscillations wherein a K^0 can become its antiparticle, a $\bar{K}^0$, and later oscillate back to a K^0. It was through this phenomenon that observation of *CP* violation first occurred. That observation, coupled to the *CPT* theorem, implies that physics is not exactly symmetric under time reversal, for example, that the probability of a K^0 becoming a $\bar{K}^0$ is not exactly the same as that in the reverse process. *See* WEAK NUCLEAR INTERACTIONS.

The possibility of neutron-antineutron ($n\bar{n}$) oscillations also exists if baryon number is not strictly conserved. While there are theories which allow, and even require, baryon number to be violated, sensitive searches for $n\bar{n}$ oscillations have found none, and place lower limits on oscillation times of $\tau_{n\bar{n}} \geq 5 \times 10^7$ s.

Antimatter in cosmology. Experimental observations, both ground- and balloon-based, indicate that the number of cosmic ray antiprotons is less than 1/10,000 that of protons. This number is consistent with the antibaryon production that would be expected from collisions of cosmic protons with the Earth's atmosphere, and is consistent with the lack of appreciable antimatter in the Milky Way Galaxy. Clouds of positrons have been observed at the center of the Galaxy. Their origin is unclear, but it is suspected that they arise from massive star formation near a large black hole, explosion of massive stars, or other such phenomena, rather than from a primordial source. *See* COSMIC RAYS.

Attempts to find antimatter beyond the Milky Way involve searches for gamma radiation resulting from matter-antimatter annihilation in the intergalactic gas that exists between galactic clusters. The null results of these searches suggests that at least the local cluster of galaxies consists mostly of matter.

If matter dominates everywhere in the universe, a question arises as to how this came to be. In the standard model of cosmology, the big bang model, the initial condition of the universe was that the baryon number was zero; that is, there was no preference of matter over antimatter. The current theory of how the matter-antimatter asymmetry evolved requires three ingredients: interactions in which baryon number is violated, time reversal (or *CP*) violation, and a lack of thermodynamic equilibrium. The last requirement was satisfied during the first few microseconds after the big bang. Time reversal violation has been observed in the laboratory in K^0 decays, albeit perhaps not of sufficient size to explain the observed baryon-antibaryon asymmetry. But the first ingredient, baryon number violation, has not yet been observed in spite of sensitive searches. Thus, the origin of the dominance of matter over antimatter remains an outstanding mystery of particle and cosmological physics. *See* BIG BANG THEORY; COSMOLOGY; THERMODYNAMIC PROCESSES.

Antiprotons in accelerators. Antiprotons are stable against decay, just as are their proton counterparts. Thus, after their production in high-energy proton-matter collisions they can be cooled, collected, and stored for an unlimited amount of time. The only condition on their continued existence is that they not come in contact with matter, where they would annihilate. Stored antiprotons have been accelerated to high energies and brought into head-on collision with similarly high-energy protons at particle accelerators at both CERN, near Geneva, Switzerland, and Fermilab, near Batavia, Illinois. Experiments with proton-antiproton collisions led to the first observation of *W* and *Z* bosons, the mediators of the weak force, and to the discovery of the top quark, the heaviest known elementary particle with a mass of about 185 proton masses. *See* INTERMEDIATE VECTOR BOSON; PARTICLE ACCELERATOR.

Antihydrogen atoms. Antihydrogen atoms, composed of an antiproton and a positron, have been created in high-energy particle physics experiments. This was accomplished by passing a stream of high-energy antiprotons through a gas jet of xenon atoms at CERN and protons at Fermilab. Occasionally a photon was created in the interaction between particles in the jet and the antiprotons. Sometimes this photon had sufficient energy to produce a subsequent electron-positron pair, and the positron became attached to the antiproton. Only a few atoms were created by this method, and they had high velocities in the laboratory.

A method of forming antihydrogen atoms with low velocities in the laboratory is to cool, decelerate, and trap antiprotons, and bring similarly cooled positrons into their proximity. Some of the positrons will be captured by the antiprotons, forming thousands of antihydrogen atoms. *See* PARTICLE TRAP.

Antiparticle nuclei. Antiparticle nuclei are difficult to form in the laboratory because their constituent antiprotons and antineutrons must be created in high-energy collisions of projectile and target particles. While several antiparticles can be created in these collisions, the probability that they will have small enough relative velocities to form themselves into nuclei without first annihilating in the rest of the debris of the collision is small. The greater the number of constituents, that is, the larger the nuclei, the smaller is the probability. Nonetheless, nuclei as large as antihelium have been observed in high-energy proton collisions with matter. Because of the difficulty of forming heavy antinuclei from particle collisions, the observation of such nuclei in cosmic rays would imply that their origin was from stars completely composed of antimatter.

Because of the short duration of balloon flights, balloon-based experiments that search for antimatter above the Earth's atmosphere lack the ability to systematically study the nature of that antimatter which is found. These experiments have never found heavier antimatter particles than antiprotons. Experiments with longer duration of observation and more sophisticated experimental instrumentation, for example, mounted on the International Space Station, can have greater sensitivity to antimatter in cosmic rays. Observation of antihelium would be indicative of primordial antimatter that survived since the big bang, and that of heavier antinuclei such as anticarbon would imply the existence of antimatter stars. *See* SPACE STATION.

Michael E. Zeller

Bibliography. Committee on Elementary Particle Physics, *Elementary Particle Physics: Revealing the Secrets of Energy and Matter*, National Research Council, National Academy Press, 1998; G. Gabrielse, Extremely cold antiprotons, *Sci. Amer.*, 267(6):78–89, December 1992; D. Perkins, *Introduction to High Energy Physics*, 3d ed., Addison-Wesley, Menlo Park, CA, 1987; *Readings from Scientific American: Particles and Fields*, Scientific American, New York, 1980; C. Schwartz, *A Tour of the Subatomic Zoo: A Guide to Particle Physics*, American Institute of Physics, New York, 1992; G. Tarlé and S. Swordy, Cosmic antimatter, *Sci. Amer.*, 278(4):36–44, April 1998; S. Weinberg, *The First Three Minutes*, 2d ed., Basic Books, New York, 1993.

Antimicrobial agents

Chemical compounds biosynthetically or synthetically produced which either destroy or usefully suppress the growth or metabolism of a variety of microscopic or submicroscopic forms of life. On the basis of their primary activity, they are more specifically called antibacterial, antifungal, antiprotozoal, antiparasitic, or antiviral agents. Antibacterials which destroy are bactericides or germicides; those which merely suppress growth are bacteriostatic agents. *See* FUNGISTAT AND FUNGICIDE.

Of the thousands of antimicrobial agents, only a small number are safe chemotherapeutic agents, effective in controlling infectious diseases in plants, animals, and humans. A much larger number are used in almost every phase of human activity: in agriculture, food preservation, and water, skin, and air disinfection. A compilation of some common uses for antimicrobials is shown in **Table 1**.

Almost 5000 years before L. Pasteur (1870) clearly enunciated the germ theory of disease, plant drugs were described by Shen Nung in *Pen Tsao* (*The Great Herbal*). The Ebers Papyrus about 1500 B.C. prescribed the use of onions and other plants for the cure of septic wounds. Through the Middle Ages the search for effective antimicrobials from plants continued, providing a major stimulus for the development of systematic botany. The herbals of this time gradually replaced superstition and incantations with precise descriptions of both the plant and the proper method of its preparation for therapeutic use. Three contemporary drugs, emetine, quinine, and chaulmoogra oil, were used as early as the seventeenth century as crude extracts of the ipecacuanha plant, cinchona bark, and *Taraktogenos kurzii*, respectively. During the next two centuries the herbal and flora evolved into the modern pharmacopoeia. A survey by E. M. Osborn (1943) of 2300 species of plants revealed that 440 species produced antimicrobial agents.

Pasteur, R. Koch, P. Ehrlich, and Lord Lister founded modern microbiology during the last three decades of the nineteenth century. Pasteur and J. Joubert (1877) discovered the phenomenon of microbial antagonism. Pyocyanase was purified from culture filtrates of *Pseudomonas aeruginosa* by

TABLE 1. Common antimicrobial agents and their uses

Use	Agents
Chemotherapeutics (animals and humans)	
Antibacterials	Sulfonamides, isoniazid, *p*-aminosalicylic acid, penicillin, streptomycin, tetracyclines, chloramphenicol, erythromycin, novobiocin, neomycin, bacitracin, polymyxin
Antiparasitics (humans)	Emetine, quinine
Antiparasitics (animal)	Hygromycin, phenothiazine, piperazine
Antifungals	Griseofulvin, nystatin
Chemotherapeutics (plants)	Captan (*N*-trichlorothio-tetrahydrophthalimide), maneb (manganese ethylene bisdithiocarbamate), thiram (tetramethylthiuram disulfide)
Skin disinfectants	Alcohols, iodine, mercurials, silver compounds, quaternary ammonium compounds, neomycin
Water disinfectants	Chlorine, sodium hypochlorite
Air disinfectants	Propylene glycol, lactic acid, glycolic acid, levulinic acid
Gaseous disinfectants	Ethylene oxide, β-propiolactone, formaldehyde
Clothing disinfectants	Neomycin
Animal-growth stimulants	Penicillin, streptomycin, bacitracin, tetracyclines, hygromycin
Food preservatives	Sodium benzoate, tetracycline

R. Emmerich and O. Low (1899) and became the first microbial antagonist to be extensively investigated as a chemotherapeutic agent. Although its action was thought to combine antimicrobial and immunizing activities, the discovery, isolation, purification, pharmacology, and clinical evaluation of pyocyanase established a pattern of investigation employed with the sulfonamides, penicillin, and all subsequently discovered antimicrobials.

Lister employed phenols and cresols as surface germicides to convert the hospitals of the late nineteenth century from pesthouses into clean, safe quarters for surgery and childbirth with a reduced hazard from cross infection. His practice is followed today not only in hospitals but in homes, food establishments, research laboratories, and wherever disinfection is essential to safe techniques. *See* ANTISEPTIC.

At the turn of the century Ehrlich, from his observations on the specific staining reactions of dyes upon certain tissue cells, conceived of highly specific dyelike molecules, “magic bullets” whose germicidal action would specifically destroy microbes without killing cells of the host. An extension of Ehrlich’s approach by G. Domagk led to the discovery in 1935 of sulfonamido-crysoidin (Prontosil), the first synthetic antimicrobial of broad clinical usefulness. J. Tréfouël and coworkers quickly established that sulfanilamide was the active moiety of the Prontosil molecule and sulfanilamide was soon modified into a series of useful derivatives, comprising the sulfa drugs, which vary in their usefulness and types of side effects. *See* CHEMOTHERAPY.

Erwin Smith (1900) established that plant diseases were as common as those of animals. Antimicrobial therapy, however, has not been equally effective. The synthetic captans and manebs are most widely used. Antibiotics, although broadly tested, have not been so useful in the control of plant diseases.

The most important antimicrobial discovery of all time, that of the chemotherapeutic value of penicillin, was made after hope of developing its clinical usefulness had been abandoned. As E. Chain (1954) recalls this dramatic period:

“Work on penicillin was begun about one year before the war (1938); it was thus not stimulated, as is so often stated, by wartime demands for new and effective methods of treating infected wounds. It is perhaps important to emphasize that the decision to reinvestigate the properties of penicillin was not motivated by the hope of developing the clinical application of an antibacterial substance which had been shown to have great practical possibilities, but for some incomprehensible reasons had been overlooked for nine years. . . . The investigation was undertaken essentially as a biochemical problem, with the aim of establishing the chemical nature of a natural, very unstable, antibacterial product active against the staphylococcus.”

The subsequent establishment of penicillin as a nontoxic drug with 200 times the activity of sulfanilamide opened the flood gates of antibiotic research. In the next 20 years, more than a score of new and useful microbially produced antimicrobials entered into daily use. *See* ANTIBIOTIC.

New synthetic antimicrobials are found today, as in Ehrlich’s day, by synthesis of a wide variety of compounds, followed by broad screening against many microorganisms. Biosynthetic antimicrobials, although first found in bacteria, fungi, and plants, are now being discovered primarily in actinomycetes.

S. Waksman (1943) was the first to recognize the

TABLE 2. Functional groups of some antimicrobial agents

Structural feature	Agents
Amino sugar	Streptomycin, neomycin, kanamycin
Polyene	Amphotericin B, nystatin, trichomycin
Polypeptide	Polymyxins, circulin, colistin, thiactin
Lactam	Penicillin
Diazo	Azaserine, DON
Coumarin	Novobiocin
Macrolide	Erythromycin, filipin, carbomycin
Phenazine	Griseolutein, echinomycin
Tetracyclic	Chlortetracycline, oxytetracycline, tetracycline, netropsin, grisein
Pyrrole	Fradicin, prodigiosin
Quinone	Xanthomycin A, cyanomycin, chartreusin
Thio-	Actithiazic acid, aureothricin, celesticetin
Alkaloid	Griseomycin
Nucleoside	Psicofuranine, puromycin, nebularine
Nitro-	Chloramphenicol, azomycin
Acetylenic	Mycomycin, nemotins
Spirane	Griseofulvins, helvolic acid
Tropolone	Puberulic acid, lactaroviolin

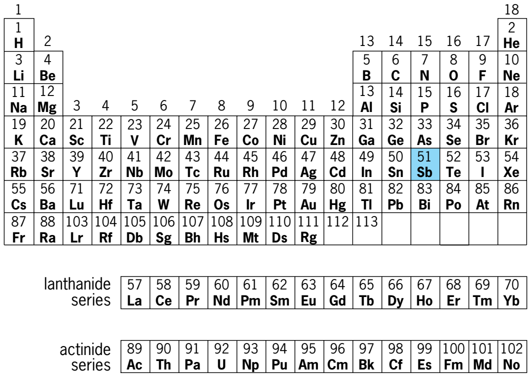
prolific antibiotic productivity of the actinomycetes. His rapid discovery of actinomycin, streptothricin, streptomycin, and neomycin inspired the intensive study of this group. Approximately three-quarters of the known antibiotics and the same fraction of all useful antibiotics are produced by actinomycetes, which live almost exclusively in the soil. Despite their habitat, a striking feature of actinomycetes is that to date it has been impossible to demonstrate the production of any antibiotics in the soil under natural growth conditions.

Antimicrobial agents contain various functional groups (**Table 2**). No particular structural type seems to favor antimicrobial activity. The search for correlation of structure with biological activity goes on, but no rules have yet appeared with which to forecast activity from contemplated structural changes. On the contrary, minor modifications may lead to unexpected loss of activity. George M. Savage

Bibliography. D. Gottlieb and P. D. Shaw (eds.), *Antibiotics*, 2 vols., 1967; S. Hammond, *Antibiotics and Antimicrobial Action*, 1978; B. M. Kagan, *Antimicrobial Theory*, 3d ed., 1980.

Antimony

The element is dimorphic, existing as a yellow, metastable form composed of Sb_4 molecules, as in antimony vapor and the structural unit in yellow antimony; and a gray, metallic form, which crystallizes with a layered rhombohedral structure. Antimony differs from normal metals in having a lower electrical conductivity as a solid than as a liquid (as does its congener, bismuth). Metallic antimony is quite brittle, bluish-white with a typical metallic luster, but a flaky appearance. Although stable in air at normal temperatures, it burns brilliantly when heated, with the formation of a white smoke of Sb_2O_3. Vaporization of the metal gives molecules of Sb_4O_6, which break down to Sb_2O_3 above the transition temperature. Some of the more significant properties of atomic and bulk antimony are given in the **table**.


1 18
1 H 2 He
2 13 14 15 16 17
3 Li 4 Be 5 B 6 C 7 N 8 O 9 F 10 Ne
11 Na 12 Mg 3 4 5 6 7 8 9 10 11 12 13 Al 14 Si 15 P 16 S 17 Cl 18 Ar
19 K 20 Ca 21 Sc 22 Ti 23 V 24 Cr 25 Mn 26 Fe 27 Co 28 Ni 29 Cu 30 Zn 31 Ga 32 Ge 33 As 34 Se 35 Br 36 Kr
37 Rb 38 Sr 39 Y 40 Zr 41 Nb 42 Mo 43 Tc 44 Ru 45 Rh 46 Pd 47 Ag 48 Cd 49 In 50 Sn 51 Sb 52 Te 53 I 54 Xe
55 Cs 56 Ba 71 Lu 72 Hf 73 Ta 74 W 75 Re 76 Os 77 Ir 78 Pt 79 Au 80 Hg 81 Tl 82 Pb 83 Bi 84 Po 85 At 86 Rn
87 Fr 88 Ra 103 Lr 104 Rf 105 Db 106 Sg 107 Bh 108 Hs 109 Mt 110 Ds 111 Rg 112 113
lanthanide series 57 La 58 Ce 59 Pr 60 Nd 61 Pm 62 Sm 63 Eu 64 Gd 65 Tb 66 Dy 67 Ho 68 Er 69 Tm 70 Yb
actinide series 89 Ac 90 Th 91 Pa 92 U 93 Np 94 Pu 95 Am 96 Cm 97 Bk 98 Cf 99 Es 100 Fm 101 Md 102 No

Antimony occurs in nature mainly as Sb_2S_3 (stibnite, antimonite); Sb_2O_3 (valentinite) occurs as a decomposition product of stibnite. Antimony is commonly found in ores of copper, silver, and lead. The metal antimonides NiSb (breithaupite), NiSbS (ullmannite), and Ag_2Sb (dicrasite) also are found naturally; there are numerous thioantimonates such as Ag_3SbS_3 (pyrargyrite).

Some atomic and bulk properties of antimony

Property	Value
Atomic weight	121.75
Electron configuration (S^4 ground state)	$[Kr]4d^{10}5s^25p^3$
Covalent radius	141 pm
Ionic radius (Sb^{3+})	90 pm
Metallic radius	159 pm
Ionization energies, 1st–6th in kJ mole^{-1}	834, 1592, 2450, 4255, 5400, 10,420
Electrode potential, Sb^{3+}/Sb	0.21 V
Electronegativity (Allred-Rochow)	1.82
Oxidation numbers	$-3, 0, +3, +5$
Specific gravity	6.691
Melting point	630.8°C (1167.4°F)
Boiling point	1753.01°C (3187.42°F)
Electrical resistivity	39.0 μohm cm (273 K)
Toxicity level	0.5 mg·m^{-3} of air

Antimony is produced either by roasting the sulfide with iron, or by roasting the sulfide and reducing the sublimate of Sb_4O_6 thus produced with carbon; high-purity antimony is produced by electrolytic refining.

Commercial-grade antimony is used in many alloys (1–20%), especially lead alloys, which are much harder and mechanically stronger than pure lead; batteries, cable sheathing, antifriction bearings, and type metal consume almost half of all the antimony produced. The valuable property of Sn-Sb-Pb alloys, that they expand on cooling from the melt, thus enabling the production of sharp castings, makes them especially useful as type metal. John L. T. Waugh

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook of Chemistry and Physics*, 85th ed., CRC Press, 2004; D. F. Shriver and P. W. Atkins, *Inorganic Chemistry*, 3d ed., 1999.

Antioxidant

A substance that, when present at a lower concentration than that of the oxidizable substrate, significantly inhibits or delays oxidative processes, while being itself oxidized. In primary antioxidants, such as polyphenols, this antioxidative activity is implemented by the donation of an electron or hydrogen atom to a radical derivative, and in secondary antioxidants by the removal of an oxidative catalyst and the consequent prevention of the initiation of oxidation.

Antioxidants have diverse applications. They are used to prevent degradation in polymers, weakening in rubber and plastics, autoxidation and gum formation in gasoline, and discoloration of synthetic and natural pigments. They are used in foods, beverages, and cosmetic products to inhibit deterioration and spoilage. Interest is increasing in the application of antioxidants to medicine relating to human diseases attributed to oxidative stress. This article aims to summarize the mechanisms by which antioxidants prevent lipid autoxidation; to present some products formed as a result of oxidation, some methods available for their detection, and the structure of some synthetic and natural antioxidants; and to emphasize the role of antioxidants in biological systems, including diseases associated with their absence.

Lipid autoxidation. The autoxidation process is shown in reactions (1)–(3). Lipids, mainly those containing unsaturated fatty acids, such as linoleic acid [RH in reaction (1)], can undergo autoxidation via a free-radical chain reaction, which is unlikely to take place with atmospheric oxygen (ground state) alone. A catalyst (L) is required, such as light, heat, heavy-metal ions (copper or iron), or specific enzymes present in the biological system [reaction (1)]. The catalyst allows a lipid radical to be formed (alkyl radical R·) on a carbon atom next to the double bond of the unsaturated fatty acid. This radical is very unstable and reacts with oxygen [reaction (2)] to form a peroxyl radical (ROO·), which in turn can react with an additional lipid molecule to form a hydroperoxide [ROOH in reaction (3)] plus a new alkyl radical, and hence to start a chain reaction. Reactions (2) and (3), the propagation steps, continue unless a decay reaction takes place (a termination step), which involves the combination of two radicals to form stable products. *See* AUTOXIDATION; CATALYSIS; CHAIN REACTION (CHEMISTRY).

$$\mathrm{RH} + \text{initiator (L)} \longrightarrow \mathrm{R\cdot} + \mathrm{LH} \qquad (1)$$

$$\mathrm{R\cdot} + \mathrm{O_2} \longrightarrow \mathrm{ROO\cdot} \qquad (2)$$

$$\mathrm{ROO\cdot} + \mathrm{RH} \longrightarrow \mathrm{ROOH} + \mathrm{R\cdot} \qquad (3)$$

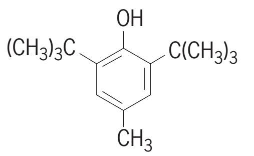
When lipid autoxidation occurs in food, it can cause deterioration, rancidity, bad odor, spoilage, reduction in nutritional value, and possibly the formation of toxic by-products. Oxidation stress in a lipid membrane in a biological system can alter its structure, affect its fluidity, and change its function. A number of assays are available for the detection of lipid autoxidation, most of them based on the detection of intermediate products, such as a lipid hydroperoxide (ROOH), or conjugated double bonds (each double bond separated by one single bond), formed during the oxidation of the unsaturated fatty acid. Other methods of detection are direct measurement of the intermediate free radicals formed (electron spin resonance methods) or the measurement of fatty acid degradation products, such as ketones and aldehydes (thiobarbituric assay). *See* FREE RADICAL.

In oxidative stress the antioxidant can eliminate potential initiators of oxidation and thus prevent reaction (1). It can also stop the process by donating an electron and reducing one of the radicals in reaction (2) or (3), thus halting the propagation steps. A primary antioxidant can be effective if it is able to donate an electron (or hydrogen atom) rapidly to a lipid radical and itself become more stable then the original radical. The ease of electron donation depends on the molecular structure of the antioxidant, which dictates the stability of the new radical; and it may be predicted by measuring the oxidation potential of the molecule, by the use of computer software, or by calculating quantum parameters of the molecule. Many naturally occurring polyphenols, such as flavonoids, anthocyanins, and saponins, which can be found in wine, fruit, grain, vegetables, and almost all herbs and spices, are effective antioxidants that operate by this mechanism. This type of antioxidant action is shown in reactions (4) and (5), with AH given as a general antioxidant and α-tocopherol (vitamin E) as an example. *See* VITAMIN E.

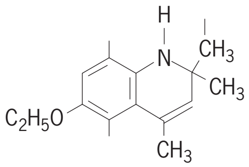
$$\mathrm{AH} + \mathrm{ROO\cdot} \longrightarrow \mathrm{A\cdot} + \mathrm{ROOH} \qquad (4)$$

$$\mathrm{ROO\cdot} + \mathrm{InH} \leftrightarrows \mathrm{ROOH} + \mathrm{In\cdot} \qquad (5)$$

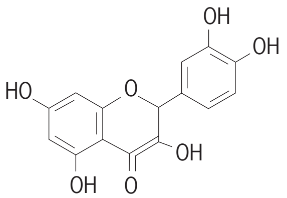
There are a number of different mechanisms by which antioxidants can prevent reaction (1) from taking place in biological and nonbiological systems. These mechanisms can be absorbing by ultraviolet light, scavenging oxygen, chelating transition metals, or inhibiting enzymes involved in the formation of reactive oxygen species, for example, NADPH oxidase and xanthine oxidase (reducing molecular oxygen to superoxide and hydrogen peroxide), dopamine-β-hydroxylase, and lipoxygenases. The common principle of action in the above examples is the removal of the component acting as the catalyst that initiates and stimulates the free-radical chain reaction. The structures of some antioxidants, synthetic and natural, are shown below. *See* ENZYME.



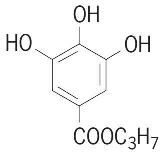
Butylated hydroxy toluene



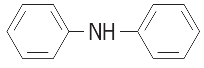
Ethoxyquin



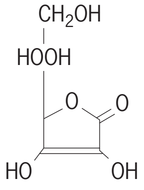
Quercetin



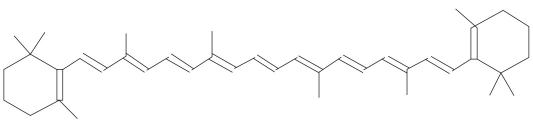
Propyl gallate



Diphenylamine



Ascorbic acid
(vitamin C)



β-Carotene

Among antioxidants, the synthetic compounds butylated hydroxyanisole (BHA), propyl gallate, ethoxyquin, and diphenylamine are commonly used as food additives. Quercetin belongs to a large natural group of antioxidants, the flavonoid family, with more than 6000 known members, many acting through both mechanisms described above. Ascorbic acid is an important water-soluble plasma antioxidant; it and the tocopherols, the main lipid soluble antioxidants, represent the antioxidants in biological systems. β-Carotene belongs to the carotenoid family, which includes lycopene, the red pigment in tomatoes; the family is known to be very effective in reacting with singlet oxygen ($^1O_2$), a highly energetic species of molecular oxygen. *See* ASCORBIC ACID; CAROTENOID; FLAVONOID; FOOD PRESERVATION.

Role in human body. Antioxidants play an important role in the human body. Each year an increasing number of diseases are attributed to oxidative stress, involving radicals and oxidants. In any biological system, the balance between the formation of reactive oxygen species and their removal is vital. New ones are formed regularly, either as a result of normal functions of the organs or of oxidative stress. Thus, superoxide ($O_2^{\bullet -}$), hydrogen peroxide (H_2O_2), and hydroxyl radical (HO) are mutagenic compounds and also by-products of normal metabolic pathways in the organs. Superoxide, which is the most important source of initiating radicals in vivo, is produced or secreted regularly from such normal cellular tissue as the mitochondria or from macrophage cells. To maintain balance, the system protects itself from the toxicity of reactive oxygen species in different ways, including the formation of antioxidants of various structures and functions. Human plasma contains a number of different enzymatic antioxidants, such as catalase, superoxide dismutase, glutathione peroxidase, reductase, and transferase, which degrade reactive oxygen species and thus protect circulating lipids, proteins, and cells from oxidative damage. Nonenzymatic protein antioxidants found in plasma, such as transferrin and albumin, are able to chelate transition metals and eliminate their catalytic effect. Other protective antioxidants are the low-molecular-weight compounds, such as vitamin C, glutathione, bilirubin, uric acid (all water-soluble), vitamin E, β-carotene, lycopene, ubiquinol-10 (all lipid-soluble), and polyphenols, found in plasma and extracellular and intracellular fluids, and in lipoproteins and membranes.

When the oxidation/reduction balance is altered toward higher oxidative stress, with more reactive oxygen species and less protecting elements, cell damage, tissue injury, and specific diseases result. Among the disorders and effects on organs associated with the proliferation of free radicals and the violation of the oxidation/reduction balance are the oxidation of lipids and proteins, with changes in their structure and function; and damage of deoxyribonucleic acid (DNA) and its contribution to the onset of cancer, cardiovascular diseases, reperfusion injury, cataracts, neurological disorders, and lung diseases. The aging process is also thought to result from the constant exposure of the organs to reactive oxygen species, with concomitant and continuous infringement of this balance, and with cumulative damage resulting from a gradual decrease in repair capacity and an increase in degenerative changes in the organs. *See* OXIDATION-REDUCTION; OXYGEN TOXICITY; SUPEROXIDE CHEMISTRY.

Jacob Vaya; Lester Packer

Bibliography. S. Baskin and H. Salem (eds.), *Oxidants, Antioxidants, and Free Radicals*, Taylor & Francis, 1997; E. Cadenas and L. Packer (eds.), *Handbook of Antioxidants (Oxidative Stress and Disease)*, 2d ed., Marcel Dekker, New York, 2001; B. Frei (ed.), *Natural Antioxidants in Human Health and Disease*, Academic Press, New York, 1994; B. Halliwell and J. M. C. Gutteridge, *Free Radicals in Biology and Medicine*, Oxford University Press, 1999; J. Pokorny, *Antioxidants in Foods: Practical Applications*, CRC Press, 2001.

Antipatharia

An order of the cnidarian subclass Hexacorallia. These animals are the black or horny corals which live in rather deep tropical and subtropical waters and usually form regular or irregularly branching

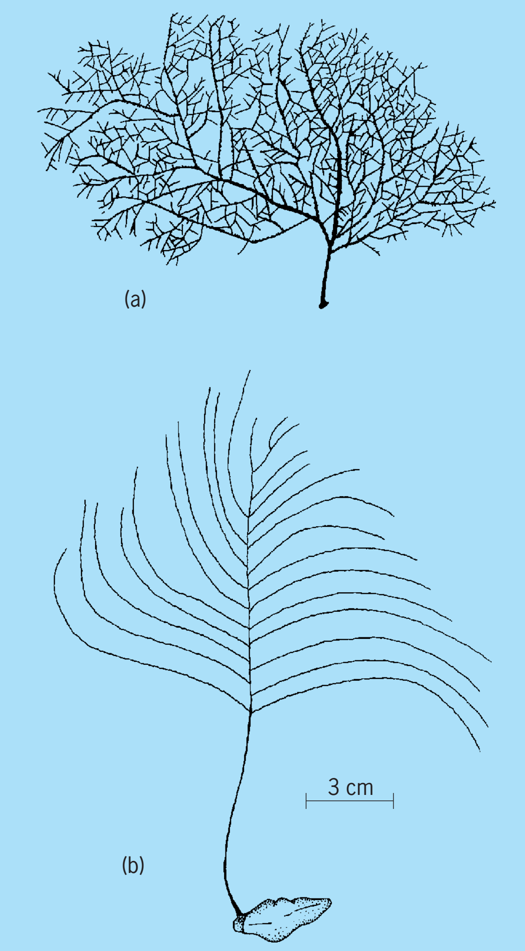


Fig. 1. Antipatharians. (*a*) *Antipathes rhipidion*. (*b*) *Bathypathes alternata*.

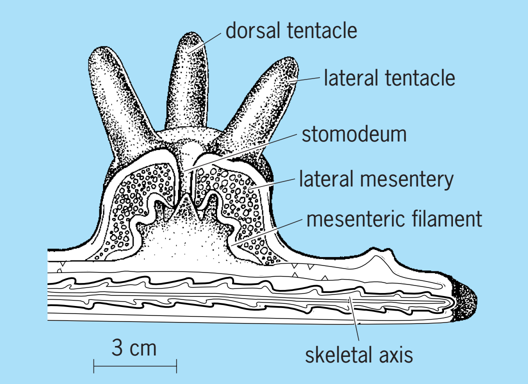


Fig. 2. Antipatharian polyp cut in vertical section.

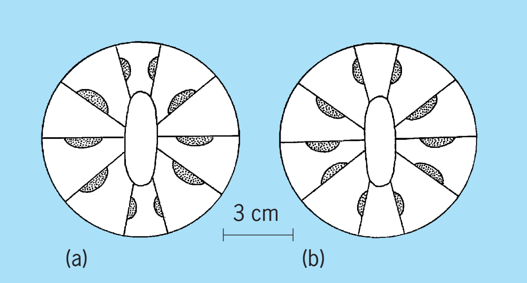


Fig. 3. Mesenteries of (*a*) *Stichopathes ceylonensis* and (*b*) *Antipathes longibrachiata*.

plantlike colonies, often 7 to 10 ft (2 to 3 m) in height, with thorny, solid lamellar, horny axial skeletons (**Fig. 1**). *Stichopathes* forms an unbranching wirelike colony.

The polyp or zooid (**Fig. 2**) has six unbranched, nonretractile tentacles with a warty surface due to the presence of nematocysts. Six primary, complete, bilaterally arranged mesenteries (**Fig. 3**) occur, of which only two lateral ones bear filaments and gonads. *Dendrobranchia*, however, has eight retractile pinnate tentacles. Both edges of the stomodeum appear to be differentiated into siphonoglyphs morphologically, but not histologically. Adjacent zooids are united by a coenenchyme, but their gastrovascular cavities have no connection. The musculature is the most weakly developed in the anthozoans.

The polyps are dioecious. Schizopathidae are dimorphic; the gastrozooid has a mouth and two tentacles, while the gonozooid, the only fertile polyp, lacks a mouth. *See* CNIDARIA; HEXACORALLIA.

Kenji Atoda

Antiresonance

The condition for which the impedance of a given electric, acoustic, or dynamic system is very high, approaching infinity. In an electric circuit consisting of a capacitor and a coil in parallel, antiresonance occurs when the alternating-current line voltage and the resultant current are in phase. Under these conditions the line current is very small because of the high impedance of the parallel circuit at antiresonance. The branch currents are almost equal in magnitude and opposite in phase.

The principle of antiresonance is used in wave traps, which are sometimes inserted in series with antennas of radio receivers to block the flow of alternating current at the frequency of an interfering station, while allowing other frequencies to pass. Wave traps are used in electric power substations to separate circuits for power-line carrier communications. *See* ANTENNA (ELECTROMAGNETISM); RESONANCE (ALTERNATING-CURRENT CIRCUITS).

John Markus

Antiseptic

A drug used to destroy or prevent the growth of infectious microorganisms on or in the human or animal body, that is, on living tissue. The legal definition is stated in the Federal Food, Drug, and Cosmetic Act, Chap. II, Sect. 201, Para. (o) as follows: "The representation of a drug, in its labeling, as an antiseptic shall be considered to be a representation that it is a germicide, except in the case of a drug purporting to be, or represented as, an antiseptic for inhibitory use as a wet dressing, ointment, dusting powder, or such other use as involves prolonged contact with the body." This means that antiseptics will render microorganisms innocuous, by either killing them or preventing their growth, according to the character of the preparation or the method of application.

Antiseptics have been in use for well over a century, first empirically in the prevention of puerperal sepsis and then more specifically in antiseptic surgery. During this period many chemical substances have been employed as antiseptics, such as certain halogens (iodine), mercurial compounds, essential oils, silver compounds, alcohols, bisphenols, and quaternary ammonium compounds. *See* HALOGEN ELEMENTS.

Iodine. This is the most important of the halogens used as an antiseptic. Although iodine solutions were used as an external application in wounds in 1839 and were admitted to the U.S. Pharmocopeia (USP) in 1840, they were not widely used in surgery until after 1860. Tincture of iodine (iodine in an alcohol solution) has been employed widely as a preoperative antiseptic and in first aid. Tincture of iodine is germicidal by laboratory test in 0.02% concentration but 2.0% solutions are usually employed in surgery and first aid. The official solutions are iodine tincture, USPXV, which consists of 2.0% iodine and 2.4% sodium iodide in alcohol; iodine tincture, strong, NFX (National Formulary X), 7% iodine and 5.0% potassium iodide in alcohol; and iodine solution, NFX, 2.0% iodine and 2.4% sodium iodide in water. The tincture is a better antiseptic than the water solution because it penetrates rapidly, is highly germicidal, and is nonspecific in antibacterial activity. Also, the tincture is not appreciably counteracted by organic matter and has a low surface tension, a low pH (which further enhances its germicidal activity) and, in usable solutions, a high degree of antimicrobial effectiveness.

Mercurial compounds. Compounds of mercury were used to prevent infection before the germ theory of disease was established, and were widely used in surgery following the work of Robert Koch on mercury bichloride in 1881. Although Koch appeared to prove that this mercurial could kill spores, the test was such that the results were actually the effect of bacteriostatic, not germicidal or sporicidal, action. In spite of this erroneous evaluation, mercury bichloride became the antiseptic of choice in place of the phenol solution used by Joseph Lister; it was employed by the professions and laypeople for many years, and still has some applications. Because of their high toxicity and severe caustic action, such inorganic mercurials as mercuric chloride, mercuric oxycyanide, and potassium mercuric iodide have been largely replaced by certain organic mercury compounds. Organic mercurial compounds are far less toxic and are nonirritating in concentrated solutions. They are highly bacteriostatic, and in concentrated solutions germicidal as well. They are also nonspecific in antimicrobial activity. The organic mercurials as well as the inorganic salts are readily counteracted by organic matter and have a high surface tension in water solutions. However, tinctures of organic mercurials (that is, solutions in alcohol) are more germicidal and, because of their lower surface tensions, are able to penetrate into tissue crevices. Tinctures of organic mercurials are effective in practice as a result of both germicidal and bacteriostatic activity. The most important organic mercurials are phenylmercuric nitrate and acetate; merbromin NF (mercurochrome), the sodium salt of dibromohydroxymercurifluorescein; thimerosal NF (merthiolate), the sodium salt of ethyl mercury thiosalicylic acid; nitromersol (metaphen), the anhydride of 4-nitro-3-hydroxymercuri-*o*-cresol; and mercresin, an alcohol-acetone solution containing a mixture of secondary amyltricresols and orthohydroxyphenylmercuric chloride.

Essential oils. Essential oils have been defined as odoriferous oily substances obtained from such natural sources as plants by steam distillation. Essential oils in alcoholic solutions also were early employed in place of the carbolic acid solution of Lister, and because of the toxic and corrosive action of mercury bichloride, they also replaced this compound. Alcoholic solution of essential oils was first developed in 1881 and was admitted as liquor antisepticus to the U.S. Pharmacopeia in 1900 and to the National Formulary IV in 1916. It was still included in the National Formulary X. While thymol is the most active ingredient in this formulation, the other essential oils included act in a synergistic manner; that is, the total effect is greater than the additive effect of the individual components. Although these essential oils are present in low concentration, the final combination is highly germicidal, in fact, equal in germicidal activity to the 2.5% solution of carbolic acid so successfully used by Lister and others in antiseptic surgery. This antiseptic is also nonspecific in its germ-killing property and kills all varieties of infectious microorganisms, with the exception of spores. Its low surface tension facilitates spreading and penetration, and its low pH enhances germicidal activity. The antiseptic is not counteracted by organic matter, is stable under all conditions of storage, and is nontoxic, noncaustic, and safe to use on all body tissues. For these reasons, the alcoholic solutions of essential oils as represented by liquor antisepticus have proved effective in a wide variety of clinical applications and in first aid.

Silver compounds. These compounds have been employed as antiseptics since 1884 and have been widely used for a variety of purposes. Because of the bland nature of most of these compounds, they have been successfully used in the eyes, nose, throat, urethral tract, and other organs. The most widely used silver compounds are silver nitrate, USP, ammoniacal silver nitrate solution, NF, silver picrate, NNR (New and Nonofficial Remedies), and certain colloidal silver preparations such as strong protein silver, NF, and mild silver protein, NF (Argyrol). These silver compounds are effective germicides of low tissue toxicity and are not counteracted by organic matter.

Alcohols. Such compounds as ethyl alcohol and isopropyl alcohol have been used as antiseptics since 1894 and are still widely employed. They are germicidal rather than bacteriostatic and are effective against the vegetative forms of bacteria and virus, but do not kill spores. Ethyl alcohol in 62.5–70% solution is most commonly used, being widely employed for disinfecting the skin before hypodermic injections and other skin punctures. It is rapidly germicidal, nonspecific in germicidal activity, nontoxic, nonirritating, and not counteracted by organic matter. Isopropyl alcohol is equal, if not superior, to ethyl alcohol and is widely used for degerming the skin and for disinfecting oral thermometers. Alcohols are also widely used in other antiseptic preparations, in which they serve to lower the surface tension and to promote spreading and penetration.

Bisphenols. Compounds such as dichlorophene and tetrachlorophene are essentially bacteriostatic agents and are weaker as germicides. The antibacterial activity of bisphenols was first demonstrated in 1906, and much new information has been developed since 1927. They have proved quite effective as skindegerming agents, when used in soaps and other detergents, and as mildew-preventing formulations. The halogenated form, such as dichlorophene, tetrachlorophene, hexachlorophene, and bithionol, is most commonly employed. When used repeatedly on the skin, as in soaps and detergents, bisphenols have a tendency to remain for long periods, thus reducing skin bacteria to a significant degree. For this purpose they are especially useful in preoperative hand washing. Bisphenols are counteracted to a variable degree by organic matter.

Quaternary ammonium compounds. These organic pentavalent nitrogen compounds have high germicidal activity. Although they are more properly classified as surface-active disinfectants, some of them are employed in certain antiseptic formulations especially suited for use on the skin and for mucous surfaces. They are counteracted to a certain extent by organic matter, especially by blood serum. Nontoxic and nonirritating, they may be used in place of alcohol after preoperative scrub-up.

Comparison of antiseptics. It is difficult, if not impossible, to evaluate the relative merits of the various classes of antiseptics under all conditions of use. However, such an effort has been made with respect to reduction of skin bacteria in the serial-basin test. The results substantiate the efficiency of iodine solutions. *See* BIOASSAY; FUNGISTAT AND FUNGICIDE.

George F. Reddish; Francis Clark

Bibliography. American Pharmaceutical Association, *National Formulary*, 12th ed., 1965; *Pharmocopeia of the United States*, 20th ed., 1980; G. F. Reddish (ed.), *Antiseptics, Disinfectants, Fungicides, and Chemical and Physical Sterilization*, 2d ed., 1957.

Antisubmarine warfare

All measures required to combat enemy submarines, including strategy, operational employment of forces, tactics, and a wide variety of equipment and weapons to find and destroy submarines and to neutralize their weapons.

Goals. The key physical facts of antisubmarine warfare (ASW) are that submerged submarines are, effectively, invisible to surface observers and that sound is the only form of energy that travels any substantial distance underwater. All forms of antisubmarine warfare are based on strategies which cause the submarines to surrender either their invisibility (or stealth) or their mobility. Generally, since the submarine can be destroyed once it is located, the effort in antisubmarine warfare goes primarily into finding it.

Whereas in the past submarines were important only as antiship weapons, some can now deliver strategic attacks. Immobilization has little effect on a strategic submarine capable of firing missiles with intercontinental range; it is an open question as to what extent antisubmarine warfare measures can be effective against such craft. Because the relative invulnerability of the strategic submarines is widely considered to be an important element of strategic stability, it is also debatable to what extent antisubmarine warfare measures directed against them are desirable. *See* BALLISTIC MISSILE.

Revised strategic emphasis. With the end of the Cold War, the emphasis has changed, at least for the major sea powers, from maintaining sea control to expeditionary operations in littoral areas, which are often relatively shallow. In the past, sea control might have been considered statistical; that is, it would have been acceptable to trade numbers of merchant ships for the elimination of an enemy submarine threat. Now the war is likely to be concentrated in an enemy's home waters, and all of the ships of the expeditionary force are quite valuable. Losses are far less acceptable. Diesel submarines, moreover, may be able to lie in ambush on the seafloor. On the other hand, likely enemies, at least minor powers, probably will not have the resources to find targets in the open ocean and cue submarines to attack them. This was much the situation the Royal Navy faced in the Falklands in 1982, when one Argentine submarine became a major and very difficult threat.

To the extent that they are intended to attack surface ships, submarines are armed with torpedoes and cruise missiles. A submarine can generally attack with homing torpedoes up to a range of about 6–12 mi (10–20 km); missiles may extend this range to 60 mi (100 km), which is about the limit of submarine-borne sensors. Attacks from greater ranges require the assistance of external sensors, and in that case the most effective antisubmarine warfare measure may be an attack on the external system. Effective antitorpedo weapons do not yet exist (although decoys can be used to negate a homing torpedo), but there are reasonably effective antimissile weapons. These considerations define the goals of tactical antisubmarine warfare: to find and destroy any submarine armed with torpedoes and approaching to within 6–12 mi (10–20 km). Ideally, any submarine armed with missiles must be dealt with beyond 60–90 mi (100–150 km), but failures to intercept torpedo-armed submarines are more likely to be devastating.

A central strategic question, then, is whether to destroy a submarine approaching its target or to destroy it in the open ocean. The first option implies a tradeoff between antisubmarine measures and weapons designed to neutralize submarine weapons. An alternative question might be whether to neutralize the submarine's offensive strategy or to destroy the submarine. If there are many potential targets, all more or less equivalent, it may be possible to destroy virtually all the submarines before their weapons can be used effectively. On the other hand, if there are a few but very important potential targets, such as carriers and major amphibious ships, then even a brief period of submarine success may be fatal to the overall war effort.

A second key question is the area over which the antisubmarine war is to be fought. During the Cold War, the area was the world ocean, and the West could take advantage of the fact that Soviet submarines had to pass through choke points on their way to patrol areas and also on the way back to rearm. Forces could be concentrated at the choke points. Had the war turned hot, anywhere a submarine was found, it could be attacked. In a modern expeditionary war, on the other hand, the combat zone will be quite restricted. The only choke point may be the entrance to a harbor, and it may be impossible to place forces there. Rules of engagement may drastically limit the circumstances under which submarines can be attacked. It may even be that more friendly than enemy submarines are in the area, because friendly submarines may be vital for roles such as reconnaissance and bombardment using cruise missiles. Thus classification of detected submarines may be much more important than it was in the past.

Sonar. Since invisibility is the key element of submarine operations, and since this is violated only by sound, sound sensor (that is, sonar) performance

-s

shapes the tactics of antisubmarine warfare. Sonars may be active, that is, the listening device may impose a signature on the target, or passive, in which case the device may listen for noise from the target. There are two limits on active sonar: water can transmit only a limited amount of energy, and active sensing will tend to alert the target submarine.

The distance that sound travels through the ocean depends on its frequency: under given conditions, the lower the frequency, the farther a sound signal will travel. Thus, very low frequency sound (typically on the order of 100 Hz or less) could be transmitted across hundreds, or even thousands, of miles, whereas very high frequency sound (hundreds of kilohertz) may travel only 1 mi (1.6 km) or less. On the other hand, for a given transducer, beam definition is inversely related to sound wavelength. Therefore, even though a very low frequency signal may be received many thousands of miles from its source, the receiver may have only a vague indication of the direction of origin. *See* UNDERWATER SOUND.

During World War II, ship-borne sonars operated at very high frequency (typically about 25 kHz), but range was limited to about 1–2 mi (2–3 km), depending on local conditions. Because wavelengths were short, even very small sonar could produce a relatively narrow beam (15° wide in typical United States sets) which gave adequate definition to control the fire of short-range weapons, such as depth charges. After World War II, sonar frequency was reduced to improve range. Ultimately, large surface ships and submarines were equipped with sonars of about 3 kHz, capable in some cases of detecting targets at a distance of as much as 30 mi (50 km); but at such ranges, target position is relatively ill-defined. Passive sonars can perform even better, but only at much lower frequencies.

The best low-frequency active sonar is the size of small boats (**Fig. 1**), which is probably about the largest size that would be acceptable within a conventional hull. Due to ship impacts, especially of lower frequency and thus ever larger sonars, there is increasing interest in towed devices. Typically the pinger and the linear-array receivers are separately towed. To hold down pinger dimensions, very wide transmitted beams are accepted, the array forming sufficiently narrow receiving beams.

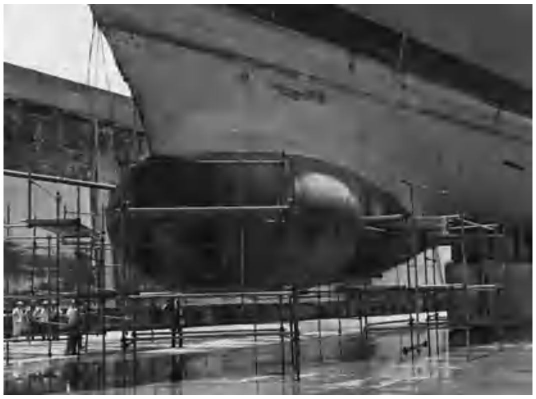


Fig. 1. SQS-26 low-frequency hull sonar, the largest U.S. hull sonar (the version onboard current surface ships, SQS-53, uses much the same sonar dome). Nearby workers give idea of scale. (*U.S. Navy*)

During the Cold War, Western navies concentrated on passive rather than active sonars for two reasons. First, using passive sonar denied a submarine information as to the whereabouts of those hunting it and, often, as to how close they were to attacking. Second, if the submarine can be detected passively, passive range often seems to be greater than active range. In addition, different types of submarines can sometimes be distinguished by the sounds they produce (sound can also indicate what the submarine is doing). The caveat is important because under some circumstances diesel submarines are extremely difficult to detect passively. Thus active sonars have become more important as attention has shifted from nuclear submarines, which must make some sound associated with their machinery, to diesel submarines, which can lie silently on the bottom with electric batteries for essential life support for a while.

Unfortunately, the virtues of low-frequency active sonars can reverse in shallow water. Sound bounces off both the bottom and the surface. The farther the sound propagates, the more it bounces and the more potential confusion. Yet long range seems essential. The main solution to date has been to use more complex waveforms which can register the small differences in path length associated with the bounces. Growing computer power makes such waveforms more usable. There have also been proposals to use low-frequency dipping sonar, whose beams would be projected more or less horizontally, grazing bottom and surface at such shallow angles that there would be very little reflection. No such sonar currently exists.

The most important current shipboard low-frequency (hence long-range) passive sonars are towed arrays. They are lines of passive transducers trailed far astern of a ship at the optimum listening depth. Although the array can detect distant targets (perhaps positioned hundreds of miles away), it cannot provide range and bearing data that are sufficient for fire control. Therefore, an airplane or helicopter is generally used to search the area defined by the sensor quickly enough to detect the submarine before it leaves the area. Passive sonar technology is also used by fixed listening systems (like the Cold War SOSUS and the current ADS) and by sonobuoys launched by aircraft (and occasionally by surface ships).

Sonar performance also depends on local water conditions and on the depth of the sonar in the water. Variable-depth sonars, which can be lowered to exploit water conditions, are towed below and behind the ship (**Fig. 2**). Submarines (**Fig. 3**) can also dive to reach favorable sonar conditions and can carry towed arrays that function as passive variable-depth sonars. The ridge running up the hull of the submarine in Fig. 3 houses such an array, which is streamed far behind the submarine, where the water is relatively undisturbed and hence better for listening, by using cable housed in the submarine's ballast tanks. *See* SONAR.

Alternatives to sonar. Sonar operates in an extremely difficult environment. For decades there have been attempts to find an alternative, particularly one which would permit very rapid search from the air or from space. The only current widely used nonacoustic sensor is the magnetic anomaly detector, which typically has a range of 1 km (0.6 mi) or less. It is generally used by aircraft as a last-minute confirmation of submarine location just before a torpedo is dropped. Magnetic detectors have also been proposed as seafloor sensors in relatively shallow water, the idea being that a submarine crossing a line of such detectors would reveal itself. A series of lines would indicate submarine course and speed.

During the Cold War the Soviet Union experimented very actively with both infrared and wake sensors. Many of their submarines had the latter either on their hulls or atop their sails. Apparently these devices sought to detect the wake of a submarine through which the Soviet submarine might pass, so that the Soviet craft could trail another submarine. Detection was based on expected differences between water in the wake and outside it. Fairly wide deployment of such sensors suggests that the Soviets had enjoyed some success, but it is not clear just how much. Since the Cold War, U.S. and British submarines have been fitted with environmental sensors which may be used either to detect submarine wakes or to distinguish water conditions in which wakes would be difficult to detect (to frustrate the Russian sensors, details of which presumably became known at the end of the Cold War). Despite claims, it seems very unlikely that an effective airborne or space-borne submarine detector has yet been perfected.

Classification of objects. A major area of effort in antisubmarine warfare has been to distinguish ever fainter (more distant) submarine noises from the surrounding random ocean noise, so that attacks are not made against nonexistent targets. Ideally it would also be possible to distinguish one kind of submarine from another (for example, to avoid accidental attacks on friendly craft). The ability to distinguish the sounds of real versus nonreal objects and friendly versus enemy submarines is generally termed classification. *See* ACOUSTIC SIGNAL PROCESSING; UNDERWATER SOUND.

Classification is often based on the sound spectrum of a detected submarine. The spectrum has two components: a series of lines (narrow-band signals) associated with rotating machinery and propeller(s), and broadband sound associated with flow over the submarine's hull and also with liquid flows inside. Much of nuclear submarine design in recent years has been associated with reducing narrow-band sounds, because the nuclear submarine cannot turn off their sources (such as turbo-generators and pumps). Moreover, a signal processor can exploit the constancy of the narrow-band lines to bring them up out of noise; they can be tracked and integrated over time. The main narrow-band signals consistently produced by a well-designed diesel submarine are associated with propeller blade rate, caused by interaction between the blades and the asymmetrical flow due to the sail and particularly the tail fins. Current sickle-shaped propellers are designed to reduce blade rate. Note that the identification of nuclear submarines has generally been associated with the narrow-band spectrum. It is difficult to identify a quiet diesel submarine. However, it is hoped that by using the details of the flow noise to visualize the shape of the submarine, identification can be accomplished. That may be possible using a different mathematical basis for analysis, perhaps wavelets rather than Fourier analysis.

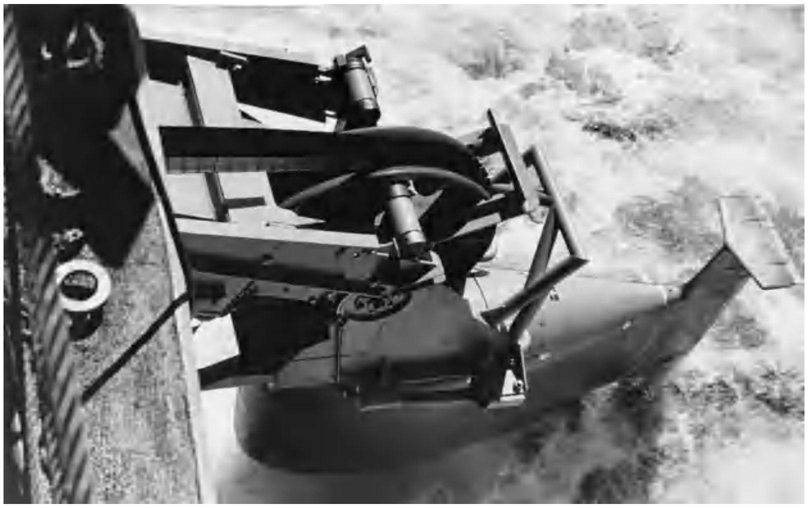


Fig. 2. Variable-depth sonar in streamlined housing (fish). (*U.S. Navy*)

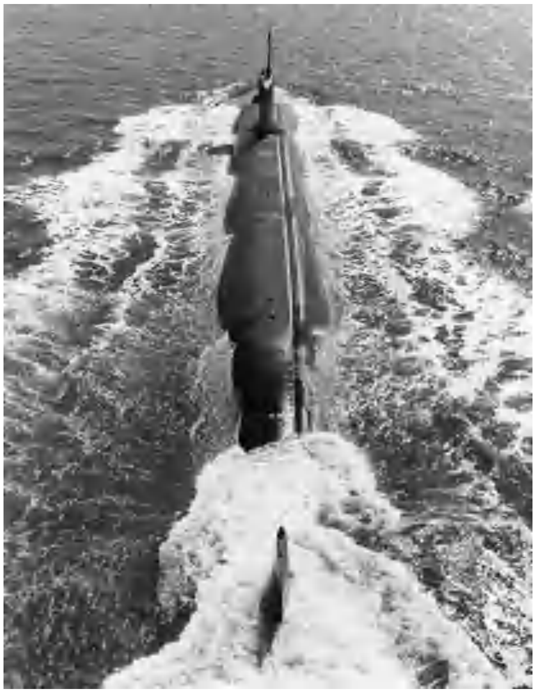


Fig. 3. Nuclear attack submarine USS *San Francisco*. The ridge running up the hull houses a towed array. (*U.S. Navy*)

Weapons. The main weapons used in antisubmarine warfare are homing torpedoes (using active or passive sonars) and depth bombs, the latter generally fused to explode at a preset depth. Most depth bombs are unguided, but in recent years the Russians have marketed a guided homing version, and some current Russian homing torpedoes behave like guided depth bombs whose propulsion is turned on only after they detect a submarine. During the

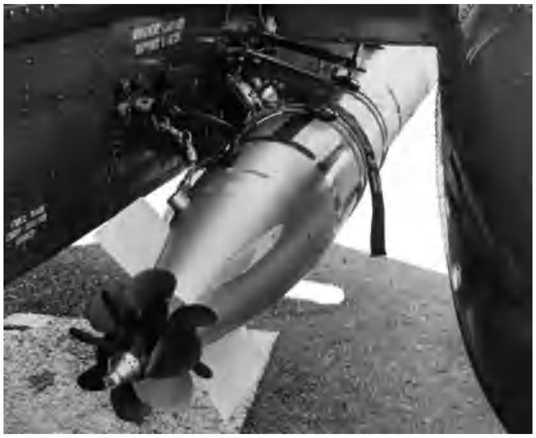


Fig. 4. Mark 46 lightweight homing torpedo aboard a helicopter. (*U.S. Navy*)

Cold War, Western and Soviet navies deployed nuclear depth bombs, but they are no longer in service. In effect the lethal radius of such a weapon could be equated with the effective search radius of a homing torpedo, and some delivery systems (such as the U.S. ASROC) had alternative nuclear depth bomb and torpedo payloads. Because sonar range may exceed the range at which a torpedo can find a submarine target, weapons are often delivered by air (several navies use ship-launched rocket, such as the U.S. ASROC to deliver torpedoes). The U.S. Mark 46 lightweight homing torpedo (**Fig. 4**) is used by many navies: as a ship-launched torpedo, as a main airplane- and helicopter-delivered attack weapon, and as the warhead of the ship-launched ASROC. The current U.S. Mk 54 lightweight torpedo combines Mk 46 propulsion with a much more sophisticated seeker and computer. A goal of ASW and torpedo research and development is increasing precision and smaller torpedoes, enabling both larger effective magazine capacity on ships and utility on developing smaller, crewless ASW vehicles. *See* ACOUSTIC TORPEDO.

Most navies have discarded conventional depth charges because a fast submarine can evade them as they sink slowly through the water. However, a few navies retain rocket-thrown depth charges because they are effective in shallow water and against submarines lying on the bottom (a homing torpedo generally needs the Doppler effect of a moving target to distinguish it from the bottom). Homing torpedoes have thus become very nearly universal antisubmarine weapons. The rule of thumb is that a torpedo needs a 50% speed margin over a submarine. Thus, a 30-knot (55-km/h) torpedo might be effective against nonnuclear submarines whose maximum speed is about 20 knots (35 km/h), but 30-knot (55-km/h) nuclear submarines require the use of 45-knot (80-km/h) torpedoes. The faster the torpedo, the more difficult it is to distinguish the echoes of a real submarine from its own flow noise. *See* NAVAL ARMAMENT.

Target characteristics. Submarines fall into two broad classes: nuclear and conventional. Nuclear submarines can operate submerged at high speed essentially indefinitely. Although nonnuclear craft may have maximum submerged speeds only slightly less than those of nuclear submarines, they can sustain those speeds for only less than an hour because they depend upon batteries with limited energy content. (Their uninterrupted underwater endurance, at very limited speed, may be as much as a week.) The batteries are recharged by air-breathing diesel engines, and recharging thus requires the submarine to draw air from the surface. A breathing pipe, the snorkel, permits the diesels to run while the submarine is operating at a shallow depth. For several years some submarines have been built with an additional air-independent power plant, which they can use to loiter for weeks at very low speed, conserving battery power for an attack and subsequent evasion. This is designed to drastically reduce snorkeling, but comes at some cost of adding complication into a craft with a very small crew.

The type of machinery used to generate power relates directly to the noise made by the submarine. The reactor of a nuclear submarine is always running, and produces noise even when it is running very slowly. A diesel-electric (conventional) submarine is noisy when snorkeling and possibly at high speed, but the electric motor is inherently fairly quiet. However, a nuclear submarine can be designed to reduce noise significantly to the point where the craft makes no more noise than a diesel. In this case the submarine would have the enormous advantage of mobility.

In water less than the maximum submarine depth (say, less than 1000 ft or 300 m deep), a diesel submarine under pursuit can hide on the bottom, almost indistinguishable from other bottom features and largely immune to attack. Nuclear submarines cannot generally do so, because their steam power plants require a constant flow of water through their condensers, and they could therefore be disabled by ingesting mud. Diesel submarines are increasingly the threat of concern due to emphasis on littoral operations (**Fig. 5**).

Operations. Most antisubmarine warfare operations fall into three distinct categories: long-range detection (which may be followed by attacks, often by aircraft), interception by submarines, and counterattack by escorts.

Long-range detection. If a submarine can be detected at very long range, it can be attacked by an airplane cued by the detection system. Until the end of the Cold War, much of U.S. antisubmarine strategy was based on the ability of fixed undersea arrays (parts of the SOSUS sound surveillance system) and strategic ship-towed arrays to detect Russian submarines at distances of hundreds or even thousands of miles in the open Atlantic and Pacific oceans. NATO navies acquired maritime patrol aircraft, such as P-3C Orions, specifically to intercept the submarines thus detected, using sonobuoys to locate them precisely enough to attack. *See* SONOBUOY.

In the aftermath of the Cold War, naval operations are likely to occur in areas not covered by the fixed systems, and possibly not covered by a limited number of slow towed-array ships. A current research trend is toward deployable submarine detection systems, both long-range (like SOSUS) and very short range (for use in large numbers in shallow water).

Interception. Because it is almost invisible, a submarine can lie near enemy submarine bases in hopes of intercepting emerging enemy ships. In the absence of effective long-range detectors, this sort of submarine blockade becomes an important tactic in possible post–Cold War confrontations.

Escort. If the submarines are likely to concentrate on ships as targets and if they cannot be located efficiently at long ranges, antisubmarine ships can be concentrated in a convoy around the potential targets. Convoys do not actively protect the escorted ships against attack. Rather, they act as a deterrent, since any submarines that attack must surrender their invisibility, inviting counterattacks from escort ships. Even with crude submarine detectors, it may be possible to track back the submarine torpedoes to find the submarines. Convoy escorts were responsible for the bulk of all German submarines destroyed at sea in World War II.

Convoy tactics have another effect. An individual submarine cannot detect and track potential targets at a very great range. Typically, submarine campaigns require some form of assistance, for example, by scouting aircraft. If targets are distributed randomly across the ocean, submarines may hunt them efficiently. If the targets are bunched in convoys, the submarines at sea require considerable assistance, and that assistance in turn requires communication with some central command. In World War II, this communication by the German U-boat force provided the Allies with the means of finding some submarines (for example, those awaiting fresh fuel and other supplies) and also of predicting the positions of submarine concentrations.

Convoy tactics were largely dropped during the 1950s and 1960s because it was felt that a single nuclear weapon could destroy an entire convoy. That threat has diminished because of nuclear deterrence and the possibility of limited and regional conflicts, but it is no longer clear that enough escorts can be provided in an emergency.

Escort is likely to be far more important in limited conflicts, not least because rules of engagement will almost certainly prohibit attacks on submarines that do not directly threaten surface ships. However, in the case of a limited conflict, as in the Falklands, the benefits to be gained by sinking or even damaging a single important ship, such as a carrier, may impel a submarine commander to chance counterattack from an escort.

It therefore becomes important to develop direct counters to a submarine's weapons, to be used to back up escort forces. Many ships tow noisemakers designed to decoy acoustic homing torpedoes, but such decoys do not address the threat of wake-following torpedoes.

Networked sensor system. Another possibility is to change the concept of underwater warfare by employing large numbers of sensors with inherently short ranges. Taken together, the sensors can create an underwater tactical picture. In theory, that is not too different from the Cold War use of very long range sensors. The difference is in the precision and response speed being aimed at. Typical Cold War practice was to use the long-range sensors to cue platforms, such as ships or maritime patrol aircraft, which in turn would use their own tactical sensors to locate a submarine well enough to become the target of a homing torpedo. The process was inherently drawn out, and it required large numbers of tactical platforms. In a littoral area, a submarine once out of port might be quite close to its targets, so time would be short. Probably there would be relatively few platforms in position to relocate and attack it. A possible future development involves large numbers of inexpensive short-range sensors, strewn over a littoral seafloor, sharing their information via underwater data links or buoyed radios. The resulting tactical picture would be quite precise, the submarine being located well enough that a very short range homing weapon would be able to attack it (the new concept is being developed in tandem with a very small homing torpedo).

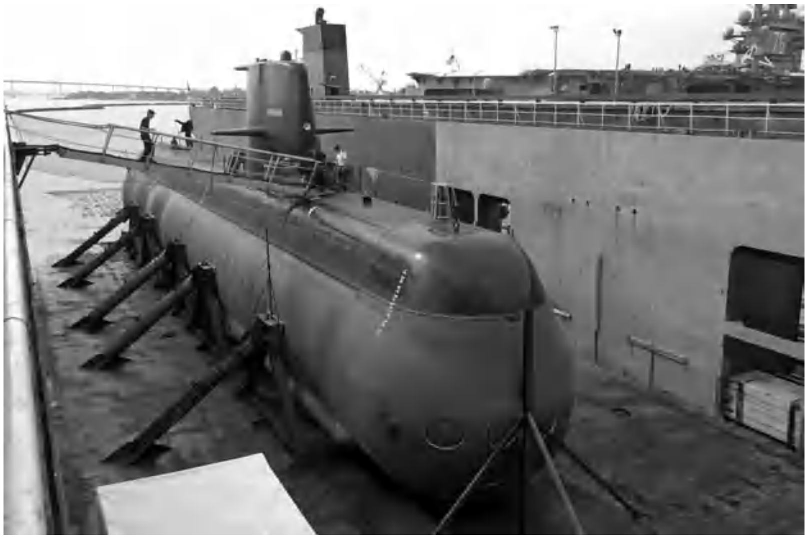


Fig. 5. Swedish diesel submarine *Gotland* in the United States on loan from Sweden to help train U.S. naval crews against the type of submarine they are likely to encounter in littoral waters. During the Cold War, the U.S. Navy concentrated mainly on the blue-water threat presented by Soviet nuclear submarines, which used very different tactics and had very different capabilities. The light-colored plating covers the submarine's passive sonar, and torpedo tubes are visible below it. The lengthwise strake along the lower hull is a low-frequency passive array. (*U.S. Navy*)

Platforms for ASW. There are four main platforms for antisubmarine warfare: surface ships equipped with sonars; airplanes equipped with sonobuoys (which they drop into the water); helicopters carrying both sonobuoys and dipping sonars; and submarines. A fifth category, unmanned underwater vehicles (UUVs), is becoming significant.

Surface ships. These are most effective for an escort or convoy strategy. Depending on their sonar range, they may fire torpedoes or rocket-propelled depth bombs directly over the side, or they may be able to employ missiles carrying torpedoes or nuclear depth bombs. Generally, surface ships equipped with long-range, low-frequency sonars, either active or passive, support aircraft for target reacquisition and attack. The aircraft may be shipboard (like U.S. LAMPS [light airborne multi-purpose system] helicopters) or they may be land-based, either fixed-wing

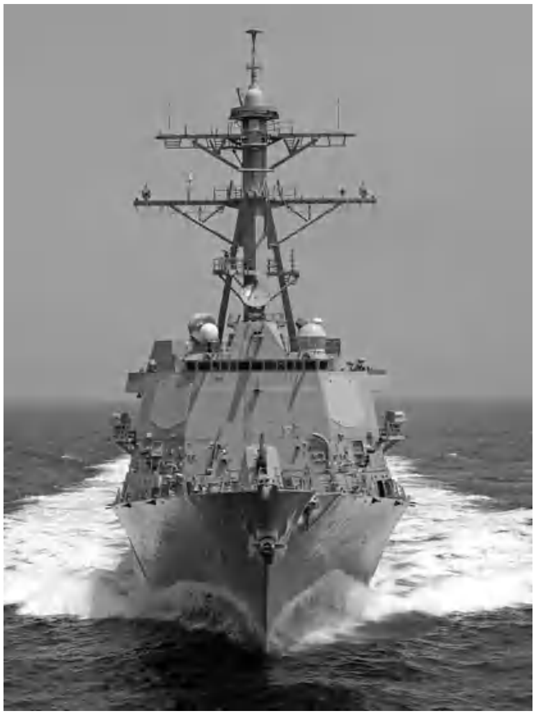


Fig. 6. Missile destroyer *Forrest Sherman*. With the demise of the Soviet Union, specialized antisubmarine ships no longer seem worthwhile. *Forrest Sherman* is a multirole destroyer designed to deal with air and underwater threats and to bombard land targets. Vertical launchers, one of which is set into the deck forward of the bridge, can fire vertically launched ASROC antisubmarine missiles as well as Standard antiaircraft missiles and Tomahawk land-attack missiles. The radome atop the mast receives the data link from the LAMPS helicopter, which is used both to deal with submarines and to attack small surface craft. The anchor is located in the bow, instead of in the traditional position on the side, to clear the big sonar dome below water. (*U.S. Navy*)

or helicopters (**Fig. 6**). *See* NAVAL SURFACE SHIP.

Aircraft. Aircraft equipped with sonobuoys are primarily a means of reacquiring submarines, not of locating them. That is because the aircraft carry only a limited number of buoys, which they must expend in order to search, and which cannot be recovered. Helicopters can lower recoverable (dipping) sonars into the water. *See* HELICOPTER.

Submarines. These are not really useful for large-scale search because their sensor range is quite limited. They can be used as escorts because they often enjoy better sonar conditions than the surface ships; however, they then have the problem of communication with surface ships. A lack of communication complicates the classification problem: antisubmarine warfare is frustrating, and surface craft or aircraft are sometimes tempted to attack any submarine they detect, without knowing whether or not it belongs to the enemy. *See* SUBMARINE.

Crewless underwater vehicles. The use of UUVs is currently of interest mainly as sensor or sensor-distribution platforms. In littoral areas, for example, they or their surface equivalents may be the best means of creating the sensor net which is intended to detect and track submarines. However, in future it is quite possible to imagine armed UUVs which might lie in wait at a harbor entrance, attacking emerging enemy submarines either under command (via an acoustic data link, which is now practicable thanks to increased computer power) or autonomously, based on rules of engagement in the vehicle's computer. A stealthy UUV might also be used to tag a submarine, even one in harbor, with a transponder which would make later detection relatively easy. Yet another possibility, given a sufficient energy source, would be for the UUV to trail the submarine, either cueing antisubmarine forces or itself attacking on command. *See* UNDERWATER VEHICLE. Norman Friedman

Bibliography. A. D. Baker III and J. L. Labayle-Couhat, *Combat Fleets of the World: Their Ships, Aircraft, and Armament*, published every 2 years (current U.S. editor is Eric Wertheim); D. R. Frieden, *Principles of Naval Weapons Systems*, 1985; N. Friedman, *Naval Institute Guide to World Naval Weapon Systems*, various editions; N. Friedman, *The Postwar Naval Revolution*, 1987; N. Friedman, *U.S. Destroyers since 1945: An Illustrated Design History*, 2d ed., 2004; N. Friedman, *Seapower as Strategy*, 2001; N. Friedman, *U.S. Submarines since 1945: An Illustrated Design History*, 1994; W. Hackmann, *Seek and Strike*, 1984; R. Harding (ed.), *The Royal Navy 1930–2000: Innovation and Defence*, 2004; J. R. Hill, *Anti-Submarine Warfare*, 2d ed., 1989.

Antitoxin

An antibody that will combine with and generally neutralize a particular toxin. When the manifestations of a disease are caused primarily by a microbial toxin, the corresponding antitoxin, if available in time, may have a pronounced prophylactic or curative effect. Apart from this, the other properties of an antitoxin are those of the antibody family (IgG, IgA, IgM) to which it belongs. *See* ANTIBODY; IMMUNOGLOBULIN.

Antitoxins have been developed for nearly all microbial toxins. Diphtheria, tetanus, botulinus, gas gangrene, and scarlatinal toxins are important examples. Antitoxins may be formed in humans as a result of the disease or the carrier state, or following vaccination with toxoids, and these may confer active immunity. The status of this can be evaluated through skin tests, for example, the Schick test for diphtheria antitoxin and the Dick test for scarlatinal antitoxin of scarlet fever, or by titration of the serum antitoxin level. Animals, especially horses, may also be immunized, and their antitoxic sera, usually refined, employed in the passive immunization of humans. *See* BOTULISM; DIPHTHERIA; GANGRENE; IMMUNITY; TETANUS; TOXIN-ANTITOXIN REACTION; VACCINATION.

Antitoxin standardization is accomplished by comparing the abilities of standard and unknown antitoxins to neutralize the toxic or serologic activities of a reagent toxin, as in the Romer skin-neutralization test in rabbits or the Ramon flocculation test. Internationally recognized standard reference antitoxins are available from various governmental laboratories, which also define procedures for their use. *See* NEUTRALIZATION REACTION (IMMUNOLOGY).

Henry P. Treffers

Bibliography. G. S. Wilson and A. A. Miles (eds.), *Topley and Wilson's Principles of Bacteriology and Immunity*, 2 vols., 6th ed., 1975.

Anura

One of the three living orders of class Amphibia, encompassing the frogs and toads. Anurans differ most obviously from salamanders (order Caudata) in lacking a tail. Usually, frogs and toads have long hindlimbs adapted for the hopping locomotion characteristic of members of the order. There can be no confusion of anurans with the limbless caecilians, members of the third order of amphibians, Apoda.

Morphology. Anurans are short-bodied animals with a large mouth and protruding eyes. The externally visible part of the ear, absent in some forms, is the round, smooth tympanum situated on the side of the head behind the eye. There are five digits on the hindfeet and four on the front. Teeth may be present on the upper jaw and the vomerine bones of the roof of the mouth, and are found on the lower jaw of only one species. Often teeth are totally lacking, as in toads of the genera *Bufo* and *Rhinophrynus*.

The short vertebral column consists of six to ten vertebrae, usually nine, and the elongate coccyx. The sacral vertebra precedes the coccyx and bears more or less enlarged lateral processes with which the pelvic girdle articulates (**Fig. 1**). A characteristic feature of anurans is the fusion of the bones in the lower arm and lower leg, so that a single bone, the radioulna in the arm and the tibiofibula in the leg, occupies the position of two in most other tetrapods (**Fig. 2**).

Distribution. Over 4200 species of anurans are known, so these animals are far more diversified than the salamanders (with less than 420 species) or caecilians (with about 165 species). Only the extreme frozen polar regions and remote oceanic islands lack native anurans; 80% of the species live in the tropics. The concentration of species in tropical regions contrasts with the distribution of salamanders, mostly found in more temperate areas.

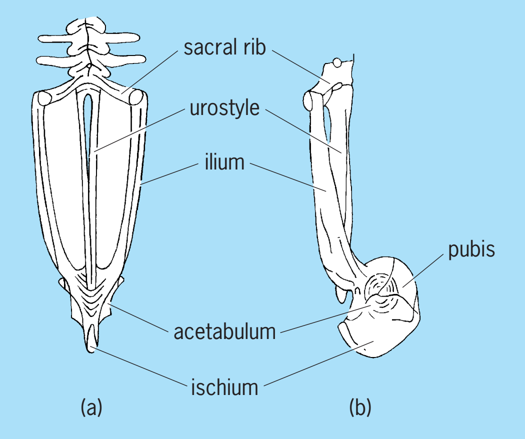


Fig. 1. Pelvic girdle of the frog in (*a*) dorsal view and (*b*) lateral view. (*After W. Montagna, Comparative Anatomy, Wiley, 1959*)

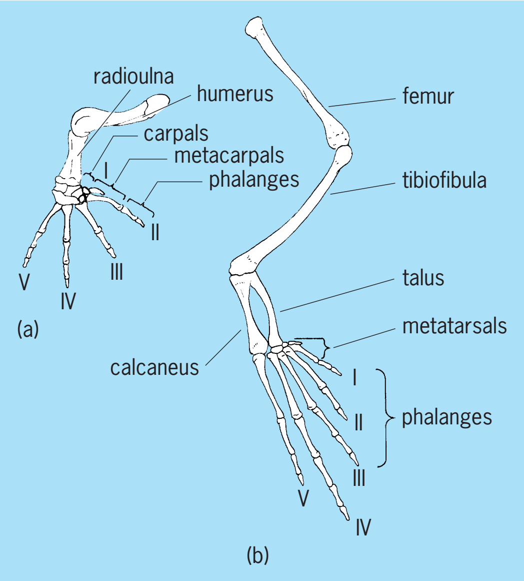


Fig. 2. Limbs of frog showing characteristic fusion of bones in (*a*) anterior limb and (*b*) posterior limb. (*After W. Montagna, Comparative Anatomy, Wiley, 1959*)

Vocalization. The one character of anurans that comes to the attention of most people, including many who may never see one, is the voice. Unlike salamanders, which are mute or nearly so, most anurans have voices and use them in a variety of ways. In the breeding season, great numbers of males may congregate in favorable sites and call, each species giving its own characteristic vocalization. Because no two species breeding at the same time and place have identical calls, it is assumed that the call is important in aiding individuals to find the proper mate (species recognition). In some species, it appears that the female is active in selecting the mate and may be responding to the mating call, but the call may not act in exactly the same way in other species. The mating call is given with the mouth closed. Air is shunted back and forth between the lungs and the mouth, so individuals can call even when submerged. Many species possess one or two vocal sacs, which are expansible pockets of skin beneath the chin or behind the jaws. The sacs (**Fig. 3**), which may be inflated to a volume as great as that of the animal itself, serve as resonators.

Other noises made by anurans include the so-called fright scream, given with the mouth open, and the warning chirp, which evidently serves as a sex recognition signal when one male contacts another. Some calls evidently serve as territorial signals. Efforts to monitor anuran populations often take advantage of the calling behavior of frogs and toads. *See* ANIMAL COMMUNICATION.

Reproduction and development. In general, male anurans grasp the female about the body with the

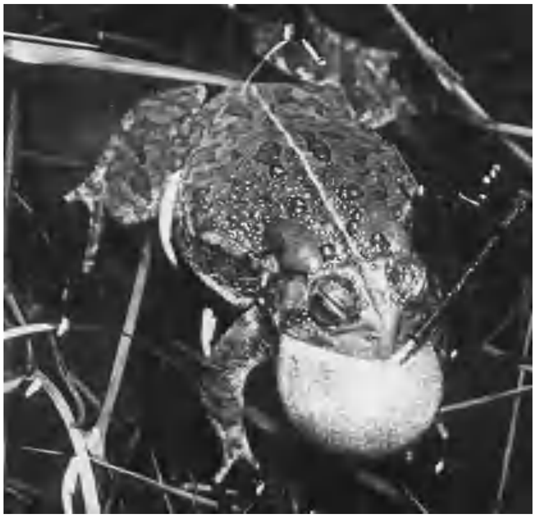


Fig. 3. Toad of the genus *Bufo* giving mating call with vocal sac expanded. (*American Museum of Natural History photograph*)

forelegs, a procedure called amplexus, and fertilizes the eggs externally as they are extruded in the water. The number of eggs may be quite large (up to 20,000 in the bullfrog or 25,000 in a common toad) or may be as few as one in some species. The eggs are each surrounded by concentric coats of jelly and may be deposited, according to the habit of the species, singly, in groups of various sizes and shapes, or in strings.

The larvae (tadpoles) are at first limbless and have external gills and a muscular tail with dorsal and ventral fins (**Fig. 4**). At hatching, there is no mouth opening present, but one soon forms that develops a horny beak and several rows of labial teeth not at all like the true teeth of the adult animal. Shortly after the tadpole hatches, the gills become enclosed within chambers and are no longer visible externally. Except for the gradual development of the hindlimbs, no additional external changes take place as the tadpole grows until the time for metamorphosis. The anterior limbs, which have been forming hidden in the gill chambers, break through the covering skin as metamorphosis begins. The tail dwindles in size as it is absorbed, while the mouth assumes its adult shape. Many other changes are taking place internally, including shortening of the intestine and its adaptation to the carnivorous diet of the adult frog.

The pattern of breeding and development outlined above (referred to as the complex life cycle) is widespread among anurans and is undoubtedly the primitive one. However, many modifications of this pattern have evolved. Many species lay eggs in moist places on land. Each egg is provided with sufficient yolk to allow the embryo to pass an abbreviated larval stage within the egg and emerge as a transformed small frog. The female marsupial frog *Gastrotheca* of South America carries the eggs in a pouch on the back, from which tadpoles or fully formed frogs emerge, according to the species. In the Surinam toad *Pipa*, also of South America, the eggs undergo development while situated in pits in the back of the mother. The male of Darwin's frog (*Rhinoderma darwini*), another South American species, has the remarkable habit of carrying the eggs and larvae in his vocal sac until metamorphosis occurs. The most highly specialized breeding habit among the Anura is seen in an African genus, *Nectophrynoides*, which is ovoviviparous; that is, the young develop within the maternal oviduct.

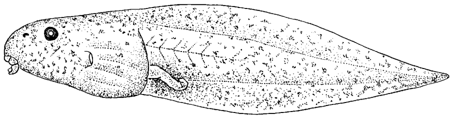


Fig. 4. Tadpole, or larval, stage of the frog *Rana pipiens*. (*After W. F. Blair et al., Vertebrates of the United States, 2d ed., McGraw-Hill, 1968*)

Nutrition. All adult anurans are carnivorous. Diet seems to depend largely upon the size of the individual, and the capacious mouth of some species permits astonishing feats of swallowing. A large bullfrog, for example, may snap up low-flying bats, ducklings, snakes, and turtles. Insects and other invertebrates form the bulk of the diet of most anurans. The tongue, moistened by a sticky secretion from the intermaxillary gland in the roof of the mouth, is used to catch smaller prey, while larger items of food may bring the front limbs into play. When swallowing, an anuran will usually depress its eyeballs into its head to aid in forcing the food down the pharynx.

In contrast to adult anurans, most tadpoles are vegetarian and feed on algae. A few are largely carnivorous or sometimes cannibalistic, and even vegetarian species will scavenge for dead animal matter. Striking feeding specializations occur, such as the funnel mouth of certain tadpoles that skim food from the surface of the water. Some are suspension feeders, like the eastern narrowmouth toad (*Gastrophryne carolinensis*), filtering organisms and organic particles out of the water column.

Ecology. The habitats of anurans are as various as the places where freshwater accumulates. Lakes and streams provide year-round habitat for many species, while others may be found in the aquatic habitat only during the breeding season. For instance, temporary aquatic habitats of the Atlantic Coastal Plain provide critical habitat for a striking diversity of anurans. The cycles of filling and drying preclude large predators from the habitat and provide ideal places for developing anurans. Any permanent source of water in the desert is likely to support a population of one or more species, and when rainstorms occur, the air around a temporary pool may be filled with mating calls for a few nights, while the animals take advantage of the water for breeding. Sometimes, the pool goes dry before the tadpoles metamorphose, and the adult frogs retreat underground to await another rain. Moist tropical regions provide an abundance of habitats little known to temperate regions, such as the air plants (bromeliads) that hold water and so provide a moist home and breeding site for anurans that may never leave the trees.

Economics. Anurans are used by humans in two important ways, as food and as laboratory animals.

Many thousands of frog legs are consumed annually, and the demand in the United States is sufficiently great that the domestic supply is supplemented by imports from overseas. Thousands more of frogs are used each year as laboratory animals, both as specimens for dissection and study in zoology classes, and as experimental animals for research on a variety of zoological and medical topics. Perhaps a more important service of anurans results from their ecological position as consumers of insects.

Classification. Although the majority of anurans fall into fairly well-defined familial categories, the arrangement of the families into subordinal groups by different authorities is not consistent, and there is controversy about the relationships of some smaller groups. The arrangement adopted here is tentative and undoubtedly will need to be changed as additional studies are made of the anatomy of anurans and are correlated with discoveries in the fossil record.

Archaeobatrachia. This suborder of the Anura is distinguished from others in having vertebrae that are concave on both anterior and posterior surfaces. Ribs are present in the adult animals, a feature confined to this group and the suborders Opisthocoela and Aglossa. Only four living species belong to the simple family Ascaphidae: three of the genus *Liopelma*, the only frogs native to New Zealand, and *Ascaphus truei* of the western United States and adjacent Canada. *Ascaphus* is called the tailed frog because of the taillike extension of the male cloaca that serves as an intromittent organ, thus permitting internal fertilization (**Fig. 5**). Small streams in forested regions are the habitat of *Ascaphus*; the flattened tadpole, with reduced tail fin and suckerlike mouth, is adapted to life in swiftly running waters.

Mesobatrachia. Some members of this suborder have trunk vertebrae that are convex anteriorly and concave posteriorly, and the adults typically have free ribs. The typical opisthocoelous family is the Discoglossidae with three genera, *Bombina*, *Alytes*, and *Discoglossus*, in Europe, North Africa, and the Orient, and a fourth, *Barbourula*, endemic to a single island in the Philippines. There are only nine species of discoglossids. Also included within the Opisthocoela at times, but of obscure relationship, is the monotypic Mexican family Rhynophrynidae.

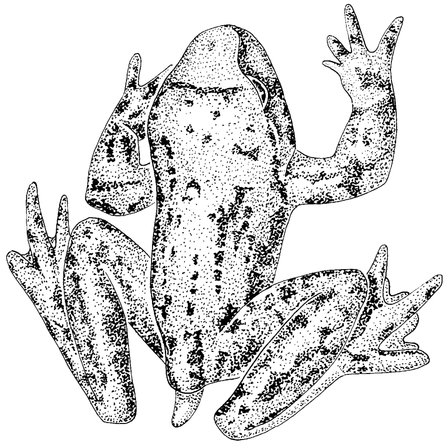


Fig. 5. Male tailed frog (*Ascaphus truei*) with a cloacal appendage, unique to the species, between the legs.

The Pipidae are also included within the Mesobatrachia. As the subordinal name suggests, a definitive tongue is lacking. Free ribs are present in the tadpole, but they fuse to the vertebrae in the adult. The vertebrae are opisthocoelous. This family of four genera and 30+ species is found in South America and Africa. A South American genus, *Pipa*, is noted for its odd breeding habits. Amid peculiar gyrations, the eggs are fertilized and deposited on the female's back, where they become implanted. Protected there, they develop in individual pouches, and eventually the miniature toads rupture the skin and emerge. An African genus is *Xenopus*, the clawed frogs, widely used in human pregnancy tests. Pelobatidae and Pelodytidae are also families assigned to this suborder. The species of the North American genus *Scaphiopus* and the European genus *Pelobates*, which is also found in northwestern Africa, are burrowing forms commonly known as spadefoot toads because of the sharp-edged digging tubercle on the hindfoot.

Neobatrachia. The suborder Neobatrachia contains most of the diversity of the order Anura and is characterized by a procoelous vertebral column, in which each vertebra is concave anteriorly and convex posteriorly, and with a free coccyx articulating by a double condyle. This suborder commonly includes 19 families, including Leptodactylidae, Bufonidae, Hylidae, Centrolenidae, Ranidae, Rhacophoridae, and Microhylidae.

Leptodactylidae. The leptodactylid frogs are most abundant in number of genera and species in the American tropics and Australia, with a very few species found as far north as the southwestern United States and New Guinea. A single genus in South Africa possibly belongs to this family. There are about 60 genera of leptodactylids with more than 600 species; about 300 species are in the genus *Eleutherodactylus* alone. Evolutionary lines within the family have undergone extensive adaptive radiation, so that one or another species lives in almost every fashion known to frogs. Species with similar habits are often similar in appearance too, so that leptodactylid frogs are often almost indistinguishable externally from a variety of species of other families.

Bufonidae. Members of the large genus *Bufo*, the true toads, are native to virtually every place in the world where anurans live, except the Australian region. The other four genera of the Bufonidae are found in Africa and the Malay region. These genera are of little numerical importance, but one of them, *Nectophrynoides* of eastern Africa, includes the only ovoviviparous anurans. All bufonids lack teeth, and the true toads of the genus *Bufo* are for the most part rather warty, short-legged, terrestrial animals that enter the water only during the breeding season.

Hylidae. The tree frogs are one of the larger amphibian families, with over 700 species known. The majority of these, about 350 species, belong to the genus *Hyla*, which is found in both Eastern and Western hemispheres. Many of the Hylidae are adapted

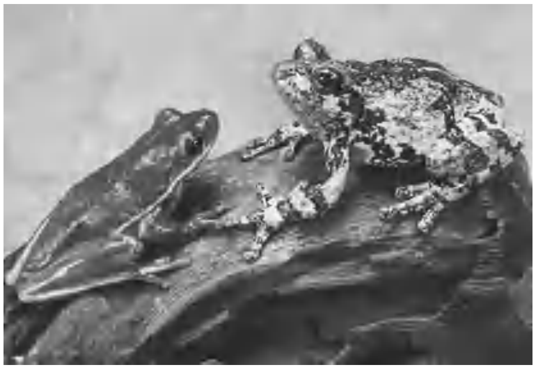


Fig. 6. North American tree frogs: green tree frog, *Hyla cinerea* (*left*) and gray tree frog, *H. versicolor*. (*American Museum of Natural History photograph*)

to arboreal life in having expanded digital disks that facilitate climbing (**Fig. 6**), but some have adopted other modes of existence and lead a terrestrial, aquatic, or even burrowing life.

The paradox frog of South America, genus *Pseudis*, and its relatives are sometimes placed within the Hylidae but may be accorded their own family, Pseudidae. The paradox frog is so called because the adult frog is smaller than the tadpole.

Centrolenidae. In the American tropics is found a small group of three genera and about 20 species of small, translucent, arboreal frogs grouped together in the family Centrolenidae. A peculiar characteristic of some species (but not restricted to this family) is that the bones are green.

Ranidae. The Ranidae number about 500 species, of which approximately 170 belong to the genus *Rana* (**Fig. 7**). This genus is fairly well represented wherever anurans live, except in South America and Australia. In each of these continents, there is only a single species of *Rana*, and the ecological position of the ranids appears to be filled by leptodactylids. *Rana* includes the abundant typical frogs, as opposed to toads, of North America and Europe. In Africa and Asia, in addition to *Rana*, there are more than 30 other genera of diverse habits and appearance that belong to the Ranidae. The poison frogs of Central and South America are small, brilliantly colored frogs whose skin secretions serve to poison the arrows of Indians (**Fig. 8**). These frogs were formerly grouped in the Procoela, but studies indicate that they may more correctly be considered as a subfamily, Dendrobatinae, of the family Ranidae.

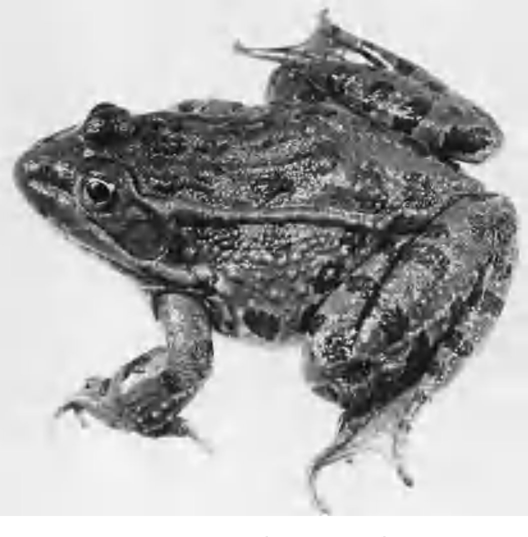


Fig. 7. Northern leopard frog (*Rana pipiens*), the most widespread species in North America, found from Hudson Bay in Canada south to Panama and from the Atlantic almost to the Pacific. (*American Museum of Natural History photograph*)

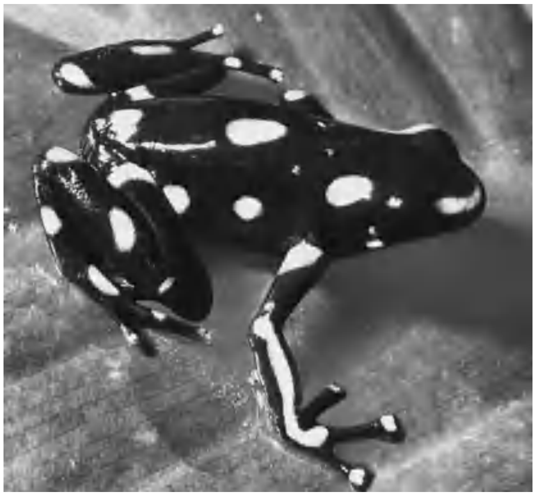


Fig. 8. Panamanian poison frog (*Dendrobates auratus*). (*American Museum of Natural History photograph*)

Rhacophoridae. Externally, hylids and rhacophorids may resemble one another very closely, but the characteristics of the vertebral column and pectoral girdle indicate their true relationships. The geographic distribution of the two families is largely mutually exclusive. Where hylids are abundant in the New World, Australia, and New Guinea, there are no rhacophorids. Over 400 species of Rhacophoridae are credited to the fauna of Africa, where only a single *Hyla* is known. About five species of *Hyla* occur in Asia among over 100 rhacophorids. The center of rhacophorid diversity is Africa, where about 20 genera are found. Only two genera, and these are doubtfully distinct, occur in Asia.

Microhylidae. The frogs of the family Microhylidae are found in three regions of the world: southern Africa and Madagascar; Sri Lanka (Ceylon), India, and southern China to New Guinea and the extreme northern part of Australia; and South America northward into the southern United States. Many of the microhylids are chunky little frogs with a pointed head and tiny mouth, and lead a terrestrial or burrowing life. In some regions, notably in New Guinea, extensive adaptive radiation has produced a variety of ecological types paralleling more typical members of other families. The specialization of microhylids is such that there are more genera recognized (58) than of ranids and rhacophorids together, though in number of species (about 230) the microhylids are inferior to either of the other groups. Investigators question the relationship of the Ranidae and Microhylidae. The two families may not be so closely related as inclusion in the same suborder would indicate; their skeletal similarities may be convergent rather than truly indicative of close phylogenetic relationship. Richard G. Zweifel; W. Ben Cash

Bibliography. W. E. Duellman and L. Trueb, *Biology of Amphibians*, 1986; M. J. Lannoo (ed.), *Amphibian Declines: The Conservation Status of United States Species*, University of California Press, Berkeley, 2005; D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, New York, 2001; R. W. McDiarmid and R. Altig (eds.), *Tadpoles: The Biology of Anuran Larvae*, University of Chicago Press, 1999; F. H. Pough et al., *Herpetology*, 3d ed., Prentice Hall, Upper Saddle River, NJ, 2004; R. D. Semlitsch (ed.), *Amphibian Conservation*, Smithsonian Institution Press, Washington, DC, 2003; R. C. Stebbins and N. W. Cohen, *A Natural History of Amphibians*, Princeton University Press, 1995; G. R. Zug, J. V. Laurie, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001.

Anxiety disorders

A group of distinct psychiatric disorders (including generalized anxiety disorder, panic disorder, obsessive-compulsive disorder, social anxiety disorder, and posttraumatic stress disorder) that are characterized by marked distress and impairment of functioning. Fear is an adaptive emotion that is seen in many animal species, while anxiety can be viewed as the subjective component of fear and, therefore, a normal emotion selected for by evolutionary processes. Anxiety disorders are among the most prevalent, disabling, and costly of psychiatric disorders. They are frequently underrecognized and undertreated. Fortunately, there is growing understanding of the mechanisms underlying these disorders, and a number of effective treatments are available. *See* NEUROTIC DISORDERS.

Generalized anxiety disorder. Generalized anxiety disorder (GAD) is characterized by excessive worry, tension, and anxiety. Accompanying physical symptoms include muscle tension, restlessness, fatigability, and sleep disturbances. Generalized anxiety disorder occurs in around 4–6% of the population and is the anxiety disorder most frequently encountered by primary care physicians, to whom sufferers may go to seek help for the physical symptoms of the disorder. This disorder is more common in women, and frequently begins in the teenage years with increasing prevalence into the fourth decade. Research on the impairment of functioning found in people with generalized anxiety disorder relative to sufferers with other medical disorders indicates that generalized anxiety disorder is one of the most disabling of all medical disorders.

Etiology and pathophysiology. Animal studies of fear and clinical studies of people with generalized anxiety disorder suggest that similar brain circuits are involved in both cases. For example, the amygdala (a small structure located in the front [anterior] part of the temporal lobes) plays an important role in fear and anxiety. Numerous complex connections to a range of other brain areas allow it to coordinate cognitive, emotional, and physiological responses to fear and anxiety. Thus in the "fight or flight" response, the organism makes cognitive-affective decisions about how to respond to the perceived danger and has a range of somatic (for example, increased heart rate and respiration rate) and endocrine (for example, release of "stress hormones") responses that act together to increase the likelihood of avoiding the danger.

Within the fear and anxiety circuits in the brain, various neurochemical systems [for example, amino acid neurotransmitter systems, such as glutamate and gamma-aminobutyric acid (GABA), and monoaminergic neurotransmitter systems, such as serotonin and noradrenaline] are responsible for mediating the communication between the functionally connected regions. Medications acting on these systems, such as benzodiazepines, selective serotonin reuptake inhibitors (SSRIs), and noradrenergic/serotonergic reuptake inhibitors (NSRIs), are effective in treating generalized anxiety disorder. *See* NEUROBIOLOGY; NORADRENERGIC SYSTEM; SEROTONIN.

Treatment. Both specific medications and particular kinds of psychotherapy have proven effective in the treatment of generalized anxiety disorder. Although benzodiazepines have often been used to treat generalized anxiety disorder, selective serotonin reuptake inhibitors (SSRIs) and noradrenaline serotonin reuptake inhibitors (NSRIs) are currently viewed as first-line options because of their favorable safety profile. Cognitive-behavioral psychotherapy focuses on using behavioral techniques and changing underlying thought patterns; it is the best studied of the psychotherapies in generalized anxiety disorder, with growing evidence for its efficacy. *See* PSYCHOPHARMACOLOGY; PSYCHOTHERAPY.

Panic disorder. Panic disorder (PD) is characterized by repeated, sudden panic attacks, which are accompanied by a range of physical symptoms, including respiratory (shortness of breath), cardiovascular (fast heart rate), gastrointestinal (nausea), and occulovestibular (dizziness) symptoms. Agoraphobia frequently coexists with panic disorder and involves anxiety about and avoidance of situations that provoke a panic attack. The prevalence of panic disorder is approximately 2% in the general population; it is more common in women and is often complicated by depression. Panic attacks mimic a number of important medical conditions that need to be carefully excluded at initial assessment. *See* PHOBIA.

Etiology and pathophysiology. The same brain circuits and neurotransmitters implicated in fear and generalized anxiety disorder are also likely to play a role in panic disorder. "Challenge" studies that activate particular neurochemical systems have been used to demonstrate the involvement of GABA, serotonin, and noradrenaline in panic disorder. This is complemented by evidence of treatment efficacy of medications that modulate these systems. Carbon dioxide provokes panic attacks, an observation that has led to the hypothesis that panic disorder involves a false suffocation alarm. As predicted by this theory, panic attacks are more common when pulmonary carbon dioxide is increased (for example, in lung disease) and is less common when pulmonary carbon dioxide is decreased (for example, after pregnancy).

As in generalized anxiety disorder, the amygdala appears to play an important role in panic. Damage to this area may be associated with loss of fear responses. Conversely, in conditions with increased activation of the amygdala, such as temporal lobe epilepsy, there may be an increased predisposition to panic. Brain imaging has also suggested a role for the parahippocampal region, which appears to be involved in remembering contextual aspects of fear (that is, where the panic attack occurred). *See* BRAIN.

Treatment. Treatment of panic disorder, as other anxiety disorders, involves the use of specific medications and psychotherapies. The first-line choice of medication is typically an SSRI or NSRI. Benzodiazepines are effective alone or in combination with SSRIs, but their use as the only medication is generally avoided due to the potential for dependence and withdrawal. Cognitive-behavioral therapy that addresses avoidance behavior and irrational dysfunctional beliefs is also effective.

Obsessive-compulsive disorder. Obsessive-compulsive disorder (OCD) is characterized by obsessions (unwanted, persistent, distressing thoughts) and compulsions (repetitive acts to relieve anxiety caused by obsessions). The disorder occurs in 2–3% of the population and often begins in childhood or adolescence. Interestingly, obsessive-compulsive disorder is also seen in the context of certain infections, brain injury, and pregnancy or the puerperium (6-week period after delivery). Unfortunately, since people are often embarrassed by their symptoms and unaware of available treatments, it is common for diagnosis and appropriate treatment to be delayed for many years. *See* OBSESSIVE-COMPULSIVE DISORDER.

Etiology and pathophysiology. A range of evidence now implicates a brain circuit (loop) between the frontal cortex, basal ganglia, and thalamus in mediating obsessive-compulsive disorder. Lesions to the basal ganglia (for example, in certain neurological disorders) may lead to obsessions and compulsions. Brain-imaging studies have demonstrated increased brain activity in the basal ganglia in obsessive-compulsive disorder, and show that both medication and specific psychotherapies are able to normalize this activity. In certain very treatment-resistant cases of obsessive-compulsive disorder, neurosurgical procedures that effect minor interruptions in communication between the implicated cortico-basal brain regions, can bring some relief of symptoms.

Key neurotransmitters in this circuit include the dopamine and serotonin neurotransmitter system. Genetic studies have suggested a role for aspects of the dopamine system in obsessive-compulsive disorder. Medications that increase dopamine in the basal ganglia increase repetitive movements in animal models of obsessive-compulsive disorder, and can increase tics in people with obsessive-compulsive disorder. Furthermore, medications that lower dopamine transmission appear to be helpful as part of the treatment of obsessive-compulsive disorder. A range of data also points to the importance of the serotonin system. This includes the selective response of obsessive-compulsive disorder to serotonergic and not to noradrenergic antidepressants.

Treatment. Countering misperceptions in the community is crucial to treating many of the anxiety disorders. A wider understanding of obsessive-compulsive disorder as a treatable medical disorder would encourage people to seek treatment sooner. SSRIs are current first-line treatments for obsessive-compulsive disorder, with dopamine blockers added in those who do not respond to these agents. Behavioral therapy focuses on exposure and response prevention, while cognitive strategies address the distortions in beliefs that underlie the perpetuation of symptoms.

Social anxiety disorder. Social anxiety disorder (SAD) is characterized by persistent fears of embarrassment, scrutiny, or humiliation. People with social anxiety disorder may avoid social situations (such as talking in small groups and dating) and performance situations (eating or speaking in front of others), resulting in marked disability. For some, symptoms are confined to one or more performance situations, while for others symptoms may be generalized to include most social and performance situations. Generalized social anxiety disorder is usually more severe, and sufferers are more likely to have a family history of social anxiety disorder.

Social anxiety disorder is particularly common, with prevalence figures in some studies upwards of 10%. While women are more likely to have social anxiety disorder, men with the disorder may be more likely to come for treatment. Onset is typically in the teenage years, with symptoms continuing for many years. Social anxiety disorder is often complicated by depression, and people with it may self-medicate their symptoms with alcohol, leading to alcohol dependence. *See* AFFECTIVE DISORDERS; ALCOHOLISM.

Etiology and pathophysiology. There is a growing search for the structures and chemicals that are involved in thoughts and feelings related to social interaction and status. Animal studies have, for example, found that serotonin and dopamine receptors differ in dominant and subordinate primates. People with social anxiety disorder who are exposed to facial expressions that invoke fear of scrutiny will selectively increase brain activity in the amygdala (previously discussed in relation to fear in panic and generalized anxiety disorder). In addition, brain-imaging studies have found that effective treatment with medication and psychotherapy normalizes activity in the amygdala and the closely related hippocampal region in social anxiety disorder.

Treatment. Many people with symptoms of social anxiety disorder view these as immutable aspects of their personality. Education of the community through advocacy groups can be helpful in changing this perception. When social anxiety disorder is identified in the clinic, it is important to begin treatment by addressing such misperceptions. SSRIs, NSRIs, and cognitive-behavioral therapy are all effective in the treatment of social anxiety disorder. Monoamine oxidase inhibitors and benzodiazepines are also known to be effective treatments for the disorder, but they have a number of disadvantages. *See* MONOAMINE OXIDASE.

Posttraumatic stress disorder. By definition, posttraumatic stress disorder (PTSD) is an abnormal response to severe trauma. It is characterized by distinct clusters of symptoms: reexperiencing of the event (for example, in flashbacks, dreams, or distress in response to reminders of the trauma), avoidance (of reminders of the trauma), numbing of responsiveness to the environment, and increased arousal (for example, insomnia, irritability, and being easily startled). Although exposure to severe trauma occurs in more than 70% of the population, posttraumatic stress disorder has a lifetime prevalence of 7–9% in the general population. Risk factors for developing posttraumatic stress disorder following exposure to severe trauma include female gender, previous history of psychiatric illness, trauma severity, and absence of social support after the trauma. *See* POSTTRAUMATIC STRESS DISORDER.

Etiology and pathophysiology. Brain-imaging studies have suggested that in posttraumatic stress disorder frontal areas (for example, the anterior cingulate cortex) may fail to effectively dampen the "danger alarm" of the amygdala (known to mediate cognitive, affective, and physiological responses to intense fear). Whereas stress responses ordinarily recover after exposure to trauma, in posttraumatic stress disorder they persist. Furthermore, brain-imaging studies have suggested that hippocampal volume may be reduced in posttraumatic stress disorder, perhaps accounting for the memory deficits that are characteristic of this disorder.

One possibility is that stress hormones are responsible for the decreased hippocampal volume in posttraumatic stress disorder. Indeed, there is growing evidence that functioning of the hypothalamic-pituitary-adrenal hormonal axis in posttraumatic stress disorder is disrupted. Other systems, such as serotonin and noradrenaline, may however also be involved. *See* STRESS (PSYCHOLOGY).

Treatment. Given that posttraumatic stress disorder may occur in the context of interpersonal trauma, establishing a relationship with a trusted therapist is important. As with many of the anxiety disorders, both SSRIs and cognitive-behavioral therapy are effective in decreasing posttraumatic stress disorder symptoms. Behavioral techniques (using different forms of exposure in the safety of the consultation room) or cognitive retraining (addressing irrational thoughts on the trauma and its consequences) can both be helpful. Paul D. Carey; Dan J. Stein

Bibliography. E. Hollander and D. Simeon, *Concise Guide to Anxiety Disorders*, American Psychiatric Publishing, Washington, 2003; Social anxiety disorder; *Acta Psychiat. Scand.*, vol. 108, suppl. 417, 2003; D. J. Stein, *Clinical Manual of Anxiety Disorders*, American Psychiatric Publishing, Arlington, 2004; D. J. Stein and E. Hollander (eds.), *Textbook of Anxiety Disorders*, American Psychiatric Publishing, Washington, DC, 2002; Update on posttraumatic stress disorder, *J. Clin. Psychiat.*, vol. 65, suppl. 1, 2004.

Anyons

Particles obeying unconventional forms of quantum statistics. For many years, only two forms of quantum statistics were believed possible, Bose-Einstein and Fermi-Dirac, but it is now realized that a continuum of possibilities exists.

In quantum mechanics, the probability that a given state will occur is calculated by adding the amplitudes for all processes leading to that state, and then squaring this total amplitude. Thus, in the case of two indistinguishable particles A and B, the amplitude for the process that leads to A arriving at point x while B arrives at point y must be added to the amplitude for the process that leads to A arriving at y while B arrives at x—the so-called exchange process—because the final states cannot be distinguished. Actually, this simple recipe is appropriate only for particles obeying Bose-Einstein statistics (bosons). The recipe for particles obeying Fermi-Dirac statistics (fermions) is to subtract the amplitude for the exchange process. *See* BOSE-EINSTEIN STATISTICS; FERMI-DIRAC STATISTICS.

The primary definition of anyons posits other possible recipes for adding exchange processes. Anyons of type θ are defined by the rule that the amplitudes for processes where A winds around B n times in a counterclockwise direction are added with a factor $e^{2in\theta}$. In this definition, half-integer values of n are allowed; half a winding is an exchange. The values $\theta = 0$ and $\theta = \pi$ correspond to the previous definitions of bosons and fermions, respectively.

The more general possibilities for quantum statistics, corresponding to other possible values of the angle θ, are defined in a mathematically consistent way only for particles whose motion is restricted to two space dimensions. However, many of the most interesting materials in modern condensed-matter physics are effectively two-dimensional, including microelectronic circuits and the copper-oxide layers fundamental to high-temperature superconductivity. The question as to whether the elementary excitations in these systems, the quasiparticles, are anyons can be settled only by a deep investigation of specific cases. The quasiparticles in fractional quantized Hall states are known to be anyons. *See* HALL EFFECT; INTEGRATED CIRCUITS; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; QUANTUM STATISTICS; SUPERCONDUCTIVITY. Frank Wilczek

Bibliography. F. Wilczek, Anyons, *Sci. Amer.*, 264(5):58–64, May 1991; F. Wilczek (ed.), *Fractional Statistics and Anyon Superconductivity*, 1990.

Aorta

The main vessel of the systemic arterial circulation arising from the left ventricle of the heart; it is divided into three parts for convenience only. The first portion, the ascending aorta, passes upward under the pulmonary artery; the coronary arteries arise at the base of the ascending aorta behind the aortic valves. The second part, or aortic arch, curves over the hilum of the left lung, giving off the innominate, left carotid, and left subclavian arteries, which supply the neck, head, and forelimbs. The third portion, or descending aorta, continues downward in the thorax on the left side of the vertebral column to the diaphragm, giving off small arteries to the bronchi, esophagus, and other adjacent tissues. Below the diaphragm this vessel, known as the abdominal aorta, descends to the level of the fourth lumbar vertabra where it bifurcates into the two common iliac arteries supplying the hindlimbs.

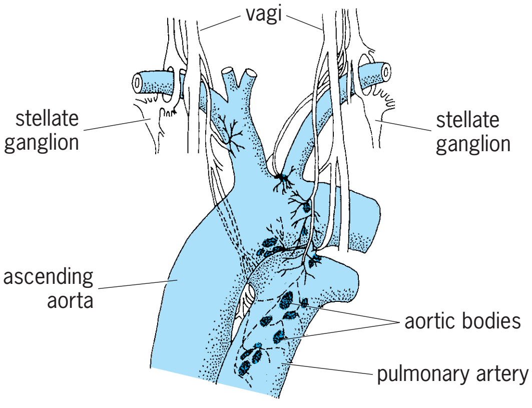
In the abdomen the major branches of the aorta include the single celiac, superior mesenteric and inferior mesenteric, and the paired renal and internal spermatic (male) or ovarian (female) arteries. In addition, many small branches go to other organs and to the body wall.

The aorta is a large elastic artery. Its endothelial lining is continuous with that of the heart and its arterial branches. Around this lining is a thin tunica intima, consisting of longitudinal elastic fibers, followed by a heavy layer of circularly arranged tunica media, which has a predominant proportion of elastic fibers arranged in dense layers between muscle tissue and collagenous fibers. The outer tunica adventitia consists of loose connective tissue which contains fat, nerves, and nutrient blood vessels called vasa vasorum. This arrangement of layers, especially the elastic fibers, enables the aorta to expand and recoil in response to each pulsation of blood through it.

Walter Bock

Aortic body

One of the cellular aggregations traditionally known as chemoreceptors which are generally found adjacent to the aortic arch and pulmonary artery and which often are adjacent to the orgin of the subclavian arteries (see **illus.**). The term is best restricted to those groups of cells that have topographic, morphological, and functional similarities. However, many aggregations of epitheloid cells have been described which lie adjacent to the aorta in the thorax, in the root of the neck, and in the abdomen, and any of these could be included, with some topographic descriptive justification, in the group called aortic bodies.



Oval-shaped aortic bodies are found lying on the anterior and posterior surfaces of the aortic arch and pulmonary artery.

The comparative physiology and anatomy of aortic bodies are not well documented because these organs are microscopic and, therefore, not readily accessible and identifiable. Their functions are not well studied, even in mammals. However, studies have shown that the morphological substrate of these bodies appears to be similar in clusters of two types of epitheloid cells, one of which is innervated, surrounded by a number of large capillaries.

There are two types of parenchymal cells in aortic bodies. Chief or glomus (type I) cells are usually polyhedral in shape, have fingerlike processes, and contain membrane-bound electron-dense granules that contain neuropetides and biogenic amines, usually dopamine and norepinephrine. These parenchymal cells are connected by gap junctions. Sustentacular (type II) cells are supporting elements which closely surround several chief cells. These cells are not innervated and do not contain vesicles, which are characteristic of chief cells. Sustentacular cells resemble the supporting Schwann cells associated with chromaffin tissues and have many ultrastructural features in common with such tissues. In addition to functional differences, the principal dissimilarities between the two types of cells are the size of the cytoplasmic granules and the nature of the contained amines, the nature of the innervation of the chief cells, and the distinct structure of capillary blood vessels that form a characteristic feature of the chemoreceptors.

The presence of catecholamines has been demonstrated by high-performance liquid chromatography. Acetylcholine and neuropeptides, which are known to be neurotransmitters in other tissues, have also been found in chief cells. Receptors for these neurochemicals have been localized in aortic body cells and nerve endings. However, the neurotransmitter which is specific for arterial O_2 and CO_2 chemoreception in the aortic bodies has not been identified. It is likely that several neurotransmitters are released in response to a single stimulus and the sensory responses are dependent on the simultaneous effects of these transmitters.

Aortic bodies are innervated by afferent nerve fibers derived from the vagus nerve and efferent nerve fibers of the sympathetic nervous system. Both types of fibers terminate in synaptic-type expansions on chief cells. The larger (afferent) endings of the vagal fibers contain a variable number of clear synaptic vesicles and mitochondria; the endings of sympathetic nerve fibers contain synaptic-type dense vesicles that are typical of postganglionic noradrenergic nerve endings. However, all chief cells are not innervated. There are no loose nerve endings between the cells, and the sustentacular cells are not innervated. The distribution and innervation of aortic bodies by the left and right aortic nerves are not symmetrical.

The aortic bodies in adults receive arterial blood supplies from small and slender blood vessels that originate from either the coronary artery or the aortic arch. In the fetal stage, oxygen-poor blood of the pulmonary arterial system supplies these chemoreceptor tissues as well. The extensive capillary network is surrounded by a continuous fenestrated endothelium.

The aortic bodies resemble carotid bodies in that they respond to systemic hypoxia eliciting chemosensory discharge. They are poorly responsive to carbon dioxide partial pressure or to hydrogen ion concentration, although these stimuli augment the response to hypoxia. The impulses that travel up the aortic nerves increase in hypoxia; under conditions of high oxygen tensions the reverse happens. The mechanism by which this effect is achieved is still unknown, but it appears that chief cells are the first chemoreceptive sensors. These cells influence the central nervous system by generating impulses through the release of neurotransmitters. The chemosensory fibers of the aortic body show a profound excitatory response to hypotension, and this response is synergistically augmented by hypoxia. Conditions such as anemia and carboxyhemoglobinemia are potent stimuli for the aortic chemoreceptors. However, the response is ultimately due to a fall in the partial pressure of oxygen in the cells. Accordingly, aortic bodies appear to have less of an oxygen supply than the carotid bodies and hence are stimulated by any factor that tends to compromise oxygen delivery to them. These chemoreceptors are also transiently stimulated by the metabolic inhibitors of oxidative phosphorylation. *See* CAROTID BODY.

As a part of the reflex arc, the aortic bodies subserve the control function of both the respiratory and the cardiovascular systems. The respiratory function of the aortic body appears to be relatively weak compared to that of carotid bodies. The effects on the cardiovascular reflex are expected to be significant owing to the fact that blood pressure effects on the chemosensory discharge are vigorous. These effects would have to be mediated by the autonomic nervous system. However, the physiological function of the aortic bodies is still under investigation.

Tumors may arise that involve either chief cells or chromaffin cells (rather than sustentacular cells). The tumor may contain high concentrations of catecholamines, usually norepinephrine. *See* CHEMORECEPTION.

Sukhamay Lahiri

Bibliography. H. Acker and R. G. O'Regan (eds.), *Physiology of the Peripheral Arterial Chemoreceptors,* 1983; N. S. Cherniack and J. G. Widdicombe, *Handbook of Physiology: The Respiratory System,* vol. 2, pp. 247–362, 1986.

Apatite

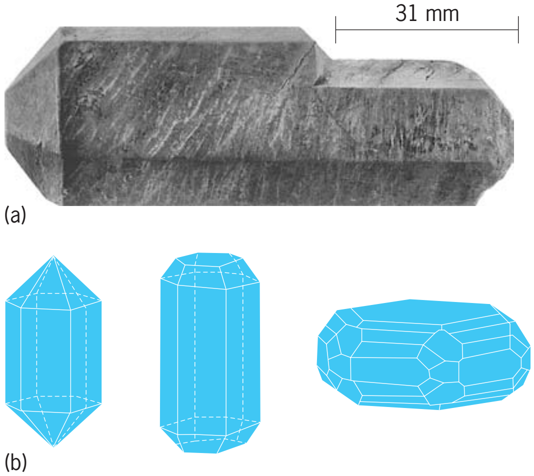
The most abundant and widespread of the phosphate minerals, crystallizing in the hexagonal system, space group $P6_3/m$. The apatite structure type includes no less than 10 mineral species and has the general formula $X_5(YO_4)_3Z$, where X is usually Ca^{2+} or Pb^{2+} or As^{5+}, and Z is F^-, Cl^-, or $(OH)^-$. The apatite series takes $X = Ca$, whereas the pyromorphite series includes those members with $X = Pb$. Three end members form a complete solid-solution series involving the halide and hydroxyl anions and these are fluorapatite, $Ca_5(PO_4)_3F$; chlorapatite, $Ca_5(PO_4)_3Cl$; and hydroxyapatite, $Ca_5(PO_4)_3(OH)$. Thus, the general series can be written $Ca_5(PO_4)_3(F,Cl,OH)$, the fluoride member being the most frequent and often simply called apatite.

The apatite series is further complicated with other substitutions. Carbonate-apatites involve the coupled substitution $(PO_4)^{3-} \rightleftharpoons (H_4O_4),(CO_3)^{2-}$; $Ca^{2+} \rightleftharpoons H_2O$; but the exact nature of this substitution is not yet known. The fluoride- and chloride-bearing carbonate-apatites are called francolite and dahllite, respectively. They are of considerable biological interest since they occur in human calculi, bones, and teeth. Dental enamels are composed in part of dahllite, and the efficacy of fluoride-bearing toothpaste partly concerns apatite crystal chemistry. Other less frequent substitutions occur in nature, involving $(AsO_4)^{3-} \rightleftharpoons (PO_4)^{3-}$ and the arsenate end members are known as svabite, $Ca_5(AsO_4)_3F$, and hedyphane, $Ca_5(AsO_4)_3Cl$. More limited substitutions can include many other ions, such as Mn^{2+} for Ca^{2+} (mangan-apatite), and Si^{4+} and S^{6+} for P^{5+}.

Apatites have been synthesized hydrothermally by the hydrolysis of monetite, $CaH(PO_4)$, or by direct fusion of the components, such as $Ca_3(PO_4)_2 + CaF_2$, which yields fluorapatite. Other apatite structures have been synthesized, usually by fusion of the components, involving the following elements and ions: Ba, Pb, Mg, Ni, Sr, Zn, Cr (3+ and 6+), Al, Fe, Na, Ce, Y, O^{2-}, $(OH)^-$, F^-, Cl^-, $(CO_3)^{2-}$, and $(SO_4)^{2-}$.

Diagnostic features. The apatite isomorphous series of minerals occur as grains, short to long hexagonal prisms terminated by pyramids, dipyramids, and the basal pinacoid (see **illus.**). The minerals are transparent to opaque, and can be asparagus-green (asparagus stone), grayish green, greenish yellow, gray, brown, brownish red, and more rarely violet, pink, or colorless. Mangan-apatite is usually blue or blue-gray to blue-green. Some apatites fluoresce in ultraviolet radiation, in particular the mangan-apatite variety. The cleavage is poor parallel to the base and fracture uneven. Apatites are brittle, with hardness 5 on Mohs scale, and specific gravity 3.1–3.2; they are also botryoidal, fibrous, and earthy.

Occurrence. Apatite occurs in nearly every rock type as an accessory mineral. In pegmatites it occurs as a late-stage fluid segregate, sometimes in large crystals and commonly containing Mn^{2+}. It often crystallizes in regional and contact metamorphic rocks, especially in limestone and associated with chondrodite and phlogopite. It is very common in basic



Apatite. (*a*) **Specimen for Eganville, Ontario, Canada (*American Museum of Natural History specimens*). (*b*) Typical crystal habits (*after C. Klein, Manual of Mineralogy, 21st ed., John Wiley & Sons, 1993*).**

to ultrabasic rocks; enormous masses occur associated with nepheline-syenites in the Kola Peninsula, Russia, and constitute valuable ores which also contain rare-earth elements. Apatite is a common accessory with magnetite ores associated with norites and anorthosites and is detrimental because it renders a high phosphorus content to the ores. Apatite occurs rarely in meteorites along with other phosphates.

Large beds of oolitic, pulverulent, and compact fine-grained carbonate-apatites occur as phosphate rock, phosphorites, or collophanes. They have arisen by the precipitation of small concretions formed by organisms and by the action of phosphatic water on bone materials or corals. Extensive deposits of this kind occur in the United States in Montana and Florida and in North Africa. The material is mined for fertilizer and for the manufacture of elemental phosphorus. *See* FERTILIZER; PHOSPHATE MINERALS; PHOSPHORUS; PYROMORPHITE. Paul B. Moore

Apes

The group of primates most closely related to humans. They include the African great apes, the gorillas and chimpanzees; the Asian great apes, the orangutans; and the lesser apes from Asia, the gibbons. The apes can be distinguished from the rest of the primates by a number of anatomical and behavioral traits, which indicate their common origin; thus they are considered a monophyletic group and are called the Hominoidea.

Apes are distinguished from other primates through such obvious features as the absence of a tail and presence of an appendix. They share a number of specializations (synapomorphies) of the skeleton, which are useful as diagnostic characters, particularly when it comes to distinguishing fossil apes, because bones and teeth are the most readily preserved parts in the fossil record. The distal end of the humerus is especially useful, both because it is one of the most robust body parts and therefore readily preserved, and because it is diagnostic of the ape condition, with a large trochlea (ulnar forearm articulation) and a well-developed trochlea ridge. The wrist is also modified for mobility of the joint. The lumbar vertebrae are distinctive in having robust pedicles, caudally inclined spinous processes, and dorsally oriented transverse processes arising from the pedicle. There are few synapomorphies of the skull, which in general retains the primitive primate condition except in individual species, but two shared specializations are the deep arched palate and relatively small incisive foramina. The teeth also are generally primitive, except in the broad, low-crowned incisors and enlarged molars. *See* DENTITION; SKELETAL SYSTEM.

Each group of apes differs from this basic ape pattern in varying degrees.

Gibbons. The gibbons retain many of the primitive ape characteristics. (Note that the specialized, or derived, conditions that characterize the apes as a group must subsequently be considered primitive when differences within the group are considered.) They have also developed a number of unique characters that are different from any other ape: They have very elongated arms (relative to body weight) with many modifications of the musculature for a brachiating form of locomotion (swinging by the arms). Their legs are also lengthened, so that they are quite efficient bipeds. They have adopted a monogamous family system, which is unusual in primates, and one of the outcomes is role sharing between males and females and lack of size variation between the sexes. They have also developed a complex system of vocal communication related both to their close social bonds and to their thick tropical forest environment, which makes visual communication difficult. At present there are 13 species of gibbon (*Hylobates*) divided into three subgenera [Hylobates (**Fig. 1**), Nomascus, Symphalangus] occupying most parts of Southeast Asia where primary forest still remains. *See* SOCIAL MAMMALS.

Orangutan. The sister group to the gibbons is the great ape and human group, which is readily distinguished by the shared presence of enlarged premolars in all its members. Within this group can be distinguished the orangutans and the African apes. The orangutans have a great many specializations that support this separation, although because of its common heritage with the chimpanzee and gorilla, and the great increase in body size of all three, the orangutan has a superficial similarity which has led in the past to all being grouped in the same family, the Pongidae. The differences separating the orangutans from other apes are in part biochemical (for example, the structure of their blood proteins) and in part morphological. The deep face, enlarged premaxilla, narrow distance between the eyes, massive zygomatic bones, smooth floor of nose, and enlarged central incisors are characters unique to the orangutan, and in all of these traits the African apes and humans retain the primitive ape condition.

The orangutan is today confined to the tropical rainforests of Borneo and Sumatra, where it exists

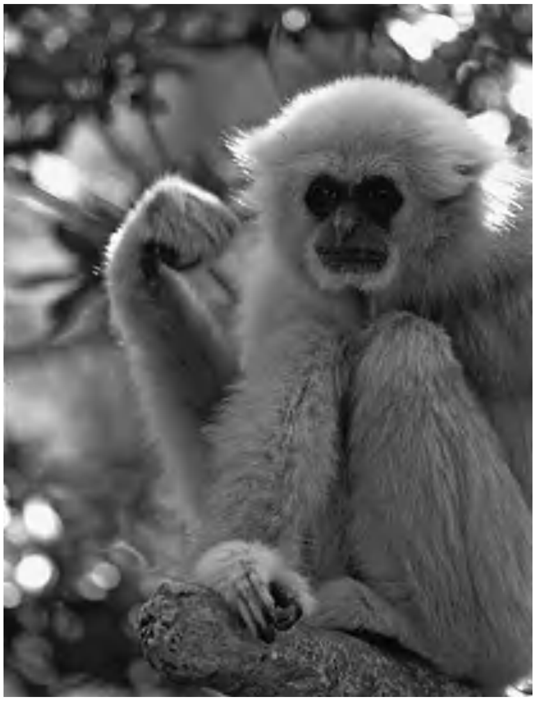


Fig. 1. *Hylobates lar* (white-handed gibbon). (*Photo by Gerald and Buff Corsi, © 2004 California Academy of Sciences*)

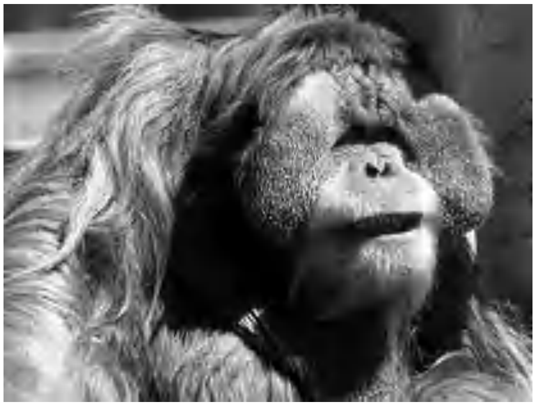


Fig. 2. *Pongo pygmaeus* (orangutan). (*Photo by H. Vannoy Davis, © 2001 California Academy of Sciences*)

as two variable species. Protein differences between them are well above the subspecies level, and there are differences in the chromosomes and in mitochondrial deoxyribonucleic acid (mtDNA) cleavage sites, and it would appear that the orangutan is a good example of recent speciation. The species from Borneo is known as *Pongo pygmaeus* (**Fig. 2**), and the one from Sumatra, as *P. abelii*. Both are largely arboreal despite their great size, which ranges in body weight from 88 to 308 lb (40 to 140 kg). They lead solitary or small-group existences, leading to the massive sexual-size variation indicated by body weight variation. Little is known of the social structure.

African apes. The other part of the great ape group consists of four species of African apes, two each of gorilla and chimpanzee. They are distinguished from the orangutans (and other primates) by specializations of the wrist and frontal sinus and the development of massive brow ridges (all of which they also share with humans), and by a further series of unique modifications of the hand that are related to its unusual form of locomotion, called knuckle walking. Their legs are reduced in length (relative to body weight) so that their arms appear long in proportion to the rest of their body. The effects of this are further increased both by elongation of the hand and by the fact that when they walk quadrupedally on the ground they support their weight on the middle phalanges (finger bones) of the hand. This raises the body to a semiupright position even when the animals are walking on all fours. The orangutan, which parallels the leg reduction and therefore the proportionally longer arms, does not use its knuckles for weight support in this way but supports itself on its clenched fists.

The common chimpanzee (*Pan troglodytes*; **Fig. 3**) inhabits much of the forested region of tropical Africa, extending beyond the forest into more seasonal wooded habitats as well. The pygmy chimpanzee, or bonobo (*Pan paniscus*), is confined to the southern loop of the Congo River, where it inhabits mainly swamp forest. The two gorilla species are also confined to tropical Africa, but the genus is divided into the eastern gorilla (*Gorilla beringei*) and the western gorilla (*G. gorilla*; **Fig. 4**). The gorillas are the largest of the primates, with body weights ranging from 165 to 396 lb (75 to 180 kg), while the chimpanzee at 88 to 110 lb (40 to 50 kg) is much

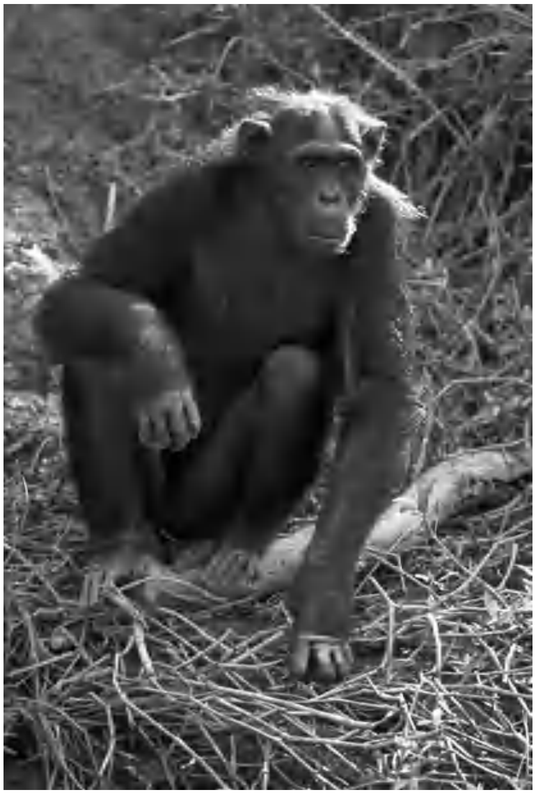


Fig. 3. *Pan troglodytes* (chimpanzee). (*Photo by Gerald and Buff Corsi, © 2002 California Academy of Sciences*)

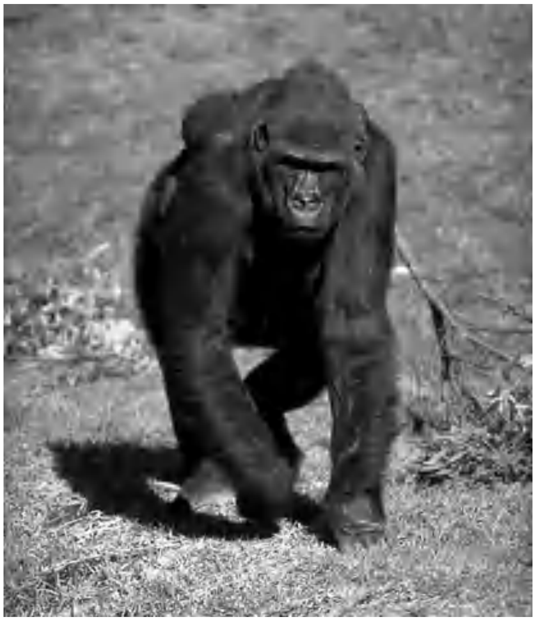


Fig. 4. *Gorilla gorilla gorilla*. (*Photo by Gerald and Buff Corsi,* © *2002 California Academy of Sciences*)

smaller. Gorilla social groups consist of a dominant male with several females and immature males and females, while chimpanzees live in fluctuating multimale or multifemale groups.

The relationships between the species of ape are shown in **Fig. 5**, with the gibbons at the top, divided into three subgenera of the genus *Hylobates*; orangutans (*Pongo pygmaeus* and *P. abelii*); the African great apes and humans at the bottom; and some fossil species that will be described below. Notice that the orangutans are put into a separate group (or clade) from the other great apes (the chimpanzees and gorillas) because the latter have been shown to be more closely related to humans in evolutionary terms than they are to the orangutan. This signifies that the idea of the "great apes" encompassing all three is not a valid taxonomic concept, and the use of a single family, the Pongidae, to include them all is not correct. The great apes, however, have many superficial similarities to each other, mostly related to their large size, so that they may be said to be similar in grade, but in terms of clade (ancestral-descendant relationship) they are different. This could be recognized by restricting the use of the family Hominidae, but other classifications are also possible based on this set of relationships.

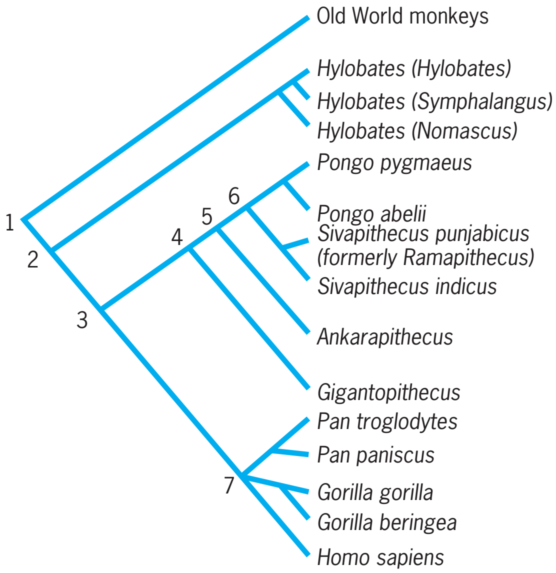


Fig. 5. Cladogram showing the relationships of fossil and living apes. The relationships are based on shared morphological characters (synapomorphies) as described in the text, and there is no time element inherent in the cladogram. The timing of divergences between groups can be inferred from the oldest data for which fossils are known: the date at which the Old World monkey split with the apes must have been in excess of 20 million years; the divergence of the orangutan must have occurred before 11 million years ago; and the divergence of humans from the African apes must have taken place before 4 million years ago. Nodes 1–7 are explained in the text.

Fossil apes. The fossil history of the apes can be traced back about 22 million years to the early Miocene. Claims have been made that the Oligocene primate *Aegyptopithecus zeuxii* from the Egyptian Fayum can be considered an ape, which could take ape history back to around 35 million years ago, but recent evidence from the postcranial bones, especially the distal end of a humerus, and from the skull, shows that *Aegyptopithecus* has none of the characters described earlier as being diagnostic of the living apes. Thus there is no reason for including them as apes, however great their superficial similarity to apes based on shared primitive characters. *Aegyptopithecus* was a small, rather heavily built arboreal primate that could be viewed as being physiologically similar to the ancestral monkeys and apes (at node 1 in the illustration).

Even by the early Miocene, for which there is a wealth of fossil evidence from many sites in Kenya and Uganda, it is far from certain that apes have been definitely identified. *Proconsul* is the best known of these primates, together with four other genera and a total of 10 species; like *Aegyptopithecus* it mainly shares primitive characters with living apes. The evidence for its being an ape comes from its postcranial morphology, which shows many similarities to ape morphology in the shoulder, elbow, and foot. These similarities have now been demonstrated for at least two of the species of *Proconsul*, and they appear to settle the question of the hominoid affinities of this fossil genus. This means that two things can be said about ape evolution: first, the apes must have differentiated from the other primates by the time *Proconsul* lived 22 million years ago; and second, the initial modifications that characterized them as apes were postcranial. This is interesting, because it will have been seen from the descriptions of the living apes that their greatest specializations, both as a group and as individuals, are postcranial ones, with all of the living apes having adopted quite extraordinary methods of locomotion: gibbons, brachiating; orangutans, fist walking; chimpanzees and gorillas, knuckle walking; and humans, bipedal walking. Perhaps the trend to this extraordinary diversity of locomotion started with the beginnings of ape evolution.

Sivapithecus. By the middle Miocene, three groups of apes had differentiated which have apparent affinities with groups of living apes. The best-documented group is *Sivapithecus* from India, Pakistan, Greece, and Turkey. These are small to medium-size apes that lived some 8–13 million years ago, and they share with the orangutan many of the specializations of the face mentioned earlier as being diagnostic of this species. The synapomorphies include the narrow distance between the eyes, concave face, smooth floor of nose, massive zygomatic bones, and enlarged central incisors (relative to the lateral incisors). One or all of these characters are known to be present in *S. indicus* and *S.* (formerly *Ramapithecus*) *punjabicus* from India and Pakistan and *Ankarapithecus meteai* from Turkey. Skulls of *S. indicus* from Pakistan and *A. meteai* from Turkey show these characters clearly, but similar fossils from Greece, which some authorities have linked with the *Sivapithecus* group, appear to lack these orangutan synapomorphies. These species are also characterized by thick enamel on their molar teeth, a feature present in orangutans and also in humans, which has led in the past to some or all of these fossils being described as humanlike or even as being actual human ancestors.

Mention can also be made of *Gigantopithecus*, which appears to share many of the characters of this node, but there is no evidence for any of the body parts characteristic of the orangutan clade and so it is not possible to place it exactly. *Gigantopithecus* is a giant ape known from India and China, and like *Sivapithecus* it has thick enameled teeth and a massively built mandibular body. Unlike the other fossils, however, it also has molariform premolars and canines and greatly reduced incisors; the most that can be said of its relationships at present is that it probably fits in the illustration between nodes 3, 4, and 6.

Dryopithecus. The second group of middle Miocene ape consists of at least three species from Europe. It includes the classic *Dryopithecus fontani*, a species comparable in size with *S. punjabicus*, and it is known mainly from France, Germany and Spain, where it has been described under numerous synonyms. There is also evidence from Hungary and Spain for several other species of *Dryopithecus* considerably smaller in size. All have the enlarged premolars, mentioned earlier to be diagnostic of the great ape and human group; and for most of the characters described above as uniquely shared by the orangutan and *Sivapithecus* species, the specimens of *Dryopithecus* from Hungary and Spain retain the primitive condition; that is, they lack the orangutan synapomorphies. They lack the thickened enamel present in the sivapithecines, but they appear to have dental enamel slightly thicker than was present in earlier apes like *Proconsul*. They have postcranial adaptations for suspensory locomotion similar to that seen in the living great apes, particularly in their limb proportions and lengthened hands, so that this group of fossil apes may now be placed near the origin of the great ape and human clade, which is at or near node 3 in Fig. 5.

A group of late Miocene apes from Greece (*Graecopithecus freybergi*, or "*Ouranopithecus macedoniensis*") have similarities in skull form to *Dryopithecus*, but what little is known of their postcranial skeleton suggests that they retain the primitive condition for hominoids, and they have thickened enamel on their teeth. Thus, their present taxonomic position is uncertain, and they are not shown in Fig. 5.

Kenyapithecus. The third group of middle Miocene fossil ape is an African group represented by the genera *Kenyapithecus* and *Afropithecus* from Kenya and *Heliopithecus* from Saudi Arabia. There is some uncertainty about the status of this group and of its phylogenetic affinities. *Kenyapithecus* shows some similarities to later specimens of *Sivapithecus*, and like this Asian genus the molars have thickened enamel, but on the other two genera the enamel is only moderately thickened. All of them, however, share massive premolar enlargement, and it is this feature that groups them. In other respects, especially their postcrania, they appear to retain ancestral characters, so that they also fall on or near node 3 in Fig. 5. One species of this genus is also known from outside Africa, in Turkey, and the date of this site (Pasalar ~15 million years ago) provides the earliest evidence of the emigration of apes from Africa to Europe and Asia. There is no fossil evidence for the later stages of African ape and human evolution (node 7 or higher in Fig. 5). *See* FOSSIL HUMANS; MONKEY.

Peter Andrews

Bibliography. C. Groves, *Primate Taxonomy*, 2001; W. Hartwig, *The Primate Fossil Record*, 2002; J. R. Napier and P. H. Napier, *A Handbook of Living Primates*, 1967; D. R. Swindler (ed.), *Comparative Primate Biology*, vol. 1, 1986; F. S. Szalay and E. Delson, *Evolutionary History of the Primates*, 1979.

Aphasia

Impairment in the use of spoken or written language caused by injury to the brain, which cannot be accounted for by paralysis or incoordination of the articulatory organs, impairment of hearing or vision, impaired level of consciousness, or impaired motivation to communicate. The language zone in the brain includes the portion of the frontal, temporal, and parietal lobes surrounding the sylvian fissure and structures deep to these areas. In right-handed persons, with few exceptions, only injury in the left cerebral hemisphere produces aphasia. Lateralization of language function is variable in left-handers, and they are at greater risk for becoming aphasic from a lesion in either hemisphere. *See* HEMISPHERIC LATERALITY.

Distinctive recurring patterns of deficit are associated with particular lesion sites within the language zone. These patterns may entail selective impairment of articulation, ability to retrieve concept names, or syntactic organization. Other dissociations affect principally the auditory comprehension of speech, the repetition of speech, or the recognition of written words. The erroneous production of

-S

unintended words in speech (paraphasia), oral reading (paralexia), or writing (paragraphia) is a feature of some forms of aphasia.

In general, injury to the speech zone anterior to the rolandic fissure (Broca's area) results in nonfluent aphasia, with effortful articulation, loss of syntax, but relatively well-preserved auditory comprehension. Postrolandic lesions typically result in fluent aphasia. Facility of articulation, grammatical organization, and rate of speech are well preserved, while the comprehension of language and word choice are most affected. Within the group of fluent aphasias three major subtypes may be distinguished. Wernicke's aphasia, associated with injury to the posterior portion of the first temporal convolution, is characterized by severely impaired auditory comprehension, and fluent, but markedly paraphasic speech, sometimes at faster than normal rate. Anomic aphasia, typically caused by injury at the temporoparietal junction, leaves the patient unaffected in comprehension, ease of articulation, or grammatical organization, but unable to recall the principal nouns and verbs for the message. Conduction aphasia features selectively impaired repetition, with relatively fluent speech production and good comprehension. This form of aphasia is often associated with damage in the white matter underlying the parietal operculum. In the transcortical aphasias both repetition and the ability to recite overlearned material are remarkably preserved, while communicative speech is virtually impossible. Anterior (transcortical motor) and posterior (transcortical sensory) forms are recognized. *See* BRAIN.

While the above syndromes often occur in clearcut form, mixed forms of aphasia, caused by multiple lesions or lesions spanning anterior and posterior portions of the speech zone, are quite common, and massive destruction of the entire language area results in a global aphasia. Further, individual variations in behavioral manifestations of similar lesions have set limits on the strict assignment of function to structures within the language area.

Preadolescent children suffering aphasia after unilateral injury usually recover rapidly, presumably by virtue of the capacity of the right cerebral hemisphere early in life to acquire the language functions originally mediated by the left hemisphere. Capacity for recovery of function decreases during later adolescence and young adulthood.

Complete recovery in adults after a severe injury is much less common, and severe aphasia may persist unchanged for the duration of the person's life. Many individuals are aided by remedial language training, while others continue severely impaired. *See* MEMORY.

Harold Goodglass

Apiales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Rosidae of the class Magnoliopsida (dicotyledons). The order (also known as the Umbellales) comprises two families, the Araliaceae, with about 700 species, and the Apiaceae, with about 3000. They are herbs or woody plants with mostly compound or conspicuously lobed or dissected leaves (see **illus.**), well-developed schizogenous secretory canals, separate petals, and an inferior ovary that has only one ovule in each locule. The nodes are generally multilacunar, and the pollen is trinucleate. Ginseng (*Panax*) and English ivy (*Hedera helix*) are well-known members of the Araliaceae.

Cow parsnip (*Heracleum lanathum*), family Umbelliferae, order Umbellales. (*Photograph by A. W. Ambler*, *from National Audubon Society*)

The Apiaceae are mostly aromatic herbs, most numerous in temperate regions. The flowers consistently have an ovary of two carpels, ripening to form a dry fruit that splits into halves, each containing a single seed. Some common garden vegetables and spice plants, including carrot (*Daucus carota*), parsnip (*Pastinaca sativa*), celery (*Apium graveolens*), parsley (*Petroselinum crispum*), caraway (*Carum carvi*), and dill (*Anethum graveolens*), belong to the Apiaceae, as do also such notorious poisonous plants as the poison hemlock (*Conium*) and water hemlock (*Circuta*). *See* ANISE; CARROT; CELERY; FENNEL; GINSENG; MAGNOLIOPHYTA; MAGNOLIOPSIDA; PARSLEY; PARSNIP; PLANT KINGDOM; ROSIDAE.

Arthur Cronquist

Apical dominance

Correlative inhibition of the growth of lateral (axillary) shoots exerted by the growing apical bud of the plant stem. Partial or complete apical dominance can also occur in the initiation of lateral roots and in the orientation of lateral organs such as branches, leaves, and stolons.

Origin of lateral meristems. In the apical meristem, cell division occurs in the young forming leaves and along the flanks of the apical bud. As the number of cells increases, they elongate, pushing the apical meristem upward and leaving a small portion of the dividing cells behind the axil of each of the laterally forming leaves. This meristem, called the lateral meristem, may remain a small group of cells with little organization or may develop into an axillary bud possessing short internodes, unexpanded leaves, and an apical meristem.

In roots, branching does not directly involve the apical meristem. Lateral roots do not form from organized buds but originate in the layer of cells between the endodermis and root vascular system (pericycle). Cell division within the pericycle results in meristematic tissue which will form a lateral root. *See* APICAL MERISTEM; LATERAL MERISTEM; ROOT (BOTANY).

Lateral plant growth. The degree of apical dominance over the lateral buds varies with the plant species. Some plants, such as pea and sunflower, exhibit strong apical dominance, causing the formation of single branchless shoots. Other plants, such as tomato, have weak apical dominance, allowing axillary bud growth and a bushy growth habit.

Apical dominance can be broken by several factors, including apical bud removal (decapitation), horizontal positioning of the plant's main axis (gravistimulation), shoot inversion, low light intensity, or short-day photoperiods. In some situations, apical dominance is weakened as the plant becomes older. For example, in sycamore and ash, strong apical dominance controls the early growth of the plant, which results in a single leader shoot. In later years, weak apical dominance causes a branching habit. In herbaceous plants, apical dominance is often weakened at the end of the growing season, especially when the terminal bud forms a terminal flower. After apical dominance is broken, the released lateral shoots have a hierarchy of dominance of their own. Frequently, the uppermost shoot becomes dominant and inhibits the growth of the lower shoots. *See* PHOTOPERIODISM; PLANT MOVEMENTS.

Plant organs other than the main shoot are under the control of apical dominance. In roots, strong apical dominance causes tap-root growth, whereas weak apical dominance results in a fibrous root system. Also, increased lateral root initiation in the pericycle occurs in detipped roots. Leaves and branches which grow at characteristic angles to the main axis of the stem will grow more upright after removal of the apical bud of the main stem. The growth form of some organs is also controlled by the presence of the apical bud. Rhizomes (underground axillary shoots lacking chlorophyll and having rudimentary leaves) of potato are one example of such control. Rhizomes will grow upright, forming green, leafy shoots if the apical bud and all the aboveground axillary buds are removed.

Lower plants, such as mosses and ferns, as well as fungi and algae, also exhibit apical dominance. In one example, the removal of the fruiting body of the fungus, *Phycomyces*, results in the formation of lateral branches which are normally absent.

Nutritional factors. One hypothesis for the mechanism of apical dominance is based on the competition of nutrients between centers of growth. According to this hypothesis, nutrients are preferentially transported to the apical bud, causing deficiencies in the axillary buds. Decapitation may shift nutrient transport from the sink (apical bud) and allow for accelerated transport to the lateral buds. Nutrient deficiencies have been found to inhibit bud growth, and nutrient supplementation can overcome bud inhibition in some cases. However, the concept that apical dominance is controlled solely by nutritional factors is unlikely. Frequently, there has been failure to relieve bud inhibition by direct application of nutrients, and the nutrient levels of inhibited buds are not always lower than the levels in growing buds. *See* PLANT MINERAL NUTRITION.

Hormonal factors. A second hypothesis for the mechanism of apical dominance proposes that a plant hormone produced in the apical bud acts as the correlative signal.

Auxin. The plant growth hormone, auxin, is synthesized in the young leaves of the apical bud and is transported down the stem. Addition of auxin to decapitated shoots restores apical dominance, thus replacing the function of the apical bud in its inhibition of lateral bud growth. Therefore, auxin is proposed to be the primary signal involved in maintaining apical dominance in stems. Accordingly, if auxin directly inhibits lateral bud growth, then high levels of auxin are also expected in the inhibited buds. However, contrary to this theory, lateral buds have been found to be low in auxin. Auxin-directed nutrient transport has been proposed to account for inhibition of lateral bud growth by auxin since some nutrients accumulate in areas of high auxin levels. Thus, tissues low in auxin, such as axillary buds, may be starved for nutrients. Auxin may also control the transport of other hormones essential for lateral bud growth. However, it is interesting to note that auxin levels actually increase in growing lateral buds. Thus, auxin may be required for shoot growth after buds are released from apical dominance.

In roots, auxin promotes the initiation of lateral roots. Therefore, auxin may not act as the correlative inhibitor in roots as it is proposed to be in shoots. However, little research has been conducted on the hormonal control of apical dominance in roots. *See* AUXIN.

Cytokinins. Cytokinins are well known as cell-division factors and promote lateral bud growth in plants under strong apical dominance. Cell division is blocked in inhibited axillary buds. In the early stages of cytokinin-promoted bud growth, cell division is stimulated within 1 h, accompanied by increases in nucleic acids. Doubtless, outgrowth of suppressed lateral buds involves activation of certain genes, resulting in changes in the rates of synthesis of key regulatory proteins. However, investigations of hormone-induced gene products involved in apical dominance have not been completed.

It is postulated that the auxin cytokinin ratio is the important factor in determining whether lateral buds grow or remain inhibited. There is evidence that auxin inhibits cytokinin transport from the roots (primary site of cytokinin biosynthesis) to the lateral bud. Supporting this observation is the fact that levels of cytokinin at the bud site are low when the buds are suppressed and increase greatly when the suppression is broken by decapitation or gravistimulation.

Cytokinin application directly to potato rhizomes results in their conversion to leafy, upright shoots. In decapitated and disbudded plants, the presence of roots is necessary for this rhizome-shoot conversion. Thus, lack of cytokinin from the roots may be responsible for the maintenance of the rhizome growth form.

In roots, cytokinin strongly inhibits root initiation in detipped roots. It has been proposed that cytokinin is the primary inhibiting agent from the root apex controlling apical dominance over lateral rooting. *See* CYTOKININS.

Other hormones. Although auxin and cytokinin are the major hormones regulating apical dominance, there is increasing evidence that other hormones (gibberellin, abscisic acid, and ethylene) contribute to this process. Gibberellin application enhances auxin's effect in maintaining apical dominance in decapitated shoots and has been found to promote auxin accumulation in the area of application. Thus, plant tissues high in gibberellin, such as the apical bud, may accumulate a high concentration of auxin. However, gibberellin also stimulates cytokinin-promoted lateral bud growth. Thus, it appears that gibberellin may contribute to the elongation of already growing buds. In rhizome growth of potatoes, addition of gibberellin and auxin together will cause aerial axillary buds to form rhizomes, whereas either hormone alone will not yield that effect. *See* GIBBERELLIN.

The growth inhibitor, abscisic acid, occurs in high levels in inhibited lateral buds under strong apical dominance in some plant species. The direct addition of abscisic acid to lateral buds will counteract the growth-promoting effect of cytokinin application or decapitation of the main stem apex. It has been proposed that a high level of auxin in the stem may trigger abscisic acid synthesis in the lateral buds. Abscisic acid also inhibits lateral rooting when applied to detipped roots. However, the role of abscisic acid in apical dominance of roots and shoots is still unclear. *See* ABSCISIC ACID.

The gaseous plant hormone, ethylene, inhibits lateral bud growth promoted by cytokinin or decapitation when it is supplied continuously. Ethylene is produced at especially high levels at the plant nodes where the lateral buds are situated. High levels of auxin in plant tissues promote ethylene biosynthesis in these nodes, and it is postulated that auxin levels in the stem stimulate ethylene production, thus suppressing lateral bud growth. Accordingly, decapitation should cause a lowering of ethylene production at the nodes, reducing bud inhibition. However, no lowering of ethylene is observed in the nodes of decapitated shoots. Therefore, an inhibitory role for ethylene in apical dominance seems unlikely. In contrast to an inhibitory effect, a short burst of ethylene promotes lateral bud elongation. Thus, under certain conditions, ethylene may contribute to bud growth after release from apical dominance. Still another hypothesis suggests that ethylene produced in inverted or gravistimulated shoots retards main shoot elongation and allows for the subsequent axillary bud outgrowth. *See* ETHYLENE; PLANT GROWTH; PLANT HORMONES.

Marcia Harrison

Bibliography. L. J. Audus and F. T. Addicott (eds.), *Abscisic Acid*, 1983; M. A. Harrison and P. B. Kaufman, Estimates of free and bound indole-3-acetic acid and zeatin levels in relation to regulation of apical dominance and tiller release in oat shoots, *J. Plant Growth Regul.*, 2:215–223, 1983; M. A. Harrison and P. B. Kaufman, The role of hormone transport and metabolism in apical dominance in oat, *Bot. Gaz.*, 145:293–297, 1984; I. D. J. Phillips, Apical dominance, *Annu. Rev. Plant Physiol.*, 26:341–367, 1975; T. K. Prasad and M. G. Cline, The role of gravity in apical dominance: Effects of clinostating on shoot inversion-induced ethylene producion, shoot elongation and lateral bud growth, *Plant Physiol.*, 83:505–509, 1987; H. E. Street and H. Opik (eds.), *The Physiology of Flowering Plants: Their Growth and Development*, 3d ed., 1992; P. F. Wareing and I. D. J. Phillips, *Growth and Differentiation in Plants*, 3d ed., 1981.

Apical meristem

Permanently embryonic tissue involved in cell division at the apices of roots and stems, and forming dynamic regions of growth. These apical meristems, usually consisting of small, densely cytoplasmic cells, become established during embryo development. Thereafter they divide, producing the primary plant body of root and shoot. Below the apical meristems, tissue differentiation begins; the protoderm gives rise to the epidermal system, the procambium to the primary vascular system, and the ground meristem to the pith and cortex (**Figs. 1** and **2**). Plant apical meristems have been the object of experiments on development similar to those carried out on animal embryos.

Roots. Root apical meristem is covered by a root cap, a region of parenchymatous, often starch-containing cells which has a protective function and is responsible for perceiving gravitational changes. In many monocotyledons, including grasses, the root cap has a separate initiating region, the calyptrogen. In many vascular cryptogams the root apex has a prominent tetrahedral apical cell (**Fig. 3**). In flowering plants, root apices have multicellular apical meristems. Root tips have been shown to possess a central region, usually hemispherical, which consists of cells which rarely divide or synthesize deoxyribonucleic acid (DNA), and have less ribonucleic acid (RNA) and protein than adjacent cells; this region is known as the quiescent center (**Fig. 4**). The cells which divide and give rise to root tissues

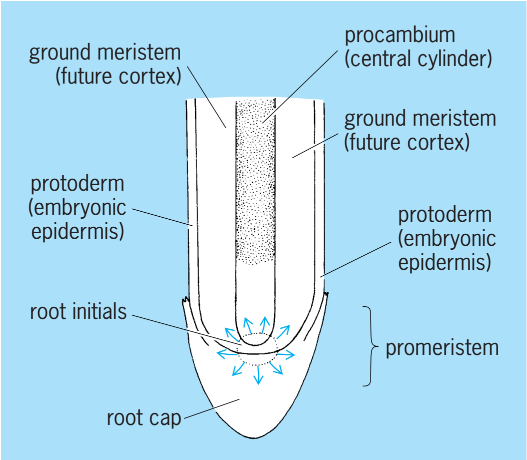

ground meristem (future cortex)
procambium (central cylinder)
ground meristem (future cortex)
protoderm (embryonic epidermis)
protoderm (embryonic epidermis)
root initials
promeristem
root cap

Fig. 1. Diagram of a root apical meristem. Cortex and central cylinder have separate initials; epidermis and root cap have a common origin.

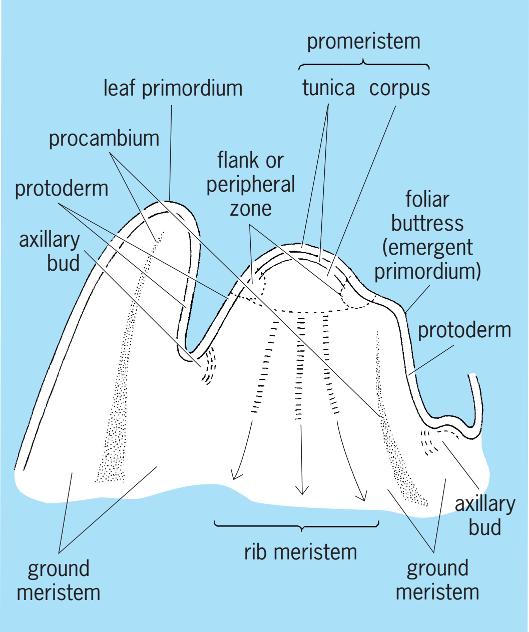

promeristem
leaf primordium
tunica
corpus
procambium
flank or peripheral zone
protoderm
foliar buttress (emergent primordium)
axillary bud
protoderm
axillary bud
rib meristem
ground meristem
ground meristem

Fig. 2. Diagram of a hypothetical shoot apical meristem with a two-layered tunica.

lie around the periphery of this region, and are sometimes described as constituting a proximal and distal meristem (Fig. 4). Cells in the quiescent center are regarded as cells that are mitotically young and genetically sound; they can renew the initial cells from time to time. Surgically isolated quiescent centers of corn roots grown in sterile nutrient medium could form whole roots.

Single apical cells of cryptogams were considered to become quiescent during development and to represent a quiescent center, but work on the aquatic fern *Azolla* demonstrated conclusively that the apical cell functions as an initial, dividing repeatedly in a very regular fashion to give rise to merophytes, structural units whose origin and subsequent differentiation can be accurately followed.

Lateral roots. Primordia of lateral roots have an endogenous (internal) origin in deep-seated tissues of the parent root, usually the pericycle, at some distance from the root apex. In cryptogams, the endodermis may contribute to lateral root formation. In some species, lateral roots arise in greater profusion, close to the apical meristem and to one another (**Fig. 5***a*, *b*); they are then usually formed according to a more or less regular pattern. Adventitious roots may be formed on stems or leaves, usually after excision from the plant; their meristems originate by dedifferentiation of mature stem of leaf tissues. This occurs in propagation by cuttings. *See* ENDODERMIS; PERICYCLE.

Shoots. Shoot apices vary greatly in size and shape. The diameter can vary from about 50 micrometers in *Zea* to 0.14 in. (3.5 mm) in *Cycas revoluta*; the shape may be elongated and conical, dome-shaped, flat, or even slightly concave. The distance from the center of the apex to the youngest leaf primordium

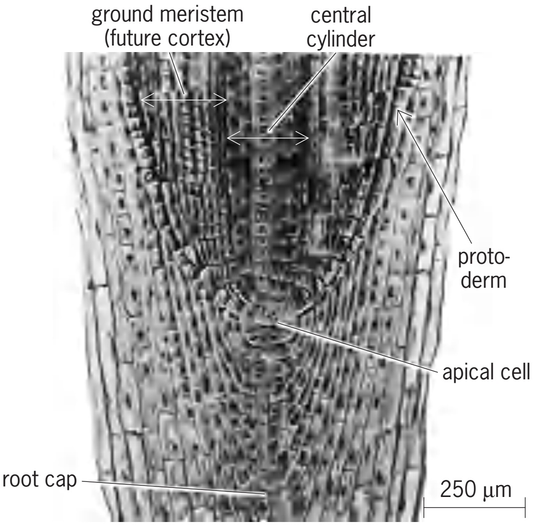

ground meristem (future cortex)
central cylinder
proto-derm
apical cell
root cap
250 µm

Fig. 3. Median longitudinal section of the root apical meristem of *Equisetum*.

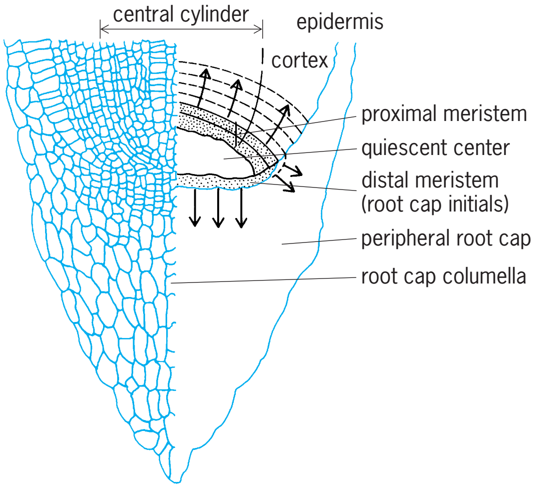

central cylinder
epidermis
cortex
proximal meristem
quiescent center
distal meristem (root cap initials)
peripheral root cap
root cap columella

Fig. 4. Longitudinal section of corn root tip, showing the quiescent center with the proximal and distal meristems at its periphery. (*After J. G. Torrey and L. J. Feldman, The organization and function of the root apex, Amer. Sci., 65:334–344, 1977*)

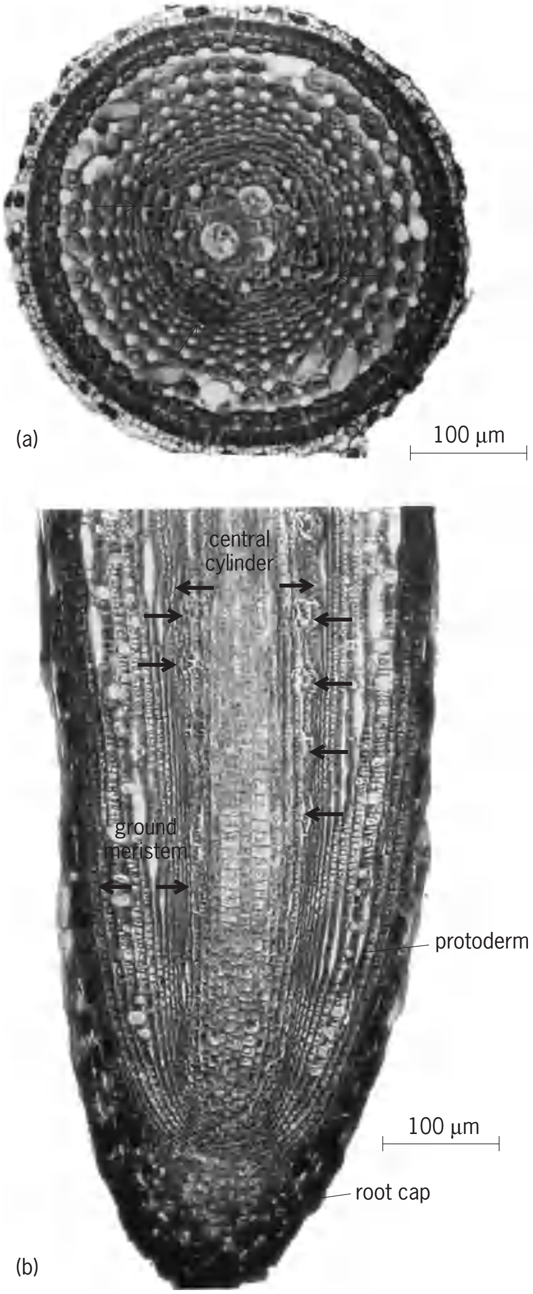


Fig. 5. Root tip of *Pontederia cordata*: (*a*) cross section and (*b*) longitudinal section showing the endogenous origin of lateral root primordia (arrows), and the various tissue regions of the root. (*From W. A. Charlton, Distribution of lateral roots and pattern of lateral initiation in Pontederia cordata L., Bot. Gaz., 136:225–235, 1975*)

also varies considerably. Apices increase in size during the development of a single plant; for example, the apical meristem of flax *(Linum usitatissimum)* increases in area 20-fold from the seedling up to the time of flowering. Apices may also change in size during the time (plastochron) between the formation of one leaf primordium and the next.

A single apical cell is present in shoot apices of bryophytes and vascular cryptogams (**Fig. 6**); however, surrounding cells are also mitotically active, and these plants have multicellular apical meristems. In flowering plants, the outer layer or layers of cells (tunica) may divide predominantly by walls at right angles to the surface; the inner tissue (corpus), in less regular planes (**Figs. 7** and **8**). Regions of the apical meristem may react differently to certain stains, reflecting so-called cytohistological zonation. One type of zonation may be superimposed upon another.

Cells in the central terminal region of the vegetative shoot apex divide less actively than those on the flanks or at the periphery, where leaf primordia are formed. Various surgical experiments, involving incision of the apical meristem, have shown that new apices can be regenerated from portions of the flank. Excised apical meristems, devoid of leaf primordia, can be successfully grown on agar nutrient medium, in the presence of auxin, and will eventually yield new plants.

Leaf primordia. Leaf primordia are formed on the flanks of the shoot apex, in an arrangement or phyllotaxy characteristic of the species. Usually, periclinal divisions (walls parallel to the surface) occur in

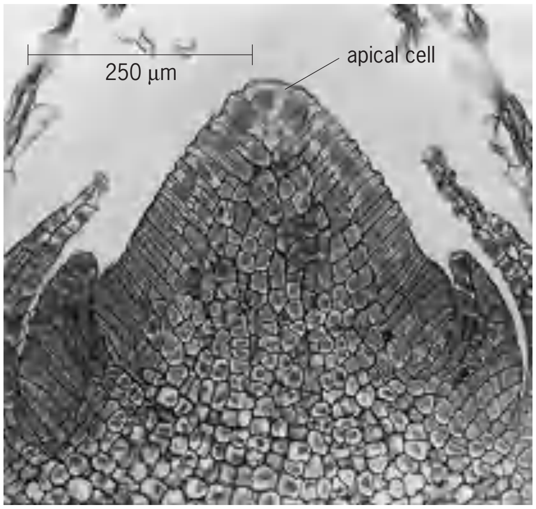


Fig. 6. Median longitudinal section of the shoot apical meristem of *Equisetum*, showing apical cell.

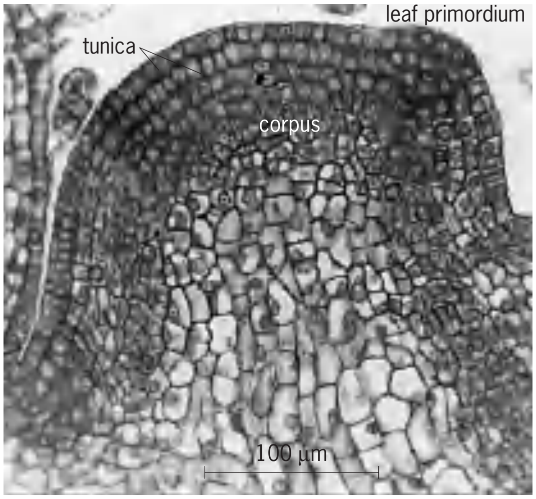


Fig. 7. Longitudinal section of the shoot apical meristem of potato, *Solanum tuberosum*, showing a two-layered tunica and a young leaf primordium on the right.

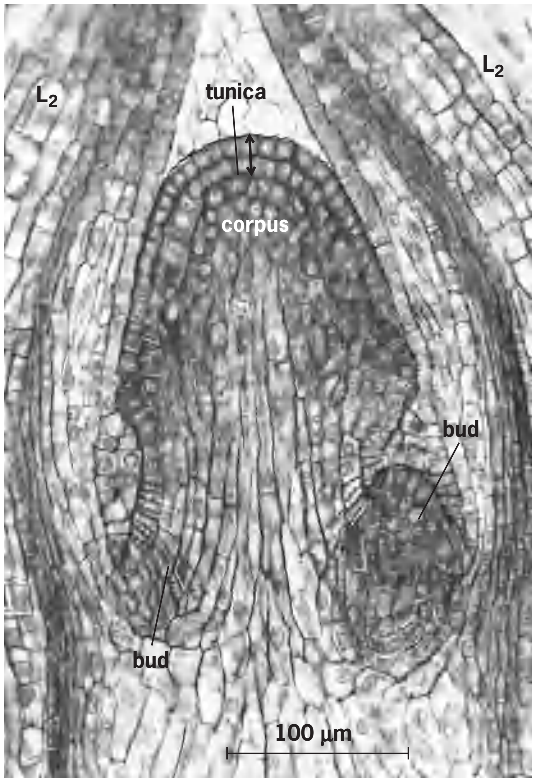


Fig. 8. Longitudinal section of the shoot apical meristem of *Alternanthera philoxeroides*, showing a two-layered tunica. Unequal axillary buds are present in the axils of the second-youngest pair of leaves, L_2. (*From E. G. Cutter, Morphogenesis and developmental potentialities of unequal buds, Phytomorphology, 17:437–445 1967*)

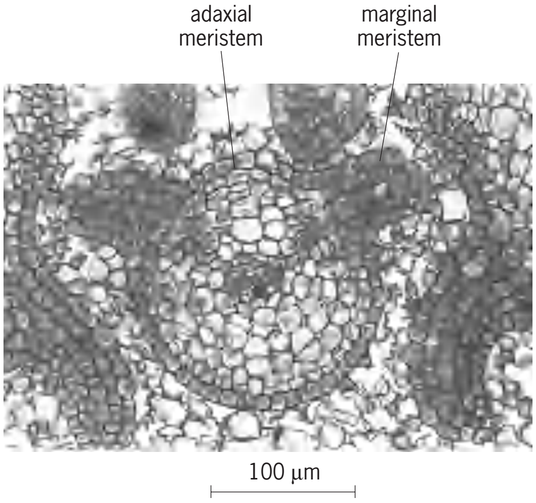


Fig. 9. Transverse section of a young leaf primordium of *Alternanthera philoxeroides*, showing marginal and adaxial meristems. (*From E. G. Cutter, Plant Anatomy, pt. 2*: *Organs, Edward Arnold Ltd., London, 1971*)

the second and perhaps deeper layers of cells, in localized positions (Fig. 7). If the leaf is sheathing when mature (as in grasses), mitotic activity spreads around the circumference of the apex. The leaf primordium grows apically at first, but in angiosperms apical growth is short-lived and is followed by intercalary growth; in fern leaf primordia, apical growth is prolonged. Marginal meristems divide to form the leaf blade or lamina (**Fig. 9**); an adaxial meristem may contribute to the thickness of the leaf. *See* LEAF.

Lateral buds. In most angiosperms, in addition to leaf primordia, the apical meristem, or its derivatives, gives rise to bud primordia. These usually occupy axillary positions (that is, above the junction of the leaf with the stem), and are exogenous (that is, originate superficially), sometimes developing from pockets of meristematic tissue separated from the apical meristem by more mature, differentiated tissue (Fig. 8). In some plants, buds occurring in the axils of a pair of opposite leaves may be of unequal size and growth potentiality (Fig. 8). Sometimes groups of buds are present. A few plants do not branch, and no bud primordia are formed. Branching may also occur by equal division, or dichotomy, of the shoot apex; this is common in cryptogams, and has now been reported in a number of angiosperms. Bud primordia are usually inhibited to a greater or lesser degree by the parent shoot apex.

Adventitious buds are formed on root cuttings (as in raspberry) or leaf cuttings (as in *Begonia*), and

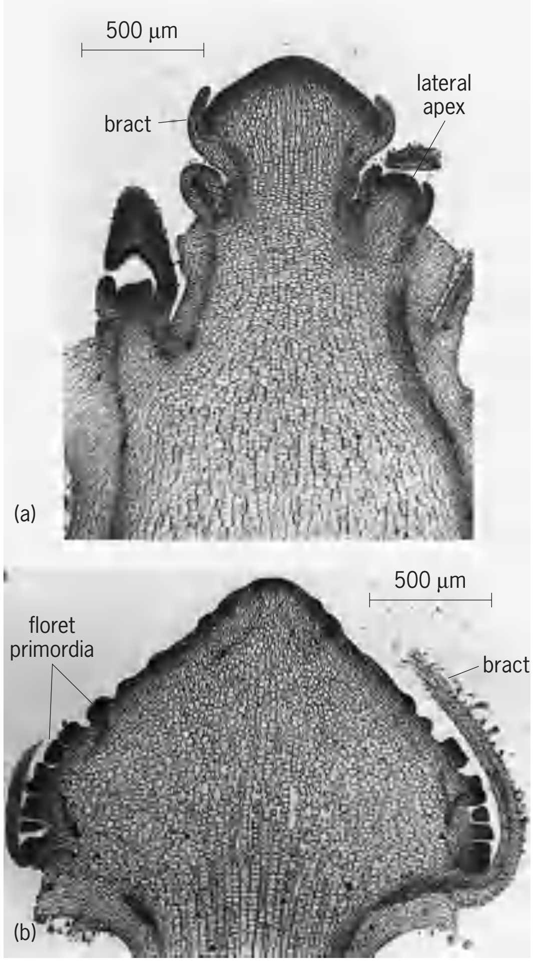


Fig. 10. Longitudinal sections of inflorescence meristems of *Calendula*. (*a*) Dome-shaped inflorescence apex with bract primordia at the periphery; a vegetative or transitional lateral apex is present on the right; note the difference in size. (*b*) Older inflorescence apex with floret primordia, at same magnification. Note increase in both width and height. (*From E. G. Cutter, Plant Anatomy pt. 2*: *Organs, Edward Arnold Ltd., London, 1971*)

they may be exogenous or endogenous in origin. *See* BUD.

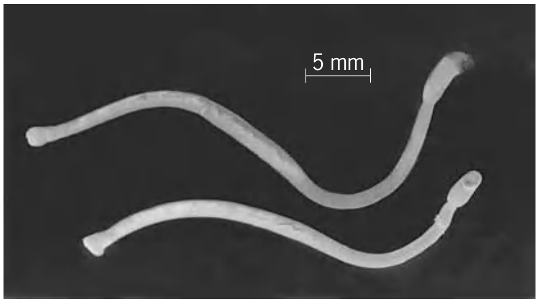
Primary thickening meristems. These meristems bring about growth in thickness of many monocotyledons, and a few dicotyledons which have little secondary growth. They consist of files of elongated cells (somewhat resembling cambium in appearance) on the "shoulders" of the shoot tip, usually below the leaf primordia; by their division, they add to the width of the shoot tip and sometimes raise the periphery of the apical region so that the apex itself lies in a depression.

Floral meristems. Under appropriate environmental stimuli (such as daylength), vegetative shoot apices may be induced to develop as apices of inflorescences or individual flowers. This involves a change from indeterminate to determinate growth. Indeed, the transition to flowering involves many changes, for example in size, shape, rate of mitosis, or amount of RNA. Inflorescence apices, such as those of the garden marigold, *Calendula*, may increase markedly in size, become more steeply domed, and finally form floret primordia (**Fig. 10***a*, *b*). These dramatic changes may occur in a matter of hours. Floral apices give rise to a sequence of primordia of floral organs (sepals, petals, stamens, carpels), many of which originate in a manner similar to leaf primordia. *See* FLOWER; INFLORESCENCE. Elizabeth G. Cutter

Bibliography. P. W. Barlow, Towards an understanding of the behaviour of root meristems, *J. Theoret. Biol.*, 57:433–451, 1976; E. M. Gifford, Jr., and G. E. Corson, Jr., The shoot apex in seed plants, *Bot. Rev.*,37:143–229, 1971; J. D. Mauseth, *Plant Anatomy*, 1988; J. G. Torrey and L. J. Feldman, The organization and function of the root apex, *Amer. Sci.*, 65:334–344, 1977.

Aplacophora

A class of vermiform marine mollusks covered by a cuticle invested with innumerable calcareous spicules; also known as solenogasters. There are 277 named species. Most species are less than 0.4 in. (10 mm) in length, but a few attain 12 in. (300 mm). The head with a mouth is poorly differentiated; there is a small posterior mantle cavity, or cloaca. Despite their specialized shape (see **illus.**), Aplacophora retain a primitive molluscan condition in the radula, integument, gonad-pericardium, muscles, and nervous system. There are two distinct subclasses, the creeping neomenioids (Neomeniomorpha) with a narrow foot within a ventral groove, and the burrowing chaetoderms (Chaetodermomorpha) without either a groove or foot, and with a cuticular shield around the mouth. Aplacophora are continental-shelf and deep-sea forms found from subtidal depths to hadal depths (30,000 ft or 9000 m), and in some rare localities may be dominant organisms. *See* CHAETODERMOMORPHA; MOLLUSCA; NEOMENIOMORPHA. Amelie H. Scheltema



Living *Chaetoderma nitidulum*. (*Courtesy of R. Robertson*)

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; A. H. Scheltema, Ancestors and descendents: Relationships of the Aplacophora and Polyplacophora, *Amer. Malacol. Bull.*, 6:57–68, 1988; A. H. Scheltema, Position of the class Aplacophora in the phylum Mollusca, *Malacologia*, 17(1):99–109, 1978; E. R. Trueman and M. R. Clarke (eds.), *The Mollusca*, vol. 10: *Evolution*, 1985.

Aplite

A fine-grained, sugary-textured rock, generally of granitic composition; also any body composed of such rock. *See* GRANITE.

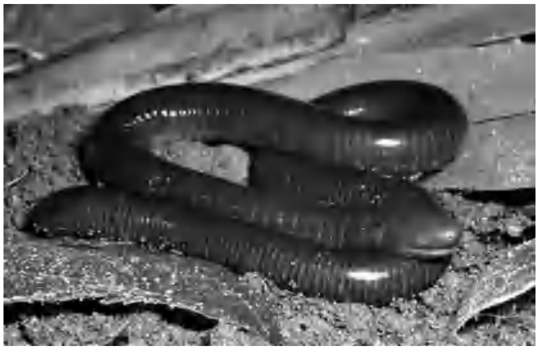
The light-colored rock consists chiefly of quartz, microcline, or orthoclase perthite and sodic plagioclase, with small amounts of muscovite, biotite, or hornblende and trace amounts of tourmaline, garnet, fluorite, and topaz. Much quartz and potash feldspar may be micrographically intergrown in cuneiform fashion.

Aplites may form dikes, veins, or stringers, generally not more than a few feet thick, with sharp or gradational walls. Some show banding parallel to their margins. Aplites usually occur within bodies of granite and more rarely in the country rock surrounding granite. They are commonly associated with pegmatites and may cut or be cut by pegmatites. Aplite and pegmatite may be gradational or interlayered, or one may occur as patches within the other. *See* PEGMATITE.

Aplites form in different ways. Some represent granite or pegmatite which recrystallized along fractures and zones of shearing. Others are of metasomatic (replacement) origin. Many form from residual solutions derived from crystallizing granitic magma (rock melt). If these fluids retain their volatiles, pegmatites may form. If the volatiles escape, a more viscous fluid may be created, and a fine-grained (aplitic) texture may be developed. *See* IGNEOUS ROCKS; MAGMA; METASOMATISM. Carleton A. Chapman

Apoda

The smallest order of class Amphibia. The order Apoda is sometimes called Gymnophiona and is known by the common name "caecilian." Caecilians



Mexican burrowing caecilian (*Dermophis mexicanus*). (*Photo © 2003 J. A. Campbell*)

are wormlike, legless amphibians with indistinct or even hidden eyes. A series of annular grooves (annuli) are usually present along the length of the body, heightening the resemblance to an earthworm (see **illustration**). Most caecilians lead a burrowing existence, though members of one genus, *Typhlonectes*, are aquatic. Some species have the eyes hidden beneath the bones of the skull and probably are blind, but others at least are able to distinguish movement. A unique feature of some caecilians among modern Amphibia is the presence of scales buried in the skin.

There are more than 165 species (6 families, 33 genera) of caecilians confined to tropical regions of both the Eastern and Western hemispheres. Due to the generally cryptic nature of caecilians, the true species richness may be much greater. Many species are less than 1 ft (30 cm) in length, but three species of the genus *Caecilia* grow to over 3 ft (90 cm). The breeding habits of caecilians are poorly known. Caecilians are internal fertilizers, and some species lay eggs, while others bring forth their young alive. The embryos of the species that bear living young are nourished in the later part of their embryonic development by "uterine milk," which is secreted by the mother. In some of the species that lay eggs, there is an aquatic larval stage. Caecilians are carnivorous, but little is known of their food habits. Captive specimens have fed on earthworms, and in the natural state caecilians have been known to eat lizards.

Fossil caecilians are rare. The earliest known were found in Arizona and date from the Early Jurassic period (206 million years before present). Among the significant characteristics of this fossil series was the existence of girdles that supported limbs. It was once proposed that caecilians share a closer relationship to salamanders (sister group) than to frogs, but recent molecular phylogenetic data support a more distant relationship of caecilians to other amphibians. New fossil representatives and continued molecular data will shed increasingly clear light on the phylogeny of Order Apoda. *See* AMPHIBIA; ANURA; URODELA.

Richard G. Zweifel; W. Ben Cash

Bibliography. D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, New York, 2001; F. H. Pough et al., *Herpetology*, 3d ed., Prentice Hall, Upper Saddle River, NJ, 2004; E. H. Taylor, *The Caecilians of the World: A Taxonomic Review*, 1968; R. Zardoya and A. Meyer, On the origin of and phylogenetic relationships among living amphibians, *Proc. Nat. Acad. Sci.*, 98(13):7380–7383, 2001; G. R. Zug, L. J. Vitt, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001.

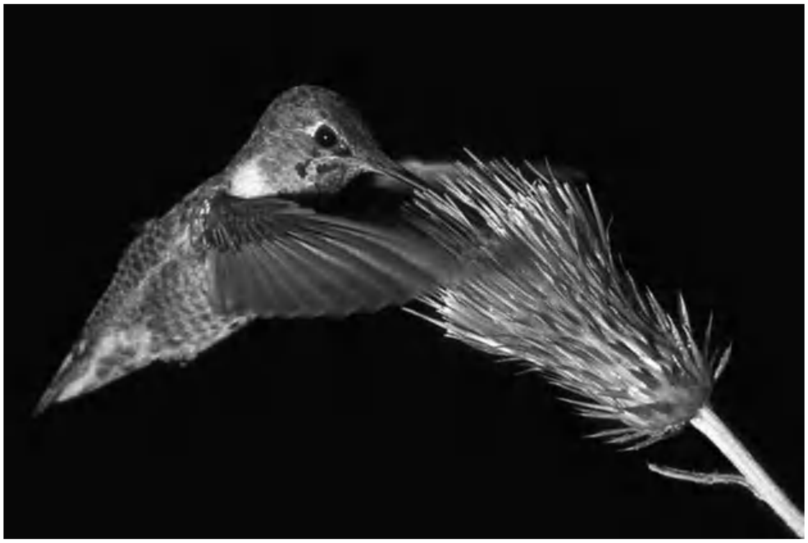
Apodiformes

An order of birds consisting of two dissimilar groups, the swifts (Apodi) and the hummingbirds (Trochili). These birds have been placed together because of several anatomical specializations of the wings and feathers, although an alternative classification based on morphological and molecular characters places the swifts and hummingbirds in two separate orders. Swifts and hummingbirds have strong wings and are excellent fliers but have small, weak feet and cannot walk. Some of these features may be convergences because of the rapid but different flight of these birds. The two groups share a unique crossover structure of a superficial neck muscle which cannot be related to their ways of life. The relationship of the swifts and hummingbirds to other birds is obscure.

Classification. The order Apodiformes is divided into the suborder Apodi, containing the families Aegialornithidae (fossils), Hemiprocnidae (crested swifts; 4 species), and Apodidae (true swifts; 83 species), and the suborder Trochili, containing the single family Trochilidae (hummingbirds; 341 species). The Aegialornithidae are primitive swifts from the Early Eocene of Europe. Otherwise the fossil record of swifts and hummingbirds is very poor.

Characteristics. Swifts are fast-flying, aerial birds with dull, hard plumage; long, curved, pointed wings; and a short, broad, weak bill with a wide gape, adapted to catching insects in flight. They rest by clinging to cliffs, hollow trees, and other vertical surfaces. Their nest is composed of sticks and other plant material glued together and to vertical surfaces, with the extreme condition being a nest built completely of their gluelike mucus. The clutch varies from one to six eggs, and the young are cared for by both sexes until they fly from the nest. Swifts are found worldwide except at high latitudes. True swifts (Apodidae) never perch crosswise on branches, but crested swifts (Hemiprocnidae), found in tropical Asia to New Guinea, are able to perch on branches.

The hummingbirds (Trochilidae) are small, brightly colored, nectar-feeding birds found only in the Western Hemisphere, with most species living south of the United States–Mexican border. The bill is slender and varies in length and shape, correlated closely with the shape of the flowers utilized by each species. They have a rapid wing beat and flight, and are able to hover in front of a flower while feeding or even fly backward (see **illustration**). Hummingbirds are attracted to the color red, and flowers specialized on hummingbirds for cross-pollination are



Anna's hummingbird (*Calypte anna*). (*Photo by Dr. Lloyd Glenn Ingles; © 1999 California Academy of Sciences*)

red. Hummingbirds feed on nectar and insects, and can aggressively defend a feeding territory. They are among the smallest birds, including the bee hummingbird (*Lisugsa belenae*) of Cuba, the smallest of all birds. The diminutive size of hummingbirds results in large loss of heat because of their small body mass compared to surface area. If they are in poor physiological condition, hummingbirds may hibernate overnight to save energy. Moreover, high-altitude species place their nests under an overhanging tree branch or in a cave to reduce radiational cooling during the night. The brilliant colors of the males serve both for courtship, including species recognition, and for defense of their feeding territories. Females of a few species also possess bright colors when they defend feeding territories. Males have a courtship flight and sometimes a "song," but are not involved in nest-building, incubation, or feeding the young. The deep cup nest is built on a branch and consists of plant down and spider web. The clutch is usually two eggs, and the young are cared for by the female until they can fly from the nest.

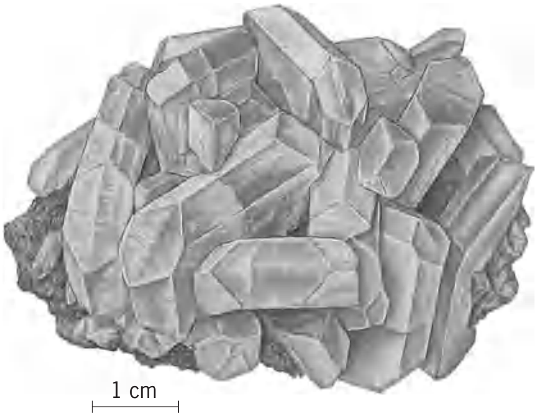
Economic importance. A few species of cave swiftlets (such as *Collocalia brevirostris*) are economically important in providing nests that serve as the base for "birds-nest soup." Prior to the early decades of the twentieth century, hummingbirds were collected extensively to use as decoration on women's hats, a practice, long over, that severely affected the status of many species. Hummingbirds are important in ecotourism and in backyard bird watching, particularly in the western United States, where many people set out hummingbird feeders. *See* AVES.

Walter J. Bock

Bibliography. P. Chantler, *A Guide to the Swifts and Treeswifts of the World*, 2d ed., 1999; K.-L. Schuchmann, Family Trochilidae (Hummingbirds), pp. 468–535, in J. Del Hoyo et al. (eds.), *Handbook of the Birds of the World*, vol. 5, Lynx Edicions, 1999.

Apophyllite

A hydrous calcium potassium silicate containing fluorine. The composition is variable but approximates to $KFCa_4(Si_2O_5)_4 \cdot 8H_2O$. It resembles the zeolites, with which it is sometimes classified, but differs from most zeolites in having no aluminum. It exfoliates (swells) when heated, losing water, and is named from this characteristic; the water can be reabsorbed. The mineral decomposes in hydrochloric acid, with separation of silica. It is essentially white, with a vitreous luster, but may show shades of green, yellow, or red. The symmetry is tetragonal and the crystal structure contains sheets of linked SiO_4 groups, and this accounts for the perfect basal cleavage of the mineral (see **illus.**). It occurs as a secondary mineral in cavities in basic igneous rocks, commonly in association with zeolites. The specific gravity of apophyllite is about 2.3–2.4, the hardness is 4.5–5 on Mohs scale, the mean refractive index is about 1.535, and the birefringence is 0.002. *See* SILICATE MINERALS; ZEOLITE.



Apophyllite crystals with basalt from French Creek, Pennsylvania. (*From Department of Geology, Bryn Mawr College*)

George W. Brindley

Apoptosis

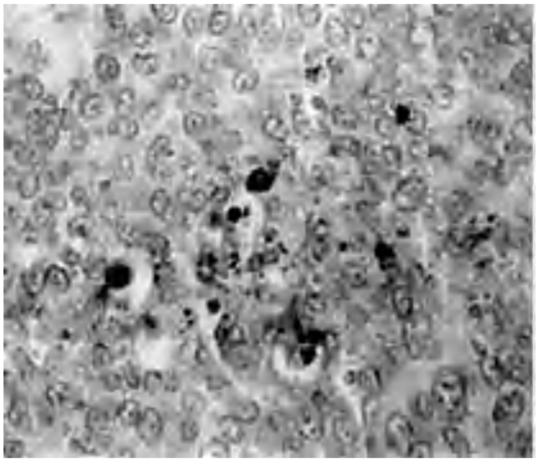
Cell death triggered by extracellular signals or genetically determined events and carried out by physiological processes within the cell. It is therefore sometimes called "programmed cell death" as opposed to uncontrolled cell death following acute injury. Apoptosis is characterized by a systematic breakdown of cellular structure, including cytoplasmic shrinkage, breakdown of chromatin and structural proteins of the cell nucleus, and blistering (blebbing) of the cell membrane.

Cell death is a fundamental concept in biology and medicine, and it is characteristic of many diseases. Cell death can take place through the mechanisms of apoptosis or necrosis. While necrotic cell death has been known for a long time, not until the 1970s was apoptotic cell death clearly recognized. Apoptosis requires energy, and it can be regulated by the cell itself. Necrosis is passive cell death, and it is usually caused by an outside factor, such as loss of blood supply. In necrosis, a large group of cells die simultaneously in the same area, and the process is often associated with inflammation and damage to the surrounding tissues. In apoptosis, individual cells die, and the fragmented cell is removed by surrounding healthy housekeeping cells, the macrophages, making the process neat; no harm is done to the surrounding tissues. *See* CELL (BIOLOGY); CELL NUCLEUS; INFLAMMATION; NUCLEOSOME.

Characteristics. Apoptotic cell death generally takes 12–24 h. It is divided into three phases: initiation, effector, and degradation. In the initiation phase, the apoptotic stimulus is introduced to the cell; depending on the strength and the nature of the stimulus, this phase may take several hours. In the effector phase, the apoptotic stimulus is modulated in the cell; this phase is still reversible (the cell can be salvaged). In the degradation phase, the cell is inevitably doomed to die; degradation of cellular proteins and fragmentation of nucleic acids in the cell (see **illustration**) characterize this phase, which normally takes 1–2 h.

Through apoptosis the organism can regulate the number of cells in a given tissue or destroy any unwanted or damaged cells without harming adjacent tissues. Apoptosis is already present at the embryonic stage, where it participates in the formation of different organs and tissues. Human digits, for instance, are formed because cells between the digit anlages die through apoptosis during embryonic development. Thus, apoptosis is the sculptor, which by deleting unnecessary cells gives rise to the appearance of organisms.

Apoptosis participates in the development of the human immune system. The immune system consists of white cells which attack and destroy foreign elements introduced into the human body. Early in development, the defense system is nonselective and consists of white blood cells which are not able to distinguish foreign molecules from self-derived molecules. When the defense system matures, white cells that cannot make that distinction are destroyed in the thymus through apoptosis. In adult tissues, a restricted loss of such a self-tolerance may sometimes take place (that is, harmful, self-directed white cells develop and fail to undergo apoptosis); the individual then suffers from an autoimmune disease. In such a disease, these white blood cells attack some of the host's own tissue components and induce apoptosis. Apoptosis thus functions to protect the host by participating in the development of the immune system, but it may also contribute to the development of an autoimmune disease.



Several apoptotic cells and fragmented apoptotic bodies in cancer cells (center area of the figure) located by the TUNEL method, which detects apoptotically fragmented nucleic acids in tissue sections.

Activation and inhibition. Apoptosis can be triggered by several factors, including irradiation, heat, several chemical drugs, and toxic compounds. Some of these factors may cause genetic alterations. If such alterations occur, the cells are often capable of correcting such damage, but if this fails, an apoptotic cell death program is stimulated. Apoptosis can also be induced through stimulation or inhibition of certain types of cell surface receptor molecules. A receptor serves as a sensor of the outside milieu of the cell. Hormone-dependent cells, such as breast or prostate cells, have hormonal receptors on their surface; the well-being of these cells depends on a constant stimulation of the receptors. If the receptors are blocked, the cells suffer an apoptotic death. For example, when receptors fail to respond to certain growth-promoting molecules (such as interleukins), apoptosis is initiated. Some cells, such as white blood cells, harbor ligand molecules, which are able to stimulate receptors that induce apoptosis. Inside the cell, two families of proteins, the bcl-2's and the caspases, regulate apoptosis. The bcl-2 family proteins operate in the effector phase of apoptosis and may either promote or inhibit apoptosis. A balance between individual bcl-2 family proteins determines whether the cell will live or die. Caspases are enzymes which operate at the degradation phase of apoptosis. Their activation leads to degradation of cellular proteins and fragmentation of cellular nucleic acids. Recently, other groups of proteins involved in apoptosis have been discovered. Inhibitor of apoptosis proteins (IAPs) are capable of inhibiting caspases. There are also proteins, Fas-associated death-domain-like interleukin-1 (IL-1) beta converting enzyme inhibitory proteins (FLIPs), which inhibit receptor-mediated apoptosis early in the signaling cascade. *See* TUMOR SUPPRESSOR GENE.

Pathology. Apoptosis takes part in the development of several diseases. These include viral and other infections, autoimmune diseases, many neurodegenerative diseases such as Alzheimer's and Parkinson's diseases, and several metabolic and endocrinologic diseases. In these diseases, apoptosis is usually increased, leading to a loss of cells. In Alzheimer's disease, loss of nerve cells in the brain leads to loss of brain matter and deterioration of the intellectual function. Alzheimer's disease is characterized by a gradual accumulation of an aberrant protein in the neuronal cells called amyloid. Genes which produce this protein also influence apoptosis, and it has been suggested that dysregulation of apoptosis plays a primary role in causing the neuronal loss seen in Alzheimer's disease.

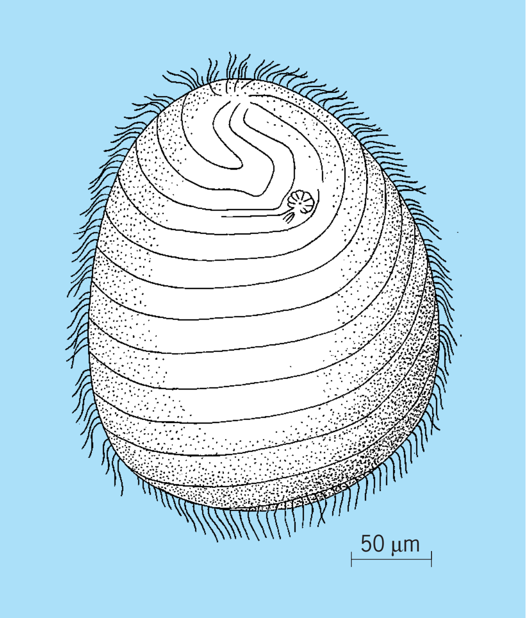
Apoptosis is also operating in vascular diseases such as ischemic heart disease. This disease is characterized by formation of plaques in the vessel walls, eventually leading to obstruction of the blood flow. A great amount of apoptosis has been shown in cells within such plaques. It has been suggested that apoptosis may contribute to the softening of plaques through deletion of cells capable of synthesizing hard collagenous tissue; this softening makes plaques more prone to rupture. A plaque rupture precipitates the formation of an obstructive clot in the blood vessel which leads to an acute heart infarct. When this happens, a part of the heart muscle nourished by the obstructed vessel is destroyed by necrosis. However, some cells at the marginal zone of the infarcted heart muscle undergo apoptosis. By different treatment modalities at an early stage of an acute heart infarct, one may diminish the size of the damaged tissue and save cells from death because at an early stage of apoptosis cells can still be revived and brought back to normal. Apoptosis also plays a role in the further development of ischemic heart disease. As a consequence of heart infarct, some people develop heart failure; one reason is that the remaining heart cells, which are under continuous strain, slowly undergo apoptosis. This apoptosis gradually leads to a decrease in the number of functioning muscle cells in the wall of the heart and development of a failing heart. Decreased apoptosis may also cause disease. For example, some types of lymphomas arise due to deranged apoptotic machinery and overproduction of a protein, which lengthens cell survival. In many other types of cancer, apoptosis is increased. The apoptotic rate in different cancer types may vary, but usually about 1–2% of cancer cells show apoptosis. Cancers arise, in part, because even though cell death in cancer tissues is high, the rate of cell division is higher. One method to treat cancer would be to increase the proportion of apoptotic cells. In fact, treatments directed at increasing apoptosis are already in use. Both irradiation of tumors and cancer chemotherapy lead to apoptosis of cancer cells, and thus the tumors grow more slowly. In hormone-dependent tumors (such as breast cancer), cancer cells are deprived of hormonal stimulus by antiestrogenic treatment, resulting in their apoptosis. Unfortunately, cancer cell populations are heterogeneous, and such treatment destroys only a part of them. The remaining cells are usually resistant to the treatment and must be dealt with in some other way. More selective treatment modalities, including transfection of viral vectors expressing apoptotic genes, are under development. It is plausible that through increased knowledge of apoptosis a cure for some types of cancer will become possible in the near future. In addition, researchers will be able to develop better treatments for other diseases where apoptosis plays a role. *See* ALZHEIMER'S DISEASE; CANCER (MEDICINE); CELL (BIOLOGY); DEVELOPMENTAL BIOLOGY; GENETICS; TUMOR.

Ylermi Soini

Bibliography. J. M. Adams and S. Cory, The Bcl-2 protein family: Arbiters of cell survival, *Science*, 281:1322–1326, 1998; B. Alberts et al., *Molecular Biology of the Cell*, 4th ed., Garland Publishing, 2002; A. Haunstetter and S. Izumo, Apoptosis: Basic mechanisms and implications for cardiovascular disease, *Circ. Res.*, 82:1111–1129, 1998; J. Jenkins and E. P. Reddy, Reviews: Apoptosis, *Oncogene*, 17:3203–3399, 1998; J. F. Kerr, A. H. Wyllie, and A. R. Currie, Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics. *Brit. J. Cancer*, 26(4):239–257, 1972; D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 4th ed., Worth, 2004; N. A. Thornberry and Y. Lazebnik, Caspases: Enemies within, *Science*, 281:1312–1316, 1998.

Apostomatida

A little-studied group of ciliates comprising an order of the Holotrichia. The majority occur as commensals on marine crustaceans. This association is generally more intimate and complex than that of the chonotrichs found on related hosts. Their life histories may become exceedingly complicated, and they appear to bear a direct relationship to the molting cycles of their hosts. Apostomes are particularly characterized by the presence of a unique rosette in the vicinity of an inconspicuous mouth opening and the possession of only a small number of ciliary rows wound around the body in a spiral fashion. *Foettingeria* (see **illus.**) is an example commonly encountered. *See* CILIOPHORA; HOLOTRICHIA; PROTOZOA.



***Foettingeria*, an example of an apostomatid.**

John O. Corliss

Appendicitis

An inflammation of the vermiform appendix. Acute appendicitis is the most common cause of emergency abdominal surgery, occurring in 5–6% of the population of the United States. Although about 80% of the individuals are between 5 and 35 years of age, the disease does occur in the very young as well as in the elderly. In the latter groups, the clinical picture is atypical, and frequently results in a significant delay in diagnosis with a higher incidence of complications. Acute appendicitis is rare in children less than 1 year of age. *See* APPENDIX (ANATOMY).

Cause. Acute appendicitis develops when the lumen of the appendix becomes obstructed, usually by fecal material, a foreign body, or hyperplasia of lymphatic tissue that is normally present in the wall of the appendix. The obstructed appendix becomes distended because of continued secretion of mucus by the lining cells. Distention eventually interferes with the blood supply of the appendix, and reduced blood supply leads to the development of areas of cell death with eventual gangrene and perforation. There is also a proliferation of bacteria within the obstructed appendix, a factor that contributes to infection as gangrene develops.

Typically, acute appendicitis progresses from obstruction of the lumen and distention of the appendix to spread of the inflammation beyond the appendix. Initially, there is localized peritonitis confined to the area of the appendix. If unrecognized and untreated, this may progress to an inflammatory mass or abscess, or to perforation of the appendix with resultant diffuse peritonitis, generalized toxic reaction, and even death. Each year, as many as 10,000 deaths in the United States are attributed to appendicitis and its complications, although it is a much rarer cause of significant morbidity and mortality since the introduction of antibiotics. *See* PERITONITIS.

The usual progression of symptoms includes pain in the region around the navel; loss of appetite, nausea, and occasionally vomiting; localization of the pain to the right lower quadrant of the abdomen; and mild fever. Although the pain typically is localized in the right lower quadrant of the abdomen, there are variations because the appendix may be located in a number of other positions within the abdominal cavity. Fever is a fairly late sign, with mild elevation the rule; a high fever increases the suspicion of perforation or of some other inflammatory process.

Diagnosis and symptoms. The diagnosis of appendicitis is generally made by history and physical examination, although laboratory and radiologic studies may be helpful in differentiating appendicitis from other conditions. The condition must be considered in any case of abdominal pain, especially involving the lower abdomen. Appendicitis is suspected when there are signs of peritoneal inflammation in the right lower quadrant of the abdomen, specifically tenderness on examination by touch and diminished bowel sounds. Pain may also accompany examination of the rectum or the pelvis. With progression of the disease beyond 12–18 h, the incidence of perforation along with secondary complications of infection and a generalized toxic reaction increases significantly. To avoid this crisis, early surgery is essential.

Treatment. The treatment of acute appendicitis is prompt surgical removal of the inflamed appendix. Prior to surgery, the patient may be given intravenous fluids to correct dehydration and electrolyte imbalances. The use of antibiotics before surgery to decrease wound infection is often recommended. Antibiotics are continued after surgery in cases where the inflammation has extended beyond the appendix. Delay in removal of the appendix increases the chance of perforation. Death following appendectomy for acute appendicitis is less than 0.5% overall, but does increase to about 15% in the elderly. This increase is related to delay in diagnosis, associated cardiopulmonary complications, and the frequent presence of unrelated major diseases. The incidence of complications in appendicitis is 10–15%, but increases to 30–40% when perforation is present. The most common complication of appendectomy for acute appendicitis is infection. To reduce the risk of such complications, prompt diagnosis and surgery are necessary. *See* ANTIBIOTIC; GASTROINTESTINAL TRACT DISORDERS. Anthony L. Imbembo

Bibliography. D. Law, R. Law, and B. Eiseman, The continuing challenge of acute and perforated appendicitis, *Amer. J. Surg.*, 131:533, 1976; F. R. Lewis et al., Appendicitis: A critical review of diagnosis and treatment in 1000 cases, *Arch. Surg.*, 110:677, 1975.

Appendicularia (Larvacea)

A class of marine, planktonic animals in the subphylum Tunicata. This class is characterized by the persistence of a tail, notochord, gill slits, and dorsal nerve cord throughout life and by a unique feeding structure, the "house." Appendicularians are free-swimming, solitary animals that are believed to have evolved through neoteny from the tadpole larvae of a bottom-dwelling, ascidianlike, ancestral tunicate. They resemble a bent tadpole with a flat, muscular tail and a trunk containing all major organs. Included are a complex digestive system, reproductive organs, two ciliated openings of the gill slits (the spiracles) leading to a mucus-lined pharynx, a mucus-producing gland called the endostyle, and a simple circulatory system with a single, muscular heart (**Fig. 1**). *See* NEOTENY.

All appendicularians, except the species *Oikopleura dioica*, are hermaphroditic; that is, each individual bears the reproductive organs of both sexes. While sperm may be liberated from the testes via tiny ducts to the outside, release of eggs from the ovary occurs only when the trunk splits open, resulting in the death of the animal. Each individual produces from about 50 to several hundred eggs, which are fertilized externally and develop into free-swimming tadpole larvae that undergo metamorphosis without settling. Life-spans range from one to several weeks.

The epidermal cells on the trunks of all appendicularians secrete a gelatinous tunic made of mucopolysaccharides. This tunic is then expanded into a nonliving, balloonlike structure, the house, used in feeding. Of the three families of appendicularians, the Oikopleuridae have the most complex houses. The animal resides within the house and forces water, by means of the muscular beating of its tail, through two incurrent filters on the external surface of the house (**Fig. 2***a*). These filters contain a fibrous mesh which removes large particles, particularly large diatoms and other algae. Water laden with smaller particles then passes through a complex internal filter, where small phytoplankton and bacteria become further concentrated. The ciliated spiracles on the animal's trunk draw these food particles in a slurry up a hollow mucous tube to the animal's mouth.

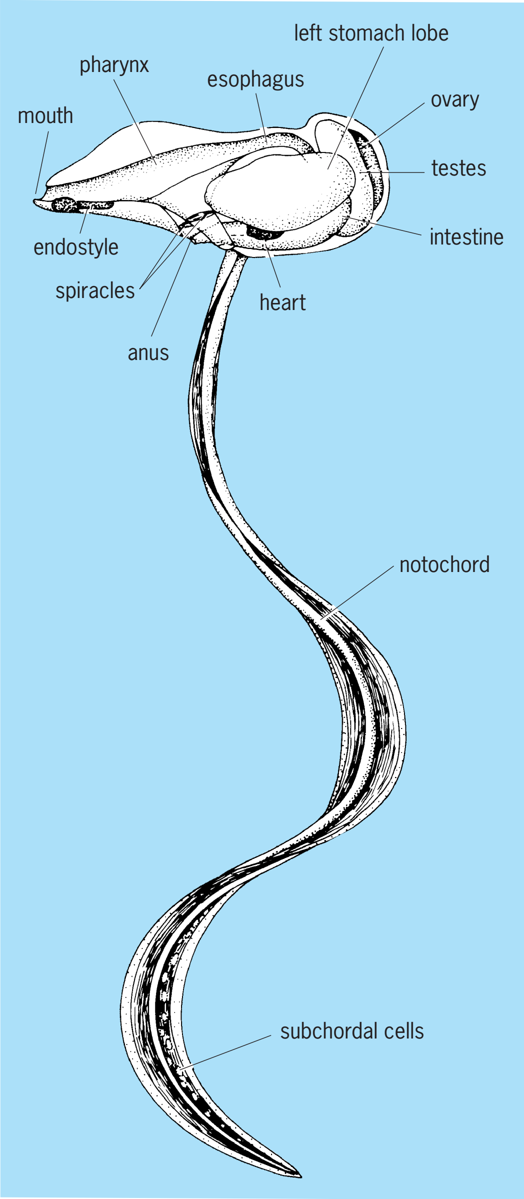


Fig. 1. Anatomy of *Oikopleura albicans*, a typical appendicularian. (*After A. Alldredge, Appendicularians, Sci. Amer., 235:94–102, 1976*)

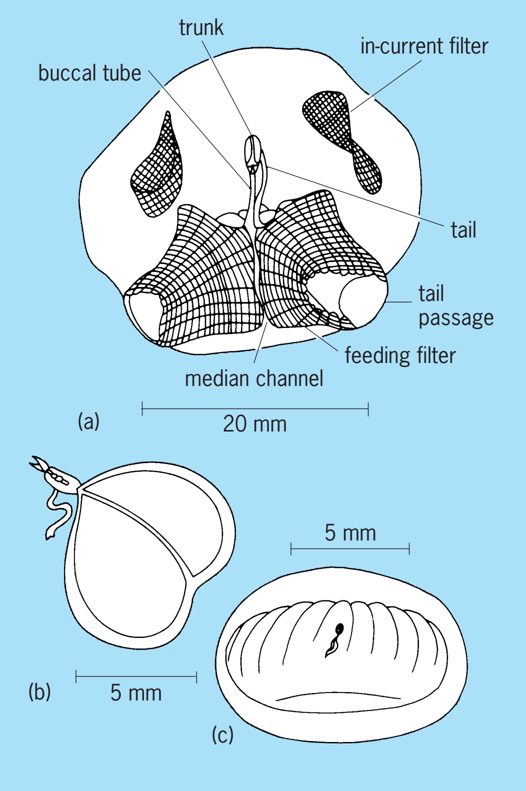


Fig. 2. Houses of the three appendicularian families: (*a*) Oikopleuridae, (*b*) Fritillaridae, and (*c*) Kowalevskiidae.

The house becomes readily clogged with the larger nonfood particles, and appendicularians may abandon the old house and expand a new one as often as six times a day. The house built by the family Kowalevskiidae is a much simpler umbrellalike structure lacking complex filters (Fig. 2*c*). Members of the family Fritillaridae deploy the house as a large, collapsible bubble in front of the mouth (Fig. 2*b*).

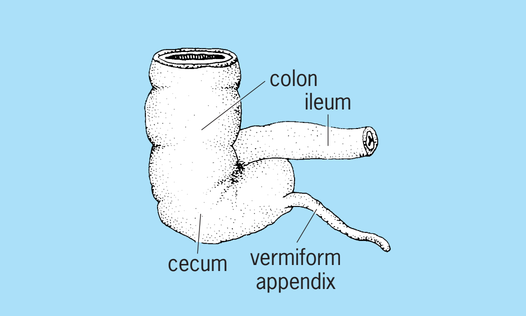
Appendicularians feed primarily on small particles from 0.1 to 30 micrometers in diameter. Larger individuals may filter up to several thousand milliliters of seawater per day. As one of the few metazoan groups capable of capturing bacteria and tiny phytoplankton, appendicularians are important marine grazers which may significantly reduce phytoplankton populations in the ocean. About 13 genera and 70 species of appendicularians are known. They are found in all oceans of the world and are a common component of plankton samples, particularly in coastal waters. The most abundant genera are *Oikopleura* and *Fritillaria*. *See* CHORDATA; TUNICATA.

Alice L. Alldredge

Bibliography. A. L. Alldredge, Appendicularians, *Sci. Amer.*, 235:94–102, 1976; A. L. Alldredge and L. P. Madin, Pelagic tunicates: Unique herbivores in the marine plankton, *Bioscience*, 32:655–663, 1982; R. D. Barnes, *Invertebrate Zoology* 6th ed., 1994.

Appendix (anatomy)

A narrow, elongated tube closed at one end, extending from the cecum, a blind pocket off the first part of the large intestine (see **illus.**). It is found in only a few mammals.The size and detailed structure of the appendix vary markedly, depending on the species and the age of the individual. Most reptiles, birds, and mammals have either a single or a paired cecum at the anterior end of the large intestine, but it is quite rare that this cecum has a thinner projection or true appendix.



Junction of human ileum and colon, showing cecum and vermiform appendix. (*From C. K. Weichert, Anatomy of the Chordates, 4th ed., McGraw-Hill, 1970*)

In humans the appendix is about 3 in. (7.5 cm) long, but it varies greatly in both size and its specific location in the lower right quarter of the abdomen. Rabbits and their relatives have a very large cecum with an elongate and rather thick appendix, which may reach a length of several inches. A few rodents, such as some of the Old World porcupines, also possess an appendix. The pyloric ceca at the anterior end of the intestine of many fishes are sometimes called appendixes, but they are quite different structures.

The inner lining (mucosa) of the appendix is continuous with the remainder of the digestive lining. In young humans the glandular epithelium is surrounded by clusters of lymphoid follicles which tend to regress with age. The next layer is a smooth muscle coat which, in turn, is surrounded by a connective tissue coat, the outermost part being covered with visceral peritoneum (serosa).

The exact function of the human appendix is unknown, and it is considered to be a remnant of a portion of the digestive tract which was once more functional and is now in the process of evolutionary regression. The appendixes presumably function in digestion in forms with larger ones. *See* APPENDICITIS.

Thomas S. Parsons

Apple

Apples (genus *Malus*) belong to the family Rosaceae (see **illus.**), which includes many other deciduous fruits such as pears, peaches, and cherries. There are about 30 *Malus* species in the North Temperate Zone. The fruits of most species are edible. Although selections of the Asiatic species *M. prunifolia* have been cultivated for their fruits for centuries in China and Japan, they have now been largely replaced by varieties of the "European" cultivated apple. *See* FRUIT; ROSALES.

More apples are consumed than any other temperate-zone tree fruit. Apples are eaten fresh, processed into jellies or preserves, cooked in pies and pastries, or made into sauces. Apple juice is drunk fresh, or at various stages of fermentation as cider, applejack, or brandy. Apple cider vinegar is popular for use in salads and in many processed foods.

Origin and breeding. The "European" cultivated apple is now thought to have been derived principally from *M. pumila*, a Eurasian species which occurs naturally from the Balkans eastward to the Tien Shan of central Asia. In the wild, some forms of *M. pumila* approach present cultivars in size and quality. Another Asian species, *M. sylvestris*, whose range extends into western Europe, grows side by side with *M. pumila* and hybridizes with it in the Caucasus Mountains. Thus *M. sylvestris* probably also had some genetic input into the cultivated apple.

Wild apples, mainly the edible *M. baccata*, grow so thickly east of Lake Baikal in Siberia that the region is called Yablonovy Khrebet ("Apple Mountains").

The success of the Delicious and Golden Delicious cultivars may be laid to the demand for better-quality fresh fruit. Both originated on farms as chance seedlings near the end of the nineteenth century. In spite of the dominance of such "chance" cultivars, cultivars such as Cortland and Idared that were produced by scientific breeding have begun to achieve prominence. Apple crosses are made by first removing the stamens from the flowers of the female parent and then applying previously gathered pollen of the selected male parent to the receptive pistil. After fruits form and ripen, seed is removed and germinated.

Most of the apple breeding programs under way earlier this century have ceased, but the few remaining ones are now introducing some exciting new cultivars. Many of the new cultivars are of

Apple varieties: (*a*) Priscilla and (*b*) Silverspur. (*Bountiful Ridge Nurseries*)

excellent quality and in addition are resistant to the most damaging diseases, such as scab, rust, and fire blight. These cultivars promise to revolutionize apply growing in the future. *See* BREEDING (PLANT).

Distribution. The apple is probably the most widely distributed fruit crop in the world, although it ranks behind grapes, bananas, and oranges in total production. There are substantial apple-growing areas on all temperate-zone continents; the United States, Italy, France, and Germany are leading producers.

Apples can be grown as far north as 60° latitude in the maritime climate of northwestern Europe. In North America, apple culture does not extend much above 50° north latitude. Away from the coasts, the buffering effects of the Great Lakes on temperature extremes facilitate heavy apple production in New York, Michigan, and Ontario. Hardier cultivars are continuing to be developed for use in colder regions. Apples can be grown in the tropics at higher elevations where sufficient chilling to break dormancy is available. Cultivars with lower chilling needs are being developed.

The principal apple-growing regions in North America are, in order of importance: the Northwest (Washington, Oregon, Idaho, and British Columbia), the Northeast (New York, New Jersey, New England, Ontario, Quebec, and Nova Scotia), the Cumberland-Shenandoah area (Pennsylvania, Maryland, Virginia, West Virginia, and North Carolina), Michigan, California, the Ohio Basin (Ohio, Indiana, and Illinois), and Colorado.

Planting systems and rootstocks. Historically apples have been grown in America as large trees on seedling rootstocks, widely spaced, with a density of 25–40 per acre (62–100 per hectare). This system has many disadvantages: the trees take many years to reach full production, they are difficult to prune and harvest because ladders are needed, and the fruit is often poorly colored and small in size.

Dwarfing rootstocks have been used by European growers for centuries to produce smaller trees that bear earlier, and produce better-quality fruit. Earlier in this century the various clonal rootstocks were cataloged into size categories (the M. series) according to the numbering system of the East Malling Research Station, England. New clonal rootstocks (the MM. series) have been developed at East Malling by breeding. There are now rootstock-breeding programs under way in several European countries and at the New York State Agricultural Experiment Station at Geneva.

In America, most apple trees are planted on clonal rootstocks which produce semidwarf trees at a density of 100–200/acre (250–500/ha). There are also some plantings of full-dwarf trees (400–800/acre or 1000–2000/ha), sometimes grown as espaliers on a trellis. These plantings are much more productive on an area basis than the older ones. Although the acreage of apple orchards has declined drastically in this century, the number of trees and the production of apples have increased. *See* PLANT PROPAGATION.

Orchardists in many areas are using sod culture with herbicide treatments in the row to prevent competition from weeds for moisture and nutrients. Nutrients are applied by broadcasting granular fertilizers or by spraying solutions on the soil near the trunk of trees. In special situations, urea in solution may be applied to foliage.

Apples are usually irrigated in western North American growing areas. The use of trickle irrigation rather than flood, furrow, or sprinkler irrigation has increased. In trickle irrigation, small amounts of water are applied over a long period of time and as needed by the trees. Fertilizer can be applied by trickle irrigation.

In certain areas, particularly in western North America, spring frosts can damage blossoms. To reduce the danger of damage, several techniques are available. Orchards can be heated with smudge pots which burn waste oil, or with gas heaters, or by burning "bricks" of petroleum. The higher prices of petroleum-based fuel and more strict pollution control regulations are making this method less popular. Large fans ("wind machines") can be used in orchards to break inversion layers by mixing warm air posed above the trees with the cold air near the ground. When trees are sprinkled in the late dormant period, evaporative cooling results in delay of bloom beyond the frost danger period. Sprinkling can also be used during periods of freezing temperatures to prevent frost injury directly. The latent heat of fusion liberated when the sprinkled water freezes provides sufficient heat to prevent the blossoms from freezing.

Pruning. To maintain the correct balance between wood, vegetative growth, and fruiting spurs, apple trees must be pruned regularly. On most cultivars, blossoms, and later fruits, are formed on spurs produced in previous years. Pruning seeks to maximize spur formation on a strong tree structure into which is permitted the penetration of adequate light for coloring and ripening. Traditionally, pruning has been carried out in the dormant season, but in recent years a system of pruning during the summer has been developed that requires less work and gives better results. Manual, hydraulic, or pneumatic pruning tools are used. Pruners gain access to trees by climbing, with the aid of ladders and self-propelled mechanical platforms.

Pollination. Most apple cultivars are self-incompatible and require pollen from a second cultivar for fertilization and, therefore, fruit formation to occur. The pollen is transferred from tree to tree by insects, usually bees. Fruit growers often plant two cultivars that bloom at the same time in close proximity, so that each will pollinate the other effectively. Triploid cultivars, such as Idared, must be planted in close proximity with two other cultivars since the triploid's pollen is infertile. In many areas natural populations of bees are insufficient to provide adequate pollination, and hives are placed in the orchard during bloom. *See* POLLINATION.

Pest control. Apples are attacked by viruses, bacteria, mycoplasmas, fungi, nematodes, mites, insects, and mammals. In most apple-growing regions precautions must be taken against all of these pests, though to different degrees. In practice, damage caused by fungi and insects is most important. Much protection can be afforded by growing cultivars that are resistant to one or more of the pests prevalent in the area. But a limited number of cultivars resistant to only a few pests are now available. Organisms that are spread solely in propagating material, like some viruses, can be controlled by maintaining virus-free nurseries. Some pests can be reduced by cultural practices that interfere with their life cycle. Thus, close mowing limits weed growth, and removing dropped fruits, which harbor pests, may reduce subsequent damage from those pests.

However, good pest control on the present commercial apple cultivars usually requires the use of chemical pesticides. Chemicals for use on apples are stringently tested for efficacy and for safety to orchard workers, consumers, and the environment. Several compounds previously used have been discontinued for safety reasons. Modern pesticides, properly used, provide excellent control of most pests without damaging the trees and fruit. However, many pests have developed tolerance to some of the best pesticides, rendering them ineffective. In an effort to forestall the development of resistance, several different chemicals are used; pests are unlikely to develop tolerance to all of them. Since the development, testing, and registration of new pesticides have become such a lengthy and expensive process, the future of chemical control is uncertain. *See* PESTICIDE.

Three approaches that have become more important for pest control are integrated pest management (IPM), the use of resistant cultivars, and the development of biological controls. In integrated pest management all available relevant methods of control are integrated in such a way that each method supplements and does not interfere with any other. Thus chemicals would be used only when pests are predicted to reach sufficient levels to result in economic losses. Cultural practices would be used that result in reduced damage from pests by encouraging the development of more resistant plants and by favoring the action of predators and parasites that are active against apple pests.

Considerable progress has been achieved on biologically based controls for apple pests. Biological control in the strict sense involves the use of predators, parasites, or pathogens to reduce pest populations and activity, thus causing a reduction in damage by the pest species of interest. Generally, biological controls are highly specific to the target organism and have little effect on other organisms in the vicinity. Once established, biological controls are likely to be long-lasting because of the basic compatibility of the agent with the pest. *See* INSECT CONTROL, BIOLOGICAL.

In many areas, the control of the European red mite is accomplished through the activity of predacious mites. Populations of the predacious mites often increase in orchards to levels that result in suppression of the apple-damaging red mite. Highly specific insecticides that do not affect predacious mites have been developed for use in orchards. Occasionally, however, certain insecticides must be used to control other apple-damaging insects, and these also affect the predacious mites, thus disrupting the biological control process of the red mites. *See* INSECTICIDE.

Great advances have been made in understanding insect pheromones, chemical substances that are produced in minute quantities by female insects and that males can sense over great distances. Several pheromones have been identified and synthesized chemically. Synthetic pheromones are used as attractants in traps to determine insect populations. If large amounts of pheromones are released in orchards, sexual activity of insects is disrupted and reproduction is reduced. *See* PHEROMONE.

As far as diseases of apple are concerned, only a limited number of biological control systems are in use. The crown gall disease caused by a bacterium is being controlled in some nurseries by treating nursery stock with suspensions of a nonpathogenic bacterium that produces an antibiotic that inhibits the pathogenic bacterium. *See* CROWN GALL.

Absolute control of a pest can be achieved by growing a cultivar that is resistant to the pest. Even partial resistance may be sufficient in areas and years when pest pressure is low. The amount of pesticide needed for pest control on partially resistant cultivars is less than that required on more susceptible cultivars.

There are several projects in the United States and other countries to breed new apple cultivars that are more resistant to pests, particularly diseases. Some scab-resistant cultivars, such as Liberty, Nova Easygro, and Priscilla, are already available. Some of these cultivars are also resistant to other diseases such as rusts, fire blight, and powdery mildew. *See* PLANT PATHOLOGY.

Harvesting. Traditionally, apples have been harvested manually by pickers standing on ladders. Two developments are changing this practice: (1) The use of mechanical tree shakers with catching frames and conveyor belts that detach and move fruit to bulk bins or trucks without individual handling. Apples so harvested usually are destined for processing. (2) The use of dwarf trees whose fruit can be picked from the ground by inexperienced personnel. *See* AGRICULTURAL MACHINERY.

Marketing. Apples may be sold fresh immediately, or after a period of storage, or they may be processed into less perishable products such as canned, frozen, or dried slices or chunks for baking, applesauce, apple juice or cider, and vinegar.

Fresh apples may be sold locally at roadside stands or at farmers' markets, or sold in wholesale quantities to supermarkets. Often, large quantities of fruit are traded by cooperatives or independent buyers for sale in distant markets. Thousands of tons of apples are shipped annually from the state of Washington to the Midwest and the East Coast. Apples can be stored for long periods at low temperatures under controlled atmosphere (CA storage), so that fresh fruits are now available year-round.

There is a large international trade in apples, particularly in Europe between the large producing countries of France and Italy and the net-importing countries Germany, Britain, and Scandinavia, and from the United States to Canada. Apples grown in the Southern Hemisphere (Australia, New Zealand, South Africa, and Argentina) are shipped in large quantities to western Europe during the northern winter. *See* FRUIT, TREE.

Herb S. Aldwinckle; Steven V. Beer

Apraxia

An impairment in the performance of voluntary actions despite intact motor power and coordination, sensation and perception, and comprehension. The apraxic person knows the act to be carried out, and has the requisite sensory-motor capacities; yet performance is defective. The abnormality is highlighted when the act must be performed on demand and out of context. The individual may perform normally in such activities as hammering or tooth brushing performed with the object in hand, especially in a natural situation, but will often fail when required to pantomime such acts.

Defects in performance vary from total inability to initiate the action, to incorrect serial ordering of elements, to partial approximations. A common apraxic behavior is the use of a body part as an object. Pantomiming the act of brushing the teeth, for example, a person may run the index finger across the teeth as though it were a toothbrush, while in normal performance, the hand assumes the posture of holding and moving the brush.

Apraxia is usually observed in both upper extremities. When it occurs unilaterally, it is usually the left arm and hand that are affected. This has been explained by assuming that the left cerebral hemisphere is specialized in the organization of voluntary movements, just as it is in language. The left hand is under the immediate control of the right hemisphere, but for skilled voluntary actions, the right hemisphere is dependent on information transmitted from the dominant left hemisphere over the corpus callosum. Callosal lesions produce apraxia of the left hand, because the right hemisphere is incapable of organizing the plan of movement independently. With an appropriately placed left-hemisphere lesion, a bilateral apraxia will result. *See* HEMISPHERIC LATERALITY.

When the left-hemisphere lesion also destroys the primary motor zone, the right arm is paralyzed and the apraxia is masked. The observable apraxia on the left side is referred to as sympathetic apraxia. This is seen in many individuals with right hemiplegia (unilateral paralysis of the body) and Broca's aphasia. Another apraxia often coupled with Broca's aphasia is nonspeech oral apraxia (or buccofacial apraxia). Individuals with this disorder can be observed to struggle to perform such simple acts as protruding the tongue or licking the lips on command or imitation, even though these movements are executed easily as part of the act of eating. *See* APHASIA.

There are several disorders that are controversial with regard to their interpretation as forms of apraxia. The nonfluent speech pattern of Broca's aphasia, often riddled with speech-sound errors, is considered as apraxia of speech by some authorities, while others view it as an integral part of the linguistic deficit of the aphasia. In dressing apraxia and in some types of constructional apraxia, the defect appears to be perceptually based. Limb-kinetic apraxia is widely interpreted today as a mild spastic paresis, while ideational apraxia, commonly associated with dementia, is likely due to conceptual confusion rather than to a disturbance of motor organization. *See* AGNOSIA.

Gerald J. Canter

Bibliography. G. R. Hammond (ed.), *Phenomenology of Will and Action*, 1967; K. M. Heilman and E. Valenstein (eds.), *Clinical Neuropsychology* 3d ed., 1993; P. J. Vinken and G. W. Bruyn (eds.), *Cerebral Control of Speech and Limb Movements*, 1991; R. W. Wertz et al., *Apraxia of Speech in Adults: The Disorder and Its Management*, 1991.

Apricot

The stone fruit *Prunus armeniaca*, thought to be native to China, and then distributed throughout Asia, Europe, and eventually to North and South America and Oceania. The species is genetically diverse and can grow in a wide range of climates depending upon the cultivar. Such diversity occurs in North America, where apricots are produced near Penticton, British Columbia, Canada, in northern New York, as far south as southern California, and even near Puebla, Mexico. Most commercial production in the world is limited to areas where temperatures do not fall below −10 to −20°F (−23 to −29°C) for extended periods; however, certain cultivars can tolerate even severer conditions. Many apricot cultivars can tolerate high summer temperatures in excess of 105°F (40°C). Some cultivars develop an internal browning of the flesh if high temperatures persist with fruit on trees. Apricots tend to bloom earlier than other stone fruit and are sensitive to frost. Frostfree areas are generally preferred.

Horticulture. Flowers from most commercial cultivars are self-fertile, but examples of self-infertility are found in commercial cultivars. In those self-incompatible cultivars, cross-pollination with another cultivar is required. Honeybees are normally used to facilitate cross-pollination. The Japanese apricot (*P. mume*) is a favorite ornamental and fruit tree in China and Japan. In Japan, for example, the *mume* is used in juice and processed and sold as a consumable fruit product. There are small orchards of *mume* in California; it has also been used in hybridization with apricot.

Fruits for commerce are generally yellow to orange in skin color and flesh (**Fig. 1**). The skin surface is generally smooth, with a thin suture line. New cultivars have very smooth skin, with very intense orange skin, sometimes with a red blush and deep orange flesh color. Apricots can also be red, or white with varying shades of yellow or orange; however,

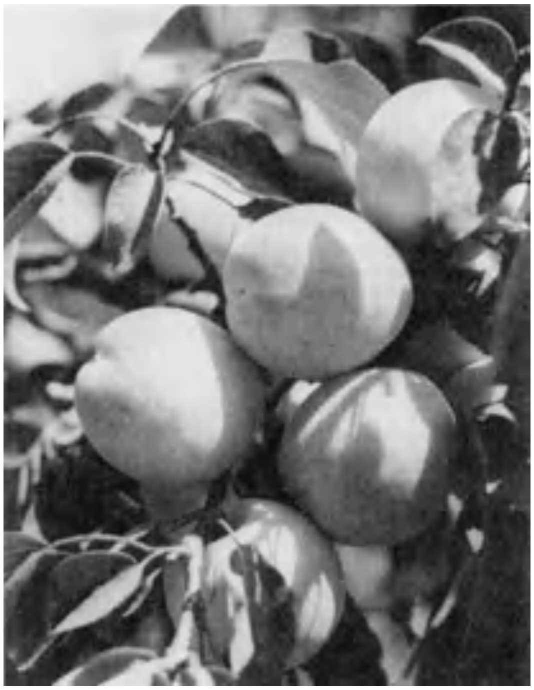


Fig. 1. Castlebrite apricot branch with a fruit cluster.

these are normally not found in commercial outlets. Apricot size ranges from small (about 25–30 g per fruit) to large (100–130 g). The fruit can be consumed fresh, dried, frozen, or processed as canned product, as juice, or as baby food (pureed).

Apricots thrive best in well-drained soils. They tolerate sandy, gravel, and clay soils, but grow best in deep alluvial, sandy loams. Fruit with increased color and taste are probably produced on soils which have high amounts of sand and gravel. Apricots are adapted to soils where the pH is 6.2–7.2. Soils with neutral pH help to minimize micronutrient deficiencies. Apricots require only a modest amount of nitrogen compared with many other fruit trees. Typically 50–100 pounds of actual nitrogen per acre per year are applied to commercial apricot orchards. Too much nitrogen can delay fruit maturity, reduce fruit firmness, and generally reduce fruit quality. Fruits from high-nitrogen fertilized soils tend to be more prone to disease and produce excessive shoot growth with ample water. Additions of potassium, zinc, manganese, and boron are normally required on a regular basis. In some regions, phosphorus- and sulfur-containing fertilizers are applied. Calcium sprays appear to increase fruit firmness.

Production and propagation. Of the total commercial production of apricots in North America, California produces 90–95%, followed by Washington and Utah. Trees are propagated by budding on hardwood cuttings of Marianna 2624 (*P. cerasifera* × *P. musoniana?*) or Myrobalan 29C (*P. cerasifera*) seedlings of apricot, peach (Lovely, Nemaguard, Nemared), and Myrobalan. Peach rootstocks are the most commonly used. Plum roots are used in sites that are poorly drained. Citation (peach × plum) is a vegetatively propagated rootstock which has gained use in commercial apricot plantings. The rootstock is thought to control tree size and improve fruit quality. It is sensitive to crown gall disease. Trees are set from 24 ft (7.3 m) apart, to 8 ft (2.5 m) between rows in the orchard. Orchard undergrowth is either clean-cultivated (free of undergrowth) or controlled by chemical herbicides along tree rows, and the remainder (row middles) are mowed. Planted cover crops or natural vegetation is mowed in middles. Trees are pruned and trained to an open-center framework by summer and dormant pruning. *See* PLANT PROPAGATION.

Harvest. Apricots are harvested mostly by hand, but some machine harvesting is done for processed fruit. Fresh fruit is picked and packed in the orchard or is graded and sorted in sheds specifically designed for fruit packing. The fruit is cooled and shipped in volume fill or tray packs to markets throughout the world. Processed fruit is generally harvested into large bins and then loaded onto trucks and carried to canneries, freezers, or facilities that can process the fruit into juice. Apricots are a very nutritious fruit high in vitamin A and antioxidants lycopene and beta carotene. They are also a good source of vitamin C and soluble fiber.

Diseases. There are several fungal diseases of apricot that require annual control. The brown rot fungus destroys blossoms and twigs (*Monilinia laxa*) in the spring and frequently causes fruit decay (*M. fructicola* and *laxa*) in the summer (**Fig. 2**). The disease can be devastating and can cause reduced yield. Sprays are applied from early in the bloom through full bloom to reduce blossom and twig blight. Typically one to three sprays are applied, and chemicals such as Iprodione and Vinclozolin (Dicarboximides), or Tebuconazole, Propiconazole, or Fenbuconazole (DMI-triazoles) are effective. New fungicides to control brown rot disease are continually under development. Protection of the fruit is achieved by sprays applied from 1 to 3 weeks

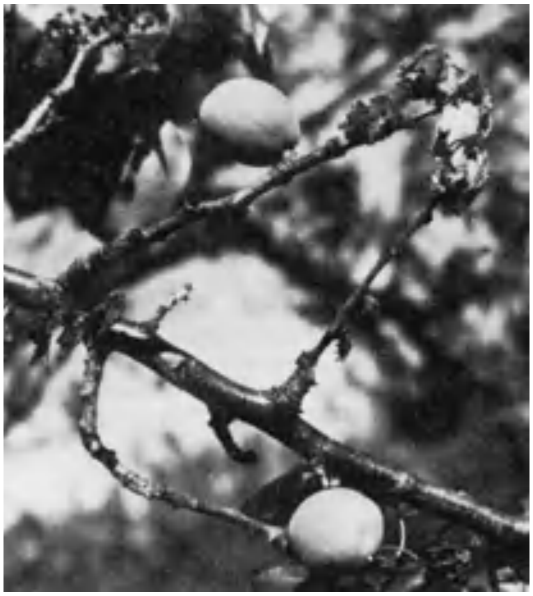


Fig. 2. Spurs and shoots infected with brown rot fungus.

prior to harvest with the aforementioned chemicals. A disease that infects during the flowering and small fruit stage is known as jacket rot, caused by *Botrytis cinerea*. Symptoms occur on small green fruit when remnants of flower parts are still attached to the fruit. The disease causes a brown discoloration on the fruit under the petal remnants. The young fruit may wither and fall within weeks. Control can be achieved with fungicides. Powdery mildew (*Sphaerotheca pannosa* and *Podosphaera tridactyla*), a spring fungal disease, appears as weblike white growth on fruit, leaves, and stems. Fruit may not be packed because of associated lesions on the surface. The fungus *Wilsonomyces carpophilius*, causing shot-hole disease, kills dormant buds, spots fruit and leaves, and, in severe cases, induces defoliation. The fungus survives within infected buds and on twigs. It is controlled by spraying a fungicide at early dormancy (fall) and during blossoming. *See* FUNGISTAT AND FUNGICIDE.

A disease destructive to the framework of trees is caused by *Eutypa lata* (gummosis, cytosporina, limb dieback). This fungus invades fresh pruning wounds when rainfall occurs 2–6 weeks after pruning, causing infections that kill branches and sometimes the entire tree. The greatest incidence of disease occurs during periods when rainfall is highest (fall and winter). Control is based largely on removal and destruction of infected branches, pruning during June, July, August, and September, or no pruning (California). No other treatments have been consistently commercially feasible and reliable.

A lethal root rot of apricot is caused by the fungus *Armillaria mellea*. It can be partially controlled by the use of Marianna 2624 rootstock. Another root and crown rot is caused by *Phytophthora* spp. Both *Armillaria* and *Phytophthora* diseases can kill trees. Site selection to avoid poorly drained soils aids in control of *Phytophthora*. Verticillium wilt (*Verticillium dahliae*) is a soil-borne fungal disease that causes branches and limbs on trees to collapse and die. Symptoms typically appear in young trees less than 6 years old. Careful water management, and tree removal followed by soil fumigation with methyl bromide can aid in the control of *Armillaria*, *Phytophthora*, and *Verticillium* spp.

Bacterial canker caused by *Pseudomonas syringae* is the most important bacterial disease. The disease is found typically in young trees less than 7 years old and is associated with trees growing in stressful conditions, such as cold, high rainfall, poor nutrition, or nematode infestation. Twigs, branches, and sometimes entire trees are killed. Disease severity usually is greatest in trees on sandy soil and in those on apricot and plum rootstocks. Reducing tree stress and budding the cultivar high on the rootstock seem to reduce the incidence of disease. Fumigation with methyl bromide may aid in controlling nematode populations. Bactericidal sprays are relatively ineffective.

A few virus and viruslike diseases of apricot are known. The most common, the plum pox virus (sharka) and chlorotic leaf spot, are found throughout producing regions of Europe. Most of these are controlled by the use of disease-free planting stock. *See* PLANT PATHOLOGY.

Insects. The most common insect pests of apricot are peach twig borer (*Anarsia lineatella*), worms (*Archips argyrospila*, *Orthosia hibisci*, and others), earwigs (*Forficula auricularia*), mites (*Tetranychus* spp.), and aphids. Integrated pest management programs have been developed to control most pests of apricots. Stephen M. Southwick

Bibliography. N. F. Childers, *Modern Fruit Science*, 10th ed., 1995; J. M. Ogawa and H. English, *Diseases of Temperate Zone Tree Fruit and Nut Crops*, 1991; J. M. Ogawa and S. M. Southwick, Introduction: Apricot, in J. M. Ogawa et al. (eds.), *The Compendium of Stone Fruit Diseases*, American Phytopathological Society, 1995; A. Rosati, T. M. DeJong, and S. M. Southwick, Comparison of leaf mineral content, carbon assimilation and stem water potential of two apricot (*Prunus armeniaca*) cultivars grafted on Citation and Marianna 2624 rootstocks, *Acta Hort.*, 451:263–267, 1997; S. M. Southwick and G. Stokes, Apricots, the fruit and its importance, in *The Encyclopedia of Food Science, Food Technology, and Nutrition*, Academic Press, London, 1991; S. M. Southwick and K. G. Weis, Selecting and propagating rootstocks to produce apricots, *HortTechnology*, 8(2):164–170, 1998; S. M. Southwick, J. T. Yeager, and K. G. Weis, Use of gibberellins on Patterson apricot (*Prunus armeniaca*) to reduce hand thinning and improve fruit size and firmness: Effects over three seasons, *J. Hort. Sci.*, 72(4):645–652, 1997; S. M. Southwick, J. T. Yeager, and H. Zhou, Flowering and fruiting in Patterson apricot (*Prunus armeniaca*) in response to postharvest application of gibberellic acid, *Scientia Horticulturae*, 60:267–277, 1995; G. Stokes et al., Apricots and apricot processing, in Y. H. Hui (ed.), *The Encyclopedia of Food Science and Technology*, Wiley, Cutten, CA, 1989.

Apsides

In astronomy, the two points in an elliptical orbit that are closest to, and farthest from, the primary body about which the secondary revolves. In the orbit of a planet or comet about the Sun, the apsides are, respectively, perihelion and aphelion. In the orbit of the Moon, the apsides are called perigee and apogee, while in the orbit of a satellite of Jupiter, these points are referred to as perijove and apojove. The major axis of an elliptic orbit is referred to as the line of apsides. *See* CELESTIAL MECHANICS; ORBITAL MOTION.

Raynor L. Duncombe

Aqua regia

A mixture of one part by volume of concentrated nitric acid and three parts of concentrated hydrochloric acid. Aqua regia was so named by the alchemists because of its ability to dissolve platinum and gold. Either acid alone will not dissolve these noble metals. Although chlorine gas and nitrosyl chloride are formed as indicated in reaction (1), the oxidizing properties of aqua regia are not believed to be increased. Instead, the metal is easier to oxidize because of the high concentration of chloride ions which form a stable complex ion as illustrated with gold in reaction (2).

$$HNO_3 + 3HCl \rightarrow Cl_2 + NOCl + 2H_2O \qquad (1)$$

$$Au + 4H^+ + NO_3^- + 4Cl^- \rightarrow AuCl_4^- + NO + 2H_2O \qquad (2)$$

See CHLORINE; GOLD; NITRIC ACID; PLATINUM.

E. Eugene Weaver

Aquaculture

The cultivation of fresh-water and marine species (the latter type is often referred to as mariculture). The concept of aquaculture is not new, having begun in China about 4000 years ago, and the practice of fish culture there has been handed down through the generations. China probably has the highest concentration of aquaculture operations in a single country today. Yet the science of aquaculture, as a whole, is far behind that of its terrestrial counterpart, agriculture. Aquaculture, though still partly art, is being transformed into a modern, multidisciplinary technology.

Aquacultural ventures occur worldwide. China grows macroalgae (seaweeds) and carp. Japan cultures a wide range of marine organisms, including yellowtail, sea bream, salmonids, tuna, penaeid shrimp, oysters, scallops, abalone, and algae. Russia concentrates on the culture of fish such as sturgeon, salmon, and carp. North America grows catfish, trout, salmon, oysters, and penaeid shrimp. Europe cultures flatfish, trout, oysters, mussels, and eels. Presently, plant aquaculture is almost exclusively restricted to Japan, China, and Korea, where the national diets include substantial amounts of macroalgae.

Aquaculture accounts for approximately 10% of the world's production of fish and shellfish, compared with an estimated 7% contribution in 1967. Roughly 65–70% of the world's aquaculture production consists of fresh-water, brackish-water, and marine fish, about 30–35% is divided equally between seaweeds and mollusks, and 0.3% crustaceans. In Japan, aquaculture supplies more than 98% of the yellowtail consumed, and in Israel about half of the finfish consumed are cultured.

Extensive to intensive culture. The worldwide practice of aquaculture runs the gamut from low-technology extensive methods to highly intensive systems. At one extreme, extensive aquaculture can be little more than contained stock replenishment, using natural bodies of water such as coastal embayments, where few if any alteration of the environment are made. Such culture usually requires a low degree of management and low investment and operating costs; it generally results in low yield per unit area. At the other extreme, intensive aquaculture, animals are grown in systems such as tanks and raceways, where the support parameters are carefully controlled and dependence on the natural environment is minimal. Such systems require a high degree of management and usually involve substantial investment and operating costs, resulting in high yields per unit area. Obviously, many aquacultural operations fall somewhere between the two extremes of extensive and intensive, using aspects of both according to the site, species, and financial constraints.

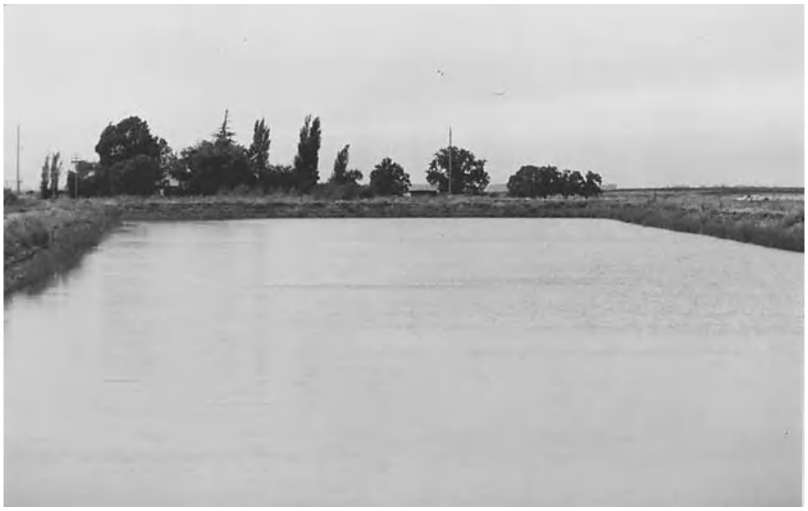


Fig. 1. Typical earthen pond as used for catfish culture.

In the Mediterranean area, culturists successfully use extensive methods to grow fish such as mullet, usually in polyculture with other species. Mullet are contained in large estuarine lagoons requiring the voluntary entrance of juveniles from the wild. These systems rely on naturally occurring feeds, and water is exchanged through arrangements of canals and sluices. Typically, the fish are harvested after 2–3 years, with relatively low production levels (88–176 1b/acre or 100–200 kg/hectare). Additional species that enter the lagoons are considered a welcome secondary crop.

An example of aquaculture further up the scale of intensity is catfish farming as normally practiced in warmer regions of the Western Hemisphere. In many areas, such as Mississippi, catfish are grown in fresh-water ponds that are often as large as 25 acres (10 ha) [**Fig. 1**]. These systems are stocked seasonally, and the animals are fed dried, milled feeds supplemented by organisms that occur naturally in the ponds. Catfish grown in this manner usually reach a marketable size of 0.7 kg within 2 years, with yields as high as 2640 lb/acre (3000 kg/ha). This yield, though far higher than that of extensive mullet culture, is restricted by the grower's lack of environmental control. For example, in densely stocked pond systems, oxygen depletion is a continual concern, and it is not uncommon for a farmer to experience massive fish deaths as a result of low oxygen levels. Aeration devices are commonly used in an attempt to avoid such disasters (**Fig. 2**).

To increase yields per unit area and limit environmental dependence, culturists such as catfish farmers are beginning to rear fish in raceway and tank systems similar to the techniques that have been

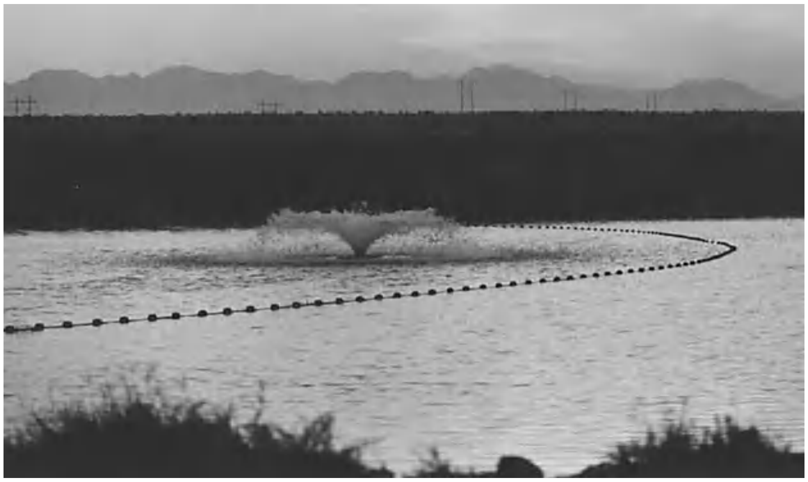


Fig. 2. Aerator operating in catfish pond.

used for years to grow trout. An example of such an operation is one in which catfish are reared in tanks ranging in size from 485 to 1450 ft^3 (45 to 135 m^3) [**Fig. 3**]. In this operation, geothermal water sources are used for year-round maintenance of optimal water temperature, and oxygen levels are carefully controlled. Such systems, when compared with pond culture, are capable of a thousandfold higher yields, providing harvestable fish of 2 lb (0.9 kg) within 12 months.

Both of the above examples of catfish-rearing techniques rely on classical brood stock and hatchery systems where mature breeding fish are maintained in ponds. Breeding pairs segregate; the females lay their egg masses in containers provided (usually milk cans); and the eggs are fertilized by the males. These fertilized egg masses are retrieved and hatched in troughs continually aerated by paddles (**Fig. 4**). Hatched larvae are maintained in small tanks or raceways and given artificial feeds until they reach a stocking size of 2.6–5.2 in. (7–15 cm).

Another outstanding example of intensive production is that of penaeid shrimp culture (the common marine prawn), as pioneered by the National Marine Fisheries Service Laboratory in Galveston, Texas, and the Universities of Arizona and Sonora. These systems consist of shallow raceways housed in greenhouselike domes made of inflated polyethylene (**Fig. 5**). The raceways may have either a closed, recirculating water source or a flow-through water system (**Fig. 6**). Penaeid species are reared on artificial feeds in high density (3.3 kg/m^2 or 0.68 $1b/ft^2$) from the postlarval stage. In such systems, environmental parameters are carefully maintained, and marketable animals can be produced within 4–6 months. The life cycle of the penaeid shrimp recently has been closed in captivity, and hatchery technology is available for the production of seed stock.

A unique combination of highly intensive and extensive aquaculture occurs in ocean ranching, as commonly employed with anadromous fish (which return from the ocean to rivers at the time of spawning). The two most notable examples are the ranching of salmon and sturgeon. In both instances, highly sophisticated hatchery systems are used to rear young fish, which are then released to forage and grow in their natural environment. The animals are harvested upon return to their native rivers.

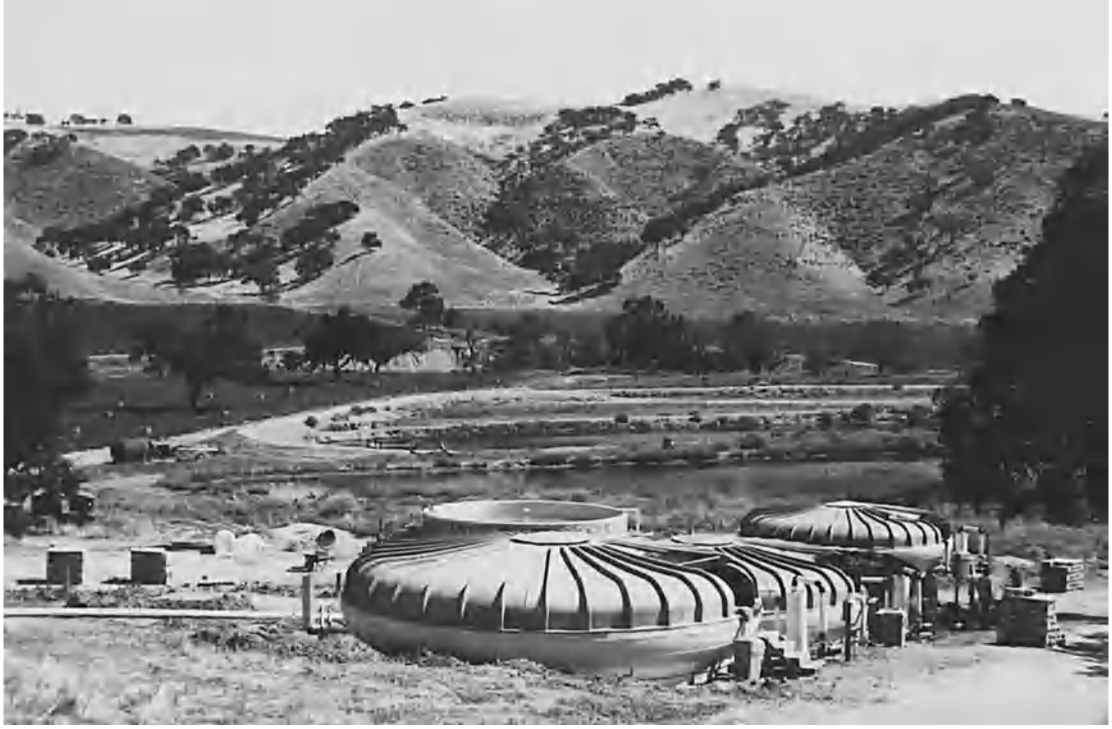


Fig. 3. Circular raceways which are used for intensive culture of catfish. This operation employs geothermal heat to ensure growth on a year-round basis. (*Calaqua, Inc., Paso Robles, California*)

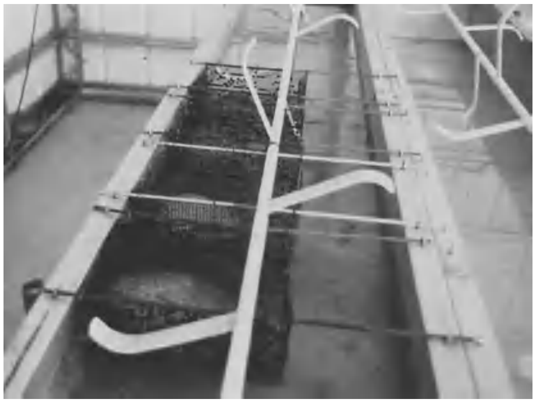


Fig. 4. Mechanized paddle troughs used in catfish aquaculture to provide aeration during egg incubation.

Industrialized aquaculture. While extremely extensive aquacultural operations are continuing, as they have been for generations, in the lesser-developed areas of the world, the more industrialized nations—where land and water resources are limited—have turned increasingly to intensive aquaculture. Intensive aquaculture brings with it high energy costs, necessitating the design of energy-efficient systems. As this trend continues, aquaculture will shift more to a year-round, mass-production industry using the least amount of land and water possible.

With this change to high technology and dense culturing, considerable knowledge and manipulation of the life cycles and requirements of each species are necessary. Specifically, industrialized aquaculture has mandated the development of reproductive control, hatchery technology, feeds technology, disease control, and systems engineering.

Improvements in each of these aspects are occurring and are critical to the future of intensive aquaculture. Reproductive control of a species is particularly important to maintain a steady supply of seed stock and to undertake genetic selection. Such control eliminates reliance on unpredictable wild stocks and is the first step toward the domestication of species for a production-oriented industry. In some fish species (notably trout and carp), reproduction has been controlled for a considerable time; in others (penaeid shrimp and lobster, for example), control has been achieved recently and is not yet routine. For a number of species (such as milkfish and eel), culturists still depend on wild stocks for replenishment.

Reproduction in captivity paves the way for geneticists to improve culture stocks through selective breeding designed to increase important traits such as growth rate, resistance to disease, tolerance to crowding, and reproductive performance. In terrestrial animal husbandry, genetic selection has substantially improved production, and the same is expected for aquaculture. *See* BREEDING (ANIMAL).

Unfortunately, many intensive aquacultural operations experience high mortalities during the animal's embryonic and larval development, hence the importance of creating efficient hatchery techniques. Many aquatic species exhibit a complicated series of larval stages, each requiring distinct environmental conditions and feeding regimes. As more biological information becomes available, hatchery systems are being carefully designed to allow high survival rates at each of these developmental levels.

In the current state of intensive aquaculture, feeds technology has been achieved only for a few species. When animals are grown in controlled systems, they are unable to forage for natural feeds, and the culturists must supply all required nutrients. To do this, however, necessitates knowledge of the animal's specific nutrient requirements and translation of that knowledge into inexpensive, formulated feeds that remain stable in the aqueous environment and are attractive to the particular animal. The lack of feed availability results in the widespread use of expensive, natural foods. *See* ANIMAL FEEDS.

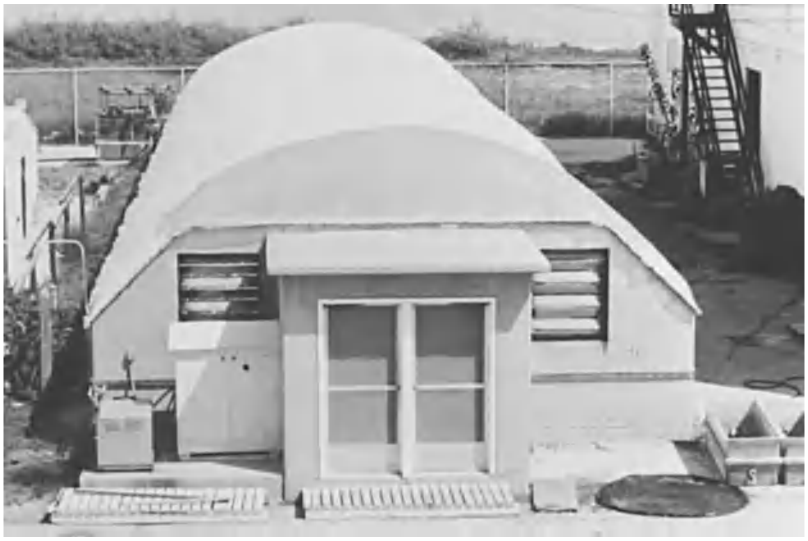


Fig. 5. Prototype inflated building housing raceways for intensive culture of marine shrimp. (*National Marine Fisheries Service Laboratory, Galveston, Texas*)

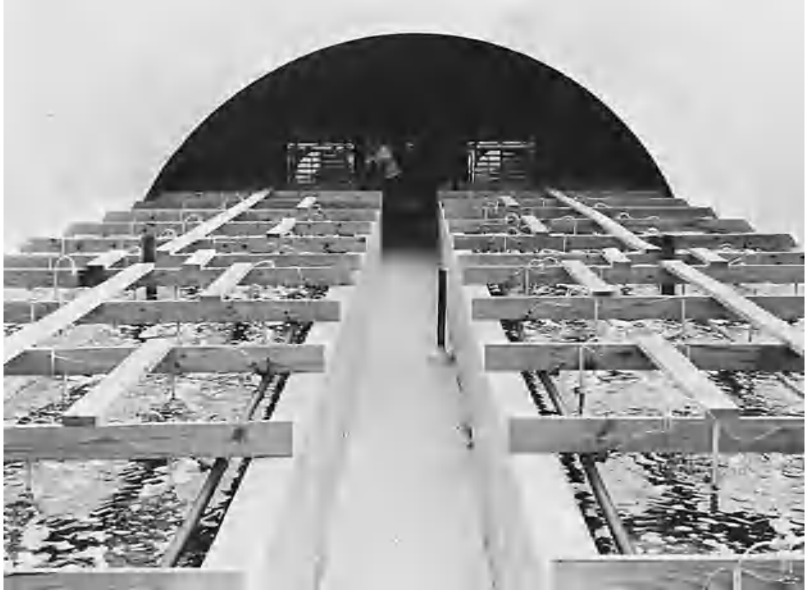


Fig. 6. Inside view of inflated building in Fig. 5 showing two long raceways used for shrimp production. (*National Marine Fisheries Service Laboratory, Galveston, Texas*)

With the trend toward more mechanized aquaculture, disease control has become increasingly important. In densely packed culture conditions, pathogens considered relatively minor in the wild or in extensive systems are capable of wiping out entire stocks. Modern aquaculture, therefore, has mandated disease-oriented research and the development of therapeutics and prophylactic techniques.

As manipulation and knowledge of the life cycles of aquatic species increases, so does the sophistication of the systems needed to support such control. The current emphasis in aquatic systems engineering is to design practical culture facilities that are appropriate to each species (including all developmental stages), are energy-efficient, provide adequate food delivery, and are suitable for the routine maintenance necessary for high water quality.

Regardless of the type of system used, aquacultural products are marketed as are fisheries products, except for some advantages. For one, fisheries products often must be transported on boats and may experience spoilage; whereas cultured products, which are land-based, can be delivered fresh to the various nearby markets. Also, intensively cultured products through genetic selection can result in a more desirable food than those caught in the wild, with uniform size and improved taste resulting from controlled feeding and rearing in pollution-free water.

Fisheries products traditionally have been the main source of animal protein in certain population centers of the world. In fact, approximately 13% of the animal protein consumed worldwide is provided by fisheries products, and the percentage is continually increasing. Coupled with the dramatic rise in world population, this growing demand for fisheries products has resulted in heavy pressure on the natural resources, which are rapidly approaching the maximum level of sustained harvest. Without the development of aquatic farming techniques, the natural fisheries would not be able to keep up with demand.

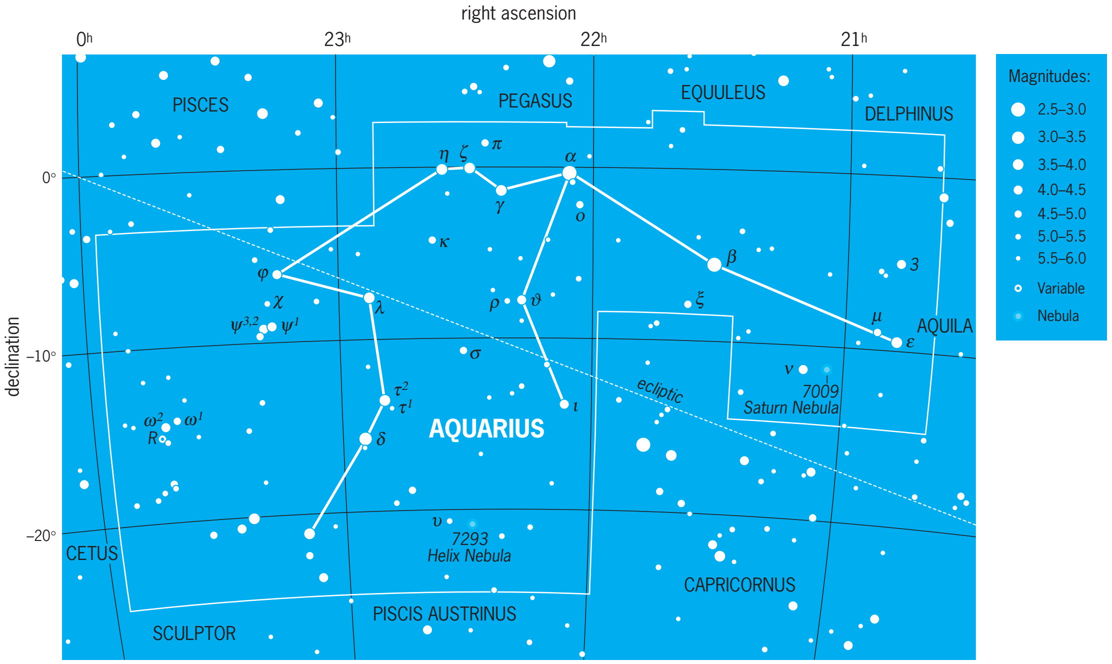
Aquaculture is following the lead of the development of agriculture, beginning with hunting, moving through stock containment and low-technology farming, and finally arriving at high-density, controlled culture. As the methodology improves, the modern techniques of aquaculture—and of agriculture—will have to be transferred to the lesser-industrialized areas of the world in an effort to offset the predicted food shortages in the coming years. *See* AGRICULTURE; MARINE FISHERIES.

Wallis H. Clark Jr.; Ann B. McGuire

Bibliography. J. E. Bardach, J. H. Ryther, and W. O. McLarney, *Aquaculture: The Farming and Husbandry of Freshwater and Marine Organisms,* 1972; Food and Agriculture Organization, *Proceedings of the FAO Technical Conference on Aquaculture, Kyoto, Japan, 1976*, FAO Fish. Rep. 188; P. Korringa, *Farming Marine Fishes and Shrimps: A Multidisciplinary Treatise,* 1976; P. R. Limburg, *Farming the Waters,* 1981; R. R. Stickney, *Principles of Warmwater Aquaculture,* 1979.

Aquarius

The Water Bearer, in Greek mythology, is a zodiacal constellation, meaning that the path of the Sun and planets passes through it. (The Sun actually passes



Modern boundaries of the constellation Aquarius, the Water Bearer. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

through 13 of the modern constellations each year, not the traditional 12.) It is almost entirely south of the celestial equator; none of its stars is especially bright (see **illustration**). Early constellation drawings show Aquarius as a man or boy pouring water from a bucket. The stream of water is oriented so as to flow into Piscis Austrinus, the constellation of the Southern Fish, pointing at the bright star Fomalhaut, which can be taken to correspond to the mouth of the fish. The constellation may have gotten its name from the Sun's position in it during the often-rainy month of February, which often led to floods in Babylonia and Egypt. The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION; ZODIAC.

The constellation boasts of two beautiful planetary nebulae, the Helix Nebula and the Saturn Nebula. The former subtends an angle in the sky larger than that of the Moon, although it is too faint to see without a telescope. *See* PLANETARY NEBULA.

Jay M. Pasachoff

Aquifer

A subsurface zone that yields economically important amounts of water to wells. The term is synonymous with water-bearing formation. An aquifer may be porous rock, unconsolidated gravel, fractured rock, or cavernous limestone. Economically important amounts of water may vary from less than a gallon per minute for cattle water in the desert to thousands of gallons per minute for industrial, irrigation, or municipal use.

Among the most productive are the sand and gravel formations of the Atlantic and Gulf Coastal plains of the southeastern United States. These layers extend for hundreds of miles and may be several hundred feet thick. Also highly productive are deposits of sand and gravel washed out from the continental glaciers in the northern United States; the outwash gravel deposits from the western mountain ranges; certain cavernous limestones such as the Edwards limestone of Texas and the Ocala limestone of Florida, Georgia, and South Carolina; and some of the volcanic rocks of the Snake River Plain in Idaho and the Columbia Plateau.

Aquifers are important reservoirs storing large amounts of water relatively free from evaporation loss or pollution. If the annual withdrawal from an aquifer regularly exceeds the replenishment from rainfall or seepage from streams, the water stored in the aquifer will be depleted. This "mining" of ground water results in increased pumping costs and sometimes pollution from seawater or adjacent saline aquifers. Lowering the piezometric pressure in an unconsolidated artesian aquifer by overpumping may cause the aquifer and confining layers of silt or clay to be compressed under the weight of the overlying material. The resulting subsidence of the ground surface may cause structural damage to buildings, altered drainage paths, increased flooding, damage to wells, and other problems. Subsidence of 10 to 15 ft (3.0 to 4.6 m) has occurred in Mexico City and parts of the San Joaquin Valley of California. Careful management of aquifers is important to maintain their utility as a water source. *See* ARTESIAN SYSTEMS; GROUND-WATER HYDROLOGY.

Ray K. Linsley

Arachnida

The largest class of the subphylum Chelicerata in the phylum Arthropoda. Most familiar of the included orders are the spiders, scorpions, harvestmen, and mites and ticks. Arachnids are mainly terrestrial and may be the oldest of the Recent terrestrial animals; scorpions are known from the Silurian (over 4×10^8 years ago).

Characteristics. The arachnid body is divided into a cephalothorax (prosoma) and an abdomen (opisthosoma). The cephalothorax has six pairs of appendages: the chelicerae (jaws), a pair of pedipalps, and four pairs of walking legs. There are no antennae. The abdomen may be either segmented or unsegmented and usually lacks appendages, or the appendages may be modified into specialized structures, for example, the spinnerets of spiders.

As in other arthropods there is an exoskeleton, the outside layer of which (the epicuticle) consists mainly of lipids; it is water-repellent and prevents water loss. Respiration is by means of book lungs or tracheae or both. Book lungs are paired invaginations of the ventral abdominal wall; their structure indicates that they are derived from the gills of aquatic animals. They consist of thin leaves that are filled with hemolymph (blood) and hang into a cavity. Air enters this cavity through a slit and flows between the blood-filled leaves; oxygen is taken up by hemocyanin in the blood and transported. Some arachnids have book lungs alone, while in others one or both pairs of book lungs may be modified into tracheae (internal tubes similar to those of insects), which open to the outside through small holes (spiracles) and carry air directly to tissues.

Arachnids have an open circulatory system. The heart lies in the anterior dorsal part of the abdomen. Blood enters the heart from the pericardial chamber through small openings (ostia). The blood is pumped into the prosoma through an anterior artery, and posteriorly through a posterior artery. It flows through vessels and chambers and around the book lungs and then into the pericardial cavity and back to the heart.

Digestion takes place outside the body. The captured prey is held by the chelicerae while digestive enzymes are secreted over it; then the broth is sucked up.

Excretory organs may be either thin-walled coxal glands that open to the outside on the basal segment (coxa) of each leg, or Malpighian tubules which enter the midgut. Arachnids excrete nitrogenous waste as guanine.

The nervous system is concentrated in the cephalothorax except in the primitive scorpions, which have ganglia in each segment. The brain encircles the esophagus; the protocerebrum lies above the esophagus, the remainder of the brain below. In most orders, the lower brain includes the ganglia of the appendages, ancestrally and in scorpions located in the thorax and abdomen.

Sense organs include simple eyes (ocelli), which consist of a cornea and lens that are continuous with the cuticle, a retinal layer with photoreceptors behind, and a vitreous body in between. Some eyes have a reflective layer (tapetum), which may improve vision in dim light. Hollow hairs located at the tips of palps and legs may be olfactory. Fine hairs (trichobothria) on the appendages are sensitive to air currents and vibrations. Membrane-covered pits in the exoskeleton occur all over the body and appendages, often in groups; called slit sense organs, they respond to tension in the exoskeleton and to some vibrations.

The genital opening and gonads are on the underside at the anterior of the abdomen. Males of many species perform an elaborate courtship. In many arachnids the male deposits a package of sperm (spermatophore). The female is attracted to the spermatophore or may be guided to it by the male, and takes it into her gonopore. Females produce yolk-rich eggs and may provide some care to their young.

The arachnids are predominantly predacious. The main exceptions are among the mites, where herbivores and parasites of plants and animals are common.

Classification. The Arachnida comprise more than 10 Recent orders. Scorpiones, the scorpions, have a segmented abdomen and a postabdomen bearing a sting; they are found worldwide. Palpigradi and Schizomida are minute arachnids inhabiting litter and soil. Uropygi, the tailed whip scorpions or vinegarones, are large (up to 6 cm or 2.5 in. in body length) with large pedipalps and a posterior flagellum; they are found mostly in warm climates. Amblypygi, the tailless whip scorpions, are flat, tropical animals with whiplike first legs; they capture insects with strong, spiny pedipalps. Araneae, the spiders, includes 34,000 known species worldwide; they are most abundant in the tropics. Spiders have a spherical spinneret-bearing abdomen attached to the cephalothorax by a thin stalk or pedicel. Solifugae (or Solpugida), the sun spiders, are quick-moving arachnids with large jaws; they inhabit deserts. Pseudoscorpionida, the pseudoscorpions, resemble scorpions but lack the post abdomen; smaller than 8 mm (0.3 in.), they respire with tracheae and are common in leaf litter worldwide. Opiliones, the harvestmen or daddy long- legs, have an eye on each side of a central anterior tubercle, and many have odor-producing glands. Males transmit sperm with a penis; females have a ovipositor. Members are found worldwide. Ricinulei, a small group living in tropical leaf litter or in caves, resemble ticks and have their jaws covered by a hood. Acari, the mites, form the largest group and are considered to comprise several orders by some workers. They are important because many plant and animal parasites are included, some of which (such as ticks) may transmit diseases. A few mites are found in fresh water and oceans. *See* AMBLYPYGI; ARANEAE; ARTHROPODA; CHELICERATA; OPILIONES; PALPIGRADI; PSEUDOSCORPIONIDA; RICINULEI; SCHIZOMIDA; SCORPIONES; SOLIFUGAE; UROPYGI.

H. W. Levi

Araeolaimida

An order of nematodes in which the amphids are simple spirals that appear as elongate loops, shepherd's crooks, question marks, or circular forms. The cephalic sensilla are often separated into three circlets: the first two are papilliform or the second coniform, and the third is usually setiform; rarely are the second and third whorls combined. Body annulation is simple. The stoma is anteriorly funnel shaped and posteriorly tubular; rarely is it armed. Usually the esophagus ends in a bulb that may be valved. In all but a few taxa the females have paired gonads. Male preanal supplements are generally tubular, rarely papilloid.

There are three araeolaimid superfamilies: Araeolaimoidea, Axonolaimoidea, and Plectoidea. The distinguishing characteristics of the Araeolaimoidea are in the amphids (sensory receptors), stoma, and esophagus. The amphids are in the form of simple spirals, elongate loops, or hooks. Although araeolaimoids are chiefly found in the marine environment, many species have been collected from fresh water and soil. The amphids in the Axonolaimoidea are generally prominent features of the anterior end, visible as a single-turn loop of a wide sausage shape. Species have been collected only from marine and brackish waters. Their feeding habits are unknown. All known species occur in marine or brackish-water environments.

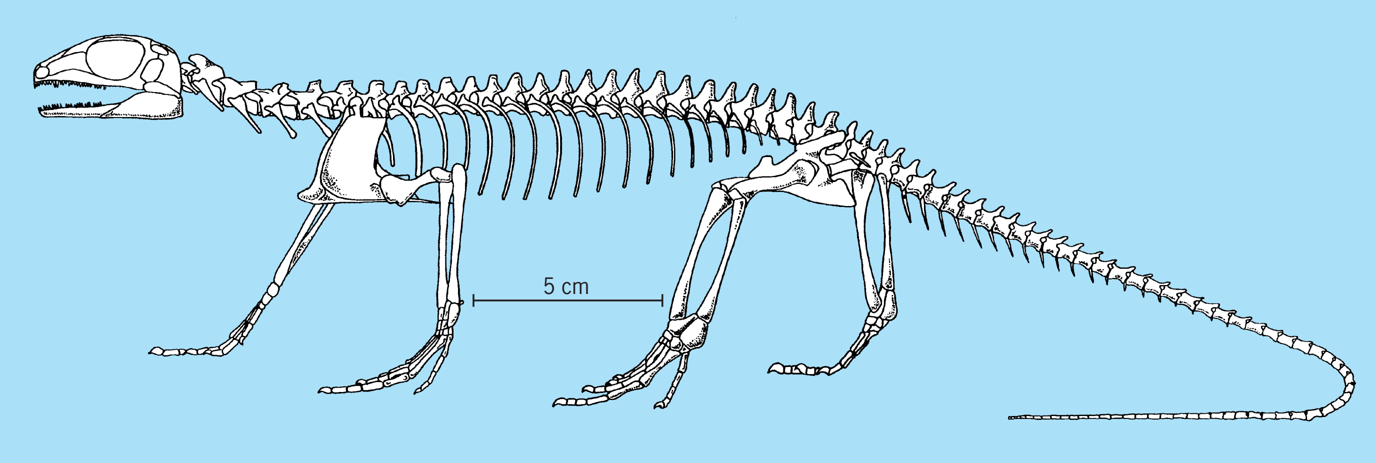
Plectoidea is a superfamily of small free-living nematodes, found mainly in terrestrial habitats, frequently in moss; some are fresh-water, and a few are marine. Those that inhabit moss cushions can withstand lengthy desiccation. For most, the feeding habit is unconfirmed; where known, they are microbivorous. Many are easily raised on agar cultures that support bacteria. *See* NEMATA.

Armand R. Maggenti

Araeoscelida

An order of Paleozoic diapsid reptiles including the families Petrolacosauridae and Araeoscelidae. Members of these families resemble primitive modern lizards, such as the green iguana, in size and most body proportions, but are distinguished by their elongate necks and distal limb elements (see **illus.**). *Petrolacosaurus*, from the Upper Pennsylvanian of Kansas, is the earliest known diapsid. The skull shows well-developed upper and lateral temporal openings and a suborbital fenestra that are characteristic of the diapsids.

Araeoscelis, from the lower Permian of Texas, lacks the lateral temporal opening, but the retention



Reconstruction of *Petrolacosaurus kansensis*. (*After R. R. Reisz, A Diapsid Reptile from the Pennsylvanian of Kansas, Spec. Pub. Mus. Nat. Hist., Univ. Kans. 7, 1981*)

of a triradiate jugal suggests that the squamosal has secondarily expanded to fill this area. The cheek may have been re-elaborated to provide additional support for the expanded jaw musculature associated with the blunt teeth in this genus.

The following derived features distinguish members of the Araeoscelida from other early diapsids: six to nine elongate neck vertebrae; a radius as long as the humerus and a tibia as long as the femur; expanded neural arches; posterior cervical and anterior dorsal neural spines with mammillary processes; a coracoid process; and enlarged lateral and distal pubic tubercles. These characters distinguish the Araeoscelida as a monophyletic group that does not include the ancestors of any of the other major diapsid groups. Primitive features of the skeleton of araeoscelids demonstrate that diapsids have a sister group relationship with the protorothyrids, the most primitive known amniotes.

In contrast to later diapsids, members of the Araeoscelida show no evidence of an impedance-matching middle ear. The stapes is massive, the quadrate is not emarginated for support of a tympanum, and there is no retroarticular process. *See* REPTILIA. Robert Carroll

Bibliography. R. R. Reisz, *A Diapsid Reptile from the Pennsylvanian of Kansas*, Spec. Pub. Mus. Nat. Hist., Univ. Kans. 7, 1981; R. R. Reisz, D. S. Berman, and D. Scott, The anatomy and relationships of the Lower Permian reptile *Araeoscelis, J. Vert. Paleontol.*, 1984.

Aragonite

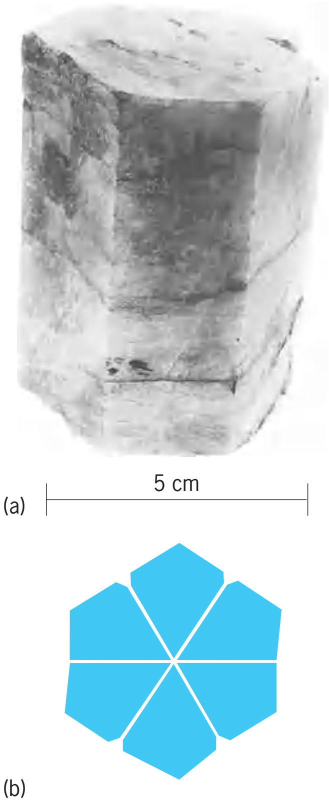
One of three naturally occurring mineral forms of calcium carbonate ($CaCO_3$). The other forms (or polymorphs) are the abundant mineral calcite and the relatively rare mineral vaterite. Still other forms of calcium carbonate are known, but only as products of laboratory experiments. The name aragonite comes from Aragon, a province in Spain where especially fine specimens occur. *See* CALCITE; CARBONATE MINERALS.

Mineralogy. Aragonite has an orthorhombic crystal structure in which layers of calcium (Ca) atoms alternate with layers of offset carbonate (CO_3) groups. A common crystallographic feature of aragonite is twinning, in which regions of crystal are misoriented as though they were mirror images of each other. This can give rise to a pseudohexagonal symmetry which is readily identified in large crystals (see **illus.**). Aragonite crystals are usually colorless or white if seen individually; however, aggregates of small crystals may exhibit different colors. Most aragonites are nearly pure calcium carbonate; however, small amounts of strontium (Sr) and less commonly barium (Ba) and lead (Pb) may be present as impurities. *See* TWINNING (CRYSTALLOGRAPHY).

Stability. At the low temperatures and pressures found near the Earth's surface, aragonite is metastable and should invert spontaneously to calcite, which is stable at these conditions. This, in part, explains why calcite is far more abundant than aragonite. However, at low temperatures the transformation of aragonite to calcite effectively occurs only in the presence of water, and aragonite may persist for long periods of geologic time if isolated from water. Increased temperature also promotes the transformation to calcite. Despite being metastable, aragonite rather than calcite is sometimes the favored precipitate from certain solutions, such as seawater, in which magnesium (Mg) ions inhibit precipitation of calcite.

Aragonite becomes more stable than calcite only at very high pressures—approximately 3500 bars (350 megapascals) at 77°F (25°C) and 7000 bars (700 MPa) at 570°F (300°C). Since it is the high-pressure polymorph, aragonite has a higher density than calcite (2.94 versus 2.71 g/cm^3 at 77°F or 25°C). This is achieved by packing nine oxygen atoms (from six CO_3 groups) around each calcium, whereas only six oxygens surround each calcium in calcite.

Occurrence. Aragonite occurs most abundantly as the hard skeletal material of certain fresh-water and



Aragonite. (*a*) Pseudohexagonally twinned specimen from Girgenti, Sicily (*American Museum of Natural History specimens*). (*b*) Arrangement of pseudohexagon twins (*after C. Klein, Dana's Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*).

marine invertebrate organisms, including pelecypods, gastropods, and some corals. The accumulated debris from these skeletal remains can be thick and extensive, usually at the shallow sea floor, and with time may transform into limestone. Most limestones, however, contain calcite and little or no aragonite. The transformation of the aragonite to calcite is an important step in forming limestone and proceeds by the dissolution of aragonite followed by the precipitation of calcite in the presence of water. This process may take more than 100,000 years. *See* LIMESTONE.

Aragonite also forms as a primary mineral during metamorphism at high pressures and moderate temperatures (330–570°F or 150–300°C). Its presence along with other high-pressure minerals in blueschist facies metamorphic rocks is regarded as important evidence for pressures in excess of 5000–7000 bars (500–700 MPa), corresponding to depths of greater than 10–15 mi (16–24 km) in Earth's crust. Examples of aragonite-bearing blueshists are found in the Franciscan Complex of the California Coast Ranges. *See* BLUESCHIST.

Other occurrences of aragonite include cave deposits (often in unusual shapes) and weathering products of calcium-rich rocks. Richard J. Reeder

Bibliography. L. L. Y. Chang, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 5B: *Non-Silicates: Sulphates, Carbonates, Phosphates, Halides*, 1995; C. Klein, *Dana's Manual of Mineralogy*, 21st ed., 1993; R. J. Reeder (ed.), *Carbonates: Mineralogy and Chemistry*, vol. 11 of *Reviews in Mineralogy*, Mineralogical Society of America, 1983.

Arales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Arecidae of the class Liliopsida (monocotyledons). It consists of two families, the Araceae, with only 1800 species, and the Lemnaceae, with only about 30.

The Araceae are herbs (seldom woody climbers) with ordinary roots, stems, and leaves (often broad and net-veined) and with the vessels confined to the roots. They have numerous tiny flowers grouped in a small to very large spadix. They are commonest in forested tropical and subtropical regions. *Anthurium* (elephant ear), *Arisaema* (jack-in-the-pulpit), *Dieffenbachia* (dumb cane), *Monstera*, and *Philodendron* are some well-known members of the Araceae.

The Lemnaceae, or duckweeds, are small, free-floating, thalloid aquatics that are generally conceded to be derived from the Araceae. Their flowers, seldom seen, are much reduced and form a miniature spadix. *Pistia* (water lettuce), a free-floating aquatic (but not thalloid) aroid is seen as pointing the way toward *Spirodela*, the least reduced genus of the Lemnaceae. *See* ARECIDAE; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM. Arthur Cronquist

Araneae

A natural order of the class Arachnida, also called Araneida, commonly known as the spiders. These animals are widespread over most of the land areas of the world and are well adapted to many different habitats. They are known to be one of the oldest of all groups of arthropods, and their remains are known from the Devonian and Carboniferous geological deposits. Through successive geological periods spiders have become adapted to insects as their chief source of food. On the other hand, certain insects consume the eggs of spiders, others parasitize the eggs, and still others capture adults and place them in their nests for food for their young.

Morphology. Spiders have but two subdivisions of the body, the cephalothorax and the abdomen, joined by a slender pedicel (**Fig. 1**). All parts of the body are covered by chitinous plates which often extend into curious outgrowths, such as spines, horns, and tubercles. Only simple paired eyes, ocelli, are present with the number varying from eight, the most common number, to none in a few species inhabiting lightless habitats.

The first of six pairs of appendages are termed the chelicerae or jaws (**Fig. 2*a***), each of which consists of a stout basal segment and a slender distal fang. A poison gland is usually associated with the basal segment, and the related duct opens near the distal end of the fang. A few spiders have lost their poison glands through retrogressive evolution. The second pair of appendages is the six-segmented pedipalps, simple organs of touch and manipulation in the female (Fig. 2*b*), but curiously and often complexly modified in males for use in copulation. The four pairs of thoracic legs consist of seven segments each, with considerable variation throughout the order (Fig. 2*c*). The spination of these appendages together with their relative lengths, associated sense organs, terminal claws, and other structures is all-important in classification and in adaptation.

Just in front of the genital opening in most females there is a more or less specific and often elaborately formed plate, the epigynum. This organ is also of importance in classification and in the reproductive activities of the female. The openings to the book lungs usually occur anterior to the epigynum. In most of the true spiders internal breathing tubules also occur with ventral openings posterior to the genital apertures. Distinctive paired ventral spinnerets occur near the posterior end of the abdomen. These vary in number from four to eight, eight being the most primitive condition. In certain families a sievelike plate lies immediately anterior to the foremost spinnerets. From this plate, the cribellum, a special kind of banded silk is extruded and used in conjunction with a comb on the fourth legs. The spinnerets and cribellum are directly associated with several types of abdominal glands and responsible for the production of the different types of silk characteristic of these animals.

Silk. Silk produced by spiders is a scleroprotein which is fine, light, elastic, and strong. At present it is used industrially only in the making of cross hairs in optical instruments. This use is diminishing as metal filaments and etched glass come into more common usage. In addition to the attractive orb webs, these animals also construct sheet webs, funnel webs, tube webs, and reticular webs. The spider's reliance upon silk extends its use to the making of egg cocoons, sperm webs by males, molting sheets, gossamer threads for ballooning, attachment disks, lining for burrows, hinges for trap doors, binding for captives, retreats, and drag lines.

Classification. No general agreement exists at present among araneologists concerning the classification and exact arrangement of the families of spiders. A. Petrunkevitch (1939) recognized 5 suborders: Liphistiomorphae containing only 2 families with a primitively segmented abdomen, and restricted to regions in the Eastern Hemisphere; Mygalomorphae with 8 families; Hypochilomorphae with a single relict family native to the southern Appalachian region; Dipneumonomorphae with 48 families; and finally the Apneumonomorphae with 3 families, which lack the book lungs and represent the most highly modified members of the order. There is now a tendency to increase the number of recognized families.

Mygalomorphae. This suborder includes the trap-door spiders, purse-web spiders, and the tarantulas of the Western Hemisphere. These tarantulas are entirely different from those of southern Europe, which are wolf spiders, Lycosidae. American tarantulas reach the largest size of any known spiders. Those from the Amazon region may attain a body length of more than 3 in. (75 mm) with a leg span of 10 in. (250 mm). About 30 species living within the confines of the United States have been recognized, most of them in the Southwest. Perhaps the true trap-door spiders are those of most interest to the general public. These creatures have a rake on their chelicerae

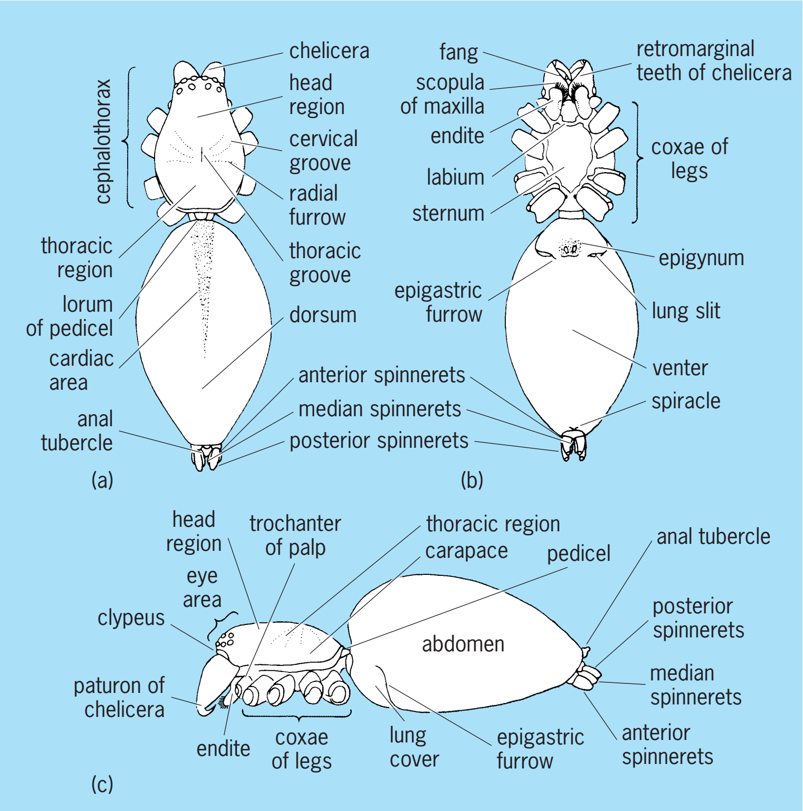

cephalothorax
chelicera
head region
cervical groove
radial furrow
thoracic groove
dorsum
thoracic region
lorum of pedicel
cardiac area
anal tubercle
anterior spinnerets
median spinnerets
posterior spinnerets
(a)
fang
scopula of maxilla
endite
labium
sternum
epigastric furrow
retromarginal teeth of chelicera
coxae of legs
epigynum
lung slit
venter
spiracle
(b)
head region
eye area
clypeus
paturon of chelicera
endite
trochanter of palp
coxae of legs
thoracic region
carapace
pedicel
abdomen
lung cover
epigastric furrow
anal tubercle
posterior spinnerets
median spinnerets
anterior spinnerets
(c)

Fig. 1. Morphology of a spider. (*a*) Dorsal aspect. (*b*) Ventral aspect. (*c*) Lateral aspect. (*After B. J. Kaston, How to Know the Spiders, William C. Brown, 1953*)

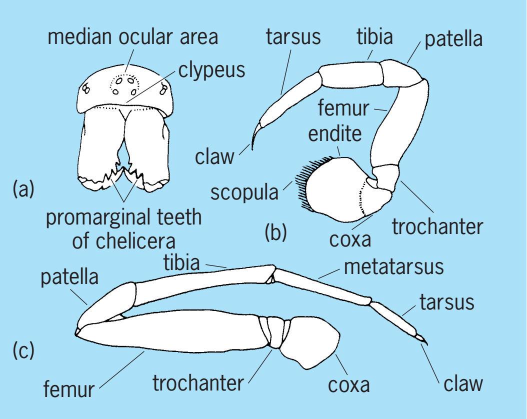

median ocular area
clypeus
(a)
promarginal teeth of chelicera
tarsus
tibia
patella
femur
endite
claw
scopula
(b)
coxa
trochanter
patella
tibia
metatarsus
tarsus
(c)
femur
trochanter
coxa
claw

Fig. 2. Appendages of spider. (*a*) Front view of face showing the chelicerae. (*b*) Pedipalp of female. (*c*) Leg. (*After B. J. Kaston, How to Know the Spiders, William C. Brown, 1953*)

with which they perform their digging operations. Several species living in the southern and western United States are remarkable in that their burrows are waterproofed, lined with silk, and finally capped with a cleverly concealing lid.

Dipneumonomorphae. These are the spiders most commonly seen in the United States. The common grass spiders, members of the genus *Agelenopsis* and closely related genera, build their silken sheets in great profusion in grassy areas. The sheet terminates in a funnel in which the spider sits in readiness for a quick dash to obtain an insect caught in the silken platform. Several families are commonly named hunting spiders because they do not usually build webs but actively hunt their prey. The genera *Lycosa, Pardosa*, and *Pirata* are typical of the family Lycosidae. These actively pursue their prey over the ground, sometimes across bodies of water, and even dive beneath the surface for a time. Some species build retreats or tunnels. Some resemble the true trap-door spiders in their ability to build hinged doors over the mouths of their burrows, and others construct cleverly arranged turrets at the entrance to their burrows. Females typically construct a carefully built silken bag, fill it with eggs, and carry it attached to their spinnerets. When the young are ready to emerge from the sac, they are liberated or escape and climb upon the mother's body to be carried by her until they are capable of caring for themselves. The crab spiders have gained this name because of their crablike lateral movements. Some of these animals have the remarkable ability of changing their color to agree with that of their background. The jumping spiders, Salticidae, are probably the most colorful of any family group, especially in the tropics where they abound. They have keen vision and rapid movements, and are active and highly efficient hunters. In stalking insects they exhibit many interesting aspects of behavior. The courtship dances of males have been studied and carefully described. Males are often provided with conspicuous plumes and other ornaments for display during these activities. A dragline is always spun out behind each of these spiders as it goes jumping or walking about. They do not spin webs, but are skillful in constructing retreats, where they may retire at night or for the laying of eggs. Some of the most successful ant mimics belong in this family.

Much notoriety has been given to the several species of the genus *Latrodectus* belonging to the Theridiidae or comb-footed spiders. In the United States the females are commonly known as black widows because of their color and the popular belief that the female always eats the male immediately after mating. Their bad reputation has arisen from the general belief that they are very aggressive and readily attack human beings with little or no provocation. There is general agreement among araneologists, however, that the spiders are in reality timid and unaggressive. Their biting is a fear reaction, and a bite is not inflicted unless the animal is strongly stimulated in some way. The venom is a potent neurotoxin, and care should be taken to avoid the animals. This is especially important, of course, where children are concerned. Fatalities are rare and competent medical care will usually carry the bitten person through to complete recovery. *See* ARACHNIDA.

Arthur M. Chickering

Bibliography. W. S. Bristowe, *The World of Spiders*. 1958; J. H. Comstock, *The Spider Book*, 1965; W. J. Gertsch, *American Spiders*, 1949; B. J. Kaston, *Spiders of Connecticut*, Conn. State Geol. Nat. Hist. Surv. Bull. 70, 1948; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; A Petrunkevitch, *Catalogue of American Spiders*, pt. 1, *Trans. Conn. Acad. Arts Sci*., vol. 33, 1939.

Arboretum

An area set aside for the cultivation of trees and shrubs for educational and scientific purposes. An arboretum differs from a botanical garden in emphasizing woody plants, whereas a botanical garden includes investigation of the growth and development of herbaceous plants, as well as trees and shrubs.

The largest of the arboretums in the United States is the Arnold Arboretum of Harvard University, founded in 1872, as the result of a bequest of James Arnold. More than 6500 kinds of woody plants are grown there. The United States National Arboretum, near Washington, D.C., was founded in 1927. Here hundreds of native and imported woody shrubs and trees are grown experimentally. Studies are made of shade and light tolerance, water requirements, temperature range, environmental influence, reproduction, and other botanical problems. *See* BOTANICAL GARDENS.

Earl L. Core

Arboriculture

A branch of horticulture concerned with the selection, planting, and care of woody perennial plants. Knowing the potential form and size of plants is essential to effective landscape planning as well as to the care needed for plants. Arborists are concerned primarily with trees since they become large, are long-lived, and dominate landscapes both visually and functionally.

Plants can provide privacy, define space, and progressively reveal vistas; they can be used to reduce glare, direct traffic, reduce soil erosion, filter air, and attenuate noise; and they can be positioned so as to modify the intensity and direction of wind. They also influence the microclimate by means of evaporative cooling and interception of the Sun's rays, as well as by reflection and reradiation. Certain plants, however, can cause human irritations with their pollen, leaf pubescence, toxic sap, and strong fragrances from flowers and fruit. Additionally, trees can be dangerous and costly: for example, branches can fall, and roots can clog sewers and break paving.

Plant selection. A plant's growth habit and its size at maturity are important characteristics in the selection process. It is important to select plants with roots that are free from kinks or circling. It is also desirable that trees be able to stand without support and have small branches along the trunk to provide protection and nourishment during establishment. Whether the plant is to be used for shade, screening, privacy, accent, or protection will determine what kind is selected. Sturdy root systems protect plants from being uprooted during storms, and if the roots are tolerant of poor soil the plants will remain healthy and can resist pests more readily. Leaves, flowers, and fruit are another consideration; not only are they visually important, but they also can have a considerable impact on maintenance.

Most plants grow best in deep, loamy soils, although some are able to withstand unfavorable soil conditions such as poor drainage, strata of different textures, extreme acidity or alkalinity, or chemical toxicities. The lowest recorded temperatures in a given region will determine which plants can survive in that region. Long-lived plants in particular should be selected to withstand the lowest temperatures expected.

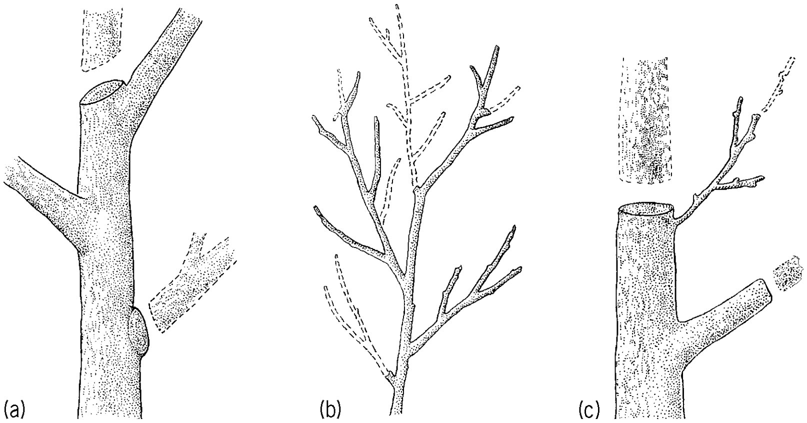
Species native to an area may not perform well, however, particularly around buildings where microclimate and soil can differ considerably from preconstruction conditions.

Planting and care. Unless the soil has been compacted, the planting hole should only be deep enough to take the roots. The soil around the roots should be firmed and watered to settle it and remove air pockets. Unless the plant is overgrown, the only pruning needed after planting is some thinning to begin tree structure development, to shape shrubs, or to select vine canes to trellis. Little pruning is necessary for central-leader trees, conifers, and some hardwoods to grow strong and be well shaped. Species that become round-headed, however, may need considerable pruning to ensure the desired height of branching and a strong branch structure.

The less a branch or tree is pruned, the larger it will become. Therefore, only large branches that are too low or will compete or interfere with more desirable branches should be removed. Permanent branches, particularly of large-growing trees, should be at least one-third smaller in diameter than the trunk where they arise and be vertically spaced at least 18 in. (45 cm) apart along the trunk.

A tree will be more open and better retain its natural form if branches are removed completely (thinned) in contrast to being headed or stubbed back (see **illus.**). Heading concentrates subsequent growth just below the pruning cut and results in dense foliage with weakly attached branches. In removing a branch the final cut should be just to the outside of the branch bark ridge in the crotch and the collar below. Such a cut minimizes the size of the wound and the possibility that the trunk will decay. Seldom is it advisable to paint pruning wounds.

Fertilization of young plants is necessary to achieve rapid growth; mature plants, however, may need little or no added nutrients. Nitrogen is almost universally deficient in soils, and it usually is the only element to which trees and large shrubs will respond. Nitrogen fertilizers are water-soluble and can be applied to the soil or lawn surface and then watered in. In alkaline soils, the availability of iron or manganese may be so low for certain kinds of plants that they exhibit the typical pale leaves with narrow (due to iron deficiency) or wide (due to manganese deficiency) darker green extending bands along the veins. Increasing soil acidity or applying chelated nutrients usually ameliorates these problems. *See* FERTILIZER.



Methods of plant pruning. (*a*) Thinning (removing a branch at its origin) of a mature tree. (*b*) Thinning a young tree. (*c*) Heading (pruning to a small lateral branch or to a stub) of a mature tree. (*After R. W. Harris, Arboriculture: Care of Trees, Shrubs, and Vines in the Landscape, Prentice-Hall, 1983*)

Irrigation can ensure establishment of young plants, the attractive appearance of foliage, and even survival. Many mature plants can endure long periods without rain or irrigation, if a large volume of moist soil is present and the plants have extensive root systems. If the water supply is limited, it is important for the soil to be fully moist at the start of the growing season. A few heavy irrigations are more efficient than frequent light irrigations; more plants do poorly from too much water than not enough. *See* PLANT-WATER RELATIONS.

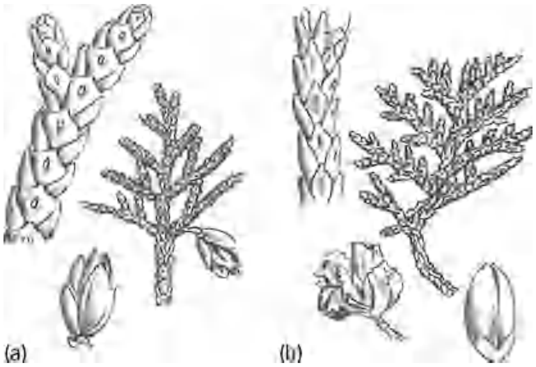
Mulch, that is, material placed on the soil surface, can control weeds, protect the soil from compaction and erosion, conserve moisture, moderate soil temperatures, provide an all-weather surface for walking, and allow plants to root in the most fertile and well-aerated surface soil. A wide range of organic and inorganic or synthetic materials can be employed. Mulch should be kept 2 in. (5 cm) away from the trunks of plants to reduce trunk disease and rodent damage. *See* LANDSCAPE ARCHITECTURE.

Richard W. Harris

Bibliography. B. Ferguson, *All about Trees*, 1982; C. W. Harris and N. T. Dines (eds.), *Time-Saver Standards for Landscape Architecture* 2d ed., 1997; R. W. Harris, *Arboriculture: Care of Trees, Shrubs, and Vines in the Landscape*, 1983; J. Hudak, *Trees for Every Purpose*, 1980; A. L. Shigo, *A New Tree Biology*, 1936; B. B. Wilson, *The Growing Tree*, 1984.

Arborvitae

A plant, sometimes called the tree of life, belonging to the genus *Thuja* of the order Pinales. It is characterized by flattened branchlets with two types of scalelike leaves. At the edges of the branchlets the leaves may be keeled or rounded; on the upper and lower surfaces they are flat, and often have resin glands. The cones, about $^1/_2$ in. (1.2 cm) long, have the scales attached to a central axis (see **illus.**). *See* PINALES; RESIN; SECRETORY STRUCTURES (PLANT).



Examples of arborvitae. (*a*) Eastern arborvitae (*Thuja occidentalis*). (*b*) Oriental arborvitae (*T. orientalis*).

The tree is valued both for its wood and as an ornamental. *Thuja occidentalis*, of the eastern United States, grows to a height of 60 ft (18 m). Known as the northern white cedar, it occurs in moist or swampy soil from Nova Scotia to Manitoba and in adjacent areas of the United States, and extends south in the Appalachians to North Carolina and Tennessee. The wood is light, soft, and fragrant. It resists deterioration from moisture, and is used for construction, cabinetmaking, and making barrels and shingles. A large number of arborvitae seedlings are sold annually as ornamentals.

Other important species include the giant arborvitae (*T. plicata*), which grows to 180 ft (54 m) in the Pacific Northwest and is highly prized for its resistance to decay; oriental arborvitae (*T. orientalis*); and Japanese arborvitae (*T. standishii*). Among the horticultural forms are the dwarf, pendulous, and juvenile varieties. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Arboviral encephalitides

A number of diseases, such as St. Louis, Japanese B, and equine encephalitis, which are caused by arthropod-borne viruses (arboviruses). In their most severe human forms, the diseases invade the central nervous system and produce brain damage, with mental confusion, convulsions, and coma; death or serious aftereffects are frequent in severe cases. Inapparent infections are common (see **table**).

Arboviruses. The arbovirus "group" comprises more than 250 different viruses, many of them differing fundamentally from each other except in their ecological property of being transmitted through the bite of an arthropod. A large number of arboviruses of antigenic groups A and B are placed in the family Togaviridae, in two genera, alphavirus (serological group A) and flavivirus (serological group B). Still other arboviruses, related structurally and antigenically to one another but unrelated to Togaviridae, are included in the family Bunyaviridae, consisting chiefly of the numerous members of the Bunyamwera supergroup—a large assemblage of arboviruses in several antigenic groups which are cross-linked by subtle interrelationships between individual members. The nucleic acid genomes of all arboviruses studied thus far have been found to be RNA; most arboviruses are inactivated by lipid solvents (ether or sodium deoxycholate) and are unstable at room temperature. *See* VIRUS CLASSIFICATION.

Arboviruses will grow in many laboratory animals and in embryonated eggs and tissue cultures, often inapparently; they are almost universally pathogenic for newborn mice. Of the groups that have been established on the basis of antigenic relationships, most of those causing encephalitis belong to group A or B or the California complex. Animals inoculated repeatedly with one virus may develop serum antibodies against other group members. Depending upon the closeness of relationship, heterologous responses range from low and transient up to responses nearly equaling the homologous response. In some instances only hemagglutination-inhibiting antibodies cross, whereas in others complement-fixing and neutralizing antibody crossings occur. Within the arbovirus groups are several important members which, although neuropathogenic in the mouse and antigenically related to those causing encephalitis in humans, invade the human central nervous system only atypically.

Members of serological group A include western equine encephalitis (WEE), eastern equine encephalitis (EEE), and Venezuelan equine encephalitis (VEE) viruses; and Mayaro, Semliki Forest, Chikungunya, and Sindbis viruses, which have nonencephalitic syndromes. Group A viruses are chiefly mosquito-borne. Serological group B viruses include Japanese B, St. Louis, and Murray Valley encephalitis viruses (mosquito-borne), and the viruses of the Russian tick-borne complex, some of which produce encephalitis (Russian spring-summer), whereas others cause hemorrhagic fevers (Omsk, Kyasanur Forest) or other syndromes, such as louping ill. Also in group B are the nonneurotropic viruses of West Nile fever, yellow fever, dengue, and other diseases. *See* LOUPING ILL; YELLOW FEVER.

Pathogenesis. Illness begins 4–21 days after the bite of the infected vector, with sudden onset of fever, headache, chills, nausea, and generalized pains; marked drowsiness and neck rigidity may ensue. Signs and symptoms are often diphasic, the second phase being more severe and involving mental confusion, speech difficulties, convulsions, and coma. Mortality case rates vary in different diseases and epidemics. Mortality in Japanese B encephalitis in older age groups has been reported to be as high as

Summary of six major human arbovirus infections which occur in the United States

Disease	Exposure	Distribution	Vectors	Infection:case ratio (age incidence)	Sequelae	Mortality rate, %
Western equine encephalitis	Rural	Pacific Mountain West Central Southwest	*Culex tarsalis*	50:1 (under 5) 1000:1 (over 15)	+	2–3
Eastern equine encephalitis	Rural	Atlantic Southern coastal	*Aedes sollicitans* *A. vexans*	10:1 (infants) 50:1 (middle-aged) 20:1 (elderly)	+	50–70
St. Louis encephalitis	Urban-rural	Widespread	*C. pipiens* *C. quinquefasciatus* *C. tarsalis* *C. nigrapalpus*	>400:1 (young) 64:1 (elderly)	±	5–10
Venezuelan equine encephalitis	Rural	South (also South America, Central America)	*Aedes* *Psorophora* *Culex*	Unknown ratio	Unknown	0.5%
California encephalitis	Rural	North Central Atlantic South	(*Aedes* sp.?)	Unknown ratio (most cases under 20)	±	Fatalities rare
Colorado tick fever	Rural	Pacific Mountain	*Dermacentor andersoni*	Unknown ratio (all ages affected)	Rare	Fatalities rare

80%. Serious aftereffects such as mental and sensory defects, epilepsy, and paralysis occur. Abortive cases may produce the early symptoms or resemble aseptic meningitis. California encephalitis is seldom fatal, and prognosis excellent, but convalescence may be prolonged.

Laboratory diagnosis. Testing for hemagglutination-inhibiting (HI) serum antibodies may be complicated by antigenic cross-reactions, but can establish the arboviral group responsible. Specific identification depends on the detection of variations in the HI, complement-fixing, or neutralizing antibody titers between appropriately timed acute and convalescent specimens. Virus isolation from the blood or spinal fluid is usually not successful.

Epidemiology. All arboviruses causing encephalitis require an infected arthropod vector for transmission. The usual natural hosts are mammals or birds, with humans accidentally infected in the cycle. Ticks may serve not only as vectors but also as a natural reservoir, because the virus may pass transovarially to offspring; in other instances birds or mammals are the probable reservoir. In highly endemic regions many persons have antibodies, chiefly from inapparent infections. Each virus is limited to definite geographic areas. The population's immunity to its local virus may cross-immunize against closely related viruses. For example, endemic West Nile fever may partially protect its areas against infections by other flaviviruses, some of which can cause encephalitis. Another hypothesis suggests a common ancestor, different descendants having adapted to various regions, each with its special vector and reservoir.

Treatment. There is no proved specific treatment. In animals, hyperimmune serum given early may prevent death. Killed virus vaccines have been used in animals and in persons occupationally subjected to high risk. A live, attenuated vaccine against Japanese B encephalitis virus, developed in Japan, has been used experimentally with some success, not only in pigs to reduce amplification of the virus in this important vertebrate reservoir but also in limited trials in humans. In general, however, control of these diseases continues to be chiefly dependent upon elimination of the arthropod vector. *See* ANIMAL VIRUS; VIRUS.

Joseph L. Melnick

Bibliography. E. Jawetz, J. L. Melnick, and E. A. Adelberg, *Review of Medical Microbiology*, 12th ed., 1976.

Arc discharge

A type of electrical conduction in gases characterized by high current density and low potential drop. The electric arc was discovered by Humphry Davy in 1808, when he connected a piece of carbon to each side of an electric battery, touched the two pieces of carbon together, then drew them slightly apart. The result is a dazzling steam of ionized air, or plasma, at a temperature of 6000°C (10,800°F), the surface temperature of the sun. A typical arc runs at a voltage drop of 100 V with a current drain of 10 A. The arc has negative resistance-the voltage drop decreases as the current increase-so a stabilizing resistor or inductor in series is required to maintain it. The high-temperature gas rises like a hot-air balloon while it remains anchored to the current-feeding electrodes at its ends. It thereby acquires an upward-curving shape, which accounts for its being called an arc.

Applications. There are many applications of such an intensely hot object. The brilliant arc and the incandescent carbon adjacent to it form the standard light source for movie theater projectors. The electronic flashgun in a camera uses an intense pulsed arc in xenon gas, simulating sunlight. Since no solid-state material can withstand this temperature for long, the arc is used industrially for welding steel and other metals. Alternatively, it can be used for cutting metal very rapidly. Electric arcs form automatically when the contacts in electrical switches in power networks are opened, and much effort goes into controlling and extinguishing them. Lightning is an example of a naturally occurring electric arc. *See* ARC HEATING; ARC LAMP; ARC WELDING; CIRCUIT BREAKER; LIGHTNING; OPTICAL PROJECTION SYSTEMS; STROBOSCOPIC PHOTOGRAPHY; WELDING AND CUTTING OF MATERIALS.

Temperature. The arc has been pushed to extremely high temperatures in the search for thermonuclear fusion, the record temperature being $4\times 10^5{}^\circ$C in a long pulse in helium. The arc temperature appears to be limited by the energy lost in intense radiation from the interface region between the intensely hot, fully ionized plasma core and the surrounding cooler, un-ionized gas.

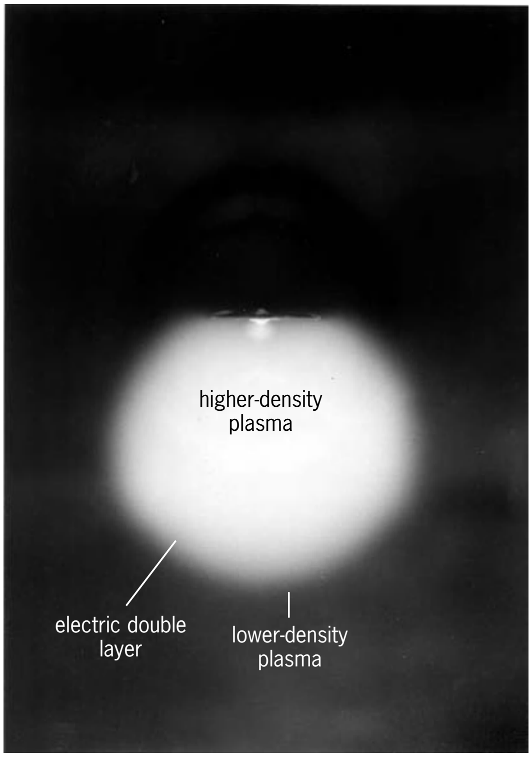
Structure. The structure of an electric arc has several components. The most prominent feature is the brilliant column of partially ionized gas, or plasma, referred to as the arc body. The arc body in general has a relatively low voltage drop along its length. The percentage of ionization can be calculated from Saha's equation, shown below. Here, n_0 and n_i are

$$\log_{10}\left(\frac{n_i{}^2}{n_0}\right) = -5040\frac{v_i}{T} + \frac{3}{2}\log_{10}(T) + 15.385$$

the neutral and ionic densities (cm^{-3}), respectively, v_i the ionization potential in volts, and T the absolute temperature in kelvins.

The electrons exciting the arc are emitted from an intensely hot spot on the negative electrode, or cathode. This spot is heated by positive ion bombardment from the plasma column. The electrons exit the arc from a second brilliant spot on the positive electrode, or anode. Ion bombardment also occurs here. The arc column automatically becomes positive with respect to both electrodes in order to contain the hot, rapidly moving, negative electrons, and this positive potential ejects ions. Both electrodes are separated from the plasma column by a thin transition region, the Debye sheath. The voltage drop between the cathode and the arc body is called the cathode fall, and is in general a major source of energy to the arc body. The voltage increase between the arc body and the anode is called the anode rise. The details of the structure of the electric arc are still not fully understood.

Anode spots. If the current to the anode is not sufficient to maintain the discharge, an interesting phenomenon can frequently be observed. Spherical luminous plasmas that are called anode spots or fireballs form on the anode (see **illus.**). Their formation and size depend on the ionization potential, the background pressure, and the possible impurities in the host gas. There plasmas usually are found at an imperfection in a smooth surface where the electric field is larger. Once formed, they are very stable and can exist for hours.



Anode soot or fireball in argon. (*From B. Song, N. D'Angelo, and R. L. Merlino, On anode spots, double layers and plasma contractors, J. Phys. D: Appl. Phys., 24:1789–1795, 1991*)

Experiments have shown that anode spots have an electric potential that is slightly greater than the ambient plasma in which they reside. The indication is that some of the higher-mobility electrons have escaped, leaving the anode spot with a slight excess positive charge. Under certain conditions, more than one anode spot can form. If a circular anode contains an edge in the shape of a sharp lip, the additional anode spots will be located as far from each other as possible on the circular lip. Two anode spots will be at the opposite ends of the diameter, three will be at the corners of a square, and so forth. The mutual repulsion of anode spots and their search for a stable equilibrium position result from the fact that each one has approximately the same amount of net positive charge enclosed within it.

Simulation of natural phenomena. Some experiments have been directed toward simulating natural phenomena in the laboratory. If a voltage pulse is applied to the anode, the high-density anode spot explodes into the lower-density ambient plasma. Voltage signals detected by small movable Langmuir probes within the plasma can be displayed by using an oscilloscope that is triggered by the voltage pulse. Hence, the spatial and temporal characteristics of the expansion of a higher-density plasma into a lower-density plasma can be monitored and investigated. The stable transition region from the higher-density spherical plasma to the ambient lower-density spherical plasma to the ambient lower-density plasma is similar to phenomena found in space. The term double layer is applied to the two adjacent laminar charge-density regions that are found within the transition region. The signs of the two charged regions are opposite, with the positively charged region being adjacent to the higher-potential side of the double layer. *See* ELECTRICAL CONDUCTION IN GASES.

Igor Alexeff; Karl E. Lonngren

Arc heating

The heating of matter by an electric arc. The matter may be solid, liquid, or gaseous. When the heating is direct, the material to be heated is one electrode; for indirect heating, the heat is transferred from the arc by conduction, convection, or radiation. *See* ARC DISCHARGE.

At atmospheric pressure, the arc behaves much like a resistor operating at temperatures of the order of thousands of kelvins. The energy source is extremely concentrated and can reach many millions of watts per cubic meter. Almost all materials can be melted quickly under these conditions, and chemical reactions can be carried out under oxidizing, neutral, or reducing conditions.

Arc heating is applied in, but is not limited to, electric arc furnaces (both open and submerged arc), plasma torches (including cutting, spraying, and gas heating), arc welders, arc machining, and arc boring. By far the greatest application in installed capacity is in electric arc furnaces. *See* ARC WELDING; ELECTRIC FURNACE; WELDING AND CUTTING OF MATERIALS.

Direct-arc furnaces. In a direct-arc furnace, the arc strikes directly between the graphite electrodes and the charge being melted. These furnaces are used in steelmaking, foundries, ferroalloy production, and some nonferrous metallurgical applications. Although an extremely large number of furnace types are available, they are all essentially the same. They consist of a containment vessel with a refractory lining, a removable roof for charging, electrodes to supply the energy for melting and reaction, openings and a mechanism for pouring the product, a power supply, and controls. The required accessory components include water-cooling circuits, gas cleaning and extraction equipment, cranes for charging the furnace, and ladles to remove the product. Because the electrodes are consumed by volatilization and reaction, a mechanism must be provided to feed them continuously through the electrode holders.

Most direct-arc furnaces operate in a batch mode whereby they are charged with raw materials (scrap steel, ores, and fluxes) which are then melted down to a two-phase liquid of slag and metal. The metal is refined by the addition of further raw materials, and finally the slag and metal products are removed from the furnace by pouring.

In a typical direct-arc furnace for steelmaking (**Fig. 1**), three electrodes are used in three-phase alternating-current operation. Furnace powers may vary from a few megavolt-amperes to over 100 MVA, and furnace inside diameters may range from less than 3 ft (1 m) to more than 20 ft (7 m). Generally, the furnace shell is cylindrical and the bottom is hemispherical; in larger furnaces the bottom is mounted on rockers for pouring. Because productivity is almost as important as energy economy in modern steelmaking furnaces, large portions of the furnace are water-cooled to reduce refractory erosion and minimize shutdown time for maintenance. *See* ELECTROMETALLURGY; FURNACE CONSTRUCTION; STEEL MANUFACTURE.

Refractory lining. The furnace lining functions to contain the liquid product within the furnace and to reduce heat losses. It must be carefully chosen for chemical and thermal compatibility with the specific process being used and must have sufficient strength and resilience to withstand the rigors of charging by heavy loads of scrap steel. In basic steelmaking, the bottom usually consists of a subhearth of magnesia bricks topped by a monolithic layer of rammed high-magnesia refractory which forms the hearth. The side walls are usually made of key-shaped magnesia bricks or magnesia—carbon bricks. The slag line is particularly susceptible to erosion, and therefore a wide variety of different materials are used. Since the water-cooled side walls are eventually coated with a layer of slag, only a thin layer of refractory is required for startup. The roof of the furnace, which generally is also water-cooled, is commonly made of 70% alumina brick. *See* REFRACTORY.

Electrical supply. Furnace transformers for direct-arc furnaces normally accept input voltages up to 69 kV, and so a step-down transformer is normally needed

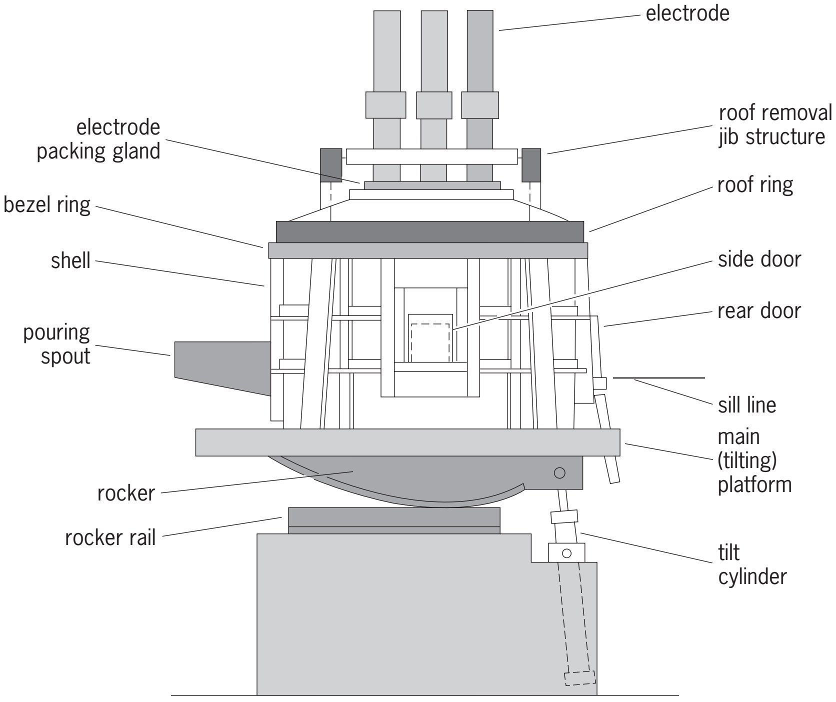


Fig. 1. Schematic representation of an electric arc furnace used in steelmaking. (*After AIME, Iron and Steel Society, Electric Furnace Steelmaking, ed. by C. R. Taylor, L. G. Kuhn publisher, 1985*)

in a substation if the utility voltage is more than this value. The primary side of the power supply consists of a disconnect switch, an oil circuit breaker (to protect equipment on the primary side of the furnace transformer), the furnace-operating vacuum switch, and the primary transformer windings with tap changers. The operating switch must be particularly rugged since it may have to interrupt furnace power up to 60 times per day. The secondary furnace circuit consists of the secondary windings on the transformer; flexible cables, which allow tilting of the furnace and movement of the roof; tubular bus-bars, which carry current to the electrodes; and electrodes. The transformer and most current-carrying lines are heavily water-cooled. The correct layout of the cables and bus tubes to the electrodes is very important to stable furnace operation so that the mutual coupling between the phases can be equalized and the currents between the phases balanced.

Instrumentation, additional reactors (which may have to be added to smaller furnace power supplies to limit the arc current during short circuit), and devices for power factor correction and harmonic suppression are also part of the electrical supply. The power levels required at different stages of the steelmaking cycle vary considerably, being low for startup, high for melt-down, and medium during refining. This power is controlled by adjusting the voltage taps on the primary of the furnace transformer to give the desired operating voltage and by continuously moving the electrodes up and down to adjust the current to the desired level. The latter operation is automated in modern furnaces by means of computer-controlled electric motors or hydraulic systems. Modern furnace transformers have at least six voltage taps; since the transformer can be changed from delta to star connection, this gives a choice of 12 voltage settings. Tap settings may be changed on load or off load; off-load changers are more common. *See* TRANSFORMER; WIRING.

Submerged-arc furnaces. The arcs are submerged below the solid feed and sometimes below the molten product. Submerged-arc furnaces differ from those used in steelmaking in that raw materials are fed continuously around the electrodes and the product and slag are tapped off intermittently. The furnace vessel is usually stationary. Submerged-arc furnaces are often used for carbothermic reductions (for example, to make ferroalloys), and the gases formed by the reduction reaction percolate up through the charge, preheating and sometimes prereducing it. Because of this, the energy efficiency of this type of furnace is high. The passage of the exhaust gas through the burden also filters it and thus reduces air-pollution control costs.

Plasma torches and devices. Although carbon arcs are plasmas, common usage of the term plasma torch suggests the injection of gas into or around the arc. This gas may be inert, neutral, oxidizing, or reducing, depending on the application and the electrodes used. Plasma torches are available at powers ranging from a few kilowatts to over 10 MW; usually they use direct-current electricity and water-cooled metallic electrodes.

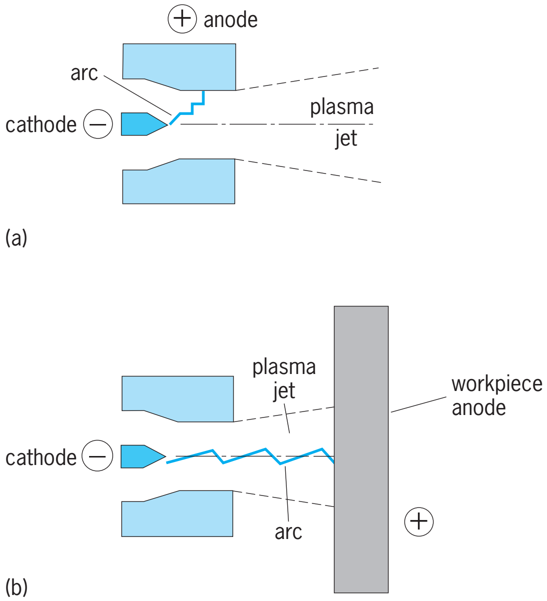


Fig. 2. **Diagrams of small-scale plasma torches. (*a*) Nontransferred-arc torch. (*b*) Transferred-arc torch.**

Small scale. Small-scale (20–150-kW) plasma torches are used in spraying and metal-cutting operations as well as in laboratory-scale research. Plasma spraying uses a plasma jet formed by striking an arc between a thoriated tungsten cathode and a cylindrical copper anode (**Fig. 2***a*). This type of plasma torch is called a nontransferred-arc torch since the arc remains within the torch and is not transferred externally. Particles of the material to be sprayed are injected radially into this jet where they are entrained, melted, and projected onto the substrate to be coated. Metallic or ceramic powders may be used to coat existing parts or to build up new parts on a mandrel. Since the plasma gas and surrounding atmosphere must be carefully controlled to give the desired coating properties, the entire process is usually highly automated and computer-controlled. Plasma spraying is widely used in the aerospace industry and in the manufacture of metallurgical and chemical process equipment. *See* WELDING AND CUTTING OF MATERIALS.

Plasma cutting is done in the transferred-arc mode where the arc is struck between the cathode and the workpiece which forms the anode (Fig. 2*b*). A large fraction of the arc power is liberated at the anode, thus melting the metal which is blown away by a high-velocity plasma jet. In both spraying and cutting operations, the plasma gas must be nonoxidizing to avoid rapid erosion of the tungsten cathode; oxidizing gases may be added downstream of the cathode in some applications.

Large scale. Large-scale plasma torches may also be operated in the transferred- or nontransferred-arc mode. In these torches, the internal electrodes are usually water-cooled cylinders. For nontransferred-arc operation, plasma gas is introduced tangentially between the two electrodes and the arc is struck across the gap. The arc length may be increased, to increase the plasma voltage, by inserting a number of electrically insulated water-cooled cylindrical spacers between the cathode and anode. Additional gas may also be injected through gaps between these spacers. The advantage of increasing arc voltage is to minimize the current required for a given arc power; this reduces power supply cost, resistance losses, and electrode erosion. Erosion of the electrodes is further reduced by the proper choice of electrode material, electrode cooling, and a combination of aerodynamic and magnetic rotation of the arc to reduce its residence time at any location.

Other applications. The nontransferred-arc operation is used for the heating of large quantities of gas which may then be used as a heating fluid or reactant. Torches of this type are used for the production of acetylene from various hydrocarbons, for the production of titanium dioxide pigment, and in a number of metallurgical operations, including ironmaking and the treatment of metallurgical dusts. An application with commercial potential is the superheating of blast furnace tuyere gases.

Transferred-arc operation is used for the heating of electrically conducting solids or liquids in which the product forms one electrode. High efficiencies can be obtained even at relatively low gas flow rates since only the internal torch electrode need be cooled. The main applications for this type of plasma torch are the smelting and remelting of ferroalloys, the melting of specialty steels and alloys, and ladle heating.

Arc remelting. In arc remelting applications, material which has been melted by conventional means is remelted for consolidation or purification. The process is used for highly reactive metals such as zirconium, titanium, and superalloys. A great variety of geometries and operating conditions are used, inlcuding vacuum-arc remelting (VAR), vacuum-arc double-electrode remelting (VADER), plasma-arc remelting (PAR), and electroslag remelting (ESR). In all cases, the product is collected into a water-cooled mold, and sometimes an ingot is continuously withdrawn. In most cases, the material to be remelted forms one electrode while the product forms the second. The differences between the systems are essentially the arc atmosphere (vacuum, plasma gas, or slag) and the geometry of electrodes used. Sometimes, particulate material is fed, and plasma torches are used as electrodes. In vacuum-arc double-electrode remelting, a vacuum arc is struck between two electrodes made of consolidated raw material, and the product drips into a water-cooled crucible below them. Because the material in the crucible is not directly heated, better control over the product macrostructure is possible.

Direct-current carbon arc furnaces. A direct-current arc is more stable than its alternating-current counterpart, and can, therefore, be operated at lower current and higher voltage by increasing the arc length. This reduces both the electrode diameter and the electrode consumption compared to alternating-current operation at similar powers. Tests have also shown that injecting gas through a hole drilled through the center of the electrode further increases stability and reduces wear. Powdered ore and reductants may be injected with this gas, reducing the need for agglomerating the arc furnace feed.

In most cases, direct-current carbon arc furnaces have one carbon electrode, with the product forming the second electrode. The current is usually removed from the furnace through a bottom constructed of electrically conducting material. Several direct-current plasma furnaces with powers ranging from 1 to 45 MW are in operation. R. J. Munz

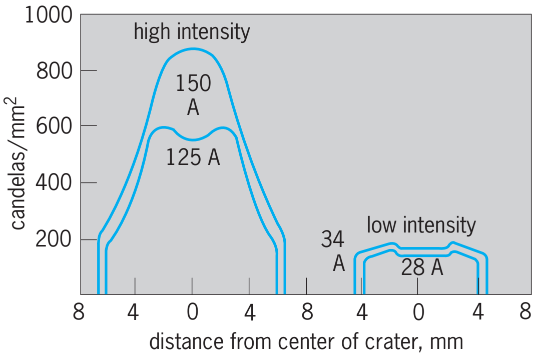
Bibliography. American Institute of Mining, Metallurgical, and Petroleum Engineers, Iron and Steel Society, *Electric Furnace Steelmaking*, ed. by C. R. Taylor, 1985; AIME, Iron and Steel Society, *Plasma Technology in Metallurgical Processing*, ed. by J. Feinman, 1987; AIME, *Proceedings of Electric Furnace Steel Conferences*, annually; A. Choudhury, *Vacuum Metallurgy*, ASM International, Materials Information Society, 1990; A. G. E. Robiette, *Electric Smelting Practice*, 1973.

Arc lamp

A type of electric-discharge lamp in which an electric current flows between electrodes through a gas or vapor. In most arc lamps the light results from the luminescence of the gas; however, in the carbon arc lamp a major portion of the light is produced by the incandescence of one or both electrodes. The color of the arc depends upon the electrode material and the surrounding atmosphere. Most lamps have a negative resistance characteristic so that the resistance decreases after the arc has been struck. Therefore some form of current-limiting device is required in the electric circuit. For other electric-discharge lamps. *See* VAPOR LAMP.

The carbon arc lamp was the first practical commercial electric lighting device, but the use of arc lamps is limited. In many of its previous functions, the carbon arc lamp has been superseded by the high-intensity mercury-vapor lamp. Arc lamps are now used to obtain high brightness from a concentrated light source, where large amounts of radiant energy are needed and where spectral distribution is an advantage. Typical uses of arc lamps are in projectors, searchlights, photography, therapeutics, and microscope lighting, and for special lighting in research. *See* MERCURY-VAPOR LAMP.

Carbon arc lamp. The electrodes of this lamp are pure carbon. The lamp is either open, enclosed, or an intensified arc with incandescence at the electrodes and some light from the luminescence of the arc. The open-arc form of the carbon arc is obsolete. It is unstable on a constant voltage supply, although on a constant current system the operating characteristics are good. In the enclosed type, a restricted supply of air slows the electrode consumption and secures a life of approximately 100 h. The intensified type used a small electrode, giving a higher-intensity output that is more white in color and more steady



Crater brightness distribution in forward direction for typical low- and high-intensity carbons. 1 mm = 0.04 in.

than either the open or enclosed types. The electrode life is approximately 70 h. In the high-intensity arc the current may reach values of 125–150 A with positive volt-ampere characteristics, and the lamp may operate directly from the line without a ballast. The **illustration** shows two basic differences in brightness characteristics of the low-intensity and high-intensity arc lamps. The brightness of the high-intensity arc is higher and depends markedly on current. *See* LUMINOUS INTENSITY.

Although carbon arc lamps are built for alternating-current operation, direct current produces a steadier and more satisfactory operation. To start the arc, the carbons are brought together for a short period and then separated to form an arc between the electrodes. Some forms of the lamps introduce a third electrode between the main electrodes for a brief period of time to start the lamp. As the carbon burns away, the arc gap must be adjusted for efficient operation. This may be done by manual adjustment; however, an automatic feed mechanism is more efficient and satisfactory.

Flame arc lamp. The flame arc lamp radiates light from the arc instead of from the electrode. The carbon is impregnated with chemicals, which are more volatile than the carbon and, when driven into the arc, become luminous. The chemicals commonly used are calcium, barium, titanium, and strontium, which produce their characteristic color of light as chemical flames and radiate efficiently in their specific wavelengths in the visible spectrum. Some flames use several chemicals, some for light and others for high arc temperature and steady arc operation. A variety of other chemicals may be introduced to produce radiation outside the visible spectrum for specific commercial or therapeutic purposes.

The flame arc is available in short-burning, medium-electrode-life type and long-burning type. With enclosure and automatic adjusting mechanism, and with consideration of the proper chemical composition, the electrode life of the long-burning type may be more than 100 h.

Metallic electrode arc lamp. In this type of lamp, light is produced by luminescent vapor introduced into the arc by conduction from the cathode. The positive electrode is solid copper, and the negative electrode is formed of magnetic iron oxide with titanium as the light-producing element and other chemicals to control steadiness and vaporization. These lamps are limited to direct-current use. A regulating mechanism adjusts the arc to a constant length.

Xenon short-arc lamp. This special type of lamp operates with the arc enclosed in a fused quartz bulb filled with xenon gas at a pressure of several atmospheres. The light emitted is continuous throughout the visible portion of the spectrum and approximates daylight in color. Lamps are available in sizes ranging from 30 to 30,000 watts electrical input with efficacies as high as 50 lumens per watt. The light is primarily emitted by the gas arc rather than from the hot electrodes. Large-wattage lamps require liquid cooling of the electrodes. Special power sources are required to operate these lamps. These lamps are used where it is important to have a source of very bright white (6000 K or 10,300°F color temperature) light radiated from a very small (0.4 in. or 10 mm or less) light source. Because the bulbs have internal operating gas pressures above 10 atm (150 lb/in.^{2} or 1.0 megapascal), the lamps must be in some form of protective enclosure. Protection is also needed from the large amount of ultraviolet radiation these lamps produce. G. R. Peirce

Bibliography. D. G. Fink and M. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; J. E. Kaufman and J. F. Christensen (eds.), *IES Lighting Handbook: Applications*, rev. ed., 1987.

Arc welding

A welding process utilizing the concentrated heat of an electric arc to join metal by fusion of the parent metal and the addition of metal to the joint usually provided by a consumable electrode. This discussion of the arc-welding process is divided into five general subjects: kinds of welding current, methods of welding, arc-welding processes, types of arc welders, and classification of electrodes. **Figure 1** presents a fundamental arc-welding circuit, and **Fig. 2** shows the elements of the weld at the arc. For the metallurgical aspects of welding *See* WELDING AND CUTTING OF MATERIALS.

Welding current. Electric current for the welding arc may be either direct or alternating, depending upon the material to be welded and the characteristics of the electrode used. The current source may

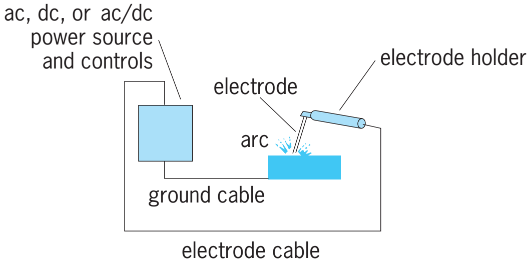


Fig. 1. Fundamental connections of electric equipment and workpiece for arc welding.

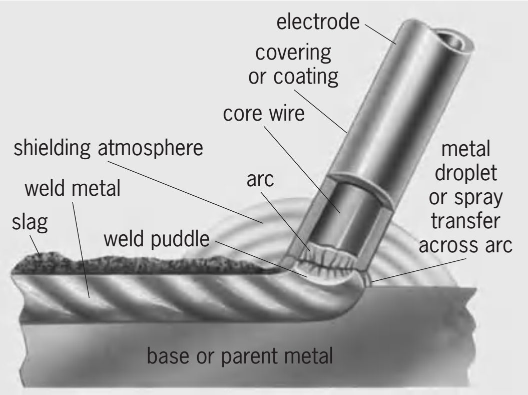


Fig. 2. Metallic welding arc.

be a rotating generator, rectifier, or transformer and must have transient and static volt-ampere characteristics designed for arc stability and weld performance.

On direct current the distribution of heat in the arc generally produces a concentration of heat at the positive terminal. The work generally has the larger mass and requires more heat for proper fusion; therefore the work is made positive and the electrode negative. This condition is known as straight polarity. When certain covered electrodes are used, the electrode is made positive and this condition is referred to as reverse polarity. With alternating current there is no polarity because of the current reversal each half cycle.

Methods. There are three basic welding methods: manual, semiautomatic, and automatic. Each has its benefits and limitations.

Manual welding. This is the oldest method, and though its proportion of the total welding market diminishes yearly, it is still the most common. Here an operator takes an electrode, clamped in a hand-held electrode holder, and manually guides the electrode along the joint as the weld is made. Usually the electrode is consumable; as the tip is consumed, the operator manually adjusts the position of the electrode to maintain a constant arc length.

Semiautomatic welding. This method is fast becoming the most popular welding method. The electrode is usually a long length of small-diameter bare wire, usually in coil form, which the welding operator manually positions and advances along the weld joint. The consumable electrode is normally motor-driven at a preselected speed through the nozzle of a hand-held welding gun or torch.

Automatic welding. This method is very similar to semiautomatic welding, except that the electrode is automatically positioned and advanced along the prescribed weld joint. Either the work may advance below the welding head or the mechanized head may move along the weld joint.

Welding processes. There are, in addition to the three basic welding methods, many welding processes which may be common to one or more of these methods. A few of the more common are described below.

Carbon-electrode arc welding. This process has been superseded, to a great extent, by other welding processes, but is still in limited use for welding ferrous and nonferrous metals. Normally, the arc is held between the carbon electrode and the work. The carbon arc serves as a source of intense heat and simply fuses the base materials together, or filler metal may be added from a separate source. A modification in which the arc is formed between two carbon electrodes is known as twin carbon arc welding, where the work does not form a part of the electrical circuit and only the heat of the arc is played against the work. The carbon electrodes are slowly consumed, and adjustments must be made while welding to maintain a constant arc length.

Shielded metal arc welding. This is the most widely used arc-welding process. A coated stick electrode, usually 14 in. (35 cm) long, is consumed during the welding operation, and therefore provides its own filler metal. The electrode coating burns in the intense heat of the arc and forms a blanket of gas and slag that completely shields the arc and weld puddle from the atmosphere. Its use is generally confined to the manual welding method.

Submerged-melt arc welding. In this process a consumable bare metal wire is used as the electrode, and a granular fusible flux over the work completely submerges the arc. This process is particularly adapted to welding heavy work in the flat position. High-quality welds are produced at greater speed with this method because as much as five times greater current density is used. Automatic or semiautomatic wire feed and control equipment is normally used for this process (**Fig. 3**). A modification of this process uses twin arcs (lead and trail) supplied from separate power sources but phased to reduce magnetic interaction of the arc stream.

Tungsten–inert gas welding. This process, often referred to as TIG welding, utilizes a virtually nonconsumable electrode made of tungsten (**Fig. 4**). Impurities, such as thorium, are often purposely added to the tungsten electrode to improve its emissivity for direct-current welding. The necessary arc shielding is provided by a continuous stream of chemically inert

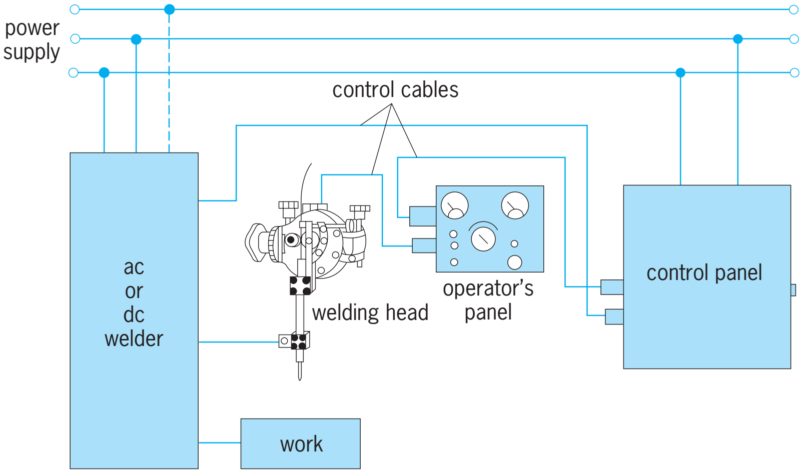


Fig. 3. Schematic diagram of automatic wire feed and control for arc welding.

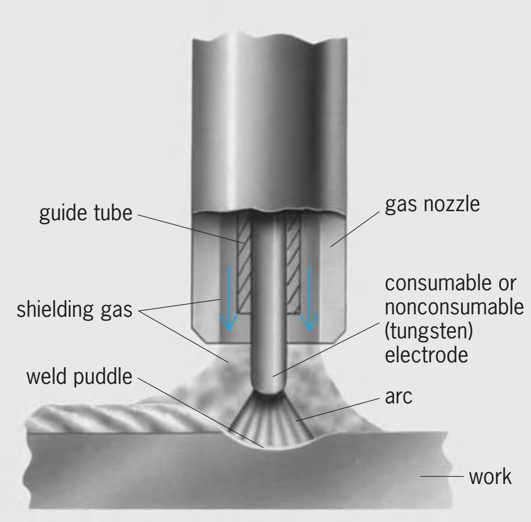


Fig. 4. Inert-gas-shielded metal welding arc.

gas, such as argon, helium, or argon-helium mixtures, which flows axially along the tungsten electrode that is mounted in a special welding torch. This process is used most often when welding aluminum and some of the more exotic space-age materials. When filler metal is desired, a separate filler rod is fed into the arc stream either manually or mechanically. Since no flux is required, the weld joint is clean and free of voids.

Metal–inert gas welding. This welding process, often referred to as MIG welding, saw its greatest growth in the 1960s. It is similar to the TIG welding process, except that a consumable metal electrode, usually wire in spool form, replaces the nonconsumable tungsten electrode. This process is adaptable to either the semiautomatic or the automatic method. In addition to the inert gases, carbon dioxide has become increasingly common as a shielding means. In MIG welding, the bare, but sometimes coated, wire is fed into the arc at current densities as high as 10 times those used for conventional shielded metal arc welding. Automatic wire feed and controls are used to drive the wire at speeds as high as 1000 in./min (2500 cm/min). This extremely versatile process produces high-quality welds at high speeds and with little or no spatter, thus eliminating cleaning.

Space does not allow describing all of the other welding processes, but a few of those most commonly referred to are electroslag welding, flux-cored welding, vapor-shielding welding, and plasma-arc welding.

Arc welders. An arc welder is an electric power generator, or conversion unit, for supplying electric current and voltage of the proper characteristics to a welding arc. Welders may be classified by the type of current supplied to the arc, alternating or direct. Most arc welders are designed for single operator use. Multiple-operator dc arc welding has limited application and is used primarily where operators are concentrated in a welding area.

Alternating-current welders. These are generally of the static type employing a two-winding transformer and a variable inductive reactor, which is connected in series with the welding circuit. The reactor may be combined with the transformer to provide adjustable leakage reactance between windings (**Fig. 5**). The transformer isolates the welding circuit from the primary line, and steps down the line voltage to an open-circuit voltage of 80 V or less for manual welding and 100 V or less for automatic welding.

The ac arc is extinguished each half cycle as the current wave passes through zero. The reactor, by reason of its stored inductive energy, provides the voltage required for reignition of the arc at each current reversal.

Several types of reactor construction are employed, including a movable core with variable air gap, a movable coil to vary leakage between windings, and a saturable reactor with dc control in combination with a tap changer.

Direct-current welders. These may be of the rotating dc generator or static rectifier type, with either constant-current or constant-voltage characteristics.

The dc generator type is usually driven by a directly coupled induction motor or internal combustion engine. The conventional design employs a generator with a combination shunt-field rheostat control and a differentially compounded series field tapped to give an adjustable welding current with drooping volt-ampere characteristic.

Other schemes of current adjustment are used, most of which employ the principle of field control. These include excitation from a third brush utilizing the effect of armature reaction or a movable core to vary the leakage between poles of the generator. Because of the extreme and rapid fluctuation of the arc voltage during welding, the welding generator must be designed for rapid response to load changes.

The rectifier-type arc welder is a combination of transformer, variable reactor, and rectifier. The transformer and reactor designs are similar to those employed in ac welders, except that dc welders are generally used on a three-phase supply (**Fig. 6**). Some single-phase welders of this type are used, particularly where a dual ac/dc output is required. The rectifiers are the selenium dry plate or silicon diode full-wave bridge type.

Both the rotating generator and rectifier types are also available with constant-voltage characteristics. On these units the output voltage is adjustable, and the welders are designed for either flat or adjustable droop volt-ampere characteristics. In addition, a direct-current inductor is often included in

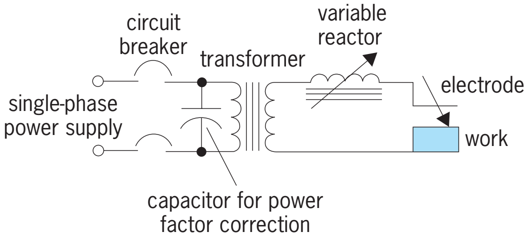

circuit breaker
variable reactor
transformer
electrode
single-phase power supply
work
capacitor for power factor correction

Fig. 5. Schematic diagram of a typical alternating-current arc welder.

-S

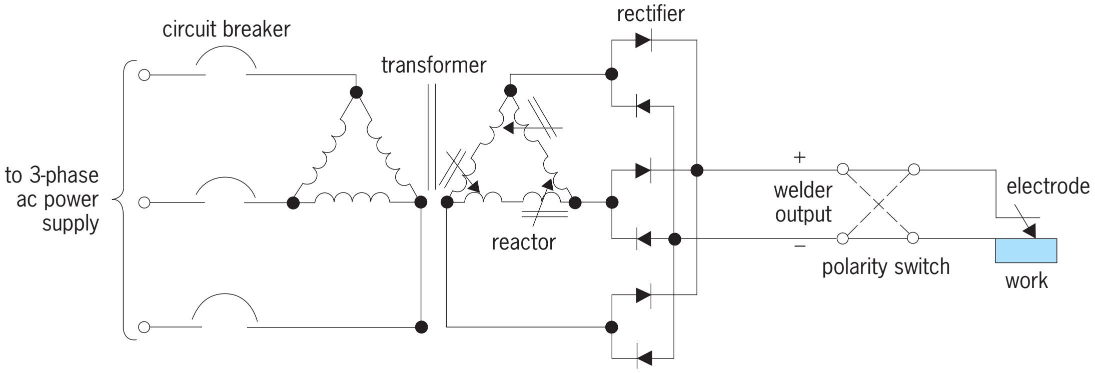


Fig. 6. Schematic diagram of a typical direct-current arc welder.

the welder output to tailor the welder's response to the process requirements. The constant-voltage type of welder is most often used with the MIG welding process where the welding current is adjusted by varying the speed of the electrode wire feed. An addition to the continually expanding line of constant-potential type of welders is a direct-current arc welder with a pulsing output superimposed on a relatively smooth background characteristic. This welder provides the power source requirements imposed by a new MIG process, the applications of which have not yet been fully explored.

Multiple-operator dc welding. This method employs one or more power units with approximately 80 load volts and relatively flat voltage regulation. The units operate in parallel to supply a number of resistor outlets, one for each operator. The resistor units provide ballast impedance and a means for adjusting the welding current. This system effects appartus economy by utilizing the load diversity of multiple welding arcs.

Electrodes. An arc-welding electrode is a wire or rod of current conducting material, with or without a covering or coating, forming one terminal of the electric arc used in arc welding. Electrodes may be classified as nonconsumable and consumable. Nonconsumable electrodes (carbon, tungsten) are consumed at a very low rate and add no filler metal to the weld. Consumable electrodes melt with the heat of the arc, and the metal is transferred across the arc to form part of the weld metal.

Consumable electrodes may be in the form of rods 9–18 in. (22.5–45 cm) long or continuous wire. Practically all welding electrodes for manual welding are coated or covered. This covering (usually extruded on the core wire) serves a number of important functions. The gas and slag shielding produced by the covering protects the molten metal from the atmosphere during the transfer across the arc and after deposit.

The materials in the electrode covering are intended to perform one or more of the following: (1) purify the weld metal; (2) control melting rate, penetration, bead shape, and cooling rate; (3) add alloying elements to the weld metal; (4) stabilize the arc; and (5) provide insulating coating to reduce shock hazard.

The covering of some electrodes includes powdered iron to increase the speed of welding by stepping up the rate of metal deposit.

Welding electrodes have been classified according to tensile strength of the deposited weld metal, type of covering, welding position, and type of current. The electrodes are identifiable by electrode classification numbers under specifications prepared jointly by the American Welding Society and the American Society for Testing and Materials. *See* RESISTANCE WELDING.

Emil F. Steinert

Bibliography. American Welding Society, *Arc Welding Safety and Health*, 1982; R. Bakish and S. S. White, *Handbook of Electron Beam Welding*, 1964; A. L. Phillips (ed.), *Welding Handbook*, 4th ed., 1957.

Arcellinida

An order of Lobosia. The shell (test) of these protozoa has a single, well-defined aperture through which slender fingerlike pseudopodia (lobopodia) can be extended. The test often has an internal chitinous layer known as a glycocalyx and an external layer composed of secreted siliceous elements, organic hollow platelets, or collected sand grains and other matter cemented together by an adhesive secretion. The secreted platelets are synthesized within membranous vesicles in the cytoplasm, where they first appear as dense granules. These granules are transported to the surface of the cell, where they are deposited on the outer surface of the cell membrane and mature into the alveolate, closely assembled platelets.

Most arcellinidans are uninucleate, but *Arcella* is binucleate. During asexual reproduction certain species prefabricate their tests before fission to some extent. During growth, the organism may assemble materials by secretion or ingestion in preparation for new test construction. At fission, when a daughter cell is produced, such materials pass through the aperture into the extruded protoplasm, where they are moved to the surface and cemented to the outer wall of the new test. The old test is kept by the sister organism. Cyst walls may be formed inside the test; the aperture is often plugged with debris before encystment. Food consists of bacteria and smaller protistans. The order includes *Arcella, Centropyxis, Cochliopodium, Difflugia*, and many other genera. *See* LOBOSIA; PROTOZOA; SARCODINA; SARCOMASTIGOPHORA.

O. Roger Anderson

Arch

A structure, usually curved, that when subjected to vertical loads causes its two end supports to develop reactions with inwardly directed horizontal components. The commonest uses for an arch are as a bridge, supporting a roadway, railroad track, or footpath, and as part of a building, where it provides a large open space unobstructed by columns. Arches are usually built of steel, reinforced concrete, or timber.

The designations of the various parts of an arch are given in **Fig. 1**. The skewback is the abutment of pier surface upon which the arch rests. Because the arch springs from its skewback, the intersection of the arch and the skewback is called the springing line. The upper surface of the arch is the extrados, the inner surface the intrados. The line passing through the center of gravity of each section is the arch axis. The crown is the highest section.

The advantage of the arch over a simple truss or beam of the same span is the reduction in the positive moment by the negative moment resulting from the horizontal thrust at the supports. The arch rib is subjected to large axial forces and small bending moments. To minimize the moments, the center line of the rib should closely approximate the funicular polygon for dead load, plus perhaps some portion of the live load. For a uniformly distributed load the funicular polygon is a parabola.

The principal dimensions of the center line of the arch are span and rise. These may be dictated by conditions at the site, if the structure is a bridge, or by architectural requirements, if the arch is to form part of a building. A rise of from one-fifth to one-third of the span may prove economical.

On the basis of structural behavior, arches are classifed as fixed (hingeless), single-hinged, two-hinged, or three-hinged (**Fig. 2**). An arch is considered to be fixed when rotation is prevented at its supports. Reinforced concrete ribs are almost always fixed. For long-span steel structures, only fixed solid-rib arches are used. Because of its greater stiffness the fixed arch is better suited for long spans than hinged arches.

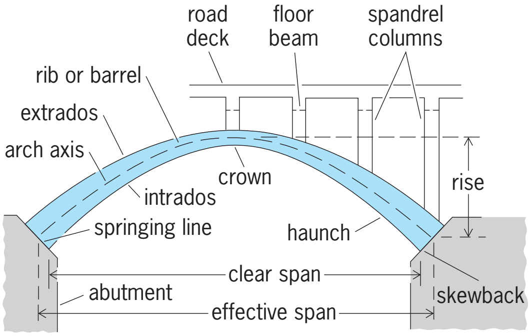

road deck
floor beam
spandrel columns
rib or barrel
extrados
arch axis
crown
intrados
rise
springing line
haunch
clear span
abutment
skewback
effective span

Fig. 1. An open-spandrel, concrete, fixed-arch bridge.

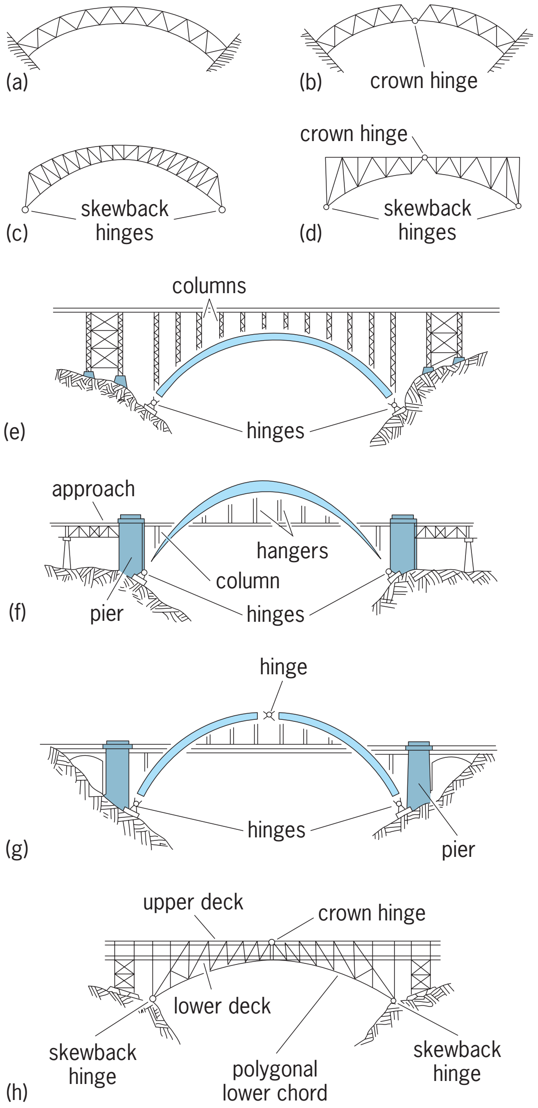

(a)
(b)
crown hinge
crown hinge
skewback hinges
skewback hinges
(c)
(d)
columns
(e)
hinges
approach
hangers
column
(f)
pier
hinges
hinge
(g)
hinges
pier
upper deck
crown hinge
lower deck
skewback hinge
polygonal lower chord
skewback hinge
(h)

Fig. 2. Various types of bridge arches. (*a*) Fixed. (*b*) Single-hinged. (*c*) Two-hinged. (*d*) Three-hinged. (*e*) Parallel curve rib arch. (*f*) Half-through, two-hinged, crescent-rib arch. (*g*) Parallel curve, three-hinged rib arch. (*h*) Double-deck spandrel-braced arch. (***After G. A. Hool and W. S. Kinne, Movable and Long-span Steel Bridges, 2d ed., McGraw-Hill, 1943***)

A hinge introduced into the arch rib produces an axis of zero moment forcing the thrust line (funicular polygon) to pass through the hinge. An arch hinged at the crown is a single-hinged arch; it is rarely used because it has no distinct advantages. The two-hinged arch has hinges at each skewback. Because foundations for two-hinged arches receive thrust forces only, the abutments are more easily planned. A hinge introduced at the crown of a two-hinged arch is statically determinate, and no stresses result from temperature, shrinkage, or rib-shortening effects. *See* STRUCTURAL ANALYSIS.

Because the horizontal component of an arch reaction is large, excellent foundation conditions must exist at the site. Abutments constructed on a steeply sloping gorge of strong, sound rock transmit the

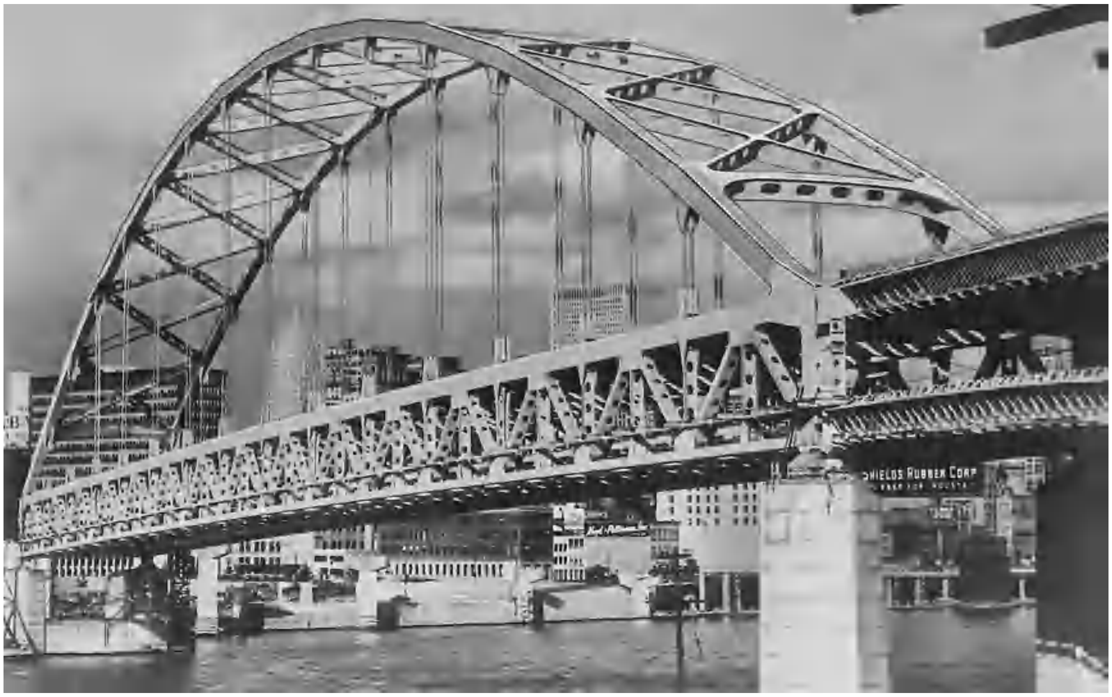


Fig. 3. Fort Pitt highway bridge, a double-deck tied-arch bridge across Monongahela River at Pittsburgh, Pennsylvania. Unusual are a truss instead of girders as ties, and box sections as arch ribs. (*Lead Industries Ass.*)

thrust directly to the bedrock. Abutments on rock can usually be designed to resist the moment developed at the skewback of a fixed-end arch. Foundations on earth or piles may rotate and relieve part of the assumed restraint. A slight yielding of foundations usually causes no serious harm to three-hinged arches. If suitable foundation conditions are not available, the horizontal reaction component may be provided by a tie between skewbacks. An arch so constructed is classified as a tied arch (**Fig. 3**). The tied arch is usually more expensive than the two-hinged type.

Concrete arches. Concrete is relatively weak in tension and shear but strong in compression and is therefore ideal for arch construction. In Fig. 1, the arch proper consists of two or more solid ribs or a single solid barrel whose width is equal to that of the bridge. The solid-rib arch supports its roadway by means of a system of beams and spandrel columns. The rib type is particularly desirable for skew crossings because each rib may be arranged to act independently and is designed as a right arch. Ribs are usually interconnected by cross struts. Hollow box sections have been used in place of solid ribs to form a hollow-rib arch.

The roadway in the barrel-type arch is supported on an earth fill, which is confined by spandrel walls, extending from the barrel to the road deck. The barrel type, frequently called a spandrel-filled arch, is generally less economical of material than the rib type. Where for architectural reasons a solid wall effect is desired for a low, flat arch, the spandrel-filled design is often adopted. Placing solid curtain walls over the spandrel openings of an open-spandrel arch achieves the same effect.

Because the rib or barrel is subject to compression and some bending, it is designed as an eccentrically loaded column, with reinforcement placed near both the intrados and extrados. Steel percentages should be kept small.

Precast reinforced concrete arches of the three-hinged type have been used in buildings for spans up to 160 ft (49 m). *See* CONCRETE.

Steel arches. Steel arches are solid-rib or braced-rib arches. Solid-rib arches usually have two hinges but may be hingeless. A parallel curved rib (with a constant depth throughout its length) is the most commonly used form of plate girder or solid webbed rib.

The braced-rib arch has a system of diagonal bracing replacing the solid web of the solid-rib arch. The world's longest arch spans are both two-hinged arches of the braced-rib type, the Sidney Harbor Bridge in Australia and the Bayonne Bridge at Bayonne, New Jersey (**Fig. 4**), which have spans of 1650 and 1652 ft (502.9 and 503.5 m), respectively. Both are of the half-through type. Other classifications according to the method by which the roadway is supported by the arch proper are through arches and deck arches. Through and half-through arches are usually of the rib type.

The spandrel-braced arch is essentially a deck truss with a curved lower chord, the truss being capable of developing horizontal thrust at each support. This type of arch is generally constructed with two or three hinges because of the difficulty of adequately anchoring the skewbacks.

Wood arches. Wood arches may be of the solid-rib or braced-rib type. Solid-rib arches are of laminated construction and can be shaped to almost any

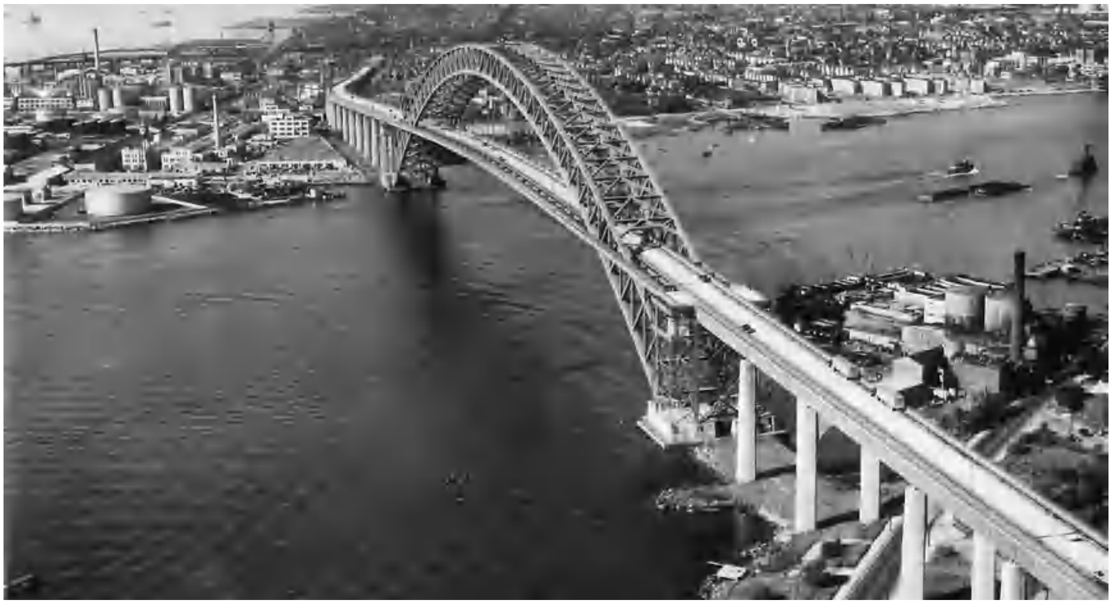


Fig. 4. Bayonne Bridge across the Kill Van Kull in New Jersey. (*Port of New York Authority*)

required form. Arches are usually built up of nominal 1- or 2-in. (2.5- or 5-cm) material because bending of individual laminations is more readily accomplished. For wide members a lamination may consist of two or more pieces placed side by side to form the required width. All end-to-end joints are scarfed and staggered. Although the cross section is usually rectangular, I-beam sections have been used. Because of ease in fabrication and erection, most solid-rib arches are of the three-hinged type. This type has been used for spans of more than 200 ft (60 m).

The lamella arch has been widely used to provide wide clear spans for gymnasiums and auditoriums. The wood lamella arch is more widely used than its counterpart in steel. The steel lamella roof for the civic auditorium in Corpus Christi, Texas, has a clear span of 224 ft (68 m). The characteristic diamond pattern of lamella construction provides a unique and pleasing appearance. Basically, lamella construction consists of a series of intersecting skewed arches made up of relatively short straight members. Two members are bolted, riveted, or welded to a third piece at its center. These structures are erected from scaffolds because no supporting action develops until a large section is in place. Construction starts at both sills, moving up until the sides meet in the center.

The horizontal thrust of the lamella roof must be resisted by steel tie rods or wood ties at wall height or by buttresses. The thrust component developing in the longitudinal direction of the building may be resisted by ties extending the length of the building or by the roof decking. *See* BRIDGE; ROOF CONSTRUCTION; TRUSS.

Charles N. Gaylord

Masonry skeletons. Masonry skeletons in the form of arches, vaults, and domes have supported bridges and buildings since ancient times. An arch, dated 1400 B.C., was discovered among Babylonian ruins. It was built of brick with clay mortar joints. Although other civilizations understood the arch concept, the Romans significantly advanced the art of masonry skeleton construction using stone and brick. Extant Roman masonry skeletons include the Pont du Gard near Nimes, France, built about 18 B.C.; the Ponte di Traiano, Alcantara, Spain, built about A.D. 105; and the Hagia Sophia in Istanbul, Turkey, built about A.D. 535. *See* BRICK; CLAY; MASONRY; STONE AND STONE PRODUCTS.

The Pont du Gard comprises three tiers of stone arches that support two roadways and an aqueduct at a height of 50 m (160 ft) above the Gard River. The longest clear arch span of the lower tiers is 25 m (81 ft), and the spans of the top story arches are 5 m (16 ft). The cut stones were laid dry, except for those forming the water channel, which were set in mortar.

The Ponte di Traiano comprises six successive stone arches of up to 30 m (100 ft) span supporting the roadway at about 62 m (200 ft) above the Tagus River. One arch was destroyed on two occasions by military forces but has been restored.

The Hagia Sophia is an assemblage of partial domes surrounding the main dome of 31 m (102 ft) clear diameter. The domes are mostly brick covered by stucco on the exterior and mosaic on the interior. The Hagia Sophia was damaged during a severe earthquake in A.D. 558 but was repaired by 563.

As the virgin forests of Europe were exhausted, masonry construction gained importance for reasons of economy, architecture, and permanence. Canal, and later railroad network, construction during the eighteenth and nineteenth centuries necessitated many bridges. For example, large-scale masonry arch construction was started in the United Kingdom about 1750 and continued through the railway construction period, ending about 1910. In the United Kingdom there are some 40,000 highway and 33,000 railway masonry arch spans, most of brick. Many have been used continuously and subjected to much heavier loading than when originally put into service, and have received only minimal maintenance. Large masonry arch bridges were also built on the Continent. Examples are the stone arch bridge completed in 1903 at Plauen, Germany, and the stone arch viaduct for the Rhaetian Railway near Filisur, Switzerland. The stone arch bridge has a clear span of 91 m (295 ft) with a low rise of one-fifth the span, and is built of local slate. The railway viaduct comprises segmental arches on very tall piers and was completed in 1904.

In the eastern and midwestern United States, virgin timber was not exhausted until after inexpensive steel and reinforced concrete became available; hence the need for labor-intensive masonry arch bridges did not develop except for early railroad viaducts. Two examples of stone arch railroad viaducts are the Carrollton Viaduct at Baltimore, Maryland, and the Starrucca Viaduct at Lanesboro, Pennsylvnia. The Carrollton Viaduct was built in 1829 of dressed granite with a 25-m (80-ft) main span and a smaller side span. The Starrucca Viaduct, completed in 1848, has 17 segmental arch spans of 16 m (51 ft) on slender piers about 34 m (110 ft) high. Only the outer voussoirs were dressed. *See* TRESTLE.

The masonry arch can provide structure and beauty, is fireproof, requires comparatively little maintenance, and has a high tolerance for foundation settlement and movement due to other environmental factors. Most arches are curved, but many hectares (acres) of floor in highrise office and public buildings are supported by hollow-tile jack (flat) arches (**Fig. 5**). The arch shape selected for a particular structure depends on architectural considerations, but must be such that the statical requirements for the anticipated loading are satisfied.

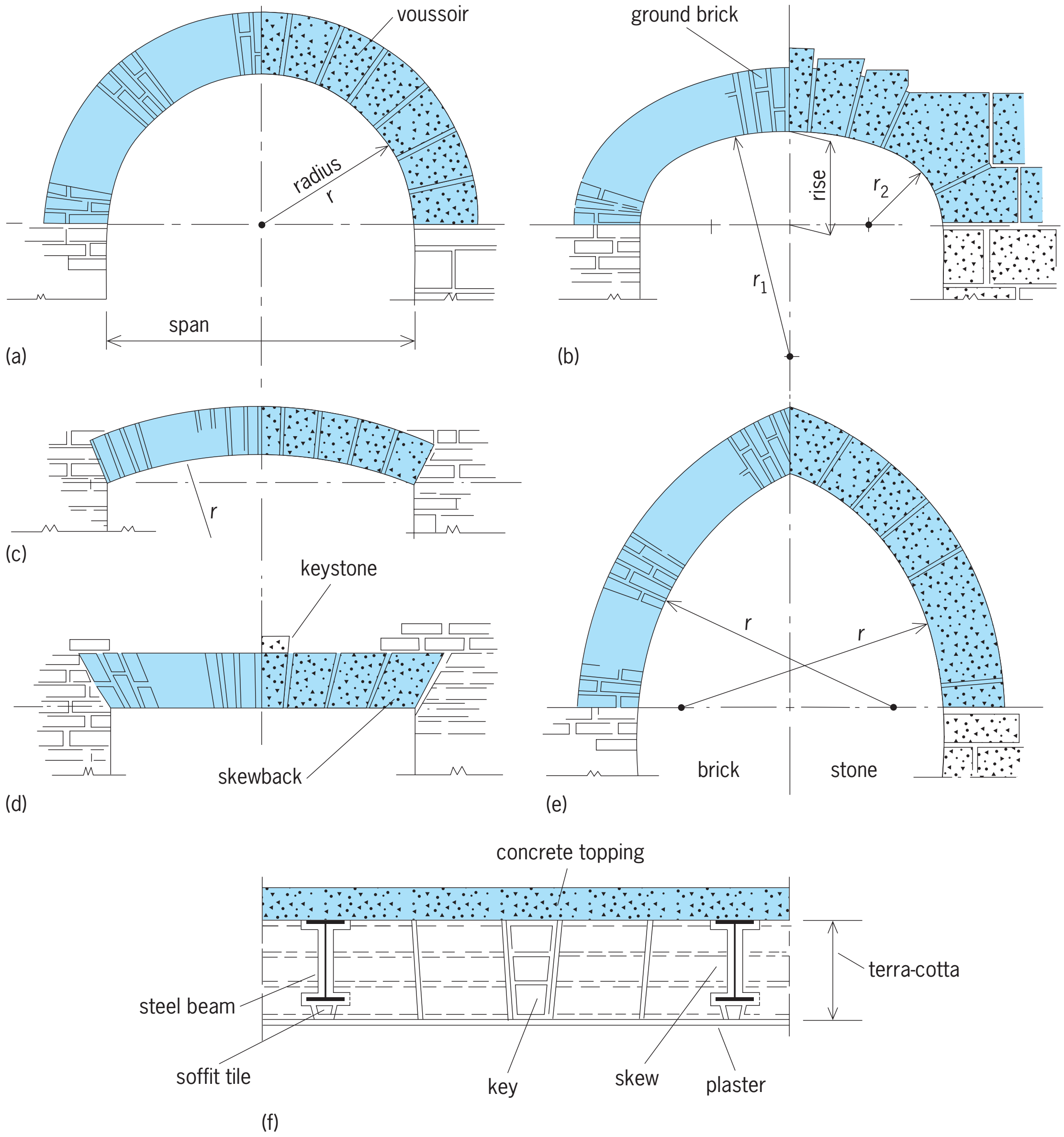


Fig. 5. Common forms of masonry arches. (*a*) Roman (semicircular). (*b*) Elliptical. (*c*) Segmental. (*d*) Jack or flat. (*e*) Gothic. (*f*) Hollow-tile jack. (*After C. G. Ramsay and H. R. Sleeper, Architectural Graphic Standards, 9th ed., John Wiley, 1994*)

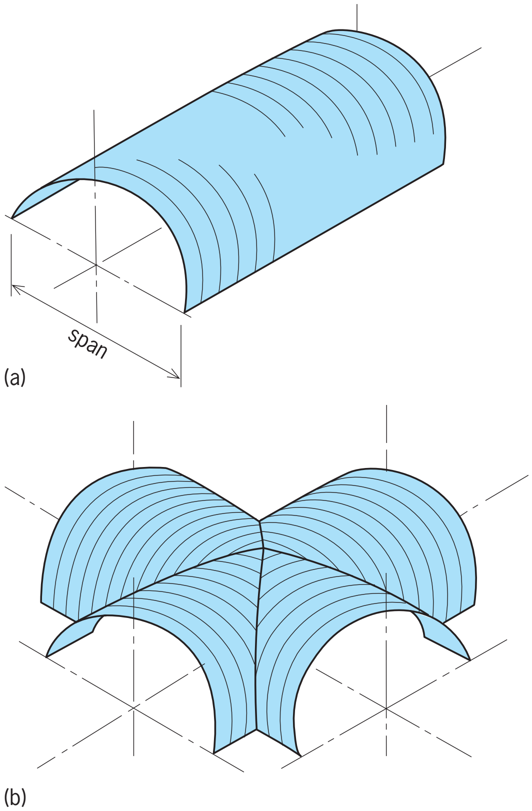


Fig. 6. Types of vaults. (*a*) Individual or barrel. (*b*) Quadripartite. (*After J. Heyman, The Stone Skeleton, Cambridge University Press, 1995*)

If a curved arch is wide (dimension normal to span), the arch is referred to as a barrel arch or vault. The vault cross section may have several different shapes (Fig. 5). Contiguous vaults may be individual, may intersect, or may cross (**Fig. 6**). A four-part vault is termed quadripartite. Contiguous quadripartite vaults that are supported at the corners by columns are masonry skeletons of large cathedrals. For example, the cathedral of Notre-Dame in Paris, France, was built as a sexpartite vault about A.D. 1200.

Stone for masonry skeletons is cut from three classes of rock; igneous (granite, traprock), metamorphic (gneiss, slate, quartzite), and sedimentary (limestone, sandstone). The physical properties vary widely within each class, and the properties of the stone actually used should be determined by test. The minimum compressive strength can range from 3000 pounds per square inch (21 megapascals) for sandstone to 15,000 psi (103 MPa) for granite. The corresponding upper limits are 20,000 psi (138 MPa) and 30,000 psi (207 MPa). *See* IGNEOUS ROCKS; METAMORPHIC ROCKS; SEDIMENTARY ROCKS.

The primary requirements for brick as a structural material are compressive strength and weathering resistance. Building brick (common brick) is now graded as Severe-, Moderate-, and No-Weathering (SW, MW, and NW) by the American Society for Testing and Materials (ASTM). The minimum acceptable compressive strength for a single brick loaded over the full bedding area varies from 2500 psi (17 MPa) for SW down to 1250 psi (9 MPa) for NW. An index of weathering durability is water absorption. The maximum acceptable absorption of brick submerged 24 h in cold water is 16% by weight for SW and unlimited for NW.

Hollow clay tiles (terra-cotta) are made by extruding a clay mixture through a die, cutting pieces from the extrusion, and then heating them to a high temperature (burning). Terra-cotta for floor arches was made semiporous in order to improve fire resistance. When the arches were loaded in the direction of the cells, the compressive strength varied from 2000 to 12,000 psi (14 to 83 MPa). Water absorption was not considered an issue for floor arches.

Masonry skeletons depend on a general state of compression for stability. Old brick or stone masonry cannot reliably transmit tensile stress, even when laid with mortar joints. With time the tensile bond between the mortar and masonry is lost because of environmental factors. Steel reinforcment can be placed in the joints between units to improve the tensile capacity of the masonry. However, the long-term viability of such construction has yet to be determined. The compressive stresses in the skeleton are usually low compared to the crushing strength of the masonry, but are sufficient to develop friction and prevent slip of adjacent units.

Until the middle of the eighteenth century the geometrical proportions of masonry skeletons were based on complex rules (closely guarded by their possessors), applied by an architect-engineer (master builder) who had served a long apprenticeship. These rules usually lead to skeletons which are in a compressive state due to gravity loads and wind and which have survived for centuries. Analytical and experimental work in the latter part of the eighteenth century led to a limit-state approach to design, but most construction still followed the old rules.

At the beginning of the nineteenth century, C. L. M. Navier advocated elastic structural analysis which offered engineers a design tool based on the working-stress philosophy. The funicular polygon of the actions on the skeleton could now be computed from statics for any arch, assuming elastic behavior and geometrical boundary conditions at the springings. This force polygon (pressure line, thrust line) must lie within the kern of the arch cross sections throughout, for all combinations of actions, if the voussoirs are not to separate at the outer or inner curves of the arch (extrados or intrados). For a solid rectangular arch cross section, the eccentricity of the pressure line at the kern limit is $e = d/6$ (**Fig. 7*a***). Because the joints between the voussoirs will not separate until $e = d/2$ (assuming no local crushing at the joints), the arch is said to have a geometrical factor of safety (GFS) = 3 for $e = d/6$. Masonry arches are usually proportioned to have GFS > 2.

Arches are often designed as if fixed at the springings. In reality, no arch is fully fixed-ended, and a slight change in the geometrical boundary conditions influences the location of the pressure line. The abutments of masonry arches displace an unknown amount in time, and the masonry is subjected

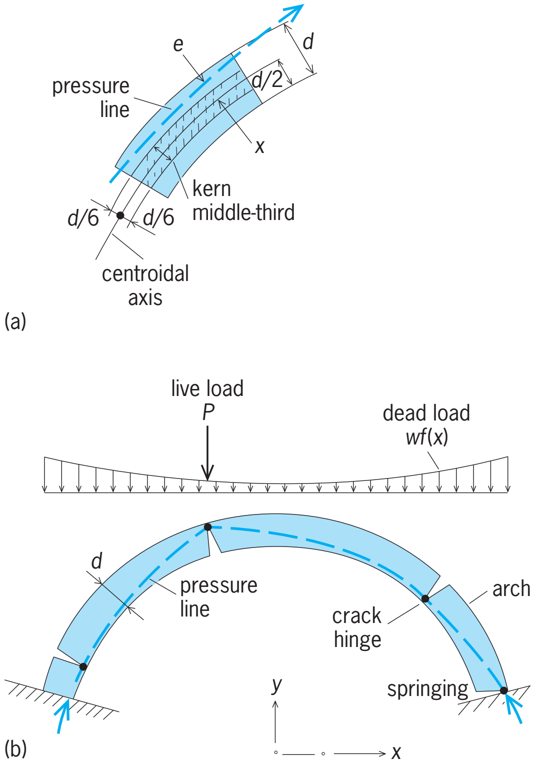


Fig. 7. Pressure line of an arch. (*a*) Notation for arches. (*b*) Arch collapse mechanism for a masonry arch subjected to a point live load.

to volume and material property changes due to the environment. Because of these effects, the validity of the conventional elastic analysis has been questioned, especially regarding long-term behavior, with the return to the limit state approach, especially for assessment.

Pipard and associates developed a limit analysis (about 1935–1962) for curved arches using a mechanism approach and assuming that the masonry has (1) no tensile strength, (2) infinite compressive strength, (3) infinite elastic modulus, and (4) no slip between adjacent voussoirs. J. Heyman related this limit-state approach to the plasticity theorems.

Analyses have been developed which are intended to predict curved masonry arch behavior prior to collapse. In 1990 a method was proposed for "fixed arches" based on the development of three crack hinges until the collapse load is reached when the fourth crack hinge develops. Procedures are available for analyzing inelastic masonry arches utilizing nonlinear finite element techniques.

Charles Birnstiel

Bibliography. J. Henrych, *The Dynamics of Arches and Frames*, 1981; J. Heyman, *The Stone Skeleton: Structural Engineering of Masonry Architecture*, Cambridge University Press, Cambridge, U.K., 1995; F. Leonhardt, *Brucken-Bridges*, MIT Press, Cambridge, MA, 1984; C. G. Ramsey and H. R. Sleeper, *Architectural Graphic Standards*, John Wiley, New York, 1994; W. Schueller, *Horizontal-Span Building Structures*, John Wiley, New York, 1983.

Archaea

A group of diverse microscopic prokaryotic organisms that constitute one of the three domains of life. Comparison of the sequences of small subunit ribosomal ribonucleic acid (rRNA) from a large variety of living organisms indicates that they may be divided into three major groups: Archaea, Bacteria, and Eukaryota (**Fig. 1**). *See* BACTERIA; EUKARYOTA.

Archaea are often found in some of the most extreme environments on this planet; specifically, anaerobic, hypersaline, extremely cold, or extremely high temperature environments. They comprise about one-third of the prokaryotic biomass in coastal Antarctic water, providing food for many marine animals. A few are symbionts in the digestive system of animals; however, none are known to cause disease in humans or animals.

General Characteristics

The Archaea are quite distinct from bacteria and eukaryotes, and within the domain there is great diversity.

Morphology. Archaea may be spherical, rod-shaped, spiral, plate-shaped, or pleomorphic (variable in shape) [**Fig. 2**]. Some are single cells, whereas others form filaments. They range in diameter from 0.1 micrometer to more than 15 μm, and some filaments grow longer than 200 μm. Although there is considerable variation in archaeal cell wall structure, all members lack peptidoglycan in their cell walls. They stain either gram-positive or gram-negative. The gram-positive archaea have a cell wall with a single thick homogeneous layer resembling that in gram-positive bacteria and thus stain gram-positive. The gram-negative archaea lack the outer membrane and peptidoglycan network and thus stain gram-negative.

Membrane biochemistry. Archaeal membrane lipids are branched-chain hydrocarbons attached to glycerol by ether links, rather than fatty acids connected by ester links, as in bacteria and eukaryotes. The ether-linked lipids confer greater structural rigidity and chemical stability, enabling these organisms to live in a more extreme environment. The membranes of extreme thermophiles (organisms that thrive at high temperatures) consist almost completely of monolayers of tetraethers, which gives these archaea thermal stability.

Physiology. Many archaea are extremophils, requiring extreme conditions of temperature, pH, and salinity to survive. Some are mesophiles, and grow optimally at intermediate temperatures ranging from 68 to 113°F (20 to 45°C); others are hyperthermophiles that can live and grow in environments above 212°F (100°C). They can be aerobic (require oxygen for growth), facultatively anaerobic (grow equally well with or without oxygen), or strict anaerobes (grow in the absence of oxygen). Some archaea obtain energy from inorganic compounds (chemolithoautotrophs), whereas others derive their nutrients from organic substances.

Reproduction is via budding, fragmentation, binary fission, or unknown mechanisms.

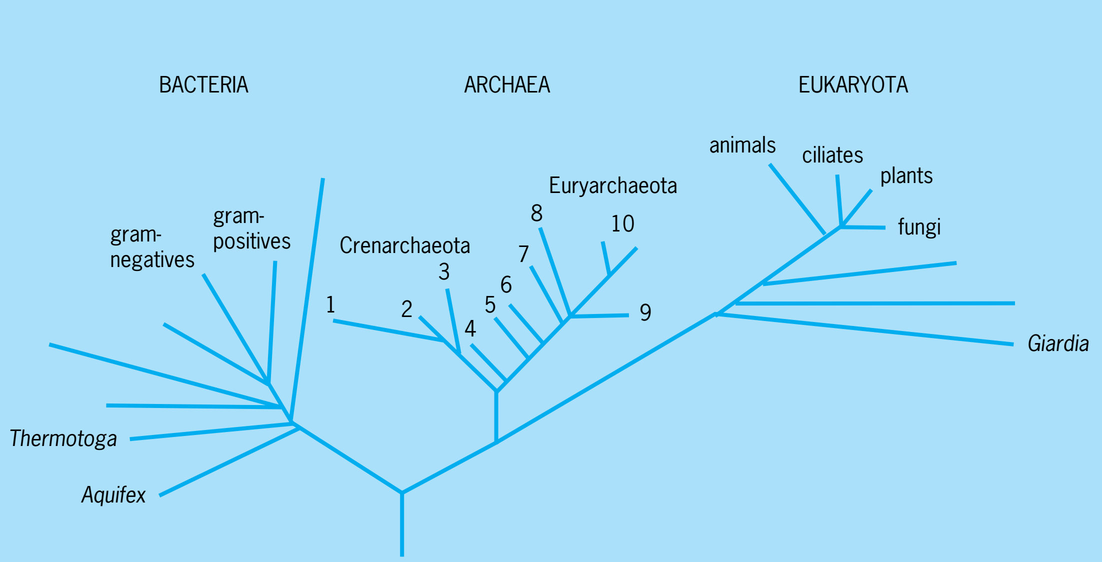


Fig. 1. Phylogenetic relationships among the three domains of living organisms: Bacteria, Archaea, and Eukaryota. The phylogenetic positions of representative archaea are indicated by numbers: (1) *Sulfolobus;* (2) *Pyrodictium;* (3) *Thermoproteus;* (4) *Methanopyrus;* (5) *Thermococcus;* (6) *Methanococcus;* (7) *Methanobacterium;* (8) *Methanosarcina;* (9) *Archaeoglobus;* (10) halophiles. (*After O. Kandler, Where next with the archaebacteria?, Biochem. Soc. Symp., 58:195–207, 1992*)

Metabolism. Although many aspects of archaeal metabolism appear similar to other organisms, the archaea differ with respect to glucose catabolism, pathways for CO_2 fixation, and the ability of some to synthesize methane. Archaea use a number of unique coenzymes in their metabolism, some of which are associated with methanogenesis (the production of methane: other coenzymes are unknown in the bacteria and eukaryotes). The enzymes involved in archaeal protein and cell wall synthesis are sufficiently different from those used by bacteria and are not inhibited by antibiotics such as penicillin, tetracycline, and streptomycin, which are effective against most bacteria.

A few archaea ferment sugars via an unusual metabolic pathway. Most bacteria and eukaryotes catabolize glucose by the Embden-Meyerhof-Parnas (EMP) pathway, which involves phosphorylated intermediates and yields energy-rich end products. Archaea use a nonphosphorylated pathway termed the pyroglycolytic pathway, which generates less energy for each sugar molecule catabolized than the EMP pathway. However, most of the EMP enzymes are present in archaea, since they synthesize sugars by reversing the reactions of the EMP pathway. It is unknown why glycolysis evolved differently in bacteria and archaea.

Very little is known about biosynthetic pathways in the archaea. Preliminary data suggest the synthetic pathways for amino acids, purines, and pyrimidines are similar to those in other organisms. *See* BACTERIAL PHYSIOLOGY AND METABOLISM.

Genetics and molecular biology. Although archaea and bacteria share some genetic similarities, there are important differences. For example, the initial amino acid in their polypeptide chains, coded by the AUG start codon, is methionine (as it is in eukaryotes), whereas bacteria use *N*-formyl methionine. Archaea are also distinguished from both bacteria and eukaryotes by their transfer ribonucleic acid (tRNA), ribosomes, elongation factors, and RNA polymerases. *See* GENETIC CODE; MOLECULAR BIOLOGY; RIBONUCLEIC ACID (RNA); RIBOSOMES.

Taxonomy

Based on rRNA data, the archaea are divided into two phyla: the Euryarchaeota and Crenarchaeota. The euryarchaeotes occupy many different ecological niches and have a wide variety of metabolic patterns. The crenarchaeotes are thought to resemble the ancestor of the archaea, and almost all of the well-characterized species are thermophiles or hyperthermophiles. *See* CLASSIFICATION, BIOLOGICAL.

Crenarchaeota. Most of the crenarchaeotes are extremely thermophilic, and many are acidophiles (organisms that have a preference for acidic environments) and sulfur-dependent. Almost all are strict anaerobes. The sulfur may be used as an electron acceptor in anaerobic respiration or as an electron source by lithotrophs (organisms that use inorganic substrates as electron donors in energy metabolism). They grow in geothermally heated waters or soils that contain elemental sulfur. Examples include the sulfur-rich hot springs in Yellowstone National Park (the orange color of these springs is due to the carotenoid pigments of these archaea), and the waters surrounding areas of submarine volcanic activity. Genera in Crenarchaeota include *Thermoproteus*, *Sulfolobus*, and *Pyrodictium*. *See* CAROTENOID; HYDROTHERMAL VENT; SULFUR.

Euryarchaeota. This is the largest phylum of archaea with many diverse classes, orders, and families. Euryarchaeota comprises five major groups that differ with respect to their physiology and ecology: methanogens, halophiles, thermoplasms, extremely thermophilic S^0 metabolizers, and sulfate-reducing archaea.

Methanogens. The methanogens constitute the largest group of archaea and are responsible for the production of all of the biologically produced methane in the atmosphere (Fig. 2). They grow in the absence of oxygen (strict anaerobes) and consume carbon dioxide and hydrogen or simple organic compounds (such as acetic acid, formic acid, alcohols, and methylaminines), from which they produce methane or methane and CO_2. They are the last organism on the food chain in the conversion of organic carbon to water-insoluble gas, most of which escapes into the atmosphere. They inhabit the digestive tract of mammals (ruminants, such as cattle, deer, elk, sheep, goats, and giraffes), sewage sludge digesters, swamps (where they produce marsh gas), hot springs, and sediments in freshwater and marine environments, and even live within anaerobic protozoa. A few methanogens can grow at very high temperatures and live alongside the extreme thermophilic S^0 metabolizers in marine hydrothermal vents. It is possible that methanogenic archaea were among the earliest organisms, as they seem well adapted to living under conditions similar to those presumed to have existed on the young Earth. *See* METHANE; METHANOGENESIS (BACTERIA).

Halophiles. The halophilic archaea are characterized by their dependence on high concentrations of salt (sodium chloride) from their environment. They carry out aerobic respiration and are chemoheterotrophic, relying on complex nutrients (typically proteins and amino acids) for growth. Halophiles are found in the Great Salt Lake in Utah, the Dead Sea between Israel and Jordan, alkaline salt lakes of Africa, marine salterns (solar evaporation ponds used to concentrate salt for use in seasoning and for the production of fertilizers), and on salt-preserved fish and animal hides. Most strains require salt concentrations near the saturation point of sodium chloride (about 36%) to maintain the integrity of their cell walls.

The most studied halophile is *Halobacterium salinarium*. This species is best known for a unique type of photometabolism that is used to derive energy from sunlight when the oxygen content of the environment declines. It uses the rhodopsin pigments (bacteriorhodopsins) as photoreceptors to produce adenosine triphosphate (ATP). *Halobacterium* also uses this ATP to rotate its flagella to propel itself to

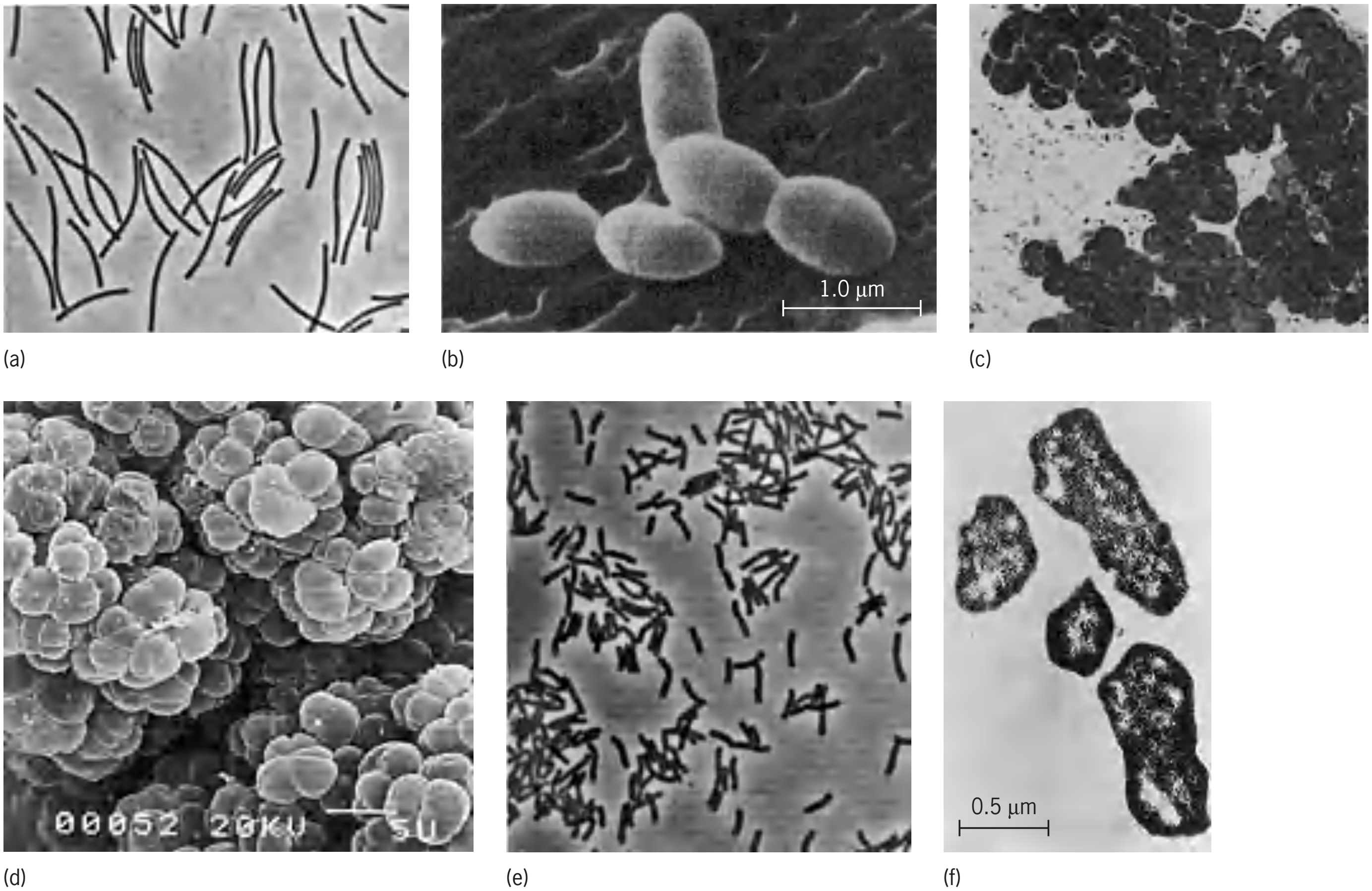


Fig. 2. Selected methanogens. (*a*) *Methanosppirillium hungatei;* phase contrast (×2000). (*b*) *Methanobrevibacter smithii.* (c) *Methanosarcina barkeri* from sewage digester; transmission electron microscope (×6000). (*d*) *Methanosarcina mazei;* scanning electron microscope. Scale bar = 5 μm. (*e*) *Methanobacterium bryantii;* phase contrast (×2000). (*f*) *Methanogenium marisnigri;* electron micrograph (×45,000). (*Reprinted with permission from L. Preslott, J. Harley, and D. Klein, Microbiology, 5th ed., McGraw-Hill, 2001*)

the proper water depth for maximum light absorption. *See* HALOPHILISM (MICROBIOLOGY).

Thermoplasms. The thermoplasms lack cell walls and are thermoacidophilic, requiring high temperatures and extremely acidic conditions to grow; their deoxyribonucleic acid (DNA), RNA, cytoplasmic membranes, and proteins do not function at temperatures below 113°F (45°C). These archaea grow in coal mine refuse piles that contain large amounts of iron pyrite (FeS), which is oxidized to sulfuric acid. As a result, these piles become very hot and acidic. *See* PYRITE.

Extremely thermophilic S^0 metabolizers. These archaea are strictly anaerobic and can reduce sulfur to sulfide. They are motile by flagella and have optimum growth temperatures around 190 to 212°F (88 to 100°C).

Sulfate-reducing archaea. These archaea are characterized not only by their unique ability to reduce sulfate but also by their possession of the methanogen coenzymes F_{420} and methanopterin. Archaeal sulfate reducers can extract electrons from a variety of electron donors (such as H_2, lactate, glucose) and reduce sulfate, sulfite, or thiosulfate to sulfide. (Elemental sulfur is not used as an acceptor.) Many species are found in marine hydrothermal vents. John P. Harley

Bibliography. T. Brock, Life at high temperatures, *Science*, 203:132–138, 1985; R. Charleobois, *Archaea*: Whose sister lineage?, in *Organization of the Prokaryotic Genome*, ed. by R. L. Charleobois, pp. 63–76. ASM Press, Washington, DC, 1999; C. R. Woese, Archaebacteria, *Sci. Amer.*, 244(6):98–122, 1981; C. R. Woese et al., Towards a natural system of organisms: Proposal for the domain Archaea, Bacteria, and Eukarya, *Proc. Nat. Acad. Sci. USA*, 87:4576–4579, 1990.

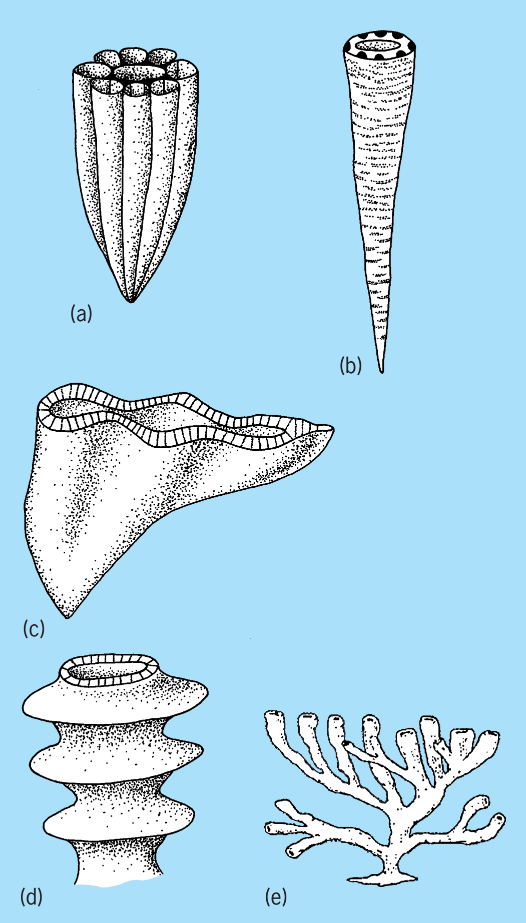


Fig. 1. Some growth forms in the Archaeocyatha: (*a*) ribbed; (*b*) tubular conical; (*c*) asymmetric, conical; (*d*) annulated; (*e*) branching.

Archaeocyatha

An extinct group of mainly Lower Cambrian marine sponges which, although lacking spicules, possessed an intricate, highly porous skeleton of calcite. It was probably a monophyletic group; that is, all representatives were derived from a single ancestor. The position of the Archaeocyatha within the Porifera is uncertain, but they were probably most closely related to the class Demospongiae. Their fossil record is well known, as archaeocyaths represent the first large skeletal animals to have been associated with reefs; they were widespread in the shallow, warm waters that surrounded the many continents that occupied tropical latitudes during the Cambrian.

Morphology. Archaeocyath sponges displayed a great variety of sizes and growth forms. They ranged from a few millimeters to over 500 mm in length and width, and included open cups, convoluted plates, and complex branching forms (**Fig. 1**). Cup-shaped and branching forms 20–100 mm in height were, however, by far the most common. A typical cup-shaped archaeocyath skeleton is composed of a double-walled, inverted cone; the area between the outer and inner walls is known as the intervallum (**Fig. 2**). The intervallum may bear radially arranged, vertical plates known as septa, and other skeletal structures such as horizontal tabulae and dissepiments, or radial rods. The inner wall encloses a central cavity, which is open to the exterior at the top of the cup.

The calcareous skeleton of archaeocyaths was densely microcrystalline, and may have been composed of high-magnesium calcite. This microstructure is remarkably similar to that found in the living calcified sponge *Vaceletia*, and so it has been inferred to have formed in a similar fashion via calcification of a series of stalked collagenous templates.

Functional morphology. The archaeocyath soft tissue was probably restricted to the intervallum, with a thin veneer of membrane around the inner and outer walls. Since archaeocyaths were suspension feeders, water almost certainly entered the animal via the numerous small pores in the outer wall, was filtered by the choanocyte chambers housed in the intervallum, and exited via the porous inner wall from the opening at the top of the central cavity. The outer skeletal wall probably acted as a sieve to exclude large, potentially clogging particles from the soft tissue.

Flume experiments have shown that the presence of septa enhances passive water flow through a model of an archaeocyath cup. Moreover, septa lacking pores appear to minimize leakage of flow from the porous outer wall. This observation might

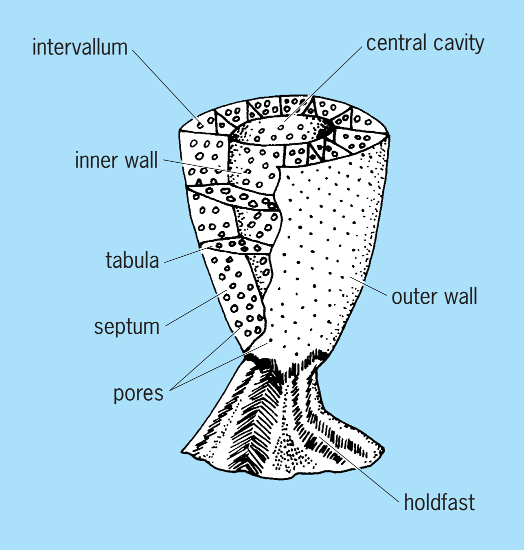


Fig. 2. Archaeocyath skeletal structure.

explain the trend found in some archaeocyaths which show a reduction in septal pores through the lifetime of an individual. As an individual increases in height, it encounters increasingly higher current velocities, necessitating greater control of water flow.

Classification. Archaeocyatha has been traditionally subdivided into two subclasses, the Regulars and Irregulars, according to differences in the early development of the skeleton. It has now been demonstrated, however, that these skeletal characters are a function of differences in soft-tissue distribution, which are independent of systematic placing. Regulars (orders Monocyathida, Ajacicyathida, Coscniocyathida, and Tabulacyathida) generally show no tabulae or dissepiments, and they are inferred to have had intervalla that were completely filled with soft tissue. By contrast, the skeletons of Irregulars (orders Archaeocyathida and Kazachstanicyathida) are believed to have borne soft tissue in their upper parts only, as they possessed abundant structures such as tabulae which served to section off abandoned areas of the skeleton as the soft tissue migrated upward.

History. Archaeocyaths appeared about 535 million years ago (Ma) on the Siberian platform with a considerable standing diversity of species, possibly indicating some preskeletal history. The group migrated from this center of origin to become globally widespread in low latitudes by about 530 Ma, and rapidly diversified to reach an acme of over 170 genera by approximately 525 Ma. However, between approximately 523 and 520 Ma, the group suffered extinction (probably due to loss of habitat) and diversity plummeted; only two species are known from later in the Cambrian.

Ecology. Archaeocyaths were a common element in Lower Cambrian reefs, where they intergrew with a varied community of calcimicrobes (possible calcified cyanobacteria). Most were attached to calcimicrobes by a holdfast or epitheca (Fig. 2), although some large solitary Regular individuals may have been rooted in lime mud. Many had buttresses which aided in stabilization and binding, as well as inhibiting the growth of competitors. Reworked archaeocyath skeletal debris was also a common component of interreef sediments.

Archaeocyaths were both upright open-surface frame builders and common crypt dwellers; indeed, up to half of the species in some communities were cryptic, living within small cavities within the reef (**Fig. 3**). Representatives from all six archaeocyathan orders were present in crypts, but in markedly different proportions. A far greater proportion of archaeocyathids, monocyathids, and coscinocyathids were represented in cryptic communities than were ajacicyathids and tabulacyathids. Ajacicyathids were a very minor component of the cryptos, even though they were the most diverse order in open surface communities.

In addition to differences in systematic distribution, cryptic and open-surface archaeocyaths display distinct morphological differences. First, all cryptic archaeocyaths have porous septa. This suggests that such forms may have been adapted hydrodynamically to low-turbulence conditions, as might be expected in crypts. Second, while cryptic individuals were often small cup-shaped and juvenile individuals, branching forms dominated open surfaces.

Some have suggested that archaeocyaths possessed photosymbionts by analogy with modern reef corals. However, the inferred absence of large areas of external soft tissue that could have been exposed to light, and the preferred cryptic niche of many representatives, does not support this hypothesis.

Evolutionary trends. Archaeocyaths show a strongly progressive trend through their history of an increasing proportion of branching forms.

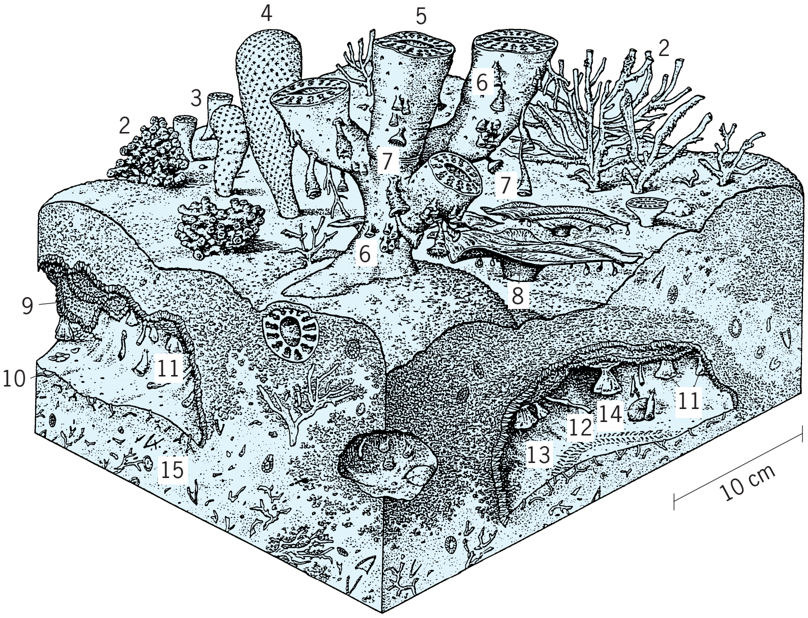


Fig. 3. Reconstruction of a Lower Cambrian archaeocyath reef community: (1) calcimicrobe; (2) branching archaeocyath sponges; (3) solitary cup-shaped archaeocyath sponges; (4) chancelloriid (problematicum); (5) radiocyaths (problematicum); (6) small archaeocyath sponges; (7) "coralomorphs"; (8) archaeocyath sponge; (9) fibrous cement; (10) microburrows (traces of a deposit feeder); (11) cryptic archaeocyaths and coralomorphs; (12) cribricyaths (problematicum); (13) trilobite trackway; (14) cement; (15) sediment with skeletal debris. (*Copyright, John Sibbick*)

Branching morphologies were derived from the more primitive solitary, cup-shaped state and were acquired independently in several lineages. Few possessed a branching habit in the early stages of archaeocyath history, but by the end of the Lower Cambrian about three-quarters of genera possessed such growth forms. This suggests that branching archaeocyaths were better competitors for space in the Lower Cambrian reef habitat. *See* PARAZOA; PORIFERA.

Rachel Wood

Bibliography. F. Debrenne and A. Yu. Zhuravlev, Irregular Archaeocyaths: Morphology, ontogeny, systematics, biostratigraphy, palaeoecology, *Cahiers de Paléontologie*, 1992; M. Savarese, Functional analysis of archaeocyathan skeletal morphology and its paleobiological implications, *Paleobiology*, 18:464–480, 1992; R. Wood, A. Yu. Zhuravlev, and F. Debrenne, Functional biology and ecology of Archaeocyatha, *Palaios*, 7:131–156, 1992; A. Yu. Zhuravlev and R. A. Wood, Lower Cambrian reefal cryptic communities, *Palaeontology*, 38:443–740, 1995.

Archaeognatha

An order of ancestrally wingless insects, commonly known as bristletails, in the subclass Apterygota; about 350 species are known in the two families, Machilidae and Meinertellidae. All have slender, spindle-shaped bodies of about 0.4–0.8 in. (1–2 cm) in length, tapered tailward to a long, multisegmented median caudal filament and paired cerci. The relatively thin integument is covered with pigmented scales after the first instar. The antennae are long and filiform, without intrinsic musculature in the flagellum, as in all true insects; and large, contiguous compound eyes are accompanied by a trio of ocelli. The mandibles are exposed and attached to the cranium by a single joint (condyle), and the maxillary palpi are extended, with seven segments, while the labial palpi have only three segments. Each leg has a large coxa, two-segmented trochanter, femur, tibia, and a two- or three-segmented tarsus terminating in a two-clawed pretarsus. The posterior pair or two pairs of coxae and most of the ventral abdominal coxites bear styles, which may represent modified legs. Outwardly projecting vesicles that are apparently water absorbing also occur on most coxites. Internally, the gut is relatively simple, with small crop, digestive ceca, and a dozen or more Malpighian tubules. The nervous system is primitive, with three thoracic and eight abdominal ganglia and twin connectives. Tracheae are well developed.

In mating, the male deposits sperm on a thread or in a spermatophore that is picked up by the female, but in at least one species sperm is placed directly on the female ovipositor. Immature forms are similar to adults in body form, though the first instar may lack cerci. Sexual maturity is attained after eight or more molts.

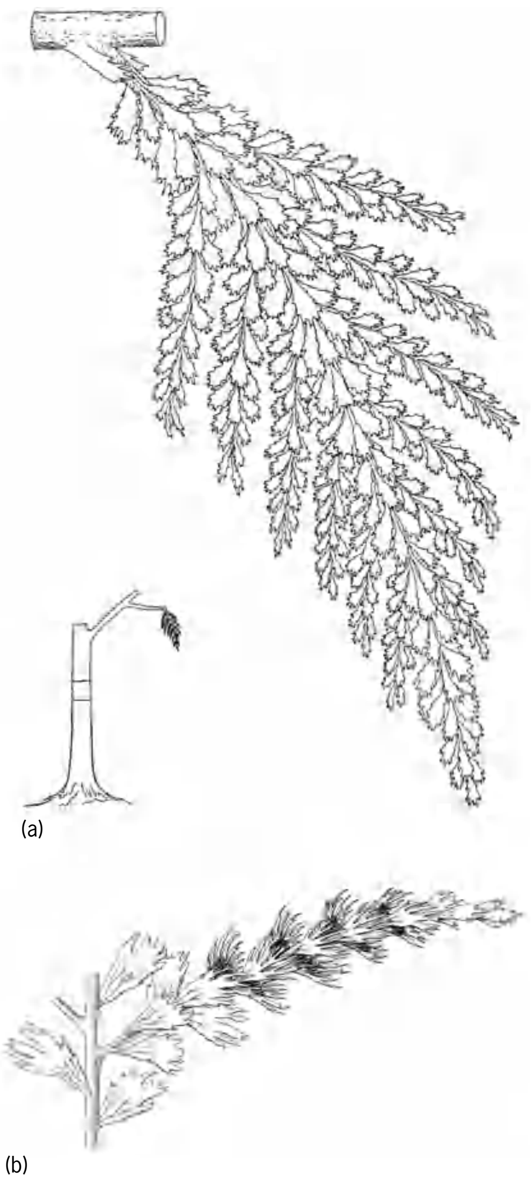
Bristletails live in a variety of habitats worldwide, often on stones or tree trunks where they find the algae and lichens that appear to be their principal sources of food. Most species are primarily nocturnal feeders. *See* INSECTA.

William L. Brown, Jr.

Bibliography. D. J. Borror, C. A. Triplehorn, and N. F. Johnson, *An Introduction to the Study of Insects*, 6th ed., 1997; Commonwealth Scientific and Industrial Research Organization Staff, *The Insects of Australia*, 1991.

Archaeopteridales

An order of extinct, free-sporing plants (cryptogams) that lived during the Late Devonian through Early Mississippian. The order includes the genera



***Archaeopteris macilenta*. (*a*) Vegetative lateral branch system bearing simple leaves (*from C. B. Beck, On the anatomy and morphology of lateral branch systems of Archaeopteris, Amer. J. Bot., 58:758–784, 1971*). (*b*) Strobilus showing adaxial sporangia (*from C. B. Beck, Archaeopteris and its role in vascular plant evolution, in K. J. Niklas, ed., Paleobotany, Paleoecology, and Evolution, 1:193–230, Praeger, 1981*).**

Archaeopteris which has been extensively investigated, *Actinoxylon*, *Actinopodium*, *Svalbardia, Eddya*, and *Siderella*.

Archaeopteris had a worldwide distribution on what today are Northern Hemisphere continents, as well as on Australia. The genus is characterized by an arborescent habit and large, determinate, deciduous, frondlike lateral branch systems that bear simple, oval to fan-shaped leaves, with dichotomous venation, which vary in different species from highly dissected to nearly entire. The main axes of lateral branch systems have radially symmetrical internal anatomy and helically arranged leaves; ultimate axes are borne suboppositely to alternately in one plane (see **illus.**). Anatomy and leaf arrangement of ultimate vegetative axes reflect a nonplanate, bilateral symmetry. Fertile leaves (sporophylls) bear microsporophylls and megasporophylls on their adaxial (upper) surfaces. Radial symmetry of fertile ultimate axes (strobili) is expressed both internally and externally in the region of sporophylls.

In large stems, a cylinder of discrete vascular bundles encloses a broad pith. Leaf and branch traces diverge from the vascular bundles along radial planes. Extensive secondary wood and lesser amounts of secondary phloem are produced by a bifacial vascular cambium. The wood is compact tissue comprising tracheids and rays of variable height and width in different species. The circular-bordered pits in the radial walls of the tracheids are arranged in groups of several to 30 or more aligned in bands at right angles to the long axes of the tracheids, a characteristic also found in isolated wood and stem fragments called *Callixylon*. *See* PHLOEM.

Archaeopteridales belongs to the class Progymnospermopsida. The progymnosperms are thought by some botanists to be ancestral to seed plants. *See* PLANT KINGDOM.

Charles B. Beck

Bibliography. C. B. Beck, (ed.), *Origin and Evolution of Gymnosperms*, 1988; S. V. Meyer, *Fundamentals of Paleobotany*, 1987; W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, 2d ed., 1993.

Archaeopteryx

A genus of fossil birds in the extinct order Archaeopterygiformes, characterized by flight feathers like those of modern birds. *Archaeopteryx* is generally accepted as the oldest known fossil bird. It represents a unique snapshot of evolution; most of its skeletal characters are shared with small carnivorous dinosaurs, yet it has fully feathered modern wings. *Archaeopteryx* is effectively a flying dinosaur—an ancient bird about the size of a modern magpie. *See* AVES; DINOSAUR.

Fossil record. *Archaeopteryx* is known from one feather impression and seven skeletons, discovered over the course of 150 years from the Altmuhl Valley in Bavaria, Germany. All have been found in lithographic limestone of uppermost Jurassic (Tithonian) age, approximately 147 million years old, and were excavated from several locations that lie within a 15-mile radius of the town of Solnhofen. The Solnhofen Lithographic Limestone preserves exceptionally fine detail, including the impressions of feathers; the animal carcases sank into fine plastic sediment on a shallow lagoon bed and were buried in anaerobic conditions. The original specimen, the impression of a single feather preserved as part and counterpart, was discovered in 1861 and named *Archaeopteryx lithographica* by H. von Meyer. A few months later, a nearly complete skeleton was discovered and subsequently sold to the Natural History Museum in London. A second, complete specimen was discovered in 1877 and eventually acquired by the Berlin Natural History Museum. It was named as a different species, *Archaeopteryx siemensii*, by W. Dames in 1897 and as a separate genus *Archaeornis* by B. Petronievics in 1925, although it is widely regarded as a smaller individual of *A. lithographica*.

Five additional well-documented skeletons of *Archaeopteryx* have been discovered since the 1950s. Not all are complete, but all show some evidence of feather impressions. The identity of three skeletons were not recognized at first. A fragment collected in 1855 showing feather impressions, claws, and a knee joint was named as a pterosaur (a flying reptile) by von Meyer in 1859; its correct identity was not realized until 1970. Two specimens had been known for many years and misidentified as examples of the small theropod dinosaur, *Compsognathus*. The largest (sixth) specimen has been described as a separate genus and species, *Wellnhoferia grandis*, on the basis of size and distinctive manual (hand) and pedal (foot) morphology. The most recent discovery in 1993 was named *A. bavarica* by P. Wellnhofer based principally on its longer limb proportions. Eighth and ninth specimens are also known to exist, but they have not been described. *See* FOSSIL.

Morphology. The most striking feature of *Archaeopteryx* is the feather imprints in the surrounding fine-grained lithographic limestone, which are preserved most completely in the London and Berlin specimens. The structure of the primary flight feathers in the wing is exactly like that seen in modern flying birds, with the midrib (rhachis) offset to give a short leading-edge vane and a longer trailing-edge vane. Furthermore, the number and arrangement of primary and secondary flight feathers is identical to that in modern birds. A pair of symmetrical feathers was attached to each tail vertebra. The skull has two temporal openings (as is characteristic of diapsid reptiles), with a large braincase. The jaws were lined with slightly curved, conical, unserrated teeth that were separated from each other by interdental plates that closely resemble those of small theropod dinosaurs.

Many features of the postcranial skeleton are shared with maniraptorian (small advanced theropod) dinosaurs. Significant features include a long scapula and coracoid that are almost equal in length. The lateral orientation of the shoulder socket is intermediate between the dorsolateral position in modern birds and the ventrolateral position in maniraptorans. The furcula (wishbone), formed from the fusion of the clavicles, is a simple boomerang shape, similar to that in maniraptorans and unlike the long, springy deeply-V-shaped one present in modern birds. There was no ossified sternum (keel); an element originally interpreted as a keel in *A. bavarica* has been reidentified recently as part of a coracoid. The forearm skeleton is long and the wrist includes a semilunate carpal. This element allowed much greater mobility of the hand, which is regarded as a prerequisite of the flight stroke. The hand has three long fingers, ending in hooked claws with long horny sheaths. The hind limb is slender with a long tibia, three walking toes, and a short reversed perching toe. *See* FEATHER.

Biology. The lifestyle of *Archaeopteryx* has been much debated, particularly the issue of whether it was basically terrestrial or arboreal. It may have foraged on the ground; the dentition suggests feeding on a range of small animals—arthropods and perhaps small vertebrates. *Archaeopteryx* had some climbing ability based on the mobility of the hands and curvature of the manual claws. The pedal claws are less curved and reflect a compromise between climbing and terrestrial locomotion.

Archaeopteryx had some flight capability since it possessed flight feathers and wings. Its brain and sensory systems were also flight ready. Computed tomography and three-dimensional reconstruction of the London specimen has shown that the brain volume of *Archaeopteryx* is within in the range for modern birds. It possessed a birdlike brain pattern, with large optic lobes (reflecting the importance of sight), large cerebellar lobes for flight control, and large floccular lobes concerned with balance. The inner ear was also birdlike, with large semicircular canals for balance and spatial orientation.

All the evidence suggests that *Archaeopteryx* was capable of limited, slow, stable powered flight. It lacked anatomical improvements seen in modern birds for maneuverable flight and ground takeoff. A combination of tree climbing and launching from a perch to fly to a safe height may have been a predator escape strategy and a stimulus for the evolution of bird flight.

Phylogeny. It is generally accepted that birds evolved from small feathered maniraptoran dinosaurs. *Archaeopteryx* shares many derived skeletal characters with maniraptorans, particularly the dromaeosaurids, the family to which it is most closely related. The discovery of several lineages of small feathered dinosaurs, including dromaeosaurs, from an exceptional preservation site of Lower Cretaceous age in Liaoning Province, China, strengthens the case for a dinosaurian origin of birds. These discoveries suggest that feathers evolved initially to provide an insulatory covering and were much later elaborated and co-opted for flight. A small minority supports the hypothesis that birds evolved from basal archosaurs. However, there is much testable evidence to support a maniraptoran origin. *See* ANIMAL EVOLUTION.

Angela C. Milner

Bibliography. P. J. Currie et al., *Feathered Dragons*, Indiana University Press, Bloomington, 2004; P. Dominguez Alonso et al., The avian nature of the brain and inner ear of *Archaeopteryx, Nature*, 430:666–669, 2004; A. Elzanoski, Archaeopterygidae (Upper Jurassic of Germany), pp. 129–159, in L. M. Chiappe and L. M. Witmer (eds.), *Mesozoic Birds: Above the Heads of the Dinosaurs*, University of California Press, Berkeley, 2002; A. C. Milner, *Dino-Birds*, Natural History Museum, London, 2002; K. Padian and L. Chiappe, The origin and early evolution of birds, *Sci. Amer.*, pp. 28–33, February 1998.

Archaic ungulate

The most diverse group of early Tertiary mammals are the archaic ungulates (condylarths). Although closely related, they are not a monophyletic group but, along with various extinct and extant mammals, form the taxon Ungulata (**Fig. 1**). They are ancestral to as many as 7 of 18 living orders of mammals: Artiodactyla, Cetacea, Hyracoidea, Perissodactyla, Proboscidea, Sirenia, and possibly Tubulidentata. Along with the Late Cretaceous zhelestids, best known from Asia, Ungulata forms the taxon Ungulatomorpha. Fossil (and molecular) evidence suggests that Ungulatomorpha separated from other placentals 85 million years ago.

Early ungulatomorphs had lower-crowned, more squared molars compared to contemporary placentals, which have a more slicing dentition, indicating a trend toward omnivory and herbivory. Although "ungulate" implies hooves, most archaic ungulates had at best rudimentary hooves or even claws.

Arctocyonidae is the sister taxon to other ungulates, with the possible exception of artiodactyls. Arctocyonids retained more primitive character states in their dentition, such as more teeth and relatively little molarization of premolars. Smaller arctocyonids were like the coati-sized *Chriacus*, a scansorial mammal probably equally adept in trees or on the ground. It had powerful limb musculature, mobile joints, flexible plantigrade pentadactyl feet with claws, a slightly divergent hallux (first digit of the hindfoot), and a long and robust tail for grasping or taking hold of something. Wolf-sized arctocyonids such as *Arctocyon* were terrestrial with an extremely long powerful tail, short stout limbs, a plantigrade manus (forefoot) and pes (foot), and an opposable hallux. The dentitions of all arctocyonids were low-crowned with rounded cusps, and moderately to well developed canines. Arctocyonids are thought to have had a bearlike diet of fruits, nuts, seeds, insects, and small animal prey. Arctocyonids reached their diversity peak in the middle to late Paleocene, disappearing in the early Eocene. They are known from North America and Europe.

Members of Hyopsodontidae were generally smaller animals. The ubiquitous, rat-sized early Eocene *Hyopsodus* was quite slender and possessed shortened limbs. The hyopsodontid dentition was similar to that of arctocyonids but showed the beginnings of molarization of premolars and a hint of crest development on molars. Hyopsodontidae ranges from early Paleocene through late Eocene, reaching its peak in late Paleocene. They are best known from North America and Europe, but with Asian and north African records.

Mioclaenidae was quite different in dental architecture. Mioclaenids had simplified biting surfaces and some premolar inflation, suggesting consumption of tougher, fibrous plants. They seem in general to have been slightly larger (up to hare size) than hyopsodontids, but similar to smaller arctocyonids. The mioclaenids are diverse in the middle Paleocene but appeared in the early Paleocene and disappeared in the late Paleocene. They occurred in North and South America and Europe.

Didolodontidae, from the middle Paleocene through Oligocene of South America, was dentally similar to Mioclaenidae. The size of didolodontids was similar to or slightly larger than that of mioclaenids. Didolodontidae gave rise to the totally extinct South American Meridiungulata, comprising Litopterna, Astrapotheria, Notoungulata, Pyrotheria, and Xenungulata.

Premolar inflation was greater in Periptychidae than Mioclaenidae. In *Periptychus*, the quite bulbous premolars were grooved. Some molar cusps were accentuated, while crests developed both on molars and on premolars. This suggests that tough, fibrous plant material was eaten by the sheep-sized periptychids. Some smaller, squirrel-sized periptychids showed similar specializations but, by virtue of size, probably ate higher-energy diets, which may have included fruits, nuts, or insects. Postcranially, periptychids are best known from the squirrel-sized *Gillisonchus* and the sheep-sized *Ectoconus*. In size and general body proportions, but not in head, tail, or pedal structure, *Ectoconus* resembled the aardvark, *Orycteropus*. *Periptychus* and *Ectoconus* were stoutly built with no reduction of the distal ulna or the fibula and no elongation of limbs, and had a pentadactylous manus and pes. At least *Ectoconus* among periptychids possessed a subplantigrade manus and pes that was short and wide with considerable padding below, wide spreading digits, and small tapirlike hooves (**Fig. 2**). Periptychidae was taxonomically diverse in the early Paleocene of North America but rapidly declined, disappearing in late Paleocene.

Triisodontines, Hapalodectidae, and Mesonychidae, along with Cetacea, constitute the Cete. Artiodactyla might be the sister taxon to Cete. In the earliest triisodonts, the suggestion of later dental trends toward carnivory can be seen. In cetans such as *Synoplotherium* and *Harpagolestes*, excessive wear on molars and incisors, massiveness of some of the cusps, and large canines point to carrion feeding. *Dissacus*, *Mesonyx*, and *Hapalodectes* lacked these specializations but probably were carnivorous, as was the slender-toothed *Hapalodectes*. These animals ranged from the raccoon-sized *Microclaenodon* and coyote-sized *Dissacus* to the wolf-sized *Mesonyx* and bear-sized *Harpagolestes* and *Pachyaena*. Based on *Mesonyx*, the skeleton of cetans developed more cursorial adaptations. Cetan skeletons tended to be more gracile, and in some such as *Mesonyx* the manus and pes were digitigrade (walking on toes) with closely aligned metapodials. Some taxa such as *Dissacus* retained the primitive plantigrade posture and hooved unguals, rather than reverting to claws as in some arctocyonids. The triisodonts were moderately diverse in the early and mid Paleocene. In North America and Asia (and a lesser extent Europe), another radiation occurred during the Eocene, before cetans disappeared.

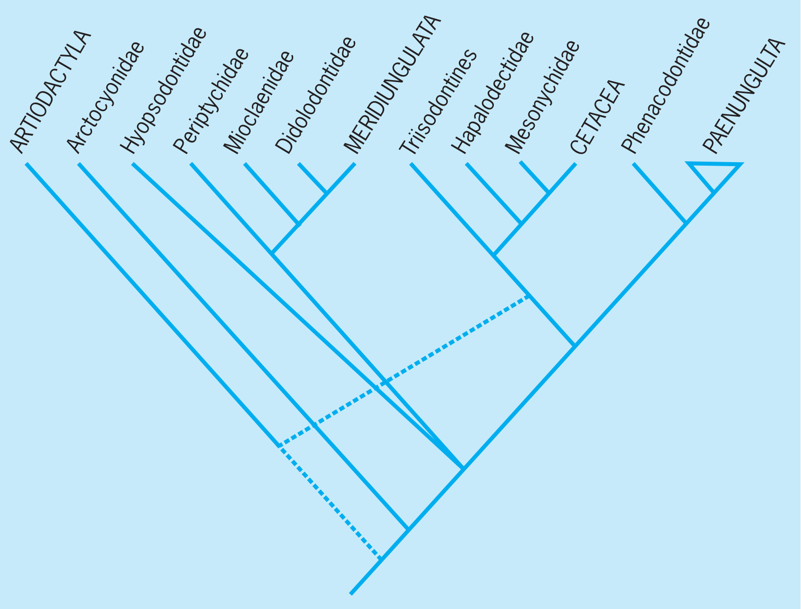


Fig. 1. Cladogram relating archaic ungulates to major clades of extant and extinct ungulates.

Phenacodontidae was taxonomically the least diverse of the archaic ungulates. Both dentally and postcranially, phenacodontids showed several trends away from arctocyonids. In *Phenacodus*, whose teeth were quite low-crowned with distinctive cusps, there was a trend toward lophs and crests on both upper and lower molars. Phenacodontids tended toward molarization of the premolars more than other archaic ungulates. The molarization of the premolars and the development of lophs and selenes (crescents) reached its extreme in *Meniscotherium*. Postcranially, the sheep-sized *Phenacodus* is one of

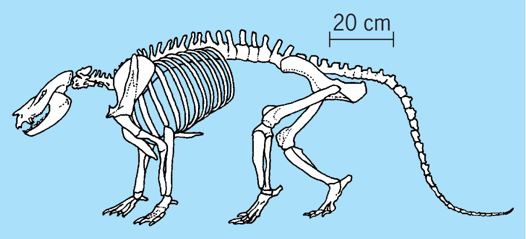


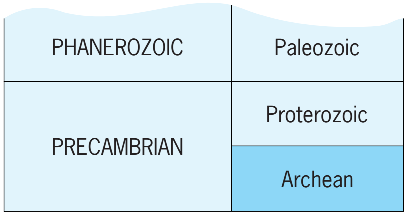
Fig. 2. Skeleton of *Ectoconus*, an early Paleocene phenacodont condylarth.

the best-known taxa. Its limbs were somewhat elongated, but flexibility was still similar to arctocyonids except possibly in the carpus and tarsus. Digits I and V were reduced. Phenacodontidae appeared in the late Paleocene and disappeared in the middle Eocene. They were from North America and Europe. Phenacodontidae is the probable sister taxon of Paenungulata, which among extant mammals includes Perissodactyla, Hyracoidea, Proboscidea, and Sirenia. J. David Archibald

Bibliography. M. J. Benton, *Vertebrate Paleontology*, Chapman & Hall, London, 1997; R. L. Carroll, *Vertebrate Paleontology and Evolution*, W. H. Freeman, New York, 1988.

Archean

A period of geologic time from about 3.8 to 2.5 billion years ago (Ga). During the Archean Eon a large percentage of the Earth's continental crust formed, plate tectonics began, very warm climates and oceans existed, and life appeared on Earth in the form of unicellular organisms.



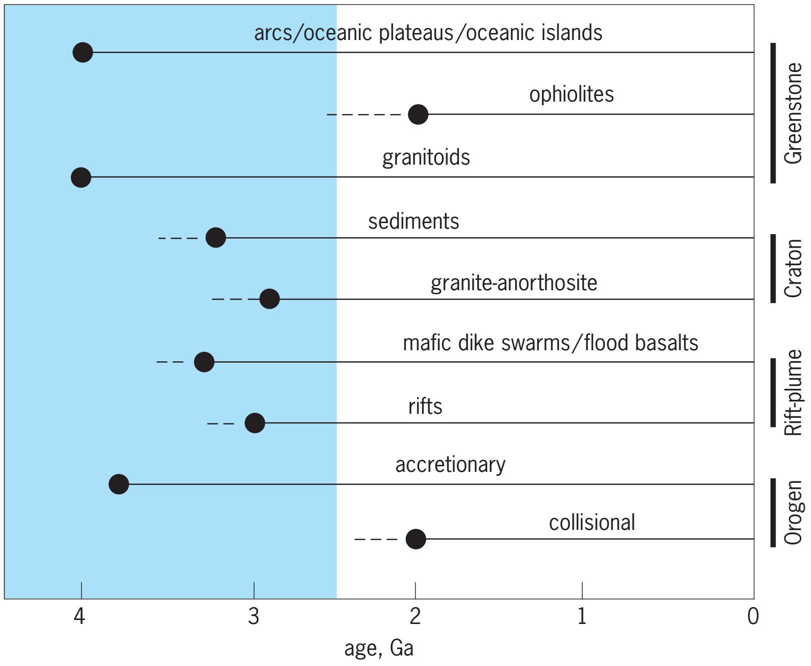


Fig. 1. Distribution of tectonic settings with geologic time. Note that most modern tectonic settings extend back into the Archean (shaded region). Ophiolites are remnants of oceanic crust.

Tectonic regimes. Plate tectonics can be tracked with time by using rock assemblages that are characteristic of certain tectonic settings. Although at one time greenstones (rock successions in which submarine basalts and related sediments dominate) were thought to be an Archean phenomenon, it is now clear that they have formed throughout geologic time. It is also clear that all greenstones do not represent the same tectonic setting: some reflect island arcs, whereas others reflect oceanic plateaus, volcanic islands, or oceanic crust. Except for ophiolite greenstones (remnants of oceanic crust), greenstones are recognized throughout the geologic record from the oldest known rocks at 4 Ga to the present (**Fig. 1**). The oldest rocks indicative of craton and passive-margin environments are sediments about 3 Ga, suggesting that cratons developed by this time, very early during the Archean, although they may not have been large in area. Giant dike swarms and flood basalts, which are probably derived from mantle plumes, are present in southern Africa and western Australia by 2.7 Ga, with the oldest known dike swarms dating to over 3 Ga (Fig. 1). The oldest continental rift assemblages, comprising basalts and immature sediments, occur in southern Africa at about 3 Ga, and the oldest accretionary orogens (linear belts of deformation resulting from plate collisions) are the Acasta gneisses (4 Ga) and the Amitsoq gneisses (3.9 Ga) in northwest Canada and southwest Greenland, respectively. Although the oldest well-documented collisional orogens are Paleoproterozoic (2.5–1.6 Ga), it is likely that late Archean collisional orogens with reworked older crust occur on some continents. *See* CRATON; GNEISS; PLUTON.

The occurrence of rock assemblages typical of arcs, oceanic plateaus, and oceanic islands and the presence of accretionary orogens in the very earliest vestiges of the geologic record at 4–3.5 Ga strongly supports some sort of plate tectonics operating on the Earth by this time. By 3 Ga, cratons, passive margins, and continental rifts were also widespread. Although plate tectonics appears to have occurred since 4 Ga, there are geochemical differences between Archean and younger rocks that indicate that Archean tectonic regimes must have differed in some respects from modern ones. The degree that Archean plate tectonics differed from modern plate tectonics is unknown; however, these differences are important in terms of the evolution of the Earth. *See* PLATE TECTONICS.

Earth's oldest rocks. The oldest rocks occur as small, highly deformed terranes tectonically incorporated within Archean crustal provinces (**Fig. 2**). Although the oldest known igneous rocks on Earth are the 4 Ga Acasta gneisses of northwest Canada, the oldest minerals are detrital zircons (zircons in sediments) from the 3 Ga Mount Narryer quartzites in western Australia. These zircons have uranium-lead (U-Pb) ages, determined by ion probe, ranging from about 4.3 to 3.5 Ga, and are important in that they indicate the presence of granitic sources for the sediments, some of which were 4.3 Ga. These sources may have been remnants of continental crust,

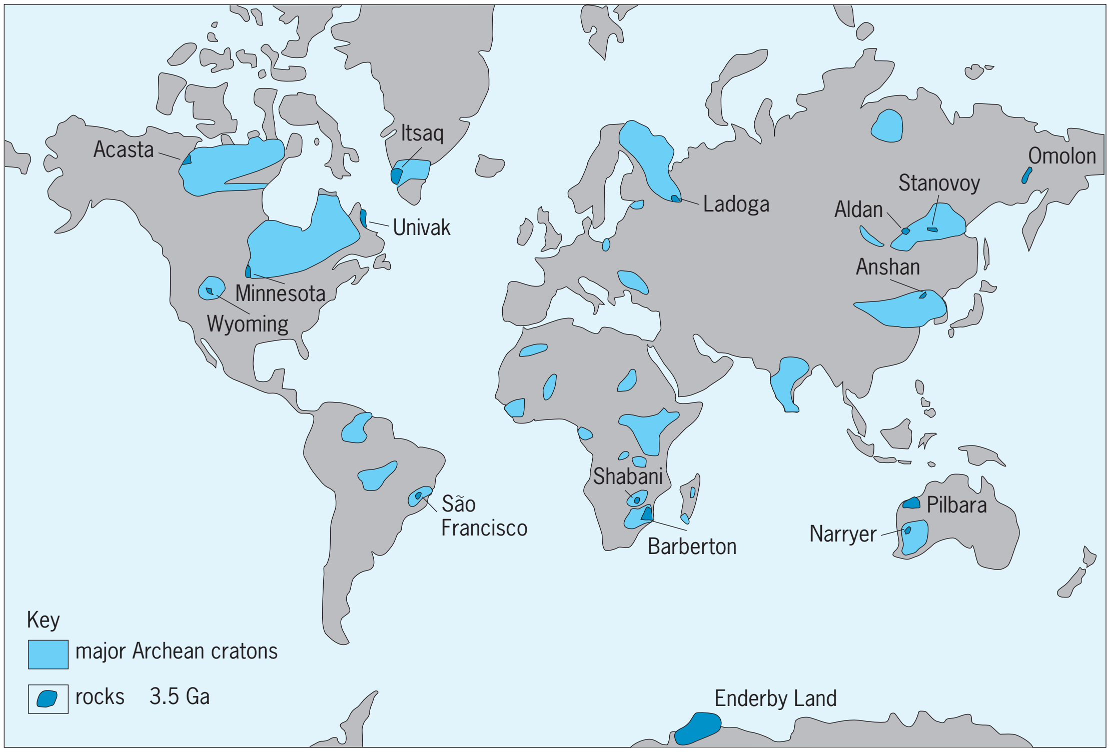


Fig. 2. Distribution of the Earth's oldest rocks.

although the volume of this crust may have been small.

The oldest isotopically dated rocks on Earth are the Acasta gneisses, which are a heterogeneous assemblage of highly deformed granitic rocks, tectonically interleaved with mafic and ultramafic rocks, and metasediments. Uranium-lead zircon ages from the granitic components of these gneisses range from 4 to 3.6 Ga, and thus it would appear that this early crustal segment evolved over about 400 million years and developed a full range in composition of igneous rocks. The chemical compositions of Acasta mafic rocks are very much like less deformed Archean greenstones representing various oceanic tectonic settings. *See* DATING METHODS; ROCK AGE DETERMINATION.

The largest and best-preserved fragment of early Archean continental crust is the Itsaq Gneiss Complex in southwest Greenland. In this area, three terranes, each with its own tectonic and magmatic history, collided about 2.7 Ga, forming the continental nucleus of Greenland. Although any single terrane records less than 500 million years of precollisional history, collectively the terranes record over 1 billion years of history before their amalgamation. The most extensively studied greenstone in southwest Greenland is the Isua sequence, composed of basalts and komatiites interbedded with banded iron formation, intrusive sheets of granite, and volcanic sediments.

The Pilbara craton in western Australia also comprises terranes that accreted between 3.46 and 3.2 Ga. Within this craton are thick successions of submarine volcanics, volcanic sediments, and chemical sediments (chert and banded iron formation). Some of these rest unconformably on an older greenstone complex with a U-Pb zircon age of about 3.5 Ga, an important observation indicating that land emerged above sea level by 3.46 Ga. The Barberton greenstone in southern Africa is one of the most studied early Archean greenstones. Together with coeval granites, the Barberton succession formed between 3.55 and 3.2 Ga and includes four tectonically juxtaposed terranes, with similar stratigraphic successions in each terrane. Each succession begins with submarine basalts and komatiites (high-temperature, magnesium-rich lavas) that may be remnants of oceanic plateaus. Overlying these rocks is a suite of felsic to basaltic submarine volcanics, fine-grained volcanic sediments and cherts, possibly representing an oceanic island arc.

Mineral deposits. The Archean is known for its reserves of iron, copper, zinc, nickel, and gold. Some of the world's largest copper-zinc deposits occur as massive sulfide beds associated with submarine volcanics in Archean greenstones in Canada and western Australia. These deposits are typically lenticular in shape with more than 60% massive sulfides of zinc and iron at the top, with a copper-rich base. They are typically associated with hydrothermal cherts and underlain by large alteration zones. These deposits are formed at the same time as surrounding volcanic rocks by the leaching of metals from the volcanics by hydrothermal fluids and deposition on the sea floor. *See* CHERT.

Nickel sulfides are associated with komatiite flows in Archean greenstones, and they are particularly important in western Australia. Like the copper-zinc deposits, they are layered deposits of finely disseminated ores within submarine komatiite flows. The komatiites appear to be lava channels following submarine topography. Typical statiform nickel ores consist of a thin rich sulfide layer ($>80\%$ sulfides) overlain by a layer of disseminated sulfides. The massive Archean nickel-sulfide ores appear to be of magmatic origin. Voluminous submarine komatiite flows assimilated sulfide-rich sediments during eruption of the sea floor, and immiscible nickel-sulfide droplets settled from the komatiitic lavas, producing the massive layered nickel ore zones at the base. *See* LAVA.

Banded iron formations are common in Archean greenstones worldwide, and many of these deposits are important sources for iron ore. Banded iron formations are sedimentary rocks formed by chemical precipitation of iron oxides in oxygen-depleted seawater. The delicate banding, which can be followed over great distances in these sediments, suggests an origin by cyclic sedimentation in an arc-related basin, or in some instances, like the banded iron formations in the Hamersley Basin in western Australia, in passive margin or intracratonic basins. *See* BANDED IRON FORMATION.

The Archean is a principal producer of the world's gold. The lode gold deposits typical of Archean greenstones vary in their geologic setting, ore minerals, and styles of mineralization. Silver is typically associated with the gold, with varying amounts of such elements as arsenic, antimony, and tungsten. Most of the Archean gold is associated with altered country rocks in which carbon dioxide (CO_2) and potassium have been introduced with salt-poor, gold-bearing fluids. Almost all Archean deposits are controlled by shear zones or faults, and mineralization occurred during the late stages of deformation or after deformation. Studies have shown that many of the deposits formed at shallow depths (<5 km or 3 mi) from hydrothermal fluids with a wide range of temperatures (200–500°C; 390–930°F).

Plate tectonics. Because the Earth has steadily cooled with time, there is reason to suspect that plate tectonics may have been different in the Archean. For instance, a hotter mantle in the Archean would have produced more melt at ocean ridges and hence a thicker oceanic crust, perhaps as thick as 20 km (12 mi). Since oceanic crust is less dense than the mantle, Archean oceanic plates would have been more buoyant and thus more difficult to subduct. A hotter mantle would also convect faster, probably causing plates to move faster, and hence oceanic plates would have less time to cool and become gravitationally unstable. Both of these factors would tend to cause plates to resist subduction in the Archean.

Because of the difficulty of subduction, Archean plates may not have been moving fast enough to remove the excess Archean heat. Although it seems that mantle plume activity was probably greater in the Archean than afterward, plumes cannot be a substitute for heat loss by plate tectonics, since they bring heat into the mantle from the core. In effect, plates cool the mantle while plumes cool the core. What is required is a mechanism in the lithosphere that promotes heat loss at the surface. Three possible explanations for excess Archean heat loss are: (1) The Archean ocean ridge system was unusually long. (2) The inversion of basalt to eclogite (a high-density mafic rock) in thick Archean oceanic crust may have increased the density of the lithosphere sufficiently for it to subduct. (3) Spreading rates of Archean ocean ridges may have been faster than at present.

Although plate tectonics undoubtedly operated in the Archean, it must have differed in some fundamental aspects from modern plate tectonics to effectively remove the excess heat from the Earth's upper boundary layer. Perhaps the most important factor leading to evolutionary changes is the cooling of Earth. If fragments of oceanic plateaus constitute a significant proportion of Archean greenstones, as seems likely, Archean mantle plume activity may have been more widespread than at present. Perhaps the first continental nuclei were oceanic plateaus (**Fig. 3**). Partial melting of the thickened roots of these plateaus would produce granitic magmas, changing the plateaus into continental nuclei. Shallow subduction around the perimeters of the plateaus would result in both lateral continental growth and thickening of the lithosphere, and collision of oceanic plateaus and island arcs around the edges would contribute to continental growth. *See* CONTINENTS, EVOLUTION OF.

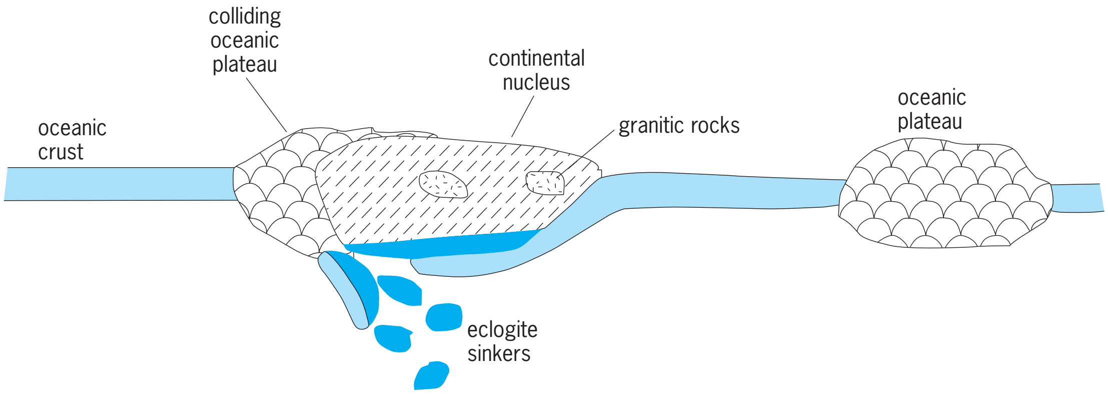


Fig. 3. Model of Archean plate tectonics. Eclogite is a heavy mafic rock that may have pulled buoyant subducting plates into the mantle.

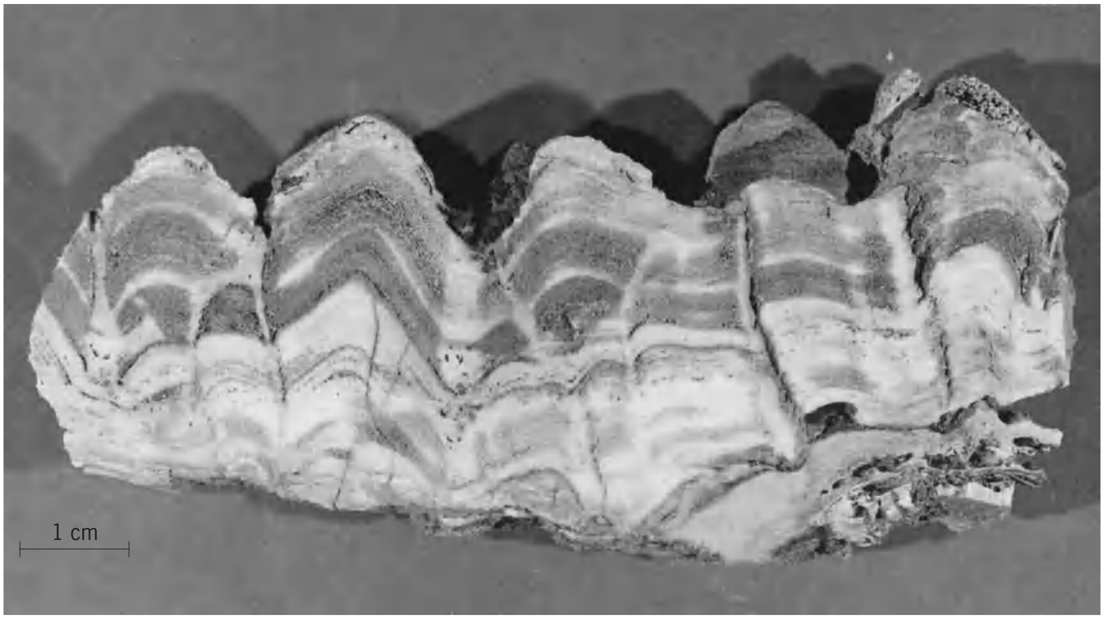


Fig. 4. Early Archean stromatolites from the Barberton greenstone in southern Africa. The laminae were probably deposited by cyanobacteria, and are among the oldest known fossils. (*Courtesy of Don Lowe*)

Atmosphere and oceans. The Earth's first atmosphere was probably composed chiefly of gases such as helium and hydrogen inherited from the solar nebula from which the solar system formed, as well as from the asteroidlike bodies that collided to form Earth. As Earth heated up from core formation, it released gases and formed a secondary atmosphere composed chiefly of CO_2, methane (CH_4), nitrogen (N_2), and water (H_2O). In support of this view, the surviving rock record includes carbonates that reflect a carbon dioxide–rich atmosphere; and also one or more greenhouse gases (carbon dioxide, methane) must have been present to prevent the surface of the Earth from freezing over. The Earth's surface could not have been frozen, because there are sedimentary rocks that were deposited in seawater by 4 Ga. If carbon dioxide and methane were very abundant in the Archean atmosphere, as seems probable, Archean climates must have been very warm due to greenhouse warming. In fact, oxygen isotopes from marine sediments such as chert indicate hot oceans, perhaps reaching 80°C (176°F). The Archean atmosphere and oceans contained very little free oxygen, since Archean sediments contain reduced detrital minerals such as uraninite (UO_2) that could not have survived in an oxidizing system. Also, ancient soil horizons preserved in Archean rocks contain iron chiefly in the reduced state (Fe^{+2}), reflecting low oxygen contents in the atmosphere. *See* ATMOSPHERE; GREENHOUSE EFFECT.

Changes in atmosphere composition in the late Archean and Proterozoic occurred in response to declining oceanic volcanism and hydrothermal input into the oceans and atmosphere after 2.4 Ga. Less carbon dioxide and methane entered the near-surface reservoirs. Also, a supercontinent formed at the end of the Archean, increasing the area of land available for weathering. The weathering extracted carbon dioxide from the atmosphere, causing a decrease in the amount of atmospheric carbon dioxide, which cooled the atmosphere and led to the first worldwide glaciations about 2.3 Ga. As this supercontinent broke up about 2.1 Ga, organic matter was rapidly buried in numerous small basins around the continental fragments, leading to a substantial increase in oxygen in the atmosphere-ocean system.

Life. There are three lines of evidence for life in the Archean: (1) fossil stromatolites, which are laminated structures deposited by microorganisms; (2) fossils of cells or cellular tissue; and (3) carbonaceous matter identifiable from its carbon isotopic composition as a product of biologic activity. Some of the oldest fossil stromatolites occur in the 3.5 Ga Barberton greenstone in southern Africa and in the 3.5 Ga Pilbara greenstone in western Australia. These have wavy or domical shapes and occur in thin chert beds (**Fig. 4**). Threadlike filamentous microfossils of bacteria also occur in some of these cherts. In the Kromberg Formation (3.4 Ga) in southern Africa, the filamentous fossils occur in association with wavy flat-laminated stromatolites composed in part of amorphous carbon. Some of the best-preserved fossil bacteria occur in the Apex chert in western Australia, where cell structure is preserved. A biologic origin for these structures is established by their morphology, characterized by well-defined, barrel- or disc-shaped cells with rounded or conical terminal cells (**Fig. 5**). The similarity of these forms to unbranched, uniseriate prokaryotic cells of cyanobacteria, both living and fossil, is striking. *See* FOSSIL; STROMATOLITE.

It is well known that carbonaceous matter of biologic origin is enriched in the light isotope carbon-12 (^{12}C), compared to carbon-13 (^{13}C). The oldest carbon isotopic evidence for terrestrial life comes from the 3.7-Ga Isua greenstone in western

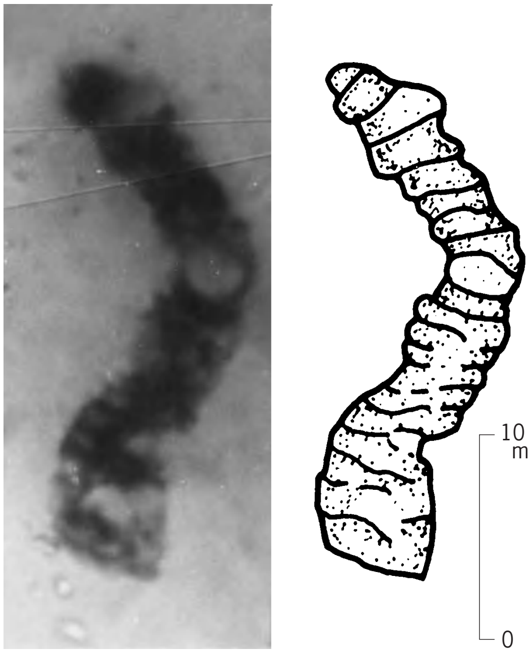


Fig. 5. Typical microfossil bacteria from the 3.46-Ga Apex chert in western Australia, shown in a photograph and interpretive drawing. (*Courtesy of J. W. Schopf*)

Greenland from carbonaceous samples taken from deformed and metamorphosed sedimentary rocks, probably deep-sea sediments. Microscopic carbonaceous globules analyzed from these rocks suggest that these globules came from planktonic, photoautotrophic microorganisms. When the fossil and carbon isotopic evidence is considered collectively, investigators conclude that autotrophic forms of primitive life were well established on the Earth by 3.5 Ga.

Catastrophic mantle plume event. An increase in the production rate of continental crust at 2.7 Ga is commonly attributed to a rapid change from layered to whole-mantle convection in the Earth. The result was a catastrophic mantle plume event, a short period (<50 million years) during which the lithosphere was bombarded by numerous mantle plumes. Evidence supporting this mantle plume event includes a large peak in the frequency of komatiites, oceanic plateau basalts, flood basalts, giant dike swarms, and large layered intrusions. The formation of perhaps the first supercontinent followed the mantle plume event and peak crustal production by 10–50 million years, as indicated by U/Pb zircon chronology of Archean rocks. A peak in gold production between 2700 and 2650 Ma also correlates with supercontinent formation.

A catastrophic mantle plume event at 2.7 Ga would also be expected to affect the ocean/atmosphere/biosphere systems and leave a permanent imprint in the geologic record. For example, (1) low sea level at 2.7 Ga may reflect direct hits of mantle plumes beneath the only two large cratons, Kaapvaal and Pilbara; (2) shales of this age derived from intensely weathered sources may record global warming caused by increased input of greenhouse gases (carbon dioxide and methane) into the atmosphere; (3) a corresponding peak in black shale deposition may reflect increased input of nutrients into the oceans and increased anoxia related to methane input; (4) a strong peak in banded iron formation deposition at 2.7 Ga may be related to enhanced hydrothermal spring activity pumping more iron into the oceans; (5) an increase in the number of occurrences of stromatolites per unit volume of juvenile continental crust at 2.7 Ga may reflect an increase in biomass related to enhanced nutrient/carbon dioxide levels caused by a mantle plume event; (6) a well-defined decrease in the carbon-12/carbon-13 ratio of kerogens at 2.7 Ga could reflect enhanced activity of methanotropic bacteria due to greater input of methane into the oceans. *See* KEROGEN.

Kent C. Condie

Bibliography. S. Bengtson (ed.), *Early Life on Earth*, 1994; K. C. Condie (ed.), *Archean Crustal Evolution*, 1994; K. C. Condie, *Plate Tectonics and Crustal Evolution*, 4th ed., 1997; E. G. Nisbet, *The Young Earth*, 1987.

Archeoastronomy

The interdisciplinary study that attempts to determine how much astronomy prehistoric people knew and how it influenced their lives. It involves multiple disciplines: astronomy to chart the heavens, archeology to probe the cultural context, engineering to survey sites, and ethnology to provide clues to the cultural past. Archeoastronomy has prompted valuable insights into the astronomy of the past, even to revolutionizing some models of prehistoric cultures. *See* ASTRONOMY.

It has been suggested that archeoastronomy and its loose family of disciplines should be subsumed under a broader study, cultural astronomy. The reason to do cultural astronomy is that the sky can perform a special role in the scheme of cultural systems. The sky then serves as a cultural resource of many uses.

The cultural context is the key to understanding the findings of archeoastronomers. Finding astronomical orientations at sites is easy; interpreting these as intentional alignments is hard. It is necessary to consider what their purpose might be (to keep a seasonal calendar? to regulate sacred time? both or neither?). The great danger is the imposition of modern astronomy and culture upon an alien culture of the past.

Naked-Eye Astronomy

The kinds of observations of key celestial cycles that can be made without a telescope are reviewed in this section. Only the Sun and the Moon will be considered.

Sun. Most people are aware that the height of the Sun in the sky at noon changes with the seasons, with the greatest height in summer and the lowest in winter. There is, however, less familiarity with the Sun's seasonal motion along the horizon. On the day of the summer solstice (around June 22), for

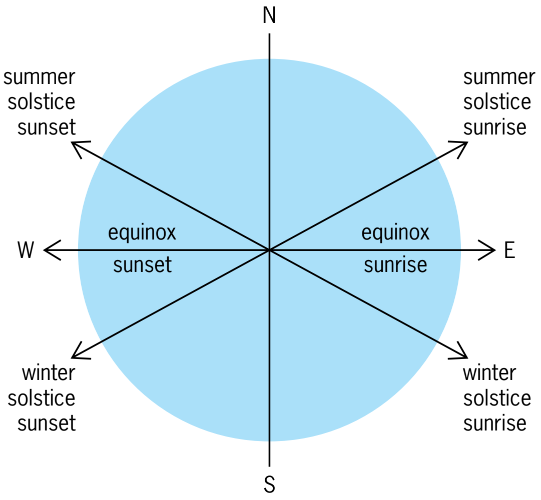


Fig. 1. Angular swing of the Sun along the horizon at rising and setting from solstice to solstice. The angle is for a latitude of 36°, that of the North American Southwest.

example, the Sun rises the farthest north of east that it will get for the year (**Fig. 1**). On the equinoxes (March 21 and September 23), it rises due east. And on the winter solstice (December 22), it reaches its farthest point south of east. The same occurs, mirror-reflected, at sunset. (This description refers to the Northern Hemisphere midlatitudes; in the Southern Hemisphere, summer and winter are reversed by date.) *See* EQUINOX; SEASONS; SOLSTICE.

Thus, from summer to winter, the sunrise point moves to the south; from winter to summer, to the north. The rate at which the sunrise point moves from day to day varies during the year. At the solstices, the sunrise points do not noticeably move for a few days. The Sun appears to "stand still" (which is the meaning of the word solstice). In contrast, at the equinoxes, the sunrise points move at their fastest rate, by almost the Sun's own diameter in a day at midlatitudes.

This seasonal voyage of the Sun along the horizon differs with latitude with respect to the size of the solstice-to-solstice swing along the horizon. At 36°, the latitude of the North American Southwest, the swing amounts to 60°, one-sixth of the total horizon circle (Fig. 1). Farther north, the arc is greater; at the latitude of Stonehenge (about 51°), it is about 80°.

The solstitial horizon points are symmetrical, and so they are also reversible. If an observer faces points that align to the summer solstice sunrise, then turning around and sighting along the points in the opposite direction reveals very closely the winter solstice sunset point (with a flat horizon).

Moon. The Moon's most obvious change is that of its phases. The month of phases (synodic month), the time from one phase of the Moon to the repetition of that phase, averages 29.5 days. *See* PHASE (ASTRONOMY).

Suppose the point of moonrise is observed for a month. It would be seen that the moonrise point varies from a point farthest south to one farthest north during the month. In other words, the moonrise motion mimics the sunrise motion but occurs about 12 times as fast. Depending on when the observations are made, the moonrise arc may be larger than, the same as, or smaller than the sunrise arc. This difference results because the Moon's path in the sky with respect to the stars is not the same as the Sun's, but is inclined at about 5°, crossing the Sun's path at two points. Thus, the Moon can appear as much as 5° below the Sun's path, 5° above it, or right on it. *See* MOON.

Complication. The matter is complicated in that the two points where the Sun's and Moon's paths cross (the nodes) move with respect to the stars, taking 18.6 years to circle the sky once. The result is that when the Moon's path reaches its highest point above the Sun's, the Moon's horizon swing is greater than the Sun's. When the two line up, the swings are the same. When the Moon's path falls below the Sun's, the total arc is less.

How much greater or less can be considerable. At a latitude of 36°, the Moon moves through a maximum arc of 70° and a minimum arc of 45° during the 18.6-year cycle (**Fig. 2**). In analogy to the Sun standing still at the solstices, the two extremes of the Moon's positions are also called standstills: major standstill for the maximum angle and minor standstill for the minimum, with 9.3 years between. Again in analogy to the Sun, these angular changes are more pronounced at more northern or southern latitudes.

When the Moon moves from minor to major standstill, its monthly arc stays within that of the Sun for about half the time. Then, an alignment that works for the Moon will work for the Sun at some time during the year. Hence, it may be that an orientation is not specifically for the Moon. However, during the time when the Moon's swing lies outside the angle of the Sun, an orientation can apply to the Moon only, and not at all to the Sun. (Again, these alignments are reversible.)

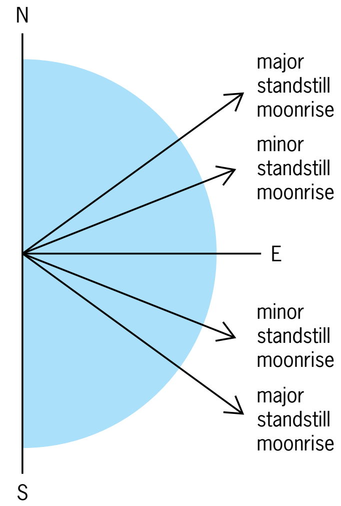

N
major standstill moonrise
minor standstill moonrise
E
minor standstill moonrise
major standstill moonrise
S

Fig. 2. Monthly angular swing of the Moon along the horizon at moonrise, for times of maximum angle (major standstill) and minimum angle (minor standstill). The angles are shown for a latitude of 36°.

Horizon-marking system. It can now be seen how a simple horizon-marking system is set up. First, a location with a clear view of the horizon must be found, with at least a few prominent features over the angular range of the sunrise (or sunset). Then it is necessary to return to this spot daily and note the rising positions of the Sun throughout the year at significant times: the solstices, the equinoxes, and, perhaps, important times to plant crops. Thus, a basic solar calendar is established. Since the Sun's positions at various dates along the arc remain fixed for a long time, once established the solar calendar will be good for many years.

A solar-horizon calendar keeps reliable track of the seasonal year. To subdivide that interval, people usually rely upon a lunar calendar that follows the phases of the Moon. The essential astronomical problem is that a seasonal cycle and a lunar-phase cycle are incommensurate. The solution can involve complicated, long-term strategies, such as that used in the Gregorian calendar, or more casual, practical ones, such as counting a "short" month of a few days at the end of each year. *See* CALENDAR.

The Old World: Stonehenge and Other Megalithic Sites

In a direct sense, Stonehenge created the interest in archeoastronomy, and the site exemplifies the problems and potential of the archeoastronomical enterprise.

Stonehenge is popularly known for the massive upright stones that form a central horseshoe and circle (some 82 ft or 25 m in diameter) in the center of the site (**Fig. 3**). Such large stones are commonly called megaliths in Great Britain; this term has come to be applied to all sites where stones, even fairly small ones, are arranged in some pattern. The horseshoe opens out on the main axis of Stonehenge, called the Avenue. Some 260 ft (80 m) from the center, within but not in the center of the Avenue, sits the tilted Heel Stone. This main axis of Stonehenge aligns roughly with the summer solstice sunrise. In fact, at summer solstice the Sun rises somewhat to the left of the Heel Stone as seen by an observer near the center of the structure.

Astronomical alignments at Stonehenge. The modern controversy concerning the astronomical use of Stonehenge developed in the 1960s. G. Hawkins, an astronomer, searched for astronomical alignments to the Sun, Moon, stars, and planets for the main features of the site. He found them for the Sun and the Moon, including moonrise and moonset during major and minor standstills. Later, he proposed that the site could even have been used to anticipate times of possible eclipses.

Radiocarbon dates indicate that Stonehenge was built over a span from 3100 to 1000 B.C. in three separate stages. The muddle over the astronomical use of Stonehenge comes, in large part, from the fact that it is a mosaic of structures, most likely built by different people, perhaps with a continuity of design. The great stones were erected between 2000 and 1500 B.C.; they attract the most attention now, but other aspects are also important.

It is the earliest parts of Stonehenge, constructed between 3100 and 2100 B.C., that have the most astronomical promise (**Fig. 4**). These comprise the outer earthwork ring and ditch (about 330 ft or 100 m in diameter and 7 ft or 2 m high) broken only in the direction of the Heel Stone; a ring of 56 holes (the Aubrey Holes) that were dug and then quickly filled with chalk; an array of postholes near the opening to the Heel Stone; and the four Station Stones that lie along the circle of the Aubrey Holes.

Lunar and solar observing can be done with these elements. The four Station Stones form a fairly good

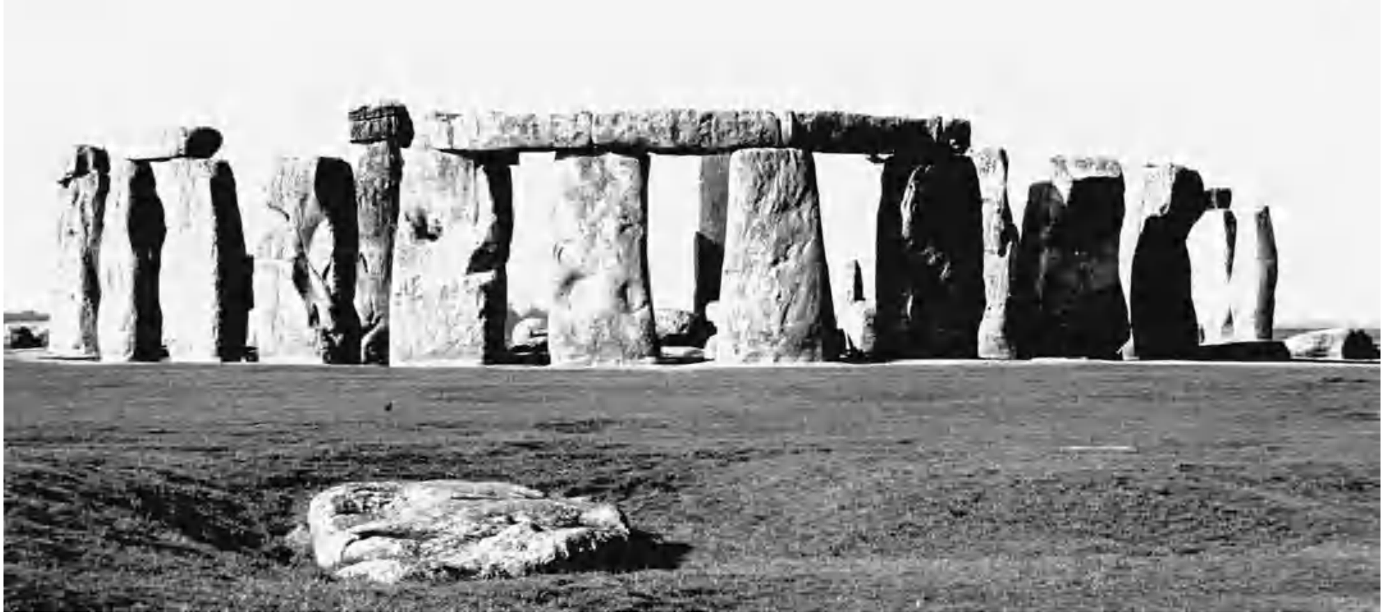


Fig. 3. Inner great trilithons of Stonehenge. (*Courtesy of O. Gingerich*)

rectangle. From its center, the summer solstice Sun rises along the opening to the Heel Stone. The short sides of the rectangle are parallel to this line, so they point to the summer solstice sunrise and winter solstice sunset. The long sides of the rectangle and its diagonals line up the moonrises and moonsets at the major and minor standstills. (These are reversible alignments.)

Hawkins also contended that the inner megaliths of the horseshoe sighting outward through the ring around them also aligned to important settings and risings of the Sun and Moon when viewed from inside the monument.

Evaluation of claims. Using positions near the inner megaliths as observing stations fails, whether for the Sun or the Moon. The gaps between the megaliths are wide, and depending on where the observer stands, the views can cover wide angles on the horizon. As noted above, even the central alignment—observing the summer solstice sunrise over the Heel Stone—does not work now, nor did it ever operate in the past 5000 years. Finally, it is not clear where the observer is supposed to stand among the massive stones.

J. North flipped the situation around so that the astronomical arguments gain considerable strength. In this interpretation, the observer stands outside the monument and uses it as the foresight. For example, an observer standing next to the Heel Stone would see standing stones and their lintels block out the glare of the western horizon at sunset. (It is again important that the extremes of the Sun and Moon are reversible alignments.) Only narrow sight lines are visible; one reveals the winter solstice setting Sun. As the Sun descends, the stones make a dark mass, pierced by sunlight through a central slit near the base of the stones. Also from the Heel Stone, the Moon's minor southern standstill appears through a central slit (not the same one as the winter solstice Sun).

The focus on the winter solstice sunset inspires a new view of the Avenue, which comes into Stonehenge from the summer solstice sunrise direction. As people in a precession around sunset of the winter solstice paced up the Avenue, they would behold first the top of the monument, and more would slide into view as they approached it. If timed right, they would reach the Heel Stone at just the moment of the final gleam of the winter sun.

The alignment of the sides of the inner rectangle holds up, in part because of its fairly large size (112 by 260 ft or 34 by 79 m). At Stonehenge's latitude, the solar and lunar extreme lines naturally cross almost at right angles; these astronomical symmetries seem to have caught the attention of the builders of this earlier part of the structure. The Station Stones fall closely around a circle with a radius and center roughly that defined by the Aubrey Stones. Using the edges of the Station Stones as backsights and the bank as an artificial horizon, the spectacle of the northern extreme moonset works out well. All told, the older parts of Stonehenge make a reasonable solar and lunar observatory.

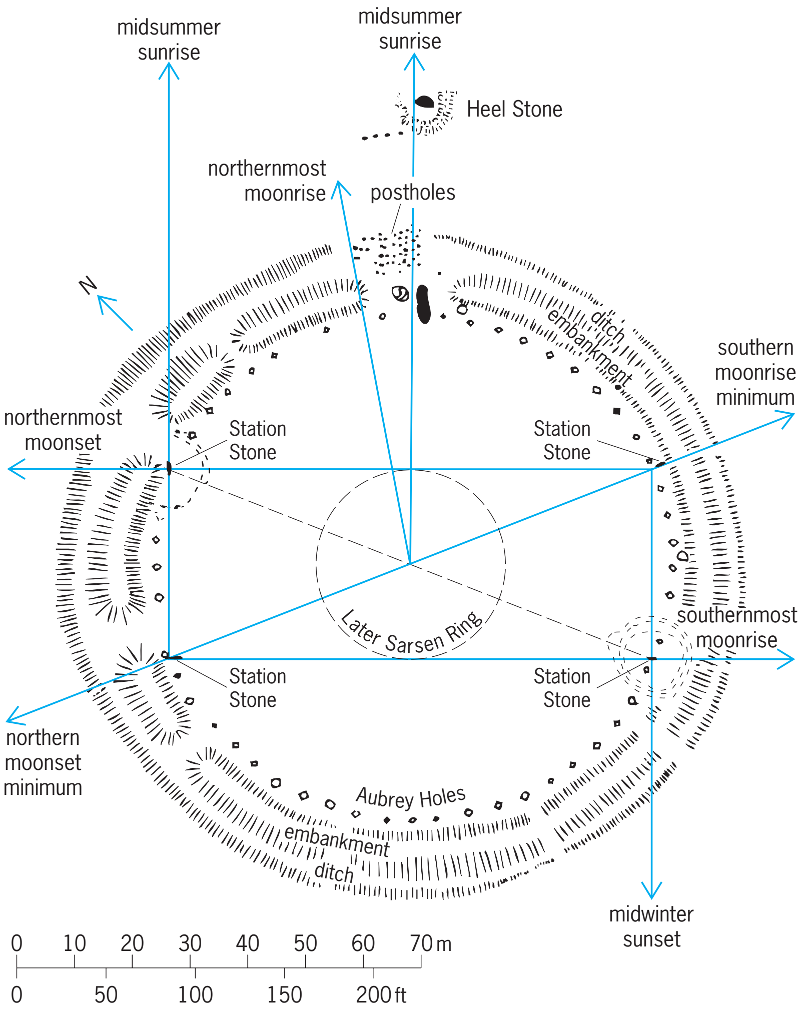


Fig. 4. Diagram of the major features of Stonehenge that may have been used astronomically. (*After O. Gingerich, in K. Brecher and M. Feirtag, eds., Astronomy of the Ancients, MIT Press, 1979*)

Other megalithic sites. The basic problem in all of this is that even if the astronomy works, the cultural context is unclear. Horizon watching can be used simply to tell the time of year or more forcefully to set a ritual calendar. One hope of guessing about the importance of astronomy in megalithic societies is to examine many sites to try to integrate astronomical and cultural evidence.

Two large burial sites, constructed about 3000–2500 B.C., also display winter solstice orientations. At Newgrange in Ireland, a long passage directs sunlight to reach the center of the tomb every morning for about a week around the winter solstice. At Maes Howe in the Scottish Orkney Islands, the setting Sun's rays illuminate the central chamber around the time of the winter solstice.

Hundreds of prehistoric sites in Great Britain and France were carefully surveyed by A. Thom, an engineer. He first found indications of alignments for the solstices and equinoxes, then for the lunar standstills. He promoted the idea that megalithic astronomers made extremely accurate observations of the Moon (using very distant foresights, tens of kilometers long) so as to pick out very small, long-term variations of the Moon's motions.

The precise lunar observations have been questioned, and different analyses lead to the conclusion that the Moon was observed, but not with the precision inferred by Thom. From the view of cultural necessity, it is unclear how such precision would benefit megalithic people in terms of simple survival value. However, more so than Hawkins's efforts, Thom's work forced archeologists to account for the astronomy in megalithic cultures.

Lacking rich cultural evidence, British investigations turn to the statistical evidence of a large number of sites, selected in an unbiased way, from ones of a similar cultural background. Such work has revealed (mostly for sites in Scotland) strong hints of rough (within a degree or so) orientations to the position of the Moon at major and minor standstills. Hence, general trends show up, even if embedded in a background of noise (that is, orientations that have no relation to astronomy).

Skywatching in the New World

Compared to the Old World, the New World archeoastronomer has the advantage of the survival of remnants of the cultures from pre-Columbian times. Even the great destruction wielded by the Spanish in Mesoamerica, especially their burning of Mayan books that contained much astronomy, could not wipe out completely the astronomy inherent in that culture.

In the Southwest, the Spanish encountered adobe villages, which they called pueblos, of the native peoples who had lived in them at least 1000 years. Many of these pueblos disappeared in historic times (from 1540 onward); those that survived are the cultural connection to the prehistoric people called the Anasazi, who occupied a vast area in the Southwest, centered on the Four Corners area (where New Mexico, Arizona, Utah, and Colorado now meet). Here stand ruins deserted from A.D. 1000 to 1400, stone and adobe constructions that provide some insight into the life of the Anasazi.

Pueblo astronomy. The Hopi (in Arizona) and Zuñi (in New Mexico) pueblos provide the best clues to the past because these villages were touched only lightly by the Spanish. Ethnographers gathered cultural information here around 1900, before the severe pressures on the part of Anglos occurred. It is inferred that the Hopi and Zuñi are cultural descendants of the Anasazi (although it is not known from which specific Anasazi sites). Thus, these pueblos preserve a remarkable cultural connection to prehistory.

Among the Hopi and Zuñi, astronomy played a central role in the agricultural and ceremonial life. The seasonal cycle of the Sun set the ritual calendar and determined the times of specific crop plantings and harvestings. The dry Southwest demands an observant farming, for raising crops is a marginal activity; in the past, failed crops could mean death.

The observing was invested in a religious office, usually called the Sun Priest. He watched daily from a special spot within the pueblo or not far outside it and carefully observed sunrise (occasionally sunset) relative to the horizon features. From past experience, he knew the points that marked the summer and winter solstices and the times to plant crops.

A crucial aspect of the Sun Priest's work was the ability to forecast ceremonial dates; he did so by making anticipatory observations about 2 weeks ahead of time. The rising (or setting) points of the Sun showed a daily change that could be reliably discerned against a horizon profile. By counting down a certain number of days, the Sun Priest could announce ahead of time the day for the ritual, allowing the people of the pueblo enough time to prepare for the ceremony. These forecasting procedures enabled the historic pueblo Sun Priest to predict the actual dates of the solstices with good accuracy, mostly within 1 day of the actual astronomical dates, as given in the historic records of the celebratory dates. The proper choice of ceremonial dates was the major responsibility of the Sun Priest, and it is likely that a prehistoric Sun Priest had the same responsibilities.

Along with horizon features, the Zuñi Sun Watcher, called Pekwin, used a natural pillar to chart the seasons. When the shadow cast by the pillar lined up in a special fashion, Pekwin knew that the summer solstice would soon occur. Also, within the pueblo, special windows and portholes allowed sunlight to hit special plates or markings on the walls at significant times of the year. Thus light and shadows, along with horizon features, made up the basis of the puebloan solar astronomy. The puebloan ancestors are believed to have done much the same as an adaptive survival strategy in a similar cultural context and environment.

Anasazi solar astronomy. Around A.D. 1000, the Anasazi prospered in the San Juan Basin and other regions of the Colorado Plateau. They built community houses that were four or five stories high, contained hundreds of rooms, and many large and small kivas—round, underground rooms used for ritual purposes.

Chaco Canyon in northwestern New Mexico grew to be a center of Anasazi culture. By 1130, eight large villages were located within 9 mi (15 km) of the central canyon. An extensive road system within the San Juan Basin, perhaps a trade network, connected Chaco to many outlying villages. Perhaps a few thousand people lived here in the large and small villages. These Chacoans faced climatic conditions similar to those of today; like the historic pueblos, they probably also had Sun Priests and seasonal solar calendars (and also lunar ones).

Three problems must be kept in mind when evaluating sites in Chaco Canyon. Many Chacoan buildings have been reconstructed, perhaps in ways quite different from the original way, so building orientations, unless based on general alignments of the original foundations, must be taken with caution. Second, few locations within the canyon offer good horizon profiles for a seasonal watch. Finally, it is not known whether each large village had its own, independent Sun Priest (as is true in the historic pueblos) or whether a central religious figure had the authority and responsibility for sun watching for the

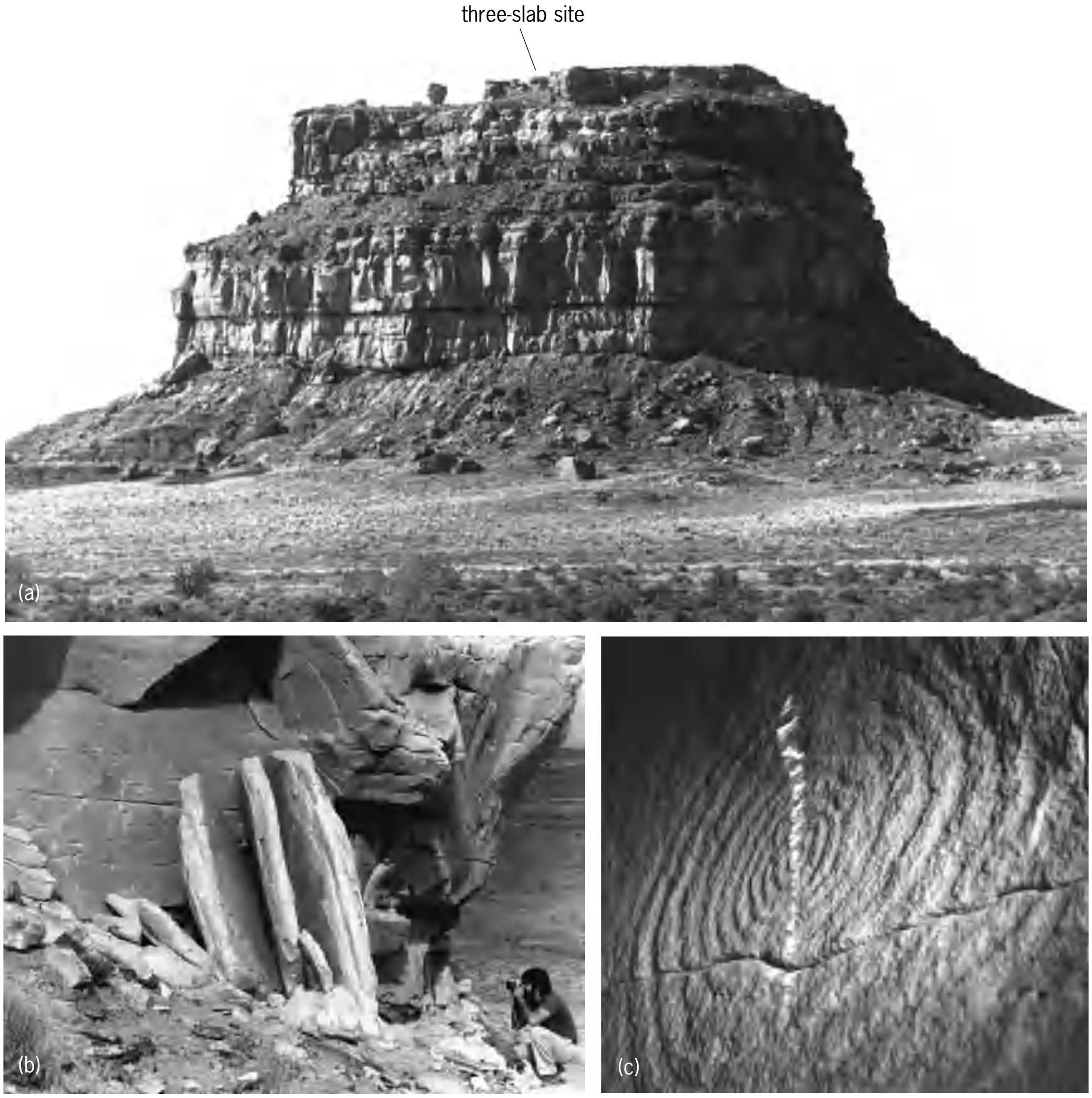


Fig. 5. Fajada Butte in Chaco Canyon. (*a*) View of entire butte (*photo by M. Zeilik*). (*b*) Rock slabs near the top of the butte. They rest against the rock surface on which two spirals are pecked. Their upper edges cause sunlight in late morning to play upon the rock face and around the spirals (*photo by R. Elston*). (*c*) Summer solstice sunlight cutting through the large spiral at about 11:13 a.m. in 1980. Due to a shift of the middle slab, the current view is somewhat different (*photo by W. Wampler*).

all the villages (including those at the outlier sites networked by the road system).

Fajada Butte. This butte (**Fig. 5*a***) thrusts upward at the eastern end of the canyon. Within 30 ft (10 m) of the summit, three rock slabs lie against the butte's southeast face (Fig. 5*b*). The slabs are a few meters long with about 4-in. (10-cm) gaps between them. They shield the rock face on which they rest from the Sun except at times before local solar noon. Then the edges of the slabs allow sunlight to strike the rock face, on which are carved two spirals: a large one (almost 1.5 ft or 0.5 m wide) right behind the slabs, and a smaller one above and to the left of the larger.

The spirals mark the Sun's yearly cycle by light patterns visible late in the morning. On the summer solstice, a shaft of light materializes above the large spiral at about 11 a.m. In about 20 min, it descends and slices through the heart of the spiral design (Fig. 5*c*). On the winter solstice, two shafts of light appear on the outside of the large spiral and pass through its outer edges at about 10 a.m. At both equinoxes, two shafts appear, one shorter and to the left of the other. The shorter shaft cuts through the center of the small spiral, while the larger one drops through one side of the large spiral.

It is possible that the play of light and shadow from a natural rock fall was noted by the Anasazi, and that they then made the spirals on the rock face. Certainly the equinoxes and solstices have distinctive patterns of light and shadow. However, it is unlikely that these patterns were used to forecast, say, the solstices with the 1-day precision achieved by the historic pueblo Sun Priests, since the horizontal motion of the main shaft of light is much too small on an average daily basis to predict the solstices any better than within a week or so. The site might well have served as a sun shrine, a place to which religious officials journeyed at important times of the year to place offerings to

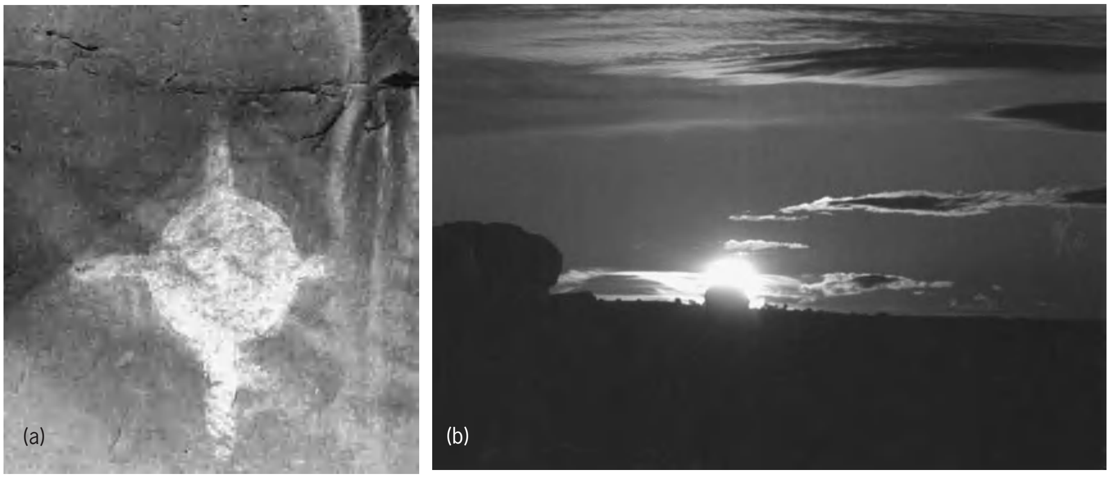


Fig. 6. **Wijiji site in Chaco Canyon. (*a*) Painted white sun symbol above the ledge near the ruin. The emblem resembles that of the Zia pueblo sun symbol and a sun symbol on a war shield from Jemez pueblo. (*b*) Winter solstice sunrise observed from the ledge at a position a few meters north of the sun symbol. This sunrise view is the same from the position at the sun symbol 16 days before the winter solstice. The angular width of the rock pillar is a little smaller than the angular diameter of the Sun. (*Photos by M. Zeilik*)**

the Sun. In the historic pueblos, sun shrines can be natural rock formations located some distance from the pueblo; a priest would visit them before sunrise.

Unfortunately, it may never be possible to validate these interpretations about the site. The middle rock slab moved in the mid-1980s, and the pattern of the light shafts has changed considerably. Hence, what is now seen is not what was seen before, and may not have been the pattern viewed by the Anasazi if erosion has shifted the slabs in the past 1000 years.

Wijiji. This is a pueblo built about A.D. 1100, late in Chaco's history. About 0.6 km (1 mi) to the east of it, a large rincon (valley) opens up in the mesa. On the northwest side of the rincon runs a narrow ledge, which can be reached by climbing a prehistoric staircase. Here on the wall is painted a large four-pointed symbol (**Fig. 6*a***) that resembles the Zia Pueblo sun sign.

North of the symbol, three boulders rest on the ledge; the largest has a double spiral carved in its surface. The design and technique are clearly Anasazi. Eastward from the ledge, a large rock pillar, across the rincon, rises above the horizon. From a spot a few meters south of the boulder with the double spiral, the winter solstice Sun rises behind the pillar (Fig. 6*b*). Because the angular width of the pillar against the sky is somewhat smaller than the angular diameter of the Sun, the winter solstice sunrise behind the pillar is a clear event when viewed from the location near the boulders. If an observer moves a few meters (yards) to either side, the shift moves the Sun to either side of the pillar. Near the sun symbol the Sun rises behind the pillar 16 days prior to the winter solstice, making possible an anticipatory observation that works extremely well to forecast the day of the winter solstice.

The area around the ledge contains much rock art, some of it Navajo, and it has been argued that the white "sun" symbol is Navajo in origin and that the site was used by Navajo (perhaps in the late seventeenth century) for sun watching. That may be the case, but there are certainly Anasazi relics here, too. Thus the site may have been first used by the Anasazi and then adopted by the Navajo for ceremonial purposes.

Anasazi lunar astronomy. There is no firm evidence so far of prehistoric monthly lunar calendars among the Anasazi. However, occasional claims have been made that the Anasazi noted the 18.6-year standstill cycle of the Moon. The most convincing case for attention to this interval has been made for the Chimney Rock Archaeological Area in Colorado, where the natural rock pillars act as a natural foresight when viewed from the site of Chimney Rock Pueblo, a Chacoan outlier, which was occupied from about 1075 to 1175. The orientation allows an anticipation and forecasting of the major lunar standstills. For about 2.5 years prior to the standstill, the Moon rises between the pillars for 1 or 2 days per month. Starting after the summer solstice, the Moon appears as a waxing crescent. Finally, near the winter solstice, the full moon stands between the pillars at around sunset. Hence, the priest in charge of the moon watching could forecast the date of the standstill and also the full moon nearest the winter solstice, an important conjunction among the historic pueblos.

Regional ceremonialism. Many models have been proposed for the development of Chaco Canyon as the regional center of the San Juan Basin. The "economic redistribution" model is fairly plausible. Here the roads serve to direct people to Chaco to trade goods, which they take back to their home pueblos.

Astronomical scheduling may have been used to tell people when to commence their travels. In a decentralized social system, each outlying pueblo might have its own Sun Priest tracking time locally. He would anticipate important ceremonial dates, such as the winter solstice, so that the people of the pueblo could prepare for a pilgrimage to Chaco Canyon. Other Sun Priests would do the same so that the arrivals at Chaco would be synchronized by the sunwatching.

The concept of regional ceremonialism scheduled by astronomy may serve as a synthesizing concept for other prehistoric sites. For instance, there is increasing evidence that Stonehenge played a central role in a network of ritual sites with the winter solstice ceremony as a cosmic organizer in sacred space and time.

Other Places and Cultures

Every prehistoric culture appears to have developed its own astronomy. The traditional navigators of Oceania needed to memorize guide stars and to employ them as the bearing markers for island-to-island travel over thousands of kilometers of water. The Carib people of northern South American developed a calendar that relied on the positions of stars relative to each other and to the Sun at times of rising and setting. A bone from the shores of Lake Edward, Zaire and Uganda, may have markings of a lunar calendar, tallied at a time over 8000 years ago, perhaps used to forecast marine activity or the weather; a focus on the Moon continues in Africa today. Chinese astronomical records cut into bones and shells may have begun as early as the twelfth century B.C., well before the Babylonians incised their earliest records on clay.

By the time the Spanish invaded Mesoamerica and South America, the use of astronomy went well beyond complex systems of cycles and calendars. For instance, in some Mayan cities knowledge of the cycles of Venus timed the onset of warfare. In others, key political events incorporated the summer solstice and perhaps conjunctions of Jupiter and Saturn. Ancient Inca city planning and politics embodied astronomy, such as the ceque system of radial lines from the Temple of the Sun in the valley of Cuzco. These lines mark the directions to sacred places as well as to specific astronomical phenomena. The ceque system had a calendric manifestation in knotted cords that tallied the days of the agricultural year. Michael Zeilik

Bibliography. A. F. Aveni, *Ancient Astronomers*, St. Remy Press/Smithsonian Books, 1993; J. M. Malville and C. Putnam, *Prehistoric Astronomy in the Southwest*, rev. ed., Johnson Books, 1993; J. North, *Stonehenge: A New Interpretation of Prehistoric Man and the Cosmos*, Free Press, 1996; C. L. N. Ruggles and N. Saunders (eds.), *Astronomies and Cultures*, University of Colorado Press, 1993; R. A. Williamson and C. R. Farrer, *Earth & Sky: Visions of the Cosmos in Native American Folklore*, University of New Mexico Press, 1992.

Archeological chemistry

The application of chemical techniques to the study of archeological finds, natural or anthropogenic, in order to ascertain their composition or, in some cases, their age. Traditional chemical analysis uses wet methods, in which a sample is brought into solution and its components are assayed by precipitation or titration. These methods were applied to ancient coins as early as the late eighteenth centuary. The obvious need to minimize damage to an irreplaceable object spurred the development of microchemical techniques. Modern analysis relies on instrumental methods that require only very small samples or are entirely nondestructive. Although these methods rely on physical phenomena rather than chemical transformation, all procedures that are capable of the qualitative and quantitative determination of the atomic or molecular composition of the object under study are usually included under the broad heading of archeological chemistry.

Methods. Various analytical methods are utilized in archeological chemistry, as described below. It should be noted that these methods of analysis are not competing but complementary. The choice of method depends on the nature of the object, on the elements to be determined, and on the accuracy required.

Optical emission spectography. The oldest instrumental method of analysis is optical emission spectrography, in which thermally excited atoms emit light at wavelengths corresponding to energy differences between electronic orbitals. A set of wavelengths identifies a given atom; the intensity of the emitted light, measured photographically or by means of a photomultiplier tube, is proportional to the concentration of that atom. Optical emission spectrography has been largely replaced by more sensitive methods covering a wider range of elements.

Atomic absorption spectroscopy. In this method the atoms in a sample flame absorb part of the energy of a reference light source containing the same atom. Accuracies range from ±1% for major components to ±15% for trace elements in the parts-per-million range.

Inductively coupled plasma. In this modern version of emission spectroscopy, an inductively coupled plasma is the means of excitation, attaining temperatures in excess of 8,000–10,000°C (14,400–18,000°F).

Neutron activation analysis. This method involves the bombardment of the sample with slow neutrons in a nuclear reactor. Some atoms are converted to radioactive isotopes whose subsequent decay produces gamma rays that are detected by a semiconductor counter coupled with a multichannel pulse analyzer to yield both qualitative and quantitative data. Small objects can be irradiated in their entirety, obviating sample removal and making the method nondestructive. Variations include instrumental neutron activation analysis (INNA), proton-induced gamma-ray emission (PIGE), and prompt gamma neutron activation analysis (PGNAA), in which the gamma rays of short-lived radioisotopes are analyzed while the sample absorbs neutrons.

X-ray fluorescence spectrometry. The sample, which may be the entire object, is exposed to x-rays that excite electrons in the lower shells to higher energy levels. When these electrons drop back to lower levels, secondary or fluorescent x-rays are emitted and analyzed for energy and intensity, thus providing identification of the elements present as well as their concentrations. X-ray fluorescence offers accuracy of ±2–5% of concentrations from 10 ppm to 100%, but since x-rays penetrate only to depths of 20–200 micrometers, only the surface of an artifact will be analyzed.

Electron microprobe analysis and proton-induced x-ray emission. The related techniques of electron microprobe analysis, which uses high–energy electrons for excitation, and proton-induced x-ray emission (PIXE), which uses protons, permit focusing the incident beam on areas of 1 μm^2 and 1 mm^2, respectively. This not only makes it possible to analyze small regions of heterogeneous surfaces, like those of painted pottery, but also allows scans to be made across the surface of a cross section to yield compositions as a function of depth, although only at the expense of cutting through the object. Both methods can be applied to bulk analysis by grinding up a sample and converting it to a glass bead or pressed pellet.

Auger electron spectroscopy. This method is similar to the electron microprobe technique, but measures electrons ejected by the absorption of x-rays in the matrix.

X-ray photoelectron spectroscopy. This method also ejects electrons by irradiation with x-rays and has the advantage of identifying the valence state of atoms. Photoelectron spectroscopy bridges the gap between elemental analysis and the determination of the valence state of ions and of the identification of molecular species, as does Mössbauer spectroscopy, which specifically reveals the oxidation state of iron. *See* ACTIVATION ANALYSIS; ATOMIC SPECTROMETRY; AUGER EFFECT; MÖSSBAUER EFFECT; PROTON-INDUCED X-RAY EMISSION (PIXE); RADIOISOTOPE; X-RAY FLUORESCENCE ANALYSIS.

Organic archeometry. Organic materials constitute only a small portion of archeological finds, but since they include such basic necessities as food, drink, and clothing, they have the potential of revealing much about past life. Because they consist of covalently bound, complex, and sensitive molecules, their study requires special methods of analysis. Organic archeometry is the newest and most rapidly expanding field of archeological chemistry. Organic dyes have long been determined qualitatively and quantitatively by absorption spectroscopy in the visible and ultraviolet ranges. The extension into the infrared range allows not only the identification of organic materials by visual or computer-aided comparison of infrared spectra (“fingerprinting”) but also some structural interpretation. Since organic residues typically consist of mixtures of dozens or even hundreds of individual compounds, the progress of organic archeometry has crucially depended on the development of chromatographic separation procedures. These include column chromatography, paper and thin-layer chromatography, gas chromatography, with or without prior pyrolysis, and liquid chromatography.

All of these techniques not only separate mixtures into individual components but permit their identification if the rate at which they travel through the chromatographic substrate, the retention time, can be matched to those of authentic reference compounds. The most powerful techniques combine a separation step with a spectroscopic identification method.

The detection of components in liquid chromatography is commonly made by ultraviolet spectroscopy, but it can now also be coupled with mass spectrometry. Gas chromatography has been coupled with infrared spectroscopy and, most powerfully, with mass spectrometry. In mass spectroscopy, individual components are converted to usually positively charged ions by electron impact or by chemical ionization. These ions are separated by electrical and magnetic fields and recorded separately. The resulting mass spectrum can identify even very complex molecules with virtual certainty. This has made possible the identification of food remains in vessels, house floors, and refuse middens.

An important adjunct to the study of prehistoric diets is the determination of stable isotopes of carbon and nitrogen by mass spectrometry, which has been applied to collagen extracted from skeletal remains. A recent advance is a single instrument that first separates organic components by gas chromatography, then converts them singly into carbon dioxide by combustion, and lastly determines the isotopic composition of each component by mass spectrometry.

Another method that is gaining use in organic archeometry is nuclear magnetic resonance spectrometry (NMR), which detects a limited number of atomic nuclei, among them ordinary hydrogen, carbon-13, nitrogen-15, fluorine-19, and phosphorus-31, by their simultaneous interaction with an external magnetic field and a radio-frequency field. The sample requirements are too high to apply this technique to the small amounts of compounds emerging from a capillary gas chromatograph (although a preparative gas chromatograph would make this possible), and the archeological applications of both proton magnetic resonance (PMR) and carbon magnetic resonance (CMR) have been to unseparated mixtures, including fossil resins and tars. *See* GAS CHROMATOGRAPHY; INFRARED SPECTROSCOPY; MASS SPECTROMETRY; NUCLEAR MAGNETIC RESONANCE (NMR); SPECTROSCOPY.

Identification. The determination of the chemical composition of an archeological find is not an end in itself, but provides the archeologist with factual evidence not otherwise obtainable and touching on many aspects of early human life. The changing elemental composition of coins detects progressive debasement and reveals economic history and fiscal policy. The metals added to copper to make bronze and brass outline the history and spread of technology in general and of metallurgy in particular. The foodstuffs consumed are indicators of the advent and progress of agriculture and animal husbandry. Together, all these paint a picture of prehistoric social, cultural, and economic stratification. *See* PREHISTORIC TECHNOLOGY.

Provenience analysis. The chemical analysis of archeological artifacts has firmly established that the composition of an object offers clues to its geographic origin, which may be far from the excavation site, as had been evident in some cases from purely typological considerations. This provides evidence of trade and exchange in commodities and raw materials. The elemental or molecular makeup of natural products differs from one deposit to another. Matching the composition of artifacts to that of all possible sources reveals extensive trade over impressive distances. For obsidian, a natural volcanic glass, this has been done by elemental analysis; for amber, a fossil resin, by infrared and carbon resonance spectrometry. The composition of manufactured objects reflects their source less directly because the transformation of clay into ceramics or of ore into metal may cause changes in composition. In addition, there is the problem that deposits of clay or ore may have been exhausted in antiquity and therefore are no longer available for comparison. Provenience analysis of metals is especially difficult because the reuse of broken implements from various sources, possibly together with new ore, will lead to the superimposition of a number of compositional indicators. Nevertheless, the use of sophisticated statistical procedures, notably cluster analysis, has made it possible to distinguish local ceramic and metal products from imports and, in many cases, to identify the sources of the latter.

Chemical dating. While the most widely used methods for dating archeological material—radioactive decay, thermoluminescence, and archeomagnetism—deal with physical processes, three depend on the progress of conventional chemical reactions. (1) Amino acid dating uses the rate of racemization of optically active organic molecules. (2) Hydration dating measures the thickness of the weathering layer produced by the action of water on natural and artificial glass, including obsidian and flint. (3) NFU dating of bone relies on the loss of nitrogen (N) from the organic collagen component and on the uptake of fluorine (F) and uranium (U) by the inorganic hydroxyapatite component. Like all nonnuclear chemical reactions, these changes are a function not only of time but also of temperature, of acidity and, in the case of fluorine and uranium uptake, of the concentrations of these elements in the surrounding soil. Chemical methods cannot produce absolute dates unless these other variables are known or can be estimated reasonably closely. They are, however, useful in establishing relative ages of finds within a single site in which the depositional characteristics are likely to have been uniform. *See* AMINO ACID DATING; ARCHEOLOGICAL CHRONOLOGY; CHEMICAL MICROSCOPY; DATING METHODS; PALEOMAGNETISM; RACEMIZATION; RADIOCARBON DATING; THERMOLUMINESCENCE. Curt W. Beck

Bibliography. J. B. Lambert, *Traces of the Past: Unraveling the Secrets of Archaeology through Chemistry*, 1997; A. M. Pollard (ed.), *New Developments in Archaeological Science*, 1992; A. M. Pollard and C. Heron, *Archaeological Chemistry*, 1996; S. U. Wisseman and W. S. Williams (eds.), *Ancient Technologies and Archaeological Materials*, 1994.

Archeological chronology

The establishment of the temporal sequence of prehistoric events. During the nineteenth and early twentieth centuries archeologists devised a number of methods to construct a relative chronology for cultural remains. These included stratigraphy (correlation of layered sediments and rocks), seriation (ordering of materials based on changes in style), and cross-dating (using temporally diagnostic ceramic and stone tool types). Beginning in the early twentieth century, dating methods developed in the natural sciences began to be applied to archeological remains to provide absolute (that is, calendar) dates. These are sometimes called chronometric dating methods, and the more commonly used ones are described here. *See* ARCHEOLOGY; DATING METHODS.

Radioactive decay methods. The decay of a radioactive material is monitored by its half-life, the amount of time it takes for half of a given amount to decay. Because half-lives for any given radioactive isotope are nearly constant, this information can be used to measure time. *See* HALF-LIFE.

Radiocarbon dating. Radiocarbon dating is the most commonly used chronometric method in archeology. The radiocarbon isotope carbon-14 is produced in the atmosphere by cosmic-ray bombardment of nitrogen-14. Carbon-14 is quickly spread throughout the biosphere beginning with photosynthesis and dissolution in the ocean and other water bodies, so that plants and animals have roughly the same concentration of carbon-14 as the atmosphere. When a plant or animal dies, it ceases to interact with the carbon cycle, and the amount of carbon-14 diminishes according to its half-life of about 5730 years. By measuring the current decay rate (conventionally by counting the number of beta particles detected over a given time) or the ratio of carbon-14 to stable carbon isotopes (by accelerator mass spectrometry, AMS), the time since the organism died can be derived. Any organic remains found at archeological sites (such as bones, shells, and preserved plant materials) can therefore potentially be dated, and AMS allows dating of very small samples. Radiocarbon dates are given in terms of radiocarbon years, which must be converted to calendar years, because the rate of cosmic-ray bombardment has varied through time. Correlations have been established between carbon-14 and tree-ring dating (dendrochronology), which are used for calibration back to about 10,000 years; correlations based on uranium series dating on corals are used for calibration back to about 30,000 years. The upper limit of radiocarbon dating is about 50,000 years. *See* BIOGEOCHEMISTRY; RADIOCARBON DATING.

Uranium-series dating. Uranium-series dating involves the use of uranium-238 (and, less commonly, uranium-235), which has a very long half-life and decays through a succession of “daughter” radioisotopes, each with a much shorter half-life. The decay series ends with lead-206, which is stable. Over a long enough time, the radioactivity of the daughters becomes the same as that of the parent uranium, a situation called secular equilibrium. However, uranium can be soluble in water, whereas most of the daughter radioisotopes cannot. Water-precipitated materials (such as stalagmites, stalactites, and travertines) or organisms such as coral, which derive carbonate material from the water, initially contain only uranium without the daughters. But after deposition, decay of the uranium leads to the buildup of the daughters until equilibrium is reestablished. This provides a clock, extending back about 500,000 years, to measure when precipitation occurred. Due to this long dating range and because much of the early archeological record is contained in limestone caves and rock shelters, uranium-series dating has been very important in dating the Paleolithic period. *See* LEAD ISOTOPES (GEOCHEMISTRY); RADIOISOTOPE (GEOCHEMISTRY).

Potassium-argon dating. Potassium-argon dating is used to date volcanic events. The radioactive isotope potassium-40, which decays to argon-40, is commonly found in volcanic rocks. During a volcanic eruption, the argon gas escapes. After cooling, the decay of potassium-40 leads to a new buildup of argon in the rocks. The ratio of potassium-40 to argon-40 determines the time since the eruption. Traditionally, potassium and argon were measured separately, but a more recent technique is to bombard the sample with neutrons, which converts potassium-40 to argon-39. Then both argon-40 and argon-39 can be measured simultaneously by mass spectrometry. This technique, known as the Ar-Ar method, provides more accurate and precise results. Although potassium-argon dating is restricted to dating volcanic events, it has been very important in dating early human evolution in East Africa, where cultural materials are often separated by volcanic ash layers. *See* GEOLOGIC TIME SCALE; ROCK AGE DETERMINATION.

Radiation-exposure methods. In some materials natural radioactivity produces free electrons that can be subsequently trapped in crystalline defects. The trapped charge accumulates at a rate governed by the natural radioactive dose rate, which is usually fairly constant. This accumulation, calibrated by laboratory irradiation as an equivalent dose, is proportional to the time since the traps were last emptied. Dividing by the dose rate yields an age.

Luminescence dating. Luminescence dating uses heat (thermoluminescence, TL) or light (optically stimulated luminescence, OSL) to empty the electron traps, resulting in the emission of light, or luminescence. The intensity of the luminescence signal is proportional to the time since the sample was last heated or exposed to light. It thus provides a way to date pottery, stone tools that have been accidentally or deliberately heated, or buried sediments. Quartz and feldspar are the two most common minerals dated, and their ubiquity at archeological sites, either in pottery or stone tools or in archeological sediments, gives luminescence dating wide applicability. Recent research on OSL, in particular, has greatly increased both the accuracy and precision of the method, and dates of up to 100,000 years and more are routinely obtained. OSL dating of single grains of quartz or feldspar is the newest development, opening up the possibility of using luminescence to study site formation processes. *See* THERMOLUMINESCENCE.

Electron spin resonance. Electron spin resonance measures the trapped electron concentration directly by inducing a change in the quantum spin of the electrons via exposure to microwave radiation in the presence of a magnetic field. This method has been commonly used to date mammalian tooth enamel (to the time of its formation), as well as shells, coral, and quartz (to time of last heating or exposure to light). It is mainly used for dating older materials, and has an upper age limit of about 2 million years, giving it a prominent role in understanding the chronology of early human evolution. *See* SPIN LABEL.

Climatic variation methods. Variations in climatic parameters such as temperature and precipitation are often recorded in biological structures and sediment accumulations. Use of pollen compositions, lake sediments, and loess deposits to track temporal climatic trends has a long history.

Dendrochronology. Dendrochronology uses climatic-induced variations in annual tree rings to produce single-year resolution dating. Each ring consists of an inner band of light-colored wood and an outer band of dark-colored wood, reflecting different parts of the growing season. The width and density of the rings in some trees is strongly influenced by temperature and precipitation, and the resulting pattern of changing thicknesses and densities is reproduced by all trees of the same species in a given region. This allows cross-referencing of the patterns from living trees to ancient wood remains until a master sequence is developed against which wood from archeological sites can be referenced and thus dated. The advantages of high resolution are countered by the disadvantage of limited applicability. Although dendrochronology has been applied all over the world, the most developed sequences are from conifers in the American southwest and oaks in northern Europe, where master sequences in both places extend more than 10,000 years. *See* DENDROCHRONOLOGY.

Oxygen isotope ratio dating. Oxygen isotope ratios recorded in deep-sea and ice cores also provide high-resolution dating. Oxygen has three isotopes, the lightest and most common being oxygen-16 and the heaviest being oxygen-18. Vapor (and thus rain and snow) above the ocean has a higher ratio of ^{16}O to ^{18}O than the ocean. During glacial periods, the lighter snow gets locked up in glaciers, so that the ocean is isotopically heavier than during nonglacial periods. These ratios are preserved in the tiny shells of foraminifera, which accumulate on the ocean floor. Cores of ocean sediments thus contain a record of past climate. Similar records are available from ice cores (with the ratios calculated on the ice itself) from polar regions. The oxygen isotope records provide only relative dating based on stratigraphy. Attempts to anchor the sequence in absolute time have most commonly used paleomagnetism, carbon-14, and uranium-series dating. Precise dates for the isotope sequence have also been obtained by alignment with astronomical data, on the justification that ice-age cycles are caused by minor periodic variations in the Earth's orbit. Application to archeological dating requires correlating the ocean data with shellfish isotope ratios or with terrestrial climatic records, which can be difficult.

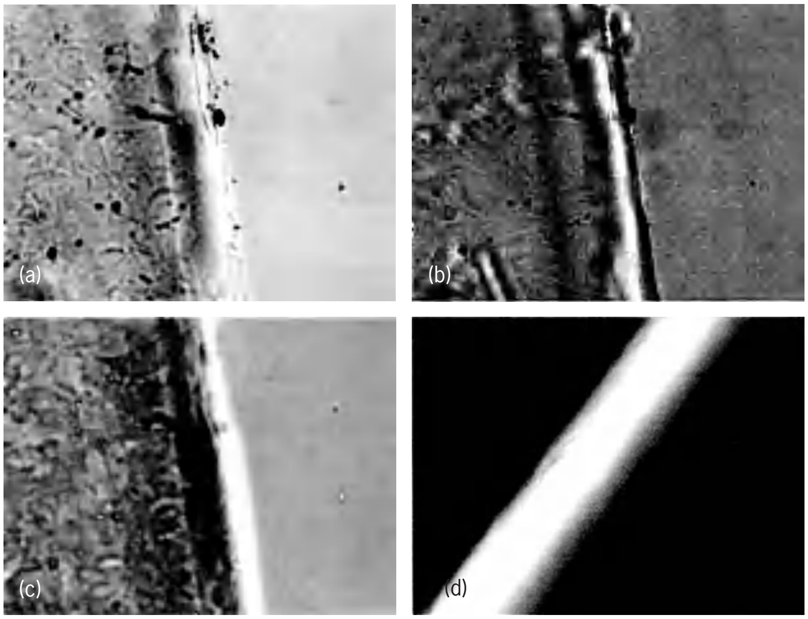
Weathering methods. Archaeologists have put considerable effort into trying to date materials by their degree of weathering. Thickness of patinas (thin, colored films) on rocks, degradation of carbon, uptake of fluorine in bones, and the extent of lichen growth on rocks have all be studied.

Obsidian hydration dating. Obsidian hydration dating is the most commonly used weathering dating method, because obsidian (volcanic rock that has cooled too rapidly to crystallize) was often used to make stone tools in prehistoric times (see **illustration**). Water diffuses into the surface of freshly fractured obsidian and forms a hydration rind. The rind thickness increases with time, but it is also a function of temperature, composition, burial depth, and other factors, all of which have to be controlled for accurate dating. Experts in the field disagree on how the rind should be measured and on the proper equation for describing the diffusion process. Most success has been achieved by comparing rind thicknesses with known-age obsidians from nearby sites or by just dating on a relative scale. *See* HYDRATION; OBSIDIAN.

Amino acid racemization. Amino acid racemization refers to the conversion of the chemical structure of amino acids, the building blocks of proteins, from "left-hand" arrangements to "right-hand" arrangements after an organism dies. The conversion rate toward an equilibrium level, at which there are equal numbers of both arrangements, is a function of time but also depends on climatic variables, particularly temperature. Early applications to bone were not successful, but recent work with chemically stable egg shells from large flightless birds, such as ostriches and emus, has shown more promise. These egg shells are common in archeological sites in southern Africa and Australia, so most work is now being done in those two areas. *See* AMINO ACIDS; RACEMIZATION.

Archaeomagnetism. In a class of its own, archaeomagnetism is based on temporal changes in the location of the Earth's magnetic pole and in the intensity of the Earth's magnetic field. When a sample is heated to a high temperature, magnetic dipoles within iron-bearing inclusions align themselves with the Earth's magnetic field. When the sample cools, the alignments are frozen in that position. Dating requires the construction of a master curve on which known past positions and intensities are plotted. The alignment recorded from an archeological sample, such as a hearth or a kiln, can then be matched with the master curve and thus dated. Dating is limited to areas for which such curves have been constructed and rarely extends further back than 10,000 years. The Earth's magnetic poles have also reversed from north to south and back several times through Earth's history, and these reversals have been recorded and dated (by potassium-argon dating) to allow coarse resolution dating on a geological scale. *See* PALEOMAGNETISM.

James K. Feathers



Photomicrographs of a hydration rim of a 150,000-year-old obsidian tool from Ethiopia taken using different types of imaging. Each image is 0.125 mm wide. (*a*) Using conventional light at optimum focus, the rim appears bounded by bright lines, demarcating it from the obsidian to the left and the mounting medium to the right. The bright lines are probably functions of varied refraction indices among the unhydrated obsidian, the hydration rim, and the mounting medium. (*b*) Reduction of the light enhances the boundary between the rim and the mounting medium. (*c*) Bringing the microscope out of focus moves both lines to the right, which is expected if the lines are refraction artifacts. (*d*) In cross-polarized light, the boundaries are shown to be gradational rather than sharp, in contrast to earlier assumptions. (*Reprinted from L. M. Anovitz et al., The failure of obsidian hydration dating: Sources, implications and new directions, J. Archeolog. Sci., 26:*735–752, *©1999, with permission from Elsevier*)

Bibliography. M. J. Aitken, *An Introduction to Optical Dating*, Oxford University Press, Oxford, 1998; M. G. L. Baillie, *A Slice through Time*, B. T. Batsford, London, 1995; S. Bowman, *Radiocarbon Dating*, British Museum Publications, London, 1990; M. J. O'Brien and R. L. Lyman, *Seriation, Stratigraphy and Index Fossils: The Backbone of Archaeological Dating*, Kluwer Academic, New York, 1990; R. E. Taylor and M. J. Aitken (eds.), *Chronometric Dating in Archaeology*, Plenum Press, New York, 1997; G. A. Wagner, *Age Determination of Young Rocks and Artifacts: Physical and Chemical Clocks in Quaternary Geology and Archaeology* (translated from German), Springer, Berlin, 1998.

Archeology

The scientific study of past material culture. The initial objective of archeology is to construct cultural chronologies, attempting to order past material culture into meaningful temporal segments. The intermediate objective is to breathe life into these chronologies by reconstructing past lifeways. The ultimate objective of contemporary archeology is to determine the cultural processes that underlie human behavior, both past and present.

Thus, archeology is both scientific and humanistic. Throughout many parts of the world, archeology is considered a subdiscipline of anthropology, focusing on the anthropology of extinct cultures. In other parts of the world, archeology is regarded as an extension of history, attempting to write a prehistory of people who lack an adequate history of their own.

The material culture of the past is of infinite variety. It is the scientific study of this evidence which differentiates the contemporary archeologist from the nineteenth-century antiquarian. So broad is the task that there is no such thing as any single archeological method, although over the past century archeologists have evolved what can be termed an overall archeological approach.

Methods. By constant confirmation, the archeologist often attempts to establish synchronism with what has already been established historically. In 1945, for instance, Mortimer Wheeler found at Pondicherry pottery shards of well-dated Roman types in close association with certain East Indian wares. Subsequently, he was able to establish a date for the unknown Indian ceramics on the basis of the historically dated Roman ceramics.

Similarly, scarabs exported and traded from the Nile Valley into Palestine, the Sudan, and further afield have been dated with reference to the Egyptian king lists, providing at least approximate dates hundreds of miles away from the place of manufacture. Even when dates in years are less readily available, some kind of relative synchronization may be attempted, such as when similar bronze axes or sword hilts were discovered with locally produced objects in Sweden, Britain, and Spain, or when daggers of Mycenaean type were found in Wiltshire, England.

Types. Archeologists use a number of types in order to categorize similar artifacts. Most common is the temporal type, a principle similar to the index fossil concept used by the geologist. A temporal type can be any kind of archeological artifact or feature, but ideally it is some object of common use in which the form is subject to change, due to either the whim of fashion or technological improvement. One example is the simple flint arrowhead with side barbs and central tang. It is typical of the British Bronze Age and was not in fashion earlier or later. Ceramic types have been established by archeologists working around the world, and a thoroughly tested ceramic chronology is invaluable as a temporal ordering device, no matter where the archeologist is working. *See* INDEX FOSSIL.

The nature of the artifact employed as a temporal type is irrelevant, and its use may not even be known. In the future, for instance, the metal springs of modern clothespins remaining after the wooden parts have decayed may serve, for the archeologist, as a means to define periods and cultural associations, even though the original use may be lost in the past.

Archeologists also establish other types. Functional types attempt to group artifacts on the basis of known or presumed functions. Function may be established through experimentation, through analogy to existing primitive groups, or through microscopic analysis of damage to the tool itself. Archeologists also establish technological types, divisions which reflect the mode of manufacture. Technological types are particularly helpful when studying stone-tool manufacture.

Cultures. The concept of culture is used in two different ways by contemporary archeologists. When dealing with cultural chronologies, the archeologist most commonly uses a modal or shared view of culture. It is this normative collection of shared ideas which cause artifacts to change in systematic ways through time, and temporal types can be established on the basis of this shared culture. When attempting to reconstruct lifeways, however, the archeologist can no longer rely on the shared aspects of culture. When transcending temporal associations, contemporary archeologists tend to view culture systematically, as people's extrasomatic (that is, learned) method of dealing with the social and cultural environment. In this sense, one does not share a cultural system—one participates in it. Participation is controlled by cultural roles, often expressed through intricate sets of status relationships.

As worldwide cultural sequences become better known, archeologists are able to abandon the modal concept of culture in favor of the more satisfactory, if more complex, systemic view of cultural phenomena.

Law of superposition. The law of superposition was formulated initially by Nicolaus Steno (1638–1687). Steno's law, simply stated, says that in any pile of sedimentary rocks that have not been disturbed by folding or overturning, the strata on the bottom were deposited first. The law of superposition thus holds that, all else being equal, older deposits will tend to be buried beneath younger ones. Thomas Jefferson is generally credited with being the first archeologist to have applied systematic principles of stratigraphy to archeology.

Mere stratigraphic equivalence, however, does not necessarily indicate contemporaneity, as there can be misleading mixtures of successive occupational debris on one surface. In sandy areas, successive occupations may have been covered by layers of blown sand. Where the covering layers are intact, no better sealing and separation of occupational surfaces can be found. If, however, wind activity is reversed, and sand is removed rather than deposited, two or more distinct layers of human occupation may become falsely associated. Archeologists must therefore study the processes of cultural deposition in order to recognize the difference between intact and disturbed strata.

Excavation. Contemporary excavation must be conducted with a plan, a firm research design that attempts to provide answers to definite questions. Archeology is one of the few sciences which destroys its own data in the process of generating them. Archeologists must therefore be extremely careful to make the appropriate observations at the time of excavation. If problems are not in mind, delicate archeological associations might be overlooked. Once a site is excavated, only the data from that excavation are available; the site itself has actually been destroyed.

Archeological technique attempts to record and clarify such relationships. The chronological relationship, for instance, between two structures lying side by side may be of interest; one structure can be established as later if it overlies the other. On a larger scale, one may wish to ascertain the duration, successive extents, economies, and purposes of occupation of a given site or area.

Contemporary archeologists are faced with the unprecedented destruction of sites because of deep plowing, quarrying, construction, and leveling for parking lots and airports. In Europe such excavation has been termed rescue archeology; in the New World the unwieldy term cultural resource management has been applied to the effort to preserve, or at least mitigate, the impact upon the survival of the archeological site.

Relation to other disciplines. The task of deciphering meaning from past material culture is so complex that the archeologist is often required to borrow from allied disciplines in the physical and natural sciences. Only a few can be mentioned.

Geology. Geological studies provide information on the natural barriers and obvious routes and the source of ores which people use. Geomorphology provides information regarding changes in sea level and the extent of glacial ice sheets and loess deposits. It serves as a basis for assigning relative dates to archeological discoveries, particularly of Paleolithic stone implements. *See* GLACIAL EPOCH; PLEISTOCENE; STRAND LINE; VARVE.

Climatology and paleobotany. The analysis of ancient plant pollen and spores—palynology—and information regarding former climates has recently become one of archeology's best methods for examining prehistoric ecological adaptations. Dates also have been determined by studying annual growth in trees. The identification of cereal grains in the surface of ancient pottery has helped in piecing together the story of the beginnings of agriculture. *See* DENDROCHRONOLOGY; PALYNOLOGY; POSTGLACIAL VEGETATION AND CLIMATE.

Paleontology. The study of fossil remains helps the archeologist to identify finds of bone, antler, and ivory and to reconstruct people's natural environment in successive periods. Paleontological studies also shed light upon the history of the domestication of animals and the development of humans' own bodily structure. *See* FOSSIL HUMANS; PALEONTOLOGY.

Land snails often have restricted habitats; a study of the shells of these creatures taken, for example, from different levels in the material which has accumulated at different times in a ditch made by humans may show that when the ditch was first excavated there were large trees growing in the immediate neighborhood. Later, pottery may have found its way into the half-filled ditch and, with it, the shells of snails typical of open grassland, therefore indicating that by the time the pottery was being made, the area had been cleared of trees.

Along with the pottery and grassland snail shells, there may be found the wing cases and other remains of beetles. The restricted habitat of beetles serves the same purpose as that of snails as far as environmental studies are concerned. There have been cases where the remains of a particular kind of dung beetle have shown that cattle were present at the time a certain deposit was made. In one case, the finding of the wings of certain ants showed that a mound had been built up at a particular time of the year, since these ants are known to fly only during their mating season.

Mineralogy. The mineralogist has been called upon to compare and identify the sources of metal ores and materials used in the manufacture of implements and to name the origin of stones. Without this information archeologists would not have known that the blue stones of Stonehenge were quarried in South Wales, nor would archeologists have found the sites in North Wales and Westmoreland where stone axes discovered in southern England were made.

Physics. The science of physics has provided the archeologist with a mean of dating finds in years, for it has made available the study of the rates of decay of radioactive substance. The study of paleomagnetism and the orientation of magnetic particles in strongly heated structures and pottery may eventually provide a means of dating archeological finds. By using the proton magnetometer, which measures minute differences in magnetic fields, the archeologist is often able to confirm the suggested existence of filled-in pits and ditches of which no surface evidence appears.

Other techniques include the use of air photographs to map known sites and to discover new ones; the use of special underground photographic apparatus, such as photographic drills, to penetrate into the earth and to record the interior of buried chambers; and the use of electrical instruments which record differences in electrical potential and thus detect irregularities below the surface. *See* AERIAL PHOTOGRAPH; RADIOCARBON DATING.

Chemistry. The study of chemistry provides knowledge of natural processes in the soil and helps in the identification of invisible evidences of occupation, such as decayed layers of buried turf. Certain chemical techniques also make possible the preservation and reconstruction of archeological finds for study and exhibition. *See* AMINO ACID DATING; ARCHEOLOGICAL CHEMISTRY.

Synthesis. The archeologist must have some understanding of all these sciences to extract from sites and materials every possible piece of information which may lead to a better understanding of prehistory. One requirement, however, is necessary for the archeologist to make any contribution to humanity's inheritance of knowledge: the archeologist must be able to record and publish every minor fact for the benefit of colleagues and successors, because the writing of prehistory requires the synthesis of all archeological discovery and interpretation. *See* ANTHROPOLOGY; PHYSICAL ANTHROPOLOGY. David Hurst Thomas

-s

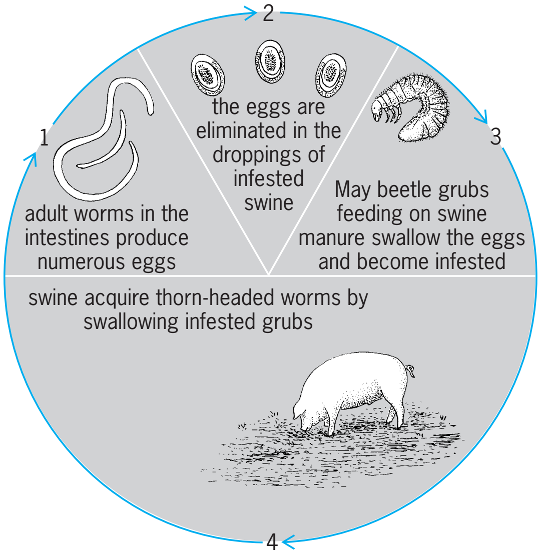
Archiacanthocephala

An order (or class, according to some classifications) of the phylum Acanthocephala. The adult worms are parasitic in terrestrial vertebrates. The body wall and lemnisci of species in this order have numerous amitotically fragmented nuclei or a few ameboid giant nuclei. The main trunks of the lacunar system are dorsal and ventral, or dorsal. Typically, there are eight separate cement glands in the male. Two ligament sacs are present in the female, one dorsal, the other ventral. They are persistent and united with the openings of the uterine bell. The eggs are elliptical and have a thick shell. The proboscis receptacle has a conspicuous ventral cleft or is a closed sac with two concentric muscle layers. The proboscis hooks occur in long rows or spiral rows, and the trunk lacks spines. The cystacanth occurs in grubs, roaches, and grasshoppers. Some common archiacanthocephalans are *Oncicola canis*, *Moniliformis moniliformis*, and *Macracanthorhynchus hirudinaceus*.

Oncicola canis is a short plump acanthocephalan, primarily parasitic in dogs and other Canidae. It occurs also in cats. The body of the adult, 0.2–0.5 in. (6–14 mm) long, is short and heavy with irregular cross furrows. The globular proboscis has six spiral rows of six hooks each. Male organs occupy more than one-half the length of the body. The arthropod intermediate host is unknown. Cystacanths have been found in armadillos and in the esophageal walls of turkeys, indicating a transport host in the life cycle.

Moniliformis moniliformis is an elongate acanthocephalan which is parasitic in house rats. The females are 4–12 in. (10–30 cm) long, whereas the males measure 2–5 in. (6–13 cm). The body of both sexes exhibits conspicuous pseudosegmentation except on the extremities. The proboscis is cylindrical with 12–15 rows of 10 or 11 hooks each. Male organs are confined to the posterior half of the body. Eight pyriform cement glands are present and the testes are long and elliptical. The egg is elliptical with a clear, thick outer shell. Cockroaches (*Periplaneta americana*) serve as the intermediate host; however, in Europe a beetle, *Blaps mucronata*, is the intermediate host. The cystacanth develops broad lateral flanges. Occasionally infections have been found in humans.

Macracanthorhynchus hirudinaceus is the giant thorn-headed worm of hogs and is probably the best known of all acanthocephalans because of its cosmopolitan distribution (see **illus.**) Females measure 10–24 in. (25–60 cm) in length and the males 2–4 in. (5–10 cm). The worms are pinkish with a transversely wrinkled body which tapers from a rather broad anterior end to a slender posterior end. The proboscis is globular with six spiral rows of six hooks each. The testes are elongate and located in the middle of the body or anterior to it. The eight cement glands are elliptical in shape, arranged in four consecutive pairs in the posterior portion of the body. Eggs are elliptical in shape with a heavy, dark-brown outer shell which is irregularly grooved and ridged, giving it a sculptured appearance. At least 25 species of scarabaeid beetle larvae have been reported as intermediate hosts. The cystacanth is cylindrical in shape. In addition to their occurrence in the domestic pig, adults have been reported to occur in squirrels, chipmunks, moles, and occasionally humans. This acanthocephalan is of considerable economic importance to the hog-raising industry.


2
1
3
4
adult worms in the intestines produce numerous eggs
the eggs are eliminated in the droppings of infested swine
May beetle grubs feeding on swine manure swallow the eggs and become infested
swine acquire thorn-headed worms by swallowing infested grubs

Life cycle of the giant thorn-headed worm of hogs.

Macracanthorhynchus ingens is an acanthocephalan found in raccoons (*Procyon lotor lotor*) which is similar in appearance to *M. hirudinaceus*. Female worms are 7–12 in. (18–30 cm) in length, and males 5–6 in. (13–15 cm). Mature eggs are elliptical in shape, with a heavy, dark-brown outer shell which is irregularly sculptured. Larvae of scarabaeid beetles serve as intermediate hosts. Two species of *Phyllophaga* and one species of *Ligyrus* have been successfully infected. Experimental work has shown that frogs may serve as transport hosts. *Macracanthorhynchus ingens* has been reported from a child who had a history of eating "bugs." Nine worms were removed from the child after the administration of an anthelminthic. *See* ACANTHOCEPHALA; COLEOPTERA; ORTHOPTERA.

Donald V. Moore

Archidiidae

A subclass of the plant class Bryopsida (mosses). The Archidiidae consists of a single genus, *Archidium*, with 26 species, occurring in ephemeral habitats, especially in wet, grassy places.

The small gametophytes consist of erect, simple or forked stems, and elongate, singly costate leaves in numerous rows. The capsules, usually terminal, are globose, immersed, and irregularly rupturing. The capsule wall consists of a single layer of cells; stomata, peristome, and columella are lacking. The spores are few, developing 1–44 spore mother cells scattered in the outermost layer of the endothecium, and are large (50–310 micrometers) and thick-walled.

The calyptra is scarcely differentiated. The haploid chromosome numbers are 13 and 26.

The subclass is unique in the scattered origin of spores in a single layer of the endothecium and in the absence of a quadrant stage in the early ontogeny of the capsule. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA; BUXBAUMIIDAE; DAWSONIIDAE; POLYTRICHIDAE; TETRAPHIDIDAE.

Howard Crum

Bibliography. J. A. Snider, A revision of the genus *Archidium* (Musci), *J. Hattori Bot. Lab.*, 39:105–201, 1975; J. A. Snider, Sporophyte development in the genus *Archidium* (Musci), *J. Hattori Bot. Lab.*, 39:85–104, 1975.

Archigregarinida

An order of the protozoan subclass Gregarinia, class Telosporea, subphylum Sporozoa. All gregarines are parasites of the digestive tract and body cavity of invertebrates or lower chordates; their large, mature trophozoites (vegetative stages) live outside the host's cells. The Archigregarinida are primitive gregarines, and live in marine worms (annelids and lower chordates—enteropneustids, sipunculids, and ascidians). Their life cycle includes sexual and asexual phases and involves three periods of schizogony (multiple fission).

There are only 28 named species in 5 genera. The most important genus is *Selenidium,* which has 24 species. Its members occur in the intestine of marine polychaete annelids, sipunculids, enteropneustids, and ascidians.

The hosts are infected by eating oocysts containing sporozoites. The sporozoites emerge, enter host intestinal epithelial cells, become ellipsoidal, grow into large schizonts, and produce many merozoites or schizozoites. These enter the intestinal lumen, become attached to intestinal cells by their anterior end, and grow into gamonts. The gamonts become free in the lumen of the intestine, and two of them (sporadins) become attached to each other in an association known as syzygy. They secrete a cyst wall around themselves, and the resultant gametocysts or gamontocysts are passed into the seawater with the feces. Within the gamontocyst, one of the gamonts produces a large number of female gametes by multiple fission, while the other gamont produces an equal number of male gametes; the two types of gamete are about the same size, but the male gametes have a flagellum. Fertilization takes place, and each zygote forms an oocyst (often called sporocyst or spore) within the gamontocyst. Then by a third multiple fission, eight sporozoites are formed in each oocyst. *See* GREGARINIA; PROTOZOA; SPOROZOA; TELOSPOREA.

Norman D. Levine

Archimedes' principle

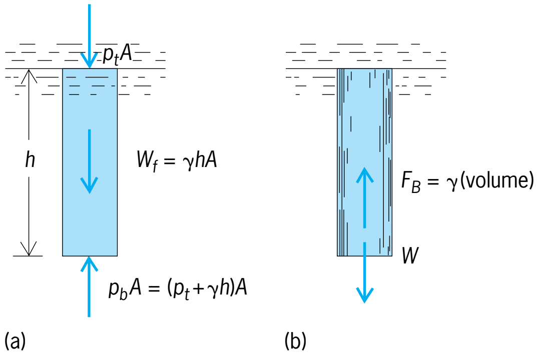


Fig. 1. Demonstration of Archimedes' principle. (*a*) Forces on a water column; p_t = pressure at top of column, p_b = pressure at bottom of column, h = height of column, A = area of column, W_f = weight of fluid column, γ = specific weight of fluid. (*b*) When the water column is replaced with a solid object, hydrostatic forces on this object are the same as on the original water column; F_B = buoyant force, W = weight of object.

The principle that there is an upward fluid force on a body submerged (or floating) in a stationary fluid that is equal to the weight of the fluid displaced by the body. This concept, perhaps the oldest stated principle in fluid mechanics, was first put forth by Archimedes in the third century B.C.

In a static fluid, the weight of the fluid causes an increase in pressure with depth. Thus, at the surface of the fluid, the pressure is atmospheric pressure (p_0 = 14.7 lb/in.^{2} = 101 kilonewtons/m^2), while at a depth h the pressure has a larger value of p_1, given by Eq. (1), where γ is the specific weight of the fluid

$$p_1 = p_0 + \gamma h \qquad (1)$$

(weight/volume). The difference in pressure force between the bottom and the top of a water column (**Fig. 1***a*) is therefore given by Eq. (2), where h and

$$(p_b - p_t)A = \gamma hA \qquad (2)$$

A are the height and area of the column, and p_b and p_t are the pressures at the bottom and top of the column. When the water is in equilibrium, this difference is precisely equal to the weight W_f of the water within the column, given by Eq. (3). If the

$$W_f = \gamma(\text{volume}) = \gamma hA \qquad (3)$$

water column were replaced with a solid object (Fig. 1*b*), the pressure forces on the object would be the same as on the original water column. That is, the net hydrostatic force on the object, termed the buoyant force F_B, would be equal to the weight of the water displaced (which is the statement of Archimedes' principle). The same concept holds for a body of arbitrary shape, which can be thought of a consisting of many small vertical columns fastened together. Archimedes' principle is valid for submerged or floating bodies in liquids or gases. *See* BUOYANCY; SPECIFIC GRAVITY.

The fact that a helium-filled balloon rises through air is explained by any of the following equivalent statements: (1) the balloon's weight is less than its buoyant force; (2) the weight of the displaced air is greater than the balloon's weight; or (3) the average air pressure on the lower surface of the balloon is

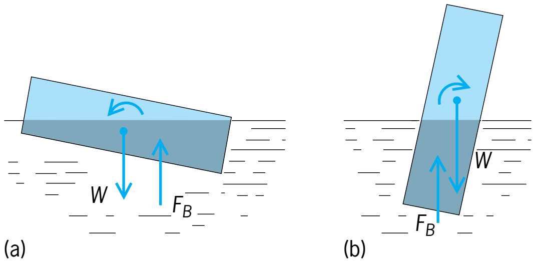


Fig. 2. Stability of a floating body depends on locations of the buoyant force F_B, which acts through the centroid of the displaced fluid volume, and the weight W, which acts through the center of gravity. Curved arrows indicate moments that result from combination of F_B and W. (*a*) Body in a stable position. (*b*) Body in an unstable position.

sufficiently greater than that on its upper surface.

A neutrally buoyant submarine displaces a volume of water precisely equal to its weight. The average specific weight of the submarine (the steel hull, power plant, occupants, enclosed air, and so forth) equals the specific weight of seawater; the buoyant force equals the total weight. In contrast, if a scale indicates a person's weight as 150 lb (667 newtons), the actual weight is approximately 150 + 0.18 = 150.18 lb (667.2 + 0.8 = 668.0 N) because of the approximate 0.18-lb (0.8-N) buoyant force of the air (specific weight of air times the person's volume) pushing upward on the person. However, since a person is much more dense than air, the buoyant force in this situation is usually negligible.

The location of the buoyant force (which acts through the centroid of the displaced fluid volume) and the location of the center of gravity of the body determine the stability of submerged or floating bodies. Specifically, a body's stability depends on what happens when the body is given a slight angular displacement (as when a ship rolls because of wave motion). If the moment (torque) caused by the buoyant force and the equal-but-opposite weight tends to rotate the body back to its horizontal equilibrium position, the body is stable (**Fig. 2***a*). However, the body is unstable if the moment causes the body to rotate further from its equilibrium position (Fig. 2*b*) *See* HYDROSTATICS.

Bruce R. Munson

Bibliography. E. N. Gilbert, How things float, *Amer. Math. Month.*, 98(3):201–216, 1991; B. R. Munson, D. F. Young, and T. H. Okiishi, *Fundamentals of Fluid Mechanics*, 3d ed., 1997.

Architectural acoustics

The science of sound as it pertains to buildings. There are three major branches of architectural acoustics. (1) Room acoustics involves the design of the interior of buildings to project properly diffused sound at appropriate levels and with appropriate esthetic qualities for music and adequate intelligibility for speech. Room acoustics is an essential component of the design of theaters, concert halls, lecture rooms, classrooms, and churches, among other building types. (2) Noise control or noise management involves the reduction and control of noise between a potentially disturbing sound source and a listener. The walls, floors, ceilings, windows, and doors in buildings reduce sound energy as it travels through them. Disturbing noise sources may be equipment and people within the building or intruding noise from external sources of sound such as amplified music, airplanes, or highways. Noise from building services, including heating, ventilating, and air-conditioning systems, lighting systems, and electrical power systems, must also be controlled so that it does not disturb people using buildings. (3) The design of sound reinforcement and enhancement systems uses electronic equipment to improve the quality of sounds heard in rooms.

Properties of building materials. When a sound wave moving through the air strikes a building surface, three things can happen to it. A portion of the sound energy is absorbed within the wall, and converted into another form of energy such as heat; a portion of the energy is reflected back into the room; and a portion is transmitted through the wall to the other side (**Fig. 1**).

Sound absorption. The sound absorption coefficient α is the portion of energy that is absorbed compared to the total energy that strikes the surface. At a given frequency, the absorption coefficient is a function of angle of incidence. In architectural acoustics it is convenient to use an average absorption coefficient which represents an average absorption over all angles of incidence. The average absorption coefficient is assumed to depend only on the physical characteristics of the material and not on the sound field (**Table 1**). The absorption of a surface can be expressed in sabins (or metric sabins), units indicating the absorption of 1 ft^2 (or 1 m^2) of a totally absorbing surface, that is, a surface having an absorption coefficient of unity (**Table 2**). *See* SOUND ABSORPTION.

Sound is absorbed by any mechanism that converts incident sound waves into other forms of energy and ultimately to heat. All materials used in buildings absorb some sound. Generally, materials that are heavy and massive such as brick and concrete will not absorb much sound energy, and instead reflect sound back into the room. Materials that are light and porous will absorb significant amounts of

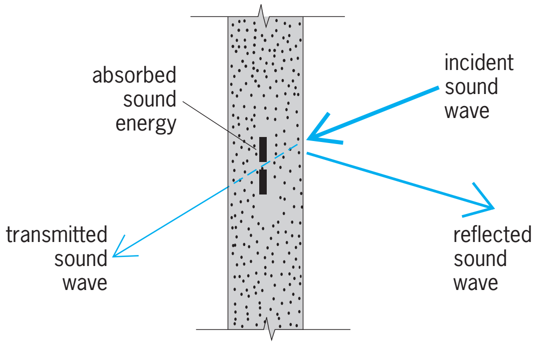


Fig. 1. Sound wave striking a building surface. Absorption of a portion of the sound energy in the wall, reflection of a portion back into the room, and transmission of a portion through the wall are shown. (*Siebein Associates*)

TABLE 1. Sound absorption coefficients of general building materials*

Material	Values of α					
	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
Brick, unglazed	.03	.03	.03	.04	.05	.07
Carpet, heavy						
On concrete	.02	.06	.14	.37	.60	.65
On 40-oz/yd^2 (1.36 kg/m^2) hair-felt or foam rubber	.08	.24	.15	.69	.71	.73
Concrete block						
Coarse	.36	.44	.31	.29	.39	.25
Painted	.10	.05	.06	.07	.09	.08
Fabrics						
Light velour, 10 oz/yd^2 (0.34 kg/m^2), hung straight, in contact with wall	.03	.04	.11	.17	.24	.35
Heavy velour, 18 oz/yd^2 (0.61 kg/m^2), draped to half area	.14	.35	.55	.72	.70	.65
Floorings						
Concrete or terrazzo	.01	.01	.015	.02	.02	.02
Linoleum, asphalt, rubber, or cork tile on concrete	.02	.03	.03	.03	.03	.02
Wood	.15	.11	.10	.07	.06	.07
Glass						
Large panes of heavy plate glass	.18	.06	.04	.03	.02	.02
Ordinary window glass	.35	.25	.18	.12	.07	.04
Gypsum board 1/2 in. (12.7 mm) nailed to 2 × 4's 16 in. (406 mm) on center	.29	.10	.05	.04	.07	.09
Marble or glazed tile	.01	.01	.01	.01	.02	.02
Plaster, gypsum or lime						
Smooth finish on tile or brick	.01	.01	.02	.03	.04	.05
Rough finish on lath	.14	.10	.06	.05	.04	.03
Plywood paneling, 3/8 in. (9.5 mm) thick	.28	.22	.17	.09	.10	.11

*Data from Acoustical Materials Association.

sound energy depending upon their porosity, thickness, density, and mounting. The sound waves are converted into heat energy by propagating through the interstices of the material and by vibration of the small fibers of the material. Another important mechanism of absorption occurs when sound waves force a panel into motion; the resulting flexural vibration converts a fraction of the incident sound energy into heat. Materials especially designed to have relatively high absorption coefficients are called acoustical materials, acoustical wall panels, and acoustical ceiling tiles.

The noise reduction coefficient (NRC) is an average of the absorption coefficients of a material in the 250-, 500-, 1000-, and 2000-Hz octave bands. The noise reduction coefficient is often used to make initial comparisons of the effectiveness of absorbent materials used in offices, atriums, corridors, and other noncritical spaces.

Sound reflection. When a sound wave is larger than the dimension of the surface it strikes, it will reflect from the surface at an angle of reflection that is equal to the angle of incidence. Diffusion or scattering occurs when sound waves strike a surface with dimensions that are the same order of magnitude as their wavelengths. In that case, the scattering of sound waves from the irregular surface randomizes the directions of the reflected sound waves. Splayed wall panels, convex curves, coffers, and specially designed diffusing panels effectively diffuse or scatter sounds (**Fig. 2**).

Sound transmission. The sound transmission coefficient τ is the portion of sound energy that is transmitted through the wall compared to the total energy that strikes the surface. The sound transmission loss in decibels is equal to 10 log $(1/\tau)$. It measures

TABLE 2. Approximate sound absorption of occupied and unoccupied auditorium chairs

Type	Absorption, sabins per chair*					
	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
Upholstered chair, occupied, depending on type of chair and spacing	3.2–4.4	3.5–5.4	4.2–6.5	4.8–7.0	5.0–7.0	4.8–7.0
Upholstered chair, unoccupied, depending on type of chair and spacing	1.8 – 3.5	2.6–5.0	3.6–5.0	3.7–5.3	3.7–6.1	4.3–5.3
Wood pews, occupied, depending on spacing	3.1	3.4	4.1	4.7	4.8	4.7
Metal or wood chair, unoccupied	0.15	0.2	0.25	0.40	0.45	

*1 sabin = 0.093 metric sabin.

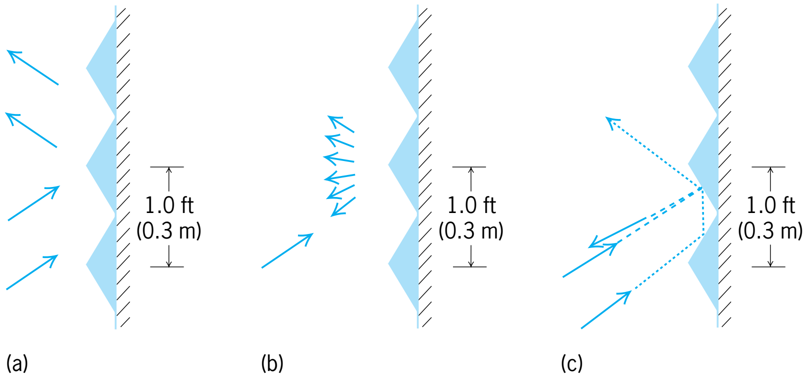


Fig. 2. Reflection of sound from a plane surface having splays with centers 1 ft (0.3 m) apart. Wavelength of sound is (*a*) much larger than the splays (frequency 100 Hz), (*b*) similar in size to that of the splays (frequency 1000 Hz), and (*c*) much smaller than the splays (frequency 100,000 Hz). (*After C. M. Harris, ed., Handbook of Acoustical Measurements and Noise Control, 3d ed., McGraw-Hill, 1991*)

how much sound is lost or how much the sound is reduced as it is transmitted through a construction assembly (an assemblage of materials that are used to construct a wall, roof, ceiling, or other surface in a building).

Frequency dependence. Sound reflection, absorption, and transmission all vary with the frequency or wavelength of sound. Therefore, it is essential to know the absorption and transmission coefficients over the entire frequency range of audible sounds. These quantities are usually measured in laboratory tests in 1/3 octave bands from 100 Hz or lower to 4000 Hz and higher.

Reverberation. Reverberation is the persistence of sound in a room after the source has stopped due to multiple reflections from the walls and ceiling of the room. Around 1900, Wallace Clement Sabine conducted pioneering research in reverberation and sound quality in rooms. He found that the reverberation time, RT, the amount of time it takes for a sound to decay to the point of inaudibility in a room after the source has stopped, was related to the volume of the room in cubic meters, V, and the absorbing power, $\Sigma_i S_i \alpha_i$, of the interior finish materials in the room by Eq. (1). Here, the absorbing power is the

$$\mathrm{RT} = \frac{0.161V}{\Sigma_i S_i \alpha_i} \tag{1}$$

sum, over the room surfaces, labeled i, of the surface area in square meters, S_i, multiplied by the respective absorption coefficient, α_i. The corresponding formula in U.S. Customary units is Eq. (2), where V'

$$\mathrm{RT} = \frac{0.049V'}{\Sigma_i S'_i \alpha_i} \tag{2}$$

and S'_i are in cubic feet and square feet respectively. *See* REVERBERATION.

Room acoustics. One essential component of room acoustics is an understanding of psychoacoustics and the qualitative evaluation of sounds heard by people in rooms. Psychoacoustics is the study of the psychology of sounds. It includes studies conducted in laboratories and in actual listening rooms of how people react to the level, frequency content, direction, and arrival time of sounds. These studies have established a set of relationships among the acoustical qualities that have been found to be important in the perception of sound, the room surfaces that contribute to these qualities, and the physical components of the sound field in a room that contribute to these properties (**Table 3**). *See* PSYCHOACOUSTICS.

TABLE 3. Summary of acoustical qualities in rooms and the architectural features of rooms that contribute to them

Acoustic quality	Architectural feature that creates acoustic quality	Physical acoustic event that creates acoustic quality
Envelopment and source width	Narrow rooms up to 70–80 ft (21–24 m) across Multiple tiers of narrow balconies	Early sound reflections arriving at the listener from the side (up to 80 ms after the direct sound)
Clarity	Sound reflecting ceiling	Sound reflections (usually from overhead) that arrive shortly after the direct sound
Reverberance	Large room volume Sound-reflecting materials	Prolonging of sound in the room after the source has stopped due to multiple reflections from the room surfaces
Loudness	Limited size of room Provision of multiple sound reflections to seats	Sound reflections from the ceiling and walls arriving shortly after the direct sound
Intimacy	Reflecting surfaces close to listeners	Arrival of the first sound reflection from a building surface shortly after the direct sound
Warmth	Heavy massive building materials	Strength of reflected low-frequency sound
Brilliance	Reflecting surfaces for high-frequency sounds	Strength of high-frequency sound
Spaciousness	Surface texture and sound-diffusing materials	Late sound energy arriving from the sides (after 80–100 ms)
Localization of sound	Clear sight and sound line between listener and source	Strength of direct sound relative to subsequent reflections
Ensemble	Overhead and side-wall sound reflecting surfaces near the performing area	Sound reflections that allow the musicians across the stage to hear each other

These relationships are based upon the concept of the impulse response of the room. The use of this concept has been an essential component in increasing the understanding of room acoustics and improving the acoustical design of rooms. The impulse response consists of the response of a room to an impulsive sound such as a single note of music or a single syllable of speech. It has four components. The direct sound is the sound that travels directly through the air by the shortest path from the source to the listener. Early reflected sounds are sounds that reflect off room surfaces and reach the ears of listeners within 50–80 milliseconds after the direct sound. These sounds will increase the apparent loudness and clarity of sounds when they come from overhead surfaces (the ceiling); increase the apparent width of the sound source and contribute to a sense of acoustic spaciousness when they come from the sides; and create a sense of acoustic intimacy if they arrive very shortly after the direct sound (less than 20 to 40 ms) regardless of direction. If the sounds contain low-frequency or bass energy, a sense of warmth will be achieved. If they contain higher-frequency or treble sounds, a sense of brilliance will be achieved. The reverberant tail is the sound energy that has reflected off multiple surfaces before it reaches the ears of listeners, gradually decreasing in amplitude or loudness and arriving at progressively longer time periods after the direct sound. It contributes to a sense of reverberance or fullness of musical sounds if it lasts as long as 2.0 to 2.3 s in a concert hall. If the reverberant sound energy arrives at the ears of listeners from the sides and the sounds that arrive at the left and right ears of listeners are configured somewhat differently, a sense of spaciousness will be enhanced. The fourth component of the impulse response is the background noise. This is the level of ambient noise in a room from building services, people moving about, equipment in the room, and intruding sounds from outside the room (**Fig. 3**).

Sound absorbent materials are strategically located in rooms to absorb unwanted sounds. They are used on the ceilings and upper wall surfaces of atriums, gymnasiums, cafeterias, and large open offices to reduce reverberant sounds that reflect off these surfaces, creating excessive noise. They are also used on the rear walls of lecture rooms and auditoriums to reduce the level of potential echoes from these surfaces. An echo is a sound reflection that arrives at a listener's ear 70 ms or more after the direct sound and is loud enough to be heard as a distinct acoustic event or a second image of the original sound. Absorbent materials are also used on the surfaces of large concave shapes such as domes. A concave shape will focus much of the reflected sound energy at its geometric center, creating an acoustic hot spot or a location where sounds are heard as much louder than at other locations. Absorbent materials are also used to vary the reflected and reverberant energy levels in rooms that are used for several acoustic functions such as music and theater. Sound absorbent banners or curtains may be lowered in a concert hall to cover wall surfaces when the room is used for theatrical performances and less reverberation is desired. *See* ECHO.

Several important design concepts are used to provide good listening conditions in rooms for speech and music. First is to provide good access to the direct sound for all people in the room. This usually involves raising the source of sound on an elevated stage, altar, or podium at the front of the room and sloping the floor surface to elevate the ears of people above the heads of those seated in front of them. The width and depth of the room should also be limited so that the natural direct sound can project from the speaker or instruments at the front of the room to the listeners. Second is to limit the background noise level in the room so that people can hear the sound they want to hear above the level of the ambient sound. Third is to limit the reverberation time in

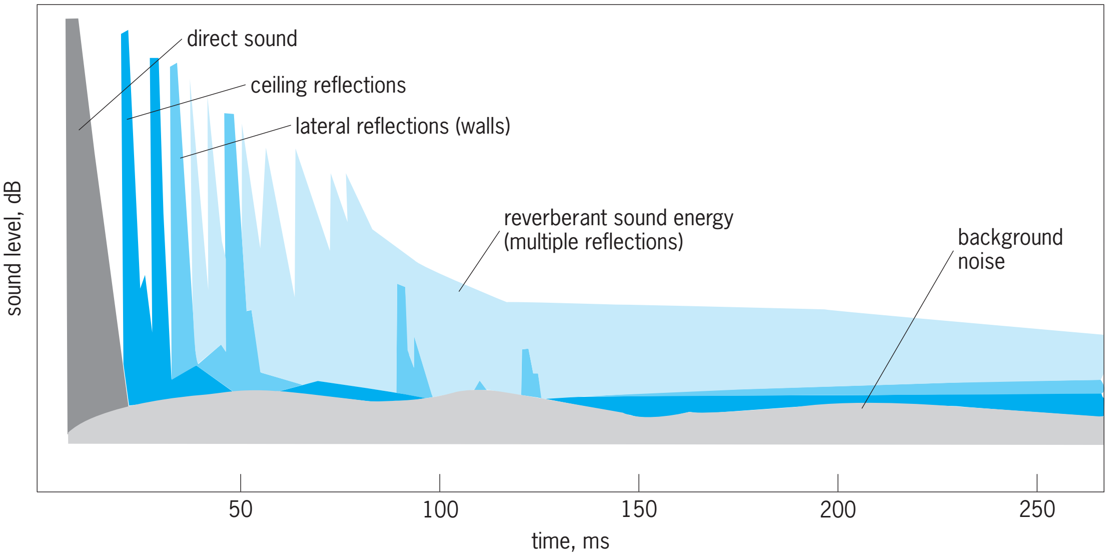


Fig. 3. Impulse response, showing four basic parts: the direct sound, early sound reflections from the ceiling and walls, the reverberant tail, and the background or ambient noise in the room. (*Siebein Associates*)

TABLE 4. Appropriate reverberation times for various rooms

Room	Reverberation time at middle frequencies (500–1000 Hz), s
Class room	0.3–0.6
Lecture hall	0.6–0.9
Music rehearsal room	0.6–1.1
Small theater	0.8–1.1
Multipurpose auditorium	1.2–2.0
Church	1.0–2.0*
Concert hall	1.8–2.3

*Depending upon size and nature of music program.

the room so that sounds are heard clearly and fully, while providing enough reverberant sound energy that sounds are heard as "full" and "live." If there is too much reverberation in a room, the persistence of an initial syllable will cover up or mask the one that follows it, making it difficult to understand what is being said (**Table 4**).

The ceilings and walls of rooms are designed to direct early reflected sounds to listeners' ears. The wall and ceiling panels in the front of the room are shaped to direct reflected sounds over the entire audience seating area. The panels in the center of the room are designed to cast reflections over the rear half of the seating area. The panels at the rear of the room cast reflections over the last rows of seating. This approach to room acoustics design uses reflected sounds to increase the apparent level of sounds toward the rear of the room where the direct sound is weakest. This helps to create more uniform sound levels throughout the room. When these panels are free-form floating panels suspended below the upper ceiling of a room, they are often called acoustic clouds (**Fig. 4**).

The Eugene McDermott Concert Hall in Dallas, TX, provides examples of several acoustic design concepts (Fig. 4). The canopy over the orchestra platform provides strong early reflections to the audience for clarity, intimacy, and envelopment. The room is surrounded by large reverberation chambers with doors that can open and close. This allows sound to enter the chambers, reflect off the walls, and reenter the room as reverberant energy, providing a rich, full sound for orchestral concerts and organ. The room has a series of narrow balconies that surround the main seating area, creating an intimate space. The undersides and faces of the balconies provide surfaces to cast sound reflections from the sides of the room to the audience for a spacious and enveloping sound. The room is built of heavy concrete with wood laminated to its surface. These materials reflect all frequencies of sound, creating a sense of wamth (reflection of low-frequency sounds) and brilliance (reflection of high-frequency sounds).

Noise control. Acoustical planning concepts for buildings include placing noisy activities away from activities that require relative quiet and locating noise-sensitive activities away from major sources of noise. Buffer spaces such as corridors or storage spaces are often used to separate two rooms that require acoustical privacy such as music rehearsal

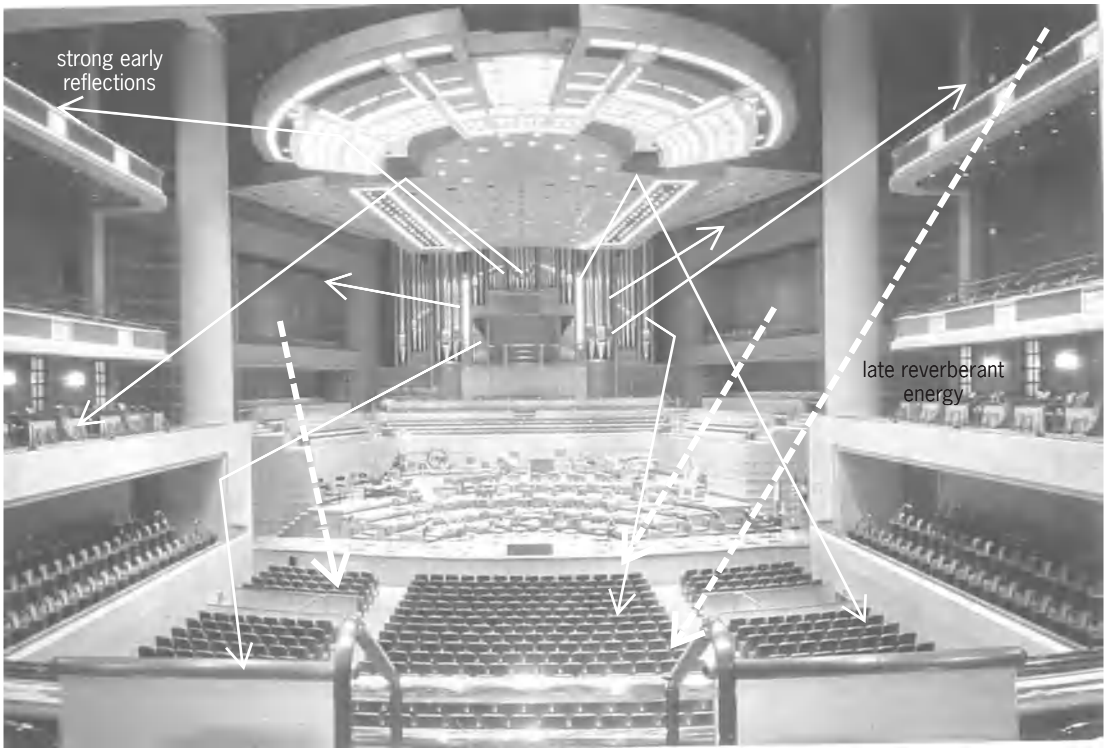


Fig. 4. Interior of the Eugene McDermott Concert Hall in the Morton H. Meyerson Symphony Center in Dallas, TX. Russell Johnson of Artec Consultants, Inc., acoustical consultants, and I.M. Pei of Pei, Cobb, Freed and Partners, architects, collaborated on the design of the room. Paths of strong early reflections and late reverberant energy are shown.

TABLE 5. STC ratings for various construction assemblies and the magnitude of perceived sounds heard by people listening on the other side of the assembly*

STC rating	Typical construction assembly	Typical application	Perception of intruding sounds on opposite side
35	0.5-in. (1.25-cm) gypsum board on both sides of metal studs	Office walls	Normal speech heard somewhat
45	0.5-in. (1.25-cm) gypsum board on both sides of metal studs with glass fiber insulation in stud space	Conference room walls	Normal speech heard faintly if at all Loud speech heard faintly
50	8-in. (20-cm) concrete masonry unit sealed on both sides	Classroom walls; mechanical rooms with air-handling units	Loud speech barely heard Music reduced in volume Low-frequency mechanical noise reduced
50	Two layers of 0.5-in. (1.25-cm) gypsum board on one side of metal studs, one layer of 0.5-in. (1.25-cm) gypsum board on the other side with glass fiber insulation in stud space	Classroom walls	Loud speech barely heard Music reduced in volume
55	Two layers of 0.5-in. (1.25-cm) gypsum board on both sides of metal studs with glass fiber insulation in stud space	Music practice room walls in schools	Loud speech generally not heard Music sounds, except loud bass, faintly heard

*Used with permission; © G. W. Siebein, *Tutorial on Architectural Acoustics*, 1992.

rooms in a school. Intruding noises from the exterior or from adjoining rooms can be reduced by using walls, ceilings, windows, and doors with appropriate transmission losses. The performance of a construction assembly used to separate rooms is often compared using a sound transmission class (STC) rating. This is a weighted average of the transmission loss values for the assembly in the 125- to 4000-Hz octave bands (**Table 5**).

Airborne sound is transmitted through a partition by forcing the partition into vibration. The vibrating partition becomes a secondary source of sound and radiates sound to the room on the other side. For a homogeneous wall, the transmission loss is related to its mass. The transmission loss increases by 3–6 dB per octave band over many frequencies. The transmission loss value of a partition is usually poor at frequencies corresponding to strongly excited natural frequencies of the partition and at certain critical frequencies where the phases of incident sound waves coincide with the phases of vibration of the wall. The transmission loss also increases by 4–6 dB in a given frequency band when the mass of the wall is doubled. A compound or double wall assembly can be used to reach a relatively high transmission loss with low mass per unit wall area. The separation between the two leaves or surfaces of the wall must be maintained as completely as possible for this to occur.

It is essential to control noise from building services. The location of air-conditioning plants on a site should be chosen so as to reduce propagation of noise to neighbors. Mechanical rooms in buildings that house air handling units, pumps, and other equipment should be located away from noise-sensitive rooms. Noise control treatments in the air-conditioning system include providing vibration isolators for equipment; providing flexible connections between ducts, conduits, and pipes to equipment; designing air ducts to operate with air velocities that will not create turbulent flow noise; and installing silencers or attenuators in the ducts to reduce noise produced by fans from traveling through the duct work. *See* MECHANICAL VIBRATION; VIBRATION ISOLATION.

Noise buildup within a room can be reduced by applying absorbent materials to the interior surfaces. The amount of noise that is reduced by adding absorbent material is estimated by the quantity 10 log (a_2/a_1). Here, a_2 is the amount of absorption in the room after absorbent materials have been added to it, and a_1 is the original amount of absorbtion in the room. The amount of absorption in a room is equal to the sum of the surface areas of each of the room surfaces multiplied by their respective sound absorption coefficients. *See* ACOUSTIC NOISE.

Sound reinforcement. Sound reinforcement systems, electronic enhancement systems, and sound amplification systems are used in many buildings. A sound reinforcement system amplifies the natural acoustic sounds in a room that is too large for people to hear with just “natural” room acoustics. This type of system reinforces the natural sounds that come from the room, increasing their apparent loudness with a series of loudspeakers.

In an electronic enhancement system, loudspeakers act as virtual room surfaces to create the perception that sounds are reflected from these surfaces at the proper times and with the proper loudness. These systems usually have a network of loudspeakers located throughout a room and connected to a microprocessor. The microprocessor can delay the signals to arrive at times corresponding to reflected sounds from the virtual room surfaces. It can also add reverberation and other special acoustic effects to create a virtual acoustic space. A sound amplification system makes all sounds played in a space louder. It is usually not designed to supplement the

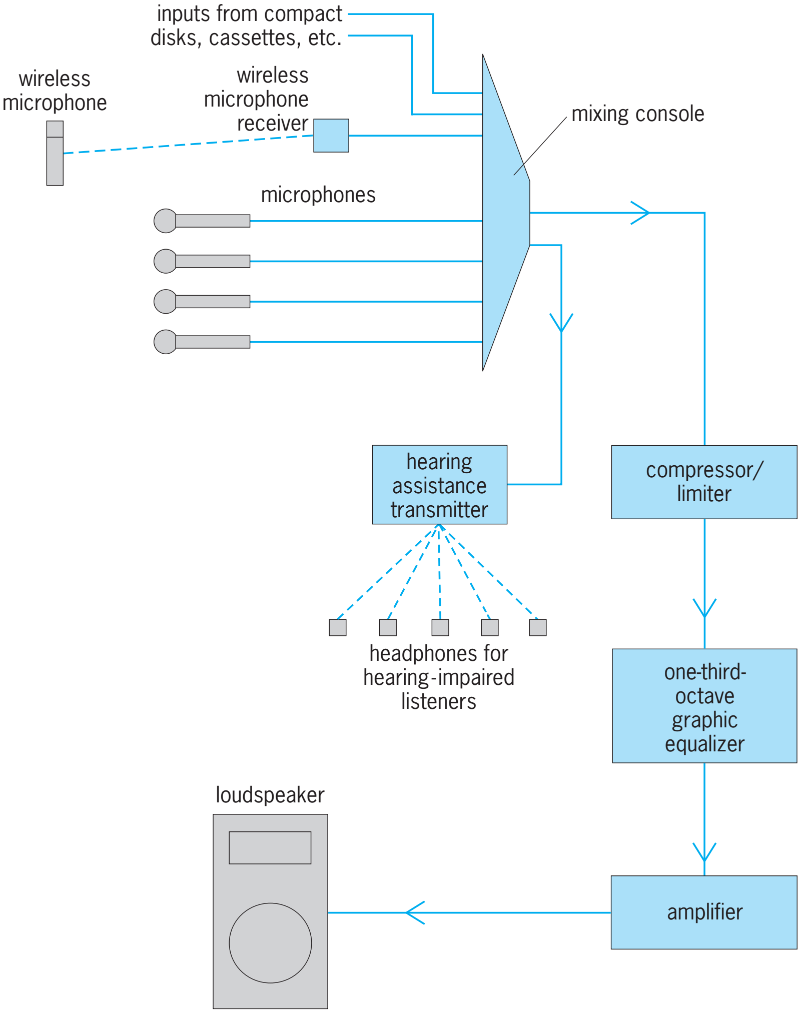

inputs from compact disks, cassettes, etc.
wireless microphone
wireless microphone receiver
mixing console
microphones
hearing assistance transmitter
compressor/ limiter
headphones for hearing-impaired listeners
one-third-octave graphic equalizer
loudspeaker
amplifier

Fig. 5. Simple sound-reinforcement system.

natural room acoustics or to provide subtle virtual room effects to the amplified sounds (**Fig. 5**).

Several basic types of loudspeaker arrangements are used in rooms. When the ceiling height of a room is 20 ft (6 m) or more, as in a church or theater, a central cluster system is often used. This consists of a group of loudspeakers located high above the sound source, which projects sound down toward the audience seating. In a room with lower ceilings, such as a convention center or hotel ball room, a distributed loudspeaker system is used. It consists of a network of loudspeakers mounted in the ceiling and spaced at even increments.

In a long, horizontal room, the sound from the loudspeaker will reach the ears of listeners before the direct sound from the person speaking because the sound waves travel in air at 1128 ft/s (344 m/s) and the sound moving through the wires as an electronic signal travels many times faster. A digital delay is added to the sound arriving from the loudspeaker so that it will arrive after the direct sound from the person speaking, maintaining its correct location in the impulse response as a reinforcing "electronic reflection." *See* SOUND; SOUND-REINFORCEMENT SYSTEM.

Gary W. Siebein

Bibliography. L. L. Beranek, *Concert and Opera Halls: How They Sound*, 1998; W. C. Cavanaugh and J. Wilkes, *Architectural Acoustics: Principles and Practice*, 1998; M. D. Egan, *Architectural Acoustics*, 1989; C. M. Harris (ed.), *Handbook of Acoustical Measurements and Noise Control*, 3d ed., 1991; M. Mehta, J. Johnson, and J. Rocafort, *Architectural Acoustics: Principles and Design*, 1999; C. M. Salter and Associates, Inc., *Acoustics: Architecture, Engineering, the Environment*, 1998.

Architectural engineering

A discipline that deals with the technological aspects of buildings, including the properties and behavior of building materials and components, foundation design, structural analysis and design, environmental system analysis and design, construction management, and building operation. Environmental systems, which may account for 45–70% of a building's cost, include heating, ventilating and air conditioning, illumination, building power systems, plumbing and piping, storm drainage, building communications, acoustics, vertical and horizontal transportation, fire protection, alternate energy sources, heat recovery, and energy conservation. In addition, to help protect the public from unnecessary risk, architectural engineers must be familiar with the various building codes, plumbing, electrical and mechanical codes, and the Life Safety Code. The latter code is similar to a building code and is designed to require planning and construction techniques in buildings which will minimize possible hazards to the occupants. *See* ELECTRICAL CODES; FIRE TECHNOLOGY.

Architectural engineering differs from other engineering disciplines in two important aspects. Most engineers work with other engineers, while most architectural engineers work or consult with architects. Furthermore, an architectural engineer not only must be fully qualified in engineering, but must also be thoroughly versed in all architectural considerations involved in design and construction. An architectural engineer designing a structural or environmental system is expected to be familiar not only with that system, but also with the multitude of architectural considerations which may affect its design, installation, and operation.

Although retaining more or less conventional forms, buildings have become increasingly complex during the last century. The need for a professional engineer to solve inherent technological problems was first recognized by the University of Illinois, which established an architectural engineering curriculum in the 1890s. There are now a number of such programs throughout the United States. All of these programs meet the same requirements as other engineering disciplines in mathematics, basic science, engineering science, humanities, and engineering design. All are accredited by the Accreditation Board for Engineering and Technology (ABET). However, there is considerable diversity in the architectural engineering curricula offered by the various institutions. Some programs are of a traditional 4-year duration. These programs meet the engineering requirements at the expense of architectural content. Other programs are 5 years in length. In general, they contain more architecturally related instruction as well as greater breadth in the environmental technologies.

A majority of the graduates of architectural engineering curricula become registered engineers. However, a sizable number elect to continue their education and become registered as architects as well. Career opportunities include partnerships with architects, consultants to architects, facilities engineers, and construction managers. Opportunities are also available as technical representatives for equipment manufacturers. *See* BUILDINGS; ENGINEERING.

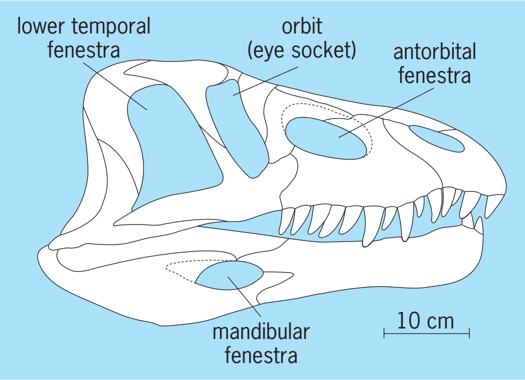
Thomas S. Dean

Archosauria

A diversified group of diapsid reptiles, comprising dinosaurs (and their descendants, birds), crocodylians, pterosaurs (flying reptiles), and various forms closely related to these groups. In the past, the primitive archosaurs from the Triassic Period were collectively referred to as Thecodontia, but this is an artificial grouping that is no longer accepted by specialists. *See* CROCODYLIA; ORNITHISCHIA; PTEROSAURIA; SAURISCHIA.

Archosaurian reptiles are readily distinguished by the presence of one or more openings (antorbital fenestrae) on each side of the skull between the nostril and eye socket, and an opening in the lower jaw (mandibular fenestra) [see **illustration**]. Among extant vertebrates, the two surviving groups of archosaurs, crocodylians and birds, differ from other reptiles and from mammals in many anatomical and behavioral characteristics. Both have a muscular foregut, or gizzard, containing stones for processing food, a heart with completely divided ventricles, and a pulmonary artery with three semilunar valves. Both crocodylians and birds use calls to mark territories and attract potential mates.

Early archosaurs employed quadrupedal locomotion, but the hindlimbs were already longer than the forelimbs. Later forms were typically bipedal, but many groups independently reverted to quadrupedal stance and gait. Most archosaurs were carnivorous, but herbivory evolved independently in a number of lineages, mainly among dinosaurs.



Skull of *Erythrosuchus africanus*, a basal archosaur from the early Middle Triassic of South Africa.

The earliest archosaurian fossils are known from the Late Permian. Archosaurian reptiles underwent a major diversification during the Triassic Period, when they became the dominant terrestrial vertebrates. Birds (with about 9000 known species) are the most diverse group of amniote vertebrates on land today, but crocodylians are represented by only about two dozen species.

The most primitive archosaurs, mainly from the Early Triassic and often grouped together as the Proterosuchia, already had the diagnostic openings in the skull but retained a sprawling limb posture with the elbows and knees sticking out sideways. Among more advanced archosaurs, the limbs moved partially or completely beneath the body, perhaps permitting more rapid locomotion. Paleontologists now distinguish two major evolutionary lineages among these advanced forms, the Crurotarsi and the Ornithodira. The Crurotarsi comprise crocodylians (Crocodylia) and various closely related forms (Crocodylomorpha), the long-snouted, superficially crocodilelike phytosaurs, the heavily armored aetosaurs, the facultatively bipedal ornithosuchids, and various large predators from the Middle and Late Triassic, commonly grouped together as Rauisuchia. The Ornithodira include dinosaurs (including birds), pterosaurs, and a few still poorly known forms probably related to the ancestry of these two groups.

Crurotarsans and ornithodirans are easily distinguished from each other by the structure of their ankle joints. In the former, the two principal ankle bones, the astragalus and the calcaneum, rotate against each other by means of a peg-and-socket joint (crurotarsal condition). In the latter, a simple hinge joint is developed between the astragalus and calcaneum (which are firmly attached to the bones of the lower leg) and the remainder of the ankle and foot (mesotarsal condition). Most ornithodirans, especially dinosaurs, show substantial modifications of the pelvic girdle and hindlimbs, which indicate erect gait. The posture of the hindlimbs in pterosaurs has been the subject of much debate. Recent discoveries of trackways suggest that they moved on all fours when walking on the ground, rather like bats. *See* AVES; DINOSAUR; REPTILIA.

Hans-Dieter Sues

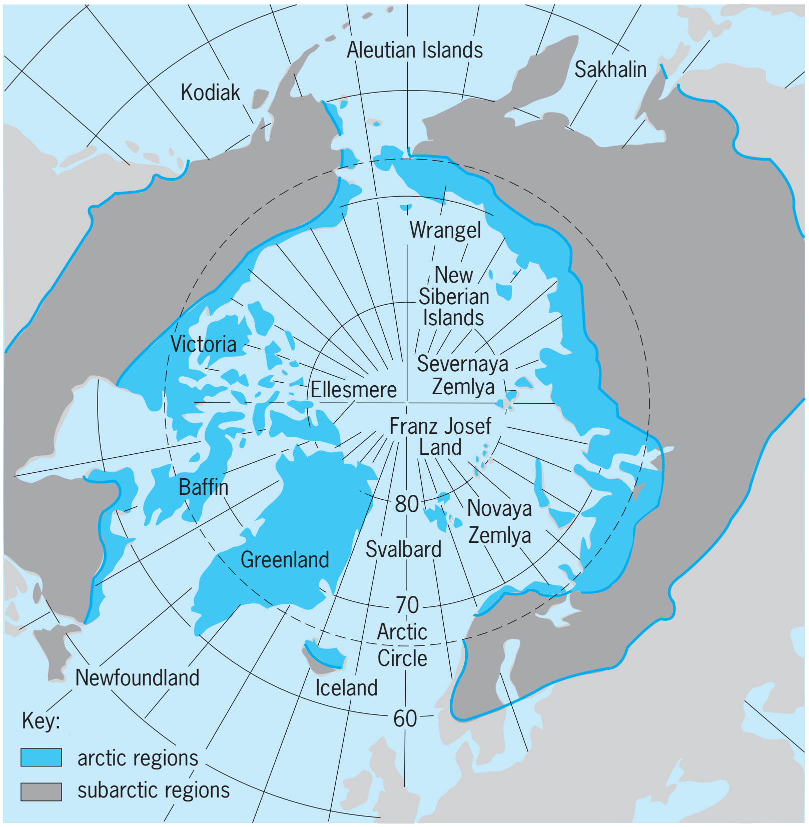
Bibliography. P. J. Currie and K. Padian (eds.), *Encyclopedia of Dinosaurs*, San Diego, 1997; P. C. Sereno, Basal archosaurs: Phylogenetic relationships and functional implications, *Mem. Soc. Vert. Paleontol.*, 2:1–53, 1991; P. C. Sereno, The evolution of dinosaurs, *Science*, 284:2137–2147, 1999; D. B. Weishampel, P. Dodson, and H. Osmólska (eds.), *The Dinosauria*, 2d ed., University of California Press, Berkeley, 2004; P. Wellnhofer, *The Illustrated Encyclopedia of Pterosaurs*, Crescent Books, New York, 1991.

Arctic and subarctic islands

Defined primarily by climatic rather than latitudinal criteria, arctic islands are those in the Northern Hemisphere where the mean temperature of the warmest month does not exceed 50°F (10°C) and that of the coldest is not above 32°F (0°C). Subarctic islands are those in the Northern Hemisphere where the mean temperature of the warmest month is over 50°F (10°C) for less than 4 months and that of the coldest is less than 32°F (0°C).

Such islands generally are in high latitudes. Distribution of land and sea masses, ocean currents, and atmospheric circulation greatly modify the effect of latitude so that it is often misleading to use location relative to the Arctic Circle as a significant criterion of arctic or subarctic. The largest proportion by area of the islands lies in the Western Hemisphere, primarily in Greenland and in the Canadian Arctic Archipelago.

Diversity of land surfaces. Physiographically, the islands include all the varied major landforms found elsewhere in the world, from rugged mountains over 8000 ft (2500 m) high, through plateaus and hills, to level plains only recently emerged from the sea. All have been glaciated except Sakhalin and some of the islands in the Bering Sea sector. R. F. Flint reported that of the 5,800,000 mi^2 (15,000,000 km^2), or 10%, of land area of the world still ice-covered, over 5,000,000 mi^2 (13,000,000 km^2) is in Antarctica, and over 700,000 mi^2 (1,812,000 km^2) lies in the arctic islands with over 600,000 mi^2 (1,554,000 km^2) of this in Greenland. Removal of the weight of ice sheets and the resultant crustal rebound has exposed prominent marine beaches and wave-cut cliffs on many of the islands. These now commonly occur at elevations of over 300 ft (90 m) above sea level. *See* ANTARCTICA.



Islands of the arctic and the subarctic regions.

Climate. The general climatic pattern of these islands is set by their location relative to the two semipermanent centers of low pressure over the Aleutian Islands and over Iceland (the Aleutian Low and the Icelandic Low). Especially during the winters these low-pressure centers affect the areas from Kamchatka to southeastern Alaska and from Newfoundland to Novaya Zemlya, respectively. The intervening areas, including the islands of the northwestern Canadian Arctic Archipelago and those of the central and eastern Russian Arctic, are much less subject to cyclonic storms; rather they are dominated by stable, dry air masses often linked across the polar basin from either continent. During the summer this pattern decreases in intensity. Most of the precipitation is cyclonic in origin. Because they are marine areas, the islands receive more precipitation than they otherwise would, yet even so this is very light for most of the arctic islands removed from the zone of cyclonic activity. Also, because they are marine areas, the islands, regions of low temperatures by definition, are not regions of extreme low temperatures. Much lower temperatures are reported from continental land areas farther south than from the most northern arctic islands. In general, the larger the island and the closer its proximity to a continental landmass, the higher are the summer temperatures and the lower its winter temperature. *See* POLAR METEOROLOGY.

Vegetation and soils. The climatic differences between arctic and subarctic islands are reflected in their natural vegetation. The arctic islands are treeless. Natural vegetation consists of the tundra—mosses, sedges, lichens, grasses, and creeping shrubs. Luxuriance and continuity of ground cover vary with such factors as moisture, insolation, and soil nutrient conditions. Bare ground is often exposed, and in places plant growth may be lacking completely except for a few rock-encrusting lichens. In such places the ground surface may consist of frost-shattered rock fragments, tidal mud flats, boulder-strewn fell fields, or snow patches and ice. Permafrost (permanently frozen ground) occurs throughout the Arctic (and in parts of the subarctic) and is reflected in impeded drainage and patterned ground. *See* PERMAFROST.

The natural vegetation of subarctic islands characteristically is the boreal forest or taiga, composed predominantly of conifers such as spruce, fir, pine, and larch with deciduous trees such as birch, aspen, and willow; the latter are especially common in regrowth of clearings in the forest. Impeded drainage because of permafrost or glaciation gives rise to numerous ponds and muskeg areas. A transitional type of vegetation, the forest-tundra, is recognized

Size and elevation of larger arctic and subarctic islands*

	Area		Highest point (elevation)		
Name	mi^2	km^2	Name	ft	m
Aleutian Islands					
Unimak Island	15,500	40,100	Shishaldin Volcano	9,978	3,040
Unalaska Island	10,800	28,000	Makushin Volcano	6,680	2,030
St. Lawrence Island	18,200	47,100	Kookoligit Mountains	2,207	670
Nunivak Island	16,000	41,400	Roberts Mountain	1,675	510
Kodiak Island	37,400	96,900	Grayback Mountain	3,317	1,000
Canadian Arctic Archipelago	500,000	1,295,000			
Baffin Island	196,000	507,000	Penny Ice Cap	6,750	2,060
Ellesmere Island	76,000	197,000	Barbeau Peak	8,600	2,600
Victoria	84,000	217,000	Shaler Mountains	2,100	655
Banks	27,000	70,000	Durham Heights	2,400	732
Devon	21,000	55,000	Treuter Mountains	6,190	1,920
Axel Heiberg	17,000	43,000	Name not available	8,400	2,560
Melville	16,000	42,000	Raglan Range	3,500	1,070
Southampton	16,000	42,000	Porsild Mountains	1,750	530
Prince of Wales	13,000	33,000	Name not available	500	150
Newfoundland	42,734	109,000	Gros Morn	2,644	806
Greenland	840,000	2,176,000	Mountain Forel	11,286	3,500
Iceland	39,961	102,000	Oraefajökull	6,955	2,120
Svalbard (archipelago)	24,100	62,000			
Vest-Spitsbergen	15,250	39,000	Newtontopp	5,630	1,720
Franz Josef Land (archipelago)	7,000	18,000	Name not available	3,000	910
Novaya Zemlya (archipelago)	36,000	93,000			
Severny Island	21,000	54,000	Name not available	3,510	1,070
Yughny Island	15,000	39,000			
Severnaya Zemlya (archipelago)	14,000	36,000	Name not available	1,500	450
New Siberian Islands	12,000	31,000	Name not available	1,033	315
Wrangel Island	2,000	5,000	Peak Berry	3,510	1,070
Sakhalin Island	27,000	70,000	Nevelskoi Mountain	6,600	2,010
Kurile Islands	6,000	16,000	Alaid Volcano	7,674	2,030

*Approximate only in some cases because of incomplete mapping.

on some subarctic islands in sectors where smaller trees are widely spaced and abundant mosses cover the ground.

The typical soils of the subarctic islands are podzols—the surface soil grayish-white beneath the raw humus layer and highly acidic in nature. The tundra soils of the arctic islands really consist only of a dark-brown peaty surface layer over poorly defined thin horizons, and much of the ground cannot properly be termed soil. *See* SOIL.

Character of major islands. Within this general description, individual islands vary considerably (see **illus.** and **table**) A brief summary of some of the larger islands and archipelagos in the Western Hemisphere is given in the following sections.

Aleutian Islands. Extending southwest for more than 1000 mi (1600 km), from the Alaska Peninsula at 163°W to Attu Island at 175°W, the Aleutians separate the Bering Sea to the north from the Pacific Ocean to the south. Their rugged, mountainous surface consists of the drowned continuation of the Alaska and Aleutian ranges. Several active volcanoes are included among them. Most of the islands were glaciated by local mountain glaciers, a few of which persist in reduced size on some of the eastern islands (Unimak, Unalaska, and Umnak). There are about 150 islands and innumerable reefs. They are grouped from east to west as the Fox, Andreanof, Rat, and Near islands. Unimak, the easternmost island, is the largest, about 65 mi (105 km) long and 25 mi (40 km) wide.

The Aleutian Islands experience extremely variable weather because of the frequent cyclonic storms, complicated by the mountainous terrain. An associated gusty wind is known as the williwaw. Swift sea currents run among the islands and there is much fog. At Dutch Harbor (Unalaska Island) the mean temperatures are January, 32°F (0°C), and July, 51°F (11°C); while at Atka they are 33°F (1°C) and 50°F (10°C). Mean annual precipitation at the two stations is 56.7 and 70.2 in. (1440 and 1783 mm), respectively. The islands are treeless but generally support a luxuriant growth of grass, willow, and alder.

Canadian Arctic Archipelago. All those islands lying north of the continental mainland and west of Greenland to 141°W are included in this group. In rough outline the archipelago resembles a triangle, from a rather irregular base at about 61°N in the east and 67°N in the west, to its apex at the northernmost tip of Ellesmere Island in latitude 83°30′N. Those lying north of 74°N are known as the Queen Elizabeth Islands.

The easternmost islands of the archipelago are mountainous, whereas those to the west and northwest are plateaulike or plains. The mountains in Baffin Island, eastern Devon Island, and southeastern

Ellesmere Island are composed of Precambrian rock and average 5000–7000 ft (1520–2130 m) elevation. Continuing the mountain line northward through Ellesmere Island and Axel Heiberg Island, summit elevation exceeds 8000 ft (2500 m), partly reflecting a different fold axis, the Innuitian. In contrast the more southerly of the Queen Elizabeth Islands are more typically plateaulike, even though they contain more of the Innuitian fold structure; here the maximum elevation is about 3000 ft (920 m) in Melville Island but is commonly less than 1000 ft (305 m). The northwestern part of the Queen Elizabeth Islands is a recently emerged coastal plain, in which salt domes occur at places. Most of the islands in the western part of the southern archipelago are low and comparatively flat.

Most of the islands were glaciated, although there still is not sufficient information to establish conditions in the northwestern islands. Relic ice caps remain upon Ellesmere, Axel Heiberg, Devon, Bylot, and Baffin islands, because of the higher elevations and greater precipitation.

The Archipelago has long, cold winters. For 3–4 months mean monthly temperatures range from −20 to −30°F (−30 to −34°C) over most of the area. February is usually the coldest month; temperatures of −35 to −40°F (−37 to −40°C) are commonly recorded then. Rarely do temperatures drop below −50°F (−46°C), although a minimum of −63°F (−53°C) has been recorded at three stations. The southeastern and eastern parts of the archipelago have milder winters, because of the proximity to open water in Davis Strait and the passage of cyclonic storms. During the cool summer, temperatures are more uniform over the archipelago, with the July mean only from 40 to 50°F (4 to 10°C). Maximum temperatures exceed 60°F (16°C) at most but not all stations.

Precipitation is light. Most of the archipelago receives less than 10 in. (254 mm) a year, with only three stations (in the southeast) recording more than 15 in. (381 mm). One station (Eureka) reported less than 2 in. (50 mm) a year over a 2-year period. Snow may fall in any month, but over most of the archipelago rain accounts for about half the precipitation. Snowfall averages from 12.5 to 100 in. (320 to 2540 mm) in the southeast. Natural vegetation is of the tundra type.

Sea ice closes archipelago waters to shipping for 9–10 months in the year. Breakup usually occurs between late June and mid-July with freezeup between mid-September and late October, depending on location. The waters around the northwestern islands are rarely clear of ice. Icebergs are confined to the eastern parts, chiefly in the waters separating the archipelago from Greenland.

Newfoundland. Generally a plateau of rolling surface, this subarctic island is tilted west to east from the Long Range Mountains (over 2000 ft or 610 m) to the Avalon Peninsula (700 ft or 213 m). The northeast-southwest grain of the island, the result of folding and faulting, is reflected in the coastal configuration and physiography. The island was completely glaciated and most of it shows the results of intense ice scour. The indented seacoasts are commonly cliffed and rugged, with marked fiords in the high sectors of the Northern Peninsula. The only significant lowland area is a coastal plain in the west.

From the standpoint of climate and natural vegetation the island is predominantly subarctic, yet it includes aspects of the arctic and of the more continental regions. The moderating marine influence on climate is lessened by the island's east (that is, lee) location relative to the continental landmass, and by the cold Labrador Current, which flows south along the east coast and swings around to affect the south and parts of the west coasts.

Winters are cold, particularly in the interior (January mean, 15 to 20°F or −9 to −7°C), but are somewhat milder on the coast, especially in the southeast (St. John's January mean, 24°F or −4°C). Conversely, the coasts remain cooler in the summer (about 55°F or 13°C in July), whereas the western lowlands exceed 60°F or 16°C in mean daily July temperatures. The entire island has abundant precipitation, well distributed through the year, with heaviest falls (over 55 in. or 1400 mm) in the south. Snowfall is abundant (over 100 in. or 2540 mm) everywhere except along the south coasts. Sea ice seals off all coasts at its maximum extent, except the south. Fog is fairly frequent both on the coast and inland. It is most common in summer and on the southeast coasts. The juxtaposition of the cold Labrador Current and the warm Gulf Stream in the offshore ocean is largely responsible.

Boreal forest covers less than half of the island, with the best stands in the western lowlands and the north central valleys. Poor drainage, resulting from recent glaciation, and elevation restrict its extent. Where the altitude exceeds 1200 ft (365 m), the forest gives way to barrens—extensive areas of tundra—in the west and southwest.

Greenland. The world's largest island, Greenland extends over 1600 mi (2575 km) from Cape Farewell (59°46′N) at the south to Cape Morris Jesup (83°39′N) at the north, the latter being the nearest land to the North Pole. Its greatest width (77°N) is just under 700 mi (1130 km). Five-sixths of the surface is buried beneath the largest remaining land ice sheet in the Northern Hemisphere, and numerous smaller ice caps and glaciers occur as separate bodies around its margin. In profile the ice sheet is similar to a flat shield rising gently to form three broad domes, the highest of which exceeds 10,000 ft (3050 m). A maximum thickness of 7000 ft (2130 m) of ice has been estimated, but more work has to be done on this as well as on the suggestion that the underlying surface may consist of several separate islands rather than just one. Occasional peaks (nunatak) project through the ice sheet near its margin. Tongues of ice from the main glacier descend to the sea at many points. Most of the icebergs in the North Atlantic originate from such glaciers in southwest Greenland.

The ice-free margins are widest in the southwest and on the northeast coasts, although access to the latter is impeded by the arctic pack ice (storis) in the southward-moving East Greenland Current. All the land has been intensely glaciated, except the northernmost (Peary Land), which probably received insufficient precipitation. The margins include alpine mountains, plateaus (particularly in areas of basalt bedrock in the central areas on both east and west coasts), and lowlands. The skaergaard, a swarm of low islands and reefs, is very prominent along the southwest coast. The shore is much indented by bays and fiords.

The full significance of the ice cap in the climatic pattern is still not known. Temperatures there usually range from 27 to −49°F (−3 to −45°C) through the year. One associated phenomenon is the outward movement of strong katabatic winds from the ice margins which often produce a foehn or chinook effect on the valleys through which they are channeled. Climatic conditions in the ice-free margins are extremely variable, and the local complex topography has a great influence; for example, the inner parts of the fiords usually are warmer in summer than the outer, and colder in winter. Mean winter temperatures range rather uniformly on the west coast, from a February mean of 18°F (−8°C) at Ivigtut in the southwest to −20°F (−29°C) at Smith Sound in the northwest. During the summer there is much less contrast, with July mean temperatures of 50°F (10°C) in the south and 35–40°F (2–4°C) in the north. Precipitation, mainly as snow, decreases rapidly from 46 in. (1170 mm) at Ivigtut in the south to less than 10 in. (254 mm) north of 69°N. Notable climatic change has occurred in Greenland and its adjacent seas within historical times. Fog is common through the summers, especially near the broken sea ice. The island becomes ice-locked in winter except for part of the southwest coast. Natural vegetation of the ice-free areas is essentially tundra. Plant growth reaches its maximum development in the inner parts of the fiords, particularly in southwest Greenland. Five varieties of arboreal growth occur in the latter area, and in the Julianehaab district some birches grow to the height of 20 ft (6 m). Copses up to 7 ft (2 m) high occur in favored areas as far north as Disko Bay, but the grasses, mosses, and stunted growth of the true tundra are more typical and commonly make up the only vegetation over much of the island. *See* ARCTIC OCEAN; ASIA; EUROPE; NORTH AMERICA.

William C. Wonders

Bibliography. V. D. Aleksandrova, *The Arctic and Antarctic*, 1980; T. Armstrong et al., *The Circumpolar North*, 1978; J. D. Ives and R. G. Barry (eds.), *Arctic and Alpine Environments*, 1974; W. C. Wonders (ed.), *The Arctic Circle*, 1976.

Arctic Circle

The parallel of latitude approximately 66°30′ north of the Equator, or 23.5° from the North Pole. Named for the northern constellation Bear, the Arctic Circle has the same angular distance from the Equator as the inclination of the Earth's axis from the plane of the ecliptic. Thus, when the Earth in its orbit is at the Northern Hemisphere summer solstice, June 21, and the North Pole is tilted 23.5° toward the Sun, the Sun's rays extend beyond the pole 23.5° to the Arctic Circle, giving that parallel 24 h of sunlight. On this same date the Sun's rays at noon will just reach the horizon at the Antarctic Circle, 66°30′ south. The highest altitude of the noon Sun at the Arctic Circle is on June 21, when it is 47° above the horizon.

Within the Arctic Circle the Sun remains above the horizon continuously for 24 h at the longest period. However, with twilight considered, it remains daylight or twilight continuously for about 5 months. Twilight can be considered to last until the Sun drops 18° below the horizon. *See* MATHEMATICAL GEOGRAPHY; SOLSTICE.

Van H. English

Bibliography. A. H. Strahler and A. Strahler, *Introducing Physical Geography*, 4th ed., 2005; H. Veregin (ed.), *Rand McNally Goode's World Atlas*, 21st ed., 2004.

Arctic Ocean

The Arctic Ocean is an enclosed ocean, connected to the Pacific Ocean through Bering Strait and Bering Sea, and to the Atlantic Ocean through Fram Strait and the Barents, Greenland, and Norwegian seas (**Fig. 1**). The Arctic Ocean is the central region including the marginal seas whose water characteristics are both arctic in nature and quite distinct from those of the Atlantic and Pacific oceans. It includes the area between Bering Strait on the Pacific side and Fram Strait on the Atlantic side together with the Canadian Archipelago and seas along the continental shelf, and generally includes the Barents Sea. Usually it does not include the Norwegian, Greenland, or Bering seas. *See* ATLANTIC OCEAN; BERING SEA; PACIFIC OCEAN.

Expeditions. At the end of the nineteenth century, F. Nansen noted that driftwood caught ice formed near Siberia and exited the Arctic Ocean into the North Atlantic Ocean. In 1893, in an attempt to learn about the oceanography of the Arctic Ocean and to reach the North Pole, Nansen allowed his vessel, *Fram*, to be frozen into the ice pack north of Siberia. *Fram* drifted across three major basins of the Arctic Ocean, though it failed to drift far enough north for Nansen to claim the North Pole. The expedition gave the first real insight into the nature of the Arctic Ocean, obtaining a vast amount of knowledge from an ocean region where virtually none had existed before. In intervening years, almost all oceanography was done from isolated ice camps. Not until 1987 did a surface research vessel, FS *Polarstern*, exceed the farthest north point in the *Fram* drift path, and then by only a few miles. On September 7, 1991, IB *Oden* and FS *Polarstern* became the first nonnuclear-powered ships to reach the North Pole. The expedition proved the feasibility of carrying out standard modern oceanographic measurements, common in other oceans, throughout most of the Arctic Basin. Since then, several expeditions have

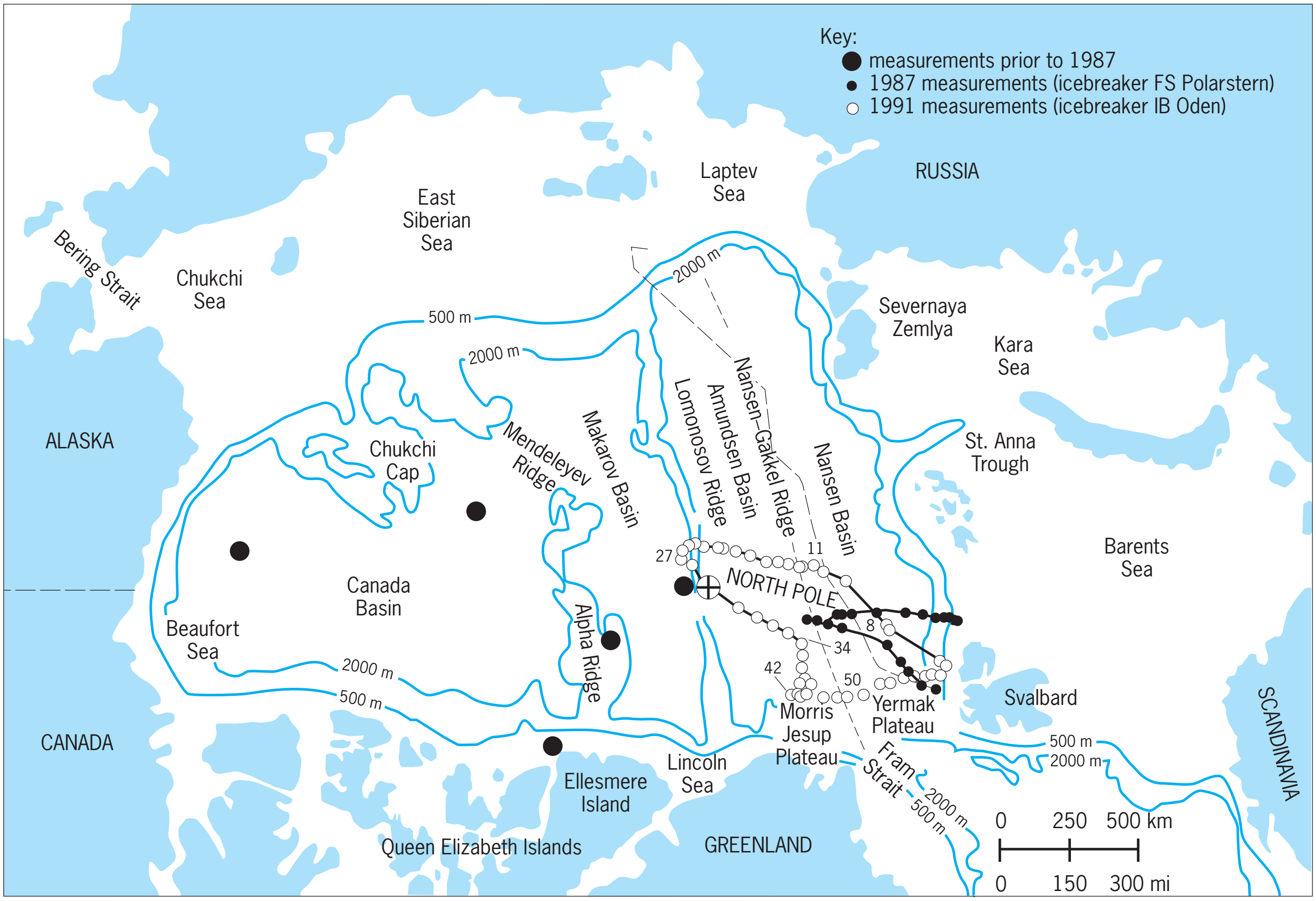

Key:
● measurements prior to 1987
• 1987 measurements (icebreaker FS Polarstern)
○ 1991 measurements (icebreaker IB Oden)
RUSSIA
Laptev Sea
East Siberian Sea
Chukchi Sea
Bering Strait
ALASKA
CANADA
Beaufort Sea
Canada Basin
Chukchi Cap
Mendeleyev Ridge
Makarov Basin
Lomonosov Ridge
Amundsen Basin
Nansen-Gakkel Ridge
Nansen Basin
Alpha Ridge
NORTH POLE
Severnaya Zemlya
Kara Sea
St. Anna Trough
Barents Sea
Svalbard
SCANDINAVIA
Yermak Plateau
Morris Jesup Plateau
Fram Strait
Lincoln Sea
Ellesmere Island
Queen Elizabeth Islands
GREENLAND
500 m
2000 m
27
11
8
34
42
50
0 250 500 km
0 150 300 mi

Fig. 1. Arctic Ocean basins.

traversed many of the key regions of the Arctic Ocean (**Fig. 2**).

General features. The Arctic Ocean, including its marginal seas and the Canadian Archipelago, has a total area of approximately 9.4 × 10^6 km^2 (3.6 × 10^6 mi^2), with the continental shelves, mostly on the Eurasian side, constituting about one-third of this area. The continental shelves, cut by several canyons, are typically 800 km (480 mi) wide on the Eurasian side and about 100 km (62 mi) wide on the North American side. The central Arctic Ocean Basin is separated by the subsurface Lomonosov Ridge, which extends from North America to the Eurasian continent almost through the North Pole, into the Eurasian Basin on the Atlantic side and the Canadian Basin on the Pacific side. Each of these two basins is subdivided into two basins by the less prominent subsurface Nansen-Gakkel and Alpha-Mendeleyev ridges, the Nansen and Amundsen basins constituting the Eurasian Basin, and the Makarov and Canada basins constituting the Canadian Basin. The deepest parts of the Arctic Ocean, which lie on either side of the Lomonosov Ridge, are more than 4000 m (13,000 ft). The depth at the North Pole is 4270 m (14,000 ft). *See* BASIN; CONTINENTAL MARGIN; MARINE GEOLOGY; SUBMARINE CANYON.

For many months of the year, daylight is weak or absent. Air temperatures over the Arctic Ocean range from a little above freezing in summer to an average of −35°C (−31°F) in January. Temperatures as low as −50°C (−58°F) have been recorded.

Biological productivity is low in most regions but is high in some parts of the marginal seas (Barents and Bering seas), where commercial fisheries exist. *See* MARINE FISHERIES.

Much of the Arctic Ocean is ice-covered throughout the year, typically with ice 2–3 m (7–10 ft) thick. In the central Arctic Ocean, ice is ever present, moving continually and mostly driven by average wind patterns (**Fig. 3**). As a result of the ice motion, large open cracks, or leads, appear and disappear as the open water freezes. As in regions where the ice is under pressure, large pressure ridges, up to 20 m (66 ft), can be formed. These pressure ridges, which can be a major barrier to icebreakers, may help mix the near-surface waters as the ice moves. Large parts of some of the marginal seas (Chukchi, Laptev, and Kara seas) and some of the Beaufort Sea become mostly ice-free for part of the summer. Much of the Eurasian coastline becomes navigable, if not completely ice-free, in the summer. It represents a major shipping route for communities in northern Russia.

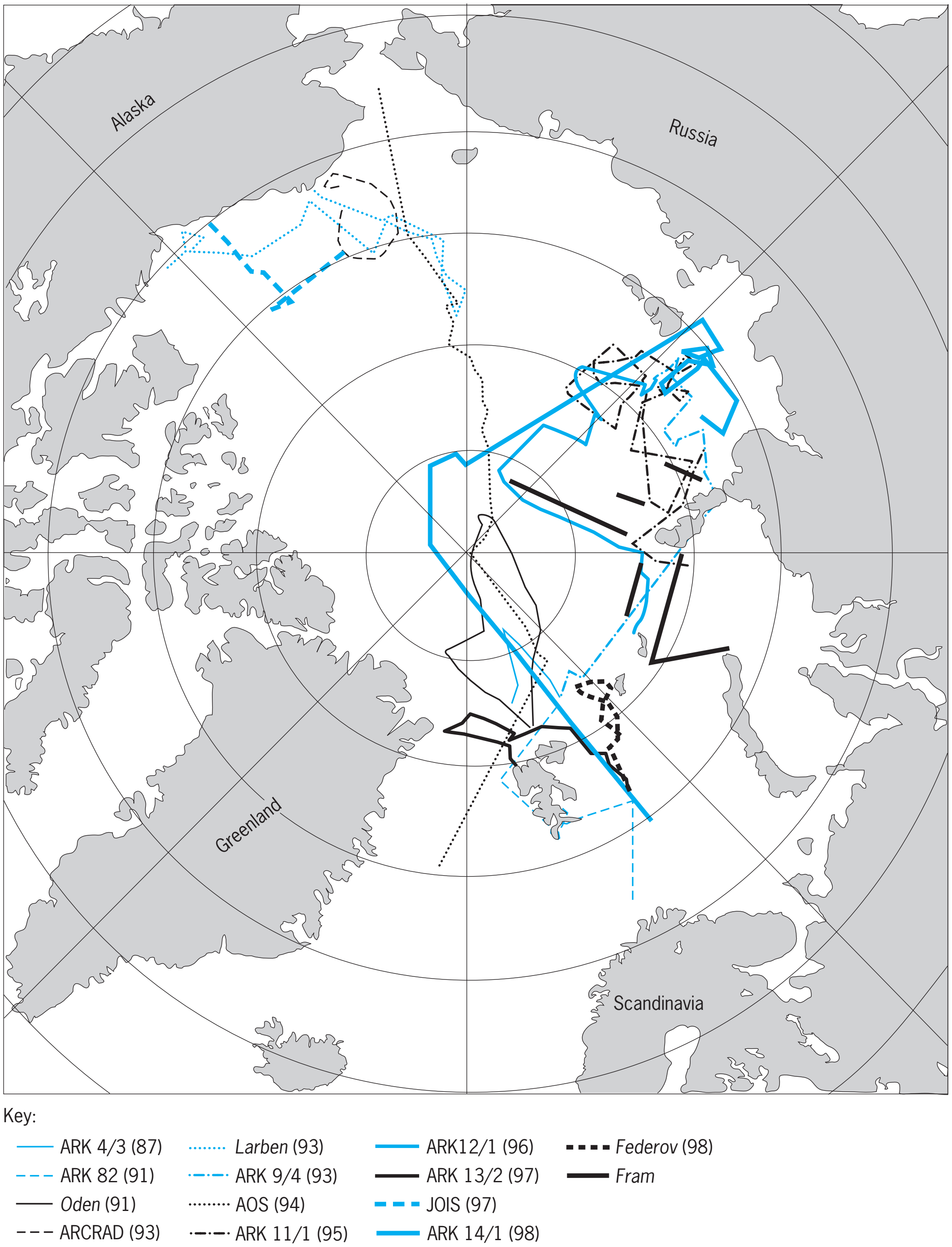

Alaska
Russia
Greenland
Scandinavia
Key:
ARK 4/3 (87)
ARK 82 (91)
Oden (91)
ARCRAD (93)
Larben (93)
ARK 9/4 (93)
AOS (94)
ARK 11/1 (95)
ARK12/1 (96)
ARK 13/2 (97)
JOIS (97)
ARK 14/1 (98)
Federov (98)
Fram

Fig. 2. Icebreaker expedition stations during 1987–1998 show that a substantial region of the Canadian Basin remains to be explored.

Seasonal production of ice over the large continental shelves strongly influences the structure and properties of the water column in the central regions.

Bering Strait is shallow (45 m; 150 ft), restricting the flow of water from the Pacific Ocean into the Arctic Ocean. Water from the Atlantic Ocean flows into the Arctic Ocean through Fram Strait and the Barents Sea. Deeper water exchanges through Fram Strait, 2500 m (8200 ft) deep, though communication with the main part of the North Atlantic is restricted to the upper 800 m (2600 ft) by the Greenland-Scotland Ridge farther to the south. Water flows out of the Arctic Ocean through Fram Strait and the Canadian Archipelago. There is a large fresh-water flow into the Arctic Ocean. About 90% of river input comes from Siberian rivers and about 10% from North

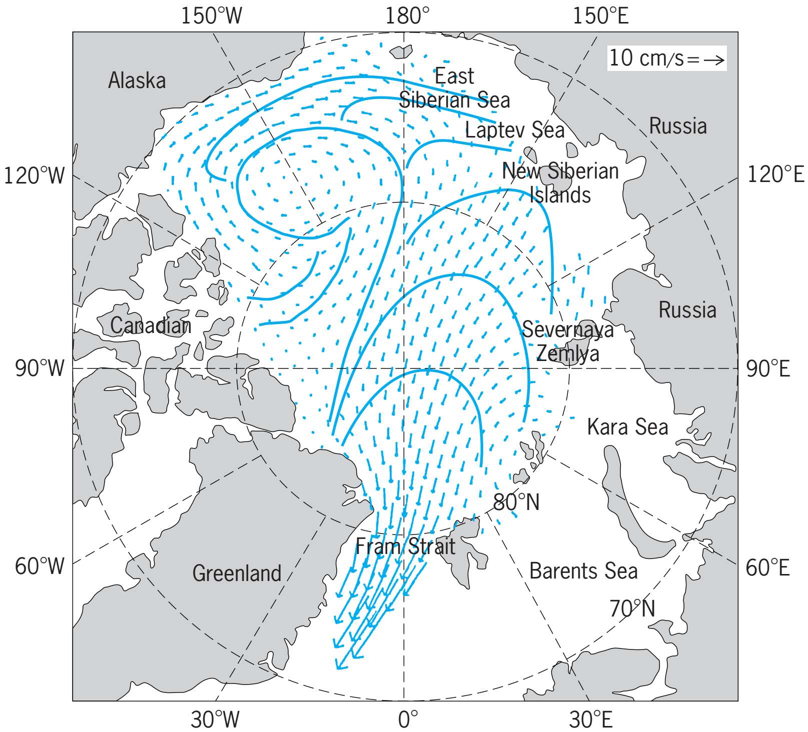


Fig. 3. Mean drift of ice in the Arctic Ocean. The data were obtained from ice station drifts and satellite tracking buoys placed on the ice. (*After S. Pfirman et al., Reconstructing the origin and trajectory of drifting Arctic sea ice, J. Geophys. Res., 102:12575–12586, 1997*)

American rivers, primarily the Mackenzie River. An unassessed amount comes from other runoff land drainage.

Water masses and circulation. Up until the early 1980s, conventional oceanography with research ships was not feasible because of severe conditions and extensive ice cover. While the former Soviet Union had an ambitious oceanographic program collecting data from ice stations and using long-range helicopters, only a limited amount of these data were available to western scientists. Icebreaker expeditions have added vastly more oceanographic data (Fig. 2), and the release of Russian historical ice station data is under way. These new data have given a much more complete picture of the Arctic Ocean and processes operating in it.

General properties. The Arctic Ocean has well-defined layers throughout most of its volume: the Polar Mixed Layer; the halocline, water with a high salinity and density gradient; the Atlantic Layer; intermediate-depth water sometimes called Upper Polar Deep Water; and deep water (see **table**).

Ice formation over the large continental shelves is responsible for many of the characteristics of the different water masses. When ice forms, brine is rejected from the ice and mixes with underlying water, producing dense (saline) water. This water flows off the shelves at many locations, forming plumes that sink through and entrain the surrounding water. The plumes influence the characteristics of all of the water masses below the Polar Mixed Layer. *See* SEA ICE; SEAWATER.

Water mass identification and circulation patterns. Salinity and temperature, as well as several natural chemical constituents (for example, nutrients, oxygen), are used to characterize water masses, to determine their sources, and to trace their circulation patterns. Anthropogenic materials (for example, chlorofluorocarbons, tritium) that were introduced into the ocean during the twentieth century are used to provide information on residence or exchange times of water masses for waters that are renewed in about 150 years or less. Naturally produced carbon-14 dates older water. *See* RADIOCARBON DATING.

The near-freezing Polar Mixed Layer lies over the entire central deep basin. Its salinity varies from relatively fresh values near Bering Strait to quite high values near Fram Strait. The salinity reflects the relative amounts of the constituent source waters (Atlantic, Pacific, river runoff, and sea ice meltwater). The mean circulation of Pacific origin waters in the Polar Mixed Layer as indicated from nutrient distributions enters through Bering Strait and exits through the Canadian Archipelago and Fram Strait (**Fig. 4**). Fresh water mainly from Siberian rivers enters this flow and exits with it. Atlantic origin water in the Polar Mixed Layer flows along the Eurasian coast toward the Canadian Basin, where it becomes overlain with fresher water from the Pacific Ocean and from rivers.

Characteristics of water masses in the Arctic Ocean*

Water mass	Depth range, m	Salinity†	Temperature	Residence time, years
Polar Mixed Layer	0–50	30–33	Near −1.8°C	5–10
Halocline				
Upper Halocline Water	50–130	Near 33.1	Near −1.6°C	5–15
Lower Halocline Water	20–200	Near 34.2	Near 1.2°C	5–15
Atlantic Layer	200–600	34.8–34.95	0–3°C	∼25
Upper Polar Deep Water	600–1700	Near 34.9	0 to −0.5°C	20–40
Deep Water				
Eurasian Basin	1700–4400	34.92	−0.94°C	50–150
Canadian Basin	1700–4000	34.94	−0.5°C	∼700

*Properties vary with depth and over the Arctic Basin. Values are typical of the respective water masses.
†Salinity values are measured in terms of the relative conductivity of seawater compared to a standard.

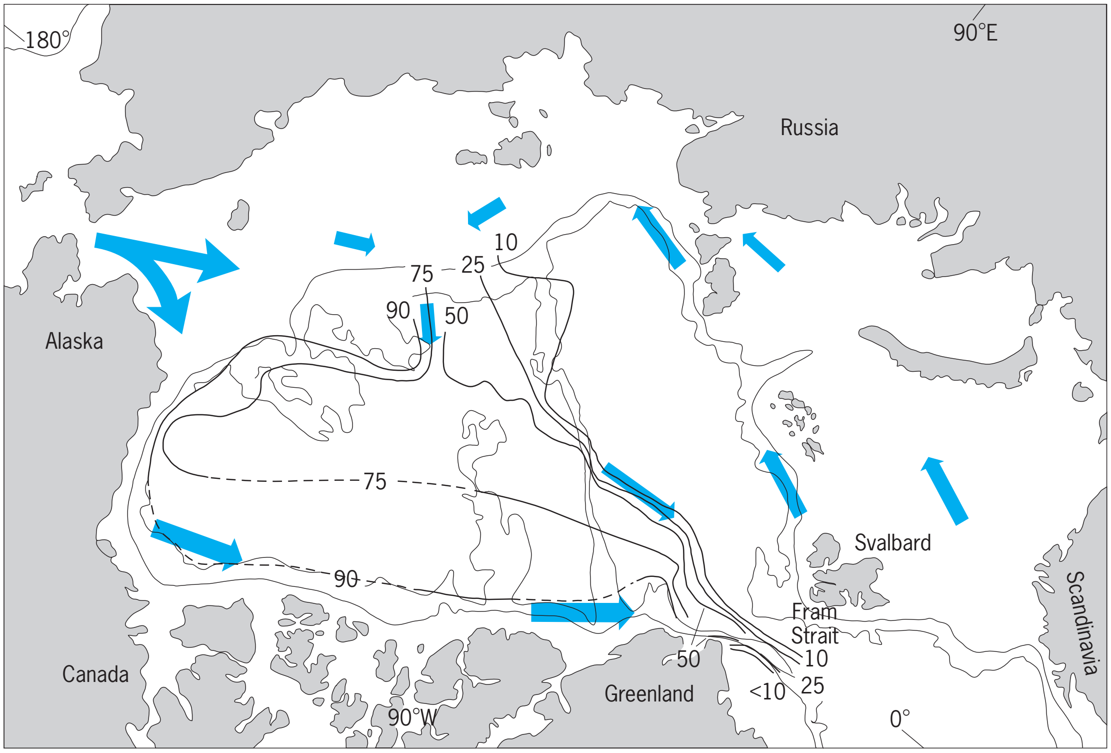


Fig. 4. Surface water (top 30 m; 100 ft) circulation determined from natural chemical tracers. (*After E. P. Jones et al., Distribution of Atlantic and Pacific waters in the upper Arctic Ocean: Implications for circulation, Geophys. Res. Lett., 25:765–768, 1998*)

Two distinct water masses make up the halocline. In the Canadian Basin, the shallower part of the halocline is characterized by high nutrient concentrations. This part of the halocline is mostly of Pacific origin but is modified probably in the Chukchi-East Siberian Sea region. It is confined largely to the Canadian Basin and flows out through the Canadian Archipelago. A narrow stream is seen flowing by the Morris Jesup Plateau and exiting through Fram Strait. Throughout both the Eurasian and Canadian basins is deeper halocline water. In the Eurasian Basin, it is found immediately under the Polar Mixed Layer; in the Canadian Basin, it lies under the less dense water originating mostly in the Pacific Ocean. This deeper halocline water is formed from water of Atlantic origin modified by melting sea ice as the Atlantic water enters though Fram Strait. Both parts of the halocline receive some fresh water from rivers and from sea ice meltwater. The circulation pattern of the whole halocline seems to generally follow that of the Atlantic Layer (see below), though it is not as definitively traced.

Warmer water, with core temperatures between 0.5 and about 2.5°C (33 and 36.5°F), has historically been denoted the Atlantic Layer. This water is warmest as it enters through Fram Strait near Svalbard. As it circulates around the Arctic basins, it cools to about 0.5°C (33°F) before returning to and exiting through Fram Strait (Fig. 4). Mixing with surface water that would otherwise allow heat exchange with the atmosphere is strongly inhibited by the density gradient of the halocline. The cooling of the Atlantic Layer is largely a result of entrainment of cold water from shelf plumes, which in turn warm and transport heat to the deeper layers as they sink.

The Upper Polar Deep Water extends from beneath the Atlantic Layer to the sill depth of the Lomonosov Ridge, about 1700 m (5575 ft) deep. It is characterized by increasing salinity and decreasing temperature with depth, both likely also influenced by shelf plumes. Its circulation is similar to that of the Atlantic Layer (**Fig. 5**).

The Eurasian Basin Deep Water is slightly colder and fresher than the Canadian Basin Deep Water. The differences are small, but have been hard to explain because the major source of the Canadian Basin Water was presumed to be cold water from the Eurasian Basin overflowing the Lomonosov Ridge. The higher temperature especially was a puzzle. An explanation again lies in the shelf plumes. Since the continental shelves of the Canadian Basin are generally shallower than those of the Eurasian Basin, shelf plumes dense enough to reach deep waters in the Canadian Basin would have to pass through the warm Atlantic Layer, entraining warmer water and carrying with them both heat and salt to the deep water. In the Eurasian Basin, the shelves are generally deeper. Shelf plumes reaching the deep waters in the Eurasian Basin bypass the Atlantic Layer by flowing under it in deep canyons such as the St. Anna Trough. Very little can be said about the circulation of the deep waters except that it is slow.

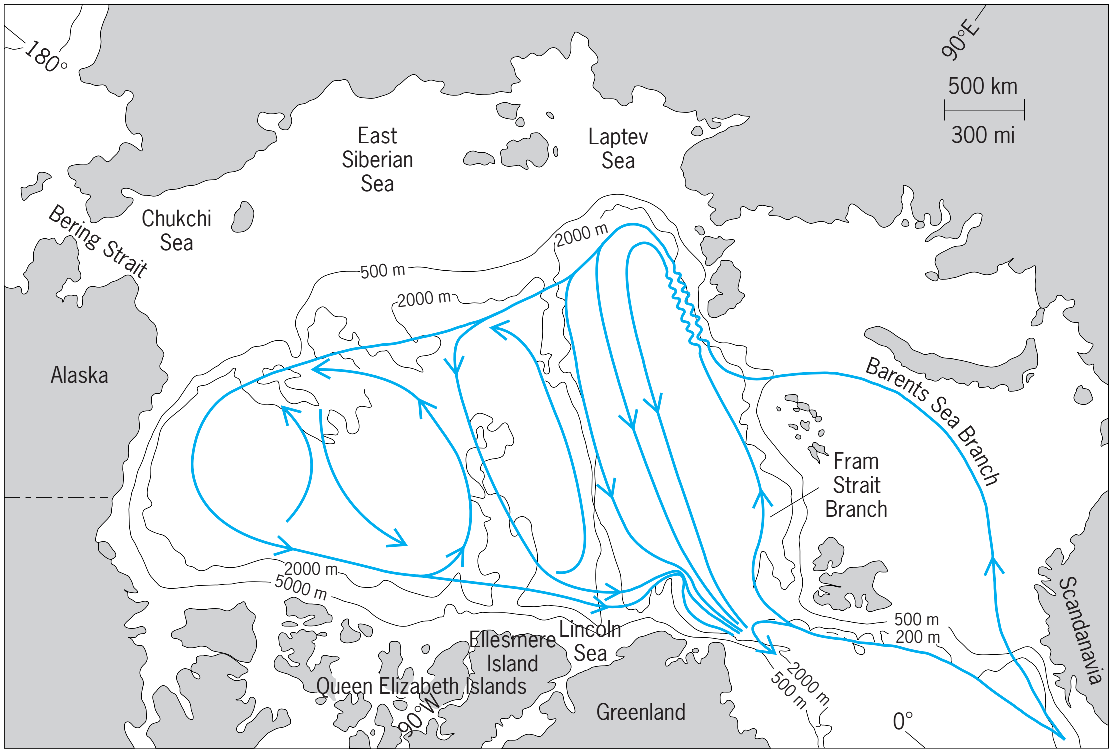


Fig. 5. Mean circulation of the Atlantic Layer and Upper Polar Deep Water. (*After B. Rudels et al., On the intermediate depth waters of the Arctic Ocean, in O. M. Johannessen, R. D. Muench, and J. E. Overland, eds., The Polar Oceans and Their Role in Shaping the Global Environment: The Nansen Centennial Volume, American Geophysical Union, Washington, DC, 1994*)

Climate change affecting the ocean. Up until the 1980s, the concept of the Arctic Ocean was of a quiet, steady-state ocean not interacting much with the global ocean. At the end of the twentieth century, a very different picture emerged. Changes include warming of the Atlantic Layer and a retreat and diminishing of the cold halocline. There has also been a major shift in the boundary between Atlantic and Pacific waters in the surface Polar Mixed Layer and halocline waters. The new observations, together with evidence of changing circulation patterns in the near-surface waters, show that the Arctic Ocean is not a quiescent environment remaining in a quasi-steady state, but exhibits as large a variability as other oceans.

In 1990, temperatures in the Atlantic Layer in the Eurasian Basin were observed to be about 1°C (2°F) warmer than those reported earlier. In 1993, warmer Atlantic Layer water (1.4°C; 34.5°F) was observed in the Canadian Basin over the Mendeleyev Ridge, far past the historical boundary of warmer part of the Atlantic Layer near the Lomonosov Ridge. At the North Pole, where the temperature maximum was historically near 0.8°C (33.4°F), the Atlantic Layer has become warmer and thicker, moving higher in the water column. By contrast, in the Canadian Basin north of Alaska, the temperature, salinity, and density were almost identical in 1985 and 1997, and are consistent with past findings. The type of change observed elsewhere had not occurred in this region.

Climate models predict that global warming as a result of the buildup of greenhouse gases (carbon dioxide, methane, and others) in the atmosphere will be enhanced in polar regions. A significant finding regarding the warming in the Atlantic Layer is that it can be related to warming in West Spitsbergen current off South Cape (Svalbard) and warming of the source waters in the Norwegian Sea. The explanation for this seems to lie in a shifting weather pattern over the North Atlantic Ocean rather than the warming predicted by climate models. *See* CLIMATE MODELING; GREENHOUSE EFFECT.

In addition to warming, two major changes have occurred in Arctic Ocean waters. There has been a diminishment of the density gradient in the halocline, and there has been a shift in the boundary between waters of Atlantic and of Pacific origin in the Polar Mixed Layer and halocline. The halocline is the barrier to heat transport from the relatively warm Atlantic Layer to the surface, where the ice is. If the halocline were to disappear or be much diminished, there is more than enough heat to melt all of the Arctic Ocean ice. Observations in the Eurasian Basin show an increase in near-surface salinity, with the halocline deepening and disappearing. The barrier depends, initially at least, on where fresh water from rivers goes. Higher salinity near surface water suggests that there has been a change in where the river runoff flows as it enters the Arctic Ocean.

Whether the observed changes are due to a large, but normal, fluctuation or are part of a trend is not yet known. Higher temperatures in the Atlantic Layer or Norwegian Sea would not, by themselves, lead to a shift in the boundary between the Atlantic and Pacific domains or to the changes in the halocline. Most striking, the Polar Mixed Layer in the interior of the Eurasian Basin is more saline than historical data (34.3 relative conductivity), and it is deeper (140 m; 460 ft) and reaches down to the thermocline at the top of the Atlantic Layer. In the near-surface water in the Eurasian and Makarov basins, there is less fresh water. This points to the river runoff into the Arctic Ocean being redistributed or decreased.

Global climate change. There is the question of whether changes in the ocean-atmosphere system of the Arctic will have global change implications. Two ways in which the Arctic Ocean could influence the global oceans and climate change are through the transport of fresh water and through changes to the global ocean "conveyor belt." Fresh-water transport is basic to global atmospheric and oceanic climate systems. The Arctic Ocean is a major pathway for the return flow of fresh water from the North Pacific Ocean, where precipitation exceeds evaporation, to the North Atlantic Ocean, where evaporation is greater than precipitation. Fresh water is returned to the Atlantic Ocean via rivers flowing into the Arctic Ocean and from the North Pacific via the relatively fresh flow into the Arctic Ocean through Bering Strait.

The global ocean conveyor belt is a highly schematic depiction of how the ocean operates in a climate context. The conveyor belt depends on processes that make surface water dense enough to sink. In polar regions, this happens in winter when surface waters cool and brine is released during ice formation. The newly formed deep water flows toward the Equator, with upwelling in tropical regions, where it becomes warm. Of the heat transported from tropical to northern regions, the atmosphere accounts for about one-half and the northward flowing warm water carries the other half. Most portrayals show the northern "terminus" of the conveyor belt to be in the Nordic seas. Recent work suggests, however, that more than 30% of the water contributing to the deep circulation from the Arctic region could be the result of processes occurring within the Arctic Ocean. *See* OCEAN CIRCULATION.

A change in the thermohaline circulation brought about by changes in fresh-water budgets or in ice formation in the Arctic Ocean could have a significant impact on this heat transport. On a longer time scale, such changes in the past are thought to have brought on ice ages, for example. In view of the fact that variability in the Arctic has long seemed to have some special involvement in global change, there is still the question of its role in global climate change. E. P. Jones

Bibliography. W. S. Broecker, Chaotic climate, *Sci. Amer.*, 273:5, November 1995; S. G. Gorshkov, *World Ocean Atlas*, vol. 3, Pergamon Press, Oxford, 1983; O. M. Johannessen, R. D. Muench, and J. E. Overland (eds.), *The Polar Oceans and Their Role in Shaping the Global Environment: The Nansen Centennial Volume*, American Geophysical Union, Washington, DC, 1994; F. Nansen, *Farthest North*, vols. 1 and 2, George Newnes, London, 1898.

Arcturus

The brightest star in the northern sky, apparent magnitude −0.05, also known as α Boötis. It is a yellow giant star of spectral type K1.5, one of the nearest giants to the Earth at a distance of 11.25 parsecs (2.16×10^{14} mi or 3.47×10^{14} km). Unlike the Sun, which is currently converting hydrogen into helium in its core, Arcturus has already exhausted its central hydrogen and has evolved away from the main sequence. It is approximately 25 times larger in diameter than the Sun, and more than 100 times more luminous. Its effective temperature is estimated to be 7700°F (4300 K). *See* BOÖTES; SPECTRAL TYPE; STELLAR EVOLUTION.

Arcturus has a large space motion relative to the Sun, and it is a member of the high-velocity group of stars known as Population II, associated with the halo of the Milky Way Galaxy. The component of its space velocity in the direction of galactic rotation is −75 mi/s (−120 km/s) relative to the local standard of rest, implying it moves around the galactic center much more slowly than the Sun. Its galactocentric orbit is highly eccentric, while the Sun's is essentially circular. The pattern of chemical abundances is also quite typical of Population II; heavy elements such as iron and nickel are about three times more deficient compared to hydrogen than they are in the Sun. while light elements such as oxygen, magnesium, and silicon are enhanced in about the same degree as in the halo stars. Water vapor has been detected in the photosphere of the star. *See* STELLAR POPULATION.

Precise measurements have revealed small variations in the radial velocity of Arcturus with a period close to 2 days and an amplitude of about 660 ft/s (200 m/s), which are believed to be due to pulsation. Variations with much longer periods of 1 or 2 years have also been reported, and could be related to convective motions in the star's photosphere or the presence of surface features moving across the disk of the star as it rotates around its axis. Similar variations have also been found in other giant stars. Brightness variations at the level of a few percent have been seen in Arcturus, and are presumably also related to pulsation. *See* STAR. David W. Latham

Area

The superficial contents of a geometrical figure of two dimensions. The area of any rectangle or square is the product of two adjacent sides, one of which may be called the base and the other the altitude. In general, any line segment that partially bounds a plane geometric figure may be called a base if its line does not separate the figure, and a perpendicular

Area formulas

Figure	Formula
Triangle	$hb/2$, where $h =$ altitude, $b =$ base; $\sqrt{s(s-a)(s-b)(s-c)}$, where $s = (1/2)(a+b+c)$, and a, b, and c are sides of the triangle
	ab, where a and b are adjacent sides
Square	a^2, where $a =$ side
Parallelogram	$ab \sin \theta$, where a and b are adjacent sides, and θ is the angle between the sides
Trapezoid	$(1/2)(a+b)h$, where a and b are the parallel sides, and h is the altitude
Quadrilateral	$(1/2)ab \sin \theta$, where a and b are the diagonals, and θ is the angle between them
Regular polygon	$(1/4)nl^2 \cot 180^\circ/n$, where n is the number of sides, each of length l
Circle	πr^2, where $r =$ radius
Ellipse	πab, where a and b are semiaxes
Sphere	$4\pi r^2$, where $r =$ radius
Spherical triangle	$(A + B + C - \pi)r^2$, where A, B, and C are angles (radians), and r is the radius

drawn to the base line from one of its points at greatest distance may be called the altitude. The area of a parallelogram is equal to the product of its base times its altitude. The area of a triangle is one-half the product of its base times its altitude. The area of a trapezoid is equal to one-half the product of the sum of its parallel sides (bases) times its altitude. Some area formulas are given in the **table**. *See* EUCLIDEAN GEOMETRY. J. Sutherland Frame

Arecales

An order of flowering plants, division Magnoliophyta (Angiospermae), of the subclass Arecidae in the class Liliopsida (monocotyledons). The name of the subclass is derived from the ordinal name. The order consists of the single family Arecaceae (Palmae), the palms, with more than 200 genera and nearly 3000 species, largely confined to tropical and subtropical regions. The order Arecales has also been called Palmales or Principes.

Most palms are trees with an unbranched trunk and a terminal crown of large leaves. Less common types are scandent, with scattered leaves, or more or less acaulescent, with the leaves arising from the ground. The leaf has a blade, petiole, and sheath; the sheath is open or closed, but in any case it fully encircles the stem at the base. The blade as a whole is usually pinnately compound, less often palmately compound or merely lobed. It is pinnately or less often palmately veined, but the individual pinnae or segments have more or less parallel veins and are plicate between the veins. The plicate structure of all palm leaves relates to a complex ontogeny shared only by some of the Cyclanthaceae (Cyclanthales), in which new tissue continues to develop along the folds after the flanking tissues have matured. *See* ARECIDAE; CARNAUBA WAX; COCONUT; CYCLANTHALES; DATE; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM; VEGETABLE IVORY. Arthur Cronquist

Arecidae

A subclass of the class Liliopsida (monocotyledons) of the division Magnoliophyta (Angiospermae), the flowering plants, consisting of four orders (Arecales, Cyclanthales, Pandanales, and Arales), five families, and nearly 6000 species. Except for the highly reduced family Lemnaceae (Arales), they have an inflorescence of usually numerous, small flowers, generally subtended by a prominent spathe and often aggregated into a spadix. Except in the Araceae (Arales), the endosperm seldom contains much starch.

More than 80% of the species have broad, petiolate leaves that do not have the typical parallel venation commonly associated with monocotyledons, and more than half of the species are arborescent (likewise an unusual character among the monocotyledons). *See* ARALES; ARECALES; CYCLANTHALES; LILIOPSIDA; MAGNOLIOPHYTA; PANDANALES; PLANT KINGDOM. Arthur Cronquist; T. M. Barkley

Arenaceous rocks

The arenaceous rocks (arenites) include all those clastic sedimentary rocks whose particle sizes range from 2 to 0.06 mm, or if silt is included, to 0.004 mm. Some arenites are composed primarily of carbonate particles, in which case they are called calcarenites and grouped with the limestones. Some oolitic iron ores and glauconite beds are properly classified as arenites. But the vast majority of arenites are commonly called sandstones, and the two words are almost synonymous. *See* CALCARENITE; GRAYWACKE; OOLITE; SANDSTONE; SEDIMENTARY ROCKS. Raymond Siever

Bibliography. H. Blatt and R. Tracy, *Petrology: Igneous, Sedimentary, and Metamorphic*, 1995; M. E. Tucker, *Sedimetary Petrology: An Introduction to the Origin of Sedimentary Rocks*, 3d ed., 2001.

Argentiniformes

An order of teleost fishes in the superorder Protacanthopterygii. These fishes, collectively called the argentines or herring smelts, were elevated from Osmeriformes (which are treated herein as a superfamily of the Salmoniformes) to ordinal rank.

Argentiniforms are identifiable by the following combination of characters: body variable, elongate and cylindrical in some to short and compressed

Argentiniformes features

Argentinoidei (families Argentinidae, Opisthoproctidae, and Microstomatidae)	Alepocephaloidei (families Platytrocidae, Bathylaconidae, and Alepocephalidae)
Eyes tubular or not tubular	Eyes not tubular
Adipose fin usually present	Adipose fin absent
Dorsal fin inserted near center of body	Dorsal fin inserted well back on body
Maxillary and premaxillary, when present, toothless	Upper jaw usually with teeth
Mouth usually small	Mouth usually large
Swim bladder present or absent	Swim bladder absent
Eggs small, development gradual	Eggs large, development direct
12 genera, 61 species	38 genera, 103 species

in others; jaw mechanism and dentition greatly reduced; premaxillary absent in some species, if present lacks teeth; maxillary usually toothless; eyes variously directed, some species with tubular eyes; some species with specialized light organs, which may be associated with tubular eyes; no serial photophores; adipose fin present or absent; caudal fin forked; swim bladder, when present, physoclistous (not attached to gut). In addition, the argentines have a crumenal organ (paired branchial pouches in which food particles are trapped by large interlocking toothed gill rakers).

Two distinct groups (suborders in this classification), Argentinoidei and Alepocephaloidei, are identified by the characters listed in the **table**.

Argentinoidei. This suborder consists of the families Argentinidae, Opisthoproctidae, and Microstomatidae.

The Argentinidae (argentines or herring smelts) is found in the Atlantic, Indian, and Pacific oceans, and consists of two genera, with about 19 species.

The Opisthoproctidae (barreleyes or spookfishes) is found in tropical to temperate Atlantic, Indian, and Pacific oceans, and consists of six genera with about 10 species.

The Microstomatidae (includes Bathylagidae, deep-sea smelts) is found in all seas, mostly ranging from subarctic to Antarctic waters, and consists of four genera, with about 32 species. *Bathylagus*, the largest genus, contains about 15 species.

Alepocephaloidei. This suborder consists of the families Platytrocidae, Bathylaconidae, and Alepocephalidae.

The Platytrocidae (searsides) is found in all oceans, but is absent from the Mediterranean Sea. It consists of 13 genera, with 37 species.

The Bathylaconidae (no common name) is found in circumtropical waters, and consists of two genera, with three species.

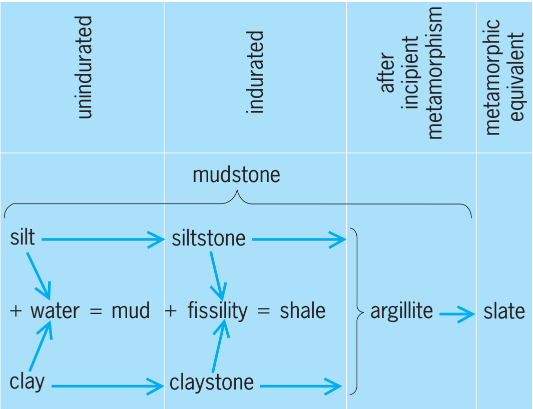
The Alepocephalidae (slickheads) is found in all oceans, and consists of 22 genera, and no less than 60 species.

Opisthoproctidae specializations. Some of the most bizarre fishes in all the seas of the world are the family Opisthoproctidae. As a family, most of the species have tubular eyes; pectoral fins on the lower side of the body; pelvic fins posteriorly placed; and parietal bones that do not meet on the midline of the skull. Most lack a swim bladder. Most species are long and subcylindrical, but one genus, *Opisthoproctus*, has a foreshortened, deep and compressed body and sole (an elongate forward projection on the underside of the abdomen that may act as a reflector for the light organ inside the body near the anus) covered with large deciduous cycloid scales. The tubular eyes are directed dorsally and topped with a gelatinous transparent tissue. The interorbital bones are reduced to a mere sliver, exposing the brain through the transparent roof of the skull; the premaxillary is absent; and the maxillary is reduced to a thin scalelike bone which is easily lost, leaving the toothless dentary to occupy almost half the length of the snout. Some of these bathypelagic fishes have never been seen alive and are known from only a few specimens. *See* AULOPIFORMES; SALMONIFORMES. Herbert Boschung

Bibliography. D. P. Begle, Monophyly and relationships of the argentinoid fishes, *Copeia*, 1992(2):350–366, 1992; D. M. Cohen, Argentinoidea, pp. 1–70 in *FWNA Fishes of the Western North Atlantic*, pt. 4, 1964; D. M. Cohen, Argentinidae, pp. 215–216 in M. M. Smith and P. C. Heemstra (eds.), *Smiths' Sea Fishes*, Springer-Verlag, Berlin, 1986; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, 2006; J. R. Paxton and D. M. Cohen, *Argentinidae*, p. 1884 in K. E. Carpenter and V. H. Niem (eds.), *FAO Species Identification Guide for Fishery Purposes: The Living Marine Resources of the WCP*, vol. 3: *Batoid Fishes, Chimaeras and Bony Fishes, Part 1 (Elopidae to Linophrynidae)*, FAO, Rome, 1999; M. Schneider, Argentinidae: Argentinas, pp. 858–859 in W. Fischer et al. (eds.), *Guia FAO para Identification de Especies para lo Fines de la Pesca*, Pacifico Centro-Oriental, 3 vols., FAO, Rome, 1995.

Argillaceous rocks

Clastic sedimentary rocks containing silt- or clay-sized particles that are less than 0.0625 mm and/or clay minerals. The argillaceous rocks (lutites) include shales, argillites, siltstones, and mudstones. They are



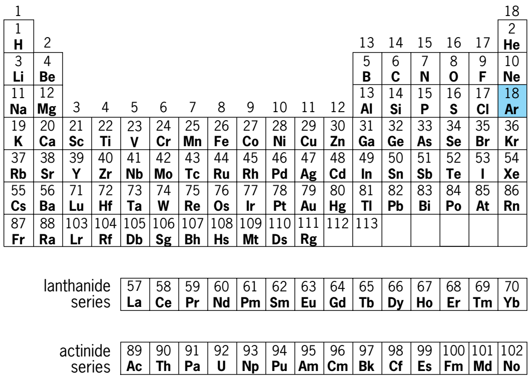
Classification of fine-grained mechanical sediments. (*After F. J. Pettijohn, Sedimentary Rocks, 2d ed., Harper & Row, 1957*)

the most abundant sedimentary rock type, varying according to different estimates from 44 to 56% of the total sedimentary rock column. Claystone is hardened or cemented (indurated) clay, which consists dominantly of fine material of which at least a major proportion is clay mineral (hydrous aluminum silicates). Shale is a laminated or fissile claystone or siltstone, in general more consolidated than claystone. Mudstone is a claystone that is blocky and massive. The term argillite is used for rocks which are more indurated than claystone or shale but not metamorphosed to slate. All these argillaceous rocks are consolidated equivalents of muds, oozes, silts, and clays (see **illus.**). Loess is a fine-grained, unconsolidated, windblown deposit. The term shale has been used by many authors generically to denote all of these types of rock. *See* BENTONITE; CLAY; CLAY MINERALS; LOESS; SEDIMENTARY ROCKS; SHALE. Raymond Siever

Bibliography. H. Blatt and R. Tracy, *Petrology: Ingenious, Sedimentary, and Metamorphic*, 1995; M. E. Tucker, *Sedimetary Petrology: An Introduction to the Origin of Sedimentary Rocks*, 3d ed., 2001.

Argon

A chemical element, Ar, atomic number 18, and atomic weight 39.948. Argon is the third member of group 0 in the periodic table. The gaseous elements in this group are called the noble, inert, or rare gases, although argon is not actually rare. The Earth's atmosphere is the only natural argon source; however, traces of this gas are found in minerals and meteorites. Argon constitutes 0.934% by volume of the Earth's atmosphere. Of this argon, 99.6% is the argon-40 isotope; the remainder is argon-36 and argon-38. There is good evidence that all the argon-40 in the air was produced by the radioactive decay of the radioisotope potassium-40. *See* INERT GASES; PERIODIC TABLE.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 **H**																	2 **He**
3 **Li**	4 **Be**											5 **B**	6 **C**	7 **N**	8 **O**	9 **F**	10 **Ne**
11 **Na**	12 **Mg**											13 **Al**	14 **Si**	15 **P**	16 **S**	17 **Cl**	18 **Ar**
19 **K**	20 **Ca**	21 **Sc**	22 **Ti**	23 **V**	24 **Cr**	25 **Mn**	26 **Fe**	27 **Co**	28 **Ni**	29 **Cu**	30 **Zn**	31 **Ga**	32 **Ge**	33 **As**	34 **Se**	35 **Br**	36 **Kr**
37 **Rb**	38 **Sr**	39 **Y**	40 **Zr**	41 **Nb**	42 **Mo**	43 **Tc**	44 **Ru**	45 **Rh**	46 **Pd**	47 **Ag**	48 **Cd**	49 **In**	50 **Sn**	51 **Sb**	52 **Te**	53 **I**	54 **Xe**
55 **Cs**	56 **Ba**	71 **Lu**	72 **Hf**	73 **Ta**	74 **W**	75 **Re**	76 **Os**	77 **Ir**	78 **Pt**	79 **Au**	80 **Hg**	81 **Tl**	82 **Pb**	83 **Bi**	84 **Po**	85 **At**	86 **Rn**
87 **Fr**	88 **Ra**	103 **Lr**	104 **Rf**	105 **Db**	106 **Sg**	107 **Bh**	108 **Hs**	109 **Mt**	110 **Ds**	111 **Rg**	112	113					

lanthanide series	57 **La**	58 **Ce**	59 **Pr**	60 **Nd**	61 **Pm**	62 **Sm**	63 **Eu**	64 **Gd**	65 **Tb**	66 **Dy**	67 **Ho**	68 **Er**	69 **Tm**	70 **Yb**
actinide series	89 **Ac**	90 **Th**	91 **Pa**	92 **U**	93 **Np**	94 **Pu**	95 **Am**	96 **Cm**	97 **Bk**	98 **Cf**	99 **Es**	100 **Fm**	101 **Md**	102 **No**

Argon is colorless, odorless, and tasteless. The element is a gas under ordinary conditions, but it can be liquefied and solidified readily. Some properties of the gas are listed in the **table**. Argon does not form any chemical compounds in the ordinary sense of the word, although it does form some weakly bonded clathrate compounds with water, hydroquinone, and phenol. There is one atom in each molecule of gaseous argon.

Properties of argon

Property	Value
Atomic number	18
Atomic weight (atmospheric argon)	39.948
Melting point (triple point), °C	−189.4
Boiling point at 1 atm pressure, °C	−185.9
Gas density at 0°C and 1 atm (101.325 kPa) pressure, g/liter	1.7840
Liquid density at normal boiling point, g/ml	1.3998
Solubility in water at 20°C, ml argon (STP) per 1000 g water at 1 atm (101.325 kPa) partial pressure of argon	33.6

The oldest large-scale use for argon is in filling electric light bulbs. Welding and cutting metal consumes the largest amount of argon. Metallurgical processing constitutes the most rapidly growing application. Argon and argon-krypton mixtures are used, along with a little mercury vapor, to fill fluorescent lamps. Argon mixed with a little neon is used to fill luminous electric-discharge tubes employed in advertising signs (similar to neon signs) when a blue or green color is desired instead of the red color of neon. Argon is also used in gas-filled thyratrons, Geiger-Müller radiation counters, ionization chambers which measure cosmic radiation, and electron tubes of various kinds. Argon atmospheres are used in dry boxes during manipulation of very reactive chemicals in the laboratory and in sealed-package shipments of such materials.

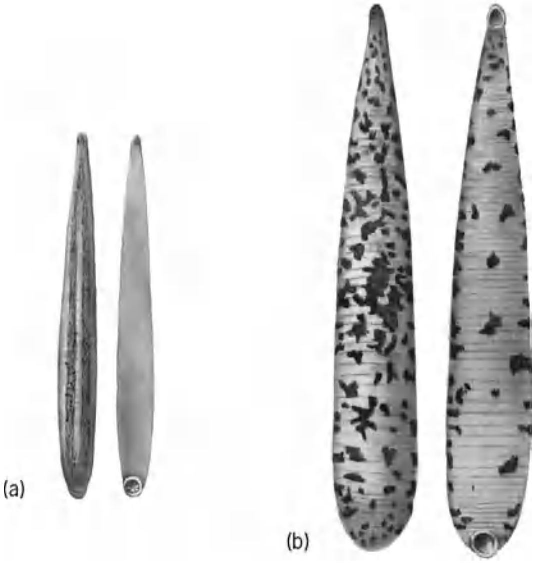
Most argon is produced in air-separation plants. Air is liquefied and subjected to fractional distillation. Because the boiling point of argon is between that of nitrogen and oxygen, an argon-rich mixture can be taken from a tray near the center of the upper distillation column. The argon-rich mixture is further distilled and then warmed and catalytically burned with hydrogen to remove oxygen. A final distillation removes hydrogen and nitrogen, yielding a very high-purity argon containing only a few parts per million of impurities. A. W. Francis

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; M. Ozima and F. A. Podosek, *Noble Gas Geochemistry*, 2001.

Arhynchobdellae

An order of the class Hirudinea (leeches) which do not have an eversible proboscis, but frequently have three jaws armed with sharp teeth. The blood of these annelids contains hemoglobin. They may be divided into the Gnathobdellae, with jaws, and the Pharyngobdellae, without jaws. *See* HEMOGLOBIN.

Gnathobdellae have bodies which are oval in cross section and have a conspicuous posterior sucker. The anterior sucker does not project beyond the sides of the body but forms a deep cup on the underside of the head in which the jaws can work to make their incision in the host. This group contains most of the important bloodsucking, leech parasites of humans and other warm-blooded animals. Some have often been used by medical practitioners such as *Macrobdella decora* in North America, *Hirudo medicinalis* in Europe and northern Asia, and *Hirudinaria granulosa* in southern and eastern Asia. The land leeches, such as *Haemadipsa*, are members of this group. They occur in great numbers on vegetation in swamp and jungle areas and attach themselves to passing warm-blooded animals. They make traveling on foot in these areas exceedingly unpleasant, and death from loss of blood or secondary infection is not uncommon. *Haemopis* is an example of a gnathobdellid leech which has lost its bloodsucking habits (**illus.** *a*). Its food consists of small invertebrates such as earthworms, which it ingests whole. The teeth are blunt.



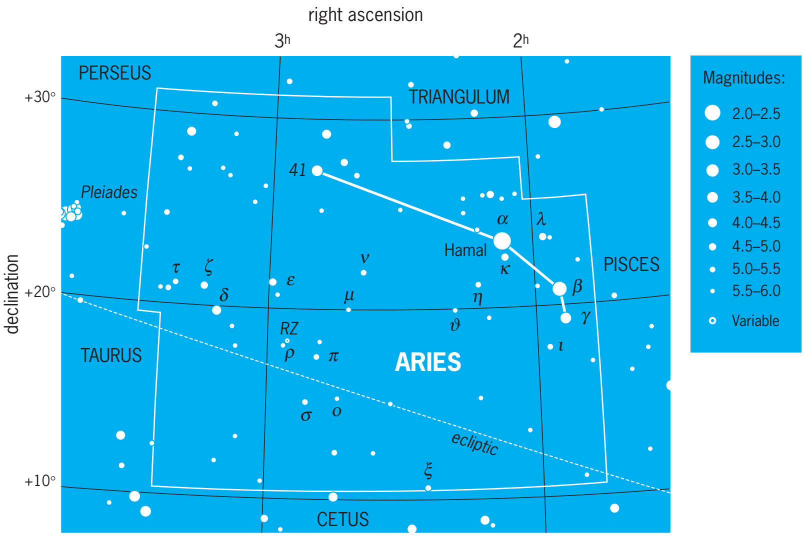
Arhynchobdellae. (*a*) *Haemopis grandis*. (*b*) *Erpobdella punctata*.

Pharyngobdellae are specialized for carnivorous diets and in many cases have completely lost the jaws. They have a strong muscular pharynx which extends nearly half the length of the body. *Erpobdella* (illus. *b*) is common in lakes and streams in the Northern Hemisphere, while *Trocheta* tends to leave the water and forage in moist soil. *See* HIRUDINEA; RHYNCHOBDELLAE.

K. H. Mann

Aries

The Ram, a northern zodiacal constellation, meaning that the path of the Sun and planets passes through it (see **illustration**). In Greek mythology, the golden fleece of this ram was pursued by Jason and the Argonauts, whose ship is memorialized in four current-day constellations in the southern sky. In Johannes Bayer's 1603 star atlas, the first three lettered stars are located in the ram's horns. *See* ZODIAC.



Modern boundaries of the constellation Aries, the Ram. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

When the constellations were organized about 2000 years ago, the point where the ecliptic and the celestial equator cross, the point of the spring equinox, was in Aries, so that point is known as the first point of Aries. (Because of the precession of the equinoxes, that point has now moved into the constellation Pisces.) *See* EQUINOX; PRECESSION OF EQUINOXES.

The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION.

Jay M. Pasachoff

Aristolochiales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Magnoliidae of the class Magnoliopsida (dicotyledons). It contains only the family Aristolochiaceae, with 8 or 10 genera and about 600 species, most of them in tropical and subtropical regions. Within its subclass the order is marked by the presence of ethereal oil cells, by its uniaperturate or nonaperturate pollen, and especially by its strongly perigynous to epigynous flowers, usually with united carpels, that typically lack petals and have the sepals joined into a highly irregular, corolloid calyx. Many of the species are climbing vines. *Aristolochia* (birthwort or Dutchman's pipe) and *Asarum* are well-known genera of the order.

Several parasitic, nonchlorophyllous families that have been included by many authors in the Aristolochiales are here treated as forming a separate order, Rafflesiales in the subclass Rosidae. *See* MAGNOLIIDAE; MAGNOLIOPHYTA; MAGNOLIOPSIDA; RAFFLESIALES.

Arthur Cronquist

Arithmetic

A branch of mathematics dealing with numbers, operations on numbers, and computation. Arithmetic is useful in solving many practical problems, such as buying, selling, budgets, sports statistics, and measurement. The usual numbers of arithmetic are whole numbers, fractions, decimals, and percents. Beyond the numbers of arithmetic are negative numbers, rational numbers, and irrational numbers. The rational and irrational numbers together constitute the real numbers.

Whole numbers. The whole numbers include the infinite sequence of counting numbers—one, two, three, four, five, . . . —and the number zero. For numbers to ten, a single symbol is used, and for larger numbers a combination of symbols.

Numbers to ten. Numbers to ten are designated with a single digit; the digits are 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Counting by ones shows "how many." To extend this counting concept and to prepare for operations on the numbers, it is helpful to recognize the numbers as made up of parts, for example to see five on dice as 4 and 1, as 3 and 2, or as 2 and 2 and 1.

Numbers ten and greater. Numbers ten and greater are expressed by using a combination of the ten digits, with the place of the digit indicating the value of the digit. This place-value system, named the Hindu-Arabic numeration system, is now used around the world. *See* NUMBERING SYSTEMS.

In a multidigit numeral, the value of each place from right to left is a successive power of ten and the total values for all places are combined or added. For example, the meaning of 1,234 is given in (1).

$$\text{1 thousand} + \text{2 hundreds} + \text{3 tens} + 4 \quad \text{or}$$
$$1 \times 10^3 + 2 \times 10^2 + 3 \times 10 + 4 \qquad (1)$$

Each three places denotes a period. The first period on the right is the ones or units period, but this name is usually omitted. The next six period names are thousand, million, billion, trillion, quadrillion, and quintillion. Numbers within each period are read as if they were in the ones period and then the period name is given. For example, 123,456,789,721,042 is read "one hundred twenty-three trillion four hundred fifty-six billion seven hundred eighty-nine million seven hundred twenty-one thousand forty-two."

Operations. The basic operations are addition (+) and multiplication (× or ·), with subtraction (−) and division (÷) defined, respectively, by using addition and multiplication.

If two numbers, a and b, are combined or added, the result is a number, c, called the sum. In the example, $4 + 6 = 10$, 4 and 6 are addends and 10 is the sum. The whole amount, 10, is the result of combining two parts, 4 and 6.

If a given number, n, sets of objects with the same number in each set, r, are combined, then multiplication of n and r is the total number of objects. In $3 \times 4 = 12$, 3 and 4 are factors and 12 is the product (**Fig. 1**).

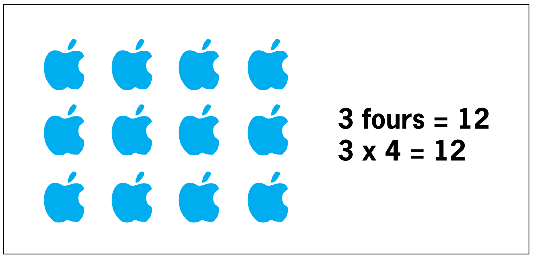


Fig. 1. One meaning of multiplication.

The multiplication 3×4 can also be thought of as the number of members in a rectangular array with 3 rows and with 4 objects in each row. Another use is to find the number of ways that 3 objects can be paired with 4 objects, for example, matching 3 pairs of trousers and 4 shirts. Yet another use of multiplication is to find the area of a rectangular region, and in turn the area of other kinds of regions. *See* AREA.

Subtraction is the inverse operation to addition, finding an addend when a sum and one addend are known. If there are 18 children on the playground and 10 are boys, then the number of girls is $18 - 10$ or 8, illustrating that the whole minus a part leaves the other part. Subtraction is used also to find the difference, for example, to see how many more are in one group of 18 children than in a group of 10 children. Subtraction is checked with addition; $18 - 10 = 8$ because $8 + 10 = 18$.

Division is the inverse operation to multiplication, that is, finding a factor when a product and a factor are known. In $12 \div 3 = 4$, 12 is called the dividend, 3 is called the divisor, and 4 is the quotient. There are two applications of division, in finding the number of threes in 12, or in finding the number in each set when 12 is separated into 3 equal groups. In division such as $13 \div 4$, the quotient of 3 and a remainder of 1 is usually written as 3r1. With fractions, the quotient can be shown as a factor, $3^1/_4$.

Basic facts. Any of the four operations with single-digit numbers is a basic fact. There are 100 basic facts each for addition, subtraction, and multiplication, but only 90 for division because 0 cannot be a divisor. Thinking strategies are useful in learning facts (see **table**).

Computation. The algorithms, or rules, for computation rely on basic facts, place value, and properties of the operations. In addition and subtraction, the total values of each of the places are added or subtracted. The sum of $345 + 67 + 162$ is found by combining the amount in the ones places, the amount in the tens places, the amount in the hundreds places, and then combining these three totals. A vertical format, as in (2), makes it easy to combine like values. Mathematical reasons for combining this way are the associative

$$\begin{array}{r} 345 \\ 67 \\ +162 \\ \hline 574 \end{array} \qquad (2)$$

Strategies for learning basic facts of arithmetic

Operation	Strategy	Example
+	Count on, for 6 + 3	Six, seven, eight, nine
	Doubles, for 7 + 8	Seven plus seven = 14; so 7 + 8 is one more, or 15
	Use 10, for 9 + 6	Nine and one make 10; 10 plus 5 = 15
−	Add on, for 12 − 9	What added to make 9 makes 12?
	Count up, for 12 − 9	Nine, ten, eleven, twelve—keeping track of the counts
×	Doubles, for 8 × 7	4 sevens = 28; so 8 sevens = 28 + 28 or 56
	Use known facts, for 7 × 7	5 sevens = 35; 2 sevens = 14; so 7 sevens = 35 + 14 or 49
	Relating to 10, for 9 × 6	10 sixes = 60; so 9 sixes = 60 − 6 or 54
÷	Use multiplication, for 72 ÷ 9	How many nines make 72? What times nine = 72?
	Count up, for 20 ÷ 5	Five, ten, fifteen, twenty; so four—keeping track of the counts

property, $a + (b + c) = a + (b + c)$, and the commutative property, $a + b = b + a$. *See* ALGEBRA.

The algorithm for multiplication depends on the distributive property for multiplication over addition, $a \times (b + c) = a \times b + a \times c$. For example, $7 \times (60 + 8) = 7 \times 60 + 7 \times 8$. The algorithm (3*a*) illustrates this property and shows why the 5 is "carried" to the tens place and why 47 is moved one place to the left in the standard algorithm (3*b*).

$$\begin{array}{r} 68 \\ \underline{\times 7} \\ 7 \times 8 \rightarrow 56 \\ 7 \times 60 \rightarrow \underline{420} \\ 7 \times 68 \rightarrow 476 \end{array} \quad (3a) \qquad \begin{array}{r} {}^{5} \\ 68 \\ \underline{\times 7} \\ 476 \end{array} \quad (3b)$$

The written algorithms for addition, subtraction, and multiplication are done right to left, but division is done left to right. For example, if six people share the $750 cost of a bus rental, the quotient for $6\overline{)750}$ is found by dividing hundreds, then the 15 remaining tens, and then the 30 remaining ones, for a quotient of 125. Each person needs to pay $125.

Most of the arithmetic that people use everyday involves either mental computation or estimation.

Mental computation means finding the exact answer in one's head without paper and pencil or calculators. For all operations, mental computation is most often done left to right. For 68 + 76, one might think "60 + 70 = 130; 130 + 8 = 138; 138 + 6 = 144."

Estimation is done when only an approximate answer is needed. For example, an estimate for the area of a lawn that measures 325 ft by 62 ft might be 300 × 60 or 18,000 ft^2. A slightly better estimate is 300 × 70 or 21,000 ft^2. There are many different ways to estimate, depending on the need in the practical problem. Beyond estimation as a practical skill, it is needed to check calculations on a calculator.

Fractions and decimals. Understanding fractions and decimals, as well as operations on these numbers, is essential for practical uses and for long-term memory.

Meanings of fractions. The initial and most basic idea is that a fraction shows "part of a whole." In the fraction $^{3}/_{4}$, read "three fourths," the 4 shows the number of equal-size pieces in each whole unit as well as the size of one piece, "fourth". The 3 shows the number of equal-size pieces being taken or considered. The top number, 3, "numbers" the parts and is called the numerator. The bottom number, 4, "names" the parts and is called the denominator. For a set of objects such as a dozen eggs, the egg carton is considered the whole and individual eggs the parts. The fraction $^{3}/_{12}$ might show the part of a carton with brown eggs if 3 of the 12 eggs are brown and the rest are white.

Fraction meaning is extended to include division of a whole number by a counting number. For example, 1 ÷ 4 could mean one whole pie divided among four people. Since each person would get $^{1}/_{4}$ of a pie, $1 \div 4 = {}^{1}/_{4}$. In a similar manner, $3 \div 4 = {}^{3}/_{4}$ (**Fig. 2**).

Fractions that represent the same part of a unit are equivalent fractions. A ruler with customary units shows equivalents such $^{1}/_{4} = {}^{2}/_{8} = {}^{4}/_{16} = {}^{8}/_{32}$. The algorithm for generating equivalent fractions is: multiply or divide the numerator and denominator by the same nonzero number. For example, the fraction $^{15}/_{100}$ can be expressed as $^{3}/_{20}$ by dividing 15 and 100 by 5.

A mixed number shows a combination of wholes and parts. For example, $3^{4}/_{5}$ means 3 wholes and $^{4}/_{5}$ of another whole. The fraction does not occupy a place, as each symbol does for whole numbers. It is sometimes useful to find mixed number and fraction equivalents. For $3^{4}/_{5}$, one whole can be expressed as $^{5}/_{5}$, so 3 wholes equals $^{15}/_{5}$; $^{15}/_{5}$ and $^{4}/_{5}$ more make $^{19}/_{5}$. Hence, $3^{4}/_{5} = {}^{19}/_{5}$.

Operations on fractions. The meaning for the operations on fractions depends on the meaning of the

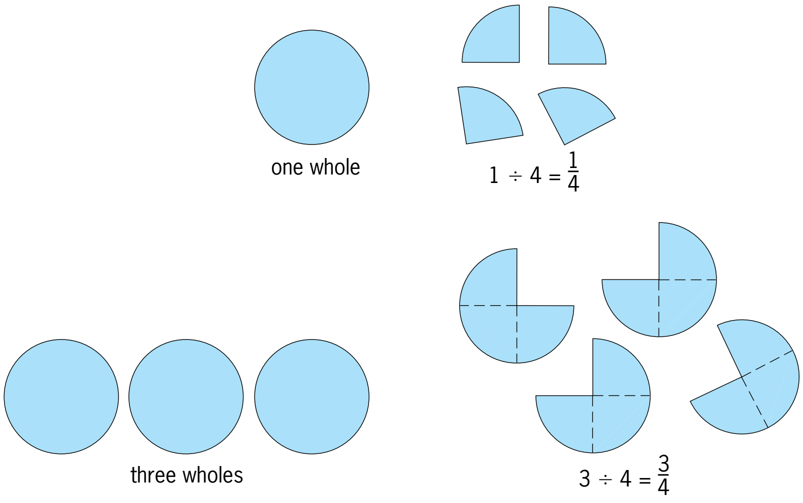


Fig. 2. Division of counting numbers as fractions.

numbers. A key idea for addition and subtraction is that equal-size parts are separated or combined. For example, each fraction in $^3/_8 + ^2/_8$ shows eights, so the parts can be combined: 3 eighths + 2 eighths makes 5 eighths. The general algorithm is given by Eq. (4).

$$\frac{a}{b}+\frac{c}{b}=\frac{a+c}{b}, \quad b\neq 0 \tag{4}$$

To add or subtract fractions with different-size pieces, both fractions are expressed as equivalent fractions with a common denominator, such as in Eq. (5). One relatively easy way to find the common denominator for 6 and 4 is to take successive multiples of the 6—that is, 6, 12, 18,. . . —until a number is found for which 4 is a factor. The first multiple, 6, does not have 4 as a factor. Since the next multiple, 12, has 4 as a factor, 12 is a common denominator, in fact the least common denominator, for 6 and 4. The least common denominator for three or more fractions is the least multiple of all of the three or more denominators.

$$\frac{3}{4}-\frac{1}{6}=\frac{9}{12}-\frac{2}{12}=\frac{7}{12} \tag{5}$$

One way to think of multiplication of fractions is to think of finding a fractional part of another number. For example, $^1/_2$ of $^6/_{11}$ is one-half of six equal pieces, each of which is one-eleventh. Finding "$^1/_2$ of the amount" is multiplication by $^1/_2$, as in Eq. (6). The usual algorithm, multiplying numerators and multiplying denominators, as in Eq. (7), gives Eq. (8). One way to multiply mixed numbers is to express each as a fraction, such as in Eq. (9).

$$\frac{1}{2}\text{ of }\frac{6}{11}=\frac{1}{2}\times\frac{6}{11}=\frac{3}{11} \tag{6}$$

$$\frac{a}{b}\times\frac{c}{d}=\frac{a\times c}{b\times d}, \quad b\neq 0, \quad d\neq 0 \tag{7}$$

$$\frac{1}{2}\times\frac{6}{11}=\frac{6}{22}=\frac{3}{11} \tag{8}$$

$$2\frac{1}{4}\times 3\frac{1}{2}=\frac{9}{4}\times\frac{7}{2}=\frac{63}{8} \tag{9}$$

One use of division with fractions is to find the number of parts contained in a number. For example, $5 \div {}^1/_8$ could mean to find how many $^1/_8$ pieces of a candy bar are in 5 candy bars. Since there are eight eighths in one whole, there are 5 × 8 or 40 in 5 wholes. Hence, $5 \div {}^1/_8 = 5 \times 8$. The general rule is: to divide by c/d, multiply by its reciprocal, d/c, as in Eq. (10), where none of the denominators b, c, or d is 0. To divide mixed numbers, each mixed number is first expressed as a fraction.

$$\frac{a}{b}\div\frac{c}{d}=\frac{a}{b}\times\frac{d}{c} \tag{10}$$

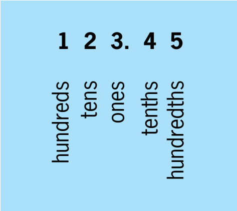


Fig. 3. Decimals with symmetry around the ones place.

Meaning of decimals. To show decimals less than one, the place value system for whole numbers is extended to the right of the ones place. The value of each place to the right is a successive power of $^1/_{10}$. The decimal point is needed to designate the ones place because the place on the right is no longer the ones place. The decimal point also separates the whole number from the decimal part. The names and values of the places are symmetric with respect to the ones place (**Fig. 3**). The 3. (decimal point included) is the center, surrounded by tens and tenths, as well as hundreds and hundredths. The meaning can be shown as $100 + 20 + 3 + {}^4/_{10} + {}^5/_{100}$. The decimal part, .45, can be thought of as 45 hundredths because $^4/_{10} = {}^{40}/_{100}$, and $^{40}/_{100} + {}^5/_{100} = {}^{45}/_{100}$.

Decimal operations. As with whole numbers, addition and subtraction with decimals is done by combining or separating equal-size parts. For example, to add 62.4 + 5.62, hundredths are first combined, then tenths, then ones, then tens for a total of 68.02.

One way to think about multiplication of decimals is to think of multiplying fractions, such as in Eq. (11).

$$0.2\times 0.3=\frac{2}{10}\times\frac{3}{10}=\frac{6}{100}=0.06 \tag{11}$$

Counting the total number of places to the right of the decimal points in the two factors shows the power of $^1/_{10}$ and consequently the number of places to the right of the decimal point in the product.

Division with decimals is usually done by solving an equivalent problem with the divisor a whole number. For example, $0.6\overline{)4.62}$ is changed to an equivalent problem, $6\overline{)46.2}$, by multiplying divisor and dividend by 10.

Percent is another way to express fractions and decimals that show hundredths. For example, 7 hundredths can be expressed as $^7/_{100}$, as 0.07, or as 7%. All three expressions show the same part of a whole. *See* PERCENT.

Joseph N. Payne

Bibliography. E. F. Krause, *Mathematics for Elementary Teachers*, 2d ed., 1991; National Council of Teachers of Mathematics, *Mathematics for the Young Child*, 1990; J. A. Van de Walle, *Elementary School Mathematics*, 2d ed., 1994.

Arkose

Arenite (rock composed of sand-size fragments) that contains a high proportion of feldspar in addition to quartz and other detrital minerals. Arkose is also known as feldspathic sandstone. Although there is no universal agreement, many geologists consider a minimum of 25% feldspar a requisite for calling sandstone an arkose. Other geologists accept a lower value.

Composition. Arkoses may contain a high proportion of other nonquartz detritus, such as igneous and metamorphic rock fragments, micas, amphiboles, and pyroxenes. Frequently the accessory heavy mineral suite consists of a variety of species. Though the arkoses are rarely as well sorted as orthoquartzites, they may be moderately well sorted. The grains are angular or poorly rounded. Clay matrix is generally subordinate in arkoses, but there may be as much as 10–15%. If there is that much, the rocks have textural similarity to feldspathic graywackes, though the mineralogy may be appreciably different in the latter. The clay is dominantly kaolinite with smaller proportions of micaceous and montmorillonitic clay. Presumably much of the kaolinite has come from the weathering of feldspar. Conglomeratic zones are common in many arkoses. Feldspathic sandstones that contain less than 25% of feldspar have been termed subarkoses. They are in general similar to the more highly feldspathic true arkoses.

Structure. Sedimentary structures of arkoses are similar in kind to those of the orthoquartzites. Cross-bedding, the major feature, may be displayed on a huge scale, some cross-bedded units being many feet thick. Bedding is crude and many times not distinguishable; many beds are thick and massive. Ripple marks may be present but are not common. Some arkoses, such as those of the Newark Series along the Atlantic Coast of the United States, contain mud cracks, frost crystal impressions, and footprints of small dinosaurs.

Occurrence. Arkoses are associated with a variety of clastic rocks, dominantly conglomerates, and reddish-colored shales. Arkoses also are found with basic lava flows. The formations occur as thick, wedge-shaped deposits, the thick end of the wedge being in close proximity to the source area and sometimes separated from it by normal faults of large displacement. Other arkoses are relatively thin formations that overlie granitic terrain. These formations grade laterally into other kinds of sandstone away from the area of underlying granite. Most arkoses are found in geosynclinal areas, but the thin, reworked, granite-wash arkoses can be found on stable continental platforms. *See* GEOSYNCLINE.

Origin. The granite-wash arkoses appear to have formed as the result of a transgression of the sea over a land area underlain by granite. The fragmented granite in the soil and mantle rock is incorporated in the basal sediment. In some areas the original granite is changed so slightly the arkose is called recomposed granite and may be almost indistinguishable from the original granite.

The origin of the arkoses is best understood in terms of the abundance of feldspar. Feldspar is unstable both chemically and mechanically as compared with quartz and, given sufficient rigors of chemical weathering at the source and abrasion during transportation, will disappear. The lack of appreciable chemical weathering at the source (which allows the contribution of much feldspar to the sediment) may be due to high topographic relief or to climatic extremes.

Since high relief and climatic extremes generally are associated with orogenic movements, arkoses are usually interpreted as sediments that result from tectonically active regions. Rift valleys formed by the divergence of two continental plates at a spreading center are now favored as the tectonic environment for arkose formation. This explains too the association with basaltic lavas. The abundant iron oxides present in most arkoses, the mud cracks, and the fanglomerates point to a mostly terrestrial mode of deposition. The thin granite washes are marine and represent different conditions. *See* FELDSPAR; GRAYWACKE; SANDSTONE; SEDIMENTARY ROCKS.

Raymond Siever

Armadillo

A heavily armored mammal in the family Dasypodidae, order Cingulata (previously Edentates or Xenarthra). Armadillos are characterized by bony dermal plates, which help to protect their bodies. Being mammals, they do have hair, but it is usually sparse and often quite inconspicuous. The anterior and posterior parts of the body are often heavily armored, whereas the midportion of the body is often encircled or partially encircled by rings. These rings are used taxonomically in describing and identifying many of the species of armadillos. The skin of the ventral part of the body is soft. The toes are clawed and used to burrow and dig for food. The hindlimbs always have five toes with claws; the forelimbs may have three, four, or five heavy claws used for digging. The snout is long, and the cylindrical viscous tongue is used in capturing food. The teeth are simple pegs with no roots and no enamel. Most species have about 14 to 18 teeth, but the giant armadillo, *Priodontes maximus*, has about 80 to 100 small teeth. It reaches a length of approximately 10 feet (3 m), whereas the smallest species, the fairy armadillo, *Chlamyphorus truncatus*, is about 5–6 in. (12–15 cm) long. When disturbed, many species roll into a ball or wedge themselves into a burrow opening, with the dermal plates helping to protect them.

Armadillos occur only in the New World. They range from the southern United States through most of South America. There are 21 species of armadillos in nine genera, but only one, the nine-banded armadillo, *Dasypus novemcinctus*, occurs in North America (see **illustration**). It is the best-known species. It is adaptable and has extended its range rapidly in the United States: it is now found from Texas and Oklahoma eastward to Florida, Georgia, and southern South Carolina. It feeds on a great variety of invertebrates, including beetles, termites, great numbers of ants, and many other insects such as centipedes, millipedes, and spiders. It also feeds on small vertebrates, including lizards and their eggs,

Nine-banded armadillo (*Dasypus novemcinctus*). (*Photo © 2005 Pam Y. Burtt*)

and on some plant foods, such as various fruits, tubers, and subterranean fungi. This species uses scent a great deal in hunting. It places its nose close to the ground and holds its breath (up to 6 min) while digging so as to avoid inhaling dust. It is of public health importance because it is a reservoir host for *Trypanosoma cruzi*, the causative agent of Chagas' disease.

The nine-banded armadillo has a very unusual life cycle. Four young are born in a den or chamber at the end of the burrow. The young are always of the same sex and are identical quadruplets, all arising from the division of a single egg. The multiple-birth condition, or polyembryony, is unique among mammals. The young are well developed at birth and are weaned at about 8 weeks of age.

The genus *Dasypus*, the long-nosed armadillos, includes six other species in addition to the nine-banded armadillo. The genera *Priodontes* and *Chlamyphorus* contain one species each. Species in other genera including only one species are *Calyptophractus retusus*, the greater fairy armadillo; *Euphractus sexcinctus*, the six-banded armadillo; and *Zaedyus pichiy*, the pichi. Other genera are *Cabassous*, the naked-tailed armadillos, with four species; *Chaetophractus*, the hairy armadillos, with three species; and *Tolypeutes*, the three-banded armadillos, with two species. *See* EDENTATA; TRYPANOSOMIASIS.

John O. Whitaker, Jr.

Bibliography. A. L. Gardner, Order Cingulata, in *Mammals of the World*, ed. by D. Wilson and D. A. Reeder, in press, 2005; G. J. Galbreath, Armadillo, pp. 71–79 in *Wild Mammals of North America*, ed. by J. A. Chapman and G. A. Feldhamer, Johns Hopkins University Press, Baltimore, 1982; G. G. Montgomery, *The Evolution and Ecology of Armadillos, Sloths and V. vermilinguas*, Smithsonian Institution Press, Washington, DC, 1985; R. M. Nowak, *Walker's Mammals of the World*, 5th ed., vol. 1, Johns Hopkins University Press, Baltimore, 1991; J. O. Whitaker, Jr., and W. J. Hamilton, Jr., *Mammals of the Eastern United States*, Cornell University Press, Ithaca, NY, 1998; D. E. Wilson and D. M. Reeder, *Mammal Species of the World*, 3d ed., Johns Hopkins University Press, Baltimore, 2005.

Armature

That part of an electric rotating machine which includes the main current-carrying winding. The armature winding is the winding in which the electromotive force (emf) produced by magnetic flux rotation is induced. In electric motors this emf is known as the counterelectromotive force.

On machines with commutators, the armature is normally the rotating member (**Fig. 1**). On most ac machines, the armature (**Fig. 2**) is the stationary member and is called the stator. The core of the armature is generally constructed of steel or soft iron to provide a good magnetic path, and is usually laminated to reduce eddy currents. The armature windings are placed in slots on the surface of the core. On machines with commutators, the armature winding is connected to the commutator bars. On ac machines with stationary armatures, the armature winding is connected directly to the line. *See* COMMUTATOR; CORE LOSS; ELECTRIC ROTATING MACHINERY; WINDINGS IN ELECTRIC MACHINERY.

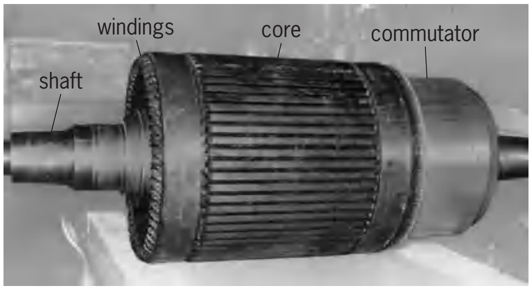


Fig. 1. A rotor armature of of a direct-current generator or motor. (*General Electric*)

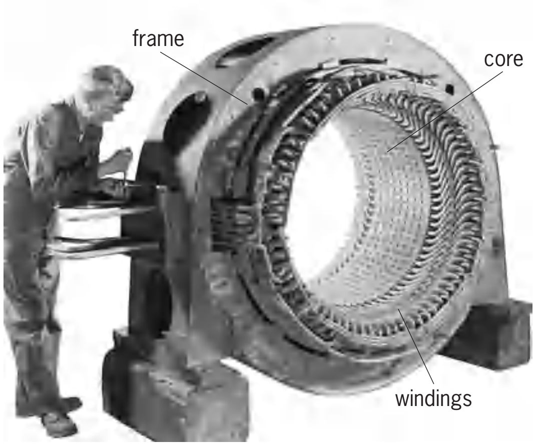


Fig. 2. A stator armature of an ac induction motor. (*Allis Chalmers*)

Arthur R. Eckels

Bibliography. S. A. Nasar (ed.), *Handbook of Electric Machines*, McGraw-Hill, 1987; S. A. Nasar and L. E. Unnewehr, *Electromechanics and Electric Machines*, 2d ed., Wiley, New York, 1983; National

Electrical Manufacturer's Association, Pub. no. MG1-1972, *Motors and Generators*, New York, 1972; M. S. Sarma, *Electric Machines: Steady State Theory and Dynamic Performance*, 2d ed., Brooks-Cole, Monterey, 1994.

Armature reaction

The effects of the magnetomotive force (mmf) of the armature on the air-gap field of direct-current (dc) and synchronous alternating-current (ac) machines. Since armature current varies directly with the electrical or mechanical load on the machine, armature-reaction effects are load-dependent. In dc machines, the armature reaction causes a distorted flux-density distribution in the air gap. **Figure 1** shows the magnetic field produced in the air gap of a two-pole dc motor or generator by (*a*) the field mmf, (*b*) the armature mmf,and (*c*) both armature and field mmf acting together. **Figure 2** shows the distorted flux distribution in the air gap, curves *a, b*, and *c* corresponding to the above three cases. Due to the saturation of the armature teeth, the flux density is decreased by a greater amount under pole tip *e* than it is increased under pole tip *f*, and therefore the armature reaction produces a demagnetizing effect, and the generated voltage or countervoltage will be reduced when the armature is loaded. In a generator this degrades the voltage regulation. In a motor it tends to increase the speed and may cause instability.

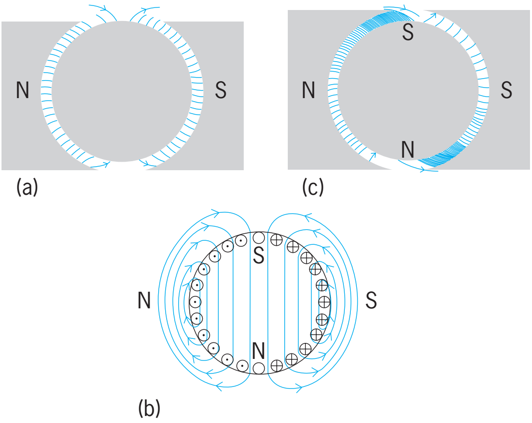


Fig. 1. **Magnetic fields in air gap of two-pole machine. (*a*) Main field. (*b*) Armature field. (*c*) Load conditions.**

For dc machines subject to heavy overloads, rapidly reversing loads, or operation with a weak field, the resultant flux-distribution distortion by excessive armature reaction will cause nonuniform distribution of voltage between commutator segments, and may result in flashover between commutator segments. A pole-face (or compensating) winding, embedded in slots in the pole face and excited by armature current, is provided to neutralize the armature mmf under the pole faces.

In polyphase synchronous machines, balanced polyphase load currents flowing in the armature winding produce a revolving mmf. This mmf rotates at exactly the speed of the rotating field poles. Hence the mmf of armature reaction reacts with the mmf of the field winding on the rotor poles to produce a displaced resultant magnetic field. The angle of the armature reaction mmf with respect to the pole axis depends upon the power factor and kilovolt-amperes of the load.

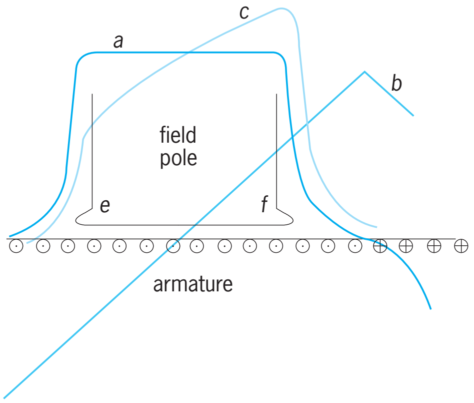


Fig. 2. **Distorted flux distribution in air gap.**

In cylindrical-rotor machines, the position and magnitude of the peak air-gap flux density are determined by the resultant mmf field, and armature reaction in such machines therefore causes a change in the amplitude and phase of the generated voltage without substantially altering its waveform. In salient-pole machines, the armature reaction tends to concentrate the flux at one side of the pole face (**Fig. 3**), and some voltage waveform distortion results.

In analyzing the effects of armature reaction in salient-pole machines, the armature mmf is often resolved into two components. That component which acts along the pole axis is known as direct-axis armature reaction and has a direct magnetizing or demagnetizing effect. That component located at right angles to the pole axis is known as quadrature-axis

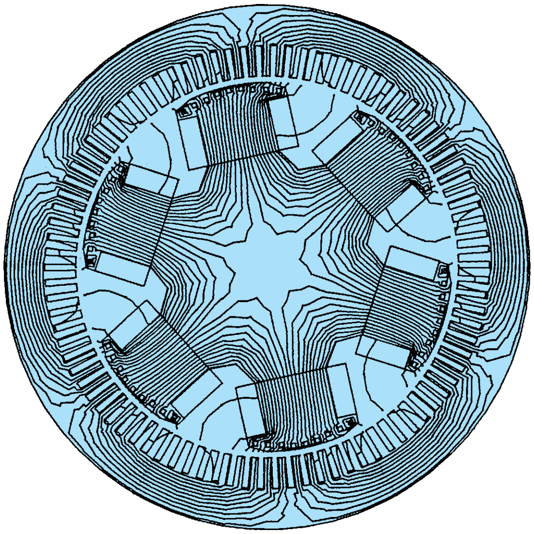


Fig. 3. **Computer plot showing skewing of pole flux effected by armature reaction in a salient-pole synchronous machine. (*E. G. Strangas, Michigan State University*)**

armature reaction. It is analogous to the armature reaction effect in modern dc machines. *See* ARMATURE; ELECTRIC ROTATING MACHINERY; MAGNETOMOTIVE FORCE. George McPherson, Jr.

Army armament

The weapons, equipment, and supplies that permit lethal or nonlethal devices to strike their targets. Armaments can be classified as field artillery, individual and crew-served weapons (infantry), armor and antiarmor, antiaircraft, helicopter armaments, and mines and countermines.

Supplies must be available to all army units to support armaments. For example, a battery of six howitzers can expend 1.5 tons (1360 kg) of projectiles, propellant, and fuzes during 1 min of firing. A logistics system which includes aircraft, ships, and a wide variety of ground vehicles and material-handling equipment (such as forklifts and cranes) is designed to maintain all forces with supplies. General-purpose ground vehicles used for supply include 0.5- and 0.75-ton (450-and 680-kg) trucks and 6-ton (5400-kg) trailers. Specialized cargo-carrying vehicles are used to accompany and supply ammunition to self-propelled artillery howitzers. For example, supporting the 155-mm self-propelled howitzer fleet is the M992A2 field artillery ammunition supply vehicle (FAASV), which has a loaded weight of 58,500 lb (26,500 kg) and a built-in automated conveyor to deliver fuzed artillery ammunition.

Field Artillery

Artillery can deliver highly lethal warheads to ranges well beyond the reach of infantry weapons. Artillery includes rockets, missiles, and self-propelled and towed howitzers. Artillery fire support is designed to meet adversarial forces by attacks on personnel and on medium and hard targets such as troop carriers and tanks. This mission is conducted at extended ranges with high-explosive projectiles, projectiles with shaped-charge submunitions, and the Copperhead guided projectile. Artillery can also channel, delay, and destroy oncoming forces by delivery of antipersonnel and antitank mines. Future goals include the use of smart munitions; increased range, lethality, and firing rates; new propulsion techniques; cartridge course correction in flight; and systems which permit much greater battlefield versatility.

Cannons (howitzers). There are two calibers of cannons in the U.S. Army's inventory, the 105-mm and the 155-mm. The army-specific name for these cannons are howitzers, and they are the primary armaments of field artillery units. They can be either towed or self-propelled systems.

105-mm howitzers. These are the M102 weapon and the M119 towed weapon. They are intended for rapid deployment of light infantry and air assault divisions. The M102 weighs approximately 5242 lb (2380 kg), and has a range of 7.0 mi (11.3 km) with the M1 high explosive projectile and 9.2 mi (14.8 km) with the rocket-assist M548 projectile. The M119A1 howitzer is derived from the British L119 light gun and is designated to replace M102 howitzers. The weapon weighs approximately 4270 lb (1816 kg) and fires the standard 105-mm artillery cartridges. Its range is 8.9 mi (14.3 km) with the M760 high-explosive projectile and 12.1 mi (19.5 km) through rocket assist.

155-mm howitzers. These consist primarily of the self-propelled, fully tracked M109 series weapon and the towed M198 howitzers.

The M109 series howitzer is the main fire support artillery system of the armored and mechanized units. The most advanced version is the M109A6 Paladin (**Fig. 1***a*), which is capable of semiautonomous operations for more rapid gun positioning and firing. Paladin has both a chassis lined with high-strength synthetic fiber material (aramid) and a pressurized compartment to ensure survivability of the crew against ballistic and nuclear, biological, and chemical threats. The combat-loaded weapon weighs 63,500 lb (28,900 kg). Its M284 cannon has a maximum range of 11 mi (18 km) with the M107 high

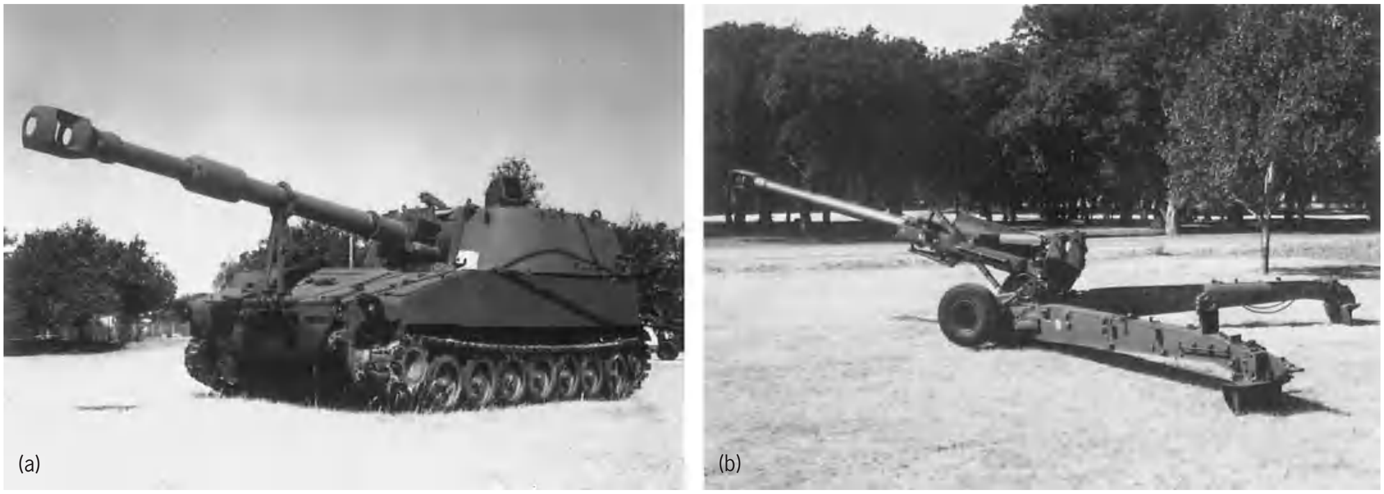


Fig. 1. 155-mm howitzers. (*a*) M109A6 Paladin self-propelled howitzer (*BMY Combat Systems*). (*b*) M198 towed howitzer (*U.S. Army*).

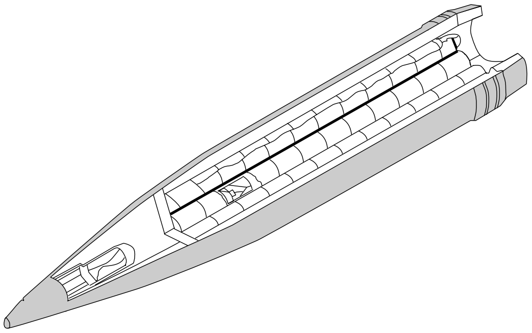


Fig. 2. M483A1 155-mm cargo-carrying projectile. (*U.S. Army*)

explosive and 19 mi (30 km) with the M549A1 rocket-assist projectile. The M483A1 projectile (**Fig. 2**), which was used effectively against Iraqi forces in the Persian Gulf War, contains 88 dual-purpose, anti-light-armor/antipersonnel M42/M46 submunitions and can be fired at a range of 10.9 mi (17.5 km). The M864 projectile contains 72 submunitions and a baseburner motor which extends the range to 17 mi (27 km). The 155-mm howitzer also fires the M712 shaped-charge Copperhead guided projectile to a range of 10 mi (16 km). The M712 uses semiactive laser guidance for the terminal flight and homing on target.

Crusader is a 155-mm self-propelled artillery system being designed to replace the M109A6 Paladin system in 2007. The Crusader will offer the most advanced technologies in automation, electronics, and software, digitization, mobility, and lethality. Through use of autonomous operations, Crusader will have unprecedented positioning and firing rate. Its XM297 cannon has a maximum range of 25 mi (40 km) with the M549A1 projectile and nearly 31 mi (50 km) with the developmental XM982 projectile.

The towed M198, air-transportable weapon (Fig. 1*b*) replaced the M114 series. It weighs 15,000 lb (6810 kg). Its M199 cannon has a maximum range of 19 mi (30 km) with the M549A1 rocket-assist projectile and 14 mi (22 km) with the M107 high-explosive projectile.

The lightweight 155-mm towed howitzer (XM777) system (**Fig. 3**) replaces the M198. The XM777 through the use of high-strength titanium maintains strength but at a significantly reduced weight (less than 9000 lb or 4100 kg). It is designed to meet or exceed the capabilities of the M198. Because it is lightweight, it has improved mobility and terrain coverage. The XM777 is also the first towed artillery system to use digitization for improved lethality and battlefield dominance.

Rockets. The multiple launch rocket system (MLRS; **Fig. 4**) is a free-flight artillery rocket weapon used primarily as a counterfire weapon with a range of up to 25 mi (40 km). It consists of a 12-rocket launcher, pod-mounted on a tracked vehicle. The warhead contains 644 of M77 dual-purpose, shaped-charge, antipersonnel, antimateriel bomblets similar to those used in the M483A1 projectile. The MLRS was used with great effectiveness against Iraqi forces in the Persian Gulf War.

The 2.75-in. rocket systems are a family of "fire and forget" rockets used primarily for air-to-ground applications by fixed and rotary wing aircraft. The rockets are fired from either a 7- or 19-tube launcher for antipersonnel, antimaterial, and antiarmor purposes. Screening smoke and both visible and infrared illumination are also available for supportive missions.

Individual and Crew-Served Weapons (Infantry)

Infantry armament is composed largely of line-of-sight, direct-fire weapons and indirect-fire mortar weapons used by nonmechanized light infantry (60-mm and 81-mm mortars) and mechanized infantry and armor divisions (120-mm mortars)

Small-caliber weapons. Many small-caliber weapons have been introduced as replacements or additions to the inventory.

M9 and M10 pistols. These 9-mm pistols will replace the M1911A1 .45-caliber pistol and various .38-caliber revolvers. The pistols are semiautomatic,

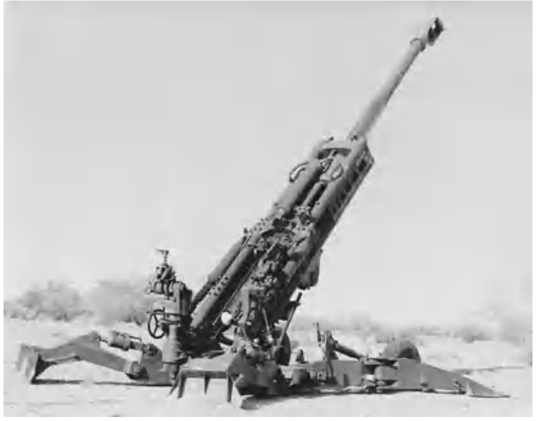


Fig. 3. XM777 lightweight 155-mm towed howitzer. (*U.S. Army*)

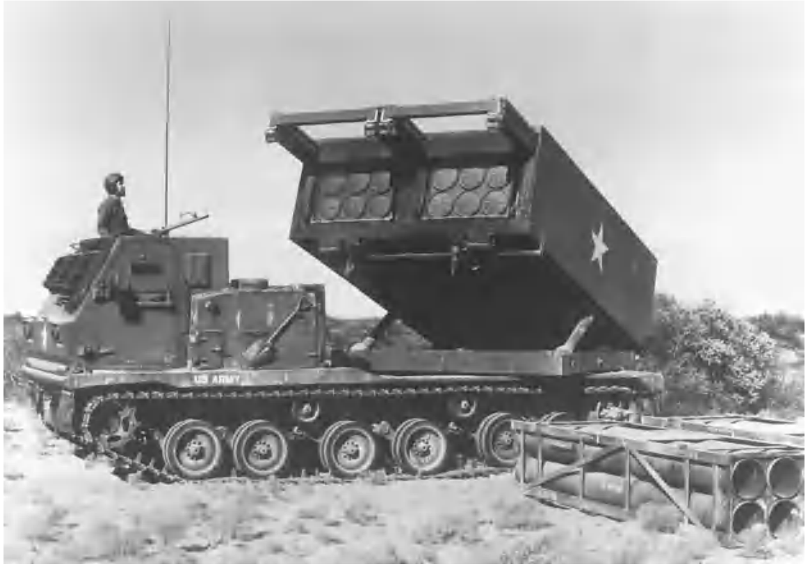


Fig. 4. Multiple-launch rocket system (MLRS). (*Vought Corp.*)

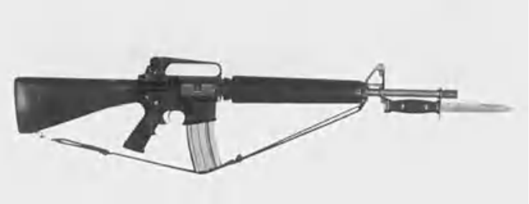


Fig. 5. M16A2 rifle. (*U.S. Army*)

magazine-fed, recoil-operated, double-action handguns with a 15-round magazine. They are chambered to accept M882 ball ammunition and the standard 9-mm NATO ammunition. The M9 and M10 weigh approximately 2.52 lb (1.145 kg).

M16A2 rifle. The rifle (**Fig. 5**) is an improved version of the 5.56-mm-caliber M16A1 rifle. It includes a muzzle brake compensator to reduce muzzle jump, a burst control to limit automatic fire to three rounds per trigger pull, and a barrel designed to accommodate NATO standard 5.56-mm ammunition. The rifle has a maximum effective range of 2900 ft (880 m) against area targets. It fires M855 ball and M196 tracer ammunition.

M4 carbine. The operating characteristics of the M4 5.56-mm carbine are similar to those of the M16A2 rifle; the same ammunition is used. However, the carbine weighs 7.2 lb (3.5 kg) with a loaded 30-round magazine versus 8.8 lb (4 kg) for the M16A2. The M4 is also almost 25 cm (10 in.) shorter. The maximum effective range against area targets is 2600 ft (800 m).

M24 sniper weapon system. This replacement for the current M21 sniper rifle is a bolt-action, 7.62-mm-caliber rifle with a detachable 10-power optical sight and a five-round magazine containing M118 special ball cartridges. The maximum effective range is 2600 ft (800 m).

M203 grenade launcher. This is a 40-mm, pump-action, single-shot weapon that attaches under the barrel of the M16 series rifles. It weighs 3.6 lb (1.6 kg) loaded, has a maximum range of 1150 ft (350 m), and fires practice, buckshot, dual-purpose, pyrotechnic, and spotting ammunition.

Mark 19 Mod 3 grenade machine gun. This weapon is an advanced version of the Mark 19 gun. It is an air-cooled, automatic machine gun that fires M430 40-mm dual-purpose antipersonnel, anti-light-materiel cartridges at a cyclic rate of 325–375 rounds per minute to a range of 1.4 mi (2.2 km) from a tripod- or vehicle-mounted position.

M60 7.62-mm machine gun. This is a general-purpose widely used machine gun mounted on a bipod (**Fig. 6**), on a tripod, or in a vehicle. It weighs 23 lb (10.4 kg) empty and fires 100 rounds per minute at a sustained rate to a range of 2.3 mi (3.7 km). An M60E3 version has been designed to be lighter and more versatile.

M249 squad automatic weapon. This is a one-person portable machine gun that fires the 5.56-mm M855 ball and M856 tracer ammunition. It is used to replace the M16 for sustained and accurate fire to approximately 3300 ft (1000 m) at a rate of 800 rounds per minute.

M231 5.56-mm firing port machine gun. This is an air-cooled, gas-operated, magazine-fed system mounted in the Bradley fighting vehicle (discussed below). It weighs 9.5 lb (4.3 kg) with a 30-round magazine, has a sustained rate of fire of 50–60 rounds per minute, and an effective range of 1000 ft (300 m).

M242 25-mm chain gun. This is the main armament system for the Bradley. It is externally powered by a 1.5-hp (1.1-kW) motor, has a dual feed, and can fire single shots and multiple shots at up to 500 rounds per minute of high-explosive M792 and armor-piercing M791 and M919 ammunition. It has an effective range of 6500–10,000 ft (2000–3000 m).

M230 30-mm Hughes chain gun. This is a major turret-mounted weapon of the AH-64 Apache helicopter (discussed below) used in air-to-air and air-to-ground roles. It weighs 118 lb (53.6 kg) and fires the M789 high-explosive, dual-purpose antitank and antipersonnel rounds at a rate of 625 rounds per minute.

Next-generation weapons. Improvements in hit probability, combat load, and target acquisition of small-caliber weapons over those of the current systems are being pursued. The long-term objective is a family of small arms that may feature composite materials, bursting ammunition, laser fire control, night-vision devices, and microelectronics. The Objective Individual Combat Weapon and the Objective Crew-Served Weapon are slated to be the next-generation small-caliber systems.

Mortars. The U.S. Army uses three mortar weapon systems: the 60-mm M224 lightweight company mortar, the 81-mm improved M252 mortar, and the 120-mm M120 mortar. The 107-mm M30 mortar has been replaced by the 120-mm M120 mortar (towed and ground mounted) and the 120-mm M121 mortar (carrier mounted). Each type of mortar fires a family of projectiles, including high-explosive, smoke, illuminating, and full-range and short-range practice cartridges. The main direction of mortar development has been to increase weapon and cartridge performance (through improvements in lethality, smoke obscuration, white-light and infrared-light illumination, range, reliability, and rate of fire and reductions in weight and unit cost; to improve fuze versatility through use of a multioption fuze; and to use waterproof combustible cases for propulsion charges. There has been significant improvement in mortar fire control through the use of the M23 Mortar Ballistic Computer, and initiatives are under way to enable mortars to be active participants in the digital

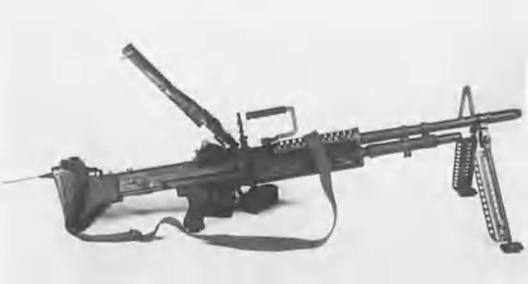


Fig. 6. M60 7.62-mm machine gun. (*U.S. Army*)

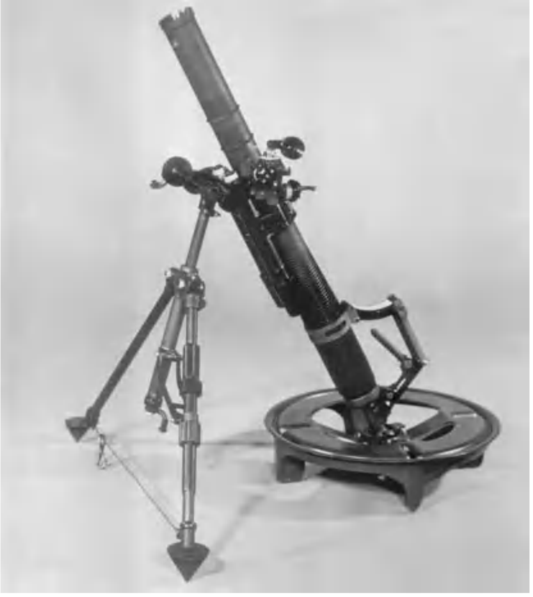


Fig. 7. M224 60-mm lightweight company mortar system. (*U.S. Army*)

battlefield via the XM30 Improved Mortar Ballistic Computer and the XM35 Digital Mortar Fire Control System. A program is also under way to make mortar cartridges safer for the user through insensitive munition design.

60-mm M224 mortar. This weapon is a lightweight, smoothbore system which can be drop- or trigger-fired (**Fig. 7**). It fires the M720 or M888 high-explosive cartridges with the M734 multioption fuze and the M935 point detonation fuze respectively to a range of 2.2 mi (3.5 km). The M224 also uses the M721 illuminating cartridge, the M722 white phosphorus marking cartridge, and the M767 60-mm short-range practice cartridge. The M767 infrared illuminating cartridge, the M720A1 high-explosive cartridge, the M768 high-explosive cartridge, and the M769 full-range practice cartridge are in development.

81-mm M252 mortar. This is the basic United Kingdom L16 smoothbore mortar with a U.S.-developed blast-attenuating device and an improved M177 bipod mount. The weapon weighs 93.25 lb (42.3 kg). It fires the M821 and M821A1 high-explosive cartridges with the M734 multioption fuze to a range of 3.5 mi (5.7 km). The M853A1 illumination, the M819 red phosphorus smoke, the M889 and M889A1 high-explosive cartridges with the M935 point detonation fuze, the M879 full-range practice cartridge, and the M880 short-range practice cartridge are standard rounds for the M252 system. The XM816 infrared illuminating cartridge is in development.

120-mm M120 and M121 mortars. This mortar weapon has replaced the 107-mm M30 mortar. It is an improved version of the Soltam K-60 Tampella mortar and is fielded in two configurations. The M120 towed version is transported on a two-wheel carriage. The M121 carrier-mounted version is mounted on the M1064 mortar carrier. The M121 weapon can be fired from the carrier or removed and fired from the ground by using an auxiliary baseplate and mount. The towed mortar weighs 318 lb (144 kg) in firing position and 642 lb (291 kg) in towed position. It fires a full family of ammunition within the range from 560 ft to 4.5 mi (170 m to 7.2 km). An improved high-energy cartridge (M933/M934), an illuminating cartridge (M930), and a smoke cartridge (M929) have been developed for use with this weapon. The 81-mm M303 insert allows 120-mm crews to live-fire-train with the older 81-mm M300 series ammunition and the 81-mm M880 short-range practice cartridge with a significant reduction in training costs. Cartridges in development include the XM983 infrared illuminating cartridges. The XM984 extended-range DPICM submunition round, and the Precision-Guided Mortar Munition (PGMM).

Shoulder-launched munitions. The primary shoulder-launched infantry munitions consist of the AT4 recoilless rifle and the Shoulder-Launched Multipurpose Assault Weapon-Disposable (SMAW-D) Bunker Defeat Munition.

M136 AT4 recoilless rifle. Infantry forces use this weapon primarily to defeat light-armored targets. It is issued as a complete system containing an 84-mm projectile with a shaped-charge warhead and a disposable composite launcher. The system weighs 14.8 lb (6.7 kg), is 39 in. (1.0 m) long, and has an effective range of between 800 and 1600 ft (250 and 500 m). It replaces the M72 LAW (light antitank weapon). The system has been employed in several conflicts, including the Persian Gulf War.

XM141 SMAW-D bunker defeat munition. Infantry forces use the SMAW-D (**Fig. 8**) to defeat bunkers, walls, and very light armor targets. It is issued as a complete system and includes a dual-mode warhead and disposable composite launcher. The system weighs 15.7 lb (7.1 kg) and is 32 in. (0.8 m) long. The system has an effective range of 800–1600 ft (250–500 m).

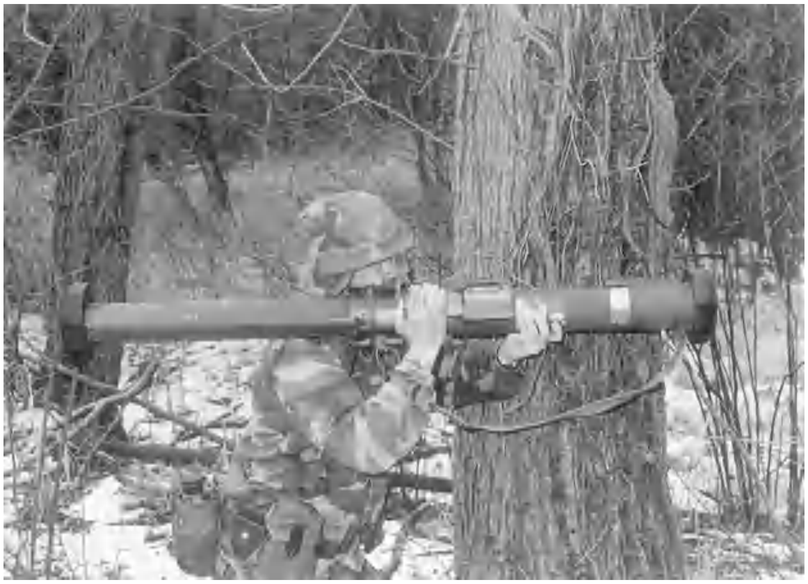


Fig. 8. XM141 SMAW-D bunker defeat munition. (*U.S. Army*)

Armor and Antiarmor

The primary armored weapon systems consist of the M1 and M1A1 main battle Abrams tanks; the more lightly armored Bradley M2 infantry and M3 cavalry fighting vehicles; and the M113 family of armored personnel carriers. Major thrusts in armament development have resulted in conversion of the rifled 105-mm cannon in the M1 tank to the 120-mm smoothbore cannon in the M1A1 tank, and the utilization of tungsten and depleted-uranium long-rod penetrators for maximum antiarmor effectiveness. Increased protection against penetration by kinetic-energy and shaped-charge ammunition has been provided by the use of laminate Chobham-type armor and reactive armor.

Main battle tanks. The main purpose of the Abrams series of main battle tanks is to provide mobility, firepower, and shock effect to overwhelm and destroy enemy forces on the battlefield. The Abrams powerful main gun, special armor, and 1500-hp (1.1-MW) engine make it particularly suitable for both attacking and defending against enemy armor forces. The M1 tank (**Fig. 9**) is equipped with a 105-mm M68 main gun, a M219/M240 coaxial weapon which fires 7.62-mm ammunition, another M240 7.62-mm loader's machine gun, and a commander's M85 .50-caliber machine gun. The M1A1 tank's main gun is the 120-mm M256 smoothbore cannon; small-caliber weapons carried are identical to those on the M1 tank. M1A1 features include increased armor protection, suspension improvements, and a protection system for nuclear, biological, and chemical attacks. These afford the tank and the soldiers greater survivability in a variety of contaminated environments. Antitank ammunition available for use in the 105-mm cannon includes the M735, M774, and M833 kinetic-energy, heavy-metal, armor-piercing, fin-stabilized, discarding-sabot projectiles; and the M456 series of high-explosive, shaped-charge projectiles. Projectiles developed for use in the 120-mm cannon include the M829A2 high-velocity kinetic-energy round and the M830A2 shaped-charge round, which has high antiarmor effectiveness and has an added proximity switch selection that provides the tank with a self-defense capability against helicopters. The propellant for the 120-mm gun is contained in combustible cartridge cases; metal cartridge cases are used in the 105-mm weapons. Both cannons fire kinetic-energy and shaped-charge projectiles.

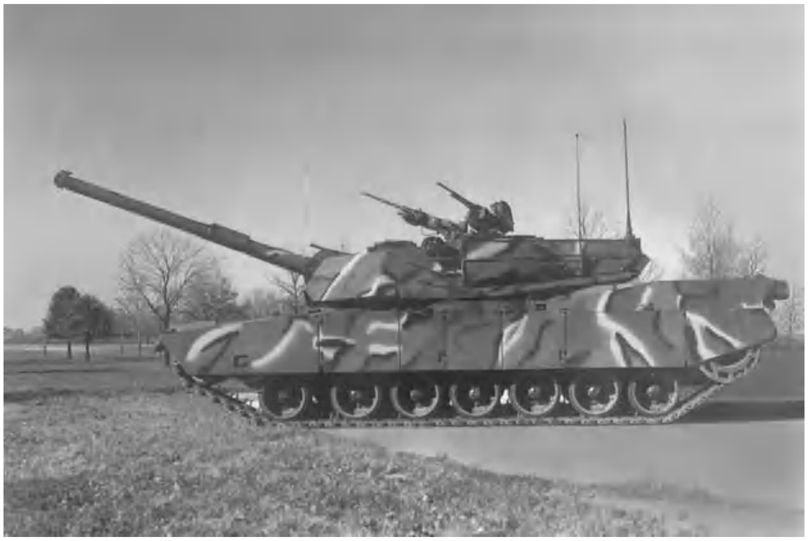


Fig. 9. M1 Abrams battle tank. (*U.S. Army*)

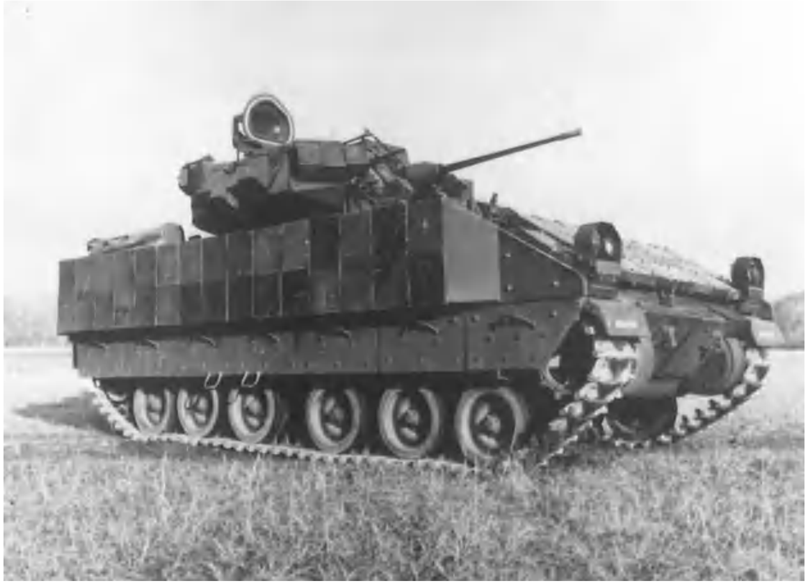


Fig. 10. M2 Bradley infantry fighting vehicle with M242 25-mm turret-mounted chain gun. (*U.S. Army*)

Infantry fighting vehicles. The Bradley M2 infantry fighting vehicle (**Fig. 10**) transports and supports infantry soldiers on the battlefield, while the M3 cavalry fighting is used to conduct reconnaissance and surveillance operations. Both vehicles have identical armament, which consists of the 25-mm M242 turret-mounted chain gun, a 7.62-mm M240C coaxial machine gun, and six M231 firing port weapons. The vehicles also carry a dual-tube TOW missile launcher (discussed below). The M3 variant carries more TOW missiles than does the M2 variant. This space is used on the M2 to carry more infantry soldiers.

M113 armored personnel carrier. This versatile vehicle (**Fig. 11**) exists in many variations to accomplish different missions. Normal armament is the .50-caliber machine gun. The vehicle may also serve as the carrier for 81-mm, 107-mm, or 120-mm mortar weapons, the M163 20-mm Vulcan air-defense gun, TOW launchers, and the Mark 19 Mod 3 machine gun.

Appliqué armor. These appliqués consist of packages that may contain a series of plates and an insensitive explosive which reacts with incoming shaped-charge jets to greatly reduce their armor penetration effectiveness. The armor tiles are flat metal boxes of various dimensions that are mounted on the front, rear, top, and sides of the vehicle. Appliqué armor has been designed for the M2A2 and M3A2 Bradley fighting vehicles and for the M60A3 tanks.

TOW. The tube-launched, optically tracked, wire-guided missile (TOW; **Fig. 12**) is a heavy antiarmor assault weapon that can be fired from a ground tripod, from helicopters, and from armored and unarmored vehicles. The missile is 6 in. (15 cm) in diameter and contains a variety of shaped-charge warheads. It has a range of up to approximately 2.5 mi (4.0 km). Following initial launch to a safe distance from the gunner, the rocket motor accelerates the missile to the tactical velocity. The initial TOW system was fielded in 1970, and since then has continually been upgraded.

Air Defense

There are a number of air defense weapons. These include the Avenger Air Defense System, the handheld Stinger rocket, and the Patriot system.

Stinger. This is a human-portable, shoulder-fired, two-stage, infrared homing missile, designed to attack and destroy at short range any low-altitude, fixed- and rotary-wing aircraft with a high-explosive fragmentation warhead. The Stinger is to be replaced by the Linebacker Air Defense System, which is in production and being fielded in the Armor and Mechanized Divisions (as of 2000).

Avenger Air Defense System. This air defense system is mounted on a Humvee (high-mobility multipurpose wheeled vehicles; HMMWV). It is lightweight and transportable and designed to counter enemy cruise missiles, crewless aerial vehicles, and low-flying fixed-wing or rotary aircraft.

Patriot. This system fires MIM-104 surface-to-air missiles (**Fig. 13**) at attacking aircraft and missiles at ranges up to 37.3 mi (62.3 km), providing medium- and high-altitude defense. It was used in the Persian Gulf War to defend against Iraqi-launched SCUD missiles. The system consists of up to eight M901 mobile launchers, each of which contains four ready-to-fire missiles. The launcher is supported by major units including a power plant, an MPQ-53 phased-array radar, a communications center, and the crewed control station. The radar searches, detects the incoming target, and provides the electronic illumination of the target, which signals the target's location to the intercepting missile's semiactive radar homing unit. The MIM-104 missile has a 200-lb (90-kg) high-explosive warhead initiated by a radar proximity fuze.

Helicopter Armaments

The major combat aircraft used in support of ground operations are the AH-64 Apache and the AH-1 Cobra attack helicopters. Both helicopters are highly mobile and are capable of destroying moving armored columns and other point and area targets on the modern battlefield.

The AH-64 Apache is armed with the Hellfire laser-guided antitank missile system, 2.75-in. rockets, and the M230 30-mm turret-mounted chain gun. The M230 gun is used in air-to-air and air-to-ground roles.

The AH-1 Cobra helicopter is armed with point and area weapon systems: the TOW missiles, 2.75-in. rockets, and the 20-mm cannon.

The RAH-66 Comanche (formerly the LHX, light helicopter experimental) is under development to perform both reconnaissance and attack missions against airborne and ground targets. Each helicopter is equipped with three-barrel 20-mm cannon and can carry up to 14 Hellfire or 28 Stinger missiles, or 56 2.75-in. rockets. *See* HELICOPTER.

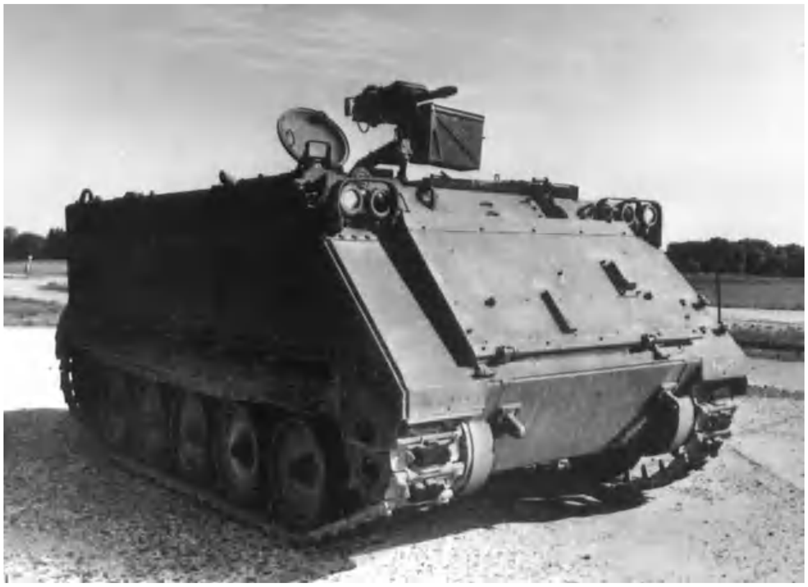


Fig. 11. M113 armored personnel carrier. (*U.S. Army*)

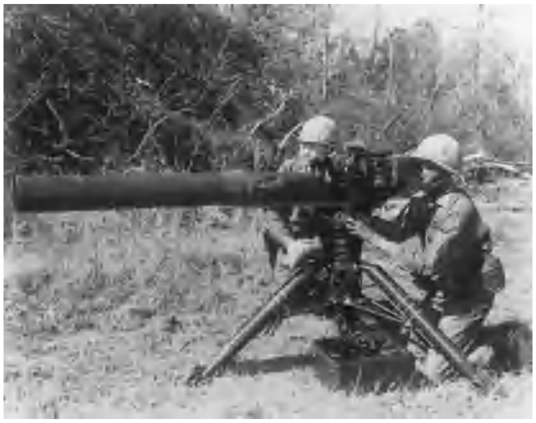


Fig. 12. TOW 2 (tube-launched, optically tracked, wire-guided) missile (tripod-mounted). (*U.S. Army*)

Mines

The U.S. Army places great emphasis on the use of mines in modern warfare. A family of mines and modes of delivery have been developed, which include scatterable antipersonnel and antitank mines (family of scatterable mines; FASCAM) that have high utility because they can be emplaced by artillery, aircraft, and ground vehicles at close-in and extended ranges. They are autonomous after deployment, are effective over a time period, and self-destruct to permit occupation of the area by friendly forces.

Artillery-delivered mines. The two 155-mm artillery projectiles which carry mines are the antipersonnel M692/M731 ADAM (area denial artillery munition) and the antivehicular M718/M741 RAAM (remote antiarmor mine) rounds. The ADAM contains 36 wedge-shaped mines, and the RAAM has 9 mines. The ADAM mine is tripwire-fired and produces a pop-up fragmenting warhead; a magnetic sensor, which detects the vehicle signature, activates the RAAM mine. The 155-mm howitzer can deploy both mine systems at ranges up to 11 mi (18 km).

Ground-emplaced scatterable mines. The vehicle-mounted GEMSS (ground-emplaced mine scattering system; **Fig. 14**) and the FLIPPER mine system emplace the magnetic-influenced antiarmor M75 and the tripline-activated antipersonnel M74 mines. They are used to rapidly establish a defensive position for friendly forces. The GEMSS consists of the M128 mine dispenser mounted on the M794 flat-bed chassis. The dispenser contains two independently operated magazines holding 800 mines and a mine launcher which can be programmed to control the density and width of the minefield and to specify either a long or a short self-destruct time. The FLIPPER is used for slower, low-volume operation. It consists of the manually loaded M138 mine dispenser which can be mounted on tracked or wheeled vehicles.

Modular pack mine system (MOPMS). This system is a four-person portable M131 mine dispenser which contains 17 antitank M78 and 4 antipersonnel M79 mines. The dispenser can be activated to disperse mines by the M71 remote control unit at a range of up to 3300 ft (1000 m). The remote control unit can also reset the mine self-destruct time or can command-destruct the minefield.

Wide-area munition (WAM). This planned mine system is designed to fire a stand-off, top-attack, antiarmor, explosively formed penetrator warhead. The WAM has a lethal range up to 330 ft (100 m), and will also have countermeasure resistance. Initial deployment will be by troop emplacement with subsequent variants being capable of delivery by helicopter and missile.

Air-launched mines. The GATOR mine cluster weapon system consists of the CBU78B/CBU89B dispenser which carries BLU91B antiarmor and BLU92B antipersonnel mines. The system is deliverable by high-speed Air Force and Navy tactical aircraft. GATOR mines are also used in the VOLCANO dispensing system. Self-destruct times are field-selectable by a switch on the dispenser. The minefield density can be controlled by the number of dispensers used and dispenser settings as well as the speed and altitude of the plane.

VOLCANO is a versatile modular mine system for rapidly deploying GATOR mines from vehicles on the ground and from UH-60A helicopters (**Fig. 15**). The system comprises a dispenser control unit and one to four launcher racks, which snap into place on mounting-beam fixtures. Each rack contains up to 40 expendable M87 mine canisters, which in turn contain five antiarmor BLU91B and one antipersonnel BLU92B GATOR mines. As many as 960 mines can be dispensed from UH-60A helicopters and M817 dump trucks. Jamie Ruffing; Gary Pacella; James Wejsa

Bibliography. M. A. Blake (ed.), *Army*, October 1999; B. Davis et al., *TACOM-ARDEC Brochure*, 1997; C. F. Foss (ed.), *Jane's Armour and Artillery 1998–1999*, 19th ed., 1999; C. F. Foss and T. J.

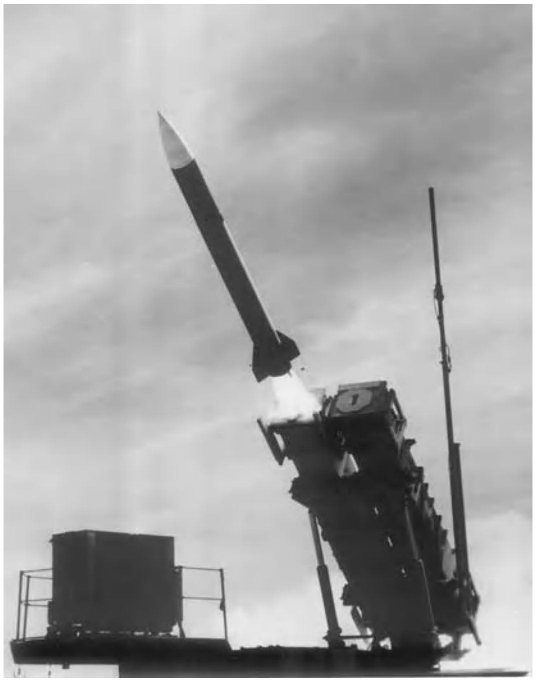


Fig. 13. Patriot system firing an MIM-104 surface-to-air missile. (*Raytheon*)

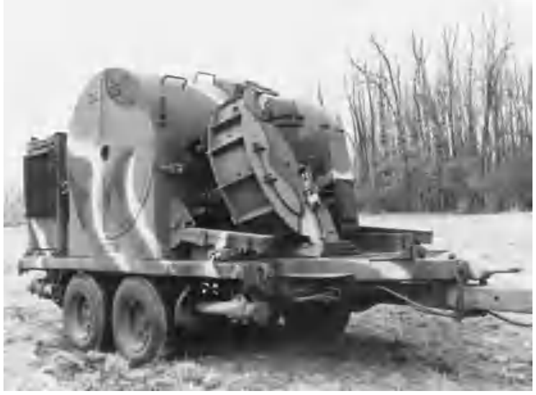


Fig. 14. Mine dispenser of GEMSS (ground-emplaced mine scattering system), mounted on flatbed chassis. (*U.S. Army*)

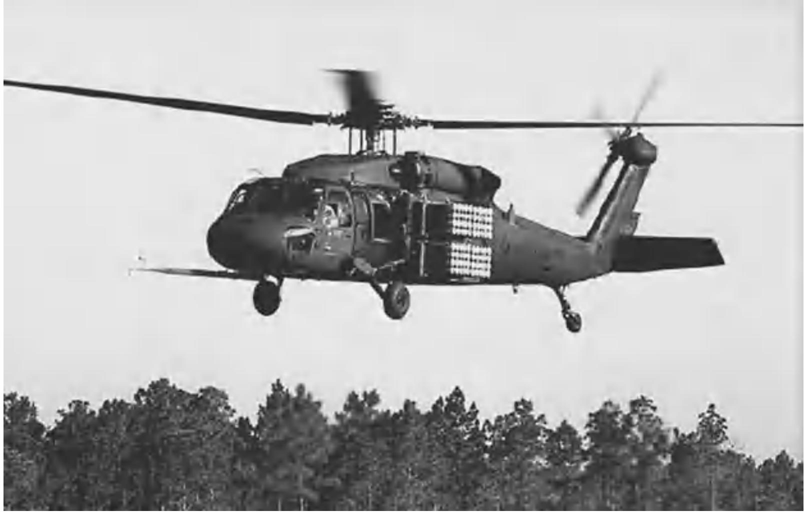


Fig. 15. Volcano modular mine-dispensing system (helicopter-mounted). (*U.S. Army*)

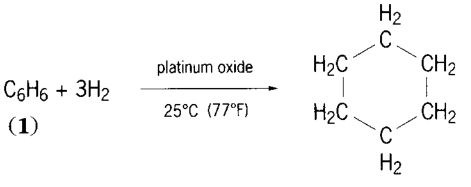
Gander (eds.), *Jane's Military Vehicles and Logistics 1998–99*, 19th ed., 1999; *Weapon Systems 1999*, U.S. Government Printing Office, Washington, DC, 1999.

Aromatic hydrocarbon

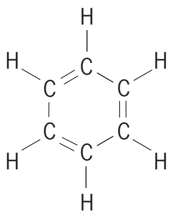
A hydrocarbon with a chemistry similar to that of benzene. Aromatic hydrocarbons are either benzenoid or nonbenzenoid. Benzenoid aromatic hydrocarbons contain one or more benzene rings and are by far the more common and the more important commercially. Nonbenzenoid aromatic hydrocarbons have carbon rings that are either smaller or larger than the six-membered benzene ring. Their importance arises mainly from a theoretical interest in understanding those structural features that impart the property of aromaticity.

Benzenoid aromatic hydrocarbons are also called arenes. Benzene itself is the prototypical arene. The properties associated with aromaticity have little to do with aroma, although the aromatic hydrocarbons were first studied in connection with naturally occurring fragrances. Instead, these compounds possess special stability; take part in certain types of reactions; and exhibit persistence of the structural integrity of aromatic rings during chemical reactions, while groups attached to those rings are chemically altered or manipulated.

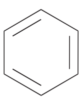
Benzene. With the molecular formula C_6H_6, benzene is highly unsaturated; it has fewer hydrogens than are required to accommodate the four valences of each carbon atom. Under appropriate conditions, benzene (**1**) can be hydrogenated to cyclohexane (**2**, C_6H_{12}), as shown in reaction (1). This result is most easily explained if the six carbon atoms of benzene are linked in a ring, with three double bonds, alternating with single bonds as in structures (**3**) or (**4**)—a structure suggested as early as 1865 by F. A. Kekulé. Quantitative study of the hydrogenation, however, showed that this structure cannot be entirely correct. For example, hydrogenation of cyclohexene (**5**) to cyclohexane is exothermic and liberates 28.6 kcal/mol of heat [reaction (2)]. Hydrogenation of 1,3-cyclohexadiene (**6**), with two double bonds, liberates approximately twice as much heat [reaction (3)]. Therefore it would be expected that benzene, if it had the 1,3,5-cyclohexatriene structure (**3**) or (**4**), would liberate about three times as much heat, or 84–86 kcal/mol. Experimentally, however, reaction (1) is exothermic by only 49.8 kcal/mol, meaning that addition of the first mol of hydrogen to benzene is actually endothermic by about 5 kcal/mol [reaction (4)]. Benzene is more stable than the 1,3,5-cyclohexatriene structure would suggest, and its stabilization energy (also called delocalization or resonance energy) is approximately 36 kcal/mol (the energy difference between the theoretical heat of hydrogenation of 1,3,5-cyclohexatriene and the experimental heat of hydrogenation of benzene, 86 − 50 = 36). Other methods for calculating the resonance energy of benzene give a similar result. *See* BENZENE.



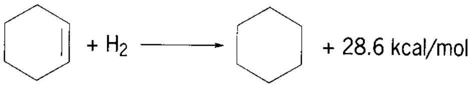
(**1**) (**2**) (1)



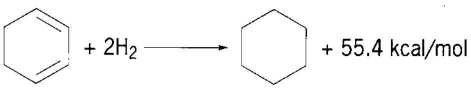
(3)



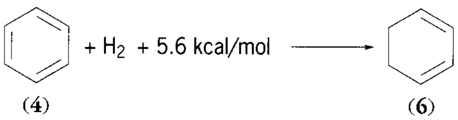
(4)



(**5**) (**2**) (2)

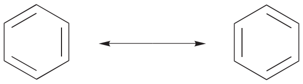


(**6**) (3)



(**4**) (**6**) (4)

Resonance. The double bonds in the benzene structure can be arranged in two ways, (**4**) and (**4′**).



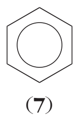
(4) (4′)

Benzene is a resonance hybrid of these two Kekulé structures; the double-headed arrow is used to signify that the benzene structure is neither (**4**) nor (**4′**), but a single structure that is a hybrid of the two. That is, the bonds between adjacent carbon atoms are neither double nor single, but of some intermediate or hybrid type.

Physical measurements bear this out. For example, x-ray determinations show that benzene is a planar, regular hexagon. Its carbon-carbon bond lengths are all identical and 0.139 nanometer longer than a normal double bond (0.134 nm) but shorter than a single bond (0.154 nm) in ethane. All the bond angles (C—C—C and C—C—H) are 120°. *See* BOND ANGLE AND DISTANCE; RESONANCE (MOLECULAR STRUCTURE).

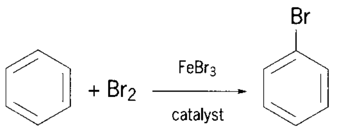
Molecular orbital view. Each carbon atom in benzene is connected to three atoms, two adjacent carbon atoms and a hydrogen atom. These three bonds lie in a single plane and use three of the carbon's four valence electrons. The fourth valence electron of each carbon is located in a *p* orbital, extending perpendicularly above and below the plane of the other three bonds. These electrons, one from each carbon atom and called π electrons, form three molecular orbitals located above and below, but parallel to, the plane of the ring.

The symbol of a hexagon with an inscribed circle (7) is often used to express the delocalized nature of the π electrons in benzene and other arenes. There is physical evidence that the π electrons circulate around the ring carbons, as implied by this formula. For example, in the nuclear magnetic resonance (NMR) spectra of arenes, the chemical shifts of arene hydrogen atoms (protons) are characteristically at lower magnetic fields than those of protons attached to carbon-carbon double bonds. This difference is due to an induced magnetic field caused by circulation of the π electrons in the molecular orbitals above and below the arene ring plane. Indeed, this chemical shift difference, due to a diamagnetic ring current, is sometimes used as evidence for aromaticity in nonbenzenoid aromatic hydrocarbons. *See* DELOCALIZATION; MOLECULAR ORBITAL THEORY.



(7)

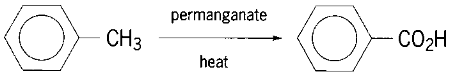
Reactions of arenes. Although benzene and other arenes react with hydrogen by addition [as in reaction (1)], they react with most other reagents by substitution, as in the formation of bromobenzene [(**8**); reaction (5)], since addition of bromine would lead to a dibromocyclohexadiene in which the aromatic stabilization would be lost.



(5)

(**8**)

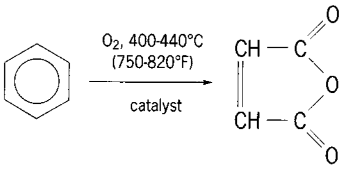
Stability of benzene ring. Whereas most unsaturated hydrocarbons are easily oxidized by, for example, potassium permanganate at room temperature, benzene is untouched by this reagent even at 100°C (212°F). The stability of the benzene ring toward oxidation is strikingly illustrated with toluene (**9**) in which the aromatic stabilization would be lost.); methylbenzene, where it is the methyl (CH_3) side chain and not the arene ring that is oxidized to yield structure (**10**), as in reaction (6). Not only is the arene ring unscathed, but the methyl group attached to the ring is more susceptible toward oxidation than would normally be the case. The methyl groups of ethane ($H_3C—CH_3$), for example, are not oxidized under similar conditions.



(6)

(**9**) (**10**)

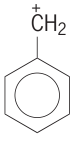
Benzene can be oxidized at high temperature to the polyester intermediate maleic anhydride (**11**), as in reaction (7), which is carried out industrially on a large scale.



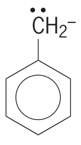
(7)

(**11**)

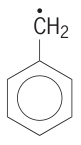
Arene rings and substituents. Arene rings stabilize reactive intermediates with a positive charge (cation), negative charge (anion), or odd electron (radical) on the atom attached to the ring, as in the benzyl cation (**12**), anion (**13**), or radical (**14**).



(**12**)

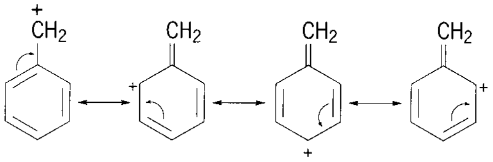


(**13**)



(**14**)

The reason is that the adjacent +, −, or · can be delocalized into the arene ring, as shown by the curved arrows for structure (**12**) in notation (8).



(8)

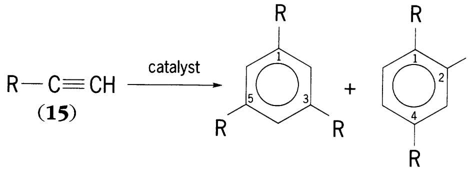
For this reason, the methyl group in toluene but not ethane is easily oxidized.

Arenes with one benzenoid ring. Besides benzene itself, several alkylbenzenes are commercially important and produced on a large scale—millions of pounds annually. Production is commonly by the cyclodehydrogenation of alkanes at high temperatures over metallic catalysts such as platinum.

Benzene, toluene, and the xylenes are added to unleaded gasoline to raise the octane number. These arenes are also essential to the petrochemical industry. Products derived from them include polyesters, polyurethanes, polystyrene, and synthetic rubber; alkylbenzenesulfonate detergents; phenol and acetone; pharmaceuticals, flavors, and perfumes; plasticizers; and many others. *See* PETROCHEMICAL.

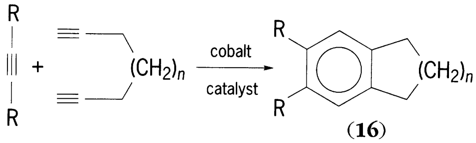
Arenes with rather specialized structures can be produced by the cyclotrimerization of acetylenes. Transition-metal catalysts are used. Monoalkylacetylenes (**15**) may give mixtures of 1,3,5- and

1,2,4-trisubstituted benzenes[reaction (9)]. With appropriate substituents, diacetylenes and monoacetylenes can be co-“trimerized” to give specifically substituted benzenes (**16**) in high yields, as in reaction (10), where R = $(CH_3)_3Si$ and $n = 0$–2.



(**15**)

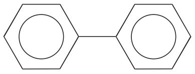
(9)



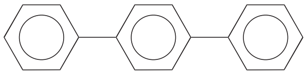
(10)

(**16**)

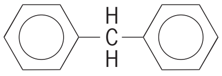
Arenes with linked benzene rings. Benzene rings can be linked, either directly as with biphenyl (**17**), terphenyls (**18**) and others, or with one or more nonaromatic carbons between, as with diphenylmethane (**19**) and triphenylmethane (**20**). *Para*-polyphenylene (**21**) is an important thermally stable polymer. Triphenylmethane provides the carbon framework for a major class of dyes and food colorings.



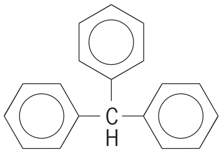
(**17**)



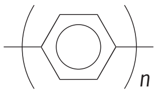
(**18**)



(**19**)

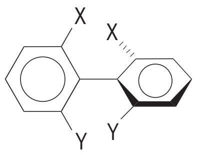


(**20**)

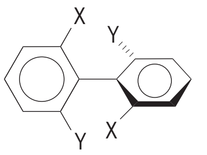


(**21**)

The two arene rings in biphenyl are not coplanar but tilted about 45° with respect to each other. Rotation around the bond that joins the two rings is facile, but if large groups are placed in the *ortho*-positions of each ring, rotation becomes difficult and leads to separable mirror-image isomers (**22a**) and (**22b**). Such molecules are chiral and differ from one another only in a left- and right-handed way. *See* STEREOCHEMISTRY.

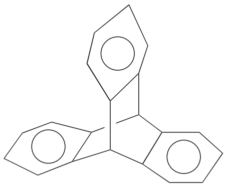


(**22a**)

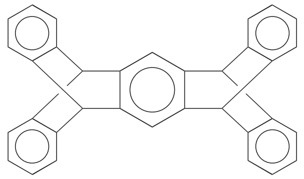


(**22b**)

Arene rings can be linked through three-dimensional frameworks. An interesting example is triptycene (**23**), which has three arene rings held rigidly at angles of 120°. The framework is useful for studying interactions of groups attached to the rings and held at fixed angles and distances from one another. Triptycene is the first of a general class known as iptycenes, another of which is pentiptycene (**24**), with five fixed arene rings.

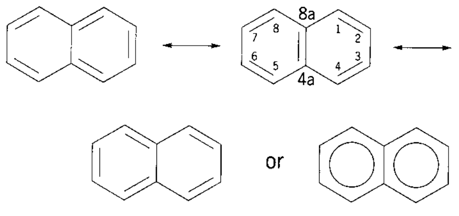


(**23**)



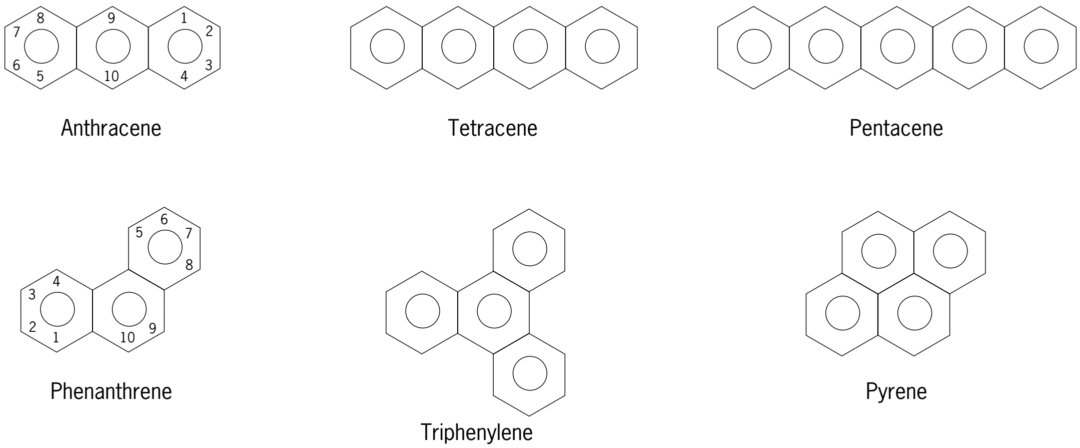
(**24**)

Fused rings. Arenes with fused rings are also known as polynuclear aromatic hydrocarbons. Rings are said to be fused when they share two carbon atoms. The simplest example is naphthalene, a colorless crystalline compound found in coal tar, best known as a moth repellent, whose structure can be represented by the various forms shown in notation (11). Though planar, not all the carbon-carbon bond lengths in naphthalene are identical. The C_1–C_2 and C_{4a}–C_{8a} bonds are shorter (0.137 and 0.139 nm) than the C_2–C_3 and C_1–C_{8a} bonds (0.140 and 0.142 nm), suggesting that the resonance contributor shown with the numbering and having double bonds at the short bond distances contributes more to the resonance hybrid than do the other two contributors. The NMR spectrum of napthalene indicates a diamagnetic ring current, although the resonance energy is somewhat less than twice that of benzene.



(11)

Additional arene rings can be fused. For example, anthracene, tetracene, and pentacene are linearly fused, while phenanthrene, triphenylene, and pyrene are angularly fused (**Fig. 1**). In general, angular fusion results in more stable systems than linear



Anthracene

Tetracene

Pentacene

Phenanthrene

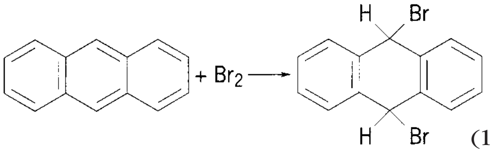
Triphenylene

Pyrene

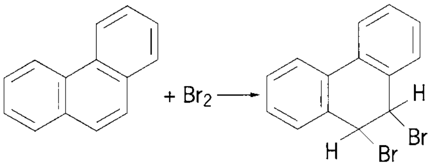
Fig. 1. Structures of some fused-ring (polynuclear) aromatic hydrocarbons.

fusion. Phenanthrene, for example, is about 6 kcal/mol more stable than its linear isomer anthracene. Stability falls off sharply in the linearly fused series, and compounds with more than seven such rings are unknown.

Anthracene and phenanthrene undergo addition reactions at the 9,10-positions [reactions (12) and (13)]. The products, each with two fully aromatic benzene rings, retain more than two-thirds of the original resonance energy. Even so, these bromine adducts readily eliminate hydrogen bromide (HBr) on heating to give the fully aromatic substitution products, 9-bromoanthracene or 9-bromophenanthrene.



(12)



(13)

Some polynuclear aromatic hydrocarbons are carcinogenic. Examples include benz[a]pyrene, 7,12-dimethylbenz[a]anthracene, and cholanthrene. Such compounds are environmental pollutants, being formed during the combustion of fuel oil, in the burning of cigarettes, in forest fires, and in roasting red meats. In the body, these hydrocarbons are enzymatically oxidized to related structures that react with deoxyribonucleic acid (DNA) bases, causing mutations that lead to cancer.

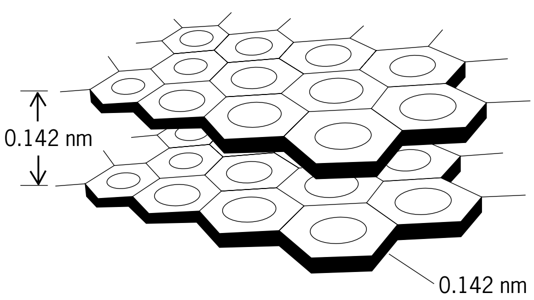
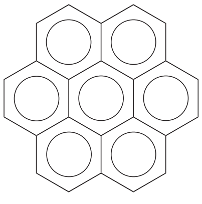


Fig. 2. Structure of graphite.

Increasing the number of ring fusions results in higher carbon:hydrogen (C/H) ratios. For example, the C/H ratio for benzene is 1.0; for naphthalene, 1.25; for anthracene, 1.4; and for pyrene, 1.6. If fusions are extended infinitely in two dimensions, sheets of carbon atoms are obtained that are arranged in hexagons, the known structure of graphite (**Fig. 2**). There are no formal bonds between the graphite layers. *See* GRAPHITE; POLYNUCLEAR HYDROCARBON.

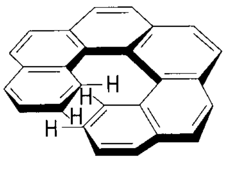
Bent and deformed benzene rings. Although arene rings are normally flat, they are more flexible than was originally thought. They can be bent substantially without losing their aromaticity.

Coronene (**25**) is a planar arene present in coal tar.



(**25**)

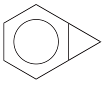
If two of the peripheral rings are disconnected and each completed, a planar geometry is no longer possible, and the molecule, called hexahelicene (**26**), assumes a spiral or helical arrangement of the



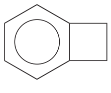
(**26**)

rings. Each arene ring is distorted a bit to accommodate the strain. The molecule can exist in two mirror-image forms, depending on whether the helix is right or left handed. Helicenes with additional rings (7–14) in the spiral have been prepared.

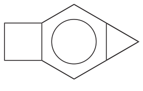
Fusion of small rings to an arene ring, as in structures (**27**)–(**31**), distorts the structure and imparts unusual reactivity to the arene ring. Benzocyclobutene (**28**), for example, undergoes thermal ring-opening to the highly reactive tetraene [(**32**); reaction (14)]; release of strain compensates for the loss of aromatic resonance energy. Despite the strain, fusion of more than one small ring to a single arene ring is possible [structures (**29**) and (**30**)]. Spectra of such compounds show that aromaticity is retained (for example, NMR spectra show the usual chemical shifts due to π-electron ring currents).



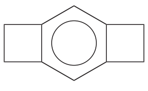
(**27**)



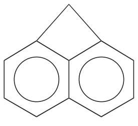
(**28**)



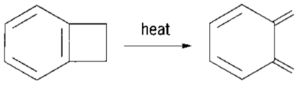
(**29**)



(**30**)



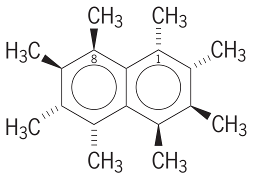
(**31**)



(14)

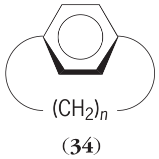
(**28**) (**32**)

Heavy adjacent substitution distorts arene rings. In octamethylnaphthalene (**33**), for example, the methyl groups and the carbon atoms to which they are attached are bent from planarity because of steric requirements of the substituents. The molecule is like a two-bladed propeller, and the angle between the C_1 and C_8 methyl substituents is about 40°. Yet spectral properties show that aromaticity is preserved.

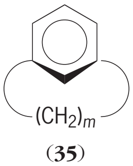


(**33**)

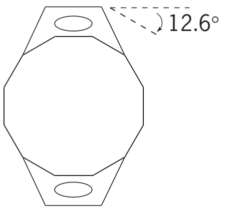
Cyclophanes are molecules in which an arene ring is part of a larger ring, as in *para*-cyclophanes (**34**) and *meta*-cyclophanes (**35**). If n or m is large the arene ring remains planar, but if small ($n = 6$, $m = 5$) the ring will distort, and if even smaller the molecule becomes impossible to construct. Two or more arene rings may be incorporated in a single cyclophane. In [2.2]paracyclophane (**36**), the "end" carbons are bent from planarity by 12.6°, yet the arene rings undergo most reactions (for example, substitution) characteristic of benzenoid compounds. Analogs of compound (**36**) have been synthesized in which from three to all six of the arene ring positions are bridged. *See* MACROCYCLIC COMPOUND.



(**34**)



(**35**)



(**36**)

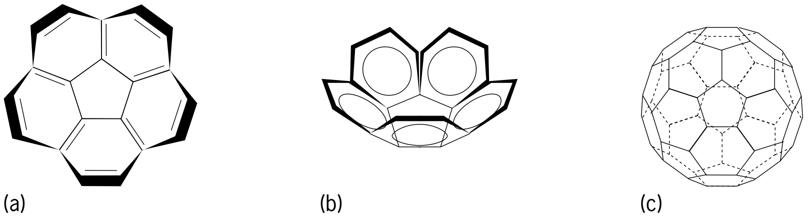
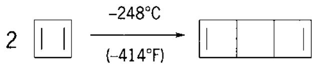


Fig. 3. Bent benzenoid structures. (*a*) Corannulene, showing the five arene rings fused to a central five-membered ring. (*b*) Corannulene, showing the cup shape. (*c*) Buckminsterfullerene, a spherical allotrope of carbon.

Corannulene (**Fig. 3*a***) has five arene rings fused to the central five-membered ring. The molecule, which cannot be planar without deforming the hexagons, is cup shaped (Fig. 3*b*). This structural unit is present in the spherical allotrope of carbon, C_{60} (Fig. 3*c*), known as buckminsterfullerene; it has no hydrogens and is not a hydrocarbon. Nevertheless, it is aromatic, since the bonding resembles that of benzene. That is, each carbon atom is covalently attached to three other carbons, and the fourth valence electron is in a *p* orbital that is perpendicular to the surface of the sphere. C_{60} has a π-electron cloud spread over the outside and inside surfaces of the sphere, just as benzene has a π-electron cloud above and below the plane of the hexagon. Though it has its own distinctive chemistry, C_{60} displays special stability associated with aromaticity. *See* FULLERENE.

Nonbenzenoid aromatic hydrocarbons. Given the special aromatic properties associated with the formal arrangement of alternating double and single bonds inscribed in a hexagon, it was natural to explore other ring sizes with similar arrangements. Cyclobutadiene (**37**) was exceptionally difficult to synthesize, and it is extremely reactive. Though it has been trapped in a rigid argon matrix at −265°C (−445°F), it dimerizes [reaction (15)] even at

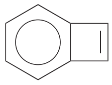


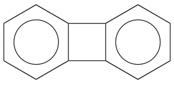
(15)

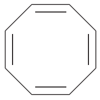
(**37**)

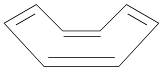
-248°C (-414°F) when the matrix softens. X-ray studies of more stable tetra-substituted analogs of cyclobutadiene show rectangular rather than square geometry, with alternate single and double bonds; that is, the π electrons are not delocalized as they are in benzene.

Even if one of the double bonds is incorporated in a benzene ring, as in (**38**), the compound resists isolation and rapidly dimerizes. Biphenylene (**39**), however, with both of cyclobutadiene's double bonds as part of a benzene ring, is a stable white crystalline compound.


(**38**)


(**39**)

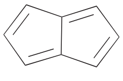

(**40a**)

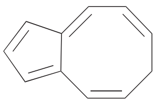

(**40b**)

Cyclooctatetraene (**40a**) has the properties of a cycloalkene with four double bonds. The structure is nonplanar and tub shaped (**40b**), and the π electrons are localized at the double bonds.

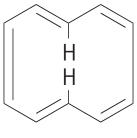
Hückel rule. Originally it was thought that cyclobutadiene was not aromatic and unstable because of the strain associated with the four-membered ring, and that cyclooctatetraene was not aromatic because it was nonplanar. But the difference between these molecules and benzene turned out to be more fundamental. From molecular orbital theory, E. Hückel derived the rule that planar, cyclic conjugated (alternate single and double bonds) systems with $4n + 2$ π electrons (n is an integer, 0, 1, 2...) will be aromatic and have substantial resonance energy, whereas those with $4n$ such electrons will not; indeed, it was later shown that $4n$ systems are often destabilized, hence antiaromatic.

Benzene is a $4n + 2$ system ($n = 1$) and aromatic; cyclobutadiene is a $4n$ ($n = 1$) system and antiaromatic. Cyclooctatetraene would be a $4n$ system ($n = 2$) if it were planar, but since such a system would be antiaromatic, there is no driving force for the molecule to become planar. As striking confirmation of these ideas, pentalene (**41**), a planar analog of cyclooctatetraene, is exceptionally reactive, unstable, and antiaromatic (a $4n$ system, $n = 2$), whereas the purple hydrocarbon azulene (**42**; a $4n = 2$ system, $n = 2$, and an isomer of naphthalene) is stable and undergoes substitution reactions analogous to those of benzenoid arenes.

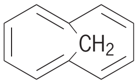

(**41**)


(**42**)

Annulenes. Monocyclic conjugated systems $(CH)_n$, where n is twice the number of double bonds, are called annulenes (benzene, where $n = 6$, is a [6]annulene). According to the Hückel rule, annulenes with 6, 10, 14, 18... π electrons, if planar, should be aromatic. In [10]annulene (**43**) with two trans double bonds, the inside hydrogen atoms must occupy the same space. Replacement of these hydrogens by a —CH_2— group as in (**44**) yields a "bridged" annulene with a nearly planar 10-π-electron system. In accord with the Hückel theory, the compound is stable and shows aromatic properties (diamagnetic ring current, substitution reactions).

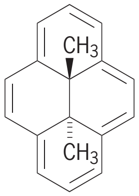


(**43**)

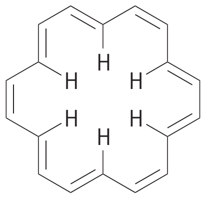


(**44**)

The [14]annulene (**45**) and [18]annulene (**46**) are among the many $4n + 2$ aromatic annulenes that have been synthesized. Despite the numerous nonplanar conformations accessible to such molecules, they tend to be planar. The 12 "outside" protons in (**46**) are strongly deshielded (NMR chemical shift 9.28) as in benzene, but the 6 "inside" protons are strongly shielded (NMR chemical shift -2.99). This result is entirely in accord with a strong ring current around the periphery of the molecule.



(**45**)

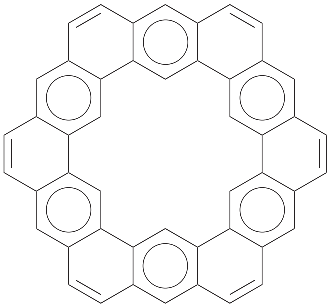


(**46**)

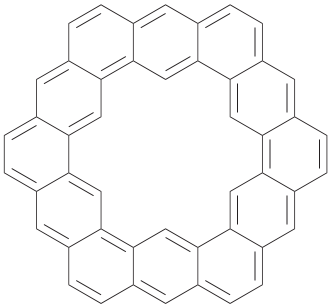
Annulenes with $4n$ π electrons have also been synthesized ($n = 3, 4, 5\ldots$), and in general they are less stable than the $4n + 2$ systems. The proton NMR spectra of the [12] through [24] annulenes alternate, being diamagnetic (aromatic) for the $4n + 2$ members and paramagnetic (nonaromatic) for the $4n$ members, consistent with the Hückel rule.

Kekulene, a cycloarene with 12 arene rings, provides an interesting comparison of benzenoid and nonbenzenoid arenes. Two of its possible π-electron arrangements are shown in (**47a**) and (**47b**). The former has six benzenoid rings connected by single and double bonds. The latter has no benzenoid rings but two $4n + 2$ annulene rings, the "inner" [18] ($n = 4$) and the "outer" [30] ($n = 7$). X-ray and NMR data all support structure (**47a**) as the best representation.

Aromatic ions. One of the most striking confirmations of Hückel theory came with its prediction of

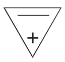


(47a)

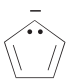


(47b)

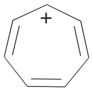
aromatic ions; the concept of aromaticity is not restricted to neutral molecules. For example, the cyclopropenyl cation (**48**), cyclopentadienyl anion (**49**), and tropylium cation (**50**) follow the Hückel $4n + 2$ rule, with $n = 0, 1,$ and 1 respectively. Each of these ions shows exceptional stability, and in each the ionic charge is delocalized over all ring atoms. A particularly striking example is seen with cyclooctatetraene. Though nonplanar and not aromatic, the hydrocarbon (**40a**) becomes planar when reduced to the dianion (**51**) which, being a Hückel system ($n = 2$), is aromatic, as in reaction (16) [K = potassium].



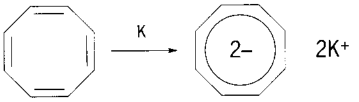
(**48**)



(**49**)



(**50**)



(16)

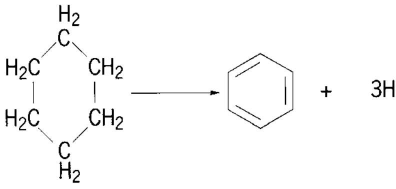
(**40a**) (**51**)

Harold Hart

Bibliography. A. T. Balaban, M. Banciu, and V. Ciorba, *Annulenes, Benzo-, Hetero-, Homo-derivatives and their Valence Isomers*, vols.1–3, 1987; D. H. R. Barton and W. Ollis (eds.), *Comprehensive Organic Chemistry*, vol. 1, ed. by J. F. Stoddard, 1979; J. R. Dias, *Handbook of Polycyclic Hydrocarbons*, 1987; F. Diederich, *Cyclophanes*, 1991; P. J. Garratt, *Aromaticity*, 1986; M. Hornby and J. M. Peach, *Foundations of Organic Chemistry*, 1993; D. Lloyd, *Nonbenzenoid Conjugated Carbocyclic Compounds*, 1984.

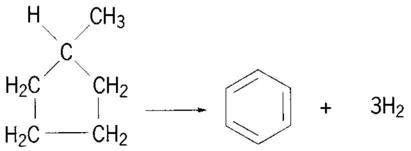
Aromatization

The conversion of any nonaromatic hydrocarbon structures, especially those found in petroleum, to aromatic hydrocarbons. There are numerous routes and means to accomplish this transformation, the simplest and most important of which are direct dehydrogenation of naphthenes to aromatics, reaction (1); dehydroisomerization of naphthenes to aromatics, reaction (2); dehydrocyclization of aliphatics to aromatics, reaction (3); and high-temperature condensation of hydrocarbons to aromatics, reaction (4).



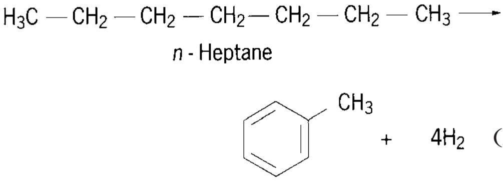
(1)

Cyclohexane Benzene Hydrogen



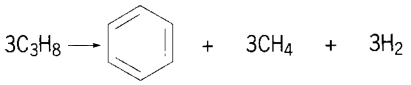
(2)

Methylcyclopentane Benzene Hydrogen



(3)

Toluene Hydrogen



(4)

Propane Benzene Methane Hydrogen

Reaction (1) was performed on a huge scale by the petroleum industry during World War II for the production of toluene for TNT, starting from methylcyclohexane and utilizing molybdenum oxide on porous activated alumina or mixed tungsten-nickel sulfides as catalysts. At that time, reaction (3), employing chromia on alumina catalysts, was investigated extensively but was found impracticable commercially.

Subsequently, reforming of naphthas with catalysts comprising small amounts of platinum on an acidified alumina support provided a means of aromatization that rapidly displaced the earlier processes, since, it accomplishes reactions (1), (2), and (3) readily and simultaneously. It is a major process for benzene, toluene, and other aromatics from petroleum sources.

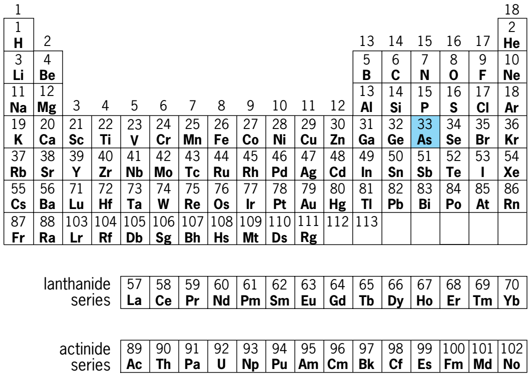
Reaction (4) merely illustrates one type of reaction that may occur in the high-temperature (600–800°C or 1100–1500°F) thermal cracking of petroleum fractions. Small hydrocarbons from the cracking of heavier components, as well as any originally present, may condense to aromatics, usually with considerable splitting to methane as well as hydrogen. Thermal cracking processes emphasizing this type of reaction have achieved only limited use in the petroleum industry. *See* PETROLEUM PROCESSING AND REFINING.

Bernard S. Greensfelder; Mott Souders

Bibliography. W. A. Gruse and D. R. Stevens, *Chemical Technology of Petroleum*, 3d ed., 1960; *Kirk-Othmer Encyclopedia of Chemical Technology*, 3d ed., vol. 1, 1978; E. I. Shaheen, *Catalytic Processing in Petroleum Refining*, 1983; J. G. Speight, *Chemistry and Technology of Petroleum: An Introduction*, 1982.

Arsenic

A chemical element, symbol As, atomic number 33. Arsenic is found widely distributed in nature (approximately 5×10^{-4} of the Earth's crust). It is one of the 22 known elements composed of only one stable nuclide, ${}^{75}_{33}As$; the atomic weight is 74.92158. There are 17 other radioactive arsenic nuclides known.



There are three polymorphic modifications of arsenic. The yellow cubic α-form is made by condensing the vapor at very low temperatures. The black β-polymorph is isostructural with black phosphorus. Both these modifications revert to the stable γ-form, gray or metallic, rhombohedral arsenic, on heating or exposure to light. The metallic form is a moderately good thermal and electric conductor and is brittle, easily fractured, and of low ductility. Some of the important atomic and bulk properties of arsenic are given in the **table**.

Arsenic is found native as the mineral scherbenkobalt, but generally occurs among surface rocks combined with sulfur or metals such as Mn, Fe, Co, Ni, Ag, or Sn. The principal arsenic mineral is FeAsS (arsenopyrite, mispickel); other metal arsenide ores are $FeAs_2$ (löllingite), NiAs (nicolite), CoAsS (cobalt glance), NiAsS (gersdorffite), and $CoAs_2$ (smaltite). Naturally occurring arsenates and thioarsenates are common, and most sulfide ores contain arsenic. As_4S_4 (realgar) and As_4S_6 (orpiment) are the most important sulfur-containing minerals. The oxide, arsenolite, As_4O_6 is found as the product of the weathering of other arsenical minerals, and is also recovered from flue dusts collected during the extraction of Ni, Cu, and Sn from their ores; it also results when the arsenides of Fe, Co, or Ni are roasted in air or oxygen. The element may be obtained by roasting FeAsS or $FeAs_2$ in the absence of air or by reduction of As_4O_6 carbon, when As_4 may be sublimed away.

Some atomic and bulk properties of arsenic

Property	Value
Electron configuration (S^4 ground state)	$[Ar]\ 3d^{10}4s^24p^3$
Covalent radius	121 pm
Ionic radius (As^{3+})	69 pm
Metallic radius	139 pm
Ionization energies 1st–6th in kJ mole^{-1}	947, 1798, 2734, 4834, 6040, 12300
Electrode potential, As^{3-}/As	0.25 V
Electronegativity (Allred-Rochow)	2.20
Oxidation numbers	−3, 0, +3, +5
Specific gravity (α, β, γ)	2.026, 4.7, 5.727
Melting point (γ)	817°C (1503°F) (3.6 MPa)
Boiling point (γ)	616°C (1141°F) (sublimes)
Electrical resistivity	33.3 μΩ cm (273 K)
Toxicity level	0.5 mg · m^{-3} of air

Elemental arsenic has few uses. It is one of the few minerals available in 99.9999+% purity, which is largely used in the laser material GaAs and as a doping agent in the manufacture of various solid-state devices. Arsenic oxide is used in glass manufacture. The arsenic sulfides are used as pigments and in pyrotechnics. Dihydrogen arsenate is used in medicine, as are several other arsenic compounds. Most of the medicinal uses of arsenic compounds depend on their toxic nature. *See* ANTIMONY; PHOSPHORUS.

John L. T. Waugh

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. F. Shriver and P. W. Atkins, *Inorganic Chemistry*, 3d ed., 1999.

Art conservation chemistry

The application of chemistry to the technical examination, authentication, and preservation of cultural property. Chemists working in museums engage in a broad range of investigations, most frequently studying the chemical composition and structure of artifacts, their corrosion products, and the materials used in their repair, restoration, and conservation. The effects of the museum environment, including air pollutants, fluctuations in temperature and relative humidity, biological activity, and ultraviolet and visible illumination, represent a second major area of research. A third area of interest is the evaluation of the effectiveness, safety, and long-term stability of materials and techniques for the conservation of works of art. Though analytical techniques appear to dominate, many other areas of chemistry, biology, physics, and engineering, including polymer chemistry, kinetic studies, imaging methodologies, biodegradation studies, dating methods, computer modeling, metallography, and corrosion engineering, play active roles in conservation science.

Examination and analysis. Methods of examination may be divided into two classes: those that provide an image of the entire object (holistic examination; **Fig. 1**) or a section of it; and those that provide an analysis at a point on the object, with or without sampling. Nondestructive methods, not requiring sampling, are always preferable. However, modern methods of analysis can be employed on such minute samples that they are in effect nondestructive. In some cases, samples must be taken for methods that are in principle nondestructive because the object is too large to fit into a sample chamber. The ability to analyze minute samples introduces the serious concern that the sample may not be representative of the composition of the artifact but may be an inclusion or contaminant introduced by the experimentalist. With specimens from painted surfaces, great care must be taken to identify areas of restoration.

A further concern arises from the differing depths from which signals originate. On a metal surface, ion scattering spectrometry (ISS) would see the initial fraction-of-a-nanometer, predominantly adsorbed species and contaminants. Secondary ion mass spectrometry (SIMS) would begin to penetrate the oxidized area; Auger electron spectrometry (AES) would examine the bulk of the oxidized layer; and x-ray-induced photoelectron spectrometry (XPS) would give data on the bulk sample some 10 nm below the specimen surface. In the analysis of silver artifacts by neutron activation, significant differences were observed in results for samples taken by drilling when compared with streak samples obtained by rubbing a quartz slide across the surface of the artifact. *See* ACTIVATION ANALYSIS; AUGER EFFECT; NONDESTRUCTIVE EVALUATION; SECONDARY ION MASS SPECTROMETRY (SIMS); SURFACE PHYSICS; X-RAY FLUORESCENCE ANALYSIS.

The most commonly employed holistic method is x-ray radiography, where variations in the density and average atomic number of the sample attenuate an x-ray beam, leaving a negative image on film. Ordinary medical x-ray equipment is routinely used with paintings and small three-dimensional objects, while high-energy 300-kV x-ray radiography is used for large stone and metal sculpture. Watermarks are detected in prints and drawings by β-radiography with a poly(methyl methacrylate) plate doped with carbon-14. Computerized tomography (CAT) scans and nuclear magnetic resonance (NMR) imaging have been applied to Egyptian mummies both to assess their condition and to conduct physical anthropological studies. *See* COMPUTERIZED TOMOGRAPHY; NUCLEAR MAGNETIC RESONANCE (NMR); RADIOGRAPHY.

Other methods, such as ultraviolet and infrared reflectance and fluorescence, are used to show areas of compositional difference indicating restoration or variation in the pigments used by the artist. Infrared reflectography has been most successfully employed in the study of preparatory drawings under the paint layers in Netherlandish paintings and has been introduced to the study of underdrawings in medieval manuscript illumination. *See* INFRARED IMAGING DEVICES; INFRARED RADIATION; LUMINESCENCE ANALYSIS; ULTRAVIOLET RADIATION.

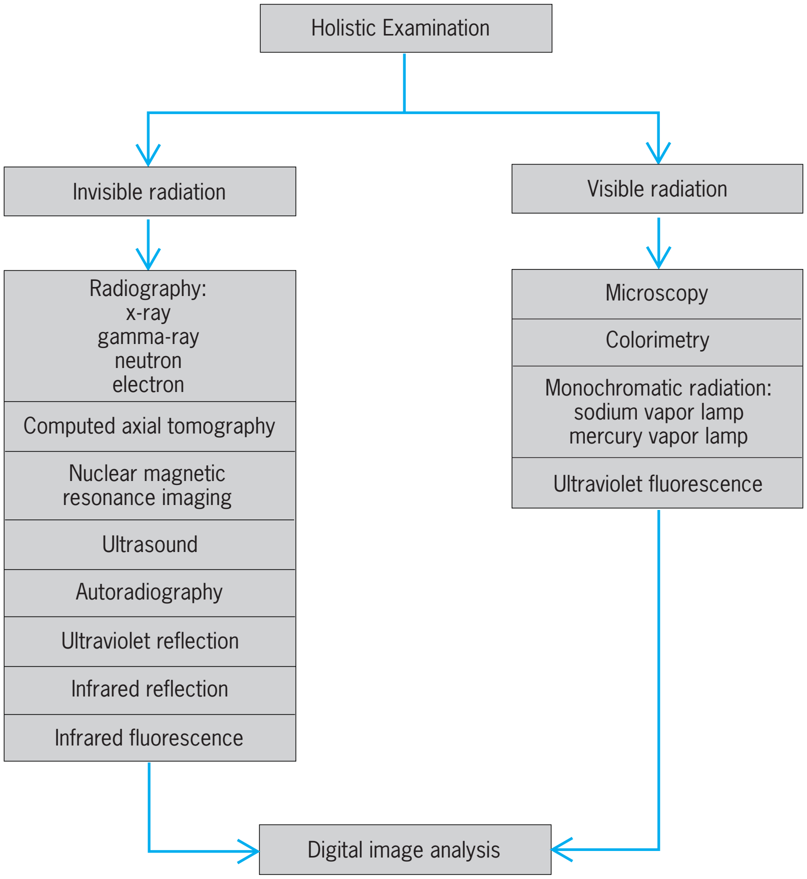


Fig. 1. **Methods for holistic investigation of art objects. (*After F. Mairinger and M. Schreiner, New methods of chemical analysis: A tool for the conservator, in N. S. Bromelle and G. Thomson, eds., Science and Technology in the Service of Conservation, International Institute for Conservation, 1982*)**

In the examination of paintings, small samples are taken under the binocular microscope, embedded in transparent resin, and polished to produce a cross section for microscopic examination. This permits a study of the artist's painting technique and shows how several layers may have been built up to achieve a desired effect. The same specimens may then be subjected to electron beam microprobe examination. The microprobe in a sequence of point analyses permits the creation of a holistic picture of the elemental distribution through the several layers of the paint specimen. A series of elemental scans can be combined by image-processing software to give a color-coded mapping of the pigments distributed throughout the specimen. *See* CHEMICAL MICROSCOPY; IMAGE PROCESSING; MICROSCOPE.

Conservation studies of composition and technique embrace the entire spectrum of modern chemical analysis (**Fig. 2**).

Authentication. The separation of the fake from the authentic is a small but often spectacular aspect

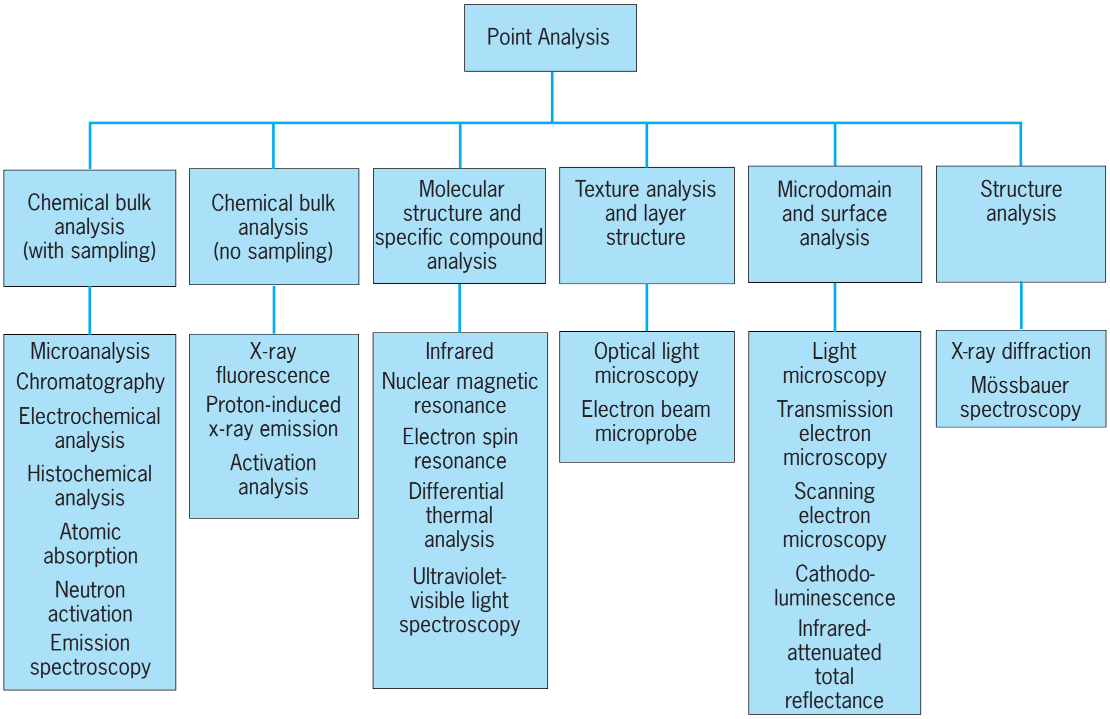

Point Analysis
Chemical bulk analysis (with sampling)
Chemical bulk analysis (no sampling)
Molecular structure and specific compound analysis
Texture analysis and layer structure
Microdomain and surface analysis
Structure analysis
Microanalysis
Chromatography
Electrochemical analysis
Histochemical analysis
Atomic absorption
Neutron activation
Emission spectroscopy
X-ray fluorescence
Proton-induced x-ray emission
Activation analysis
Infrared
Nuclear magnetic resonance
Electron spin resonance
Differential thermal analysis
Ultraviolet-visible light spectroscopy
Optical light microscopy
Electron beam microprobe
Light microscopy
Transmission electron microscopy
Scanning electron microscopy
Cathodo-luminescence
Infrared-attenuated total reflectance
X-ray diffraction
Mössbauer spectroscopy

Fig. 2. Methods for point analysis applied to the examination of works of art. (*After F. Mairinger and M. Schreiner, New methods of chemical analysis: A tool for the conservator, in N. S. Bromelle and G. Thomson, eds., Science and Technology in the Service of Conservation, International Institute for Conservation, 1982*)

of the technical examination of artifacts. In some cases, direct age determination (dendrochronology for panel paintings, fission track dating for uranium glass, radiocarbon dating for organic materials, thermoluminescence dating for ceramics) is possible. The development of direct counting methods for radiocarbon dating promises to expand the application of this technique as the minimum sample size has been reduced from grams to milligrams. In 1988, the Shroud of Turin was declared a medieval forgery based on radiocarbon dates obtained by three independent laboratories. *See* ARCHEOLOGICAL CHRONOLOGY; DATING METHODS; DENDROCHRONOLOGY; FISSION TRACK DATING; RADIOCARBON DATING.

More commonly, the issue of authenticity turns upon anachronisms in composition or technique when the artifact in question is compared to accepted artifacts of the period. Thus, the greater part of the work in the conservation laboratory concerns the building of databases of analyses of composition, trace-element distributions, and studies of technique. For example, the analysis of 1000 Sassanian silver coins shows gold impurity levels (**Fig. 3**) in modern fake coins to be distinct from those of authentic specimens. In some cases, the silver used was so pure (less than 0.001% gold) that it could only have been refined in the twentieth century.

Anachronisms in the pigments used have unmasked many forgers, including Hans Van Meergeren (1889–1947), who forged and sold a series of paintings in the style of Vermeer to major Dutch collections. The presence of one of these paintings in the collection of the Nazi leader Hermann Goering led to a notorious trial that relied heavily on technical analysis, including pigment identification, x-ray radiography, and microchemical demonstration that the paint medium was a twentieth-century phenol-formaldehyde resin, selected to imitate the reduced solubility of aged drying oil media. In 1972, pyrolysis

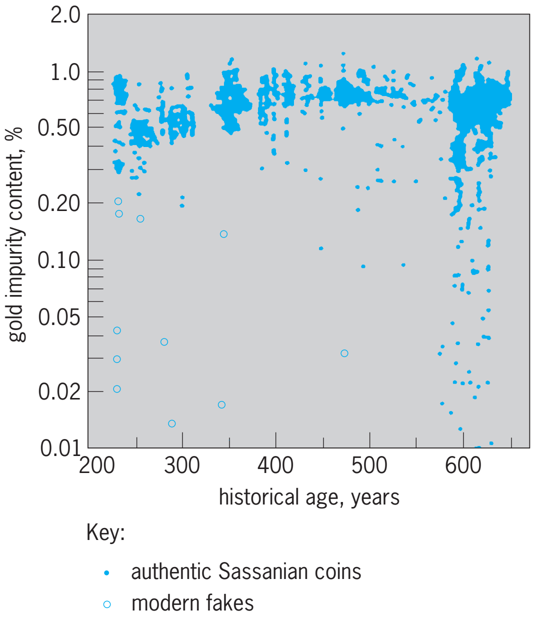


Fig. 3. Gold impurity levels versus historical age for Sassanian (Near East, A.D. 224–641) coins. Data points representing modern fakes indicate date of the originals that the fake coins intended to imitate. (*After S. J. Fleming, Authenticity in Art: The Scientific Deduction of Forgery, Crane Russak Co., 1976*)

gas chromatography was used to identify the specific resin used in the most famous of these faked Vermeers, *Christ at Emmaus*, sold to the Museum Boymans in Rotterdam. *See* PIGMENT (MATERIAL).

A demonstration of the limits of technical examination is provided by a comprehensive study of the painting *Deborah Kip, Wife of Sir Balthasar Gerbier, and Her Children* (*The Gerbier Family*) by Peter Paul Rubens, acquired by the National Gallery of Art, Washington, D.C., in 1971. Among the techniques employed were x-ray radiography, optical microscopy, emission spectroscopy, and electron beam microprobe study of the pigments and paint layer structure. Transmission and scanning electron microscopy were applied to the chalk white in the ground layer of the painting in order to identify coccoliths, calcareous remains of unicellular marine algae. Lead isotope mass spectrometry was used to develop a database of the distribution of lead isotope ratios in lead-white samples from all paintings examined: all paintings examined painted before 1800; Dutch, Flemish, and Belgian paintings painted before 1800; and different portions of *The Gerbier Family* (**Fig. 4**). The ever-narrowing distribution of lead isotope ratios reflecting ever-narrowing sources of supply and the pigments observed were highly consistent throughout the painting. Yet, despite the wealth of technical data, it could not be stated with certainty that Rubens had painted the entire picture, or that the six portions of canvas that had been sewn together to produce this large painting were painted at the same time by the same hand. *See* ELECTRON MICROSCOPE; EMISSION SPECTROCHEMICAL ANALYSIS; LEAD ISOTOPES (GEOCHEMISTRY); MASS SPECTROMETRY.

Pigment analysis by optical light microscopy, x-ray diffraction, and ion microprobe of ultramicro samples from the Vinland Map owned by Yale University identified significant amounts of anatase, a modern titanium dioxide pigment first available in the early twentieth century. The map, thought to be a fifteenth-century copy, was purchased as the earliest map of the New World providing evidence that Leif Ericsson had visited North America some 500 years before Columbus. Researchers using proton-induced x-ray emission (PIXE) have attempted to challenge these pigment findings, claiming that their elemental analyses show only small amounts of titanium in the map. These results are less persuasive than those of the earlier analysis, since in that work the microscopic pigment particles were positively identified while in the proton-induced x-ray emission PIXE analyses an average elemental analysis over a much wider area was obtained. *See* PROTON-INDUCED X-RAY EMISSION (PIXE).

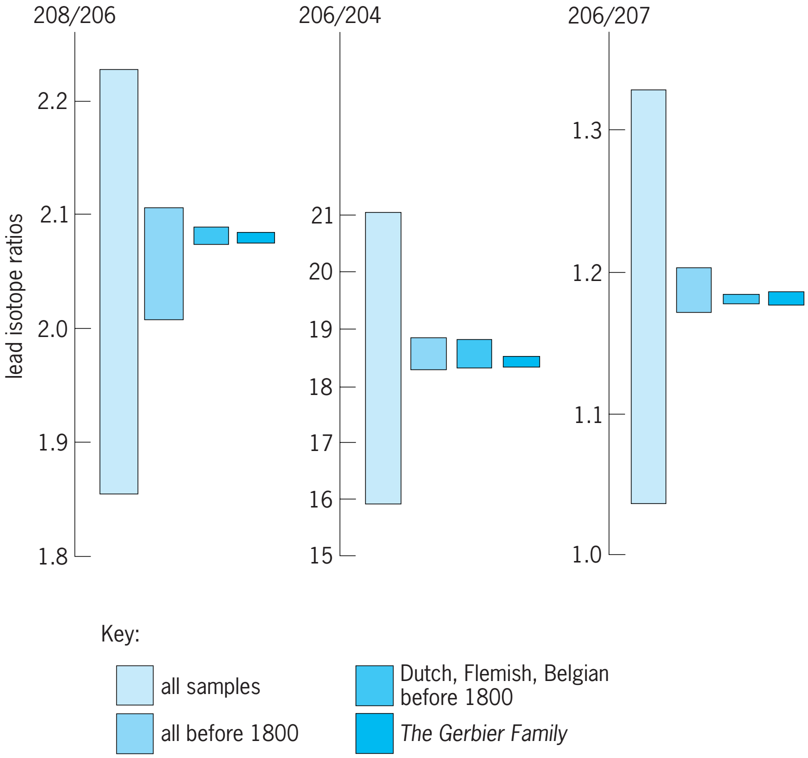


Fig. 4. Distribution of lead isotope ratios in lead-white pigment samples taken from paintings. (*After B. Keisch and R. C. Callahan, Rubens' The Gerbier Family: Investigation by Lead Isotope Mass Spectrometry, National Gallery Studies in the History of Art, pp. 75–78, 1973*)

Substantial collections of notable occurrences of pigments in paintings have been developed, permitting estimates of the probability of finding a specific pigment in a painting of a given period. Typically, in paintings from 1300 to 1800 there is almost 100% certainty that lead white $[2PbCO_3 \cdot Pb(OH)_2]$ will be detected, while in paintings of around 1900 estimates of the probability of finding certain pigments are as follows: lead white, 90%; zinc white, 50%; and barium sulfate, as an extender in white pigments, 50%. The data show that several pigments of the same hue will frequently appear in a single painting.

Other studies have used thin-layer chromatography to follow the introduction of natural dyes such as cochineal from the New World in Flemish tapestries of the fifteenth and sixteenth centuries. Cochineal, derived from scale insects, was discovered in Mexico but was extensively imported to Europe shortly after, since its properties were superior to those of kermes, the principal insect dye used in Europe up to that time. Similar studies on yellow dyes taken from textiles of the period 1500–1850 demonstrated that 80% were weld (dyers' weed), which was used in both European textiles and Anatolian carpets. *See* ARCHEOLOGICAL CHEMISTRY; DYE; TEXTILE CHEMISTRY.

Museum environment. Many artifacts are sensitive to destructive agents in the museum atmosphere. Rapid changes in relative humidity will cause dimensional changes in wood furniture, polychrome sculpture, and panel paintings, leading to cracking and splitting of the wood with loss of painted surface decoration. High relative humidity can lead to mold growth and foxing on books and prints, while low relative humidity will cause photographic prints and films to become brittle. Chemists have developed treated silica gel to stabilize the relative humidity where sensitive artifacts are displayed in sealed cases. *See* HUMIDITY; HUMIDITY CONTROL.

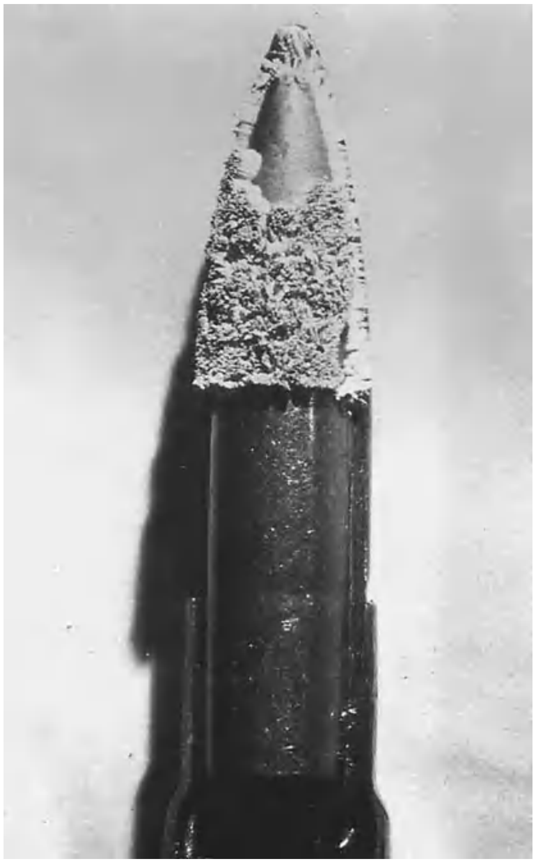


Fig. 5. Lead bullet with lead formate corrosion product formed by reaction with formic acid released from display case materials. (*Smithsonian Institution*)

Oxidation of iron objects, tarnishing of silver plate, and the development of corrosion products on lead artifacts by the action of formic and acetic acids emitted by wooden display cases have regularly been observed in museums. Lead formate was identified by infrared spectroscopy as the corrosion product (**Fig. 5**) found on a lead bullet in a plywood-backed display case. Some artifacts are made of unstable materials and decompose if left untreated. For example, cellulose nitrate film has, through spontaneous combustion, caused major fires in film repositories. Books and works of art on paper supports made of acid paper rapidly degrade by acid hydrolysis. Methods for treating acid paper with an alkaline reserve have been developed, including the Library of Congress diethyl zinc vapor phase process for books and a process using a solution of methyl magnesium carbonate to deposit magnesium carbonate in acid paper. *See* PAPER.

The common air pollutants sulfur dioxide (SO_2) and ozone (O_3) have been monitored at elevated levels in museums, libraries, and archives. These pollutants cause the degradation of leather, spotting of photographic prints, and fading of dyes and pigments. Chemical methods of analysis are used to identify degradation products and to study the kinetics of degradation mechanisms. Specialists in air-pollution monitoring use analytical instrumentation to measure ambient pollution levels in museums. Some of the museums remove SO_2 and O_3 from the air by using filter systems containing activated carbon or aluminum oxide impregnated with potassium permanganate in their heating, ventilation, and air conditioning (HVAC) systems.

With a few highly important artifacts, inert atmospheres have been used. Thus, the Declaration of Independence, the Constitution, and the Bill of Rights are displayed, in the National Archives of the United States, in a helium atmosphere within a metal frame bonded to glass. To prevent contamination and oxidation, some Moon rocks are kept in inert atmospheres.

Conservation treatment. The most frequently employed techniques of the museum conservator involve chemicals and chemical reactions. Adhesives are used to mend broken pots, glass, metal, and so forth; benzotriazole to stabilize bronzes suffering from so-called bronze disease; bleaches such as hydrogen peroxide to remove stains and foxing from prints and drawings; diethyl zinc or methyl magnesium carbonate to deacidify paper; ethylene oxide and sulfuryl fluoride to fumigate infested textiles to arrest biodeterioration; and solvents to remove selectively discolored varnishes or overpainting from paintings on canvas. Friable materials such as waterlogged wood must be freeze-dried; or, if too large to fit into a vacuum chamber, they must be consolidated with resins or poly(ethylene glycol). Consolidants are also used with friable stone, leather, archeological ivory, and any other material that has been severely degraded. These processes are developed by collaboration of the scientist and the conservator to assure that the techniques are effective, safe, and, wherever possible, reversible.

Part of the evaluation process for conservation materials and techniques involves the examination of long-term stability. Two major approaches are

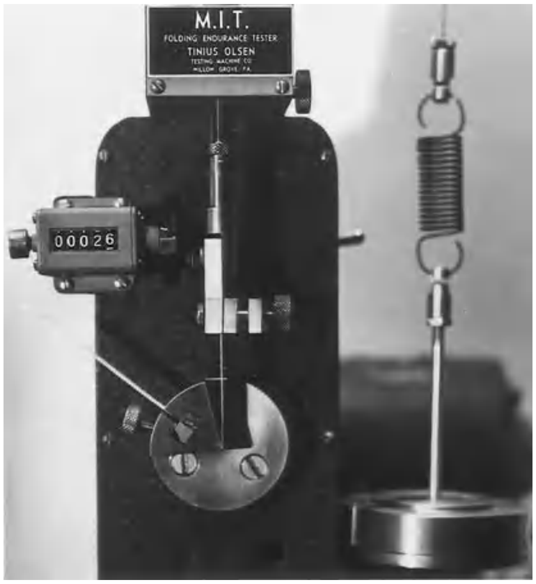


Fig. 6. Folding endurance tester used to evaluate the increase in strength of paper that is treated with the impregnating agent poly(vinyl alcohol).

used: light exposure and accelerated thermal aging. With the former, the natural or treated specimen is subjected to intense illumination for extended periods, commonly 2000 h. With thermal aging, the system under evaluation is heated for 1–50 days at temperatures of 122–212°F (50–100°C). In both cases, the aged samples are then studied by observing changes in physical properties [tensile strength, folding endurance (**Fig. 6**), color, molecular weight, compressive strength] and chemical properties [solubility, pH, change in composition]. Many materials which have ideal properties in commercial applications are unacceptable in conservation treatments, because they cannot be expected to remain stable for the extended periods required in museum practice. Much of the work of museum conservators involves removing materials used in the past which have discolored or become unstable or brittle. Mending tape on prints and drawings, shellac varnishes and pottery mends, and linseed oil consolidation are a few examples of poor materials used in attempts to repair works of art. Often such toxic solvents as dimethyl formamide and tetrahydrofuran must be used to undo the damage caused by materials. *See* ANALYTICAL CHEMISTRY.

Norbert S. Baer

Bibliography. N. S. Bromelle and P. Smith (eds.), *Adhesives and Consolidants*, 1984; N. S. Bromelle and G. Thomson (eds.), *Science and Technology in the Service of Conservation*, 1982; R. L. Feller (ed.), *Artists' Pigments: A Handbook of Their History and Characteristics*, 1994; S. J. Fleming, *Authenticity in Art: The Scientific Detection of Forgery*, 1975; J. S. Mills, *The Organic Chemistry of Museum Objects*, 2d ed., 1999; G. Thompson, *The Museum Environment*, 1986.

Arteriosclerosis

The name given a group of degenerative diseases of arteries characterized by thickening and hardening of their walls. The group includes three types of lesions: (1) atherosclerosis involves the aorta and its major branches; (2) medial sclerosis involves the muscular arteries of the legs; and (3) arteriolosclerosis involves the small branches of the arterial tree, called the arterioles.

Atherosclerosis. Atherosclerosis is by far the most common and important form of arteriosclerosis, and the two terms are often used interchangeably. Some degree of atherosclerosis is found at autopsy in persons of all ages, even children, but it increases in severity and extent with advancing years. One study on 22-year-old American soldiers in the Korean War revealed the presence of atherosclerotic lesions in the coronary vessels in 77% of them. It is more common in males than in females, especially females before the age of menopause. After menopause the sex incidence and severity are about equal.

Incidence. Atherosclerosis has global distribution and occurs in virtually epidemic proportions in the industrialized nations of Western civilization. There is an extremely high prevalence and severity of the lesions in the Northern European countries, the United States, Australia, and New Zealand. There is a relatively low occurrence in economically deprived areas such as India, Africa, and South America. Factors associated with the high incidence and severity of atheromas (anatomic lesions of atherosclerosis) include a high total caloric intake, high fat intake, sedentary living, aggressive personality, emotional stress, and cigarette smoking. Hypertension (high blood pressure) does not induce arteriosclerosis, but it augments its development and accelerates the progress of the disease if it is present. The excessive incidence of myocardial infarction (heart attack) in cigarette smokers has been clearly documented. High blood lipid levels especially of cholesterol and triglycerides are also associated with higher incidence. The strong supposition thus arose that environmental and nutritional factors are of prime etiologic importance in the development of the disease, although the specific cause has not been identified. Although virtually ubiquitous, atherosclerosis fortunately does not cause serious disease symptoms in most individuals. *See* HEART DISORDERS; HYPERTENSION.

Pathology. Atheromas begin within the lining layer of the aorta (intima) or its branches and subsequently extend into the middle layer (media) of these vessels. The basic lesion has been shown to be a focal overgrowth of the smooth muscle cells of these layers. There is evidence that the initial lesion is a monoclonal overgrowth of smooth muscle cells, in some respects resembling a benign smooth muscle tumor such as a leiomyoma. This monoclonal overgrowth of smooth muscle cells, forming the initial atherosclerotic plaque, represents a mutation; its most likely cause, therefore, is a chemical mutagen from the environment or viruses. The cells subsequently degenerate, producing an accumulation of lipids in their cytoplasm, then necrosis, and finally calcification and scarification. As these lesions enlarge in the intima, the overlying endothelial cells may become disrupted, and fibrin clots are deposited on the surface. The lumen of the vessel is diminished in diameter (stenosis) by both the atheromatous plaque and the overlying clot. Subsequently, the fibrin clot may be incorporated into the fibrous scar in the intima. The occurrence of successive layers of clots and enlarging plaque produces narrowing of the lumen. When this process happens in large vessels such as the thoracic or abdominal aorta, it may not produce occlusion. However, if the process involves all layers of the arterial wall, it may weaken the wall, and an aneurysm may result. This may rupture, and the person may die from hemorrhage into the body cavity. *See* ANEURYSM.

When the atherosclerotic process occurs in the smaller branches of the aorta, a variety of consequences may ensue. Complete occlusion of the lumen by clot or hemorrhage into the plaque may occur. If the process takes place in the vessels supplying blood to the heart muscle (coronary arteries), the blood supply may be restricted or stopped completely, producing a myocardial infarction. If the process involves the arteries supplying the brain, a similar occlusion may produce an infarction of the brain. Sometimes hemorrhage into the infarction may ensue. This is called a cerebrovascular accident (stroke). If arteries leading to the legs, arms, or internal organs are occluded, gangrene may result. Because of these sequelae, atherosclerosis assumes awesome importance as the major cause of death in the United States. *See* HEMORRHAGE.

Causes. Although the vast majority of cases of atherosclerosis appear to be principally caused by environmental factors, there are some specific genetic defects associated with the genesis of the process. Individuals with primary diabetes mellitus develop severe arteriosclerosis at an earlier age than nondiabetic individuals do. Familial hypercholesterolemia, homocystinuria, and hypothyroidism are other examples of metabolic defects associated with arteriosclerosis. *See* DIABETES; THYROID GLAND DISORDERS.

Medial sclerosis. Medial sclerosis is an uncommon type of arteriosclerotic lesion of little clinical significance. It affects the arteries of the arms, legs, and genital tracts of both sexes equally. It rarely occurs under the age of 50, and its cause is obscure. The disorder is characterized by ringlike calcifications within the media (middle layer) of affected vessels. The endothelial lining of the vessel is not altered, and the size of the lumen not impinged upon. Therefore, occlusion of the vessel seldom occurs.

Arteriolosclerosis. Arteriolosclerosis is an increased generalized thickening of the walls of arterioles related to hypertension. The change is often most prominent in the kidneys, although other internal organs may be similarly affected. Individuals with diabetes mellitus have an increased incidence of the lesion. *See* CIRCULATION DISORDERS.

N. Karle Mottet

Bibliography. J. D. Wilson et al. (eds.), *Harrison's Principles of Internal Medicine*, 14th ed., 1998; R. W. Wissler (ed.), *Arteriosclerosis Plagues: Advances in Imaging for Sequential Quantitative Evaluation*, 1992.

Artesian systems

Ground-water conditions formed by water-bearing rocks (aquifers) in which the water is confined above and below by impermeable beds. These systems are named after the province of Artois in France, where artesian wells were first observed.

Because the water in the intake area of an artesian system is higher than the top of the aquifer in its artesian portion, the water is under sufficient head to cause it to rise in a well above the top of the aquifer. Many of the systems have sufficient head to cause the water to overflow at the surface, at least where the land surface is relatively low. Flowing artesian wells were extremely important during the early days of the development of ground water from drilled wells, because there was no need for pumping. Their importance has diminished with the decline of head that has occurred in many artesian systems and with the development of efficient pumps and cheap power with which to operate the pumps. When they were first tapped, many artesian aquifers contained water that was under sufficient pressure to rise 100 ft (30 m) or more above the land surface. Besides furnishing water supplies, many of the wells were used to generate electric power. With the increasing development of the artesian aquifers through the drilling of additional wells, the head in most of them has decreased and it is now from a few feet to several hundred feet below the land surface in many areas of former artesian flow. A majority of artesian wells are now equipped with pumps. *See* AQUIFER.

Perhaps the best-known artesian aquifer in the United States is the Dakota sandstone, of Cretaceous age, which underlies most of North Dakota, South Dakota, and Nebraska, much of Kansas, and parts of Minnesota and Iowa at depths ranging from 0 to 2000 ft (0 to 600 m). The water is highly mineralized, as a general rule, but during the latter part of the nineteenth century, when these areas were being settled, the Dakota sandstone provided a valuable source of water supply under high pressure. Few wells in this aquifer flow more than a trickle of water today. The St. Peter sandstone and deeper-lying sandstones of early lower Paleozoic age, which underlie parts of Minnesota, Wisconsin, Iowa, Illinois, and Indiana, form another well-known artesian system. Formerly, wells on low ground flowed abundantly, but now wells have to be pumped throughout most of the area. Some of the water is highly mineralized, but in many places it is of good quality. In New Mexico, in the Roswell artesian basin, cavernous limestone of Permian age provides water to irrigate thousands of acres of cotton and other farm crops. Although the head has been steadily declining, many wells still have large flows and others yield copious supplies by pumping. Among the most productive artesian systems are the Cretaceous and Tertiary aquifers of the Atlantic and Gulf Coastal plains. These provide large quantities of water for irrigation and industrial use and supply large cities, such as Savannah, Georgia; Memphis, Tennessee; and Houston and San Antonio, Texas. Numerous artesian basins are found in intermontane valleys of the West. Some of the best known are in the Central Valley, California, where confined aquifers provide water to irrigate millions of acres of farmland, and the San Luis Valley, Colorado. Numerous other lesser artesian systems are found in all parts of the United States. *See* GROUND-WATER HYDROLOGY.

Albert N. Sayre; Ray K. Linsley

Arthritis

A group of diseases affecting joints or their component tissues. Several types of arthritis are recognized, and these can be divided into groups by their clinical course and pathologic appearance.

Types. There are four basic types of arthritis: inflammatory arthritis, degenerative joint disease, nonarticular rheumatism, and miscellaneous arthritis.

Inflammatory arthritis. This type is characterized by inflammation of tissues associated with joints. Connective tissue diseases, crystal deposition diseases, infectious arthritis, and spondyloarthropathies are examples of inflammatory arthritis.

Connective tissue diseases are a group of acute and chronic diseases characterized by involvement of joints, connective tissue, serosal membranes, and small blood vessels. These diseases are divided into acquired disorders (for example, rheumatoid arthritis, systemic lupus erythematosus, scleroderma, polymyositis, vasculitis) and rare hereditary diseases (for example, Ehlers-Danlos syndrome).

Rheumatoid arthritis is the most common variety of inflammatory arthritis. It occurs in younger and middle-aged persons and is characterized by noninfectious inflammation of the synovium (joint-lining membrane) frequently associated with extraarticular manifestations other than in the joints. The etiology is unknown, but genetic, immunologic, infectious, and psychologic disturbances have all been suggested. The systemic disease follows a variable but slowly progressive course, marked by spontaneous flares and remissions. Symmetrical involvement of the smaller joints of the hands and feet are often the initial clinical signs of disease, with wrists and other large joints affected later. Destruction of joints is common with advancing disease, resulting in a high incidence of disability.

Other connective tissue diseases, such as systemic lupus erythematosus, scleroderma, and vasculitis, frequently include joint symptoms. Other manifestations—involving soft-tissue structures such as kidneys, muscle, heart, and lungs—are the primary source of morbidity and mortality in these conditions. *See* CONNECTIVE TISSUE DISEASE.

There are three groups of crystal deposition disease classified according to type of crystal involvement: gout (monosodium urate), pseudogout (calcium pyrophosphate), and calcific tendonitis (hydroxyapatite). Gout, a disorder of uric acid breakdown, has long been associated with a violent arthritis, particularly of the first toe. Deposits of monosodium urate crystals in the affected joints and irregular episodes of symptoms are typical features. *See* GOUT; URIC ACID.

Infectious arthritis is an inflammatory joint disease caused by the invasion of the synovial joint by living microorganisms such as gonorrheal, streptococcal, and staphylococcal bacteria. Such arthritis usually results from a generalized infection but may appear following local spread or after trauma. The outcome varies considerably with the person's age and immune status, the adequacy of the treatment, and the nature of the organism. The most common type of infectious arthritis occurs in disseminated gonococcemia, but the gonococcal organisms are frequently difficult to isolate from the involved joints. Rheumatic fever, a sequel to streptococcal infection, tends to include a migratory arthritis that may involve several joints in succession; permanent damage to the joints is minimal. Tuberculous arthritis is less prevalent now than before the advent of adequate therapy for tuberculosis. *See* GONORRHEA; RHEUMATIC FEVER; STAPHYLOCOCCUS; STREPTOCOCCUS; TUBERCULOSIS.

The spondyloarthropathies are types of inflammatory arthritis characterized by involvement of the axial (central) skeleton (for example, the spine rather than the limbs). Ankylosing spondylitis (sacroiliitis) and Reiter's syndrome (clinical triad of arthritis, conjunctivitis, and urethritis) are examples of the spondyloarthropathies. All are hereditary to some extent through association with the histocompatibility system.

Degenerative joint disease (osteoarthritis). This is a ubiquitous joint disease characterized pathologically by deterioration of cartilage lining the joints and new bone formation beneath the cartilage. The disease is very common in older persons and is thought to be inherent in the aging process. In addition, various other factors are involved, including joint injury and the presence of other forms of arthritis. Degenerative joint disease is marked by a progressive stiffness, loss of function, and destruction of the larger, weight-bearing joints of the body. With advancing age, the continued slow damage causes increasing disability. *See* AGING.

Nonarticular rheumatism. This group of diseases, also called soft-tissue rheumatisms, includes tendonitis, bursitis, tenosynovitis, and fibrositis. Calcific tendonitis and nonseptic bursitis usually follow local trauma, muscle strain, or excessive exercise but can result from deposition of hydroxyapatite crystals (crystal deposition disease). Septic bursitis may also occur from local invasion by microorganisms. Fibrositis is a complex of musculoskeletal pains that frequently involves the neck, shoulder girdle, and extremities and is associated with multiple trigger points. The etiology is unclear, but the disorder may relate to psychobiologic or sleep disturbances or muscular and soft-tissue abnormalities. *See* BURSITIS; RHEUMATISM.

Miscellaneous arthritis. Systemic diseases of other or unknown etiology may produce arthritis or joint destruction. There are neurologic, blood, and endocrine examples of these unusual rheumatic diseases.

Treatment. Disability can often result from arthritis but can be curtailed by general health maintenance, rest, and rehabilitation. Occupational and physical therapies can be helpful. Diagnosis of the particular type of arthritis is extremely important in choosing drug therapy. Rheumatoid arthritis can be treated with nonsteroidal anti-inflammatory drugs such as aspirin. Disease-modifying antirheumatic drugs such as gold, antimalarial drugs, and immunosuppressive drugs are also frequently used. Osteoarthritis is treated with anti-inflammatory and analgesic drugs. Cortisone compounds are also used to treat arthritis, but the dosage and duration of treatment must be carefully monitored because of significant side effects. Surgical treatment includes arthroscopic surgery and joint replacement.

Arthritis affects an estimated 11 million persons in the United States. The annual cost of treatment and care and the losses in time, money, and productivity are incalculable. With an increasing survival time in the population, arthritis constitutes one of the greatest medical, social, and economic problems. *See* AUTOIMMUNITY; JOINT DISORDERS. Robert Searles

Bibliography. R. G. Lahita (ed.), *Systemic Lupus Erythematosus*, 3d ed., 1998; D. J. McCarty, *Arthritis and Allied Conditions*: *A Textbook of Rheumatology*, 13th ed., 1996; A. Soren, *Arthritis and Related Affectations*: *Clinic, Pathology, and Treatment*, 1993.

Arthropoda

The largest phylum in the animal kingdom, including the well-known insects, spiders, ticks, and crustaceans, as well as many smaller, less well-known groups, and an abundance of strange forms known only as fossils. Arthropods make up about 95% of all species of animals. The estimated number of known species exceeds 1 million; of this number, most are insects. No one really knows the exact number of arthropod species. Some authorities estimate that it may range as high as 10 million.

Arthropods vary in size from the microscopic mites to the giant decapod crustaceans, such as the king crab with an appendage span of 5 ft (1.5 m) or more. The adult arthropod typically has a body composed of a series of ringlike segments as well as a pair of many-jointed limbs on each segment, movable on each other by means of muscles. However, some parasites, such as the pentastomids or the rhizocephalans, as adults have virtually no sign of segmentation. The integument of arthropods secretes a multilayered cuticle containing chitin. Often this cuticle is sclerotized by the deposit of hardening proteins. To accommodate growth, this exoskeleton must be shed periodically in a process called molting, or ecdysis. Young stages may be quite different from the adults, and some parasitic species differ very radically in body form from their nearest relatives. These characteristics, taken together, distinguish the arthropods from all other animals.

Classification

At one time, it was thought that the arthropods consisted of several phyla, that is, that they had a multiplicity of origins from different ancestors (polyphyly). However, recent investigations have successfully demonstrated that Arthropoda is a coherent group that arose from a single ancestor in the distant past (monophyly). Nevertheless, there are as many ways of classifying the arthropods as there are arthropod researchers. This is especially true when attempting to incorporate the wide array of extinct arthropods seen in the fossil record. The taxonomic system given here is generally accepted and places all the living groups in a logical scheme along with only a few of the fossil forms. The asterisk indicates extinct forms. The classification is modified from R. C. Brusca and G. Brusca, *The Invertebrates*, 2d ed., 2003.

Generalized Higher Classification of Phylum Arthropoda*

- Subphylum Fuxianhuiida*
    - Trilobitomorpha* (including Trilobita)
    - Cheliceriformes
- Subphylum Cheliceriformes
    - Class Chelicerata
        - Subclass Merostomata
        - Arachnida
    - Class Pycnogonida
- Subphylum Myriapoda
    - Class Diplopoda
    - Chilopoda
    - Pauropoda
    - Symphyla
- Subphylum Hexapoda
    - Class Entognatha (including Parainsecta)
    - Insecta
- Subphylum Crustacea
    - Class Remipedia
    - Cephalocarida
    - Branchiopoda
    - Malacostraca
        - Subclass Phyllocarida (Leptostraca)
        - Hoplocarida
        - Eumalacostraca
    - Class Maxillopoda
        - Subclass Thecostraca (including Cirripedia)
        - Tantulocarida
        - Ostracoda
        - Copepoda
        - Branchiura (including Pentastomida)
        - Mystacocarida

The phylogenetic relationships between these subphyla are still not certain. The hexapods are believed by many to be closely related to the crustaceans, but as to which group of crustaceans is still strongly argued. However, depending on whether or not one includes fossils in the analysis, hexapods can emerge as more closely related to the myriapods (a traditional affiliation). Conversely, recent molecular evidence seems to indicate that the myriapods by themselves might be closely allied with the cheliceriforms, but strong arguments also ally them with other mandibulate groups such as the hexapods and crustaceans. Trilobitomorphs, long thought to bear some relationship to the cheliceriforms, recently have been suggested to have common origins with mandibulates. The phylogenetic relationships within each of these subphyla are only beginning to be settled. A massive, worldwide effort (called "Assembling the Tree of Life") is under way to utilize information from a variety of sources [comparative anatomy,

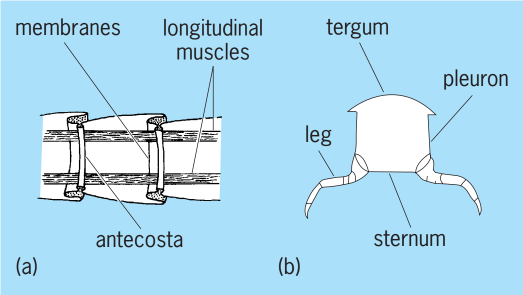


Fig. 1. Arthropod segmentation. (*a*) Segmental plates. (*b*) Cross section of a body segment. (*After R. E. Snodgrass, Principles of Insect Morphology, McGraw-Hill, 1935*)

embryology, developmental genetics, molecular deoxyribonucleic acid (DNA) sequences] to elucidate kinship connections.

Metamerism

Body segmentation, or metamerism, is one of the most fundamental characteristics of the arthropods (**Fig. 1***a*). However, it is a feature shared with several other phyla, such as the annelid worms, kinorhynch worms, and even chordates. The link with annelid worms has been formalized into a superphylum group called Articulata. In this respect, metamerism would be viewed as the key character. However, as striking as this feature is, other aspects of the biology suggest that the link between annelids and arthropods is not so close. The molecular sequencing of deoxyribonucleic acid (DNA) has advanced an alternative view in which arthropods are united into a superphylum called Ecdysozoa, which also includes kinorhynchs, priapulids, nematodes, onychophorans, and tardigrades. In this respect, molting would be viewed as the key character. The ecdysozoans are contrasted to another superphylum, called Lophotrochozoa, in which the annelids, mollusks, and some other phyla are united.

The arthropod cuticle may be sclerotized continuously around the segments. More typically, however, it forms discrete segmental plates, or sclerites. A dorsal plate of a segment is called a tergum or notum; a ventral plate is a sternum; lateral plates are pleura (Fig. 1*b*). The consecutive tergal and sternal plates, unless secondarily united or fused, are connected by infolded arthrodial membranes, and are thus movable on each other by longitudinal muscles attached on anterior and posterior marginal ridges on the underside of the plates. Since nearly all the body and limb muscles are attached to either integumental sclerites or to skeletal elements infolded from the surface sclerites called endophragms, there is effectively no limit to the development of skeletomuscular mechanisms. Furthermore, the body segments, associated with specialized sets of appendages, tend to become consolidated or united into segment groups, or tagmata, forming differentiated body regions, such as the head, thorax, and abdomen.

The segmentally arranged limbs of all modern arthropods develop in the embryo from small lateroventral outgrowths of the body segments that lengthen and become jointed. One phylum, the onychophorans, appears to present a model for an ancestor to arthropods, whereby rudimentary limbs could have given rise to more elaborate appendages. Whether this ancestor was a crawling or swimming form is still widely debated. The fossil record provides examples of both walking and swimming forms that could serve equally well as models for an arthropod ancestor. *See* ONYCHOPHORA; ANNELIDA.

The first segmental limbs may have been all leglike appendages (as they were in the trilobites), which served for walking across the bottom of the sea. On the other hand, some authorities have suggested that the earliest limbs were flagellar (very long with many small segments) in form with the ancestral scuttling along the bottom or swimming. In their later evolution, however, arthropod legs became modified in structure for many other purposes, such as feeding, grasping, swimming, respiration, silk spinning, egg laying, and sperm transfer. Thus, modern arthropods possess legs that are highly modified and more specialized than those seen in any other group of animals. *See* METAMERES.

Development

A compelling set of arguments for the unity of Arthropoda arises from the relatively new field of evolutionary developmental genetics, or "evo-devo." The exploration of how genes control various aspects of development in both animals and plants has yielded new information relative to assessing the origin of anatomical structures and systems. For example, a series of maternal genes set the polarity of the animal egg even before development begins. Other genes in the zygote quickly specify numbers and polarity of individual segments by actually segregating distinct anterior and posterior halves of segments in arthropods. A critical gene in this regard, known as *engrailed*, specifies the anterior portions of what in the embryo are called the parasegments, the first manifestations during ontogeny of any metamerism. As development progresses, the two halves of the parasegment uncouple from each other and associate with the tissues of the adjacent parasegment. As a result, the *engrailed* gene comes to be expressed in the posterior portion of the adult segment, a pattern seen in all arthropods.

The *Hox* genes, another family of genes responsible for setting the anterior-posterior polarity in the animal body, have become specialized in the arthropods into a family of genes that also help specify the identity of segments within body regions, or tagmata. The genes *labial*, *proboscipedia*, *Deformed*, and *Sex combs reduced* play a role in determining the identity of segments in the arthropod head (**Fig. 2**). The genes *Antennapedia*, *Ultrabithorax*, and *abdominal A* are expressed in various ways in the thorax. *Abdominal B* and *caudal* are genes that are expressed in the tail end of the body, for example, in the pleon of a crayfish (Fig. 2). The *Hox* genes are located on

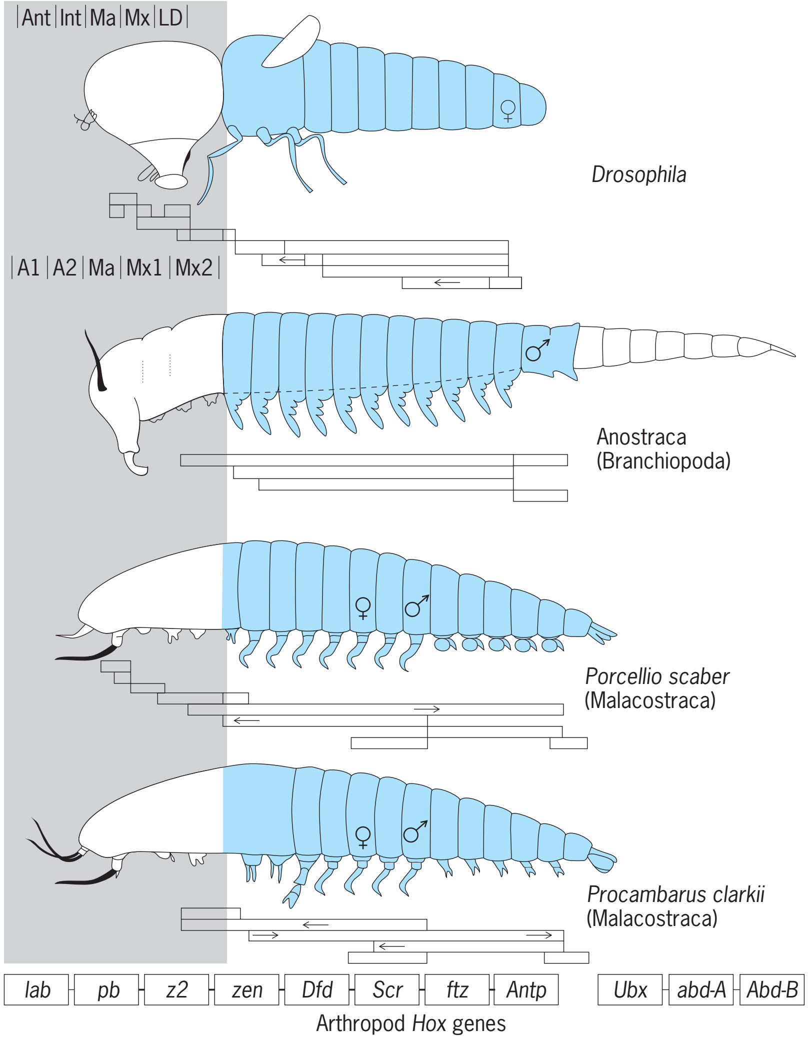


Fig. 2. Areas of expression of the various *Hox* genes in selected mandibulate arthropods. *lab* = labial; *pb* = proboscipedea; *z2* = zerknuellt; *zen* = zen; *Dfd* = deformed; *Scr* = sex combs reduced; *ftz* = fushi terazu; *Antp* = Antennapedia; *Ubx* = Ultra biothorax; *abd-A* = abdominal-A; *AbdB* = abdominal-B.

the chromosome in the order in which they are read and expressed from anterior to posterior. Variations occur, but the underlying pattern is identical across all arthropod groups.

In addition, the genes that control specification of the limbs are employed in a distinctive way characteristic of all the arthropods examined so far. Two systems operate. The basalmost parts of the limb, the coxopodites, are controlled by a gene called *extradenticle*. This system is quite separate from the complex array of genetic interactions that specifies various aspects of the more distal portions of the arthropod limb. Here, the *engrailed* gene triggers a gene series called the *hedgehog* cascade, wherein the genes engaged in a specific sequence specify various quadrants of the developing limb bud, culminating in the tip of the appendage where a gene called *distalless* is expressed.

No matter how different the body plan of an adult arthropod may be, the same genes are present in all arthropods and are expressed in the same areas of the body. While most of these genes are also found in a wide array of animal phyla, it is the distinctive combination outlined above that is unique to arthropods. In other words, all arthropods share a common pattern of developmental control and body region specification. *See* DEVELOPMENTAL GENETICS.

Internal Organs and Physiology

The arthropods, from lobsters to the tiniest gnat, have all the internal organs essential to any complex animal (**Fig. 3**).

Gut. A gut extends either straight or coiled from the anterior, ventrally directed mouth to the terminal anus. There are three regions in the gut. The anterior stomodeal region (foregut) is formed by an ingrowth of the outer germ layer of the embryo, the ectoderm. This stomodeum is often elaborately developed with teeth and plates of hardened chitin, and setose sieves and filters. The middle part of the gut, or mesenteron, is formed from the endoderm and contains the digestive diverticula as well as a region for absorption. This middle region can be quite short in some arthropods. The posterior gut region is the proctodeum (hindgut), also formed from ingrowths of ectoderm.

Nervous system. The nervous system includes a brain and a subesophageal ganglion in the head, united by connectives around the stomodeum, and the system generally displays paired ventral nerve cords with segmental, interconnected ganglia. Some of the successive ganglia, however, may be condensed into composite ganglionic masses. Peripherally directed nerves proceed from the ganglia. Internal proprioceptors and surface sense organs of numerous kinds (chiefly tactile, olfactory, and optic) are present. The sensory range can be quite remarkable: Insects can see into the ultraviolet range; mantis shrimp have color perception that far exceeds the abilities of humans; and spiders and scorpions "see" by vibratory excitement of minute hairs on their legs.

Heart. The dorsal heart is often tubular, but it also can be short and compact. Arteries can distribute blood in a number of different ways. The heart can pump either directly through arteries to the various tissues and organs (a more or less closed system) or from the anterior end of the tube directly into the body cavity (an open system). Conversely, blood can reenter the heart through openings (valves) along its sides from the body cavity, or return to the heart by means of veins from the body tissues.

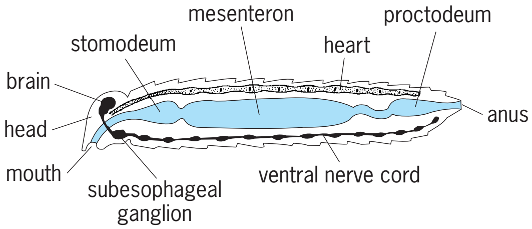


Fig. 3. Diagram of position of principal internal organs in a typical arthropod.

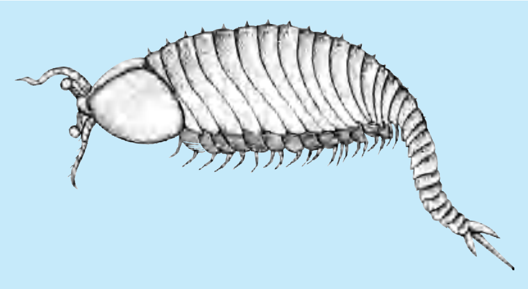


Fig. 4. The Cambrian proto-arthropod, *Fuxianhuia protensa*, from China.\vskip8pt

Respiration and excretion. Aquatic arthropods breathe by means of gills. Most terrestrial species have either flat air pouches, or tubular tracheae opening from the outside surface; some have both. A few small, soft-bodied forms respire directly through the thin cuticle. Excretory organs open either at the bases of some of the appendages or into the midgut.

Reproduction. Most arthropods have separate sexes, but some are hermaphrodites. Parthenogenesis is known but can be said to be common only in certain groups of insects. The genital openings differ in position in different groups and are a key component in defining the distinct body plans found in the phylum. The gonopores usually (but not always) are on the same body segment in the two sexes.

Fuxianhuiida

This is an extinct group of arthropods known from the Cambrian more than 550 million years ago. The currently recognized forms had rather large and long bodies (**Fig. 4**). However, the most interesting body part is the head. It appears to have only two segments: one that contained the eyes and one with an antennalike limb between which was the mouth. What is perhaps significant is that in modern arthropods these same two segments form the anteriormost part of the head and are the only segments that do not have any *Hox* gene expression. The mouth was anteroventral but had no special limbs to assist in capturing or rendering prey, so it is possible the fuxianhuiids merely plowed through bottom sediments and ate the mud.

Trilobitomorpha

The trilobitomorphs are extinct and are among the oldest known arthropods. Trilobita (or trilobites) were the dominant group, yet they are not the actual ancestors of any modern arthropods. *See* TRILOBITA.

A typical trilobite has a flattened oval shape with an elevation, the rachis, running lengthwise along the middle of the back, giving it a three-part appearance from which the trilobites get their name (**Fig. 5**). More important, however, is the crosswise division of the body into a head, a thorax, and a pygidium. The dorsal aspect of the head is not overtly segmented, but indentations of the median elevation suggest a primitive metamerism. On its upper surface, laterally, it bears a pair of compound eyes. The antennae extend forward on the undersurface, and a posteriorly directed mouth is located just posterior to a ventral sternite between the antennae called the hypostome. The posterior half of the ventral head bears four pairs of limbs. The thorax is completely segmented, and each metamere bears a pair of limbs similar to those on the head. The minimal number of thorax segments is eight, but there are often many more than that. The pygidium is a rounded plate at the posterior end of the body showing evidence of former segmentation.

The trilobite ancestors were probably fully segmented primitive arthropods. All of the postoral head and trunk limbs were alike and consisted of a long series of seven-segmented legs. The similarity of these limbs suggests that the diversified appendages of other arthropods were once jointed legs used for locomotion. Furthermore, each trilobite leg (**Fig. 6**) has a fringed or setose branch, perhaps a gill, arising laterally from the basal leg segment. Many trilobites have medially directed, spiny gnathobases on the basalmost limb segment, and these structures probably served for grasping food and passing it forward to the mouth, similar to what occurs in the living horseshoe crabs. The position of the reproductive

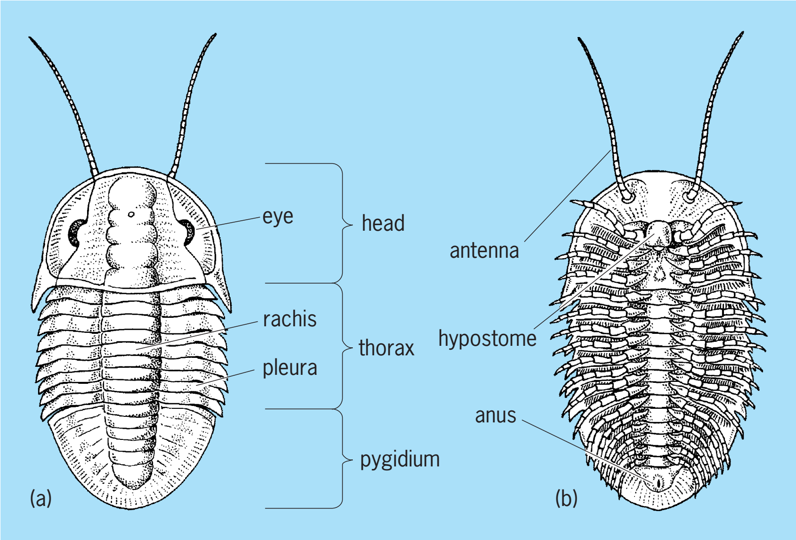


Fig. 5. Generalized trilobite. (*a*) Dorsal aspect. (*b*) Ventral aspect. (*After R. E. Snodgrass, Textbook of Arthropod Anatomy, Cornell University Press, 1952*)

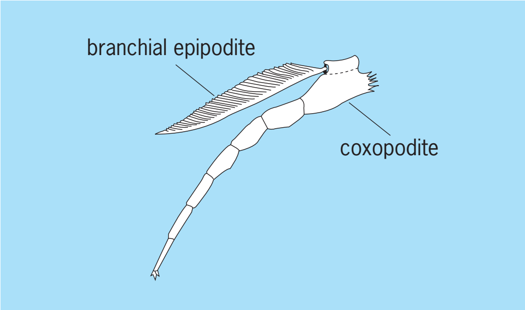


Fig. 6. Generalized trilobite leg.

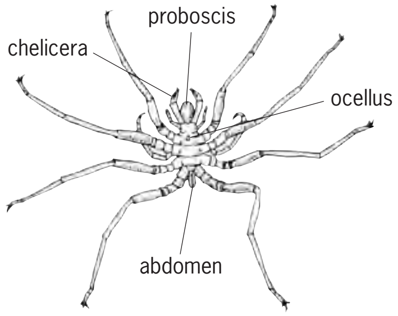


Fig. 7. ***Nymphon hirtipes*, a pycnogonid, an inhabitant of ocean shores. (*After R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*)**

openings is not known. The anus is at the end of the pygidium.

Cheliceriformes

The cheliceriformes get their name from the fact that their first appendages are a pair of small pincers, known as chelicerae. They have no antennae, but it is now clear from developmental genetic studies of *Hox* genes that the chelicerae are on the same segment that bears the antennae of other arthropods. Cheliceriforms can be viewed as arthropods that had a distinctive *Antennapedia* mutation early in their history, leading to the formation of chelicerae rather than antennae.

Two classes are recognized: Pycnogonida and Chelicerata.

Pycnogonida. The pycnogonids are generally small, long-legged creatures that live mostly along the shores of the ocean, or in the deep sea. Their slender legs give them the superficial appearance of spiders. They differ, however, in that the abdomen in living forms is an insignificant lobe projecting from between the posteriormost legs (**Fig. 7**), and the mouth is situated on a large, anteriorly directed proboscis. Pycnogonids feed particularly on the tentacles of sea anemones, which they reduce to a pulp and swallow in liquid form. Although most pycnogonids have eight walking legs, some deepwater forms in Antarctica have 10 and 12 legs. Pycnogonids have a set of gonopores at the base of each walking limb. *See* PYCNOGONIDA.

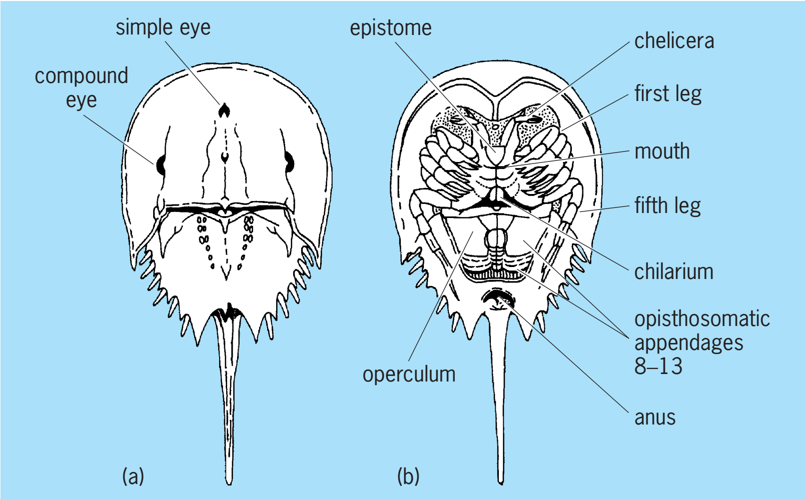


Fig. 8. Horseshoe crab (*Limulus*). (*a*) Dorsal aspect. (*b*) Ventral aspect. (*After R. E. Snodgrass. A Textbook of Arthropod Anatomy, Cornell University Press, 1952*)

Chelicerata. The body of the chelicerata is divided into two parts: a prosoma (or cephalothorax) and an opisthosoma (or abdomen). The prosoma carries the chelicerae, a second pair of appendages (termed pedipalps), and usually four pairs of walking legs. Appendages are usually absent on the opisthosoma, but when present they are never leglike. The genital aperture in each sex is always on the eighth postoral segment. The chelicerates include the subclasses Merostomata (which includes the orders Xiphosurida and Eurypterida) and Arachnida. *See* CHELICERATA.

Merostomata. The subclass Merostomata includes the order Xiphosurida (horseshoe crabs), which are not crabs at all (**Fig. 8**), and Eurypterida (extinct sea scorpions). As in the trilobites, the basal segments are very spiny and directed medially. These spiny lobes work against each other to "chew up" food before it is passed forward toward the mouth. The xiphosurids feed on solid food. All the other modern chelicerates are liquid feeders.

The eurypterids are extinct aquatic merostomes that lived in the Paleozoic Era. A typical eurypterid has an elongate body divided into an unsegmented prosoma or cephalothorax carrying six pair of appendages (**Fig. 9**) and a segmented opisthosoma or abdomen without appendages. Among the prosomal appendages, the first pair are the chelicerae, while the second pair are the homologs of the pedipalps of the arachnids. In some species, the opisthosoma ends in a tail bearing a large terminal spine and so resemble scorpions, such that they are

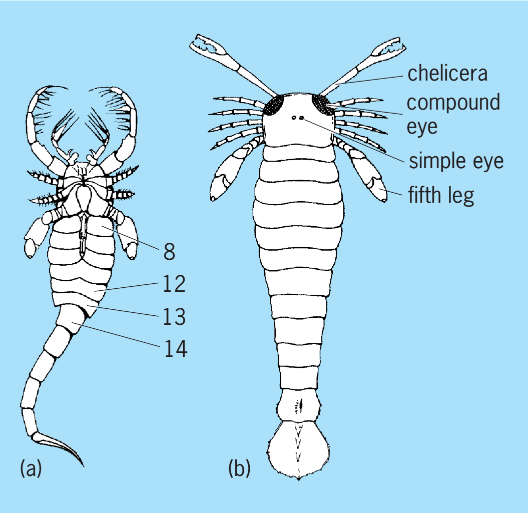


Fig. 9. Eurypterids. (*a*) *Mixopterus kiaeri*. (*b*) *Pterygotus buffaloensis*. (*After R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*)

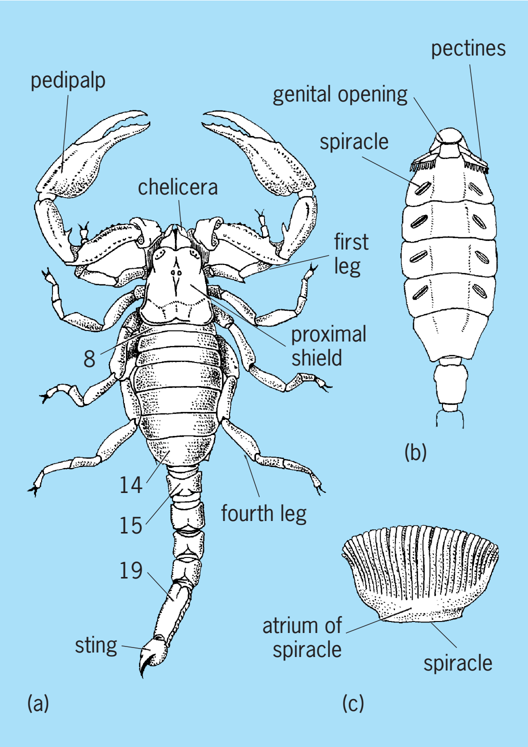

pedipalp
pectines
genital opening
spiracle
chelicera
first leg
proximal shield
8
(b)
14
15
fourth leg
19
sting
atrium of spiracle
spiracle
(a)
(c)

Fig. 10. Scorpion (*Chactas vanbenedeni*). (*a*) Entire animal, dorsal aspect (*after R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*). (*b*) Anterior abdomen, ventral aspect. (*c*) Book lung, diagrammatic.

presumed to be the ancestors of modern scorpions (Fig. 9*a*). Although most of the eurypterids were 6–15 in. (15–38 cm) long, some attained a length of 6–8 ft (1.8–2.4 m). A scorpion of such a size would indeed be a formidable animal. *See* EURYPTERIDA; MEROSTOMATA; XIPHOSURIDA.

Arachnida. The familiar subclass Arachnida includes the modern scorpions (**Fig. 10**) as well as the spiders (**Fig. 11**), daddy longlegs, ticks (**Fig. 12**), mites, and others, besides a large number of fossil species. The first leg posterior to the chelicerae is modified as a grasping appendage, or pedipalp. The prey is pierced (or torn open by the chelicerae), digestive juices are discharged into it from the stomach, and the predigested tissue is then ingested by the sucking pharynx. These corrosive digestive secretions are why the bites of spiders are often so painful and difficult to heal.

Arachnida includes the orders Palpigradi, Schizomida, Scorpiones (scorpions), Uropygi (tailed whip scorpions), Amblypygi (tailless whip scorpions), Araneae (spiders), Solfugae (sun spiders), Pseudoscorpionida (pseudoscorpions), Opiliones (harvestmen), Ricinulei, and Acari (ticks and mites). *See* ARACHNIDA.

Mandibulata

The mandibulate arthropods include the myriapods (centipedes and millipedes), hexapods (insects), and crustaceans. There is great diversity within the group, but they may be readily distinguished from the chelicerates by the presence of antennae and the possession of mandibles, which are typically a pair of biting and chewing jaws, but these limbs may be drawn out into cutting blades or piercing stylets. The antennae arise from the front of the head and are never leglike in form. The crustaceans have a pair of second antennae that arise behind the mouth and belong to the series of segmental limbs. In place of a second pair of antennae, the myriapods and hexapods lack a distinct appendage on that segment, but appear to modify limb rudiments as an upper lip, or labrum. The labrum forms an anterior lip or shield over the mouth. The mandibles posterior to the second antennae are the second postoral appendages. These have evolved from a pair of primitive legs by the development of teeth or masticatory surfaces on the basal segments and the complete elimination, or reduction to a small palp, of the rest of the limb. Posterior to the mandibles are two accessory feeding

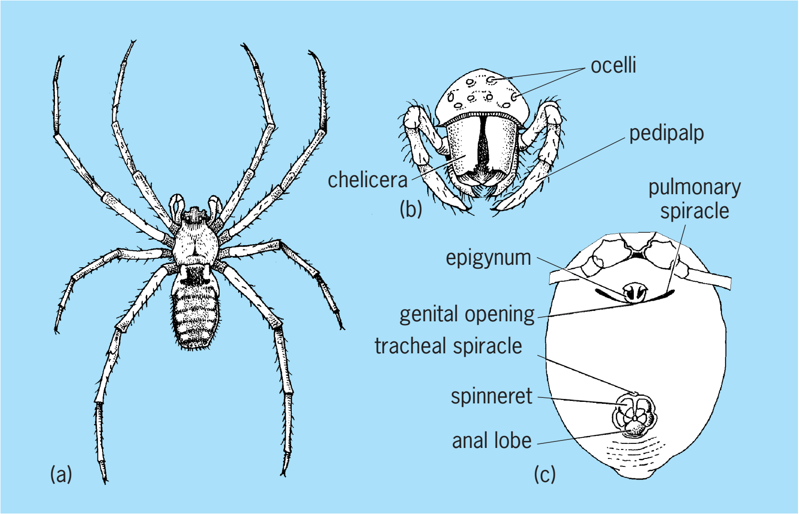

ocelli
pedipalp
chelicera
(b)
pulmonary spiracle
epigynum
genital opening
tracheal spiracle
spinneret
anal lobe
(a)
(c)

Fig. 11. Orb-weaving spider. (*a*) Entire animal, dorsal aspect. (*b*) Anterior view of head. (*c*) Undersurface of abdomen.

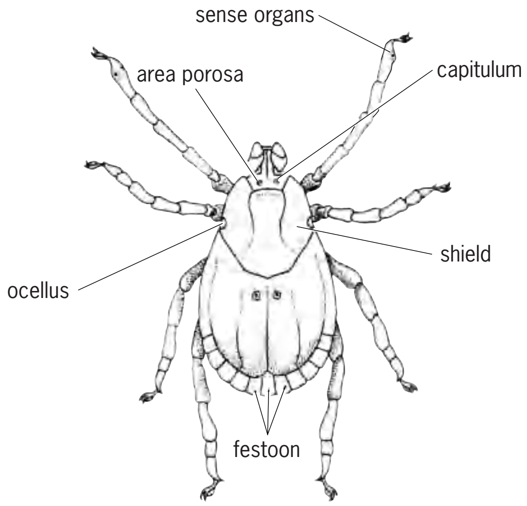

sense organs
area porosa
capitulum
shield
ocellus
festoon

Fig. 12. Tick (*Dermacentor*). (*After R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*)

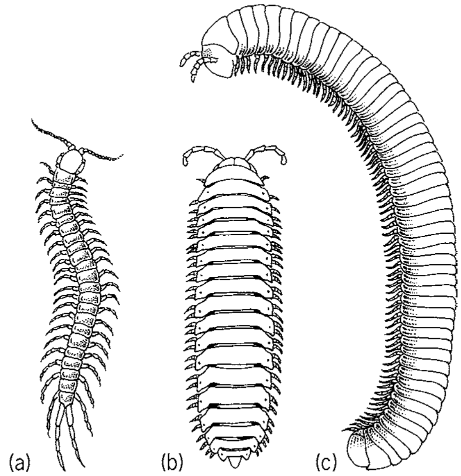


Fig. 13. Chilopoda and Diplopoda. (*a*) Chilopod, or centipede (*Scolopocryptops*; formerly *Otocryptops*). (*b*) Polydesmoid diplopod. (*c*) Julid diplopod. (*After R. E. Snodgrass, A Textbook Anatomy, Cornell University Press, 1952*)

limbs of various forms known as the first and second maxillae. The trunk limbs, which vary in number in the several mandibulate groups, are used mostly for walking or swimming, but some are commonly specialized for other purposes. The trunk may consist of only a head and body, or it may be divided into head, thorax, and "tail region." This tail can be either a limbless abdomen or a limb bearing pleon. *See* MANDIBULATA.

Myriapoda. The myriapods include the common centipedes and millipedes, and some other groups of a similar nature. There are four rather distinct living classes: Chilopoda, Diplopoda, Pauropoda, and Symphyla. Most of them develop after hatching by anamorphosis, in which the young hatch with only five to seven body segments and three pairs of legs. Further development takes place in a subterminal zone of growth, in which new segments and legs are generated and added to those already present. Two groups of the chilopods, however, develop after hatching by epimorphosis. Their development (in terms of segment number) is completed within the egg; they increase only in size after hatching. The adult myriapod consists of only two parts: a head and a segmented body.

Chilopoda. The chilopods are often called centipedes (**Fig. 13*a***). Though they include four groups of species, these groups differ mainly in the number of body segments, which varies from 15 to 100 or more. In some species, the number of tergal plates is less than the number of segments. Except for the house centipede that prefers human dwellings, they all live in similar places in the ground, under rocks or logs, or beneath loose bark. The appendages of the first body segment are developed as poison fangs that are directed forward beneath the head; a venom gland opens on the tip of each fang. The other appendages, except those of the reduced last two segments, are legs. Centipedes move quickly, as one would expect of a rapacious obligate carnivore.

Eyes, when present, are in lateral clusters that may look like compound eyes but do not have a compound-eye structure. They detect prey not so much by sight, however, as by tactile detection of vibrations. The reproductive openings are at the posterior end of the body. *See* CHILOPODA.

Diplopoda. The diplopods, or millipedes, differ from all other arthropods in that each body segment, except for the first three or five, bears two pairs of legs (Fig. 13*b* and *c*)—thus their name, Diplopoda. The body is either flattened somewhat or cylindrical and is ideally suited to bulldoze through leaf litter and decaying woody material that they consume for food. The mouthparts consist of a pair of mandibles with a large flat lobe, known as the gnathochilarium, forming a lower lip. They respire by means of tracheae arising from small ventral plates that carry the legs. The reproductive openings of each sex are for the most part on the third body segment. However, in the males of the common diplopods, one or both pairs of legs on the seventh segment (called the gonopods) are modified to receive the sperm from the genital outlets on the third segment and transfer it to the sperm receptacles on the third segment of the female. One group of diplopods have the gonopods at the rear end of the body, but these posterior gonopods are used for holding the female, whereas the male uses his mandibles to transfer the sperm. *See* DIPLOPODA.

Pauropoda. These are minute creatures, having generally not more than 12 body segments and 9 pairs of legs (**Fig. 14**). They are distinguished by their branched antennae. The mouthparts consist of a pair of simple undivided mandibles and a gnathochilarium resembling that of the diplopods. The genital ducts open on the ventral plate of the third body segment. *See* PAUROPODA.

Symphyla. The symphylans look like small, white, soft-bodied centipedes with long antennae, 12 pairs of legs, and two tapering unjointed appendages at the rear end of the body (**Fig. 15**). A characteristic feature of the symphylans is the presence of

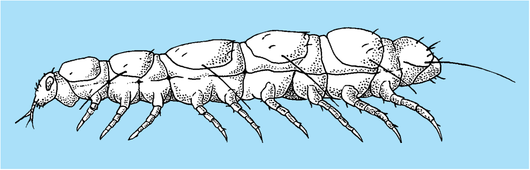


Fig. 14. Pauropod (*Pauropus silvaticus*). (*After R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*)

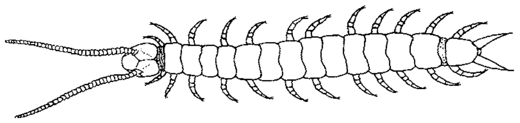


Fig. 15. Symphylan, *Scutigerella immaculata*, with long antennae, 12 pairs of legs, and two unjointed appendages. (*After R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*)

small styli on the undersurface of the body medial to the legs. They live mostly in damp places under stones, in rotting logs, or in the ground, but one species, known as the garden centipede, sometimes becomes a destructive pest of garden crops and greenhouse plants. The mandibles resemble those of the diplopods in being composed of two parts, of which the distal ones are the functional jaws. The head, however, bears two pairs of maxillae, and those of the second pair, as in the insects, are united in a labiumlike organ. The reproductive openings are ventral on the fourth body segments. *See* SYMPHYLA.

Hexapoda. Insects and their allies may or may not have wings, but are easily recognized by the division of the body into head, thorax, and abdomen and by the presence of only three pairs of legs carried by the thorax (**Fig. 16***a*). The head always includes the segments of the mandibles and two pairs of maxillae. Most adult insects have both compound and simple eyes. There is only one pair of antennae. The second maxillae are united in a single organ, the labium. The mouthparts undergo endless modifications in different orders by which they are adapted to biting, chewing, piercing, and sucking, according to the nature of the food. Respiration is by means of tracheae. The reproductive openings in each sex are near the end of the abdomen.

Entognatha. This class of hexapods lacks wings and is characterized by having the jaws partially enclosed within the head capsule. There are three groups; none of them is large, either in species numbers or in body size (Fig. 16*b*). The entognaths are generally only a few millimeters in size and among the cryptofauna that inhabits leaf litter and rooting vegetation.

Insecta. There are both wingless and winged forms in this class. The wingless insects include those such as silverfish and bristletails (Fig. 16*c*). The winged insects, or Pterygota, include some groups that have lost their wings. The pterygotes not only make up the vast majority of the Insecta but also represent the dominant form of animal life in terms of existing species, with probably 900,000 described forms (as a conservative estimate). The reason for their abundance and diversity rests on their ability to disperse by flight, small size, very specialized feeding habits and ecological requirements, and the niche segregation of larvae from adults. *See* INSECTA.

Crustacea. The familiar crustaceans are the aquatic shrimps, crayfish, lobsters, crabs, and the terrestrial sow bugs. The oceans and fresh waters, however, swarm with lesser members of the clan, which constitute the basic food for most of the larger animals. *See* CRUSTACEA.

While crustaceans share many characters with other arthropod groups, it is the crustacean head (cephalon) that is distinctive. The cephalon consists of five, usually fused segments. Each cephalic somite has a pair of appendages, which include the antennules and antennae (also referred to as first and second antennae), mandibles, maxillules, and maxillae (the last two also referred to as first and second maxillae). It is the first antenna that is the homologue of the antennae of other arthropods. One problem with studying the phylogenetic relationships of crustaceans is that while the body is typically divided into three distinct regions—cephalon, thorax (trunk), and abdomen (pleon)—the gonopores can be located in a number of places. In other groups of arthropods, the location of the gonopores is fixed for each group, and this variability in crustaceans suggests that the subphylum Crustacea is actually composed of several disparate groups. The crustaceans can be subdivided into five classes, though there is much doubt whether all of these are true monophyletic groups, that is, with a single shared origin: Cephalocarida, Remipedia, Branchiopoda, Maxillopoda, and Malacostraca.

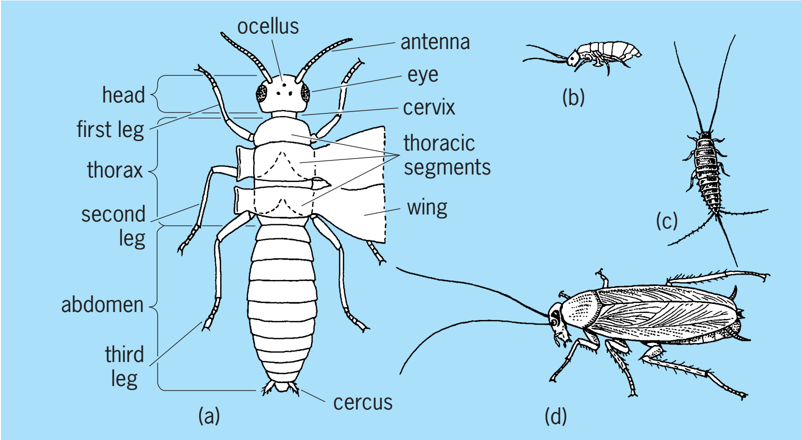


Fig. 16. Insecta. (*a*) Diagram of insect structure (*after R. E. Snodgrass, Principles of Insect Morphology, McGraw-Hill, 1935*). (*b*) Collembolan. (*c*) Thysanuran. (*d*) Pterygote, cockroach, *Periplaneta* (*parts c and d after R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*).

Cephalocarida and Remipedia. The cephalocarids are a group of very small (2.0–3.6 mm) bottom-dwelling crustaceans, which at one time were regarded as the most primitive of crustaceans. The structure and function of the serially homologous appendages was the focal point of the cephalocarid theory, which postulated that evolution in the Crustacea came about through specialization, simplification, and reduction of cephalocarid-like appendages (**Fig. 17***a*). However, the 1981 discovery by cave divers of the Remipedia (Fig. 17*b*) has challenged this hypothesis. The lack of trunk regionalization in remipedes and their simple, biramous appendages have caused some carcinologists to propose that the Remipedia more correctly reflect the crustacean ancestral type. *See* CEPHALOCARIDA; REMIPEDIA.

Branchiopoda. The branchiopods for many years have been divided into four major groups. Although the Anostraca (fairy shrimps; **Fig. 18***a*) and Notostraca (tadpole shrimps; Fig. 18*b*) have been retained as monophyletic taxa, the names Conchostraca (clam shrimps) and Cladocera (water fleas) have been replaced. The former has been split into the orders Laevicaudata and Spinicaudata, which both have bivalved carapaces. The cladocerans, while still a monophyletic group, are subdivided into

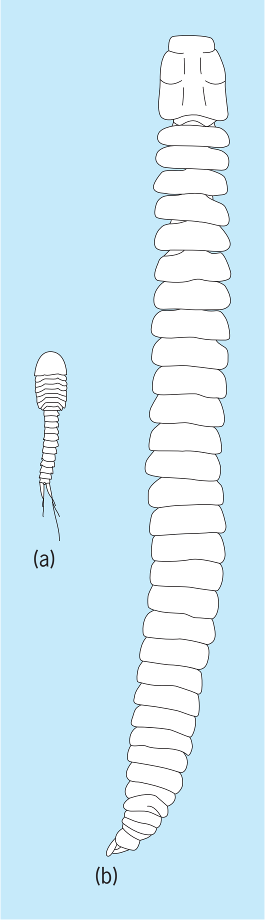


Fig. 17. Two primitive living crustaceans suggested as representing something close to an ancestral type. (*a*) Cephalocarid. (*b*) Remipede.

the orders Ctenopoda, Anomopoda, Onychopoda, and Haplopoda (Fig. 18*c–f*). *See* BRANCHIOPODA.

Maxillopoda. The class Maxillopoda has not received universal acceptance among carcinologists. One of the principal threads of continuity seen throughout the group is the ways in which the maxillae and anterior thoracic appendages are used in food gathering and locomotion. However, there are probably at least two quite separate groups included in the maxillopodans. One group, the largest by far, generally conform to a body plan with five segments in the head, seven segments in the thorax, and four segments in the abdomen. This would include the copepods, thecostracans, and perhaps some of the ostracodes. Maxillopoda contains the subclasses Thecostraca, Tantulocarida, Ostracoda, Copepoda, Branchiura, and Mystacocarida. *See* MAXILLOPODA.

Copepoda (**Fig. 19**) include some 8000 to 10,000 species; they are by far the largest taxon assigned to the Maxillopoda. Copepods are found in virtually all aquatic habitats and a few moist terrestrial habitats, but a large number are parasitic and infest a vast variety of vertebrate and invertebrate hosts. *See* COPEPODA.

The second largest maxillopodan subclass is Thecostraca (**Fig. 20**), which includes the stalked and sessile barnacles (order Thoracica; Fig. 20*b–d*). The cirripedes also include two parasitic orders, Rhizocephala and Ascothoracica (collectively known as the cirripedes). The rhizocephalans are noted for their "parasitic castration" of crabs, but they have been found to infest a variety of crustacean hosts. As adults, they completely lack any similarity to other crustaceans, being merely a bag of gonads attached to a root system that permeates the host body. The ascothoracicans (Fig. 20*a*) parasitize coelenterates and echinoderms. *See* CIRRIPEDIA.

The Ostracoda is one of the largest of crustacean taxa. It is divided into two subgroups: the Myodocopa, in which the antennae are strongly developed for swimming, and the Podocopa, in which the antennae are specialized for walking. These are generally minute forms that occur in just about every aquatic environment. *See* MYODOCOPA; OSTRACODA; PODOCOPA.

Tantulocarida is a small subclass comprising ectoparasites of other crustaceans and are less than 0.3 mm in body length. However, they appear to be related as a sister group to the thecostracans. *See* TANTULOCARIDA.

The remaining maxillopodan subclasses, Mystacocarida and Branchiura, represent relatively few species. Both groups have gonopores on the fourth segment of the trunk. All mystacocarids are small, free-living meiofaunal animals with a preference for wave-washed, sandy, intertidal beaches; they have a small limbless abdomen extending posteriorly from the gonopores. The branchiurans, or fish lice, are ectoparasites of the gills of fish and occasionally amphibians. The branchiurans now contain another group, Pentastomida, which was long thought to be a separate phylum. However, the ultrastructure of their sperm and DNA sequence analysis indicates they are closely related to the fish lice. The pentastomids are all parasites of the respiratory systems of vertebrates. However, the ultrastructure of the sperm as well as DNA molecular sequence data clearly ally the pentastomids with the argulid branchiurans. *See* BRANCHIURA; MYSTACOCARIDA; PENTASTOMIDA.

Malacostraca. When most people think "crustacean," it is probable that it is a malacostracan that they have in mind, such as a shrimp, lobster, crayfish, or crab. The Malacostraca constitute a clearly definable, monophyletic group. However, it is very diverse and is subdivided into three subclasses: Phyllocarida, Hoplocarida, and Eumalacostraca, the last with the superorders Syncarida, Peracarida, and Eucarida. The most distinctive characters of the malacostracans include eight thoracic segments, each with a pair of appendages (pereiopods), and a pleon with six pairs of appendages. The first five pleopods are biramous limbs used in swimming, while for the hoplocaridans and eumalacostracans the sixth pair (uropods), together with the telson, form the tail fan. The female genital openings are invariably on the sixth thoracic segment, while male genital openings are on the eighth. Females may be provided with sperm receptacles, and males often have the first two

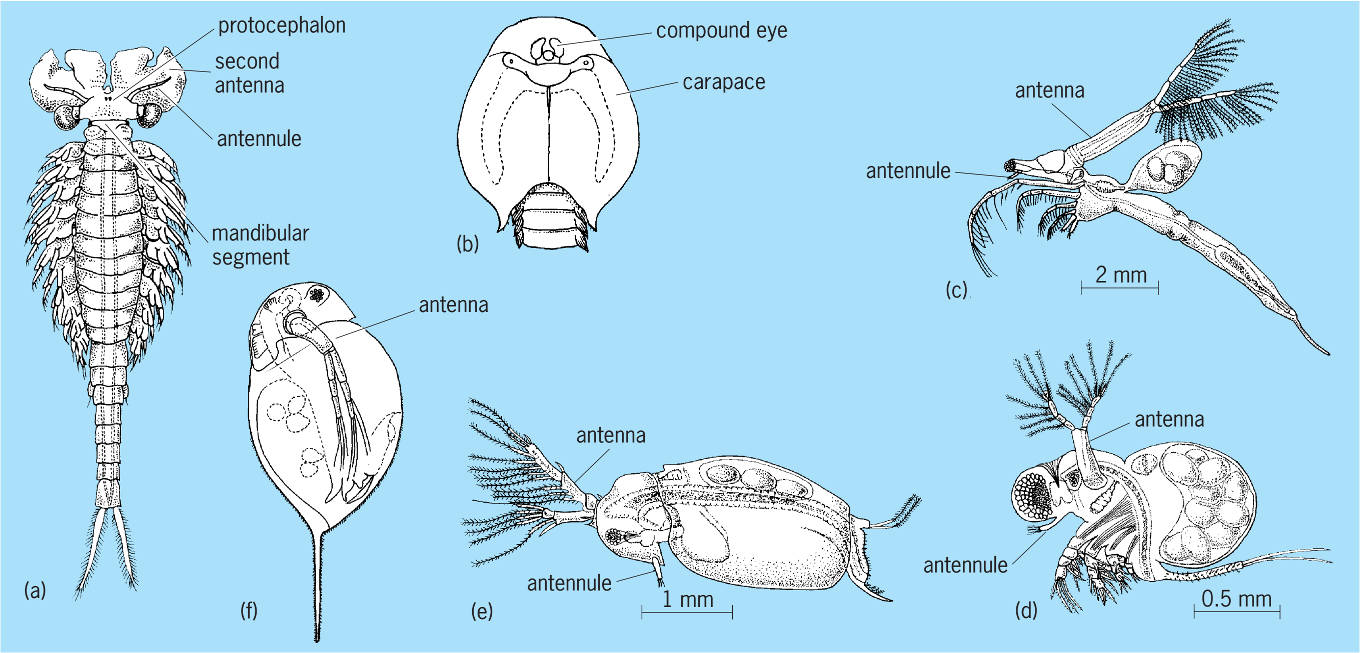

protocephalon
second antenna
antennule
mandibular segment
(a)
compound eye
carapace
(b)
antenna
antennule
(c)
2 mm
antenna
(f)
antenna
antennule
(e)
1 mm
antenna
antennule
(d)
0.5 mm

Fig. 18. Examples of branchiopods. (*a*) Anostraca (*Branchipus*). (*b*) Notostraca (*Triops*). (*c*) Hoplopoda (*Leptodora*). (*d*) Onychopoda (*Polyphemus*). (*e*) Ctenopoda (*Sida*). (*f*) Anomopoda (*Daphnia*). (*Parts c–f after W. T. Edmondson, ed., Freshwater Biology, John Wiley and Sons, 1966*)

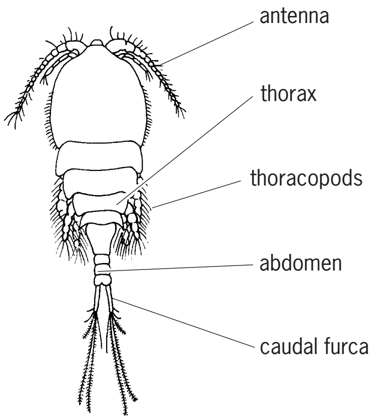

antenna
thorax
thoracopods
abdomen
caudal furca

Fig. 19. *Cyclops*, a copepod.

pairs of pleopods modified to serve in sperm transfer (gonopods). *See* MALACOSTRACA.

The most primitive of the Malacostraca are the Phyllocarida, the only group of higher crustaceans to have the carapace modified into a bivalve shell. Extant representatives of the Hoplocarida are the highly specialized and predatory stomatopods, or mantis shrimp. Within the Eumalacostraca, the simplest body plan is exhibited by the syncarids, who lack a carapace (**Fig. 21**), while most other forms generally possess a carapace to some degree (**Fig. 22**). These latter include the eucarids, that is, the lobsters, crabs, shrimp and their allies, and the peracarids, an extremely diverse group among which are some of the few crustaceans adapted to true terrestrial conditions. *See* EUCARIDA; EUMALACOSTRACA; HOPLOCARIDA; PERACARIDA; PHYLLOCARIDA; SYNCARIDA.

Fossils

In recent years, studies of the arthropod fossil record have made significant contributions to our understanding of the origin and early evolution of the subphyla.

Cheliceriformes. The cheliceriform arthropods have a fascinating record. For example, while living and many fossil Pycnogonida (or sea spiders) lack an abdomen, a few Paleozoic fossils actually possess a long tubular abdomen. Furthermore, within the fossil record of Chelicerata we can note the presence of a variety of the eurypterids and scorpions, horseshoe crabs, and true spiders and their allies.

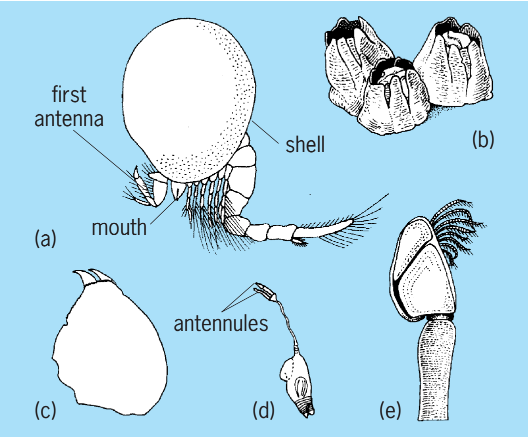

first antenna
shell
(b)
(a)
mouth
antennules
(c)
(d)
(e)

Fig. 20. Some types of Cirripedia. (*a*) Ascothoracica (*Synagoga*). (*b*) Thoracica (*Megabalanus*). (*c*) Acrothoracica (*Lithoglyptes*), female and (*d*) male. (*e*) Thoracica (*Lepas*). (*Parts c and d after J. T. Tomlinson and W. A. Newman, Lithoglyptes spinatus, a burrowing barnacle from Jamaica, Proc. U.S. Nat. Mus., 112:517–526, 1960*)

Eurypterids, with lengths of up to 6 ft (2 m), were the largest known arthropods. In addition, there are many fossils of the class Chelicerata that have no living analogs; these include the chasmataspids and some peculiar proto-chelicerates.

Myriapoda and Insecta. Although the insects and myriapods have more species than any other animal group, their fossil record is relatively poor. The earliest insects are Devonian wingless forms, but most insect fossils are little more than wings. The Late Paleozoic insect fauna was composed of so-called archaic types that—with a few notable exceptions such as cockroaches (Fig. 16*d*) and dragonflies—went extinct.

Crustacea. Most crustacean classes are small to microscopic in size and thus have a poor to modest fossil record. However, the fossil record of the crustaceans has greatly improved in the past two decades. The malacostracans (crabs, lobsters, shrimp, and their allies) are known from abundant fossils from the early Paleozoic to the present. Many of the malacostracan groups that were common in the late Paleozoic either went extinct in the Permian, or are today restricted to geographic or ecologic refuges, such as the southern continents, which once formed a single supercontinent called Gondwanaland, or groundwater habitats. However, the most noteworthy fossil arthropods are microscopic forms from the Cambrian of Sweden, in which minute details are preserved and which lend themselves to rather detailed stories as to the origin and early evolution of the group.

Other forms. Finally, several other "near-arthropods" have interesting fossil records, especially in the Cambrian. The Burgess Shale and similar deposits in China and Greenland contain peculiar animals that were once classified among the trilobitomorphs, but that researchers now realize may not even be arthropods *sensu stricto* (in the strict sense). These animals, such as the unique creature *Anomalocaris*, have wormlike bodies but possess one or more arthropod-like structures such as compound eyes or jointed limbs. The Onychophora, generally conceded to represent proto-arthropods, have a very limited fossil record; only one species is known from the Pennsylvanian Mazon Creek concretion faunas of Illinois, but a host of truly strange forms are known from the Cambrian. Finally, there are several fossil groups that are definitely arthropods, but that do not easily fit into the four major groups cited earlier. These include unusual creatures such as the euthycarcinoids and the thylacocephalans; much more infromation about them is needed before their relationships to other arthropods can be judged.

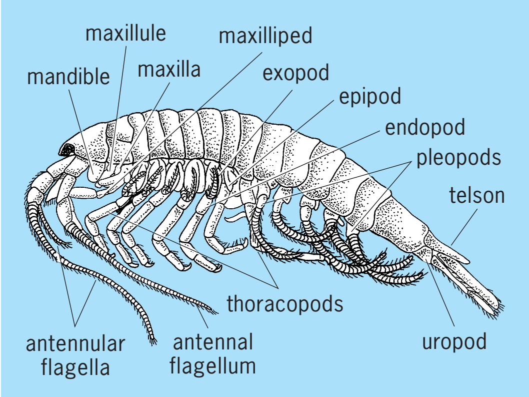

maxillule
maxilliped
mandible
maxilla
exopod
epipod
endopod
pleopods
telson
thoracopods
antennular flagella
antennal flagellum
uropod

Fig. 21. Primitive malacostracan, *Anaspides tasmaniae* (Anaspidacea). (*After P. A. McLaughlin, Comparative Morphology of Recent Crustacea, W. H. Freeman, 1980*)

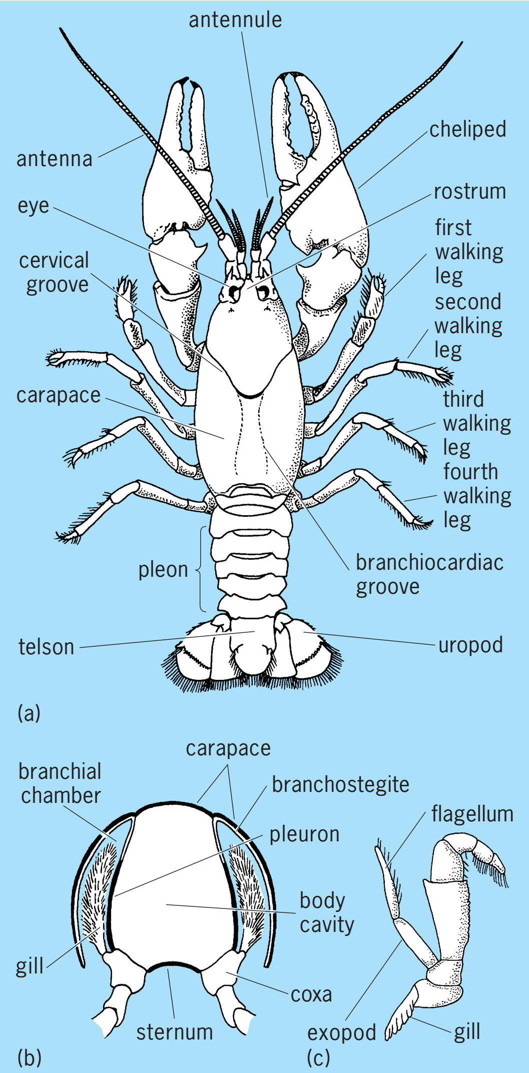

antennule
antenna
cheliped
rostrum
eye
cervical groove
first walking leg
second walking leg
third walking leg
fourth walking leg
carapace
pleon
branchiocardiac groove
telson
uropod
(a)
carapace
branchial chamber
branchostegite
pleuron
flagellum
body cavity
gill
coxa
sternum
exopod
gill
(b)
(c)

Fig. 22. Decapoda. (*a*) *Cambarus*, a fresh-water crayfish. (*b*) Cross section of thoracic segment of a decapod. (c) Third maxilliped. (*After R. E. Snodgrass, A Textbook of Arthropod Anatomy, Cornell University Press, 1952*)

Frederick R. Schram; Jerome C. Rozen, Jr.;
Patsy A. McLaughlin

Bibliography. R. C. Brusca and G. Brusca, *The Invertebrates*, 2d ed., 2003; J. Cracraft and M. J. Donaghue (eds.), *Assembling the Tree of Life*, 2004; R. A. Fortey, *Trilobite! Eyewitness to Evolution*, 2000; R. A. Fortey and R. H. Thomas (eds.), *Arthropod Relationships*, 1998; S. Koenemann and R. A. Jenner (eds.), *Crustacea and Arthropod Relationships*, Taylor and Francis, 2005; F. R. Schram, *Crustacea*, 1986.

Artichoke

Cynara scolymus, a herbaceous perennial plant, in the family Compositae; also called globe artichoke. Its origin is in the Mediterranean region. Artichoke requires a mild winter and cool summer with fog and little bright sunshine. It is a delicacy in Europe, Africa, and North and South America. Artichoke is also a medicinal plant; it is rich in the cynarin and orthophenol constituents. Of the total acres planted worldwide, approximately 56%, 14%, and 12% are planted in Italy, France, and Spain, respectively.

The marketable portion of the plant, the so-called bud, is actually the immature flower head, made up of numerous closely overlaid bracts or scales (see **illus.**). The edible portion consists of the tender bases of the bracts, the young flowers, and the receptacle or fleshy base upon which the flowers are borne. The bud can be various shapes, from round to oblong to flat, and the color can be light green to dark green, often with purple or red.

Green Globe artichoke. (*Burpee Seeds*)

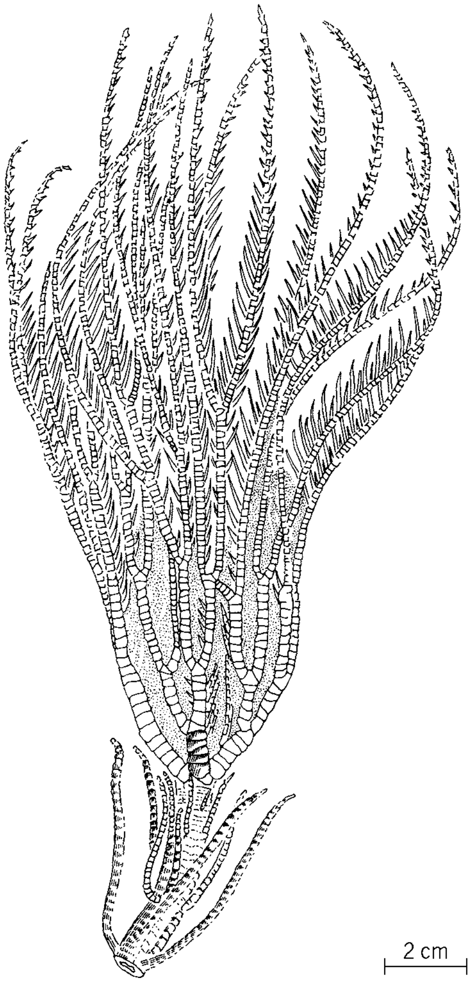
Artichoke is grown readily from seed, but the seedlings are highly variable and plants tend to throw back to the wild; hence, vegetative propagation by means of stumps (rootstocks), offshoots, or ovoli is recommended. The latter two are believed to produce earlier crops. The plant grows to a height of 3 or 4 ft (0.9 or 1.2 m), and sends up seasonal shoots from a permanent crown—as many as 12 or more in plants 4–5 years old. Each shoot forms a cluster of large basal or rosette leaves, from the center of which the stem grows. Buds are produced terminally on this elongated stem and on the lateral branches. If the buds are not removed, they develop into purple-centered thistlelike flowers and viable seeds are produced.

In the United States, artichokes are grown in the Pacific Coast area between south San Francisco and Los Angeles, mainly Monterey County, on about 10,000 acres (4050 hectares). The production per acre fluctuates according to weather conditions.

Aly M. Ibrahim

Articulata (Echinodermata)

The only surviving subclass of the Crinoidea, differentiated during Triassic times. The calyx is dicyclic, but considerable reduction of the infrabasals and basals may occur. The uniserial arms bear pinnules and usually branch, and the arm retains its movable articulation with the radial plate, despite the incorporation of the lower brachial ossicles into the calyx. Extant stalked forms with nodal rings of cirri (*Metacrinus*) are included in the order Isocrinida (see **illus.**). They do not tolerate turbulant waters and live at depths below current action, although they do inhabit shallow water when the conditions



***Metacrinus cyaneus*, a living, articulate stalked crinoid from Bass Strait, at 200 fathoms (1200 ft or 365 m).**

are suitable. These forms may trail the stem during temporary free-swimming phases, using the cirri for reattachment. The feather stars, of the order Comatulida, discard the stem when young, and thereafter remain free, either as swimming animals or as creeping benthic forms. They prefer shallow, clear water, rich in nutrients, and therefore abound on tropical coasts and in polar seas rather than in temperate waters. Four other orders have been defined; one of them includes extinct floating forms. *See* CRINOIDEA; ECHINODERMATA. Howard B. Fell

Bibliography. A. H. Clark, *Monograph of the Existing Crinoids*, U.S. Nat. Mus. Bull. 82, 1915–1967.

Artificial intelligence

The subfield of computer science concerned with understanding the nature of intelligence and constructing computer systems capable of intelligent behavior. Many activities, such as perception, learning, planning, natural language, and understanding, involve intelligent behavior and include an immense diversity of phenomena. Fields such as psychology, linguistics, and philosophy share scientific concern for these phenomena. *See* COGNITION; INTELLIGENCE; INFORMATION PROCESSING (PSYCHOLOGY); LINGUISTICS.

Artificial intelligence is primarily concerned with representations of knowledge and heuristic methods of reasoning using common assumptions and rules of thumb. Two examples of problems studied in artificial intelligence are (1) planning how a robot, or person, might assemble a complicated device or move from one place to another and (2) diagnosing the nature of a person's disease or of a machine's malfunction from the observable manifestations of the problem.

Approach. Approaches of artificial intelligence research can be experimental or theoretical. As in other experimental fields, investigators build devices (in this case, computer programs) to carry out their investigations. Programs are created to explore ideas about how intelligent action might be attained, as well as to test hypotheses about concepts or mechanisms involved in intelligent behavior to confirm theories.

Much of the inspiration for the organization and details of artificial intelligence programs comes from observing and studying the behavior of people in different situations.

Foundations. The four foundations of artificial intelligence are representation, search, reasoning, and learning. These four foundations are typical ingredients of any artificial intelligent system and are addressed to some degree in most artificial intelligence research.

Representation. A computer system must have an internal representation of the problem or related knowledge. Examples of representation include the symbolic description of a room for a moving robot and a description of a person with a disease for a classification program. A representation also includes all the knowledge, including basic programs for testing and measuring a structure in question, plus all the programs for transforming the structure into another one in ways appropriate to the task. The representation of a problem typically implies an architecture on which it can be manipulated. Two examples of architectures are the programming language used for the representation (often high-level languages, like LISP) and the database used to store and query knowledge. Changing the representation used for a task can make an immense difference, potentially turning a problem from impossible to trivial. *See* DATABASE MANAGEMENT SYSTEM; PROGRAMMING LANGUAGES.

Search. An important aspect of most artificial intelligence systems is that they use a problem-solving methodology that involves search. Artificial intelligence has gradually built up a stock of general search methods that can be used. In the worst case, there is no information to guide the search and the search is blind. In heuristic search, extra information is used to guide the search. *See* GENETIC ALGORITHMS.

Search methods include blind breadth-first and depth-first (as used in game-tree search); generate-and-test (a sequence of candidates is generated, each being tested for a solution); hill climbing (a measure of progress is used to guide each step); means-ends analysis (the difference between the desired situation and the present one is used to select the next step); impasse resolution (the inability to take the desired next step leads to a subgoal of making the step feasible); planning by abstraction (the task is simplified, solved, and the solution used as a guide); and matching (the present situation is represented as a schema to be mapped into the desired situation by putting the two in correspondence).

Reasoning. Given the representation and knowledge of a problem, many forms of reasoning exist to turn the knowledge into a solution. Examples of reasoning include expert systems (typically rule-based systems developed using human expertise to identify the rules of the problem); case-based reasoning (a database of previous problems and solutions are searched for the closest match to the current problem); deductive and inductive reasoning (problem knowledge is used to deduce a set of possible solutions, or used to build a hypothesis that best explains the existing knowledge and possibly the current problem); and resolution in logic systems (tools used to determine if the current hypothesis or problem is true or not given the facts we know about the universe in which the problem exists). *See* LOGIC; PROBLEM SOLVING (PSYCHOLOGY).

Learning. Most artificial intelligent systems use the ability to adapt, or learn, based on the history, or knowledge, of the system. Learning takes the form of updating knowledge, adjusting the search, reconfiguring the representation, and augmenting the reasoning. Some of the most common learning methods used in artificial intelligence systems are statistical learning (using the number of the different types of historical events to bias future action or to develop inductive hypotheses, typically assuming that events follow some known distribution of occurrence); neural networks (the updating of weights in a network, where networks are first trained on existing knowledge and then applied to the problem, outputting a value that can be interpreted as a solution); and reinforcement learning (the actions taken by an artificial intelligent system are rewarded or penalized based on their ability to solve the problem more or less accurately). *See* NEURAL NETWORK.

Examples. The following are examples of research in artificial intelligence.

Games. Games are an important application area in artificial intelligence. The initial efforts were in writing general problem-solving programs that would learn how to play games. After discovering that learning was not sufficient to produce game-playing programs that could compete against moderately good human players, the emphasis changed to writing sophisticated programs that were excellent at playing a specific game.

A good example of this approach is chess programs. In 1997, IBM's Deep Blue chess program beat Garry Kasparov, the world champion. A typical chess program uses a representation of a chessboard to store the current positions of the chess pieces, plus representations of standard chess openings, favorable chess positions, and possibly significant games. The program also stores representations of the possible chess moves. A chess program will often use a search function to find a good move. A chess program will take the current board and try all the program's possible moves (a half-step) and then all the possible responses that the opponent could make (the second half-step). The sequence of two such half-steps is a called a ply. A simple chess-playing program might investigate two or three plys ahead. Since the number of such positions in a chess game will be huge, the program may have an evaluation function (heuristic) for the "goodness" of a possible set of positions. This function will be used to decide which possible moves to investigate. The evaluation function may be used to eliminate (or prune) low-value (to the computer) positions and to expand the high-value positions to more depth.

Recently, efforts have been made to refocus the game-playing research back to general game players which accept the game rules as input and can play a variety of games, instead of programs that are carefully designed to play just one game.

Natural language processing. Extracting information from natural language input is an important capability, since so much human knowledge is stored in natural language (for example, textbooks) and much interaction (for example, commands and advice) occurs as spoken language. While an easy task for humans, understanding natural language is a very difficult task for computers. One important part of a natural language processor is the representation of the syntax and semantics of the language. Computer science has learned how to deal with formal languages effectively. However, natural languages inherently involve ambiguity. Approaches to natural language processing often try to build databases of previous examples and meanings of natural language, from which statistically driven hypotheses can be generated and applied to new examples. A good example of natural language processing is found in audio data-entry systems, such as those used by telephone-based help systems. Using a restricted set of expected word utterances at a given time, the system compares the acoustical signatures of the users spoken responses to those it expects to hear. By repeating the understood response to the user and asking for a confirmation, the system can update its database of acoustical signatures for that response. In doing so, the system is able to learn from specific users and all users in general. *See* NATURAL LANGUAGE PROCESSING.

Perception. Perception is the extraction, from a sensory input, of an internal representation of the external world suitable for intelligent processing. Though there are many types of sensory signals, the richest sensory input is vision. A common task in vision-perception programs is object recognition. The perception program has a representation of the objects to be identified. These representations of objects can often be sets of features that distinguish each object from other objects. Features can be edges, surfaces, or shapes. The object-recognition program then tries to match the features in the input with the objects and to identify which objects are present. However, this task is very difficult. Most objects in the real world are three-dimensional. From different angles, an object can look very different. Human vision is good at compensating for the different angles; computer programs are not. *See* COMPUTER VISION; PERCEPTION.

A different approach to object recognition is the use of artificial neural nets (networks). Each object to be identified is handled by a neural net. The neural net is trained to recognize its object. With sufficient training, the set of neural nets will be able to identify which of the set of objects appears in the scene. This approach is very useful in situations such as the military application of recognizing enemy vehicles in surveillance images.

Autonomous mobile robotics. Popular literature has many examples of autonomous robots that make human life easier by doing those tasks that are repetitious, dangerous, or unpleasant. Although there have been some successes, such as rovers on Mars and industrial robots that help in manufacturing tasks, there are many more possibilities in which autonomous robots would be very useful. Mobile robots are devices that have means of locomotion (wheels), sensors (sonars, lasers, and cameras), communications (wireless), and often actuators (arms or grippers).

A typical autonomous robot would have a representation of its world, possibly a map that shows the locations of walls, doors, obstacles, and other features of interest. Localization, the task of determining the robot's exact position in such a representation, is a major problem in robotics. Probabilistic methods have been used very successfully in matching sensor readings to the representation to determine the most likely position of the robot. An autonomous robot will have a goal to be achieved, such as exploring a disaster area looking for victims. A major task of the program will be planning the actions for achieving the goal. This will include path planning to search the whole area, sensing and moving around obstacles, and identifying possible victims. *See* ROBOTICS.

Expert systems. Even when no firm scientific base exists to make a decision, extensive experience enables people to accomplish many tasks successfully. A class of artificial intelligence programs, called expert systems, attempt to exhibit equivalent performance by acquiring and incorporating the same knowledge that human experts have. Many attempts to apply artificial intelligence to medicine, government, and other socially significant tasks take the form of expert systems. Even though the emphasis is on knowledge, all the standard ingredients are present, including the representation of the task, representation of the knowledge, and reasoning or search methods. *See* EXPERT SYSTEMS.

A typical example of an expert-system task is a medical diagnosis program, which would have "if-then" rules that a doctor would use. The program might prompt for answers to questions very similar to what a doctor would ask. With a sufficient number of correct if-then rules, the program might be able to determine a correct diagnosis. One advantage of the expert-systems program is that it can use many more if-then rules than a human could keep in mind. Very rare diseases or conditions could be covered with the rules.

Obtaining and codifying implicit expert knowledge in if-then rules is a critical aspect of developing artificial intelligence expert programs. That entire bodies of knowledge can be sets of active if-then rules, even though they did not yet form a coherent scientific theory, has been an important discovery. An important variation of the expert-systems program is the inclusion of fuzzy logic, which allows uncertainty to be handled explicitly. *See* FUZZY SETS AND SYSTEMS.

In careful tests, a number of expert systems have shown performance at levels of quality equivalent to or better than average practicing professionals on the restricted domains over which they operate. Nearly all large corporations, and many smaller ones, use expert systems. A common application is to provide technical assistance to persons who answer customers' trouble calls. Computer companies use expert systems to assist in configuring components from a parts catalog into a complete system that matches a customer's specifications, a kind of application that has been replicated in other industries by tailoring assembled products to customers' needs. Troubleshooting and diagnostic expert systems programs are also commonplace.

Expert systems have sparked important insights in reasoning under uncertainty, causal reasoning, reasoning about knowledge, and acceptance of computer systems in the workplace. They illustrate that there is no hard separation between pure and applied artificial intelligence, since finding what is required for intelligent action in a complex applied area makes a significant contribution to basic knowledge.

Scope and implications. Research in artificial intelligence explores the full range of intellectual tasks. In addition to the subject areas mentioned, significant work has been done on puzzles and reasoning tasks, induction and concept identification, symbolic mathematics, theorem-proving in formal logic, natural language understanding and generation, vision, robotics, chemistry, biology, engineering analysis, computer-assisted instruction, and computer-program synthesis and verification, to name only the most prominent. As in any developing technology, there occurs both a reinforcement of the basic ideas (about representation, search, reasoning and learning) and the discovery of new mechanisms that both extend existing ideas and reveal limits.

Artificial intelligence has close ties to several surrounding fields. As part of computer science, it plays the role of expanding the intellectual sophistication of the tasks to which computers can be applied. Various subfields, once viewed as part of artificial intelligence, have become autonomous fields, most notably symbolic mathematics and program verification. Work on vision and speech overlaps with the field of pattern recognition in electrical engineering in the concern with efficient signal processing.

Numerous applications of artificial intelligence have been deployed that demonstrate savings in money or time to business, industrial, military, scientific, and health organizations. The success in expert systems, robotics, and perception has precipitated substantial commercial activity, including the formation of many small venture firms. As computers become smaller and less expensive, more intelligence is built into automobiles, appliances, and other machines, as well as computer software, in everyday use. *See* COMPUTER; CONTROL SYSTEMS; DIGITAL COMPUTER; INTELLIGENT MACHINE; SOFTWARE.

Steven M. Gustafson; David A. Gustafson

Bibliography. T. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997; S. Russell and P. Norvig, *Introduction to Artificial Intelligence*, Prentice Hall, New Jersey, 2003; S. C. Shapiro (ed.), *Encyclopedia of AI*, Wiley, New York, 2000.

Artificially layered structures

Manufactured, reproducibly layered structures having layer thicknesses approaching interatomic distances. Modern thin-film techniques are at a stage at which it is possible to fabricate these structures, also known as artificial crystals or superlattices, opening up the possibility of engineering new desirable properties into materials. In addition, a variety of problems in solid-state physics can be studied which are otherwise inaccessible. The various possibilities include the application of negative pressure, that is, stretching of the crystalline lattice; the study of dimensional crossover, that is, the transition from a situation in which the layers are isolated and two-dimensional in character to a situation in which the layers couple together to form a three-dimensional material; the study of collective behavior, that is, properties that depend on the cooperative behavior of the whole superlattice; and the effect and physics of multiple interfaces and surfaces. These structures serve as model systems and as a testing ground for theoretical models and for other naturally occurring materials that have similar structures. For instance, ceramic superconductors consist of a variable number of conducting copper oxide (CuO_2) layers intercalated by various other oxide layers, and therefore artificially layered structures may be used to study predictions of the behavior of suitably manufactured materials of this class. A variety of applications have also been proposed or discovered. Of course, one of the most exciting prospects is the discovery of new, as yet unpredicted phenomena. For a discussion of semiconductor superlattices *see* CRYSTAL STRUCTURE; SEMICONDUCTOR HETEROSTRUCTURES

Preparation. Artificially layered structures have been prepared since the beginning of the century, using mostly chemical methods for deposition. With the advent of sophisticated high-vacuum preparation techniques, there has been a significant increase in activities in this field since the early 1970s. The preparation techniques can be conveniently classified into two groups: evaporation and sputtering. In an evaporation system, two or more sources of particles (which end up being part of the final structure) are aimed at a heated substrate where the artificially layered structure is grown. The rates are precisely controlled by using rate monitors, and the various beams are chopped by using shutters in various configurations. The ultrahigh-vacuum (UHV) version has been commonly designated as a molecular beam epitaxy (MBE) apparatus. The sputtering method relies on bombarding targets of the proper materials with an inert gas, such as argon, thus producing the beams of the various elements. Commonly, in this case, the substrate is held against a rotating, heated table and so moved from one beam to the next. *See* CRYSTAL GROWTH; MOLECULAR BEAMS; SPUTTERING.

Structure. Once the artificially layered structure is prepared, it is necessary to characterize whether the layer structure is stable at the growth temperature. This is of considerable importance, since the interdiffusion of the constituents in many cases eliminates the layered growth. A direct image of the layers in gallium arsenide–aluminum arsenide superlattices, using electron micrography, is shown in **Fig. 1**. One of the most successful methods of characterizing layered growth has been x-ray diffraction. In this method the intensity of x-rays elastically reflected from a sample is measured as a function of incidence angle. The layered growth is indicated by the existence of many superlattice peaks, which are due to the Bragg reflection by the superlattice planes. The spacing of these peaks is related to the thickness of the layers, whereas the amplitude is related to the admixture of one constituent into the other, the interfacial roughness, and the amount of strain present. *See* ELECTRON MICROSCOPE; X-RAY DIFFRACTION.

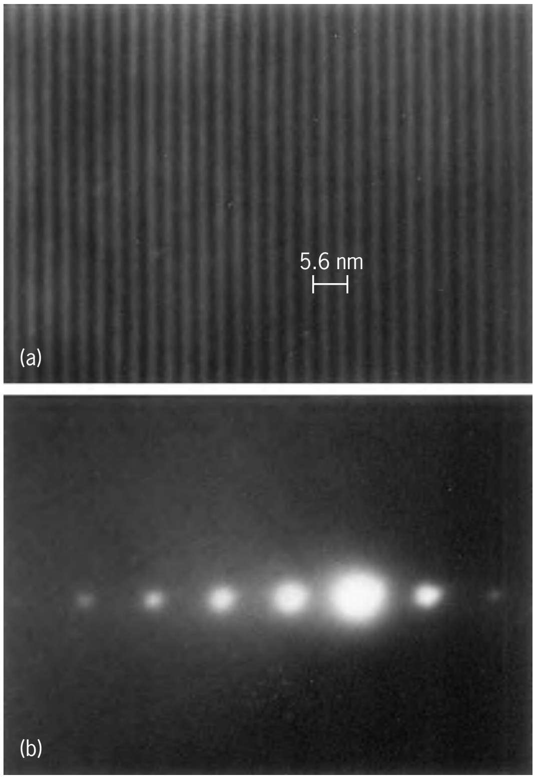


Fig. 1. Ordered superlattice consisting of alternating layers of gallium arsenide (GaAs) and aluminum arsenide (AlAs), each four atomic layers thick, deposited by molecular beam epitaxy. (*a*) Transmission electron micrograph of the superlattice. *(b)* Electron diffraction pattern of the superlattice. New diffraction spots around the central diffraction peak are the result of Bragg scattering from the long periodicity of the superlattice. (*From P. M. Petroff et al.*, *Crystal growth kinetics in* (GaAs)$_n$-*(AlAs)*$_m$ *superlattices deposited by molecular beam epitaxy*, *J. Cryst. Growth*, *44:5–13, 1978*)

Physical properties. The normal-state properties of non-lattice-matched metallic superlattices show anomalies at small thicknesses of the order of 2.0 nanometers periodicity (that is, repeat distance; **Fig. 2**). The lattice expands slightly (approximately 2%) in the direction perpendicular to the layers, the shear elastic constant decreases markedly (approximately 35%), and the temperature coefficient of resistivity changes sign, as in a metal–nonmetal transition.

Artificially layered structures that are superconducting exhibit dimensional crossover, as is shown (**Fig. 3**) beautifully by the temperature dependence of the superconducting critical field. When the superconducting layers are thin and well separated, the critical field shows a characteristic square-root-like behavior, as is expected for two-dimensional superconductors. If the superconducting layers are close together, the layers are strongly coupled and show a linear temperature dependence typical of a three-dimensional material. At intermediate separation the behavior can change from three-dimensional at high temperatures to two-dimensional at low temperatures, a phenomenon known as dimensional crossover. Since the high-temperature ceramic superconductors are also layered, many of their properties are similar to those of artificially layered superconductors. *See* SUPERCONDUCTIVITY.

Because of a short-range coupling mechanism that is not yet understood, the ferromagnetic layers in ferromagnetic-normal superlattices order with antiparallel (antiferromagnetic) alignment. As a consequence, the resistivity can exhibit changes as large as 150% with applied field (giant magnetoresistance). This amount is much larger than the ordinarily observed magnetoresistance of 2–3% in magnetic materials. In some cases a variety of magnetic coupling mechanisms have been observed across nonmagnetic layers, giving rise to new magnetic structures (in dysprosium-yttrium superlattices, for instance) and collective spin waves in molybdenum-nickel superlattices. *See* MAGNETORESISTANCE; MAGNON.

Applications. Artificially layered structures are especially useful for the construction of mirrors for soft x-rays since there are no suitable, naturally occurring crystals for this purpose. This application was one of the main motivations for research on artificially layered structures. Mirrors and polarizers for neutrons have also been manufactured and are currently used in the United States and Europe.

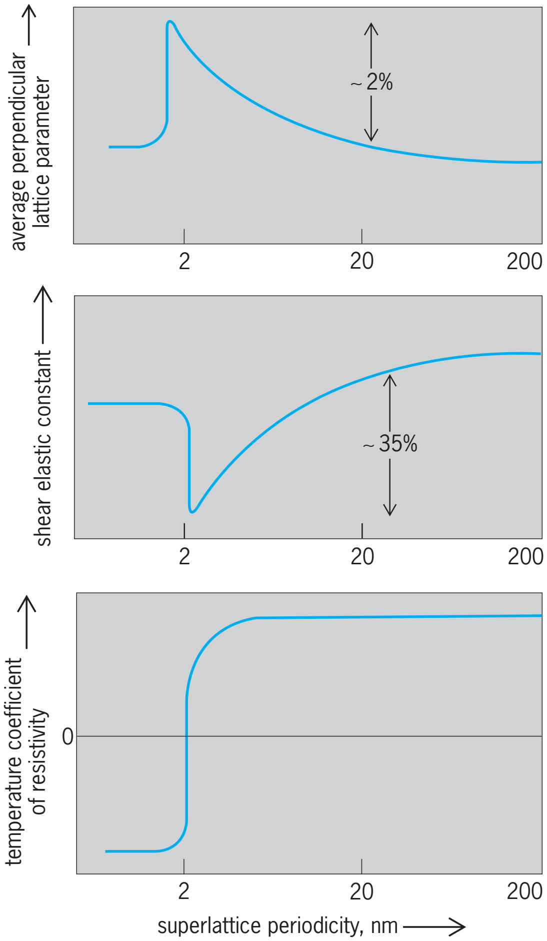


Fig. 2. **Average perpendicular lattice parameter, shear elastic constant, and temperature coefficient of resistivity versus superlattice periodicity. (*After M. R. Khan et al., Structural, elastic and transport anomalies in molybdenum-nickel superlattices, Phys. Rev., B27:7186–7193, 1983*)**

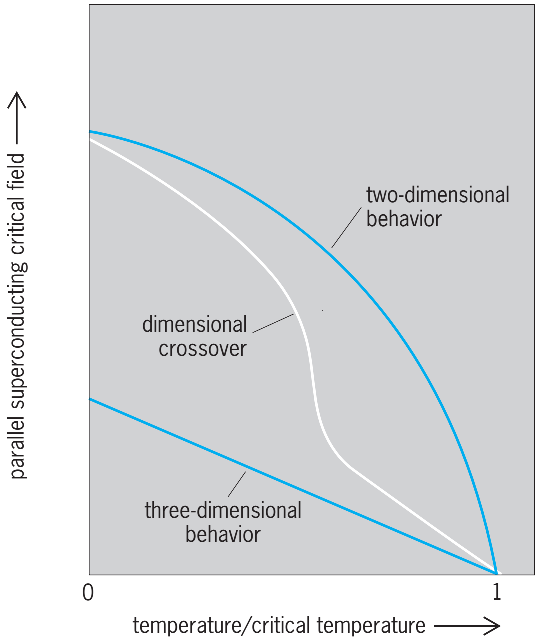


Fig. 3. **Parallel superconducting critical field versus temperature at various layer thicknesses. (*After S. T. Ruggiero, T. W. Barbee, and M. R. Beasley, Superconductivity in quasi-two-dimensional layered deposits, Phys. Rev.Lett., 45:1299–1302, 1980*)**

Superlattices with zero temperature coefficient of resistivity are useful as resistor material, and high-critical-field-magnet tapes using superconducting-insulator superlattices have been proposed. Magnetic superlattices that exhibit giant magnetoresistance are incorporated into magnetoresistive recording heads. *See* ELECTRICAL INSULATION; NEUTRON OPTICS; X-RAYS.

Ivan K. Schuller

Bibliography. M. Bartusiak, Architects in the laboratory, *Discover*, 2:60–61, 1981; P. Dhez and C. Weisbuch (eds.), *Physics, Fabrication and Applications of Multilayered Structures*, 1988; J. D. Dow and I. K. Schuller (eds.), *Interfaces, Superlattices and Thin Films,* 1987; S. Maekawa et al. (eds.), *Metallic Multilayers*, 1993; I. K. Schuller, E. E. Fullerton, and Y. Bruynseraede, Quantitative x-ray diffraction from superlattices, *Mater. Res. Soc. Bull.*, 12(12): 33–38, 1992.

Artiodactyla

An order comprising the even-toed ungulates (hoofed mammals). There are two main radiations: the predominantly omnivorous Bunodontia, including suoids (such as pigs, peccaries, and hippos); and the more herbivorous Selenodontia, including camels and ruminants (such as deer, giraffe, cattle, sheep, and antelope). Artiodactyla contains about 213 living species, making it the fifth most speciose order of mammals (exceeded primarily by rodents and bats, and slightly by carnivores and primates). First known from the early Eocene, artiodactyls have proliferated during the last 55 million years to reach great diversity (especially among the family Bovidae). Their radiation is often contrasted with that of the odd-toed ungulates, or Perissodactyla (horses, rhinos, and tapirs). Artiodactyls are also important for human economy and agriculture, comprising most of the domestic animals, providing milk, wool, and most of the meat supply. *See* PERISSODACTYLA.

Morphology . Artiodactyls are defined by a unique morphology of the ankle joint, possessing a "double-pulley" astragalus. In the general mammalian condition this ankle bone has a single pulleylike articular surface on its dorsal surface (for articulation with the tibia, or shin bone), whereas in artiodactyls it also has a second pulleylike articulation on its ventral surface, forming an articulation with the lower row of ankle bones (the cuboid and the navicular) [**Fig. 1**]. This morphological feature is frequently considered to be a key innovation of artiodactyls, but its precise functional significance is poorly understood.

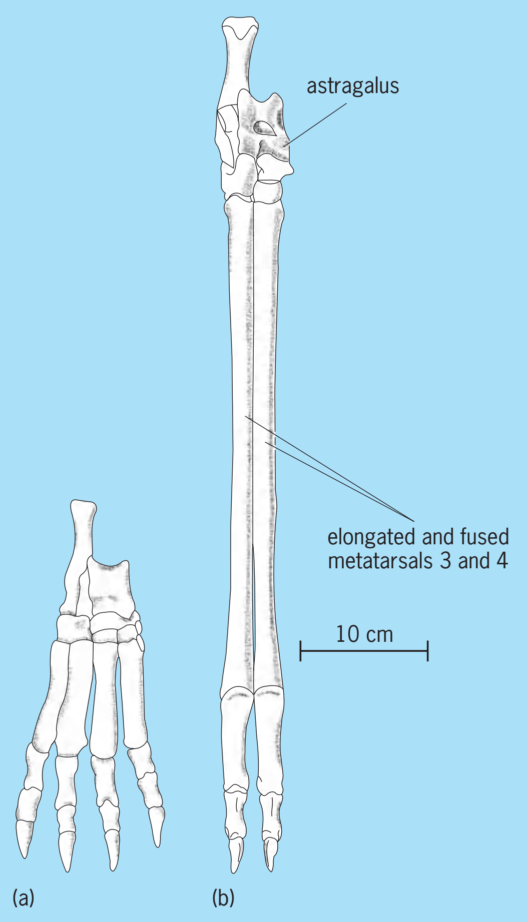


Fig. 1. Representative left hindfeet of artiodactyls, showing the double-pulley astragalus. (*a*) Primitive condition, as in the Oligocene oreodont *Agriochoerus* (although the clawed condition in this animal is a secondary one). (*b*) Derived condition, as in the Miocene camelid *Oxydactylus*. (*After A. S. Romer, Vertebrate Paleontology, 3d ed., University of Chicago Press, 1966*)

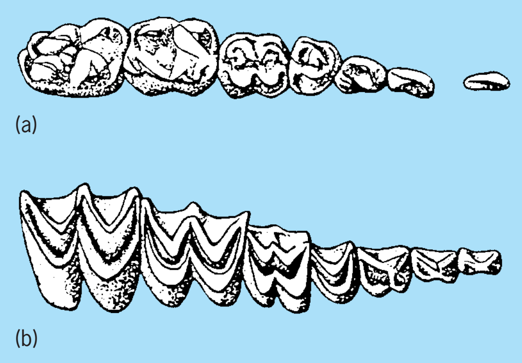


Fig. 2. Representative right upper cheek teeth of artiodactyls. (*a*) Primitive, bunodont condition in *Chleuastochoerus*, a Pliocene suid. (*b*) Derived, selenodont condition in *Merycochoerus*, a Miocene oreodont. (*After A. S. Romer, Vertebrate Paleontology, 3d ed., University of Chicago Press, 1966*)

Artiodactyls also have a paraxonic foot structure, where the axis of limb support passes between the third and fourth digits. The first toe is usually completely lost. The foot posture is usually unguligrade (standing on the most distal phalanx, like a ballerina on point). Some primitive extinct artiodactyls retain a more generalized, digitigrade stance (standing on the digits, like a person on tiptoe). Digitigrady is also secondarily adopted in derived camelids, where the hooves are replaced by a padlike foot. In primitive artiodactyls and in suoids, the foot is four-toed, and the metapodials (hand and foot bones) are short and unfused as in the generalized mammalian condition (Fig. 1*a*). In more derived artiodactyls, digits 2 and 5 have been reduced or lost entirely, and metapodials 3 and 4 are lengthened and fused to form a "cannon bone" (Fig. 1*b*). It is this fusion of the metapodials with digits 3 and 4 free that gives ruminants their cloven-hoofed appearance (in contrast with the single and solid hoof of horses).

In the primitive artiodactyl condition, retained in most living suoids, the cheek teeth have a bunodont crown morphology (**Fig. 2*a***). This is the generalized morphology of low, rounded cusps seen in many omnivorous mammals, including humans. In more derived artiodactyls, which have a herbivorous rather than an omnivorous diet, the cheek teeth are selenodont: the crown pattern is for the individual cusps to be run together into crescentic lophs, one pair on the buccal (outer) side of the each tooth and another pair on the lingual (inner) side (Fig. 2*b*). These lophs provide the teeth with high-standing ridges, better adapted to shred fibrous food than the bunodont teeth, which are adapted to pulp fruits and roots. The primitive condition is for these teeth to be brachydont, or low-crowned, but in some selenodont artiodactyls (especially in grazers) the teeth are hypsodont (high-crowned), which renders them more durable for a highly abrasive diet. *See* DENTITION; TOOTH.

The skulls of suoids and selenodonts also reflect their differences in diet (**Fig. 3**). In ruminants (and also convergently camelids and protoceratids, an extinct related family) the upper incisors and canine are reduced or absent. However, an enlarged upper canine may be retained for display in some hornless males, as seen in living mouse deer and musk

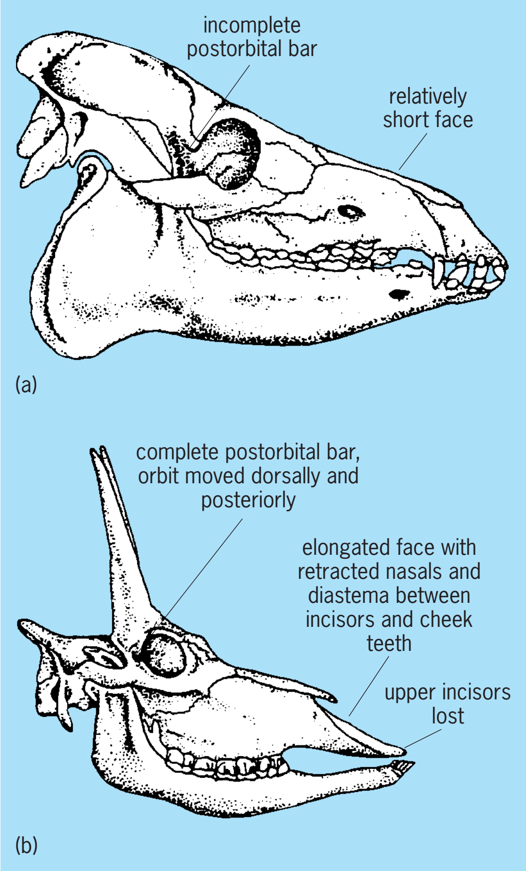

incomplete postorbital bar
relatively short face
(a)
complete postorbital bar, orbit moved dorsally and posteriorly
elongated face with retracted nasals and diastema between incisors and cheek teeth
upper incisors lost
(b)

Fig. 3. Representative artiodactyl skulls. (*a*) Suoid condition, in the Oligocene anthracothere *Bothriodon*. (*b*) Selenodont condition, in the Miocene giraffe *Samotherium*. (*After A. S. Romer, Vertebrate Paleontology, 3d ed., University of Chicago Press, 1966*)

deer (**Fig. 4*a***). The lower canine has become incisiform in shape, and these lower incisors now meet against a horny pad in the upper jaw. Horns or other types of cranial appendages (such as antlers and ossicones) are a typical feature of ruminants, although they also evolved in protoceratids, and a single median horn on the forehead is known in the extinct pig *Kubanochoerus*. Different types of cranial appendages are not homologous in their structure and development, and most likely they evolved independently in the different families (Fig. 4). Horns are usually known only in the males, and are used for intraspecific display and combat. Horns have evolved in the females in species which form year-round associations of mixed sex herds, such as cattle, some grazing antelope (such as the wildebeest), giraffe, and reindeer. *See* ANIMAL COMMUNICATION; TERRITORIALITY.

Selenodont artiodactyls, like other herbivorous mammals, require a gut fermentation chamber to house microorganisms to ferment the cellulose of plant cell walls. Selenodonts have a forestomach site of fermentation, situated before the true stomach. This site comprises three chambers in ruminants (the rumen, omasum, and abomasum), but only two in camels (where a distinct reticulum is absent). Selenodonts differ from other forestomach-fermenting mammals in their remastication of the food, or chewing of the cud. Selenodonts retain their food for a long time and extract the maximal amount of nutrients out of the cellulose. This gives them an advantage where food is of limited quantity, but they are not able to bulk-process food of low quality, in contrast with less efficient hindgut fermenters such as horses.

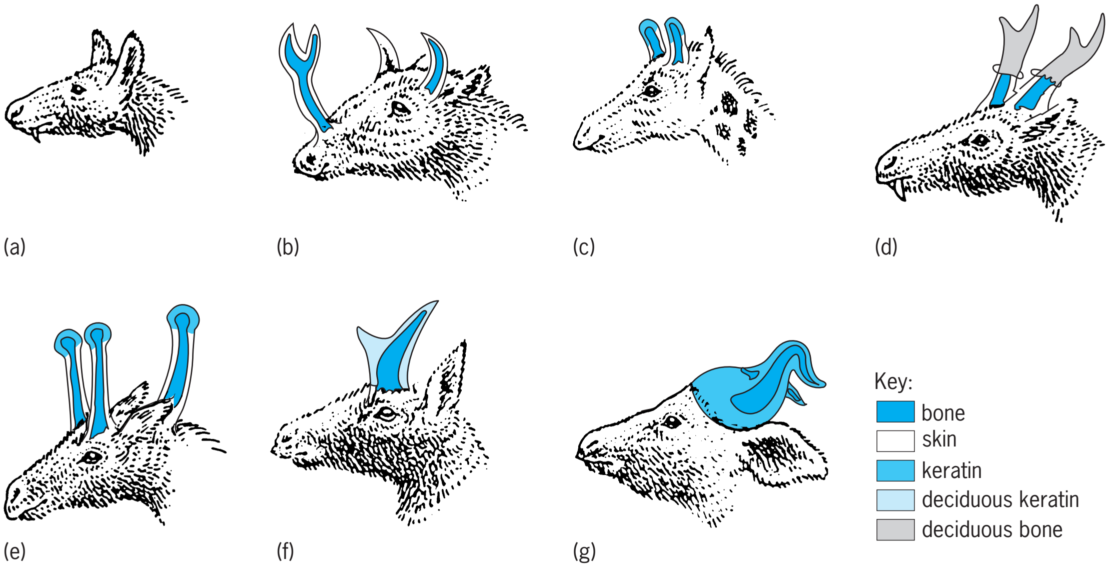


Fig. 4. Heads of male selenodont artiodactyls, showing different types of cranial appendages. Relative thickness of skin or keratin cover is not to scale. (*a*) Tragulid, showing primitive hornless condition with enlarged upper canines. (*b*) Protoceratid, an extinct tylopod convergent with ruminants in horn acquisition, with a forked nasal horn in addition to postorbial ones. (*c*) Giraffid (giraffe), with ossicones. (*d*) Cervid (muntjac deer), with antlers. (*e*) Dromomerycid, extinct cervid-related group with an occipital horn in addition to postorbital ones. (*f*) Antilocaprid (pronghorn), with deciduous keratinous horn sheath. (*g*) Bovid (cow), with true horns. (*After C. M. Janis, Evolution of horns and related structures in hoofed mammals, Discovery, 19(2):8–17, Yale Peabody Museum of Natural History, 1986*)

Classification, evolution, and distribution. Despite their host of unique (apomorphic) features, in terms of branching patterns artiodactyls represent an early divergence from the main stem of ungulate evolution. The earliest artiodactyls, rabbit-sized animals found in the Eocene of North America and Europe, are commonly known as dichobunids. By the late middle Eocene the earliest members of the two modern radiations, the Bunodontia and the Selenodontia, had arisen from among different groups of dichobunids. The three main types of living artiodactyls—suoids, camelids, and ruminants—can trace their roots back to this Eocene divergence (**Fig. 5**). There has been much debate as to whether whales (order Cetacea) should be included among the Artiodactyla. Whales are clearly the sister taxon to artiodactyls among modern mammals, but some molecular studies imply that they are closely allied with the hippos. Morphologists and paleontologists would generally prefer to keep whales separate from artiodactyls. *See* CETACEA.

Suborder Suiformes. All living members of the Bunodontia are contained within the Suiformes (suoids) comprising the families Suidae (pigs), Tayassuidae (peccaries), and Hippopotamidae (hippos). Suoids are primarily omnivorous and live in forest, woodland, or brushland. Some suoids, such as the warthog (Suidae) and the hippo, live in grasslands and eat more fibrous herbage (such as grass). Such suoids have evolved lophed teeth and some type of forestomach fermentation convergently with selenodont artiodactyls, although they do not chew the cud. Suoids are primarily an Old World radiation, except for the New World peccaries. Fossil suoids include the Anthracotheriidae, piglike forms with lophed teeth known from the early Tertiary of North America and until the Pliocene of the Old World. Anthracotheres may have been ancestral to hippos. Other extinct bunodonts (possibly grouped within the Suiformes) include the European small to medium-size Cebochoeridae and Choeropotamidae, and the Entelodontidae of North America and Eurasia, which were a radiation of giant hogs, some as large as bison, that may have been carrion scavengers. *See* HIPPOPOTAMUS.

Suborder Tylopoda. The only living selenodont artiodactyls that are not members of the suborder Ruminantia are the camelids. The suborder Tylopoda is an assemblage of nonruminant selenodont families that may not be closely related to each other. Today, the family Camelidae is known from the large, desert-dwelling camels of the Old World and the antelope-size llamas (including alpacas and vicunas) of the mountains and pampas of South America. Present-day camelids are mixed feeders, living in herds in open habitats. However, until around 2 million years ago, camelids were an exclusively North American group. They first appeared in the late Eocene, and during the later Cenozoic they were a highly diverse radiation, paralleling the African ruminants in their morphological diversity (although none appeared to be true grazers). Their diversity declined in the Pliocene, and they became extinct in North America around 10,000 years ago at the end of the Pleistocene. *See* ALPACA; CAMEL; LLAMA.

Extinct tylopods can be divided into three main groups: (1) small to medium-size European forms of the early Tertiary, including the sheeplike Anoplotheriidae, the rabbitlike Cainotheriidae, and the llamalike Xiphodontidae; (2) some North American forms probably closely related to camelids, the rabbitlike Oromerycidae and the deerlike Protoceratidae, whose later members sported horns both over the eyes and on the nose (Fig. 4*b*); (3) and the North American oreodonts (families Agriochoeridae and Merycoidodontidae), the most common North American fossils of the Oligocene, that resembled a cross between a pig, a sheep, and a hyrax.

Suborder Ruminantia. Ruminants can be divided into the more primitive Tragulina (traguloids), an assemblage of small, hornless forms, and the more derived Pecora, which usually comprise larger, horned forms. Traguloids are first known from the late middle Eocene of Eurasia and North America, and are a mostly early Tertiary radiation. They include the families Amphimerycidae, Hypertragulidae, and Leptomerycidae. Traguloids usually have relatively short legs with unfused metapodials, and the males have saberlike upper canines. Surviving traguloids are the members of the family Tragulidae, including the mouse deer (*Tragulus*) of southeast Asia and the chevrotain (*Hyemoschus*) of central Africa. Tragulids are solitary, tropical-forest-dwelling animals that eat mainly nonfibrous leaves and fruit and do not appear to chew the cud. *See* CHEVROTAIN.

The earliest known pecorans were the Eurasian Gelocidae, first appearing in the late Eocene. These animals were rather like traguloids in general appearance but had more derived features of the limbs, skull, and dentition. The living musk deer (family Moschidae) has a similar appearance: small-size, with the males possessing large upper canines rather than horns. Modern moschids are known only from Asia, but are also recorded from the Miocene of North America (the blastomerycines) and the Oligocene of Europe. The modern, horned pecoran families date from the early Miocene. Pecorans also have more derived limb, with longer, fused metapodials; the reduction or loss of digits 2 and 5; and a reduced ulna and fibula. These limbs are better adapted for fast, sustained locomotion.

The family Giraffidae is an exclusively Old World group, known today only from Africa, but also from Asia until the Recent. Giraffids have ossicones, simple skin-covered cranial appendages (Fig. 4*c*). There are two living species of giraffids, both browsers (although some extinct giraffids had a more mixed-feeding diet): the forest-dwelling, solitary okapi (*Okapia*) and the savanna-dwelling, herd-forming giraffe (*Giraffa*). There was also a Plio-Pleistocene radiation of short-necked mooselike giraffids called sivatheres. *See* GIRAFFE.

The family Cervidae is a primarily Eurasian group, although cervids first reached North America in the Pliocene (5 million years ago), and later invaded South America to produce an extensive radiation

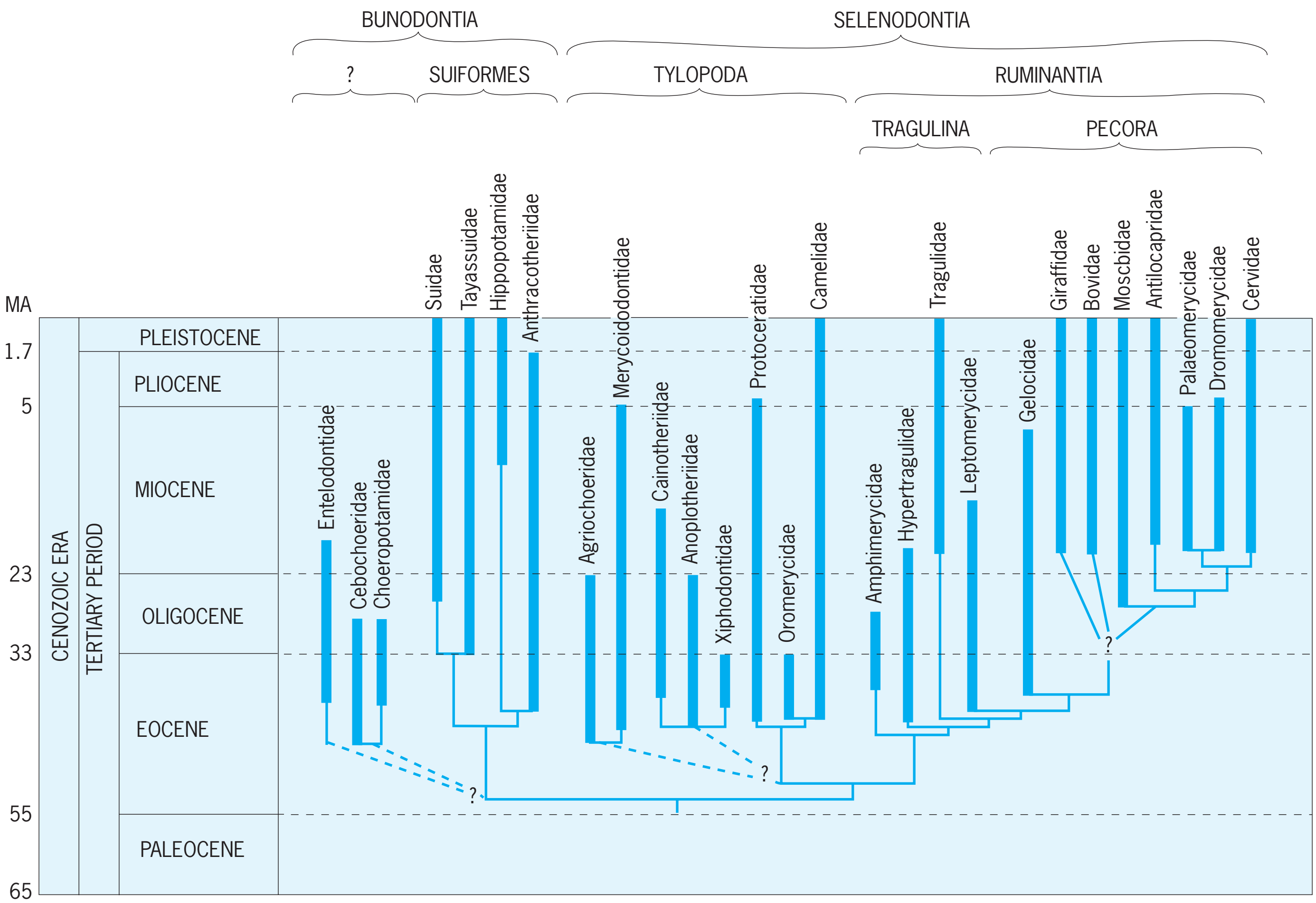


Fig. 5. Consensus of the interrelationships of the families of artiodactyls, including all living families and selected extinct ones (the basal Eocene radiation of dichobunids is not shown). Ma = million years ago.

of New World tropical and temperate forms. Most cervids possess antlers, a branched bony cranial appendage that is shed and regrown annually. Cervids that lack antlers, such as the Chinese water deer (*Hydropotes*), or that have only short antlers, such as the muntjac (*Muntiacus*), may have an enlarged upper canine in the males (Fig. 4*d*). Most cervids are browsers or mixed feeders, living in woodland or brushland, and forming small groups rather than permanent large herds (the reindeer or caribou is the one exception, and this deer is also the only one where antlers are present in the female). Extinct related groups, which were also apparently woodland browsers, include the Eurasian Palaeomerycidae and the North American Dromomerycidae. Some members of these groups sported a third horn on the occipital region as well as horns over the eyes (Fig. 4*e*). *See* DEER; MOOSE; REINDEER.

The family Antilocapridae, surviving only as the pronghorn "antelope," is an exclusively North American group. Antilocaprids are highly convergent on the antelope in anatomy and behavior, and all appear to have been savanna- or plains-dwelling mixed feeders, but they are probably more closely related to cervids than to bovids. The pronghorn combines a permanent bony horn core with a deciduous sheath that is shed annually (Fig. 4*f*). Extinct antilocaprids probably did not shed the sheath, or had horns that resembled nondeciduous antlers made from naked bone. *See* PRONGHORN.

The family Bovidae is today the most diverse of the artiodactyl families, containing cattle, sheep, goats, and antelope. Bovids are a primarily Old World radiation: they did not reach North America until the Pleistocene, and never reached South America (except as forms introduced by humans). Bovids vary greatly in their behavior and ecology, ranging from tragulid-like solitary tropical-forest selective browsers (such as duikers), to deerlike tropical-woodland browsers (such as the bushbuck and bongo), to savanna-dwelling herd-forming grazers (such as the wildebeest and buffalo), to cold or temperate-dwelling mixed feeders (such as sheep, goats, and musk-ox). Bovids are characterized by true horns, unbranched cranial appendages that consist of a bony horn core covered by a nondeciduous keratin sheath (Fig. 4*g*). The great radiation and diversification of bovids is a primarily recent (Plio-Pleistocene) phenomenon in association with the spread of African savannas. *See* ANTELOPE; BISON; BUFFALO; MUSK-OX. Christine Janis

Bibliography. G. A. Feldhamer, B. C. Thompson, and J. A. Chapman (eds.), *Wild Mammals of North*

America: Biology, Management, and Conservation, 2d ed., Johns Hopkins University Press, 2003; *Grzimek's Encyclopedia of Mammals*, vol. 5, McGraw-Hill, 1990; D. MacDonald, *The Encyclopedia of Mammals*, Andromeda Oxford Limited, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

Asbestos

Any of six naturally occurring minerals characterized by being extremely fibrous (asbestiform), incombustible, and having high tensile strength. Historically they were utilized in commerce for fire protection; for fiber reinforcing material in tiles, plastics, and cements; for friction materials; and for thousands of other uses. Currently the vast majority of asbestos used worldwide is chrysotile type which is used for asbestos-cement production, while use continues to be strong only in friction products, coating and compounds, and roofing products. Because of great concern over the health effects of asbestos, many countries have promulgated strict regulations or bans on its use. The six minerals designated as asbestos also occur in a nonfibrous form. In addition, there are many other minerals which occasionally occur naturally with a highly fibrous morphology similar to asbestos. These ambiguities generated a controversy over whether the mineralogical definition of asbestos fibers should be applied when doing health hazard evaluations for legal and regulatory purposes. The U.S. Occupational Safety and Health Administration (OSHA) has reviewed this matter and concluded that the medical and scientific information available does not support an expansion of the definition of asbestos to include minerals not mineralogically defined as asbestos in the regulations.

Minerals. The six naturally occurring minerals exploited commercially for their desirable physical properties, which are in part derived from their asbestiform habit, are chrysotile asbestos—a member of the serpentine mineral group; and anthophyllite asbestos, grunerite asbestos (known historically by the commercial name amosite), riebeckite asbestos (known historically by the commercial name crocidolite), tremolite asbestos, and actinolite asbestos—all members of the amphibole mineral group. Populations of these mineral fibers, however processed, can be demonstrated to be asbestos if the length varies independently of the diameter.

Chrysotile asbestos. Chrysotile has the ideal chemical formula $Mg_3[Si_2O_5](OH)_4$ [Mg = magnesium, Si = silicon, O = oxygen, and OH = hydroxy]. Small amounts of aluminum (Al), iron (Fe), manganese (Mn), calcium (Ca), potassium (K), and sodium (Na) may appear in the bulk chemical analyses of this mineral (see **table**). The chrysotile fibers consist of long hollow tubes, and each fiber is approximately 25 nanometers in diameter, with lengths varying from well under 1 micrometer to well over 10 cm. The fibers are characteristically curly and ropelike. Chrysotile is identified by a combination of quantitative elemental composition; distinctive shape, size, and hollow-tube morphology of the fibers (**Fig. 1**); and characteristic x-ray or electron diffraction pattern. *See* SERPENTINITE.

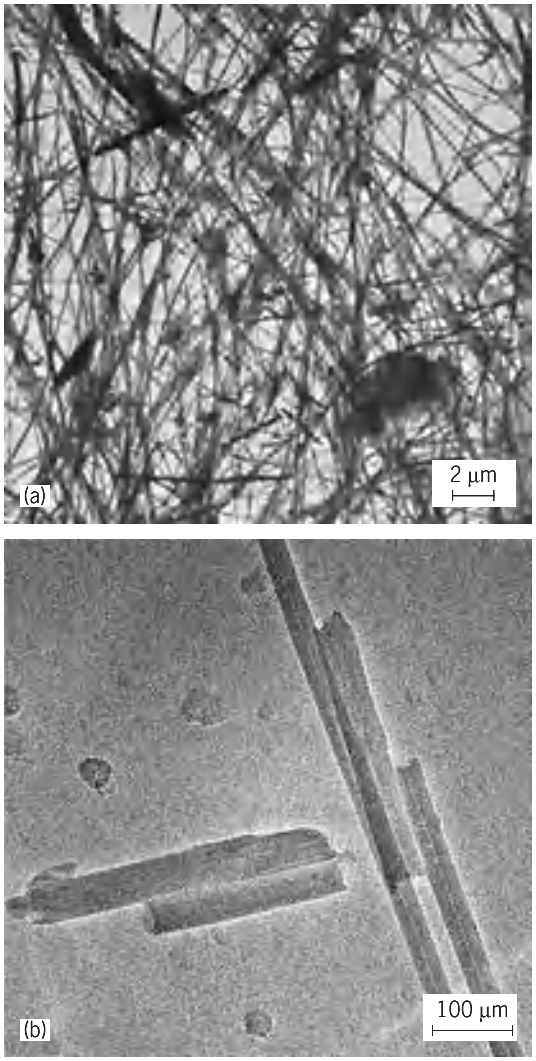


Fig. 1. Transmission electron photomicrographs of chrysotile asbestos from the Bazhenovskye Deposit in the Russian Federation. (*a*) At 1000 × magnification. (*b*) At 40,000 × magnification, where the hollow-tube morphology is visible.

Amphibole asbestos. The five regulated forms of amphibole asbestos and their ideal chemical formulas are grunerite asbestos, $(Fe^{2+},Mg)_7[Si_8O_{22}](OH)_2$; riebeckite asbestos, $Na_2Fe^{3+}_2(Fe^{2+},Mg)_3[Si_8O_{22}](OH)_2$; anthophyllite asbestos, $Mg_7[Si_8O_{22}](OH)_2$; tremolite asbestos, $Ca_2Mg_5[Si_8O_{22}](OH)_2$; and actinolite asbestos, $Ca_2(Mg,Fe^{2+})_5[Si_8O_{22}](OH)_2$. A considerable amount of chemical substitution takes place in these asbestos minerals, as indicated in representative analyses presented in the table. The amphibole asbestos minerals are characterized by their various elemental compositions and distinctive diffraction patterns, and generally vary in width

Anhydrous wt % of the major elements of representative specimens of the six regulated asbestos minerals

Element	Amosite (Penge, RSA[*])	Anthophyllite (Paakila, Finland)	Chrysotile (Coalinga,[†] CA)	Crocidolite (Kuruman, RSA[*])	Actinolite (Wrightwood, CA)	Tremolite (Korea)
Si	24.3	28.8	22.3	24.4	27.3	27.8
Fe	31.1	5.2	2.7	28.3	4.9	0.7
Mn	0.5	—	—	—	—	—
Mg	4.0	19.2	29.2	2.1	14.5	15.1
Ca	—	—	—	—	8.0	10.5
Na	—	—	—	4.3	—	—
O	40.1	46.9	45.7	40.8	45.3	45.9

[*] Republic of South Africa.
[†] Coalinga is part of Diablo Mountain Range.

from 0.1 to over 1 μm and in length from 1 μm to over 10 cm (**Fig. 2**). *See* AMPHIBOLE.

Geology. The major deposits of commercial chrysotile asbestos are found in massive serpentinite bodies formed by alteration of magnesium- and iron-rich igneous rocks such as dunite, peridotite, and iherzolite. The world's largest deposits of this type are found in Québec Province in Canada, in the Ural Mountains of the Russia Federation, in the Troodos Mountains of Cyprus, and in the Diablo Range of

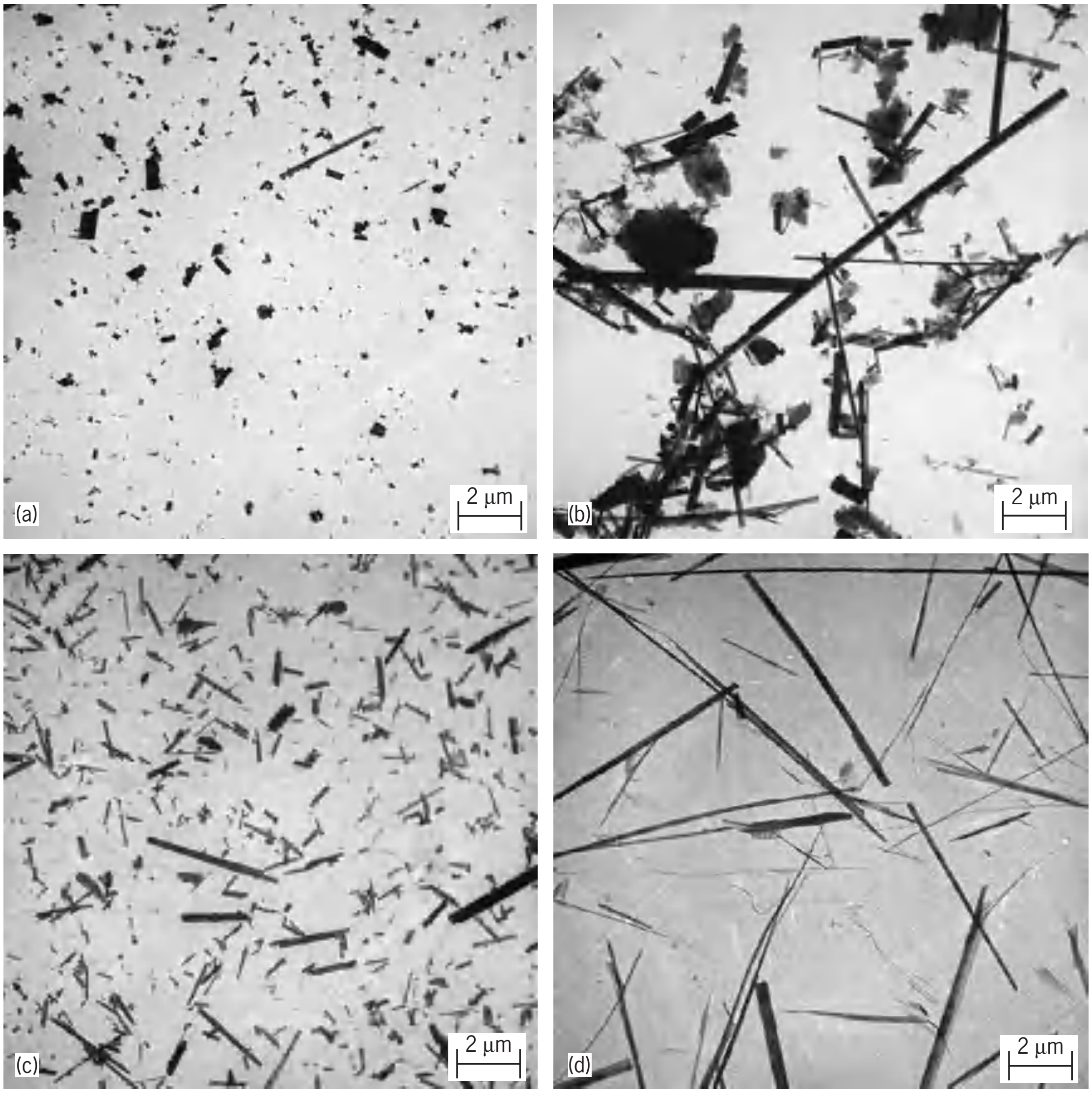


Fig. 2. Transmission electron photomicrographs of (*a*) amosite asbestos, Penge, Republic of South Africa; (*b*) anthophyllite asbestos, Paakila, Finland; (*c*) crocidolite asbestos, Koegas, Republic of South Africa; and (*d*) tremolite-actinolite asbestos, Jamestown, CA. Note that tremolite and actinolite are part of a solid solution series and differ only in iron content.

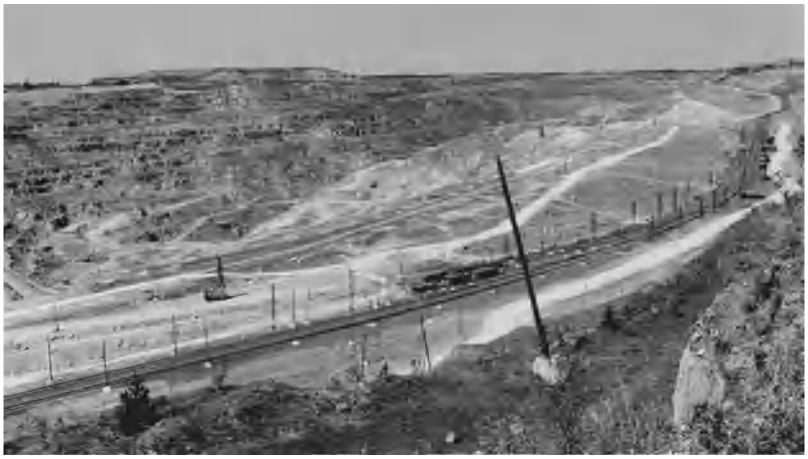


Fig. 3. Bazhenovskoye chrysotile asbestos deposit located in the Urals Region of the Russian Federation. The deposit was discovered by Alexey Ladizhensky in 1889.

California. The asbestos in these serpentinite deposits is usually mined in large open pits (**Fig. 3**). Minor amounts of fibrous and nonfibrous anthophyllite, tremolite, and talc are also found in serpentinites. Serpentinized limestones also contain minor deposits of chrysotile asbestos. All the important commercial deposits of grunerite and riebeckite asbestos occur in Precambrian banded iron formations located in three areas—in the Cape and Transvaal provinces of the Republic of South Africa and in Western Australia. Anthophyllite asbestos, found within serpentinized periodotites of the Karelian Mountains, East Finland, was mined in modest commercial quantities from 1918 to 1975. Asbestos fibers crystallize from aqueous fluids that penetrate the open cracks, fissures, faults, and slip planes within tectonically stressed host rock. The fibers that form oriented perpendicular to the walls of the open spaces are known as cross fibers, while the fibers that form oriented parallel to a slip or fault plane are known as slip fibers. *See* DUNITE; PERIDOTITE; SERPENTINITE.

Sources. Asbestos, mainly the chrysotile variety, is mined throughout the world. The two most important sources of chrysotile asbestos are the Urals Region of the Russian Federation and Québec Province in Canada. Lesser amounts of amosite and crocidolite asbestos have been mined in the Republic of South Africa, and anthophyllite in East Finland. Commercial production of tremolite and actinolite asbestos has been small and sporadic. The countries of the former Soviet Union are the major producers of chrysotile asbestos, followed by Canada, China, Brazil, Zimbabwe, and South Africa. The only producing asbestos mine in the United States is located within the New Idria serpentinite at the south end of the Diablo Range in California; ore containing 50–70% short-fiber chrysotile asbestos is excavated from an open pit and then trucked to King City, CA, for wet processing.

Uses. At present more than 99% of the asbestos used in the world is chrysotile, although small regional amphibole asbestos operations are thought to continue; however, the amphibole asbestos minerals essentially have left international commerce. Chrysotile asbestos continues to be used for many types of products because of its chemical and thermal stability, high tensile strength, flexibility, low electrical conductivity, and large surface area. Past uses of asbestos, such as sprayed-on insulation, where the fibers may become easily airborne, have been generally abandoned. Chrysotile asbestos is used predominantly for the construction industry in the form of cement sheets, coatings, pipes, and roofing products. Additional important uses are for reinforcing plastics and tiles, for friction materials, and for packings and gaskets.

Health effects. Historically, unaware of the dangers, workers in the asbestos trades and in mining and milling industries were excessively exposed to needlessly high levels of asbestos, which caused a very high incidence of morbidity and mortality. The three principal diseases associated with exposure to the asbestos minerals are lung cancer; mesothelioma, a rare cancer of the pleural and peritoneal membranes that enclose the chest and abdominal cavities; and asbestosis, a nonmalignant disease characterized by a diffuse interstitial fibrosis of the lung, which causes the lung tissue to become stiff and exchange oxygen poorly. Excessive exposure to all the asbestos fiber types is associated with asbestosis and increased risk of lung cancer. Mesothelioma, a rare tumor accounting for approximately 1 in 10,000 deaths in the general population, can be dramatically increased by exposure to amosite, crocidolite, or tremolite asbestos. These last two fiber types are strongly associated with an increased incidence of nonoccupational mesothelioma and therefore are thought to present a risk at rather low exposures. As the current commercial use of amphibole asbestos is very limited, the risk of environmental mesothelioma associated with naturally occurring asbestos outcroppings, particularly tremolite asbestos, has become of concern. The ingestion of asbestos has not been proven to cause disease in either humans or animals. Workers employed in the friction materials, cement, and roofing materials industries who were exposed only to chrysotile asbestos, generally at levels below two fibers per cubic centimeter, do not appear to have an excess of either malignant or nonmalignant disease. *See* MUTAGENS AND CARCINOGENS; ONCOLOGY; RESPIRATORY SYSTEM DISORDERS.

Malcolm Ross; Robert P. Nolan

Bibliography. *Asbestos in Public and Commercial Buildings*, Health Effects Institute, 1991; B. Levadie (ed.), *Definitions for Asbestos and Other Health Related Silicates*, American Society for Testing and Materials, STP 834, 1984; G. D. Guthrie and B. T. Mossman (eds.), *Health Effects of Mineral Dusts*, Mineralogical Society of America, 1993; R. P. Nolan et al. (eds.), *Health Effects of Chrysotile-Asbestos: Contribution of Science to Risk Management Decisions*, Can. Mineralog. Spec. Pub., 2000; H. C. W. Skinner, M. Ross, and C. Frondel, *Asbestos and Other Fibrous Materials*, 1989; R. L. Virta, *Minerals Yearbook*, vol. 1: *Metals and Minerals (Asbestos)*, U.S. Bureau of Mines, 1989.

Ascaridida

An order of nematodes in which the oral opening is generally surrounded by three or six labia; in some taxa labia are absent, but the cephalic sensilla are always evident. Usually there are eight cephalic or labial sensilla; the submedians may be fused and then only four sensilla are seen. The stoma varies from being completely reduced to spacious or globose. The esophagus varies from club shaped to nearly cylindrical, never rhabditoid. There may be posterior esophageal or anterior intestinal ceca. The collecting tubules of the excretory system may extend posteriorly and anteriorly. Males generally have two spicules; however, in some taxa there may be none or only one. The gubernaculum may also be present or absent. Though females generally have two ovaries, multiple ovaries do occur. The number of uteri is also variable: two, three, four, or six. Phasmids are sometimes large and pocketlike. Reportedly, the larvae lack a stomatal hook or barb.

The order probably comprises seven superfamilies: Ascaridoidea, Seuratoidea, Camallanoidea, Dracunculoidea, Subuluroidea, Dioctophymatoidea, and Muspiceoidea (*incertae sedis*).

Ascaridoidea. This superfamily of about 65 genera comprises large parasitic roundworms, whose adult stages usually occur in the stomach or small intestine of terrestrial and aquatic mammals, birds, reptiles, and fishes (see **table**); the parasitic larval stages of many species occur, either temporarily or indefinitely, in other parts of the host's body.

General morphology. These roundworms are mostly medium to large and thick bodied, with three large lips, interlabia, and other cuticular specializations of the head. Eight incompletely fused submedian cephalic papillae are located on the external circle, and a lateroventral pair is present. The buccal capsule is usually more or less rudimentary. The esophagus is more or less cylindroid, is sometimes wholly muscular, and sometimes terminates posteriorly in a glandular ventriculus with or without one or more appendices. The intestine sometimes has one or more anterior ceca. Females are oviparous, sometimes with more than two uteri; the eggs are generally more or less spherical and thick shelled. Males have two spicules, but rarely caudal alae or a precloacal sucker.

Life cycles. Many species have a direct life cycle. Others, mainly species with marine mammals, birds, and fishes as definitive hosts, require an intermediate host, such as a fish, amphibian, insect, crustacean, or small mammal. Distribution is cosmopolitan, and these worms are the causative agents of ascariasis, or ascaridosis, occurring in many agricultural animals (and occasionally in humans). Infestation is typically characterized by pulmonary damage and distress initially, and digestive disturbances later. Damage may also occur during larval migration to other parts of the body, including the liver and brain.

Common ascarids. *Ascaris lumbricoides*, the large roundworm of humans, has a life cycle similar to that of *A. suum*. Infestations are very common in tropical and temperate regions where sanitation is poor or lacking. These worms often are as big as an ordinary pencil.

Ascaris suum, the pig ascarid, usually is considered distinct from *A. lumbricoides*, on certain biological criteria. The eggs of this worm pass with feces onto the ground, where they develop until each contains an infective larva, capable of hatching in the gut of another pig. The liberated larvae penetrate into the intestinal wall, enter the bloodstream, break out of capillaries into the air spaces of the lungs (causing "thumps," or ascaroid pneumonia), and via the trachea reenter the gut where they develop into adults. In a mouse, the same larval migration occurs, but the rest of the cycle is not completed.

Lung damage from the larval invasion may lead to a fatal pneumonia in young pigs. Reportedly, hog cholera virus becomes pathogenic under this pulmonary stress. The virus is innocuously present in nematodes of the genus *Metastrongylus* that commonly infest the lungs of pigs. Intestinal infestations are debilitating and cause economic loss.

Heavy infestations of *Ascaridia galli*, the large intestinal roundworm of chickens, can cause intestinal blockage and death; lesser infestations stunt growth and interfere with good bone development. The life cycle resembles that of *Ascaris suum*, but with limited migration of the larvae into and out of the intestinal wall. Turkeys also are subject to infestation.

Toxacara canis, one of the common ascarids of dogs, is very injurious to young puppies and can be lethal in heavy infestations. Its life cycle is like that of *Ascaris suum*, but infection also occurs prenatally, and by ingestion of rats with encapsulated *T. canis* larvae in the tissues.

The exposure of humans to ascarids of lower animals, mainly those associated with dogs and cats, can result in a very serious condition known as visceral larva migrans. As with normal human *Ascaris*, migration to ectopic foci often results in serious and sometimes fatal conditions. In human tissues and organs such as the liver, lungs, brain, eye, viscera, or muscles, the larvae become encapsulated by granulomatous tissue. Prognosis varies, but in the majority of cases recovery is complete. However, blindness, paralysis, and death are not uncommon.

Dracunculoidea. This superfamily of parasitic nematodes comprises obligate tissue parasites of fishes, reptiles, and mammals. All known species require an intermediate host in order to complete their life cycle, and that host is always a water flea (*Cyclops*). Most often the stoma is reduced and vestibular. In mature females, the vulva is equatorial and atrophied; egg and larval deposition often results from the rupturing of the female. The most widely known example is *Dracunculus medinensis*, the guinea worm. Knowledge of this worm and its lesion extends back to antiquity.

Ingestion of water containing infective *Cyclops* is the only known source of infection. The encysted nematode larvae are released from *Cyclops* by the digestive juices of the duodenum. Then the larvae burrow through the intestinal wall, and upon reaching

Some Ascaridoidea of domesticated mammals and birds in the United States[*]

Name	Host[†]	Remarks
Toxocara cati	Cat	Occasional parasite of humans; *Toxocara* larvae can cause human visceral larva migrans; cats can become infected by eating mice with the larvae in their tissues
Toxascaris leonina	Cat, dog	Occasional parasite of humans; normally develops without passage of larvae through the lungs
Neoascaris vitulorum	Cattle	Not generally distributed; infestations are injurious to calves; infection occurs prenatally, perhaps exclusively
Parascaris equorum	Horse	Larval migration of tracheal type damages liver and lungs; heavy infestations are sometimes fatal to foals
Ascaridia columbae	Pigeon	Heavy infestations can cause anemia, droopiness, and diarrhea
Ascaridia dissimilis	Turkey	Exact information on pathogenicity is not available

[*] All have a direct life cycle.
[†] Some of the species also occur in other hosts.

the loose connective tissue, usually retroperitoneal in position, they develop to adulthood in 8 months to 1 year.

The gravid females, 28–48 in. (70–120 cm) long, migrate from the site of development to the subcutaneous tissues. At that time the ovaries are postfunctional, and the uteri are highly coiled, distended, and filled with masses of rhabditoid larvae. The females make their way to the surface of the skin, and a papule is formed. The papule continues to develop and a blister forms, usually on the lower extremities. When the blister comes in contact with fresh water, the uterus bursts through the anterior part of the nematode's body, and the worm also bursts, releasing cloudlike swarms of motile larvae. These larvae are then filtered from the water by *Cyclops* and subsequently ingested. Development to the infective stage within *Cyclops* takes about 10 days.

The formation of the blister and subsequent rupturing of the female produce a profound allergic reaction. This reaction results from the release of large amounts of toxic by-products from the worm. Upon discharge of larvae in fresh water, much of the allergic reaction abates. The reactions and systemic prodromes include erythema, urticarial rash, pruritus, vomiting, diarrhea, and giddiness. Septicemia, suppurating cysts, and chronic abscesses are not uncommonly associated with these infections. The worms can be removed surgically, or in the native manner of winding upon a stick. Chemotherapy is also available.

Control in endemic areas includes keeping infected persons from wading or bathing in water used for drinking purposes, and the education to avoid drinking suspect water. *See* NEMATA.

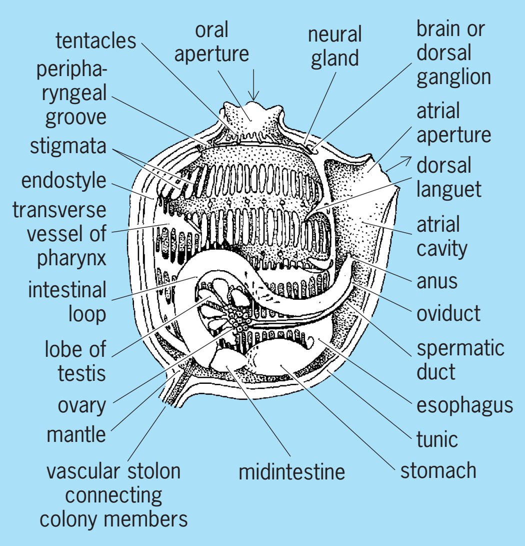
Armand R. Maggenti

Bibliography. R. Muller, *Dracunculus* and dracunculiasis, in *Advances in Parasitology*, vol. 9, 1971; F. A. Neva and H. W. Brown, *Basic Clinical Parasitology*, 6th ed., 1996.

Ascidiacea

A class of Tunicata which occurs as solitary zooids or, by a process of asexual budding, develops into colonies. Zooids vary in length from a few millimeters to 10 in. (25 cm). Individuals or colonies are invested by a protective covering, the tunic or test, made of polysaccharide material structurally close to cellulose. Beneath the test is the body wall or mantle. Each zooid has two apertures: inhalant (oral) and exhalant (atrial; see **illus.**). Water currents, created by cilia on the margins of stigmata in the pharyngeal wall, draw water into the branchial sac, where it is filtered and passed out through the exhalant aperture. The entrance to the branchial sac is guarded by tentacles which prevent large particles from entering. Filtration takes place on a mucous sheet secreted by the endostyle; the sheet is passed across the internal face of the branchial sac by ciliary or muscular action and is then rolled into a cord by the dorsal lamina (or by languets). Digestive enzymes are secreted into the stomach, and a pyloric gland, of unknown function, enters at the junction of stomach and intestine. Gonads are hermaphroditic, and may be situated in the loop of the intestine or in the mantle wall.

The tadpole larva does not feed and settles and metamorphoses within a few hours; the tail tissues are resorbed at this time into the trunk and provide



Left side of a zooid of the colonial ascidian *Perophora*. The tunic, mantle, and anterior wall of the pharynx have been removed from the left. The two arrows indicate direction of water intake and expulsion.

a nutrient reserve for further development of the adult. Asexual reproduction giving rise to colonies may occur by budding from stolons, by budding from tissues of the mantle wall, or by dedifferentiation of zooids, which then constrict into a number of buds.

Common examples of solitary ascidians are *Ciona*, *Molgula*, and *Ascidia*; examples of colonial forms are *Botryllus*, *Perophora*, and *Amaroucium*. Three orders of ascidian are recognized, Aplousobranchia, Phlebobranchia, and Stolidobranchia, differing from one another by characters of the brancial sac and position of the gonads. Ascidians occur throughout the seas of the world and at all depths, including the abyssal region. Most species feed on minute particulate matter, but a few are carnivores and engulf small zooplankters. *See* TUNICATA. Ivan Goodbody

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; F. S. Russell and C. M. Yonge (eds.), *Advances in Marine Biology*, vol. 9, 1971, vol. 12, 1974; W. G. Van Name, The North and South American ascidians, *Bull. Amer. Mus. Nat. Hist.*, vol. 84, 1945.

Ascomycota

A phylum in the kingdom Fungi, representing the largest of the major groups of fungi, and distinguished by the presence of the ascus, a specialized saclike cell in which fusion of nuclei and reduction division occur. The resulting nuclei are delimited by membranes through free-cell formation and a wall to form ascospores inside the ascus. In most ascomycetes, each ascus contains eight ascospores, but the number may vary from one to several hundred.

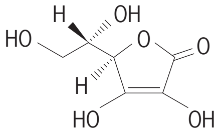
In the simplest ascomycetes (yeasts), the vegetative body (thallus) is unicellular; however, in the majority of ascomycetes, the thallus is more complex and consists of a tubular, threadlike hypha with cross walls which grows in or on the substrate. These hyphae eventually form structures called ascomata (ascocarps), on or in which the asci are formed. In addition to their sexual reproduction, most ascomycetes reproduce asexually by means of conidia. Traditionally, the structure of the ascoma and ascus has served as the basis for subdividing the Ascomycota into five classes: Hemiascomycetes, Plectomycetes, Pyrenomycetes, Discomycetes, Loculoascomycetes. The recent introduction of molecular data, however, is changing concepts of the relationships of different groups of ascomycetes and will eventually lead to a much-revised classification scheme. *See* DISCOMYCETES; HYMENOMYCETES; LOCULOASCOMYCETES; PLECTOMYCETES.

As a group, the ascomycetes occur throughout the world in all types of habitats and on both living and dead substrates. An estimated 33,000 species of ascomycetes are arranged in about 3300 genera, with new species being described regularly. Ecologically ascomycetes function as primary decomposers of plant materials, but they also are important as plant and human pathogens; in baking, brewing, and winemaking; in enzyme and acid production; and as sources of antibiotics and other drugs. *See* EUMYCOTA; FUNGI; PLANT PATHOLOGY; YEAST. Richard T. Hanlin

Bibliography. C. J. Alexopoulos, C. W. Mims, and M. Blackwell, *Introductory Mycology*, 4th ed., John Wiley, New York, 1996; J. Breitenbach and F. Kränzlin, *Fungi of Switzerland*, vol. 1: *Ascomycetes*, Edition Mycologia, Lucerne, 1984; R. W. G. Dennis, *British Ascomycetes*, rev. ed., J. Cramer, Vaduz, 1981; R. T. Hanlin, *Illustrated Genera of Ascomycetes*, vols. I and II, APS Press, St. Paul, 1990, 1998.

Ascorbic acid

A vitamin also known as vitamin C. It is a white crystalline compound, highly soluble in water, a stronger reducing agent than the hexose sugars, which it resembles chemically. Vitamin C deficiency in humans has been known for centuries as scurvy. The compound has the structural formula shown below.



L-Ascorbic acid

The stability of ascorbic acid decreases with increases in temperature and pH. This destruction by oxidation is a serious problem in that a considerable quantity of the vitamin C content of foods is lost during processing, storage, and preparation. Biological methods for estimating ascorbic acid are rarely used. The vitamin is determined chemically by making use of its reducing properties.

While vitamin C is widespread in plant materials, it is found sparingly in animal tissues. Of all the animals studied, only the guinea pig, the red vented bulbul bird, the fruit-eating bat, and the primates, including humans, require a dietary source of vitamin C. The other species studied are capable of synthesizing the vitamin in such tissues as liver and kidneys. Some drugs, particularly the terpenelike cyclic ketones, stimulate the production of ascorbic acid by rat tissues.

Vitamin C–deficient animals suffer from defects in their mesenchymal tissues. Their ability to manufacture collagen, dentine, and osteoid, the intercellular cement substances, is impaired. This may be related to a role of ascorbic acid in the formation of hydroxyproline, an amino acid found in structural proteins, particularly collagen. People with scurvy lose weight and are easily fatigued. Their bones are fragile, and their joints sore and swollen. Their gums are swollen and bloody, and in advanced stages their teeth fall out. They also develop internal and subcutaneous hemorrhages.

The biochemical role of ascorbic acid is obscure. It seems reasonable to expect that it functions metabolically in oxidation-reduction systems, since it has been shown that ascorbic acid and dehydroascorbic acid are readily interconverted in plant and animal tissues. It may also act as an antioxidant, protecting hydrogen carriers from destructive oxidation. Ascorbic acid has a role in tyrosine metabolism. It also appears to function in the conversion of folic to folinic acid. *See* FOLIC ACID.

There has been great difficulty in establishing the human requirements for vitamin C. Usually, vitamin requirements are based on data obtained from dietary surveys accompanied by blood or urine analyses and often by saturation tests. There is evidence that vitamin C may play roles in stress reactions, in infectious disease, or in wound healing. Therefore, many nutritionists believe that the human intake of ascorbic acid should be many times more than that level which produces deficiency symptoms.

The recommended dietary allowances of the Food and Nutrition Board of the National Research Council are 30 mg per day for 1–3-month infants, 80 mg per day for growing boys and girls, and 100 mg per day for pregnant and lactating women. These values represent an intake which tends to maintain tissue and plasma concentrations in a range similar to that of other well-nourished species of animals. The Accessory Food Factors Subcommittee of the British Medical Research Council does not believe that a dietary intake of more than 30 mg per day has beneficial effects. *See* VITAMIN.

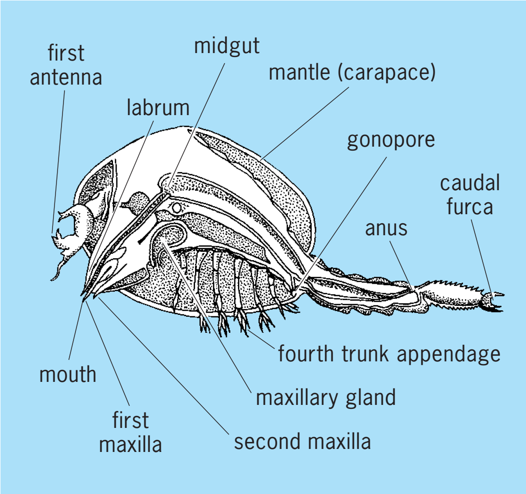
In most processes for the large-scale synthesis of ascorbic acid, technical dextrose is hydrogenated to sorbitol. Biological oxidation converts sorbitol to sorbose. The next step is diacetone sorbose, which is oxidized to diacetone gulosonic acid. Diacetone gulosonic acid is hydrolyzed to gulosonic acid. The methyl ester of the latter is reacted with sodium methylate to form the sodium salt of ascorbic acid. Stanley N. Gershoff; Werner A. Lindenmaier

Bibliography. S. N. Gershoff and W. A. Lindenmaier, Ascorbic acid, in AccessScience@McGraw-Hill, http://www.accessscience.com, DOI 10.1036/1097-8542.054400.

Ascothoracica

An order of the subclass Cirripedia. Members are ecto- or endoparasites of coelenterates and echinoderms. The body is enclosed in a voluminous saclike mantle up to 0.8 in. (20 mm) long, whereas the body is only about one-fourth of this length (see **illus.**). Ascothoracica are not attached permanently to the host by the antennular region, but the antennules may be modified as a clasping organ. The mouthparts are also modified as a clasping organ and for piercing and sucking. Up to six pairs of thoracic appendages are present, reduced in number and development in the endoparasites. Unlike all other Cirripedia, Ascothoracica retain a more or less segmented abdomen in the adult. There are no cement glands. Diverticula of the alimentary canal extend into the mantle.

The sexes are separate, males being smaller than females. Larval development takes place mainly in the parental mantle. The nauplii have no frontal horns. There are said to be two cypris stages, the first being passed in the mantle, the second free-swimming and most probably the dispersal and infective stage.



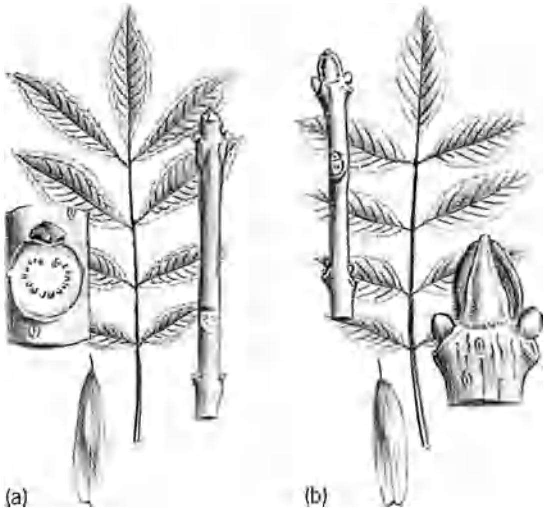
***Ascothorax ophioctenis*, a parasite in the bursae of brittle stars. (*After R. D. Barnes, Invertebrate Zoology, 2d ed., Saunders, 1968*)**

The virtually motile existence of individuals, the possession of an abdomen, and discrepancies in the larval cycle have prompted some writers to separate the Ascothoracica from the Cirripedia. The presence of a cypris stage indicates a relationship nearer to the Cirripedia than to any other subclass of the Crustacea. *See* CIRRIPEDIA. H. G. Stubbings

Ash

A genus, *Fraxinus*, of deciduous trees of the olive family Oleaceae, order Scrophulariales, which have opposite, pinnate leaflets, except in one species, *F. anomala*, which has only a single leaflet. There are about 65 species in the Northern Hemisphere. This tree occurs in America south to Mexico, in Asia south to Java, and in Europe. *See* SCROPHULARIALES.

The white ash (*F. americana*), of the eastern United States, is the most important species, and attains a height of 120 ft (36 m). This species has stalked leaflets, rusty-colored winter buds, and an erect trunk that is valuable for lumber. The wood is light, strong, and flexible, and is used for oars, baseball bats, furniture, motor vehicle parts, boxes, baskets, and crates. The black ash (*F. nigra*; **illus.** *a*) grows 75 ft (23 m) in height in wet soils in the northeastern United States and Canada and has sessile leaflets and friable outer bark. The wood of black ash is used for the same purposes as that of white ash. The red ash (*F. pennsylvania*), also in the eastern United States and adjacent Canada, grows to 60 ft (18 m) and has pubescent (hairy) twigs and leaf stalks. The uses of the wood of this species are also similar to those of white ash. Some species of ash are ornamental trees, such as the flowering ash (*F. ornus*) with gray winter buds and white flowers, and the



Two species of ash. (*a*) Black ash (*Fraxinus nigra*). (*b*) European ash (*F. excelsior*).

European ash (*F. excelsior*; illus. *b*) with black buds and sessile leaflets. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Asia

The largest of the world's continents. With its peninsular extension, commonly called the continent of Europe, it is the major portion of the broad east-west extent of the Northern Hemisphere landmasses. In many ways Asia is more a cultural concept than a physical entity. There is no logical physical separation between Asia and Europe, and even Africa is separated from Asia merely by the width of the Suez Canal. For convenience, however, the Eurasian landmass is considered to be divided by the Ural Mountains into Europe in the west and Asia in the east. Thus restricted Asia has an area of about 17,700,000 mi^2 (45,800,000 km^2), about one-third of the land area of the Earth. In the north, mainland Siberia reaches past the 77th parallel. Southward, India and Sri Lanka (Ceylon) reach nearer than 10° north of the Equator, while the Indonesian islands extend more than 10° south of the Equator. The continental heart of Asia is more than 2000 mi (3200 km) from the nearest ocean. These vast distances have great significance in the climates and vegetation of Asia and thus in the erosional and depositional patterns of the land (**Fig. 1**). *See* CONTINENT; EUROPE.

As the world's largest continent, Asia has been shaped by the processes of cratonization, geosynclinal development, and associated orogenesis. By accretion several major stable structural units were formed, and these collectively form the basement for much of the continent (Fig. 1). The three northern cratons of Asia—the Russian, the Siberian or Anggaran, and the Sino-Korean—were evidently stabilized as early as 1800 or 1700 million years ago (Ma), whereas the more southerly ones—the Chang Jiang, the Tarim, and the Arabian—appear to have stabilized and been consolidated into the Chinese protoplatform much later (around 700–500 Ma). The Indian craton, though formed early, has been in collision with the Eurasian platform since about 40 Ma.

Geosynclinal development occurred in the central Asian region on continental margins. As a consequence, sedimentary deposits now blanket much of the basement, and these sediments have been involved in a series of orogenic episodes. In the early Paleozoic (500 Ma) there were several areas of activity—Kunlun, Qinling, Beishan, and Tianshan—and the closing of this system merged the Tarim platform with the Eurasian platform. In the Mesozoic (200 Ma), disintegration and geosynclinal development occurred on the southern flank (the Tethys-Himalayan Domain, in what is now central-northern Tibet), and in the Marginal Pacific Tectonic Domain, covering present-day eastern China, Korea, eastern Russia, and the adjacent marine margins. From these events arose the major structural alignments which persist to the present, running west-east throughout central Asia, and southwest-northeast through the Pacific margins of the continent. The collision of the Indian craton with the Eurasian landmass added the predominantly west-east alignment of the Himalayan system, as well as forcing the Tibetan plateau uplift, and the reactivation of preexisting west-east mountain structures in central Asia.

It has been postulated that the Marginal Pacific Tectonic Domain of Asia may have been in the Paleozoic, an Atlantic-type margin (tensional, stable, and relatively quiescent) but that in the Mesozoic it was converted to a Pacific type, with strong compression, and a highly active Benioff zone situated to the east of the island (and quasi-island) arcs—Kamchatka, Japan, Taiwan, and the Philippines. It has also been postulated that southeast Asia was accreted as a series of relatively small fragments detached from the Australian landmass and propelled northwestward until they became part of the Asian landmass.

Tectonic and geosynclinal processes continue at the present time. Volcanism and seismic activity are widespread in the island arcs and indeed on parts of the Asian mainland affected by geosynclinal activity. In the marine basins and the lowlands behind the Pacific Rim arcs, sedimentation is proceeding actively. The convergence of India with Tibet sustains the mountain ranges of the Himalayas, with elevations of over 25,000 ft (7600 m) as well as the 16,400-ft (5000-m) plateau to the north. In the same manner the Arabian plate in collision with Iran sustains elevations in excess of 13,000 ft (4000 m). *See* OROGENY; PLATE TECTONICS.

Mountain chains. Asia's great mountain chains are related to the tectonic structures. The Pamir knot marks the northern penetration of the Indian craton, while the ranges encircling Assam in northeast India represent its northeasterly wing. Elevations in the vicinity of the Pamirs reach 24,600 ft (7500 m); ranges then radiate in several directions. The Hindu Kush range curves southwestward into Afghanistan.

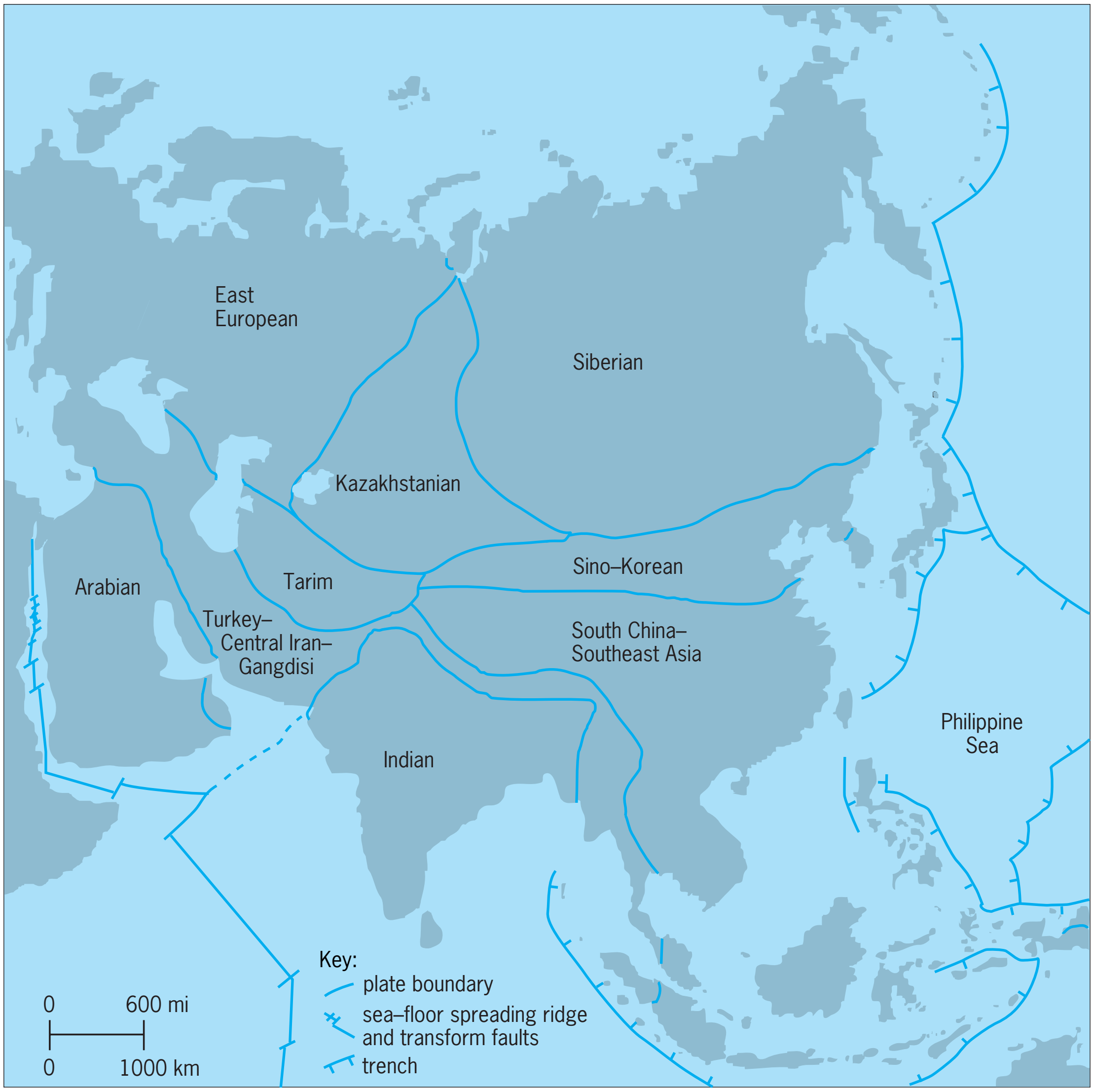

East European
Siberian
Kazakhstanian
Tarim
Sino–Korean
Arabian
Turkey–Central Iran–Gangdisi
South China–Southeast Asia
Philippine Sea
Indian
Key:
plate boundary
sea–floor spreading ridge and transform faults
trench
0 600 mi
0 1000 km

Fig. 1. Major cratons that have been involved in the Eurasian tectonic plate. (*After The Research Institute of Geology, Chinese Academy of Geological Sciences, Scheme showing the subdivisions of paleo and present plates in Asia, Cartographic Publishing House, Shingxa, 1982*)

The mountain trendline continues, after a jog northwestward, in the Elburz of northern Iran and thence in the Armenian highlands and the Caucasus, each with elevations reaching 18,000 ft (5500 m), decreasing thereafter to the Pontus and Taurus ranges of northern and southern Turkey. In western and southern Iran are the massive Zagros and Makran ranges.

Southeastward from the Pamir knot run the three most imposing mountain chains on Earth: the Karakorum, which continues the line of the Hindu Kush eastward in an arc convex to the north; the Himalaya in an arc convex to the south; and the shorter Trans-Himalaya, north of the Himalaya, with higher average elevations but peaks of lesser height. In all of these, the average elevations exceed 4 mi (6400 m), with several scores of peaks reaching a height in excess of 25,000 ft (7600 m) above sea level. Everest, 29,141 ft (8882 m), and Kinchinjunga, 28,146 ft (8579 m), lie in the Himalaya, while the peak designated as K2, 28,250 ft (8611 m), rises in the Karakorum. The few passes range between 14,000 and 15,000 ft (4300 and 4600 m), although the Karakorum Pass lies at 18,317 ft (5583 m).

In eastern Tibet the Himalaya and Nyanqen-tanglha curve sharply around the northeast extremity of the Indian craton, the Himalaya being cut by the great gorge of the Brahmaputra River. From the bend zone, great ridges divided by deep gorges run south to form the Burma-China frontiers and the mountain backbones of the Malay peninsula and Vietnam. The Nan Ling system of southern China diverges eastward to divide the Chang Jiang from the Xi drainage.

From the western Himalaya, the 11,000-ft (3400-m) Sulaiman Range runs south and, together with the Kirthar Range, divides West Pakistan from Afghanistan.

Branching off from the Karakorum south of Kashgar, the Kunlun mountains run eastward across China. In the west they have been elevated along with the uplifted Tibetan plateau, and reach elevations of 21,000 ft (6100 m). In the east, lacking the elevated Tibetan base, they nonetheless reach altitudes of 12,000 ft (3700 m) in the Qin Ling and lesser Dabie mountains. In eastern China they reach almost to the Pacific coast. Together with the northeastward arc of the Altyn Tagh and the Nan Shan branching from it, the Kunlun forms the northern wall of the Tibetan plateau. Near the eastern end of the Kunlun proper lie the A'nyemaqen Mountains, with peaks up to 25,000 ft (7600 m).

Trending east and west is a series of great ranges to its north, with mutual connections in the west. These include the Altai-Sayan, the Tannu Ola, and the Hentyin which form natural boundaries for Outer Mongolia. They continue the systems of young mountains crossing central Asia; farther northeast, they extend further in the Stanovo Mountains of eastern Siberia.

Hill lands and plateaus. Sedimentation over the ancient craton surfaces as well as in the intercraton zones has in some instances been followed by deformation, erosion, and the formation of hill regions. In other cases, where deformation has been minor, plateau surfaces have resulted. Most of southern China, which lies upon the Chang Jiang craton, consists of hill areas with local relief of up to 1640 ft (500 m), though absolute elevations above sea level are considerable. The elevated western and eastern edges of the southern portions of the Indian craton are marked by the Western and Eastern Ghats, and in the south the Nilgiri Hills rise to elevations of 8000 ft (2400 m). The Cambodian craton in its eastern regions is buried under younger geological materials that make up the Laotian and Vietnamese hill zones; and in the south it is overlain by the Cardamom Mountains.

The Tibetan plateau is maintained at its overall elevation of around 16,000 ft (4900 m) by the constant and continuing motion of the colliding Indian plate. Though much of the surface of the Tibetan plateau is flat, mountain ranges traverse it, and deep river gorges have been cut into it.

Other plateau regions represent areas in which sedimentary materials laid down on the craton surfaces have been left relatively undisturbed. In eastern Asia lies the subhumid loess plateau of northern China, the humid Yunnan and Guizhou plateaus of southern China, and the Kaima plateau of northern Korea. Much of the Siberian craton is covered by the Siberian plateau. In southern Asia most of peninsular India is a plateau block on the Indian craton, tilted from west to east, and covered in its northeasterly portions by recent volcanic lava. Saudi Arabia is likewise a tilted plateau block, with sand and other materials overlying the basement of the Arabian plate.

In southeastern Asia the Shan plateau is an undulating surface, deeply incised by the Salween river in the east. In contrast, the Korat plateau in eastern Thailand represents sedimentary rock formations very little disturbed since they were deposited on the craton surface in Mesozoic times.

Lowland plains. The most significant topographic units of Asia are the great alluvial plains and river deltas; in some of these live the vast majority of the world's population. The gross drainage pattern of Asia is radial; the rivers flow from the highlands in the heart of the continent and run outward in all directions. Only in the south, east, and north sectors of the continent do the rivers reach the sea. Flowing into the peripheral seas of the Pacific are such mighty rivers as the Mekong, the Xi, the Chang Jiang, the Huai, the Yellow, and the Amur, each building large, heavily populated plains and, with the exception of the Amur, densely settled deltas. The Yellow Plain (North China Plain) with some 125,000 mi^2 (324,000 km^2) of area, and the Chang Jiang Plain, with about 75,000 mi^2 (194,000 km^2), are among the most extensive alluvial plains of the Earth. In the shallow South China, East China, and Yellow seas, the deltas of the Mekong, Xi, Chang Jiang, Huai, and Yellow rivers are pushing steadily seaward. *See* DELTA; FLOODPLAIN.

North of the Amur, all the great Asian rivers drain across Siberia into the Arctic, with plains supporting only boreal forests and tundra. Westward, the radial rivers end in internal drainage basins and saltwater bodies such as Lake Balkash and the Aral Sea. These include the Ili, flowing into Lake Balkash from Chinese Xinjiang, and the Amu and Syr Darya, running into the Aral Sea. On the frontier of Iran and Afghanistan lies the seasonal Lake Helmand with the Helmand River emptying into it. The long Tarim River flows in the desert basin of southern Xinjiang until it dries up in its easterly course or empties into shifting salt lakes such as Lop Nur. *See* BASIN; PLAINS.

From the glaciers of eastern Tibet and the Himalaya run meltwaters to which are added the heavy summer monsoon rains of southern and southeastern Asia. These waters collect in such great streams as the Indus and Ganges-Brahmaputra of India, and the Irriwaddy and Salween of Burma, all emptying into the Indian Ocean, and all except the Salween have constructed great delta plains which support some of the densest populations of southern and southeastern Asia. India's east coasts have such great rivers as the Kistna, Godavari, and Cauveri; on the west coast are the Narbada and the Tapti. Most streams of the Indian peninsula are dry in winter.

Most rivers in southwestern Asia also are intermittent seasonal streams. The only large rivers here are the Tigris-Euphrates streams, whose sediments have formed the Mesopotamian plain extending inland from the head of the Persian Gulf. Their waters, however, are derived neither from the Arabian deserts nor the dry Mesopotamian region, but from the uplands of Armenia and southern Turkey. Smaller streams flow out of northern and western Turkey into the Black, Aegean, and Mediterranean seas. In Palestine the small Jordan flows down a rift valley to the Dead Sea, 1292 ft (394 m) below sea level. The only other such low-lying surface of Asia is the Turfan depression of Xinjiang, 928 ft (283 m) below sea level. *See* MEDITERRANEAN SEA; RIFT VALLEY; RIVER.

Islands. Important sectors of Asia, containing some 400,000,000 people, are completely insular. The most important are the Japanese, Philippine, and

Indonesian islands and Taiwan. Almost all of Asia's islands lie in great volcanic arcs bounding large seas off the continent's Pacific coast. At least 160 active volcanoes are found here and in Kamchatka. With such tectonic activity accompanied by periodic fractures, uplifts, and downfaultings, it is not surprising that the geological structures and topographic features of many of these islands are exceedingly complex and that the soils are generally immature. It also is to be expected that level alluvial land must be scarce, for rivers are short and, although their swift drop brings large loads of sediments, many of the islands rise from such great oceanic depths that delta extension seaward is slow. *See* OCEANIC ISLANDS; VOLCANO.

Few islands lie along the Asiatic coasts of the Indian Ocean, although the Sunda chain of Indonesia has perhaps more of a claim to Indian Ocean frontage than to Pacific frontage. The 400-mi (640-km) Andaman and Nicobar chains in the Bay of Bengal form genetic connections between the Arakan Yoma of western Burma and the mountains of northern Sumatra. Sri Lanka, however, is the only significant island in the northern part of the Indian Ocean west of Sumatra. In the Persian Gulf off the north coast of Arabia lies the small island Bahrein, which is rich in oil. *See* INDIAN OCEAN.

Few islands lie off the alluviated coastlands of northern Siberia. Some moderately large ones are included in the barren and rocky Severnaya Zemlya group, the New Siberian Islands, and Wrangel Island. The Commander Islands and Karaginski Island lie in the Bering Sea only a short distance from the Aleutians.

Climates. Five major climatic types may be distinguished in the Asian region: (1) the monsoonal system of eastern Asia, (2) the monsoonal system of southern Asia, (3) the equatorial regions of southeastern Asia and their extension into the Southern Hemisphere as they are influenced by the Australian monsoon, (4) the winter rainfall areas of southwestern Asia, and (5) the cyclonic and convectional storm systems of central and northern Asia.

Fundamental to understanding the climates of Asia are the vastness of the unbroken landmass and the long latitudinal stretch from the polar realm to south of the Equator. These are responsible for the great temperature and humidity extremes that occur. The greatest ranges of temperatures in the world have been recorded in interior Asia. Continentality, therefore, is the outstanding feature of climates of interior Asia. In coastal and insular areas of east Asia, however, winds moving over the warm, northward-flowing Japan Current and the western Pacific waters moderate the coastland and island climates. *See* MARITIME METEOROLOGY.

Monsoonal circulations. The term monsoon derives from an Arabic word referring to seasonal winds affecting southern and southeastern Asia that are predominantly onshore, warm, and wet during the summer and offshore, cool, and dry during the winter. While both southern Asia and eastern Asia have monsoonal regimes that are related to the large-scale western Pacific and Eurasian circulation systems, they differ somewhat in their characteristics. In each instance, however, the summer monsoon represents a vast penetration of tropical influences and humidity into more northerly latitudes.

The winter in central Asia is characterized by extremely low temperatures, associated with surface divergence. Outflow of air at ground level is blocked to the south by the high relief of the Tibetan plateau and the Himalayan mountain system. Bursts of cold air break out periodically across central Russia, northern China, and Korea, and reach as far as Japan, where they deposit large quantities of snow derived from moisture captured in passage across the Sea of Japan. Over China and the western Pacific this air makes its way gradually south, eventually passing as the northeast monsoon into southeastern and southern Asia. It is, however, a dry period in much of southern and southeastern Asia.

In summer the Asian landmass, and especially the Tibetan plateau, becomes a heat source for the atmosphere. In response, surface-level circulation patterns are established. One pushes a surge of humid warm air across the Indian Ocean from the southwest and into the Indian subcontinent, extending itself steadily during May and June all the way up to the northwestern part of Pakistan, where its progress is blocked by the mountain ranges. Some of this air passes across northern southeastern Asia to join with Pacific maritime air, and advances with a series of sudden steps into central and northern China, Korea, and the southern two-thirds of Japan. The initial onset of these monsoonal surges is accompanied by heavy rains, which are followed by intermittent bursts of rainfall as long as the dynamics of this southwest flow are sustained. By September both of these northward surges of energy have diminished, and gradually the monsoon conditions retreat to the south.

While southern and eastern Asia typify the classic monsoon pattern, parts of northern southeast Asia share in this seasonal reversal. Burma, Thailand, and Indochina receive the bulk of their precipitation from the summer monsoon. Their cool seasons (in the Northern Hemisphere winter) are dry, except for some coastal locations in Vietnam, but the three remain sheltered from the extreme cold that is associated with the outflow of Siberian air in regions farther to the north. *See* MONSOON METEOROLOGY.

Equatorial regions of southeastern Asia. Across southeastern Asia, and especially in parts of Malaysia and Indonesia, equatorial conditions prevail. At no time of the year is this region without substantial rainfall and high humidity. However, in moving eastward across Indonesia it becomes apparent that there is an effect derived from the "monsoonal" regime of the Australian landmass. Less pronounced than the monsoon of the Northern Hemisphere, that of Australia does nonetheless show similar features, with high rainfall and humidity in the "high sun" months (November through February) and cooler drier conditions in the "low sun" months (May through August).

Southwestern Asia. Southwestern Asia extends the desert zone of northern Africa eastward, with low precipitation and extremely high summer temperatures. During the winter months cyclonic storm systems track through the Mediterranean trough and occasionally reach as far east as the Persian Gulf area. *See* DESERT.

Central Asia. Continentality is a predominant feature of the climates of this region. There are few sources of accesible moisture to provide humidity and precipitation. As is the case with southwestern Asia, midlatitude cyclonic storms penetrate this area, particularly in the summer months, to bring small amounts of precipitation. Winters are cold and dry.

Typhoons. Parts of southern, southeastern, and eastern Asia lie in the path of typhoons. Typhoons develop from initial disturbances or "waves" in the tropical easterly surface flows (flows sometimes referred to as the trade winds), usually at some distance north or south of the Equator. Such waves are not uncommon, but a portion of them, especially in the later summer months, develop into full-scale typhoons, which travel westward and then gradually curve away from the Equator in a northward or southward direction, depending on the hemisphere. Massive energy release comes from the high humidity of the tropical maritime air entrained in them, and the typhoons strike with ferocious winds and torrential downpours of rain. On making a landfall they usually lose their energy source and gradually die out. Those in the Bay of Bengal may cause extensive damage and bring heavy rains to the east coast of India. In the Pacific, the destructive typhoons may cross the northern Philippines and hit parts of the continental coast, usually between Amoy and central Vietnam. Many of them, however, follow a path gradually curving northward, past or across Taiwan, Okinawa, and the southern part of Japan as well as southern coastal Korea. These typhoons spread wind, wave, and flood destruction but also bring heavy beneficial rainfall over areas 100–300 mi (160–480 km) in diameter along their paths. *See* TROPICAL METEOROLOGY.

General rainfall patterns. The driest portions of Asia include the vast areas of southern Mongolia, Xinjiang, central Asia, and southwestern Asia. Except for small, favored mountain areas, most of this region from the Gobi to the Red Sea gets less than 10 in. (25 cm) of precipitation per year. With the exception of southern Arabia, which is subtropical desert, these are midlatitude desert and dry steppe regions. Favored with higher rainfall are the Yemen Mountains and the coastal mountains of Turkey, together with Lebanon, Syria, and northern Israel. The highlands of Armenia and the Elburz of Iran are favored also with more abundant rainfall, which ranges from 25 to 50 in. (64 to 127 cm) or more per year.

The northeastern Siberian mountains and the arctic coastal lands also receive meager rainfall, less than 8 in. (20 cm), but are not dry because evaporation is low and the water table is high. Most of Siberia has permafrost below a few feet of surface soil, so that rainwater does not filter far down. Between the arid belt of central Asia and the northeast Siberian low-precipitation zone, the annual rainfall ranges between 10 and 18 in. (25 and 45 cm).

In eastern Asia the precipitation increases in a southeasterly direction from interior Asia to the coast. The annual maximum seldom exceeds 80 in. (203 cm) in the wetter southeast coastal regions, whereas this drops to less than 30 in. (76 cm) in the North China Plain and less than 15 in. (38 cm) at the Great Wall. In some mountainous parts of Japan and Taiwan, the yearly average may be more than 100 in. (254 cm).

In the Indian subcontinent rainfall is heaviest along the western plateau fringe and in East Bengal, where it averages over 100 in. (254 cm) per year. The interior of the peninsula is relatively dry. Northwestern India and Pakistan share the drought of southwestern Asia. With the exception of the extreme north, Sri Lanka generally has abundant rainfall.

With the exception of some small areas in Assam, southeastern Asia has the heaviest rainfall of the entire Asiatic region. The mainland mountains facing the southwest summer monsoon crossing the Bay of Bengal, and parts of the Vietnamese and Laotian cordilleras facing the humidified northeast winter monsoons of eastern Asia regularly get average rainfalls of 120–150 in. (305–380 cm) or even more. Equally heavy rainfalls occur in the southwestern half of Sumatra, southwestern Java, the northwestern half of Borneo, and the Pacific fringe of the Philippine Islands. In general, and with only a few minor exceptions, southeastern Asia has no areas that are subject to severe drought.

General soil patterns. This large continent presents a variety of zonal soils and accumulations of soil material, particularly alluvium.

Podzolic soils (spodosols). In cold to cool regions of sufficient soil moisture, such as most of Siberia, soils are acid, podzolized, coarse-textured, and relatively sterile. Soils here generally are overlain with a semicarbonized coniferous leaf layer in the forest belt. The tundra ground layer may be bare rock, coarse sediment, or sterile sands. Sphagnum peat bogs occupy depressions. Elsewhere a peaty mat of little-decayed organic matter forms a "soil." This supports only tundra mosses and lichens and some scrub bushes. The greater part of Siberia, however, is in taiga, or boreal coniferous forests, scrublike toward the tundra belt but taller in the south.

Brown, coarse, acidic, and somewhat podzolized soils are characteristic of most forested areas in temperate regions of Asia because of the acidity of the leaf matter and the leaching of finer soil particles from the top soil horizon. In the higher mountains of southern and middle Asia, podzolic soils associated with coniferous forest belts are common. With warmer latitudes and lower altitudes, deciduous broad-leafed forests replace the coniferous forests, and the more rapid decomposition of leaf and litter on the forest floor produces a richer surface soil.

Calcareous soils. As ultisol-oxisol continua, the calcareous soils contrast sharply with the humid podzolic soils of the northlands and high-altitude slope lands of Asia. They include the arid and subhumid calcareous, chestnut-colored soils of the southern

Siberian, central Asian, and southwest Asian steppelands, and of the drier agricultural areas of China and India. Meager precipitation in these areas provides too little groundwater to support deep-rooted plants such as trees. Shallow-rooted grasses and herbs form a ground cover, which in time, if left uncultivated, becomes a thick layer of dead plant matter. Because of the low humidity, leaching of either organic matter or soluble minerals is much reduced. In fact, capillary action tends to bring dissolved minerals back to, or near, the surface after rains have ceased. The soils have a high percentage of soluble calcium and nitrogenous matter and are therefore of great agricultural richness. These are the soils of the west Siberian wheatlands and the steppes surrounding the central Asian deserts and such areas as southern Inner Mongolia, China's loess plateau, and Turkey's Anatolia. With sufficient water, either through rainfall or irrigation, they are exceptionally productive of cereals and other crops.

Toward the desert margins of these soils the humus thins with the thinning grass stands and the soil turns to light chestnut, but mineral plant foods are abundant. In the bottoms of poorly drained basins here, alkaline accumulations may inhibit vegetation growth. By contrast, where rainfall is somewhat more abundant, as in the west Siberian plain, the central Manchurian plains, and the Amur-Ussuri lowlands, high stands of thick grass may lead to deep black earths or chernozems.

Loess soils. The loess soils (various orders of cumulic soils) are a specialized type resulting from the accumulation of dust blown from desert regions where strong outblowing winds occur. The most significant and widespread deposits of this type in Asia are found in China's northwest loess plateau. Depths of fine-textured soils with a high calcium content may reach 100–300 ft (30–90 m). Incorporated with the loess are decayed grass roots and other organic matter. Vertical root channels with a cementation of soluble mineral salts produce a vertical structure so that columnar erosional cliffs or remnants are often found. This columnar pattern of calcification permits the Chinese in this region to tunnel into cliffs and carve out underground rooms for homes with comparative ease. Heavy rainfall results in sheet and gulley erosion, however, and much of the area has developed badland characteristics. Nevertheless, the soil is rich for agriculture. Grasses tend to be the natural vegetation where cultivation is not carried on.

Humid lateritic and leached soils. As ultisol-oxisol continua, these humid tropical and subtropical soils involve different processes in formation and have differing characteristics from the podzolized soils of the boreal and temperate forests. They also are acid and are formed under forest conditions, generally under broad-leafed forests, either deciduous monsoon forests or evergreen rainforests. Bacterial decay of the forest litter is so rapid under the warm, humid conditions and rainfall so heavy during part or all of the year that soluble minerals and comminuted organic matter are either leached away rapidly to deep soil horizons or washed away to rivers and the sea. The chemical weathering and solution of silica and the lack of carbonization of leaf matter to form carbonic acid leave iron and aluminum oxides in surface layers as red and yellow earths and clays. In extreme and long-term development, the subsurface red earths change to true laterite.

Most of the humid slope lands of Japan, southern China, southeastern Asia, and southern Asia have iron-rich red earths. In humid, mist-shrouded upland parts of southern China such as Guizhou Province, aluminum-rich yellow earths predominate. The fertility of these slope lands is quickly lost when stripped of trees and used for agriculture. Erosion then cuts the hillsides into ravines, while heavy sediment loads are carried by the numerous streams to the plains and the sea.

In the humid tropical and subtropical forest realms the fertility of the soil also differs considerably according to the nature of the parent materials. The older soils overlying the ancient rock complexes of the stable platforms such as are found in Borneo, southeastern Sumatra, large parts of southeastern Asia, and southern peninsular India are rather infertile compared with younger soil materials based upon more recent volcanic lavas and ash (andosols). The areas with numerous volcanoes and basic lava and ash have relatively fertile soils and support dense populations. Java, southwestern Sumatra, Sulawesi, and the Japanese islands demonstrate this relationship. Borneo, although large, is without volcanoes, and its infertile soils support a population only one-thirtieth as large as that of smaller Java. In peninsular India the most fertile soils are associated with the basaltic lava of the Deccan that produces the black waxy regur.

Alluvial soils. As entisol-inceptisol continua, the youngest, or least maturely developed, soils of Asia are the great areas of alluvium, which are intensively cultivated wherever water and temperature conditions are suitable. Their fertility in part derives from frequent inundations by floods which bring new layers of sediment containing organic debris and minerals in solution and suspension. Where irrigation water is available or can be brought to these alluvia, the soil is utilized for wet-rice cultivation in most parts of Asia except Russia (**Fig. 2**). *See* SOIL.

Major vegetation patterns. Asia's vegetation belts and zones follow, in general, the climatic patterns from desert lands through tropical to Arctic margins.

Tundra zone. A belt of tundra comprises the islands in the Arctic Ocean as well as a strip of territory on the Siberian lowlands. Its north-south extent is perhaps greatest in the northeast. In addition, it extends south into elevated portions of the hills and mountains. The frozen subsoil permits the growth of little more than mosses, lichens, dwarfed trees, and scrub. Because the flow of the great rivers here is northward, the mouths often remain blocked with ice even after their upper reaches have already melted. The flow downstream thus floods and spreads over immense regions of the tundra zone, making large areas impassable on foot. *See* PERMAFROST; TUNDRA.

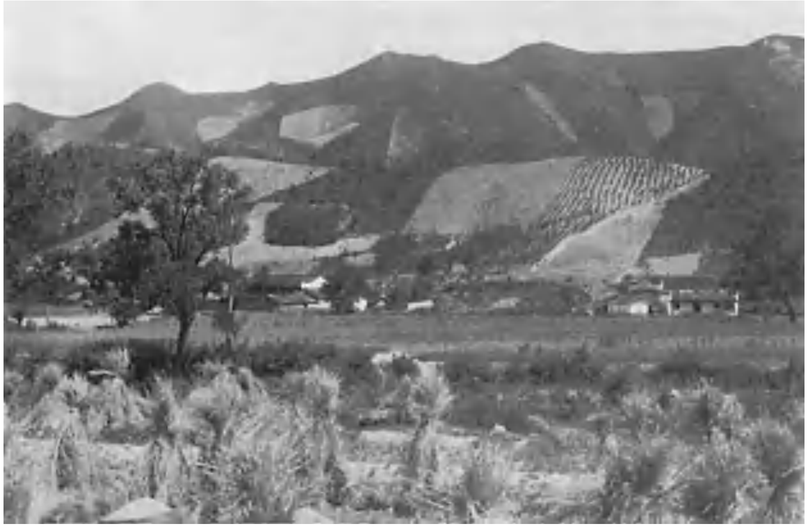


Fig. 2. Alluvial bottomlands, long heavily cultivated for rice, in the hill country of Zhejiang Province, China, have been subject to damaging sand and gravel deposits washed from unprotected clearings for tung trees and sweet potatoes. (*From G. B. Cressy, Asia's Lands and Peoples, 3d ed., McGraw-Hill, 1963*)

The bush tundra comprises willow and dwarf birch, which hug the ground. Farther southward, fir, birch, and larch trees in stunted form make their appearance. All tundra vegetation is perennial. Tundra also is characteristic of the alpine heights of mountains of interior Asia.

Taiga (boreal coniferous forest). One of the largest expanses of forest left in the world is the Siberian taiga, a dominantly coniferous forest of larches, spruce, fir, and pines, with such deciduous trees as birch and aspen occurring intermixed with the conifers or taking over as a secondary growth in burnt-over areas. The width of this belt in Siberia is more than 1000 mi (1600 km) and it stretches about 4000 mi (6400 km) from the Sea of Okhotsk to the Urals (**Fig. 3**). *See* TAIGA.

West of Lake Baikal, the taiga reaches to the Alta-Sayan, and with little transition, changes to the steppelands of central Asia.

Midlatitude mixed forests. Various admixtures of coniferous and deciduous trees compose this vegetation. Large areas in eastern Asia either have or had this kind of forest prior to clearance. It occurs in Hokkaido and northern Honshu in the Japanese archipelago, and over much of the Korean peninsula. Much of the far eastern parts of Russia is also under this vegetation type, as in the Amur-Ussuri basin, and in the mountainous regions that surround the plains of Manchuria. From the plains themselves forests have long since been cleared.

Mixed midlatitude deciduous and coniferous forest areas of a similar type occupy the hill lands surrounding the Yellow Plain, as well as the Qin Ling Mountains and on into the hill country of the Chang Jiang river valley. This type of forest also may at one time have covered the North China (Yellow) Plain, but the plain has been under cultivation so long that only a guess can be made as to the original vegetation cover. It needs to be pointed out, however, that little commercially usable forest remains in the accessible areas of China and Korea whereas Japan has splendid forests of this so-called temperate type.

Mixed mountain and highland systems. A similar kind of forest makes up an element of the mixed mountain and highland forest systems. These systems occur in the Iranian uplands, in the Altai highlands, in the Pamir-Tian Shan highlands, in the Hindu Kush highlands, and in the Himalayan highlands at lower elevations. Similar forests also surround the Sichuan basin in southwestern China, and spread onto some of the adjacent hill and plateau regions.

Central and east Asian steppelands. From the forests of the west Siberian plain southward, an increasingly dry steppeland is encountered. It extends for 400–500 mi (640–800 km) in a belt about 1000 mi (1600 km) long between the Urals and the Altai-Sayan and associated uplands. The northern half of this belt with its higher annual precipitation of 12–16 in. (30–40 cm) [comprising the Ob dry plain and the northern part of the Kazakh upland running southeastward from it] is the agricultural heart of the plain. The southern part gradually changes to desert steppe and then to desert along about the 50th parallel north. A well-defined zone of dry mountain steppeland occurs in the Tian Shan flanks at elevations of 5500–6000 ft (1680–1830 m), below which desert prevails.

Eastward of Lake Baikal a broadened steppe zone occupies the Trans-Baikal region extending southward to the Gobi Desert of southern Mongolia and eastward to the Great Xingan Mountains, where the zone, about 200 mi (320 km) wide, runs southward in Inner Mongolia. The steppe zone in Inner Mongolia widens with the increasing moisture south of the Great Wall to include most of China's loess

Fig. 3. Winter view of boreal forest (taiga) in Siberia. Hunting and trapping continue to be characteristic activities in great portions of this forest. (*Sovfoto*)

plateau. Grasses also form the natural vegetation of the Manchurian plain, with tall grass in the eastern portion thinning out to short-grass steppe in the Xingan Mountain flanks. The Gobi Desert is flanked by steppelands to its north, east, and south, as well as by mountain steppe zones in the eastern Altai and eastern Tian Shan. From the heart of the Gobi both northward and southeastward, the grasses become richer and taller with the increasing average annual precipitation, and they are at their best in northern Mongolia. In southern Mongolia the richest grasslands have been preempted by Chinese agriculturalists for cultivation.

The typical grasses of the central and east Asian steppelands are feather grass, fescues, and koeleria. In moister parts of the steppe, dicotyledons such as sickle alfalfa, milk vetches, and other legumes are prominent herbs. A grazing economy occupies much of the steppelands, but in the Russian and Chinese spheres the richer grasslands have been put to the plow. *See* GRASSLAND ECOSYSTEM.

Mixed evergreen forests. This tropical and subtropical forest type appears to be limited in Asia mostly to interior southern China and to Japan from the Kwanto Plain southward. South of the Chang Jiang valley it extends from the coast at Shanghai to the gorge lands of eastern Tibet. It includes most of the Chang Jiang drainage and the western part of the Xi drainage, as well as the Yunnan plateau. The coast south of Shanghai in an ever widening zone inland, however, is warm and moist, finally becoming part of the subtropical rainforest.

In Asia the characteristic trees of the mixed forest include broad-leafed evergreen trees such as banyans and camphor, and coniferous trees such as pines, cedars, and cypresses, as well as varieties of bamboo. On the karst limestone hills of Guizhou and Guangxi the tree growth is mainly deciduous, but pines predominate in the leached soils of other hill areas. Horsetailed pines are very prominent. In Japan this forest zone is characterized by live oak, laurel, and other hardwoods such as camphor, intermixed with stands of Japanese cedar, pine, and bamboos. These forests occupy most of the areas of red and yellow earths in China and Japan. The Japanese mountains are well covered with forest, but large parts are in secondary growth of genya or fuel wood. In China most of the original forest has long been cut down and the hills are largely barren of large trees. Much of the cut-over slope land is covered with coarse grasses, ferns, and various scrub, although a beginning has been made in reforestation.

Tropical and subtropical rainforest. The rainforest type is restricted to warm or hot regions of southern and southeastern Asia which get ample rainfall the year round or get so much rain during a large part of the year that a high ground-water table is maintained during the short dry season. The subtropical sectors are found along the southeastern China coast, in Taiwan, and in northern Burma; they merge with the tropical rainforest farther south, where rainfall and temperature increase. Most of southeastern Asia lies in this forest zone, but there are large areas which have such a long dry winter season that they do not support evergreen rainforests. These exceptions are in monsoon forest or savanna grassland described in subsequent sections on Asia's vegetation. In southern Asia the rainforest zones include Assam and the west slopes of the Chindwin, Chin, and Arakan hills, the southern half of Sri Lanka, and the Western Ghats and associated lowlands of southwestern India. *See* RAINFOREST.

The variety of trees and plants in the rainforest is enormous; 11,000 species of flowering plants have been recorded from Borneo alone. Trees are seldom in pure stands, making lumbering a selective and costly process. Camphor and sweetgum, bamboo and palm, citrus, and banyan are common trees in the subtropical coastal areas of China. In the southern tropical rainforest, the coastal forests in areas of tidal mudflats and brackish ground water are formed by varieties of mangrove. Since these are not so dependent upon rainfall, they are found also in coastal areas having seasonal droughts, for example, in eastern India. Mangrove is the characteristic forest in all the muddy tidal soils of the tropics. Next inland and in brackish water, the nipa palm may form dense stands in river estuaries. Sandy coastal strips may have *Casuarina*, *Pandanus*, *Barringtonia*, and dense plantations of coconut palms.

Dipterocarps occupy a prominent place in the lowland rainforests of the interior, because of both their height and the large amount of space they cover. They often form the top story of the forest, with their crowns towering up to 200 ft (60 m) or more. Below an altitude of 2600 ft (800 m) in Malaya they cover 50% of the wooded area, in the Philippines 75%, and in Indonesia as a whole about 50% of the forest stands. Camphor and species of *Shorea* and *Kompassia* are among the giant trees of this forest in Indonesia. Other prominent trees of this forest are gumwoods, mahogany, and ebony. Many banyan-type trees send down aerial roots from their lower branches to take root upon reaching the ground, forming a forest of smaller boles from a single tree. Typical features of the taller trees of this forest are great flaring buttresses that anchor them to the ground. Below the larger trees, a closed canopy up to 60 ft (18 m) high may be formed by secondary trees. Tangles of lianas, which in the case of the rattan palm may reach 500–600 ft (150–180 m) in length, and the abundance of epiphytes and parasitic plants growing on other trees give the impression of an utter confusion of plant life.

Clearing of the tropical rainforest and subsequent abandonment results in a succession of ferns, tropical grasses, and scrub which may be transitional to ultimate reforestation. Often, however, coarse, tough, tall savanna grasses get established; these are more permanent and stifle other growths. Burning may destroy their surface parts, but underground runners quickly send up new sprouts. Forest clearance for agriculture, for settlement, and for timber harvesting has gone on rapidly in southeastern Asia since the end of World War II, and the effective tropical forest area has been reduced by well over half. In addition, three decades of warfare in Indochina resulted in heavy damage to forest ecosystems.

Monsoon tropical deciduous forest. In the tropical parts of Asia, which have a moderately high rainfall but a long dry season (usually in the low-sun period or winter), broad-leafed trees drop their leaves during the season of drought. Most of India is in this vegetation zone, with the exception of the dry northwest and the south-central peninsular areas, the rainforest zone of the Western Ghats and southwestern coast, and the Himalaya mountain zone. In southeastern Asia this zone is found in central Burma, the Shan plateau, northwest Thailand, and in a small part of Assam. It also is found in southern Vietnam and southern Laos together with adjoining areas of Cambodia. Similar monsoon forests occur in the Lesser Sundas, east Java, and Sulawesi.

The monsoon deciduous forest consists mostly of mixed species, but sometimes a single species becomes dominant as a result of selection from frequent burnings. Teak, which has especially attracted commercial attention, is one of these. Its seeds are hard-cased and fire-resistant; even its saplings have a high degree of fire resistance. Sometimes pure stands of teak develop.

Perhaps even more widespread in monsoon forests are bamboos, which frequently grow in pure stands over wide areas. Where burnt over during the dry season, their root clumps quickly send up new shoots during the rainy season. In the Lesser Sundas, Sulawesi, and the Philippines, areas of eucalyptus forests resembling teak forests in origin have developed as migrants from Australia. *See* BAMBOO.

The monsoon dry forest, an open woodland of very few species, is associated with the poor soils and low rainfall—generally not much more than 40 in. (100 cm)—found in large areas of Burma, Thailand, and Vietnam. In densely populated regions such as India, cultivation has made large parts of the monsoon forest zone treeless.

Hot deserts. Much of the lower Indus valley in Pakistan, along with extensive regions in southwestern Asia—in Arabia and in the lands around the Persian Gulf—are areas of hot desert. Vegetation is sparse and strongly drought-resistant. Except where irrigated, these deserts are lightly populated, although land-use systems involving grazing have placed considerable stress on the land.

Cold deserts and semidesert regions. Large areas of central Asia are desert, but unlike the deserts in southwestern Asia they experience cold winters. Plant life is very restricted, and there are large areas of bare rock. Shifting sand dunes are also common. The Gobi and Taklimakan are at lower elevations. The Tibetan and Iranian plateaus are likewise desert, and because of altitude even show an Arctic character in places.

In the Gobi, precipitation may be in the form of showers or protracted drizzles, but the tropical desert areas generally receive their meager rainfall in torrential downpours on rare occasions. Sudden floods rush through the dry washes known as wadis in Arabia but soon drain away or sink into the sands.

After such rains numerous herbs may spring to life and flower, while the bunch grass here and there may become green for a short season. Xerophytic shrubs such as wormwood and saxaul and varieties of sage are common in the desert. Stunted willows occur where fresh ground water is near the surface. In the wetter climatic fringes, a thin grass cover appears.

The preceding descriptions serve to demonstrate that even in broad characterizations the Earth's largest landmass exhibits the most varied of physical characteristics.

Herold J. Wiens; Jerome D. Fellmann; W. D. McTaggart

Bibliography. C. Hutchison, *Geological Evolution of Southeast Asia*, 1989; R. Jishun et al., *Geotechtonic Evolution of China*, Science Press, Beijing, and Springer-Verlag, Berlin, 1987; A. A. Meyerhoff et al., *China: Stratigraphy, Paleogeography, and Tectonics*, Kluwer, 1991; V. Pokshishevkii, *Geography of the Soviet Union*, 1974; K. Takahashi and H. Arakawa (eds.), *Climates of Southern and Western Asia*, 1981; T. R. Tregear, *An Economic Geography of China*, 1970; G. Trewartha, *Japan: A Geography*, rev. ed., 1965; U.S. Department of the Interior, Board on Geographic Names, *Gazetteer of the People's Republic of China: Pinyin to Wade-Giles, Wade-Giles to Pinyin*, 1979; R. K. Verma, *Geodynamics of the Indian Peninsula and the Indian Plate Margin*, 1991; S. Zhao, *Physical Geography of China*, 1986.

Asparagales

A large, widespread order of petaloid monocotyledons consisting of 29 families and about 26,000 species. Asparagales are clearly circumscribed in deoxyribonucleic acid (DNA) sequence analyses but are difficult to define morphologically, and separation from Liliales has proved particularly problematic. Phytomelan, a dark seedcoat pigment present in most families of Asparagales but not found in other plants, is an obvious characteric of this order, although seeds of the largest family, Orchidaceae, and a few other taxa lack phytomelan. Most Asparagales are herbaceous perennials, but there are some vines including some species of asparagus (Asparagaceae), and woody taxa including aloes (Asphodelaceae). The order contains many horticultural taxa, including members of Orchidaceae (orchids), Amaryllidaceae (daffodils, belladonna lilies, and others), Iridaceae (irises, gladiolus, freesias), Convallariaceae (lily of the valley, Solomon's seal), and Hemerocallidaceae (daylilies), in addition to food crops including onions, leeks, and garlic (Alliaceae). *See* FLOWER; LILIIDAE; LILIOPSIDA; MAGNOLIOPHYTA; ORCHID; ORCHIDALES; PLANT KINGDOM.

Michael F. Fay; Mark W. Chase

Asparagus

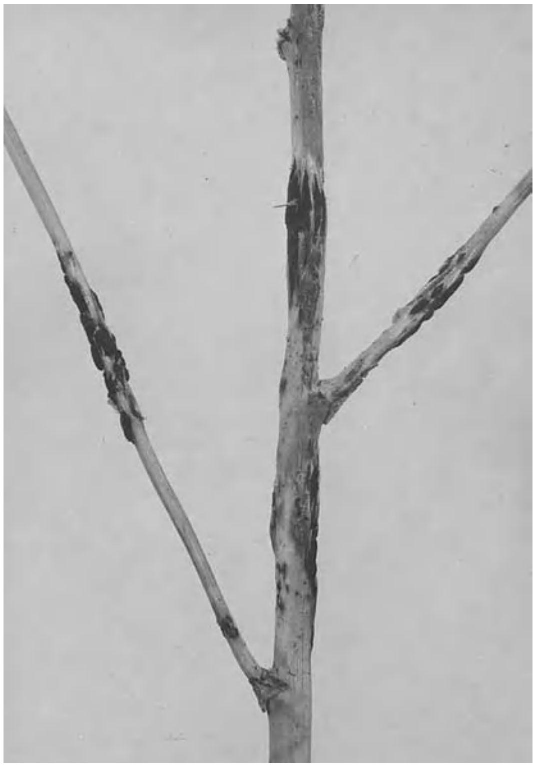
A dioecious perennial monocot (*Asparagus officinalis*) of Mediterranean origin belonging to the plant order Liliales. Asparagus is grown for its young shoots or spears, which are canned, frozen, or cooked fresh as a vegetable. These aerial stems arise from rhizomes (underground stems). The rhizomes and the fleshy and fibrous roots constitute the massive underground part of the plant. *See* LILIALES.

Propagation. Asparagus is propagated by seed with 1-year-old crowns transplanted to the field and spaced 8-18 in. (20–45 am) apart in 4–8 ft (1–2 m) rows. Although male female plants, separating crowns on the basis of sex has not been economical. Mary Washington is the principal variety cultivated; several new varieties developed by the University of California are widely planted. Blanched or white asparagus is grown by ridging soil over the rows and cutting the spears beneath the soil surface. Chemical weed control is commonly weed.

Harvesting. The length of the annual harvest season varies with age of the crowns and with climatic conditions. Generally, spears are cut for 8–10 weeks each spring after the crowns are 3–4 years old. In area with longer growing seasons, such as California, harvesting begins earlier and continues 10–12 weeks. Commercial plantings are often harvested for 12–16 years. In most areas special knives are used for harvesting; however, spears for canning and freezing are sometimes snapped off by hand above the ground level. Several types of mechanical harvesters are now in commercial use.

Commercial production is limited to areas where crowns will have a dormant period of 3–5 months each year. Dormancy in the northern states is induced by low temperatures and in California by withholding irrigation. California, New Jersey, and Washington are important asparagus-producing states.

H. John Carew



Asparagus rust. (*Courtesy of Arden F. Sherf*)

Diseases. Asparagus yields can be reduced by two fungal diseases, rust and *Fusarium*-induced wilt and root rot complex.

Asparagus rust (see **illus**.) is caused by a long-cycle autoecious rust fungus, *Puccinia asparagi*. The uredinial stage of the fungus is the most damaging and appears in the summer on stalks that grow after the harvest season. Uredinial lesions appear as oval reddish-brown pustules. Severe infection for 2 or 3 consecutive years reduces the vigor of the crowns and results in lower yields. Growers are advised to plant resistant varieties. Applications of fungicidal sprays in summer and fall may give partial control.

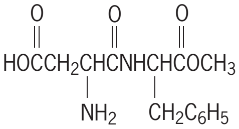
The Fusarium wilt and root rot complex is caused by two species of *Fusarium*. The fungi soil-borne, and infections occur in the roots or crowns. In addition, airborne spores can infect through insect injuries or other lesions near the base of the stalks. The fungi are also seed-borne. Affected plants become wilted, stunted, and yellowish-brown in color. Red streaks occur in the crowns, and the roots rot. In an established planting, individual diseased plants may be scattered among healthy ones, but eventually entire fields may be destroyed.

Control practices for Fusarium wilt and root rot include seed treatment, the planting of disease-free crowns, and planting in fields where asparagus has not previously been grown. *See* PLANT PATHOLOGY.

Thomas H. Barksdale

Aspartame

A white, crystalline compound, 1-aspartyl-1-phenylalanine methyl ester (APM), with formula (**1**). It is slightly soluble in water. Its sweetening properties were discovered accidentally in 1965 when the compound, a dipeptide, was produced as an intermediate in the synthesis of the C-terminal tetrapeptide of gastrin. Aspartame is the L,L-diastereoisomer; the three other possible diastereoisomers are not sweet. The taste of aspartame would not have been predictable based on its component amino acids, aspartic acid and phenylalanine.

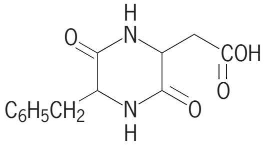


$$\mathrm{HO}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{CH_2}\underset{\mathrm{NH_2}}{\underset{|}{\mathrm{CH}}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{NH}\underset{\mathrm{CH_2C_6H_5}}{\underset{|}{\mathrm{CH}}}\overset{\mathrm{O}}{\overset{\|}{\mathrm{C}}}\mathrm{OCH_3}$$

(**1**)

The sweetness of aspartame relative to sucrose is a function of the latter's concentration, and is also dependent upon the presence of other flavors and materials. In a number of applications, such as chewing gum and various fruit-flavored products, aspartame favorably extends and enhances the flavor perception, and it shows synergy with other sweeteners. The sweetness perception may also last longer with aspartame than with sucrose or other sweeteners. *See* SUCROSE; TASTE.

Aspartame is metabolized to its component amino acids, which are further metabolized by the usual metabolic pathways. Under certain conditions of heat and pH in aqueous solution, aspartame is transformed into its diketopiperazine derivative, 3,6-dioxo-5-benzyl-2-piperazineacetic acid (**2**), which is



(**2**)

tasteless. This property limits the use of aspartame when it is exposed to high temperatures, such as in baking. The stability of aspartame in aqueous solution is pH-dependent; it is most stable at a pH of approximately 4. The rate of conversion (its half-life is 262 days at 77°F or 25°C) is sufficiently slow under the conditions of normal use that aspartame has found an increasing number of applications in various food products, and is particularly successful in soft drinks. The safety of aspartame has been established by studies in animals and human beings. Aspartame has been approved in many countries for uses in both dry and wet applications. *See* FOOD ENGINEERING.

Daniel L. Azarnoff

Asphalt and asphaltite

Varieties of naturally occurring bitumen. Asphalt is also produced as a petroleum by-product. Both substances are black and largely soluble in carbon disulfide. Asphalts are of variable consistency, ranging from a highly viscous fluid to a solid, whereas asphaltites are all solid. Asphalts fuse readily, but asphaltites fuse only with difficulty. Asphalts may, moreover, occur with or without appreciable percentages of mineral matter, but asphaltites usually have little or no associated mineral matter. *See* BITUMEN.

Natural occurrence. Many asphalts occur as viscous impregnations in sandstones, siltstones, and limestones. Most such deposits are thought to be petroleum reservoirs from which volatile constituents have been stripped by exposure of the rock.

The asphaltites (gilsonite, graphamite, and glance pitch) were probably derived from a saline lacustrine sapropel and owe their variable properties to differences in environment of deposition. These substances occur on a large scale in the Uinta Basin of northeastern Utah, where they are derived from upper Eocene Green River sediments, most of which are oil shales high in carbonate content. *See* OIL SHALE; SAPROPEL.

Asphalt. Relatively pure asphalt occurs in Kern, San Luis Obispo, and Santa Barbara counties, California. The asphalt, or tar seep, may contain up to 30% mineral matter and occasionally, as in the La Brea tar pits, has remains of insects or animals that became entrapped in the viscous material. Many of these occurrences are associated with the Miocene Monterey shale, which is thought to have been a source bed for the local crude oil. Occurrences of asphalt are also known in Kentucky and Oklahoma. Although asphalt seeps have long been known in France, Greece, Russia, Cuba, and other countries, the best known and largest are those of Venezuela and Trinidad.

Venezuelan asphalt has been recovered from the Bermudez Pitch Lake covering over 1000 acres (400 hectares) and averaging 5 ft (1.5 m) in depth. The lake is thought to be fed by asphalt springs. The asphalt at the edges of the lake becomes hard enough to walk on. After treatment to remove water and volatile constituents, the asphalt fuses at 130–140°F (54–60°C), is 92–97% soluble in carbon disulfide, and contains about 83% carbon, 11% hydrogen, 6% sulfur, and 1% nitrogen.

The Trinidad Asphalt Lake covers 115 acres (47 ha) and is 135 ft (41 m) deep at its center. The asphalt is softest at the center, where it is probably fed by underground springs, and hardest at the edges. Even at the center, however, the material is hard enough to support a person and can be broken into blocks. Gas is evolved from the asphalt, and rainwater tends to collect in depressions forming an emulsion of asphalt and water containing gas, sand, and clay. The crude material is 39% soluble in carbon disulfide, contains 27% mineral matter, and loses 29% of its weight as water and gas on heating to 100°F (38°C). After the refining process which drives off gas and water, the product fuses at about 190°F (88°C), is 56% soluble in carbon disulfide, contains 38% mineral matter, and consists of 80–82% carbon, 10–11% hydrogen, 6–8% sulfur, and less than 1% nitrogen, all on an ashfree basis. The Trinidad lake contains about 10,000,000–15,000,000 tons (9,000,000–14,000,000 metric tons) of asphalt.

Sandstones impregnated with asphalt occur in Oklahoma, Kentucky, Arkansas, Alabama, Utah, California, and other states. Many of these deposits have been developed as a source of paving material. Similar deposits also occur in Canada, South America, Europe, Asia, and Africa. One of the most extensive occurrences is that of Asphalt Ridge near Vernal, Utah, where asphalt saturates sandstones of the Uinta formation of Eocene age and of the Mesaverde formation of Upper Cretaceous age. This deposit crops out over a distance of 11.5 mi (18.5 km) and the sandstone contains 8–15% asphalt. It has been estimated that over 1,000,000,000 tons (900,000,000 metric tons) of high-grade material is readily available. The asphaltic sandstone is scooped from the deposit and used directly in road construction. If necessary, it is mixed with sand to attain the desired consistency.

Most asphalts are of marine origin and consist of the high-molecular-weight compounds normally present in petroleum residues. Asphalts and

TABLE 1. Major physical differences of asphaltite groups

Group	Specific gravity at 77°F (25°C)	Softening point, °F (°C)
Gilsonite	1.03–1.10	230–350 (110–177)
Glance pitch	1.10–1.15	230–350 (110–177)
Grahamite	1.15–1.20	350–600 (177–316)

TABLE 2. Asphalts and their uses

Asphalt type and % of production	Manufacturing process	Properties	Uses
Straight-run, 70–75%	Distillation or solvent precipitation	Nearly viscous flow	Roads, airport runways, hydraulic works
Air-blown, 25–30%	Reacting with air at 400–600°F (204–316°C)	Resilient; viscosity less susceptible to temperature change than straight-run	Roofing, pipe coating, paints, underbody coatings, paper laminates
Cracked, less than 5%	Heating to 800–1000°F (427–538°C)	Nearly viscous flow; viscosity more susceptible to temperature change than that of straight-run asphalt	Insulation board saturant, dust laying

asphaltites often contain unusually high percentages of vanadium.

Asphaltites. These substances are divided into three groups: gilsonite, glance pitch, and grahamite. The major physical differences in these substances are in specific gravity and softening point, as shown in **Table 1**. All three substances are nearly completely soluble in carbon disulfide. *See* IMPSONITE; WURTZILITE.

Differentiation of the asphaltites into three groups is based only on physical properties and not on a genetic basis. For this reason, similarly categorized substances may have somewhat different origins and variable compositions.

Glance pitch occurs on Barbados, and material from this deposit has been marketed as manjak. Other veins of glance pitch, some of which contain up to 27% mineral matter and up to 7.4% sulfur, also occur in Haiti, Cuba, Mexico, Argentina, Colombia, Chile, the Baltic states, and the Near East. Glance pitch has been used to make lacquers.

Grahamite occurs in West Virginia, Texas, Oklahoma, and Colorado. It is also known in Mexico, Cuba, Trinidad, Argentina, and Peru. The Peruvian grahamite is particularly rich in vanadium, and some vanadium minerals are associated with it. In general, most deposits are relatively small and are no longer of commercial interest. Irving A. Breger

Petroleum by-product. Asphalt is derived from petroleum in commercial quantities by removal of volatile components. It is an inexpensive construction material used primarily as a cementing and waterproofing agent. Over 27,000,000 tons (24,000,000 metric tons) of asphalt is used in the United States annually, of which more than 98% is derived from petroleum. *See* PETROLEUM PRODUCTS.

Asphalt is composed of hydrocarbons and heterocyclic compounds containing nitrogen, sulfur, and oxygen; its components vary in molecular weight from about 400 to 5000. It is thermoplastic and viscoelastic; at high temperatures or over long loading times it behaves as a viscous fluid, at low temperatures or short loading times as an elastic body.

The three distinct types of asphalt made from petroleum residues and their uses are described in **Table 2**.

In the construction of pavement surface for major roads and airport runways, hot asphalt is mixed with hot graded-stone aggregate. The mixture is spread on a dense, compacted stone base and rolled while still hot to give a smooth surface.

Roads having only light traffic are often given a thin, inexpensive wearing surface by spraying fluid asphalt on the road base and covering it immediately with stone. The fluid asphalt may be hot paving asphalt or a liquid asphalt. Liquid asphalts are produced by blending asphalt with various petroleum distillates or by emulsifying hot asphalt with water containing a small amount of soap. The liquid asphalts are fluid at ambient temperatures but harden as the solvent or water evaporates. *See* PAVEMENT.

Air-blown asphalt is used mainly for roofs. Hot asphalt may be mopped on the roof and covered with decorative gravel, or prefabricated asphalt shingles may be nailed onto the roof. *See* ROOF CONSTRUCTION.

Asphalt is also used in hydraulic works to line canals and reservoirs, to face dams and dikes, and to bind together the rocks in breakwaters.

Thomas K. Miles

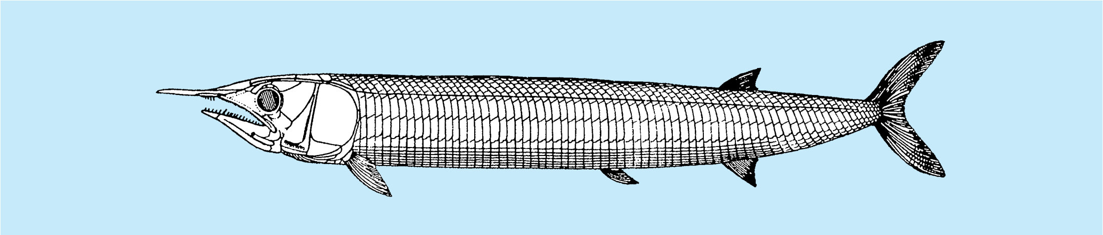
Aspidogastrea

A group of entoparasites considered to be a subclass or order of the Trematoda. They have strongly developed ventral holdfasts whose alveolations are taxonomically important. Two families, Aspidogastridae and Stichocotylidae, are recognized. Aspidogastridae, which are the commonest, occur in various cavities of mollusks and in digestive tracts of fishes and turtles. Development is usually direct, involving nonciliated juveniles and one host. Fishes and turtles probably acquire infestations of adult worms by ingesting parasitized mollusks. Elongate Stichocotylidae have a single row of alveoli and occur in the digestive tracts of skates. Juvenile *Stichocotyle nephrops* encyst in lobsters and develop to maturity after ingestion by skates, thus approaching the digeneid life cycle. Little is known of the physiology of aspidogastreids, but they appear less host-specific than other trematodes. *See* DIGENEA; TREMATODA.

William J. Hargis, Jr.

Aspidorhynchiformes

A small order of specialized holostean fishes which are first recorded from Middle Jurassic deposits of Europe. Later records indicate that they were a very successful group in terms of geographic range, and



***Aspidorhynchus acutirostris*, Upper Jurassic, Bavaria; a typical fish of the order Aspidorhynchiformes.**

probably had a worldwide distribution in the warm seas of the Cretaceous Period. The order contains one family, Aspidorhynchidae, and two genera, *Aspidorhynchus* and *Belonostomus*. These fishes, some of which reached a length of over 3 ft (0.9 m), are characterized by a ganoid scale covering with much deepened scales along the flank, by an elongate fusiform body and head with long slender snout, and by an externally symmetrical tail. All the fins are small, and fringing fulcra are reduced or absent. The dorsal and anal fins are positioned opposite one another far back on the body, and the pelvic fins are inserted closer to the anal than to the pectorals (see **illus**.). The vertebral column shows a fair amount of ossification with thin ringlike or half centra. *See* HOLOSTEI.

Noteworthy features of the skull are the small supraoccipital in the neurocranium, a large preopercle, a single series of plates behind the eye which cover the cheek region, and a mandible bearing closely set sharp teeth, with a single median (predentary) bone in front of the symphysis.

Aspidorhynchiforms are an isolated group and their relationships are not clear. They do not appear to have arisen from the earlier but superficially similar chondrosteans of the family Saurichthyidae because of substantial differences in the opercular series between the two groups. A relationship with the Pholidophoriformes has been suggested by some authors because of several skull characters that are shared. *See* CHONDROSTEI; PHOLIDOPHORIFORMES.

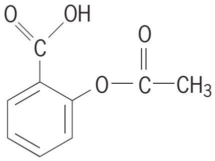
In the body form, fin position, and elongated snout, the aspidorhynchiforms resemble some of the living teleostean Exocoetoidei (needlefishes and sauries). It seems likely that these two widely separated and unrelated groups of fishes shared a similar mode of life, being predaceous open-water forms that utilized their long snouts and strong swimming ability in capturing prey. *See* OSTEICHTHYES; TELEOSTEI.

Ted M. Cavender

Bibliography. R. M. Alexander, *Chordates*, 2d ed., 1981; D. V. Obruchev, *Fundamentals of Paleontology*, 1967; A. S. Romer, *Vertebrate Paleontology*, 3d ed., 1966.

Aspirin

The acetyl ester of salicylic acid, also known as 2-(acetyloxy)-benzoic acid and acetylsalicylic acid. Aspirin is prepared by the acetylation of salicylic acid with acetic anhydride. Its structural formula is below.



Pain relief by the bark of the willow tree has been known since antiquity. In principle, therefore, aspirin is a natural remedy. Various techniques to extract the active principle, salicin, from the bark and to prepare salicylic acid were developed during the first half of the nineteenth century. During the second half, salicylates were used to preserve food, to control pain and fever, and to relieve gout. Aspirin was introduced into practice around 1899, mainly through the efforts of chemist Felix Hoffmann.

Measurable amounts of the acetylated compound are absorbed from the stomach, but most of it is absorbed (after hydrolysis) from the small intestine; none is absorbed from the large intestine. Once absorbed, aspirin is deacetylated; the peak salicylate level occurs within 1 or 2 h after absorption. Salicylic acid binds to plasma protein. Aspirin is excreted as a glucuronide, salicyluric acid, or gentisic acid. Many individual differences exist for the absorption, tolerance, and therapeutic usefulness of aspirin.

Aspirin is effective as an analgesic, antipyretic, and antiinflammatory drug. It prevents the aggregation of platelets, and there is some evidence that it can prevent stroke. Aspirin, if tolerated, is the preferred drug for the treatment of rheumatoid arthritis, and it has been used in the treatment of osteoarthritis. Aspirin lowers fever, probably by acting on the hypothalamus. Salicylates inhibit aldose reductase in the lens; it has been suggested that they might retard the development of cataracts. Aspirin might encourage the development of Reye's syndrome, an acute encephalopathy which occurs in children who recover from viral disease, but this cause-and-effect relationship remains to be confirmed. *See* ANALGESIC; ARTHRITIS.

Arachidonic acid is a cyclic fatty acid which is the raw material for prostaglandins and leukotrienes. Transformation of arachidonic acid into prostaglandins requires cyclooxygenase; transformation into leukotrienes requires lipoxygenase. Aspirin is known to inactivate cyclooxygenase by acetylation, but has no effect on lipoxygenase. It has been suggested that the inhibition of cyclooxygenase by aspirin offers lipoxygenase a more-than-normal access to arachidonic acid; and thus could induce an exaggerated production of leukotrienes. Since some leukotrienes are potent bronchoconstrictors, they have been thought to explain the violent reactions to aspirin in aspirin-sensitive patients, but some research has cast doubt on this hypothesis.

Intolerance to aspirin is not uncommon. It tends to develop in middle age and involve the skin or the respiratory tract, or both. In the skin, it causes urticaria. In the respiratory tract, nasal polyps precede the development of aspirin sensitivity. Intermittent (but occasionally progressive) bronchial asthma might occur. Death rarely ensues because people rapidly become aware of their intolerance.

The clinical symptoms of aspirin sensitivity are almost identical with those encountered in clinical allergy and anaphylaxis. Consequently, reactions caused by aspirin have been thought to represent an aspirin allergy, that is, an antibody-mediated sequence, although aspirin-sensitive individuals do not have an increased titer of immunoglobulin E (IgE). Intolerance to aspirin may be the result of a specific disease which alters receptors. Aspirin-sensitive individuals also react to nonsteroidal antiinflammatory drugs, and, occasionally, to the food color FD&C yellow no. 5, which is a pyrazolone derivative. *See* ALLERGY; ANAPHYLAXIS; IMMUNOGLOBULIN; PAIN.

Max Samter

Bibliography. L. Lasagna and F. G. McMahon (eds.), New perspectives on aspirin therapy, *Amer. J. Med.,* June 1983; M. Samter, Intolerance to aspirin, *Hosp. Pract.,* 8:85–90, 1973; M. Samter (ed.), *Immunological Diseases,* 1978; M. Samter and R. F. Beers, Jr., Intolerance to aspirin: Clinical studies and consideration of its pathogenesis, *Ann. Int. Med.,* 68:975–983, 1968.

Assembly machines

Machines that take discrete components as come into an assembly department and bring them together so as to produce a configuration of some practical value. Such machines differ from packaging machinery in two ways: in assembly machinery, components must be inserted in specific sequence and spatial attitude; and they must often be tested functionally as part of the assembly process.

Assembly machinery was originally conceived for situations where volume or hazard of production, parts size, or availability of labor made manual assembly impractical from an operational or economic viewpoint. The early applications of assembly machinery and the majority of modern applications are found in the automotive industry, consumer products, manufacturing, or hazardous assembly.

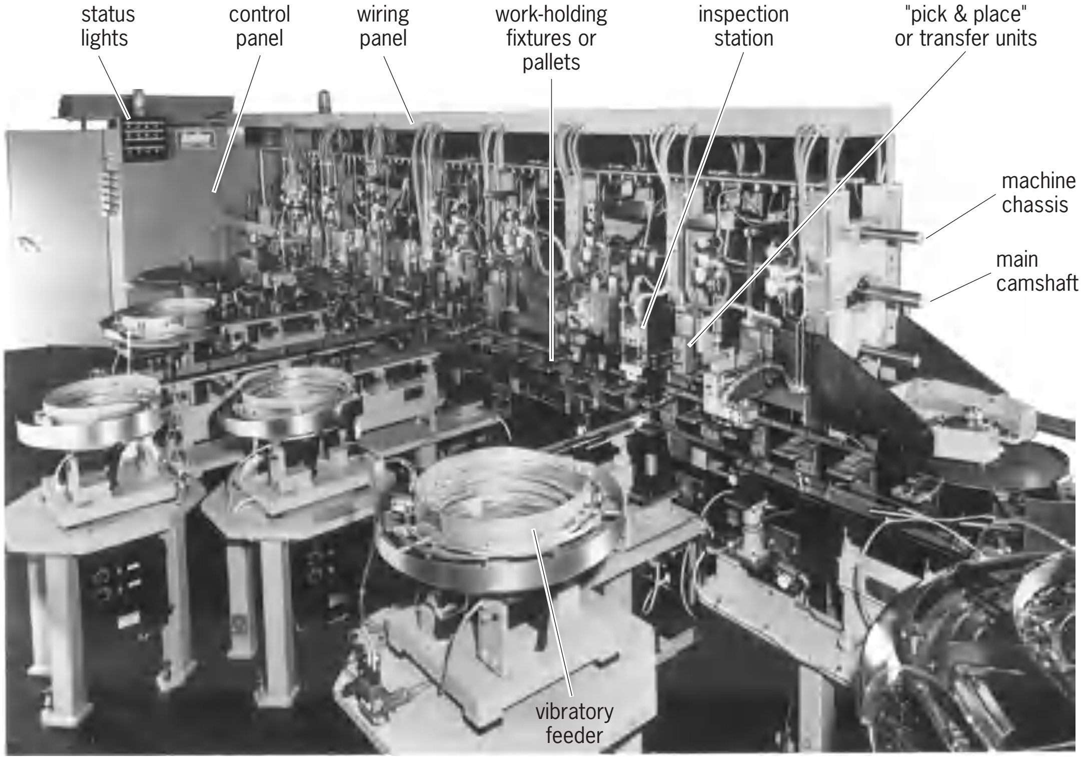
Although automatic assembly machinery competes with inexpensive manual labor in emerging countries or depressed areas in mature industrial countries, manufacturers have different attitudes toward justification of assembly machinery because of single world economy, rising costs in developing countries, and consumer attitudes toward product quality. Assembly machinery is often justified on its increased quality, reduced in-process inventory, and rapid response to customer specification, not on labor savings alone.

Types. Assembly machinery can be classified in several ways, including work path (rotary, carrousel, or linear); index–dwell ratios (continuous motion, intermittent motion, or power and free); actuation (mechanical, fluid power, or electronically programmable); work nest configuration (pallet or walking beam); and design (special or standard modular). The engineers who specify assembly machinery will usually select from one or more of these categories based on product size and weight, volume of production, product life cycle, future and present flexibility needs, human resources, and return on investment. Choice among these types of machines is restricted by choice of builders, industry backlog, and importance of prior operating experience.

Work path. Rotary dial machines have a number of pallets fastened to a rotating dial or ring. Transfer devices are usually mounted to the machine base, while feeders are placed outside the periphery of the dial. Advantages are low initial cost, high cyclic speed, low inertia, and good index accuracy (the ability of a machine to present the work-holding fixtures or pallets to a specific operating station). Disadvantages include limited number of work stations, lack of future flexibility, congested work stations, and poor access for maintenance. Rotary machines usually have high operating costs, primarily due to intensive use of pneumatic components and the high downtime required for maintenance.

Carrousel machines are usually configured in race track shape. Like rotary machines, they usually have fixed index–dwell ratios. Work-holding nests (the machine elements used to hold all of the parts being assembled) are secured to one another (a configuration known as precision link) or fastened to a chain. Where high index accuracy is required, they are fastened to stressed steel bands. Modern carrousel machines usually have a high level of mechanically actuated and synchronized work stations, driven from one or more lateral main camshafts. Highly modular, they can be reconfigured readily to meet product modifications or even product changes. Advantages include a larger number of work stations and good maintenance characteristics. Disadvantages include high capital cost, limited line-balancing ability and, with the exception of stressed band machinery, modest index accuracy. They have good to excellent operating costs. Carrousel machines may have pallets in cantilever form around the periphery of the machine or may be built in an over-under configuration, a system in which the fixtures pass the various functional stations on the machine and then go over a drum or wheel for return to start a new cycle.

Linear machines (see **illus.**) have an open-loop configuration in which pallets transport the product being assembled along one or more linear paths,



Linear synchronous assembly machine.

rather than in a closed circuit such as rotary or carrousel machines. Most linear machines are power and free systems in which each individual station is activated by the presence of the work-holding pallet before it.

Index–dwell ratios. Motion of a pallet or nest is called index, and lack of motion is called dwell. Together they form one machine cycle. In most machines, the index portion of the cycle is used by transfer devices to acquire components from feeding mechanisms. During the dwell portion of the cycle, parts are inserted, and joining operations and inspections are performed. In general, the dwell should exceed the index portion of the cycle. The ratio of dwell to index is often 1:1 at high cyclic rates, but the ratio will be much higher at low cyclic rates.

In any form of automatic assembly, production rates will be controlled by the single longest operation. If new parts can be added to the assembly without the fixture being stopped, the machine can be operated in a continuous motion. Usually such equipment is suitable for assemblies in which component parts do not need radial orientation, for example, deodorant dispensers and simple aerosol valves. Continuous-motion machines are capable of extremely high rates of production, but their application range is very restricted.

Power and free machines are sophisticated conveyor systems in which work pallets are transferred from work station to work station by free-flow conveyors, walking beam systems, or motorized pallets. These transfer machines are specifically indicated where line-balancing requirements are severe and fixture/workpiece weight is high (for example, above 11 lb or 5 kg). Primary disadvantages include high procurement costs and high engineering costs. Advantages include high index–dwell ratios (since independent stations operate only when a fixture pallet is present) and good ability to interface people among automatic sequences; line balancing can be achieved by parallel paths or redundant stations.

Actuation. Initiation of movement of machinery can be achieved by pneumatic or hydraulic fluid power, cams or electric driving motors, solenoids, or other rotary or linear actuators. Fluid power is the least costly to build, but it has the highest operating cost and often has high noise levels. Such machines usually have sequential controls. Cam-operated machines are quiet and highly reliable, and over prolonged periods they have low operating costs. They require more engineering and have a higher purchase price. Electrically actuated machines are common in Europe and Asia. They have very complex control systems including significant interlocking systems. These machines are often specified because they are perceived to have a high level of flexibility.

Machine design. There are significant differences in philosophy of machine design between European, Asian, and North American builders and between in-house and commercial builders. In-house builders (more typically found in Europe and Asia) tend to consider each assembly machine project a unique task, while commercial builders consider assembly a set of generic tasks capable of being done by generic modular stations. Commercial builders claim greater operational up time because of the reliability of proven modules, and greater flexibility for product modification because of the generic station design.

Tasks. Assembly machines perform several tasks: parts feeding (including orientation, separation, and transfer); parts joining (for example, welding, riveting, and soldering); parts inspection (condition, parameters, presence, and position); functional testing (for example, capacitance, torque, and pressure decay); marking (date coding, model number, and operational characteristics); and ejection (in controlled or uncontrolled positions). Many of the tasks of assembly machinery such as joining, functional testing, and product marking use commercial units identical to those used on manual lines. Commercial screwdrivers, orbital riveters, ultrasonic welders, vision systems, leak testing, printers, laser markers, and other commercial units can be adapted to automatic assembly machinery with little or no modification.

The unique task of automatic assembly machinery is that of parts feeding, which consists of accepting the component parts as they come to the assembly area and taking one of these components, separating that component from other parts, changing its spatial orientation to a usable insertion attitude, and transferring it from the orientation device into the work-holding fixture or into a partially completed assembly.

Some parts can be fabricated directly on the assembly line. This is true of paper or film components such as gaskets, diaphragms, and labels. Material is brought to the machine in reels and is fabricated directly on the machine. Some fragile parts are often preformed and rewound in reels or bandolier devices for cutoff or separation at the assembly machine. Transistors, resistors, and capacitors are often fed this way, although many are fed from special tubes. Many fragile mechanical parts or subassemblies usually stored in X-Y coordinate trays can also be stored more efficiently in plastic chainlike bandoliers.

Magazines. These are often used to hold flat parts or to store fragile subassemblies, and take many shapes. Certain parts that are not easily damaged can be fed by pushing one part from the stack with a technique referred to as coin changing. Parts such as motor or transformer laminations are usually fed in this mode. Other more fragile parts are fed from elevator-type magazines in which one part is lifted from the top of the stack without lateral motion relative to the other parts in the stack.

All magazine feeding systems are dependent on several technical and economic conditions. Of the economic questions, the most critical is proximity to the fabrication operations. Magazine racks, tubes, and containers are expensive, and the quantity required is proportional to the distance from fabrication to assembly. Technically, magazines are the most desirable option when the component can be ejected from a primary operation directly to the assembly machine magazine storage device without any secondary operation. Parts requiring secondary operations or manual loading of magazine tubes do not enjoy the same savings as those in which magazine loading is done at the primary fabrication operations.

Most parts coming to the assembly floor do not come in bandoliers or in convenient magazine carriers. Fragile parts are usually stored in X-Y coordinate trays or boxes requiring some level of coordinate programming (including robots) for parts retrieval. This storage is very expensive and adds greatly to machine development costs. Larger parts such as engine blocks, transmission housings, and motor frames are usually stored between fabricating, surface finishing, and assembly in overhead storage conveyors. Retrieval of such parts is often manual but is increasingly done by robots.

Parts feeders. Most of the parts coming to the assembly floor arrive in unoriented bulk condition. While ideally such parts should be captured at fabrication and carried in an oriented condition, economic reality usually makes this impractical. Many parts require a variety of secondary operations. Vast differences between fabrication cycle rates and assembly rates often dictate multiple fabricating machines, molds, or dies. Components often come to the assembly floor from outside vendors quite remote from the assembly area. When such parts are durable and the pipeline of supply is extensive, bulk unoriented shipment of components is the normal mode of delivery.

To be able to feed such parts, a variety of feeders have been developed over the years. Basically such parts feeders, operating by mechanical, vibratory, or centrifugal means, move a large number of unoriented parts and discharge them onto selective devices in feeder bowls or discharge tracks. Those parts with usable attitudes are captured and passed along to the discharge point of the feeder; those without are returned to the storage area of the feeder to try again.

Most early feeders used mechanical rotors or oscillating blades to move the parts. In some specialized applications, centrifugal forces move the parts. In most modern feeders, vibratory action is used to produce the motion of components. Vibratory feeders consist of bowls, trays, or tracks mounted on a series of springs at an angle to the plane of the bowl, tray, or track. Electromagnets are used to pull down the feeders, compressing the springs until the magnet is turned off, freeing the spring, and causing the parts container to move upward. This repeated cycling causes the components to be thrown upward and forward in either a circular or a linear unidirectional path. The parts, if properly oriented, proceed through a series of capturing and retaining devices until they are fed to the discharge point. The parts to be discharged are moved sequentially into an escapement device. Escapements are used to separate the part that is to be transferred and inserted from subsequent parts in feeder tracks. Escapements must accurately locate parts for pickup and transfer.

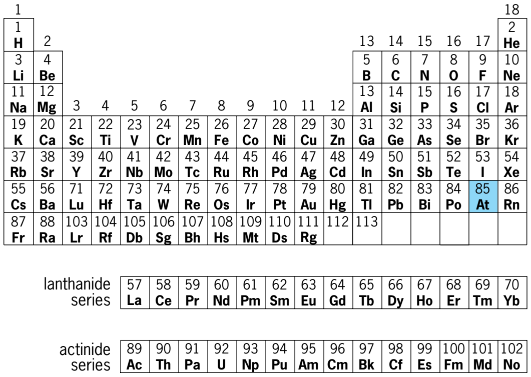
Unique aspects. Unlike fabricating machines, assembly machinery must handle discrete parts with significant variations in dimensional tolerance. All assembly machinery must be debugged prior to operation. This debugging is a normal component of assembly machine development, ensuring that the machine is capable of accepting a given spectrum of variations in component parts.

Assembly machines are usually one-of-a-kind systems, and they need extensive sensor networks to ensure that the machine is functioning properly. Few products are manufactured in such volumes as to permit full development of process reliability. The prototype nature of most assembly machinery requires that machine controls contain a large element of process verification not usually found in other industrial machinery. *See* MANUFACTURING ENGINEERING; PRODUCTION METHODS. Frank Riley

Bibliography. P. Niland, *Management Problems in the Acquisition of Special Automatic Equipment*, 1961; F. J. Riley, *Assembly Automation: A Management Handbook* 2d ed., 1996.

Astatine

A chemical element, At, atomic number 85. Astatine is the heaviest of the halogen groups, filling the place immediately below iodine in group 17 of the periodic table. Astatine is a highly unstable element existing only in short-lived radioactive forms. About 25 isotopes have been prepared by nuclear reactions of artificial transmutation. The longest-lived of these is ^{210}At, which decays with a half-life of only 8.3 h. It is unlikely that a stable or long-lived form will be found in nature or prepared artificially. The most important isotope, used for tracer studies, is ^{211}At. Astatine exists in nature in uranium minerals, but only in the form of trace amounts of short-lived isotopes, continuously replenished by the slow decay of uranium, The total amount of astatine in the Earth's crust is less than 1 oz (28 g).



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

In aqueous solution, astatine resembles iodine except for differences attributable to the fact that astatine solutions are of necessity extremely dilute. Like the halogen iodine, when astatine exists as a free element in solution, it is extracted by benzene. The element in solution is reduced by agents such as sulfur dioxide and is oxidized by bromine. It is more electropositive than the other halogens. It has oxidation states with coprecipitation characteristics similar to those of the iodide ion, free iodine, and the iodate ion. Powerful oxidizing agents produce an astatate ion, but not a perastatate ion. The free state is most readily obtained and is characterized by high volatility and high extractability into organic solvents. *See* HALOGEN ELEMENTS. Earl K. Hyde

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. F. Shriver and P. W. Atkins, *Inorganic Chemistry*, 3d ed., 1999.

Asterales

An order of flowering plants, division Magnoliophyta (Angiospermae), which gives its name to the subclass Asteridae in the class Magnoliopsida (dicotyledons). The Asterales have often been included in the order Campanulales, but they are perhaps more closely allied to the Rubiales and Dipsacales. The order consists of only the single very large family Asteraceae (Compositae), with about 20,000 species, occurring in nearly all parts of the world but most abundant and conspicuous in areas which are not densely forested. *See* CAMPANULALES; DIPSACALES; RUBIALES.

The Asterales are marked by their inferior ovary, single basal ovule, specialized pollen presentation mechanism, and pseudanthial, centripetally flowering heads which often have specialized marginal flowers with a strap-shaped corolla resembling the petal of an ordinary flower. The anthers are introrsely dehiscent and connate (or at least connivent) into a tube around the style, which grows up through the anther tube and pushes out the pollen. A more or

New England aster (*Aster novae-angliae*), a characteristic member of the order Asterales. The numerous apparent petals around margin of flower head are corollas of individual flowers, called ray flowers in contrast to disk flowers which occupy the center of head. (*Courtesy of Alvin E. Staffen, National Audubon Society*)

less similar pollen presentation mechanism occurs in many of the Campanulales and some of the Rubiales, in the same subclass.

Most members of the order are herbaceous, but some, such as the sagebrush (*Artemisia tridentata*), are shrubs, and a few tropical species are trees. Many well-known garden ornamentals, such as aster (see **illus.**), chrysanthemum, dahlia, daisy, sunflower (*Helianthus*), and zinnia, belong to the Asterales. A few garden vegetables, for example, lettuce (*Lactuca*) and artichoke (*Cynara*), and some common weeds, such as dandelion (*Taraxacum*), thistle (*Cirsium*), and ragweed (*Ambrosia*), also belong to the order. *See* ARTICHOKE; ASTERIDAE; LETTUCE; MAGNOLIOPHYTA; MAGNOLIOPSIDA; ORNAMENTAL PLANTS; PLANT KINGDOM; SUNFLOWER.

Arthur Cronquist; T. M. Barkley

Asteridae

A large subclass of the class Magnoliopsida (dicotyledons) of the division Magnoliophyta (Angiospermae), the flowering plants, consisting of 11 orders, 49 families, and more than 60,000 species. The Asteridae are mostly sympetalous with unitegmic, tenuinucellate ovules and with the stamens usually as many as, or fewer than, the corolla lobes and alternate with them. Most of them have two carpels, but a few have as many as five or even more carpels, and a few others are pseudomonomerous. The largest orders of the group are the Asterales (about 20,000 species), Scrophulariales (about 11,000 species), Lamiales (about 7800 species), and Rubiales (about 6500 species). Other orders are the Gentianales, Plantaginales, Solanales, Callitrichales, Campanulales, Calycerales, and Dipsacales. See individual articles on each order. *See* MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM.

Arthur Cronquist; T. M. Barkley

Asteroid

One of the many thousands of small planets (minor planets) revolving around the Sun, mainly between the orbits of Mars and Jupiter. The presence of a gap in J. A. Bode's empirical law of planetary spacings motivated a search for the missing planet. G. Piazzi discovered Ceres on January 1, 1801, and three other small planets were discovered in the next few years. Visual and photographic searches for additional asteroids have continued to the present day, augmented by electronic detection technology. Newly discovered ones are assigned a catalog number and name (such as 433 Eros) only after they are observed often enough to compute an accurate orbit. The Minor Planet Center in Cambridge, Massachusetts, maintains a file of 21 million measurements of asteroid positions. The Institute for Theoretical Astronomy in St. Petersburg, Russia, publishes an annual ephemeris of predicted asteroid positions for the over 73,000 cataloged asteroids. *See* PLANET.

Orbits. The majority of asteroids have semimajor axes (mean distances to the Sun; symbolized a) between 2.2 and 3.2 astronomical units (1 AU = distance from Earth to the Sun = 1.496×10^8 km = 9.3×10^7 mi). However, numerous small asteroids orbit between Venus and Mars, and two large groups, the Trojan asteroids, orbit at Jupiter's distance from the Sun. *See* TROJAN ASTEROIDS.

In 1992, the first of the trans-Neptunian "asteroids" was discovered. Called Kuiper Belt Objects (KBOs) and Scattered Disk Objects, nearly 800 had been found by early 2004. They represent a population of bodies much more numerous than the main-belt or Trojan asteroids, but are more properly thought of as comets. They originally accreted from the primordial solar nebula (the disk of gas and dust from which the planets formed) beyond the orbit of Neptune, but never were incorporated into an outer planet. As a few of these objects slowly leak away and enter the inner solar system closer to the Sun, their ices sublime and they develop the heads and tails characteristic of comets. There are also a modest number of minor planets orbiting the Sun in temporary orbits beyond Jupiter but well inside the Kuiper Belt; they are termed Centaurs. Although there may be no sharp distinctions in physical properties between asteroids and outer solar system small bodies, and some old short-period comets may be counted among the near-Earth asteroids (NEAs), this article deals with classical asteroids that originated in the inner parts of the solar system: the Trojans, the main-belt asteroids, and the NEAs, which approach the Sun to within 1.3 AU. *See* COMET; KUIPER BELT.

Most asteroid orbits are more elliptical and inclined to the plane of the ecliptic than the orbits of major planets. Eccentricities (e) average about 0.15, and inclinations (i) about 10°; occasionally they exceed 0.5 and 30°, approaching the characteristics of short-period comet orbits. Near-Earth asteroids are of three different orbital types: Amors orbit beyond Earth's orbit and do not cross it; Apollos cross Earth's orbit but generally do not intersect it; and Atens orbit inside Earth's orbit.

Asteroids are not uniformly distributed in a, e, and i. Vacant lanes in the main asteroid belt, known as Kirkwood gaps (**Fig. 1**), occur at distances where the periods of revolution would be a simple fraction (such as $^1/_3$ or $^2/_5$) of Jupiter's period of 11.86 years. Asteroids in these gaps may have been preferentially removed when Jupiter's powerful gravity sent them into wild, chaotic orbits; in these orbits, they may have collided with the Sun, the Earth, or another planet. Most asteroids originally beyond the $^1/_2$ resonance were directly ejected from such orbits by nearby Jupiter early in the history of the solar system, except those grouped near the stable $^2/_3$ (Hilda group) and $^3/_4$ (Thule) resonances. Asteroid fragments continue to be ejected from the $^1/_3$ gap and from other resonant orbits. These chaotic processes, combined with forces due to asymmetric solar heating, are believed to be responsible for bringing most of the Earth-approaching asteroids and smaller meteorites close to the Earth. *See* CHAOS.

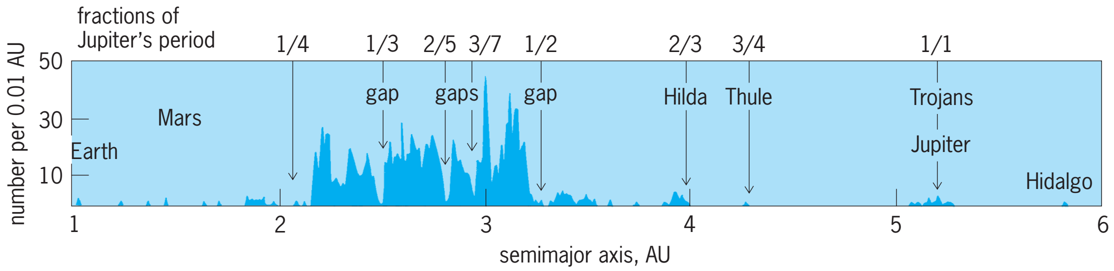


Fig. 1. Distribution of the brighter asteroids, with distance from the Sun, between the Earth and Jupiter. Major fractions of asteroid orbital periods to Jupiter's period are shown. The arrows point to associated clusters or Kirkwood gaps.

Clusterings of asteroids with similar *a, e*, and *i* are known as Hirayama families, named for the Japanese astronomer K. Hirayama who discovered some of the major ones in 1918. A few dozen families can be recognized. Some of the larger families are composed of asteroids that have the same colors, and are presumably made of the same minerals. Such a family probably consists of fragments of a larger, precursor asteroid that was broken apart by a catastrophic collision with another asteroid or comet. Some of these families are immersed in huge, torus-shaped dust belts, discovered by the *Infrared Astronomical Satellite* (*IRAS*). The colliding asteroids are gradually grinding each other down to dust, which then spirals into the Sun or is blown away by light pressure from the Sun. *See* ORBITAL MOTION.

Shapes, spins, and satellites. When the brightnesses of most asteroids are measured, they vary in a periodic manner, occasionally by more than one magnitude (a factor of 2.5), but more commonly 0.3 magnitude or less. The light curves usually are double-peaked (**Fig. 2**), characteristic of an irregularly shaped body spinning in space rather than the more complex curves that would result from albedo differences (spots) on a spherical body. Light curves have been measured for many hundreds of asteroids.

The spin periods are typically a few hours to about a day, but range from about 2 h to roughly a month. Asteroids smaller than about 500 ft (150 m) across spin much more rapidly, indicating that they are solid rocks. The fact that larger asteroids never spin faster than 2 h indicates that they are weak bodies, perhaps rubble piles held loosely together by gravity. If they were to spin faster, the centripetal acceleration at their equator would exceed their weak gravity and they would tend to fly apart. A few large asteroids spin very slowly; they are believed to have been slowed by asymmetric solar heating and reradiation. Collisions among asteroids are not so important in controlling the spins of medium-sized asteroids as was once thought. A few asteroids have variable periods or they wobble and tumble chaotically; some of them may have been struck recently by another asteroid.

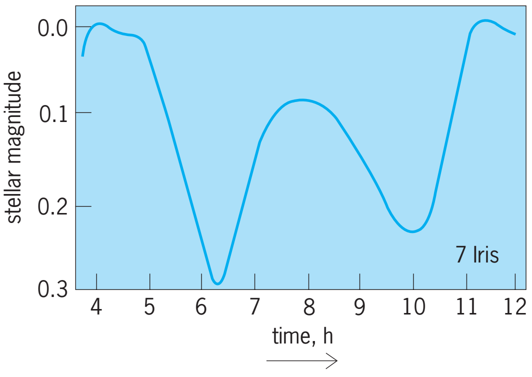


Fig. 2. Variation with time in the brightness of 7 Iris due to its irregular shape and 7.1-h rotation. (*After G. Kuiper*)

The very largest asteroids, especially those composed of weak materials, cannot maintain highly irregular shapes against the compression of their own gravitational fields, and are roughly spherical. Some of the larger, rapidly spinning asteroids have large light-curve amplitudes. They may be weak "rubble piles," previously fractured and partly disrupted by collisions and immediately reaccumulated into a pile of boulders. Some small asteroids, notably 4179 Toutatis, which was studied intensively by radar when it passed close to the Earth in 1992, have been found to have dumbbell shapes. They may actually be two bodies, lightly resting against each other.

In August 1993, the *Galileo* spacecraft flew past the main-belt asteroid 243 Ida and discovered that it possesses a mile-wide satellite, which was subsequently named Dactyl. Five years later, a second asteroidal satellite was discovered around 45 Eugenia with a telescope in Hawaii, using new adaptive optics technology. Since then (through early 2004), 13 more main-belt asteroids and one Trojan have been found by adaptive optics either to be a double object (binary) or to possess a satellite. Also, 10 near-Earth asteroids have been observed by radar to be binary or have a satellite, and another 10 near-Earth asteroids (and two more main-belt asteroids) show unusual light curves that strongly suggest a binary configuration. Only a few percent of main-belt asteroids have substantial satellites, which were probably produced by catastrophic collisions or impact cratering events. The fraction is much higher, maybe 20%, for near-Earth asteroids, most of which may result from tidal disruption during close passes to the Earth or Venus. *See* ADAPTIVE OPTICS; BINARY ASTEROID.

Sizes and densities. Improvements in radar technology make it possible to image small asteroids that pass close to the Earth almost as well as by spacecraft flybys. With the exception of these observations, no asteroid shows a disk large enough to measure

Noteworthy asteroids

Size rank	Number and name	Spectral type	Diameter, mi (km)*	Spin period, h	Orbital elements		
					a, AU	*e*	*i*, degrees
1	1 Ceres	G (C-like)	590 (949)	9.1	2.77	0.08	10.6
2	2 Pallas	B (C-like)	331 (533)	7.8	2.77	0.23	34.8
3	4 Vesta	Achondrite	329 (529)	5.3	2.36	0.09	7.1
4	10 Hygiea	C	253 (407)	27.6	3.14	0.12	3.8
5	511 Davida	C	200 (322)	5.1	3.18	0.17	15.9
6	704 Interamnia	F (C-like)	197 (317)	8.7	3.06	0.15	17.3
7	52 Europa	C	183 (295)	5.6	3.09	0.11	7.5
8	87 Sylvia	P	172 (277)	5.2	3.48	0.09	10.9
9	65 Cybele	P	167 (269)	4.0	3.43	0.11	3.6
10	15 Eunomia	S	161 (259)	6.1	2.64	0.19	11.8
11	16 Psyche	M	155 (249)	4.2	2.92	0.14	3.1
12	31 Euphrosyne	B (C-like)	154 (248)	5.5	3.15	0.23	26.3
13	451 Patienta	B (C-like)	153 (247)	9.7	3.06	0.07	15.2
14	3 Juno	S	150 (242)	7.2	2.67	0.25	13.0
15	324 Bamberga	C	149 (240)	29.4	2.68	0.34	11.2
16	13 Egeria	G (C-like)	139 (224)	7.0	2.58	0.09	16.5
17	624 Hektor	D	186 × 93 (300 × 150)	6.9	5.15	0.03	18.3
18	532 Herculina	S	137 (220)	9.4	2.77	0.17	16.3
19	107 Camilla	C	137 (220)	4.8	3.49	0.07	10.0
20	423 Diotima	P?	135 (217)	4.8	3.07	0.03	11.2
21	121 Hermione	C	135 (217)	5.6	3.46	0.14	7.6
22	45 Eugenia	F (C-like)	133 (215)	5.7	2.72	0.08	6.6
23	19 Fortuna	C	130 (210)	7.4	2.44	0.16	1.6
24	24 Themis	C	129 (207)	8.4	3.13	0.13	0.8
25	7 Iris	S	127 (204)	7.1	2.39	0.23	5.5
26	6 Hebe	S	126 (202)	7.3	2.42	0.20	14.8
27	702 Alauda	C	126 (202)	8.4	3.19	0.03	20.5
28	88 Thisbe	C	124 (200)	6.0	2.77	0.16	5.2
			Other interesting asteroids				
	41 Daphne	C	116 (187)	6.0	2.77	0.27	15.8
	44 Nysa	Aubrite?	42 (68)	6.4	2.42	0.15	3.7
	165 Loreley	C	99 (160)	7.2	3.14	0.07	11.2
	216 Kleopatra	M	87 (140)	5.4	2.79	0.25	13.2
	243 Ida	S	18.6 × 7.9 × 5.8 (29.9 × 12.7 × 9.3)	4.6	2.86	0.04	1.1
	250 Bettina	M	53 (86)	5.1	3.14	0.14	12.9
	349 Dembowska	Achondrite?	90 (145)	4.7	2.93	0.09	8.3
	433 Eros	Chon.	21.4 × 7.0 × 7.0 (34.4 × 11.2 × 11.2)	5.3	1.46	0.22	10.8
	747 Winchester	P	116 (186)	9.4	3.00	0.34	18.2
	951 Gaspra	S	11.3 × 6.5 × 5.5 (18.2 × 10.5 × 8.9)	7.0	2.21	0.17	4.1
	1566 Icarus	Chon.?	1 (2)	2.3	1.08	0.83	22.9
	1620 Geographos	S	1.3 (2.1)	5.2	1.24	0.33	13.3
	4179 Toutatis	S	~2 (4)	130	2.51	0.64	0.5

*Diameters are accurate to around 10%; size rankings may vary.

accurately by using Earth-based telescopes. For more distant asteroids the chief technique used to measure asteroid diameters is radiometry, which compares the brightness of reflected visible sunlight from an asteroid with the brightness of the asteroid's emitted thermal radiation in the infrared. In the mid-1980s, *IRAS* measured the diameters and albedos of over 2000 asteroids. Radiometric diameters have been checked by timings of occasional disappearances of stars when asteroids chance to pass between the Earth and a star. *See* ALBEDO; INFRARED ASTRONOMY; OCCULTATION.

There are about 30 asteroids larger than 124 mi (200 km) in diameter (see **table**); about 75% of them are soot black (geometric albedos of 3–5%). Asteroids are much more numerous at smaller sizes, generally following a size distribution characteristic of fragmentation processes, as would be expected if the asteroids were smashing into each other. Indeed, there are so many asteroids confined in the volume of the asteroid belt that collisions sufficient to fragment medium-sized asteroids occur every few billion years, and much more often for smaller ones. Thus all asteroids have been extensively battered, and most of the smaller ones are collisional fragments.

A few of the largest asteroids are massive enough to detectably change the orbits of other asteroids and even the orbit of Mars, from which the masses of the large asteroids can be determined. The total mass of all asteroids is less than 5% that of the Moon. For asteroids with measured masses and diameters (volumes), their bulk densities can be calculated. The resulting densities of Ceres, Pallas, and Vesta are roughly 2.1, 2.7, and 3.5 times that of water, respectively. Vesta's density is similar to that of ordinary rocks, but Ceres is as underdense as carbonaceous chondritic meteorites, which contain a large fraction of volatiles. Perturbations on the orbit of Mars, combined with estimates of densities from the orbital periods of some asteroidal satellites, suggest that most C-type

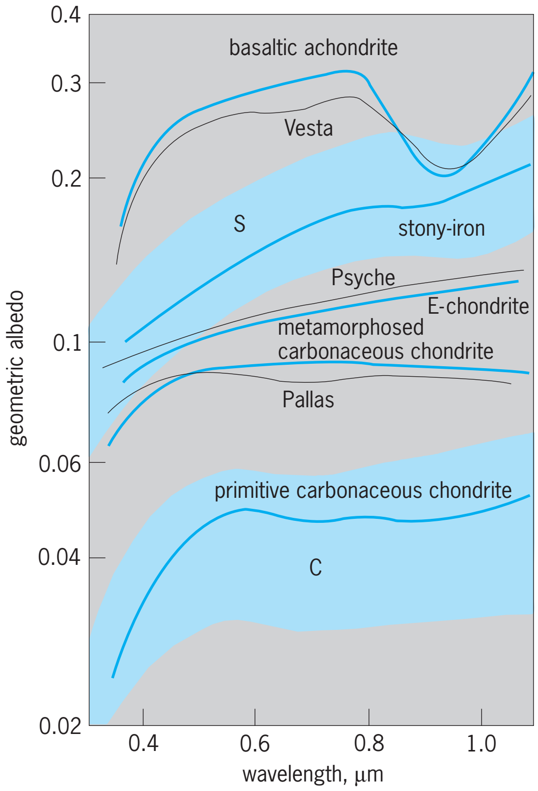


Fig. 3. Visible and near-infrared reflection spectra of the major S and C compositional classes of asteroids, plus several unusual ones. Laboratory spectra for some similar meteorites are shown with colored lines.

asteroids have densities less than 1.5, even more porous than carbonaceous chondrites. Whether this is due to inherently lightweight or porous materials or due to large voids in a rubble-pile structure, or a combination, is not yet known.

Surface compositions. Spectra of sunlight reflected from asteroids have shapes, including absorption bands, characteristic of different rock-forming minerals (**Fig. 3**). Combined with the albedo data from radiometry, the spectral colors of surfaces of thousands of asteroids show that more than three-quarters of them have very low albedos and are composed of carbon-rich material (often with hydrated, or water-rich, minerals). The black asteroids located in the middle and outer parts of the belt (called C type) resemble carbonaceous meteorites, which are believed to be among the most primitive materials in the solar system, little altered since the planets were forming. The black asteroids near the outer edge of the main belt, and most of the Hildas, have a reddish tinge and are not represented by known meteorites on the Earth; they are called P types, and may be even richer in organic components. Still farther out, many of the Trojans are even redder and more mysterious; they are termed D types.

Closer to the inner edge of the belt, most asteroids are so-called S types, characterized by moderately high albedos and by absorption bands due to the common silicate minerals pyroxene and olivine. They also contain considerable metal, and probably are akin to either the stony-iron meteorites or the ordinary chondritic meteorites. The spectral data alone favor stony-iron compositions, but such compositions would leave no other bodies remaining as sources for the ordinary chondrites, which constitute approximately 85% of all meteorite falls. It appears that the spectral properties of the surfaces of asteroids are altered by a process called space weathering. Micrometeorite impacts and other effects alter the colors of the surface regoliths (fragmental soils), changing the spectral signature of the meteoritic bed rock. Only asteroids too small to retain appreciable regoliths—less than about 1 km in diameter—look sensibly like ordinary chondrite meteorites, but it may be presumed that many of the larger S-types are also ordinary chondrites. *See* METEORITE; REGOLITH.

The general progression of asteroid compositions, from S types in the inner belt, to C's, to P's, and then to D's at Jupiter's distance, is thought to reflect the variation with distance from the Sun in the composition of the original nebular dust from which the planets were formed. *See* COSMOCHEMISTRY; SOLAR SYSTEM.

There are also some less common asteroid compositions, which are nevertheless interesting because they imply that some of the asteroids were somehow heated to the point that they melted. The asteroid Vesta is apparently mostly covered with lava flows, made of the same minerals as the basaltic achondritic meteorites. When a planetary body is heated to the point that volcanic lavas flow onto its surface, metal sinks to the center, forming a core. A number of asteroids, notably 16 Psyche, have so-called M-type spectra and radar echoes suggesting they are made of solid metal. Astronomers believe that Psyche's parent body was melted into a body that looked like Vesta does today, with a basaltic surface, an olivine-rich mantle, and a metallic core. Then collisions fragmented and stripped away the rocky crust and mantle, exposing the core.

There are questions about this picture, however. If there have been enough powerful collisions to strip some asteroids down to their metallic cores, it is strange that Vesta's crust has been so well preserved. In fact, Vesta is in the middle of a family of small, basaltic asteroids, apparently pieces of its crust blasted away by a cratering impact. Pictures of Vesta taken by the Hubble Space Telescope have revealed the huge crater created by this impact. But other asteroids should have suffered collisional fragmentation intermediate between Vesta and Psyche, and the fate of the pieces of olivine mantles of such asteroids remains a mystery. Indeed, only a few small, so-called A-type asteroids have been found that show the expected spectral signature of the supposedly abundant olivine.

Surface conditions and geology. Subsolar surface temperatures of main-belt asteroids mostly range from −80 to +10°F (210 to 260 K). The low gravity of asteroids prevents them from retaining any atmospheric gases, so their surfaces are exposed directly to interplanetary space where they are

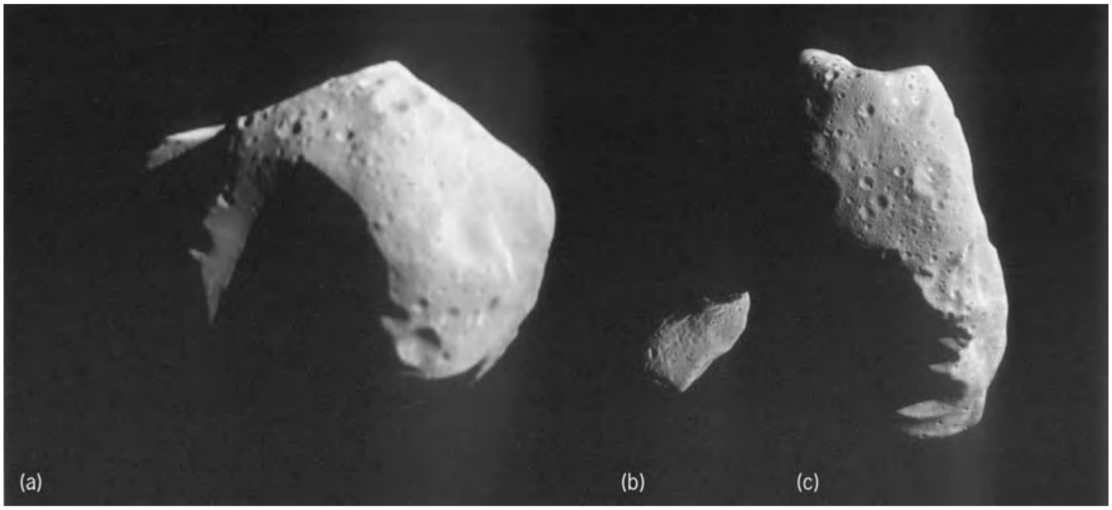


Fig. 4. Main-belt asteroids imaged by spacecraft. (*a*) Mathilde imaged by *Near Earth Asteroid Rendezvous* (*NEAR*) spacecraft on June 27, 1997. (*b*) Gaspra imaged by *Galileo* spacecraft on October 29, 1991. (*c*) Ida imaged by *Galileo* on August 28, 1993. All three objects are presented at the same scale, but relative brightnesses have been altered. Mathilde is actually much darker than the other two asteroids and has more large craters. (*NASA*)

impinged by the solar wind and micrometeorite bombardment. Polarimetric measurements suggests that asteroid surfaces are dusty, and the larger asteroids are expected to retain a fraction of the material ejected from cratering impacts, which accumulates into thick deposits of regolith.

Two S-type main-belt asteroids have been studied close-up by the *Galileo* spacecraft, which flew past them on the way toward its 1995 rendezvous with Jupiter. The encounter of 951 Gaspra on October 29, 1991, revealed an irregularly shaped body, 11.3 × 6.5 × 5.5 mi (18.2 × 10.5 × 8.9 km) in size (**Fig. 4**). Excellent data were obtained from the encounter on August 28, 1993, with the somewhat larger asteroid 243 Ida, a member of the populous Koronis family (Fig. 4). Ida's surface is covered by many more craters, especially moderate- and large-sized ones, than is Gaspra's, implying that Ida's creation as a fragment of the Koronis parent body may have happened billions of years ago. *Galileo*'s magnetometer observed magnetic anomalies as the spacecraft flew near both Gaspra and Ida, revealing unexpected interactions of the asteroids with the solar wind, possibly due to remanent magnetization of the asteroids.

A very different asteroid is the dark, 32-mi (52-km) main-belt object 253 Mathilde (Fig. 4). It is pockmarked by giant craters, five of which have diameters similar to or larger than the radius of Mathilde itself. They may result from crushing of Mathilde's porous material, rather than explosive excavation by impacting projectiles.

Earth-approaching asteroids. Apollo, Amor, and Aten asteroids are of special interest, particularly because they stand a chance of striking the Earth. Indeed, Meteor Crater (Arizona), and other craters on the Earth and the Moon, testify to the potential for collisions with near-Earth asteroids. Many scientists believe that just such a collision 6.5×10^7 years ago rendered most species of life, including the dinosaurs, extinct. A huge, eroded crater of that age on the Yucatán Peninsula in Mexico must have been caused by the impact of an asteroid or comet about 10 mi (16 km) in diameter.

Impacts by near-Earth asteroids are not confined to the past. In 1908, a small asteroid, perhaps 160 ft (50 m) across, exploded over the Tunguska region of Siberia with energy equivalent to 15 megatons of TNT. It has been estimated that the impact of a mile-wide asteroid could loft enough particles into the stratosphere to dim the Sun, disrupt agriculture, and cause worldwide famine. Such an impact has a 1-in-500,000 chance of occurring each year. Only about one-half of such potentially threatening objects have been discovered so far.

These small asteroids have not always been in Earth-approaching orbits. After perhaps 10 million years, most of the current crop will have struck the Earth, the Moon, the Sun, or one of the other inner planets, or will have been ejected from the solar system. Most are probably fragments of main-belt asteroids, traveling in chaotic orbits, just like their smaller cousins, the meteorites. (In fact, some meteorites are probably chips broken off Earth approachers.) A few near-Earth asteroids are dead comet nuclei, which have lost their ices after thousands of years spent close to the Sun in the inner solar system.

One of the largest near-Earth asteroids, 433 Eros, was the first to be visited by a spacecraft. *NEAR-Shoemaker* orbited around Eros for a year, before landing on its surface, providing the most detailed information ever obtained about the chemistry, mineralogy, and geology of an asteroid. Eros has the composition of ordinary chondrites. Seen up close (**Fig. 5**), the asteroid's surface lacks small craters but is covered with boulders. In places, there are mysterious, flat, dry "ponds".

Origin and evolution. Current cosmogonical models for the origin of planets involve accretion from

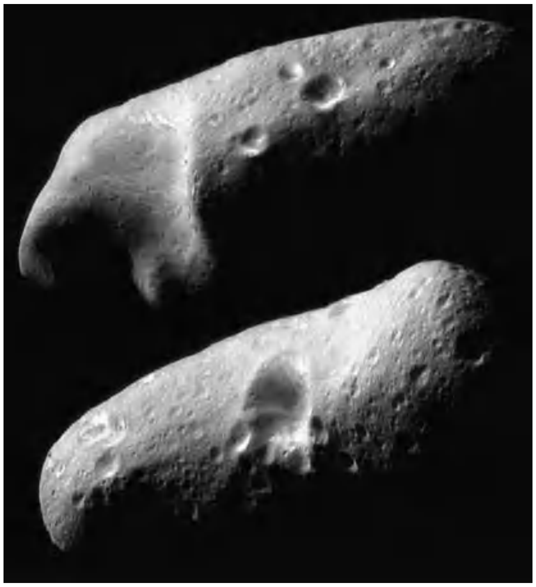


Fig. 5. Views of the two opposite sides of the asteroid 433 Eros, imaged from the *NEAR-Shoemaker* spacecraft. (*NASA*)

myriads of asteroidlike planetesimals. It is likely that asteroids are a remnant of the planetesimals that failed to accrete into a planet between Mars and Jupiter. Perhaps shifting positions of the young planets, including massive, nearby Jupiter, increased the relative velocities of asteroids to the present value of 3 mi/s (5 km/s) so that asteroids fragment rather than accrete when they meet each other. Instead of forming a planet, the asteroids have been smashing each other to bits.

Evidently some asteroids of primitive, nonvolatile solar composition were heated within the first few hundred million years after the origin of the solar system, perhaps by the solar wind or extinct radionuclides, and they melted. While the unmelted, weak, C-type asteroids may have been depleted by a large factor by collisions, most of the strong stony-iron cores of the melted proto-asteroids have survived; perhaps they are among the M- and S-type asteroids observed today. The asteroids still collide and fragment. The fragments drift due to solar forces and enter resonances, which ultimately send some of them into the inner solar system, where they produce craters or fall as meteorites.

This picture of the evolution of the asteroids is incomplete and tentative. As new data are collected, research continues on alternative interpretations.

Clark R. Chapman

Bibliography. J. K. Beatty, C. C. Peters, and A. Chaikin (eds.), *The New Solar System*, 4th ed., Sky Publishing Corp. and Cambridge University Press, 1999; W. F. Bottke, Jr., et al. (eds.), *Asteroids III*, University of Arizona Press, 2002; H. Y. McSween, Jr., *Meteorites and Their Parent Planets*, 2d ed., Cambridge University Press, 1999; J. L. Remo (ed.), Near-Earth Objects: The United Nations Conference, *Ann. N.Y. Acad. Sci.*, vol. 822, 1997.

Asteroidea

A subclass of the class Stelleroidea in the subphylum Asterozoa; asteroids are known as starfish (**Fig. 1**). The arms are not sharply demarcated from the rest of the body. The ambulacral ossicles never fuse to form vertebrae. The tube feet are locomotor organs, usually suctorial, emerging from an open ambulacral groove. The dominant growth gradients are such as to cause the skeletal ossicles to lie in longitudinal rows known as series (for example, adambulacral series, ventrolateral series, and inferomarginal series).

For many years the Asteroidea and the Ophiuroidea were accorded separate class status, but in 1951 W. K. Spencer showed that they have a common origin. Most differences between these groups disappear when fossil forms are considered. Thus both groups are now ranked as subclasses of Asterozoa. The 1700 or so known fossil and recent species of Asteroidea are now grouped into six orders: Platyasterida, Paxillosida, Valvatida, Forcipulatida, Notomyotida, and Trichasteropsida. Asteroids range in size from about 0.4 to 40 in. (10 mm to 1 m) across. Many starfish are brightly colored and attractive animals, but some are dowdy and cryptic. Their conjugated carotenoid pigments fade on preservation. Like most echinoderms, starfish have a lifespan of about 5 years. *See* OPHIUROIDEA.

Economic importance. Several species of starfish are regarded as pests, particularly in oyster, mussel, and scallop fisheries. Beyond this they are of little significance, with the notable exception of the coral predator *Acanthaster planci*, which occurs in the Indo-Pacific. When present in large numbers, it may kill extensive areas of reef-building corals. However, large aggregations of this species are occasional phenomena, and coral killed by it is capable of regenerating more quickly than was hitherto supposed.

Ecology. Starfish inhabit all types of bottom throughout the world's seas and oceans. Some burrow in sand and mud, and others live on rocks and

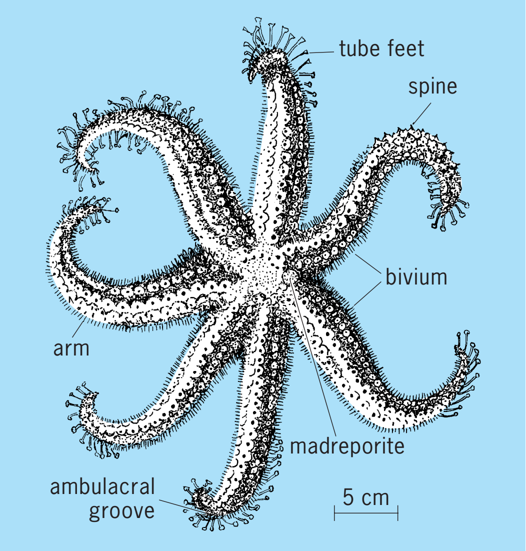


Fig. 1. A representative asteroid, *Astrostole scabra*.

coral reefs, where they are at their most diverse. *Calvasterias* may live at the surface, attached to floating weeds. Most families show well-defined bathymetrical preferences, but there are exceptions, for instance, the offshore fauna of New Zealand. There are 12 families which occur at depths greater than 2 mi (3 km), and several genera extend below 4 mi (6.5 km).

Parasitism and commensalism. Asteroids are parasitized by a variety of organisms. Small sea snails of the genus *Stylifer* bore into their tissues. A polychaete worm *Acholoë astericola* lives in the ambulacral groove of *Astropecten*; a cirripede castrates *Cascinasterias* by destroying the gonad; and a slender fish, *Fierasfer*, inhabits the body cavity of *Culcita*. This appears to enter the mouth and then bores through the stomach wall. A remarkable insect-asteroid association has been reported from Australia, where a marine caddis fly lays its eggs in the tissues of *Patiriella*.

Anatomy. The outer surface is coated with a thin layer of ciliated epithelium. Below this lies connective and muscle tissue which forms the body wall and in which the skeletal ossicles are embedded, forming the test. In some species the skeletal component is very considerable, making them heavily armored and relatively rigid; in others it is less, so they appear lighter and more flexible. Some of the ossicles, notably the spines, protrude to the exterior (**Fig. 2**). The arms can be slowly moved by the body wall musculature. Although 5 arms are common, up to 12 occur in a number of genera, and more than 12 may occur, as in *Acanthaster*.

Pedicellariae occur in some starfish, but these have evolved independently from echinoid pedicellariae. They are small seizing organs used to grip intruders and, according to one report, even to catch food. They probably assist in protecting the delicate epithelium from predators and sediments. Pedicellariae are formed from two or four modified spines which are snapped together by adductor muscles like minute tongs. Their shape and type is important in classification because different orders of starfish have evolved different forms of pedicellariae. In many species they are sessile, that is, attached directly to the test. They may be simple and spinelike, recessed like minute clamshells, or felipedal like miniature cat's paws. In the Forcipulatida they are stalked and may be crossed or straight. Their activities are controlled by their own nerves and the basiepithelial plexus of the test (**Fig. 3**).

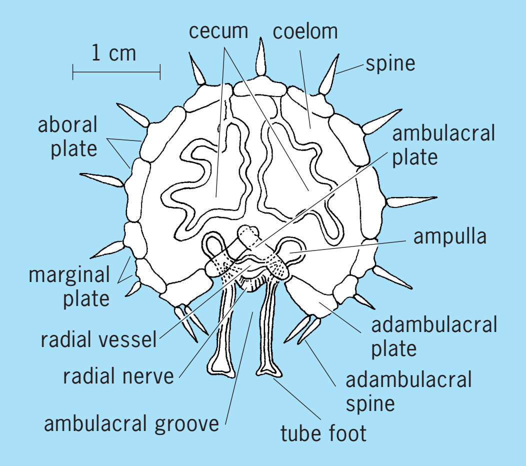


Fig. 2. Transverse section through arm of asteroid.

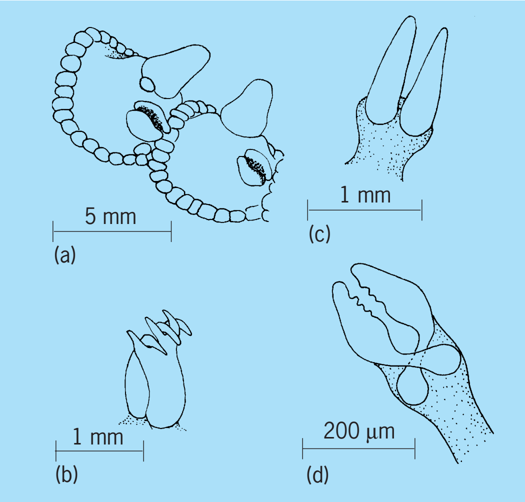


Fig. 3. Pedicellariae of asteroids. (*a*) Valvate type on marginal plates of *Hippasteria trojana*. (*b*) Felipedal type in *Cosmasterias dyscrita*. (c) Straight type. (*d*) Crossed type, as in Forcipulatida.

The well-developed water-vascular system follows the pattern for the phylum, with the following features. With the exception of some Paleozoic genera, the madreporite is always on the upper (aboral) side. There is usually only one madreporite situated interradially, but *Allostichaster* has 2 to 4 and *Acanthaster* up to 16 or more. The tube feet may be peglike and suckerless or columnar and suckered. The former occur in burrowing species and in species which inhabit soft deposits, where they are presumably used like oars or stilts. The latter are used for climbing over obstacles and are better at adhering to hard substrates like rocks. Suckered tube feet can be used to pull open the bivalve prey.

The nervous system also follows the pattern for the phylum. The radial nerves retain a superficial position just below the epithelium lining the ambulacral groove. The basiepithelial nerve plexus envelops most of the body and coordinates the activities of peripheral organs such as pedicellariae, papulae, and paxillae. Starfish cannot see, but they clearly detect changes in light intensity such as those caused by shadows. They have photosensitive eyespots located at the tip of each arm. The terminal tube feet of each ambulacrum appear to be especially sensitive to water-borne chemicals which may be emitted by prey species, and they are thus used in food detection.

Feeding habits. Starfish are voracious feeders. The prey is detected by water-borne odors or by direct contact as the result of random movement. Small food, such as amphipods and young mollusks, may be engulfed whole. *Acanthaster* everts part of its

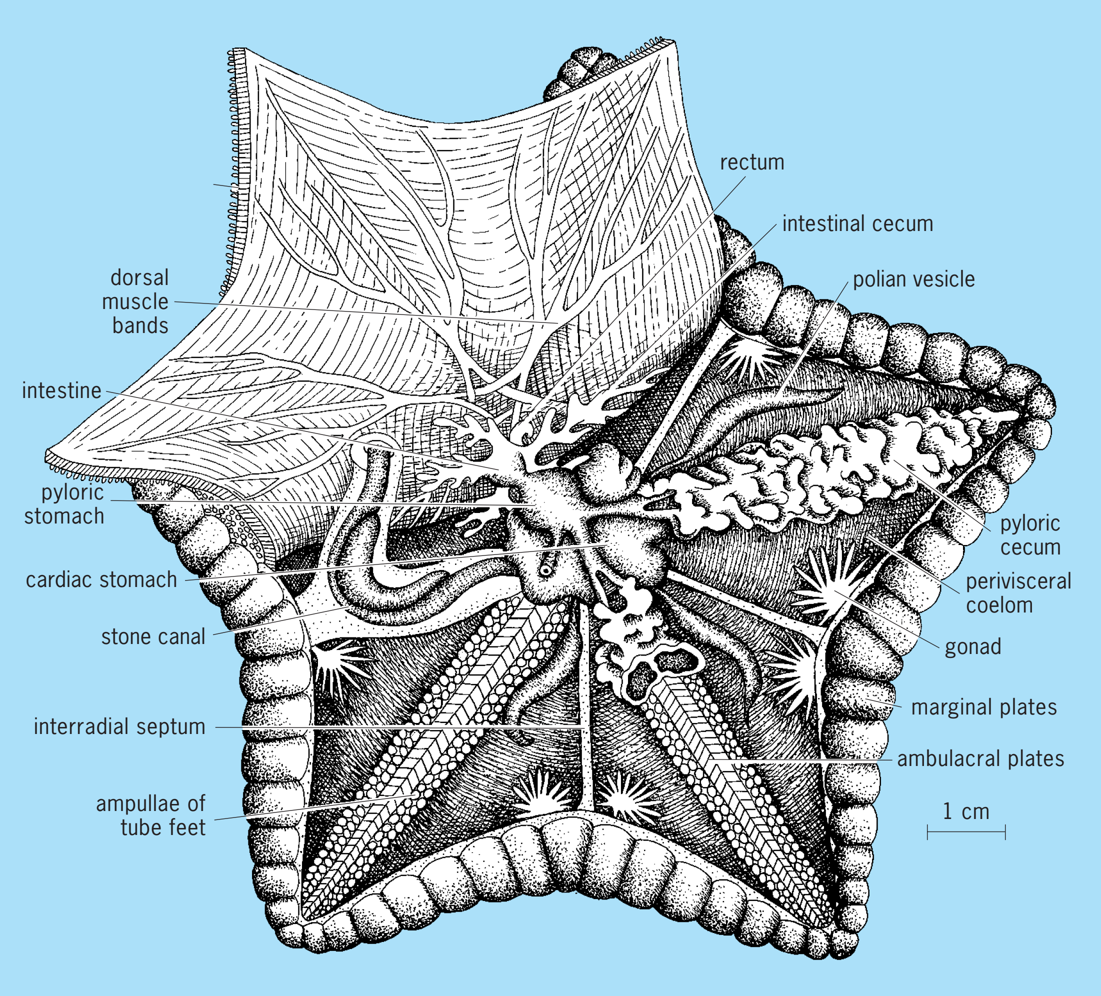


Fig. 4. Anatomy of a starfish, *Asterodon robustus*, as viewed from aboral surface.

stomach out of the mouth and wraps it over coral polyps, which are thus digested outside the predator's body. Some of the Forcipulatida, for example, *Asterias*, can open large clam and other bivalve shells by applying a continuous force which overcomes the molluscan adductor muscles. The starfish uses both tube-feet muscles and body wall muscles. The mechanical stresses applied by a starfish are sufficient to create the shell gap necessary for the insinuation of the folds of the everted starfish stomach. Toxic secretions are not used to paralyze the prey.

The mouth is in the middle of the lower (oral) surface of the disk and leads through a short esophagus to the cardiac stomach (**Fig. 4**). The number of lobes of this organ correspond to the arm number of the starfish. It is these lobes that can be everted through the mouth to carry out extracorporeal digestion. Above the cardiac stomach lies the pyloric stomach, again with as many branches as there are arms. Within each arm, the branch forks to give two blind digestive ceca. In most species, a short intestine leads upward to the rectum and anus, generally situated near the middle of the upper (aboral) surface. In starfish which lack an anus, the feces are extruded through the mouth.

Respiration and water balance are aided by small contractile outgrowths of the test epithelium called papulae. Coelomic fluid circulates on their inside and seawater on their outside, so that exchange can occur across their then membranous surface.

Reproduction. Asteroids are generally mature and able to reproduce at 1 year. They normally continue to grow for about 4 years, growth taking place when food is abundant. The sexes are usually separate, but in a few genera, such as *Fromia*, hermaphrodites occur, the sex changing with age. The sex of adults is not discernible externally. The gonads lie in the interradii, sometimes in pairs, and sometimes extending into the arms. Each has a separate duct to the exterior, and sperm and eggs are shed into the sea, where external fertilization occurs. If a larva develops, it is a bipinnaria or derived type. Many species lack a larval stage. Some brood the eggs, and hatching may take place in the stomach, the mouth, or a special marsupium. Some species reproduce by transverse asexual fission. Regeneration of lost or damaged parts is a characteristic of most genera. In *Linckia* this is most marked, and a whole new individual can be regenerated from a fragment of an arm. *See* ECHINODERMATA.

Andrew C. Campbell

Bibliography. *Asteroidea of the North Pacific and Adjacent Waters*, U.S. Nat. Mus. Bull. 76, pts. 1–3, 1911–1930; M.E. Downey, Starfishes from the Caribbean and the Gulf of Mexico, *Smithson. Contrib. Zool.*, vol. 126, 1973; H. B. Fell, Phylogeny of sea stars, *Phil. Trans. Roy. Soc. London*, ser. B, 246:381–485, 1963; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; C. Schuchert, *Revision of Paleozoic Stelleroidea*, U.S. Nat. Mus. Bull. 88, 1915.

Asteroxylales

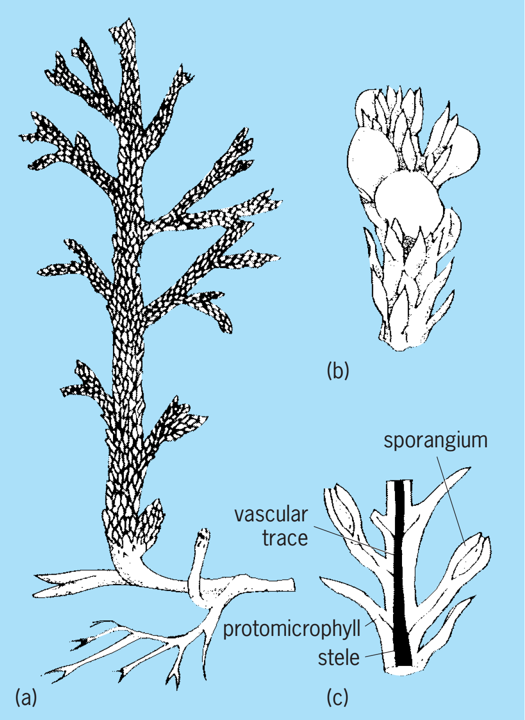
A small extinct order of the class Lycopsida that bridges the evolutionary gap between the primitive Zosterophyllopsida and relatively advanced Lycopsida; hence the asteroxylaleans are often termed prelycopsids. The best-known asteroxylalean species are of Early Devonian age, although similar forms survived to the Late Devonian. All possessed at least some hydrophytic features; together with a paucity of strengthening tissues, this characteristic suggests that they largely relied on hydrostatic pressure for structural support. *See* ZOSTEROPHYLLOPSIDA.

With the exception of *Asteroxylon*, asteroxylalean fossils are not well preserved. The asteroxylaleans are regarded as the most primitive lycopsids, but they can also be considered as the most advanced zosterophyllopsids; some authorities prefer to disperse the prelycopsids among other taxa.

The pivotal plant in this order is *A. mackiei*, which originated from the remarkable biotic communities petrified in Lower Devonian volcanigenic cherts at Rhynie, Scotland (see **illus.**). This species had naked horizontal rhizomes that produced primitive roots and aerial branches up to 20 in. (50 cm) high and 0.4 (1 cm) in diameter. The radially symmetrical actinostele (a star-shaped protostele) is typical of primitive lycopsids rather than zosterophyllopsids. Vascular traces extended toward, but did not enter, the densely packed clasping tissue outgrowths termed protomicrophylls.

The sporangia were homosporous, kidney shaped, and distributed singly and randomly among the leaves on the more distal axes. Like the zosterophyllopsids, *Asteroxylon* bore sporangia on stalks with a vascular supply that was independent of, and more extensive than, that of the surrounding enations. These reproductive structures are generally regarded as homologous with vegetative lateral branches.

Other relevant but less well understood Devonian plants are *Drepanophycus* and *Kaulangiophyton*. They were geographically widespread in the Lower and Middle Devonian, when they probably formed extensive clonal mats of low diversity in water-rich habitats. Although typically larger than *Asteroxylon*, they had the same basic growth architecture. However, the unvascularized enations were sparser and smaller, and the scattered large sporangia were egg shaped. The anatomy of these plants is not well known, and the position of the sporangia is ambiguous. For both genera, it is uncertain whether the sporangia occupied separate stalks or were positioned on the upper surfaces of sporophylls, like true lycopsids. *See* LYCOPHYTA; LYCOPSIDA.



The most completely known asteroxylalean, *Asteroxylon mackiei*, from the Lower Devonian Rhynie Chert of Scotland. (*a*) Reconstruction of vegetative parts of the plant. (*b*) Exterior of the fertile region (*after W. G. Chaloner and P. MacDonald, Plants Invade the Land, HMSO, 1980*). (*c*) Longitudinal section of the fertile region, contrasting the partial vascularization of the protomicrophylls with the complete vascularization of the larger sporangia (*after A. G. Lyon, The probable fertile region of Asteroxylon mackiei, Nature, 203:1082–1083, 1964*).

Richard M. Bateman; William A. DiMichele

Bibliography. P. G. Gensel and H. N. Andrews, *Plant Life in the Devonian*, 1984; W. H. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, 2d ed., 1993.

Asthenosphere

A layer of variable thickness within the upper mantle of the Earth that is mechanically weaker (less rigid) than both the overlying lithosphere and the underlying mesosphere (**Fig. 1**). The depth of transition from lithosphere to asthenosphere can be as little as 20 km (12 mi) in oceanic settings and as great as 250 km (155 mi) in continental settings. In both cases, the asthenosphere may extend in depth to over 300 km (186 mi). The presence of a mechanically weak layer in the upper mantle can be attributed to the temperature and pressure conditions being such that the ratio between the temperature of mantle material and its melting point (solidus) is a maximum. *See* EARTH INTERIOR; LITHOSPHERE.

The existence of the asthenosphere has primarily been inferred from geophysical measurements,

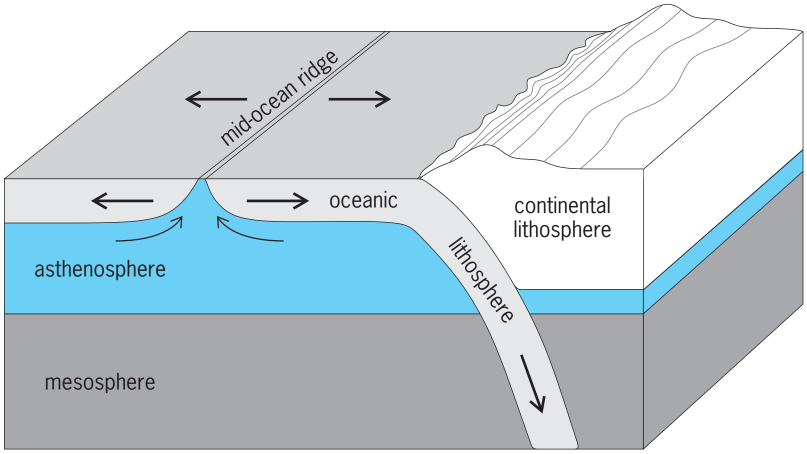


Fig. 1. Location of the asthenosphere in the upper mantle (not to scale). The thickness of the asthenosphere varies laterally due to the dynamics of plate tectonics.

including seismic, magnetotelluric, glacial rebound, and lithospheric stress distribution. The seismic low-velocity zone (LVZ) is a well-known feature of the upper mantle that is often directly correlated with the low-rigidity asthenosphere. Electromagnetic observations indicate the presence of a high-conductivity zone (HCZ) beneath the lithosphere that is consistent with partially molten upper-mantle material. The melting of large ice sheets at the end of the last ice age has significantly changed the surface load distribution throughout much of the Earth; consequent rebound rates can be related to the viscosity of the underlying mantle. Glacial rebound models indicate the presence of a low-viscosity layer beneath the lithosphere that is associated with the asthenosphere. In addition, the forces and motions associated with plate tectonics require the presence of convecting mantle material beneath the lithosphere. The predominant view is that the asthenosphere is unlikely to completely decouple the lithosphere from the mesosphere with its own flow regime; rather, the lithosphere, asthenosphere, and mesosphere are all part of a single convecting system. *See* EARTH, CONVECTION IN; ISOSTASY; PLATE TECTONICS.

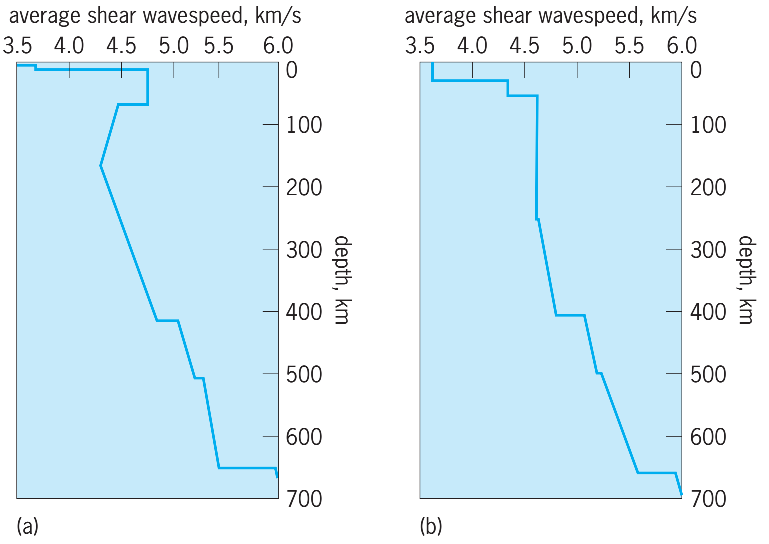


Fig. 2. Variation of seismic shear wave speed with depth in (*a*) suboceanic and (*b*) subcontinental settings. The LVZ is present beneath oceanic lithosphere but absent beneath stable continents.

Although the asthenosphere has a strict definition as a mechanically weak layer constituting the top part of the sublithospheric mantle, its meaning has broadened in recent times to such an extent that many geophysicists avoid using the term altogether. For example, many now appear to equate the asthenosphere with the LVZ, HCZ, or even the entire upper mantle. In addition, geochemists have given chemical connotations to the asthenosphere, and often describe it as a homogeneous, convecting, and depleted layer beneath the lithosphere. However, from a physical point of view, the asthenosphere could be enriched or depleted, homogeneous or inhomogeneous, and yet still adhere to its original definition.

Mechanical properties. The asthenosphere is thought to be largely composed of the mineral olivine ($\sim$60%), with lesser proportions of garnet, spinel, and pyroxenes. In the presence of stresses applied at geological time scales, mantle rocks in the asthenosphere will flow as a result of diffusion creep, a mechanism which involves the thermally activated migration of crystal defects due to stress. The ability of a material to flow can be quantified by its viscosity, which is a measure of a fluid's resistance to flow. In the lithosphere, viscosities range from 10^{21} pascal seconds at its base to infinite at shallower depths as the rheology (that is, the deformation and flow characteristics of the material) enters a brittle-elastic regime. In the asthenosphere, viscosities can be as low as 10^{18}–10^{19} Pa·s, but will increase to 10^{21} Pa · s or more in the underlying mesosphere. It is probable that some parts of the asthenosphere are molten and have extremely low viscosities, but the volume and lateral extent of such material is poorly understood. The viscosity of mantle rock and its mechanical strength depends strongly on the ratio between its temperature and melting temperature (solidus). Above the asthenosphere, this ratio decreases rapidly, resulting in a change from a ductile regime to a brittle-elastic regime. Below the asthenosphere, the mantle solidus increases with depth at a greater rate than the mantle temperature, which increases the mechanical strength of the material. The temperature in the asthenosphere is generally thought to be between 1300 and 1500°C (2372 and 2732°F), but caution is required in attributing a temperature range to the asthenosphere. For example, the mechanical strength of upper-mantle rocks can be greatly reduced by hydration, with no change in temperature. *See* GEODYNAMICS; RHEOLOGY; VISCOSITY.

Seismic properties. The measurement and interpretation of seismic waves within the Earth provides some of the strongest evidence for the existence of the asthenosphere. The speed of seismic shear (S) waves can be significantly reduced in the

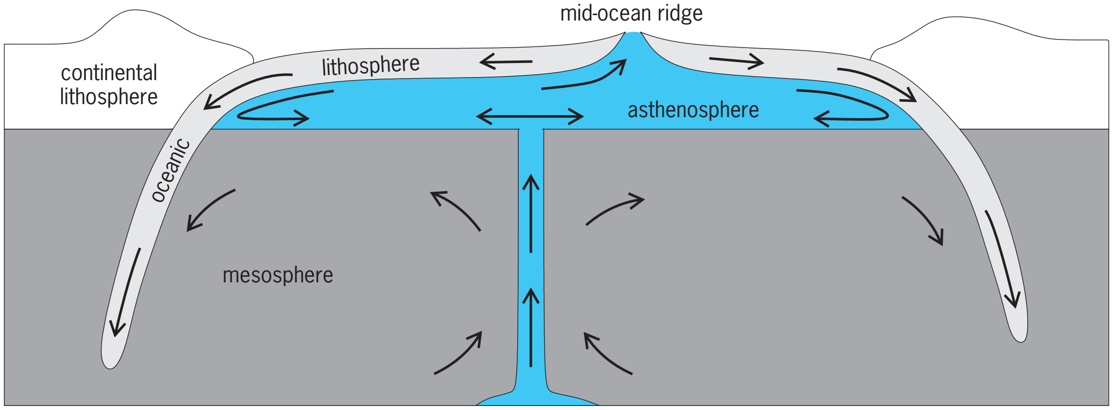


Fig. 3. Proposed mantle convection paradigm (not to scale). In this scenario, the asthenosphere accommodates large lateral flow from hotspots toward ridges and possibly a partial return flow of asthenospheric material from subduction zones. Plate motions and deep mantle flow couple directly at subduction zones and beneath the roots of continental shields.

asthenosphere due to the presence of low-viscosity material. Regional and global studies of the Earth consistently reveal the presence of the LVZ as a pronounced drop in shear-wave speed relative to the lithosphere and mesosphere. However, it appears that the LVZ is largely confined to the suboceanic lithosphere, with little evidence of slower wave speeds beneath inactive continents (**Fig. 2**). Seismic waves (especially shear waves) are also strongly attenuated in regions of relatively low viscosity. Seismic models of mantle attenuation indicate high attenuation beneath continents and low attenuation beneath oceans, a result that correlates well with studies of seismic-wave speeds. *See* SEISMOLOGY.

Another indicator of a weak zone beneath the lithosphere comes from seismic-wave observations showing pronounced changes in the fast orientation of seismic wavespeed (anisotropy) with depth in the upper mantle, which may be related to a change in the preferred alignment of olivine crystals due to variations in the stress conditions (magnitude and direction) over time. In the lithosphere, anisotropy is caused by past deformation that has been “frozen” in, while anisotropy in the asthenosphere is related to contemporary plate motion.

The relationship between the various seismic properties discussed above and mechanical strength as measured by viscosity are not unique. For example, although most of the LVZ is probably due to a change in the physical property of the mantle (that is, reduced viscosity), changes in composition can also affect seismic-wave speeds. Therefore, terms such as LVZ or “high attenuation zone” should not be used synonymously with asthenosphere.

Electromagnetic properties. Electromagnetic studies of the asthenosphere show it to be partially molten beneath mid-ocean ridges with high conductivities (0.1 S/m). Beneath stable continental shields, the conductivity is much less (0.02 S/m), which suggests the absence of an electrical asthenosphere. In regions of active continental tectonics, the depth to the top of the high-conductivity zone (HCZ) can range from 70 to 200 km (43 to 124 mi) in a little over 100 km (62 mi) laterally. Recent studies suggest that the HCZ may also be electrically anisotropic due to the contribution of hydrogen diffusivity to the conductivity of olivine. As in the seismic case, one should not directly associate the HCZ with a mechanically weak asthenosphere, since the HCZ is characterized by the presence of melt and/or volatiles and may be distinct from the LVZ and the rheological asthenosphere. *See* MID-OCEANIC RIDGE; ROCK, ELECTRICAL PROPERTIES OF.

The usefulness of electromagnetic observations in the upper mantle for inferring physical properties has improved considerably in the past decade or so, with major developments in laboratory studies of mantle mineral conductivities. Future work in this area is likely to greatly improve our understanding of the asthenosphere.

Paradigms. Several early models of plate tectonics decouple the motion of the lithosphere from convection in the mesosphere and instead suggest that flow in the mechanically weak asthenosphere could entirely account for the motion of tectonic plates. Subsequent models, which are still popular today, require the lithosphere, asthenosphere, and mesosphere to be part of a unified convective system. An alternative paradigm has been proposed in which the asthenosphere is a layer that exists predominantly beneath the oceanic lithosphere and is fed entirely by upwelling plumes from the lower mantle. The mass balance of the system is preserved via accretion to the overlying lithosphere and subduction (**Fig. 3**). The asthenosphere flow model is one of several paradigms that seek to explain the precise role of the upper mantle in plate tectonics, and helps illustrate the difficulty in trying to define the structure and dynamics of the Earth’s interior in the absence of direct measurement.

Nick Rawlinson

Bibliography. D. L. Anderson, Lithosphere, asthenosphere and perisphere, *Rev. Geophys.*, 33:125–149, 1995; J. B. Gaherty, M. Kato, and T. H.

Jordan, Seismological structure of the upper mantle: A regional comparison of seismic layering, *Phys. Earth Planet. Int.*, 110:21–41, 1999; G. Heinson, Electromagnetic studies of the lithosphere and asthenosphere, *Surveys Geophys.*, 20:229–255, 1999; I. Jackson (ed.), *The Earth's Mantle*, Cambridge University Press, 1998; J. P. Morgan, W. J. Morgan, and Y.-S. Zhang, Observational hints for a plume-fed, suboceanic asthenosphere and its role in mantle convection, *J. Geophys. Res.*, 100:12,753–12,767, 1995; J. Plomerová, D. Kouba, and V. Babuška, Mapping the lithosphere-asthenosphere boundary through changes in surface-wave anisotropy, *Tectonophysics*, 358:175–185, 2002.

Asthma

An allergic inflammatory disease of the airways, involving mast cells, eosinophils, macrophages, lymphocytes, and neutrophils. Such inflammatory changes are associated with widespread airflow obstruction, which is variable and improves (reverses) spontaneously or with appropriate therapy. Inflammation progresses to increased airway irritability (hyperresponsiveness) induced by the inhalation of allergens, cold air, and occupational factors. Although bronchospasm can be induced immediately after exposure to a specific allergen in an appropriately sensitized recipient, it is the late allergic response that most resembles the inflammatory reaction occurring in asthma. Central to this reaction is the release from mast cells, eosinophils, and lymphocytes of chemical mediators such as histamine, leukotrienes (potent bronchoconstricting agents), and various cytokines which perpetuate the response. Potent neurohumoral agents derived from neural pathways contribute further to the bronchospasm. *See* CYTOKINE; HYPERSENSITIVITY.

Diagnosis. Wheezing, nocturnal breathlessness, coughing, and chest tightness often relieved by expectoration are highly suggestive of asthma. Episodes of breathlessness which result from exposure to an irritant (such as cold air) or an allergen (such as dust mites) following exercise or a viral infection and which are reversed spontaneously or with therapy are diagnostic of asthma. A physical examination may not reveal anything. However, eczema and edema in the folds of the nasal chambers are suggestive of a hereditary allergy, the major predictor of asthma. Nasal polyps may indicate aspirin intolerance in an asthma sufferer. Objective measures of airflow obstruction which improved spontaneously or with therapy are also central to establishing an asthma diagnosis.

Atopy, the genetic predisposition for developing an immunoglobulin-E (IgE) mediated (allergic) response to inhaled environmental allergens, is the strongest predisposing factor for developing asthma. The demonstration of an allergen-specific IgE by the presence of specific IgE antibodies using skin tests or serum analysis allows for easier identification. Utilizing the suspected sensitizing allergen, visual examination by fiber-optic bronchoscopy, or even sputum analysis may provide greater insight into the pathogenesis of asthma. *See* IMMUNOGLOBULIN.

Asthma may be classified, therefore, according to severity, etiology, or pattern of airflow obstruction. It is helpful to differentiate those factors that induce inflammation from those that incite acute bronchospasm in susceptible individuals. The association of an elevated serum IgE and the occurrence of asthma in all age groups, including those who are not atopic, makes antigenic stimulation causal in all instances of asthma. The severity of asthma can best be defined in terms of peak-flow monitoring (monitoring the severity of the allergy). Such evaluations (see **table**) as mild, moderate, and severe are useful in applying therapy in a stepwise manner contingent on severity.

Peak-flow meter assessment provides severity classification and permits early detection of deteriorating condition, often prior to the onset of an asthma attack. Factors associated with a higher risk of developing a severe asthmatic condition and possible fatality include a history of acute life-threatening attacks, hospitalization for asthma within the previous year, a history of respiratory failure, recent reductions in corticosteroid (anti-inflammatory) therapy, and failure to follow a recommended therapy. Asthma sufferers with low incomes who receive inadequate medical care and who are inner-city residents are at especially high risk for fatal asthma.

Therapy. Successful management of asthma requires education of the sick individual coupled with the development of a partnership with an asthma management health-care team; assessing and monitoring the severity of asthma, with utilization of objective parameters of assessment (for example, the peak-flow meter, a device that measures the amount of air that enters and leaves the lungs); environmental management to avoid asthma triggers; and establishment of a drug regimen that controls asthma, as well as a written plan to prevent the condition from becoming worse. Adequate management of asthma should control the symptoms, prevent asthma attacks, return and maintain pulmonary function as close to normal as possible, maintain normal activity levels including exercise, avoid adverse side effects from the drugs, reduce and prevent irreversible airway changes, and prevent mortality.

The treatment of asthma requires that the individual and the family participate with the health-care provider in care. Individuals with moderate-to-severe or difficult-to-manage mild asthma should be trained to measure their morning pulmonary status. Monitoring evaluates effectiveness of the treatment, provides warning of early deterioration in asthma status, and provides critical assessment regarding treatment changes. Measuring the amount of air that enters and leaves the lungs (peak flow rate) for all initial and follow-up examinations provides the most objective assessment, because peak-flow determination is effort-dependent and is not suitable in all instances of asthma. The peak flow rate permits assessment of severity, airway hyperresponsiveness,

Classification of asthma severity

Asthma severity	Clinical features before treatment	Lung function*	Regular medication usually required to maintain control
Mild	Intermittent, brief symptoms less than 1 or 2 times per week Nocturnal asthma symptoms less than twice a month Asymptomatic between exacerbations	PEF >80% predicted at baseline PEF variability <20%	Intermittent inhaled short-acting beta-agonist (taken as needed) only
Moderate	Exacerbations more than 1 or 2 times per week Nocturnal asthma symptoms more than twice a month Symptoms requiring inhaled beta-agonist almost daily	PEF 60–80% predicted at baseline PEF variability 20–30% PEF normal after bronchodilator	Daily inhaled anti-inflammatory agent Possibly a daily long-acting acting bronchodilator, especially for nocturnal symptoms
Severe	Frequent exacerbations Continuous symptoms Frequent nocturnal asthma symptoms Physical activities limited by asthma Hospitalization for asthma in previous year Previous life-threatening exacerbation	PEF <60% predicted at baseline PEF variability 20–30% PEF normal after bronchodilator	Daily inhaled anti-inflammatory agent at high doses Daily long-acting bronchodilator, especially for nocturnal symptoms Frequent use of systemic corticosteroids

*PEF = peak expiratory flow.

response to therapy, symptomatic deterioration, and identification of asthma triggers. Peak flow rate improvement, following bronchodilator therapy, is suggestive of asthma. However, in occasional cases of asthma where there is severe obstruction of the airway, lack of a response to bronchodilators may occur until effective anti-inflammatory therapy has restored the response.

Environmental management with avoidance of indoor and outdoor asthma triggers is central to the successful management of asthma. Tight buildings with inadequate natural ventilation often lead to indoor pollutants as well as dust mites. Such environmental allergens are a major factor in the increase of airway hyperresponsiveness. A reduction in allergens reduces airway hyperresponsiveness and leads to asthma improvement. However, it may take 6 months for complete removal of allergens to be effective. Tobacco smoke and wood smoke always accentuate asthma and must be avoided. Also, drugs such as beta-blocking agents, angiotensin-converting enzyme inhibitors, and aspirin and nonsteroidal anti-inflammatory agents should be avoided in susceptible individuals.

Asthma management requires a step-by-step introduction of medications, dependent upon the standard peak flow rate zone system applicable to all cases of asthma. Medications include bronchodilators (which act as relievers) and anti-inflammatory agents (which act as preventers). Anti-inflammatory agents reduce airway inflammation, whereas bronchodilators relax smooth muscles, thus leading to airway dilatation. Higher drug concentrations can be delivered by aerosol with an inhaler or by dry powder, and there will be little systemic effect. The use of a metered dose inhaler should be repeatedly reviewed. Spacers (a direct drug delivery system) improve delivery of the drug to the airways and reduce deposition of the drug both on the mucous membranes in the mouth and in systemic adsorption.

Bronchodilators are airway smooth-muscle relaxants that enhance microciliary clearance, reduce vascular permeability, and appear to modulate most mediator release. Short-acting beta-agonists are the most predictable and potent bronchodilating agents available and are the drug of choice in reversing acute exacerbation. If beta-agonist therapy is required more than three times per week, therapy with anti-inflammatory agents should begin. If there is no prompt, sustained effect from using beta-agonists, immediate medical attention is required as well as administration of systemic steroids.

Theophylline in sustained-release form has a prolonged bronchodilator effect which is useful in reducing nocturnal asthma. Occasional toxicity requires periodic blood monitoring.

Anti-inflammatory agents are fundamental to the pharmacotherapy of asthma. Corticosteroids are the most potent and effective agents, but at higher concentrations bone metabolism and adrenal function may be inhibited. Such agents inhibit the cytokine effect as well as the synthesis of leukotriene and prostaglandin. Nonsteroid inhaled anti-inflammatory agents inhibit both the immediate and late allergic reactions.

A stepwise approach to pharmacologic therapy is based on the severity of the asthma condition and the current therapy. The number and frequency of drug therapies is increased with increasing severity. Once control is sustained for several weeks or months, a reduction, or step-down in drug therapies can be carefully developed to identify the minimum required for controlling asthma. *See* ALLERGY; RESPIRATORY SYSTEM DISORDERS.

Albert L. Sheffer

Bibliography. J. Brownell and T. B. Casale, Anti-IgE therapy, *Immunol. Allergy Clin. N. Amer.*, 24(4):551–568, 2004; National Asthma Education and Prevention Program Expert Panel Report, NIH Pub. no. 97–4051, 2002; National Asthma Education and Prevention Program Expert Panel Report: Managing asthma during pregnancy: Recommendations for pharmacologic treatment—2004 update, *J. Allergy Clin. Immunol.*, 115(1):34–46, January 2005 [Erratum, *J. Allergy Clin. Immunol.*, 115(3):477, March 2005]; National Heart, Lung, and Blood Institute, *International Consensus Report on Diagnosis and Treatment of Asthma*, 1992; U.S. Department of Health and Human Services, National Heart, Lung, and Blood Institute, *Guidelines for the Diagnosis and Management of Asthma*, 1991.

Astomatida

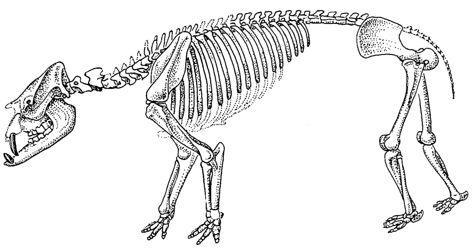
An order of protozoans, subclass Holotrichia, in which all species are mouthless. All species are parasitic in other animals, typically oligochaete annelids. Many astomatids possess an elaborate holdfast organelle, of value in attachment to the cells or tissues of the host's alimentary tract. Some authorities have considered the species of this rather large and ill-defined group to represent nothing more than degenerate, astomatous members of certain other groups, such as thigmotrichs, apostomes, and possible hymenostomes. Still, there seems to be value in recognizing the evolutionary distinctiveness of the group as a whole. *Anoplophyra* is a typical example. *See* CILIOPHORA; HOLOTRICHIA; OLIGOCHAETA; PROTOZOA.

John O. Corliss

Astrapotheria

A relatively small group of extinct South American ungulates, ranging from the late Paleocene to the late Miocene. They are customarily divided into two suborders: the late Paleocene–Eocene Trigonostylopoidea, and the early Eocene–late Miocene Astrapotheroidea. Eocene members of the order, such as *Albertogaudrya*, tended to be large animals and exhibited the enlargement of the canines characteristic of later astrapotheres.

The order, as recognized today, was probably derived independently from the ancestral condylarths, which were very likely present in South America at the end of the Cretaceous or early Paleocene. One of the most spectacular and advanced members of the order was *Astrapotherium* (see **illus.**). This animal, known from the late Oligocene to the late Miocene, averaged 9–10 ft (2.7–3 m) in length, although some other forms grew even larger. The anterior part of the skull was striking with the huge, persistently growing, curving canines. Although the upper incisors were lost and the premaxillae were drastically reduced, the lower incisors were present along with large lower canines. The cheek teeth were composed of relatively small premolars and of very large molars. The retracted, chopped-off appearance of the snout region strongly suggests that this animal had a moderately large trunk. It is likely that, as in living cows and deer, the well-developed lower incisors cropped against a horny pad above. All this, however, leaves the function of the large canines unexplained. The front legs were somewhat more strongly constructed than the hind ones, and this makes the habits of this animal a puzzle. *See* ARCHAIC UNGULATE; MAMMALIA.



Astrapotherium magnum **skeleton. Genus is known from late Oligocene to late Miocene.**

Frederick S. Szalay

Astrobiology

The scientific discipline that studies the origin, evolution, distribution, and future of life in the universe. It is an interdisciplinary science integrating contributions from biology, geology, astronomy, paleontology, and planetary science, among others. The first use of the term appears to have been in 1941 when it was defined as the subject of life in the universe other than on Earth. The U.S. National Aeronautics and Space Administration (NASA) adopted the term in 1996 and expanded the meaning to include the origin and history of life on Earth, and the future of life in the universe. In 1998, the NASA Astrobiology Institute was established, solidifying this new discipline within the science community.

Astrobiology is perhaps best understood by starting with one of its most tangible areas of study, the multiplicity of living organisms that surround us here on Earth. It is natural to ask how such diversity evolved, and when during the Earth's early history did life itself begin? To address these questions one must understand the composition of prebiotic chemicals which were present during this period, as well as the physical and environmental conditions of the early Earth in which these chemicals interacted. Investigating these areas will naturally lead one to pose broader questions, "How and when during life's origin and earliest evolution were these critical prebiotic materials formed or delivered to the Earth?" To understand the origin and evolution of life on the early Earth, one must also look to other examples of planet formation both within our solar system and as part of other solar systems throughout our galaxy. What are the factors that make a planet such as the Earth habitable? Are habitable planets common or quite rare in the universe? If life does exist on such habitable planets, how can it be detected? Such is the thread that ties all of the cosmos together, both inanimate and living, into the new discipline that is astrobiology.

Life on early Earth. The first record of life on Earth has been observed in the form of cyanobacteria-like microfossils believed to have been laid down approximately 3.5 billion years ago during the Archaean (3.8 to 2.5 billion years ago). This stage of life on Earth appears to have already evolved to the level of complex interacting microbial communities, and seems to have done so in a relatively brief period of time. There is compelling evidence indicating that until about 3.8 billion years ago the Earth was subject to an intense meteorite bombardment of debris left over from the formation of the solar system, and the intensity of these impacts would have likely sterilized the planet—destroying any nascent life. The geologic record suggests, therefore, that only about 300 million years was required for the first life to appear and then to evolve to a high level of complexity. This initial life was microbial, which continued as the only form of life for the next billion years of the Earth's history. Much of the current research in astrobiology is involved with understanding the mechanisms of adaptation and evolution which shaped this microbial life on Earth and tracing (mostly) microbial life back to it earliest forms. Comparing the genetic information carried within each organism, in the form of the nucleic acids RNA (ribonucleic acid) and DNA (deoxyribonucleic acid), astrobiologists have been able to resolve the "tree of life" to great detail, and have begun to understand the nature and metabolism of the earliest microbes. This biological information, together with geological evidence of the physical Earth during this early period of its history, sets the conditions for understanding the origin of life on Earth. *See* ARCHEAN; BIOSPHERE; CYANOBACTERIA; HADEAN.

Chemical origin of life. While a rigorous definition of life has proved elusive, there is general agreement among astrobiologists that three conditions must be present for life to exist: liquid water, organic compounds as a source of nutrients, and a biologically useful source of energy. There are a variety of potential sources of energy that astrobiologists are examining, including radiation, photochemical products, minerals, and reduced gases. Life seems quite opportunistic in this regard; it has been found in nearly every possible habitat on Earth, including the polar regions, beneath the subsurface, in desert environments, and even associated with the extreme conditions of hydrothermal vent systems. Liquid water is the one factor that is present in all of these environments, and to the best of our knowledge it appears to be a required factor, one which may limit the distribution of life elsewhere in the universe.

Given that these conditions were met on early Earth, how then did life begin? To begin to address this, researchers have acknowledged that all life on Earth shares a common set of biochemical machinery to orchestrate the processes of life.

These processes involve the use of DNA to store genetic information, a universal genetic code enabling a DNA sequence to specify a protein, protein enzymes to act as chemical catalysts which drive the cells complex chemical processes, and additional protein synthesis machinery based on RNA. Identifying the evolutionary origin of DNA and its associated proteins has proven very difficult, since each requires the other for its own synthesis. Researchers have been faced with the dilemma of which came first? One possible scenario which has been proposed that could bypass this circularity is the concept of an entirely "RNA world" that might have been the basis for the first forms of life on Earth. RNA molecules can replicate and carry information (like DNA), but they can also act as catalysts (ribozymes) much as proteins do. These abilities suggest that RNA could have originally provided both the genetic code and the catalytic function, and these roles could subsequently have been taken over by DNA and proteins. This stage of life's early history, however, continues to be controversial as well as an area of intense study. *See* DEOXYRIBONUCLEIC ACID (DNA); ENZYME; NUCLEIC ACID; PROTEIN; RIBONUCLEIC ACID (RNA); RIBOZYME.

Although the concept of an RNA world has reached some level of acceptance, it is unlikely that RNA itself, with the associated complexity of its four nucleotide bases and a ribose phosphate backbone, represents the earliest prebiotic molecule. Research on nucleic acids with hexoses (six-carbon monosaccharide sugars), for example, suggests that a wide variety of related informational macromolecules are possible. Nonsugar alternatives have also been proposed, such as peptide nucleic acid (PNA). *See* MONOSACCHARIDE; PEPTIDE.

In any case, it is clear that the prebiotic chemistry which led to life at some point must have been segregated from the changing surrounding environment to achieve the stability necessary for continued evolution. Among present-day organisms this segregation is achieved by means of a membrane composed of phospholipids and proteins. A fundamental area of research related to prebiotic chemistry, therefore, concerns the origin of the first cell membrane, which made possible a regulated vesicle for the chemistry and metabolism of life. *See* PREBIOTIC ORGANIC SYNTHESIS.

Source of life's building blocks. To fully understand the biochemical mechanisms leading to the origin of life it is necessary for astrobiologists to search back further in time, and identify what building blocks, in the form of the earliest organic compounds, were present on the early Earth. In characterizing this early material, researchers must address the related question, "How were these critical compounds delivered to our planet?" A significant fraction of this precursor material is believed to have been acquired as a late-accreting layer from impacts of C-type asteroids and comets during the period of heavy bombardment of the inner solar system 4.5 to 3.8 billion years ago. In addition to simple volatile molecules such as water and short-chain hydrocarbons, these C-type asteroids and comets are also rich in complex organic chemicals. *See* ASTEROID; COMET; COSMOCHEMISTRY; METEORITE.

An important part of the puzzle, the idea of a cometary source for prebiotic organics on Earth, in some respects simply moves the question out into the cosmos, leading astrobiologists to ask, "What is the ultimate source and distribution of organic matter in the universe?" One place they are looking at are the dense interstellar clouds, the birth sites of stars and star systems which include planets. Interstellar ices are composed primarily of water, but they have also been shown to contain some ammonia, carbon monoxide, carbon dioxide, and the simplest alcohol, methanol. It has been known for some time that when similar icy solids are exposed to conditions of ultraviolet irradiation, more complex chemicals can be produced than those originally present in the ice, and astrobiologists speculate that some of these chemicals might have played an important role in the chemistry of the early Earth and in the prebiotic chemistry of planets in general. *See* INTERSTELLAR MATTER; MOLECULAR CLOUD.

Life in extreme environments. Life on Earth has been found in nearly every environment in which water is present, including extremes of temperature, oxygen, pH, salinity, radiation, light intensity, and pressure. The study of life in these extreme environments is one of the most active areas of research in astrobiology, focusing on the environmental limits at which life can be found, the types of microorganisms which can be found under these conditions, how they are able to adapt, and where they occur in the "tree of life." Viable organisms have been discovered frozen within permafrost over geological time scales as well as associated with hot springs and hydrothermal fluids venting from Earth's subsurface crust onto the deep seafloor at temperatures near 120°C (248°F). Other organisms, such as the primitive unicellular red alga *Cyanidium caldarium* or the acidiphilic archaebacterium *Sulpholobus acidocaldarius*, can tolerate acidic conditions down to a pH of 1–2. Halophiles such as *Halobacterium* and *Halococcus*, which can tolerate high salinity, have been recorded at salt concentrations as high as 1.5–4 M (moles per liter of solution). This aspect of astrobiological research has greatly expanded our view of the versatility of life. The results of this work have broadened the scope of possible environments within our solar system which could harbor life as well as formed the basis for searches for particular sites, for example on Mars or Europa, which should be the focus of future space missions. *See* ARCHAEBACTERIA; HALOPHILISM (MICROBIOLOGY); MARINE MICROBIOLOGY.

Biosignatures. The chemistry of life is inherently different from abiotic chemistry, and all organisms leave evidence of this in the environments with which they interact. Physical properties of an environment indicative of previous life are termed biosignatures, and their characterization and use for detecting possible life elsewhere in the universe is a major component of astrobiological research. Biosignatures are generally of two types, geologic or astronomical. Those associated with geological samples include the distribution of organic compounds, the chirality of the molecules (whether they are left or right handed), the isotopic composition of biological elements, biomineralization, and even the presence of microfossils. Astronomical biosignatures are also being investigated now that it is possible to characterize the chemical composition of planets and other extraterrestrial bodies both within and outside our solar system. Such biosignatures are often associated with atmospheric composition but can also reflect the chemical make-up of the surface as well. Spectral biosignatures of the Earth, for example, would indicate the presence of liquid water, chlorophyll on the surface, and abundant oxygen in the atmosphere. More generally, the simultaneous presence of both strongly reduced gases and oxidized gases which are not in chemical equilibrium (such as oxygen and methane in the Earth's atmosphere) is considered a reliable biosignature for many types of planetary atmospheres. *See* ASTRONOMICAL SPECTROSCOPY; MOLECULAR ISOMERISM.

Habitability of other planets. One of the fundamental questions in astrobiology asks, "Where among what is likely to be a vast number of planets and moons are conditions suitable for life as we know it on Earth?" In other words, "What parts of the universe are habitable?" Recalling the three ingredients upon which life depends—liquid water, organic molecules (or carbon), and a suitable energy source—leads to the conclusion that life is a planetary phenomenon. Initially, the habitable zone of planets around stars was thought to encompass planets whose orbits are close enough to their sun for solar energy to drive the chemistry of life, but not so close as to boil off water or break down the organic molecules on which life depends. Within our own solar system this region would encompass perhaps Venus and Mars in addition to the Earth. More recently, however, astrobiologists have come to realize that the habitable zone may be much larger than originally conceived, since other localized sources of heat and energy could be present at great distances from the Sun. The strong gravitational pull caused by large planets, for example, may produce enough energy to sufficiently heat the cores of moons orbiting planets at great distances from their star. This situation is believed to be occurring between Jupiter and its moon Europa, where evidence suggests that a large, salty ocean is present beneath the surface. *See* JUPITER; SATELLITE ASTRONOMY.

The likelihood of planets or other bodies being present in the habitable zone of solar systems other than our own is dependent on, among other factors, the frequency with which planets form around stars in general. After centuries of speculation, scientists have confirmed just since 1995 that there are indeed planets orbiting other stars, and more than 160 extrasolar planets had already been discovered by 2005. While almost all of these extrasolar planets seem to be gas giants, like Jupiter, it is likely that Earth-like worlds also orbit other stars, and instrumentation with sufficient precision to detect a world as small as Earth should be available by 2014. *See* EXTRASOLAR PLANETS.

Edward M. Goolish

Bibliography. L. J. LaFleur, Astrobiology, *Astronomical Society of the Pacific Leaflet Series*, Leaflet no. 143, pp. 333–340, 1941.

Astrometry

That part of astronomy dealing with the position, distance, and motion of celestial objects, including solar system objects, stars, radio sources, and galaxies.

New astrometry. Astrometry has changed since the 1980s. Very long baseline interferometry (VLBI) observations of extragalactic radio sources at milliarcsecond accuracies have provided the basis for a fixed, epoch-independent reference frame. The *Hipparcos* astrometric satellite has provided an optical catalog at milliarcsecond accuracy. Charge-coupled devices (CCDs) have replaced photographic plates and visual observing. The old, single-star position, observing programs have been replaced by programs for observing large numbers of stars and detecting near-Earth objects and extrasolar planets. Large telescopes with CCD mosaics are being used for astrometric surveys, and space missions are planned. The purpose of astrometry is no longer just to determine star positions; rather it is also to obtain positions, motions, and distances in support of astrophysics and the discovery of new objects.

Reference systems. In 1994 the International Astronomical Union (IAU) adopted the International Celestial Reference Frame (ICRF), based on about 400 extragalactic, distant, radio sources, which have no apparent proper motions, as the fundamental reference frame. The individual sources may be subject to changes in the source structure, but the reference frame should be quasi-inertial. This kinematically based system replaced the dynamical system based on the solar system motions. The ICRF replaced the previous fundamental catalogs, such as the FK5, which were based on positions and proper motions of bright stars. *See* CELESTIAL REFERENCE SYSTEM.

The *Hipparcos Star Catalog*, from observations by that astrometric satellite, provides an accurate optical catalog based on the ICRF. Other star catalogs, such as the Tycho and USNO CCD Astrographic Catalog (UCAC), provide a denser coverage of the sky and can reach fainter magnitudes, but with reduced accuracies. The positions and proper motions of the stars provide a two-dimensional map of the sky for a given time. To provide the three-dimensional aspect, the parallaxes (which give the distances to the stars) and the radial velocities are necessary. *See* PARALLAX (ASTRONOMY).

In addition to a reference frame, the International Celestial Reference System (ICRS) includes the definition of time, algorithms (such as the precession-nutation model), constants, and the relationship between the celestial and terrestrial reference frames. *See* ASTRONOMICAL COORDINATE SYSTEMS; ATOMIC TIME; DYNAMICAL TIME; EARTH ROTATION AND ORBITAL MOTION; NUTATION (ASTRONOMY AND MECHANICS); PRECESSION OF EQUINOXES; TIME.

Positional catalogs. Astrometric data are provided in the form of catalogs. One kind is an observational catalog, which can be the publication, or a data set, of individual observations. The observations can also be reduced to a common instrumental system with the individual observations of a given star combined to form a single position for each star. Such observational catalogs that determine their own reference frames are known as absolute catalogs.

Observations can be made at different wavelengths, such as radio, optical, and infrared. The transformations between the ICRF, dynamical reference frames, and optical, infrared, x-ray, and other catalogs must be determined from observational data. *See* ASTRONOMICAL CATALOGS.

Observing programs. Observations can be divided into those involving large-angle and small-angle measurements. Large-angle measurements determine the difference in position between objects over large angular distances in the sky. Historically, such observations were made by transit circles, which record, at the time of transit of an object, both its right ascension and its declination. *See* ASTRONOMICAL TRANSIT INSTRUMENT.

Interferometers, observing at radio, optical, or infrared wavelengths, combine the reception of the emission from a source at two separate detectors. By measuring the time difference between the two detections, a very accurate measurement of the angle to the source can be determined. Thus, at radio wavelengths, VLBI uses antennas located as far apart as possible to maximize the baseline between the two instruments and to develop the most accurate observations. The data are recorded on tapes and shipped to a correlator for analysis. At optical wavelengths, the signals must be combined immediately, and the differences in the atmospheric effects must be minimized. However, since optical wavelengths are so much shorter than radio wavelengths, milliarcsecond accuracies can be achieved over baselines of tens of meters, while for radio-frequency measures, distances of thousands of kilometers are required. *See* RADIO ASTRONOMY.

The *Hipparcos Astrometric Satellite* used a technique for observing pairs of stars separated by approximately 60° to form a catalog of stars located throughout the sky. The *Hipparcos Catalog* provides the positions and proper motions for stars, measured over a 3-year period in 1990–1993, at accuracies of approximately 1 milliarcsecond at the observational epoch and degrading in accuracy at 1 milliarcsecond per year, because of the accuracy limits of the proper motions.

Small-angle measurements are made by individual-exposure techniques, for example, measuring the positions of all the objects on a CCD exposure. They provide accurate relative positions of the observed objects. They can also provide, by means of multiple observations, the parallaxes and motions of the stars with respect to reference stars in the field. CCDs, which can be mounted in mosaic cameras, offer the advantages of a very linear and efficient response to detected photons, and an instantaneous digital readout of the field. *See* ASTRONOMICAL IMAGING; CHARGE-COUPLED DEVICES.

Overlapping exposures and global solution techniques can be used to develop the small-angle measurements into large-angle solutions. Also, combining many exposures can provide surveys of the skies. *See* ASTRONOMICAL ATLASES.

For accurate positional work, multiple stars must be identified, since it is necessary to take into account the motion of the center of light of a multiple star as the components move with respect to each other, or to observe the individual moving components. Speckle interferometers take very rapid exposures (approximately 30 per second) to freeze atmospheric effects. These short exposures can then be added together to measure the separation, relative position, and magnitude of pairs of stars that could not otherwise be observed with such accuracy through the atmosphere. The atmosphere is the primary limitation on ground-based astrometric accuracy. *See* BINARY STAR; SPECKLE; TWINKLING STARS.

P. K. Seidelmann

Bibliography. E. Hoeg and P. K. Seidelmann (eds.), *Astronomical and Astrophysical Objectives of Submilliarcsecond Astrometry*, IAU Symp. 166, 1995; G. H. Kaplan, *The IAU Resolutions on Astronomical Reference Systems, Time Scales, and Earth Rotation Models*, USNO Circ. no. 179, 2005; J. Kovalevsky and P. K. Seidelmann, *Fundamentals of Astrometry*, 2004; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, 1992; P. K. Seidelmann and A. K. B. Monet (eds.), *Astrometry in the Age of the Next Generation of Large Telescopes*, 2005.

Astronautical engineering

The engineering aspects of flight and navigation in space. The latter is also known as astronautics. The terminology thus parallels aeronautical engineering and aeronautics as applied to aviation. Astronautical engineering deals with vehicles, instruments, and other equipment used in space, but not with the sociological or economic aspects of space flight, except as they influence the equipment.

Space. The boundary between the atmosphere and space is a matter of debate. Few contend that, for practical purposes, it lies much more than 100 mi (160 km) above the surface of the Earth; some would put the boundary lower. For purposes of record, the Fédération Aeŕonautique Internationale has established the line of demarcation between the atmosphere and outer space at 62 mi (100 km) above the surface of the Earth. Over 99% of the Earth's atmosphere lies within 20 mi (32 km) of its surface, but the air thins slowly above that height, merging imperceptibly with outer space, which even in its farthest reaches may contain a molecule or two per cubic inch. With the boundary of space so indeterminate, a more precise definition of a space mission is on the basis of velocity, the space mission being at or above the velocity required to circle the Earth completely as a satellite.

Because there is no sharp line of demarcation between the atmosphere and space, some overlap exists between aeronautical and astronautical engineering. Nevertheless, astronautical engineering is unique in many aspects; there are far more differences than similarities in the two disciplines. *See* AERONAUTICAL ENGINEERING.

Aircraft and spacecraft contrasts. There is a lack of parallelism between current astronautic and aeronautic vehicle terminology. An aircraft is a self-contained vehicle, having within its structure essentially all the equipment required to transport its payload from one place to another. A spacecraft, in the more restricted sense, is the container for the payload. Sometimes the word is used to denote the container and payload. Most spacecraft, to date, have had either very limited propulsion or none at all.

Since enormous speeds are the hallmark of all astronautic missions, unpowered spacecraft require a booster, or launch vehicle, usually a rocket many times as large as the spacecraft. The weight of the spacecraft, in fact, seldom exceeds 5% of the total launch vehicle weight.

The distinction between spacecraft and launch vehicle has been further clouded by the NASA space shuttle, which uses as its orbital stage a winged airplane. This so-called orbiter is capable of launching spacecraft in the more traditional sense, but also of remaining in orbit for several days and performing useful missions during its orbital stay. The shuttle also carries a crew. Its wings and other airplanelike features are used only in the descent and landing. *See* SPACE SHUTTLE.

Aircraft structures are designed to house the payload and to obtain lift and control by interaction with the atmosphere. Spacecraft structures must properly house the payload in both the atmospheric and space environments. Space vehicles are usually expended on one flight; aircraft are designed for many years of service. Spacecraft are lifted by the thrust of their rocket engines, steered by directing the thrust, and usually employ no wings or fins. Where aerodynamic effects are encountered, such as in leaving and reentering the atmosphere, the speeds involved usually differ from those of airplanes by an order of magnitude. *See* SPACECRAFT STRUCTURE.

Airplanes are most frequently propelled by air-breathing engines, which use the oxygen of the air for combustion. They produce their thrust by changing the momentum of the air. Spacecraft use rocket engines exclusively, although some proposals have been made to use air-breathing engines for the initial phases of a spaceship's ascent. Rockets contain within themselves both the energy (that is, the fuel and oxidizer) and the mass whose momentum is changed for propulsion. *See* ROCKET PROPULSION.

Airplanes are usually guided by a human pilot, spacecraft by automatic or remote controls. The precision required in space navigation makes it probable that spacecraft will always depend heavily on automatic control, although people may serve to monitor and adjust the equipment. *See* SPACE NAVIGATION AND GUIDANCE.

Electric power for auxiliary purposes is easily generated in aircraft. The propulsion engines for airplanes must operate continuously during flight to overcome the ever present aerodynamic drag. Electric generators, driven by the main engines, supply the required power. In contrast, space vehicles coast freely during practically all of their useful lives. Solar cells and nuclear reactors are two obvious means of supplying electric power. Fuel cells using hydrogen and oxygen have also been used for shorter-duration applications. *See* SPACECRAFT PROPULSION.

Cost and reliability. It is extremely expensive to put a pound of payload into Earth orbit. Thus designers have been justified in going to great lengths to convert a pound of structure into a pound of payload. Great improvement appears possible in this respect; only the cost of the propellant seems to be irreducible.

In view of the high cost of space operations, it is especially important that space vehicles operate long enough to successfully fulfill their missions. A severe reliability requirement is thus imposed upon vehicles and equipment intended for missions, such as journeys to the planets, which may require up to a year or more to accomplish.

The requirements for high reliability and low weight add tremendously to the cost of the payloads themselves, to the extent that their cost approaches that of the launch vehicle. In the Apollo lunar missions the cost of the last three stages, which included the lunar module and the service module, came very close to that of the first three stages comprising the Saturn V launch vehicle, even though the weight of the payload was only about 1.5% that of the launch vehicle. In their eagerness to minimize the cost of the launch vehicle, the designers may have overreached themselves. Obviously, when both expendable launch vehicles and expendable spacecraft are used, the objective should be minimum combined cost. This minimum does not necessarily occur when the costs of launch vehicles and spacecraft are equal.

All space missions through 1975 used expendable launch vehicles. The reusable space shuttle was expected to reduce the costs of Earth-to-orbit transportation. However, preliminary indications are that the shuttle will not prove nearly so economical as originally expected, largely because it is only partially reusable, and because the recovery provisions are heavy and complex.

Designing for minimum cost is of extreme importance in astronautical engineering because of the very high unit costs involved and because frequently only one approach is selected. The wrong approach could therefore waste billions of dollars and still not be recognized as an inferior one.

Because astronautical engineering is a transportation science, it is convenient to discuss it in terms of vehicle, environment, and interaction between the two.

Escape velocity. Gravity is a dominating influence in the design of space launch vehicles. To achieve even a low orbit, a vehicle must expend an amount of effort equivalent to climbing out of a well some 4000 mi (6400 km) deep. The task is roughly equivalent to driving a car from San Diego, California, to Bangor, Maine, while dragging a cement block as heavy as another automobile, with no gas stations en route. As a result, a space vehicle as it sits on a launching pad consists mostly of propellant.

The pull of the Earth's gravity on a body extends indefinitely far into space and varies inversely with the square of the distance to the center of the body. More generally, gravitational attraction exists between any two bodies. Isaac Newton expressed the quantitative relation as $F = g(m_1 m_2/r^2)$, where g is a universal constant, m_1 and m_2 are the masses of the bodies in question, and r is the distance between their centers of mass.

Despite the fact that the pull of gravity extends to infinity, it is nonetheless possible to escape permanently from the Earth's gravity in the sense of never being drawn back to the ground. The key is speed. Circular velocity is the minimum at which a space vehicle can remain permanently above the Earth. At low altitudes, this velocity is about 25,000 ft/s (7.9 km/s). *See* SATELLITE (SPACECRAFT).

As the speed is increased above the circular velocity, the path of a vehicle becomes a larger circle or an elongated ellipse. When the speed reaches 37,000 ft/s, or about 7 mi/s (11.2 km/s), the path becomes a parabola and the vehicle will travel along one of the legs to infinity without further propulsion. *See* ESCAPE VELOCITY.

Propulsion requirements. These velocities are tremendous by any previous standard. To reach them, a vehicle must carry the corresponding amount of energy in the form of propellant.

Even with the most energetic propellants and the lightest structures, it has not yet been possible to reach orbital velocity with a single rocket. To overcome this seemingly insurmountable obstacle, one rocket is carried as the payload of a larger one. When the larger burns out, the second is ignited and adds its velocity to that of the first. This is known as the step-rocket or staging technique. *See* ROCKET STAGING.

For lunar and planetary missions, lightweight vehicles, powerful propellants, and many stages are used. The lunar orbit rendezvous method required a total of six stages to take the Apollo astronauts to the Moon and back. The launch involves the full thrust of the first and second stages and the partial burning of the third stage of a Saturn V. The third stage is reignited to achieve escape velocity. The fourth, the service module (SM), uses some of its propellant to enter an orbit around the Moon. The fifth stage, the lunar module (LM), descends to the surface of the Moon, and the sixth returns to lunar orbit to rendezvous with the service module, at which time the astronauts and payload of the lunar ascent module are transferred to the command module. The service module then propels the command module out of lunar orbit and on a highly elliptical return to Earth. The command module accomplishes braking in the Earth's atmosphere and finally lands by parachute in the ocean.

Three different propellant combinations are used for primary propulsion. The first stage, which comprises 78% of the liftoff weight, uses dense, cheap

-s

propellants—liquid oxygen and kerosine. The second and third stages use liquid oxygen and liquid hydrogen. Hydrogen gives high performance but has a very low density and boiling point. The service module and the lunar module, which must coast in space for several days, use noncryogenic nitrogen tetroxide and mixed hydrazines.

The main engines of the space shuttle have pushed chemical propulsion technology almost to its limit. Hydrogen and oxygen are the propellants, and the combustion pressures have been raised to some 3000 lb/in.^{2} (20.7 megapascals). Very little, if any, further reduction in fuel consumption can be had through additional increases in pressure. Many people, in fact, question whether the small gains achieved in going higher than the pressures used in the Saturn V upper stages have justified the increased expense, reduced reliability, and increased maintenance.

Future vehicles may use nuclear power. Fission reactors were used to heat hydrogen gas in the Nuclear Engine Rocket Vehicle Application (NERVA) Program, and propellant consumption amounting to a third less than that achieved by the best chemical combination was demonstrated. Costs, political pressure, and the presence of intense harmful radiation have thus far prevented any practical application of nuclear power for rocket propulsion. Controlled nuclear fusion represents a longer-term, potentially very attractive possibility.

Guidance and control. Although propulsion is the key to space flight, other elements are essential and present numerous new problems. One such element is guidance and control. For the ascent phase of space vehicle flight, guidance systems similar to those used for ballistic missiles are employed.

Another control requirement of many types of space vehicles is that of maintaining the desired vehicle attitude over long periods of time. Displacement gyroscopes, even excellent ones with very low drift rates, cannot provide an accurate reference for days or weeks. Such devices must be corrected frequently by an external reference.

At least two such references are available: sources of electromagnetic radiation, and the gravitational gradient. The first might be used by such devices as a Sun seeker, a star tracker, or a horizon scanner. All of these detect sources of emitted or reflected radiation. Of special interest is the horizon scanner, which indicates the direction of the vertical with respect to the Earth. This direction is vital for orbital missions that relate to occurrences on the ground such as meteorological observation.

In the vicinity of the Earth (or any large celestial body) the difference in the pull of gravity between points on the craft having different distances from the Earth can be usefully employed.

Reaction wheels or other devices capable of storing angular momentum may be used to provide the torque to effect or maintain a given orientation. Such devices are very efficient, both from a weight and an energy standpoint, where disturbing torques on the spacecraft are small, random, and long continued. At the opposite end of the torque spectrum, torques that are large and uncompensating, rocket engines are the most suitable. For many applications reaction wheels are useful to absorb the small random torques, and rockets handle torques and total momenta which exceed the capacity of the wheel system.

Auxiliary power. Vehicle and payload equipment require electric power. This power, generally speaking, must be provided in rather generous quantities, and, in many cases, for long periods of time. The energy required can be carried along, or it can be supplied by the Sun. For small amounts of energy, chemical sources, such as batteries or chemically fueled generators, may be used. A great deal more energy can be obtained from a nuclear reactor. Energy also comes continuously from the Sun but at a fairly low density at Earth's distance.

Nuclear reactors no larger than a basketball appear possible. Large power outputs are possible, but at the expense of considerable weight. An additional, stringent limitation to the use of nuclear reactors is the radiation that they give off. For human crew members or sensitive payloads, heavy shielding is required.

Where small amounts of power are needed, but for long periods of time, solar energy presents an ideal source. Solar batteries are particularly attractive, because they have no moving parts and thus offer good dependability. Present solar cells operate at 8–10% efficiency with some improvement in prospect. In space near the Earth, each square foot of surface normal to the Sun's rays receives about 150 W of solar energy. Of this, perhaps 10 W can be delivered to the payload. Various schemes for obtaining the required intercept area have been proposed. If power requirements are low, the solar cells are mounted on the spacecraft. If the entire external area of the craft does not provide sufficient area, special panels are provided that unfold once a space environment is reached. *See* SOLAR CELL; SPACE POWER SYSTEMS.

Communications equipment. The communications equipment comprises an essential item of nearly all space vehicles. This equipment is designed for light weight, low power consumption, and, usually, long life. It may be designed especially to operate in a strong radiation environment. The equipment transmits data generated within the space vehicle, receives commands from the Earth, or emits signals to permit the vehicle to be tracked by a ground station. *See* SPACE COMMUNICATIONS.

Payload. Nearly all space vehicles have the items previously mentioned: propulsion, guidance, power supply, communications. In addition, there will be specialized payload items, depending on the function the vehicle is to perform. These may include scientific instruments, telescopes, communications relay and storage equipment, or human beings. There must also be a space frame to house all of these and to provide the required environment in the vacuum of space.

Ground communication and support. Although a large percentage of the problems and most of the romance of space flight are associated with the vehicles, it would be a mistake to assume that these constitute even a major fraction of the total operating system. Indeed, the cost of overcoming the Earth's gravity is so great that any portion of the total operation which can be performed on the ground should be done there.

The supporting ground equipment consists of the preparation and launching equipment, and the tracking, communications, and payload-oriented equipment for turning the received data into usable form. For missions which involve return of space vehicles or booster rockets, recovery equipment may also be required.

Launch and checkout equipment closely resembles (and sometimes is identical with) that used for large ballistic missiles. Ground-to-space communications equipment is generally distinguished by the use of large, high-gain antennas, coupled with extremely sensitive receivers. These receive the low-power signals from the spacecraft. Precise radio and optical tracking systems, together with electronic computers, permit rapid and accurate determination and prediction of the position of space ships. *See* LAUNCH COMPLEX; SPACECRAFT GROUND INSTRUMENTATION.

Reentry. In their interaction with the terrestrial and atmospheric environment during reentry, space vehicles resemble ballistic missiles. However, although ballistic reentry techniques have been proved successful, the use of winged vehicles also has certain attractive aspects. There is a basic difference in these two methods in respect to the way atmospheric heat is handled. The ballistic approach absorbs the heat in the reentry body or rejects it back to the air by mass transfer. The winged vehicle dissipates the heat by radiation. Considerable research has been done on compromise reentry vehicles, such as the lifting body approach. *See* ATMOSPHERIC ENTRY.

The orbital stage of the shuttle is a winged craft designed to land like an airplane. It will utilize a combination of techniques to overcome the reentry heating problems: lift, temperature-resistant materials, and local ablative cooling.

In addition to the problem of reentry into the Earth's atmosphere, some missions require entry into the atmospheres of other planets, notably Mars, Venus, and Jupiter. These atmospheres are strikingly different from that of the Earth in composition, density, and density gradient. The same general techniques will probably apply, but optimal designs will differ quite sharply from Earth atmosphere reentry systems. *See* HYPERSONIC FLIGHT.

Environment. Astronautical engineering must contend with the unique environment of space outside the Earth's atmosphere. Here it is necessary for the engineer not only to learn and apply new information, but to dismiss many old concepts and rule-of-thumb procedures.

Although gravity is present in space, whenever a vehicle is coasting unpropelled, the shell and everything in it are acted on equally by gravity and therefore appear weightless. Loose internal objects do not "fall" relative to fixed items. Fluids do not flow naturally, but must be confined and extruded. Liquids exposed to the vacuum of space evaporate or freeze. External transfer of heat takes place only by radiation.

Metals exposed to the ultraviolet rays of the Sun emit electrons. Small particles of cosmic dust strike external surfaces at extremely high velocities and gradually erode them. Cosmic radiation creates a spectrum of secondary radiation that may reach levels damaging to equipment or personnel.

Starting rotating equipment within the vehicle may set the latter into counterrotation so the momentum may be conserved. In dealing with this environment the modern engineer must think twice to avoid being misled by earthbound experience. *See* SPACE FLIGHT. Robert C. Truax

Bibliography. V. S. Avduevskii and G. Uspenskii, *Scientific and Economy-Oriented Space Systems*, 1988; L. Friedman, *Star-Sailing: Solar Sails and Interstellar Travel*, 1988; N. Johnson and D. McKnight, *Artificial Space Debris,* 1987; V. L. Pisacane and R.C. Moore (eds.), *Space Systems*, 1994; O. Zarrouati, *Space Trajectories*, 1987.

Astronautics

The application of scientific principles and engineering techniques to flight in space. Astronautics deals with space vehicles in the sense that aeronautics deals with aircraft. A distinguishing feature between astronautics and aeronautics is the extent to which the vehicles are influenced by the Earth's atmosphere. *See* AERONAUTICS; ASTRONAUTICAL ENGINEERING.

The subject matter of astronautics is flight at altitudes where a vehicle overcomes gravitational attraction and controls its course by reactive propulsion. Aeronautics concerns flight in regions where a vehicle resists gravitational attraction and controls its course by aeromechanical forces. The distinction is convenient but not clear cut. Rockets, by their reaction, assist airplanes to take off. Space vehicles, such as the shuttle, may glide back to Earth. *See* INTERPLANETARY PROPULSION; ROCKET PROPULSION.

During the twentieth century, aeronautics progressed from a human ambition to a commercial and military reality. Astronautics looks forward to similar progress during the twenty-first century. *See* INERTIAL GUIDANCE SYSTEM; NAVIGATION; RELIABILITY, AVAILABILITY, AND MAINTAINABILITY; SPACE; SPACE BIOLOGY.

Thousands of successful space missions have been flown. Crewed Earth orbital missions included two practice missions under Project Apollo and three Skylab missions. The former Soviet Union completed many crewed Earth orbital missions with its Soyuz and Salyut spacecraft. *See* SPACE FLIGHT; SPACE STATION.

A total of nine crewed lunar voyages have been made, seven of them involving landing. Two included considerable surface travel with a small, jeep-sized vehicle, the lunar rover. All were accomplished without fatalities. One launch pad accident resulted in the death of three Apollo astronauts, V. I. Grissom, E. H. White, and R. R. Chaffee. Four Soviet cosmonauts died during reentry from Earth orbital missions; there was one fatality in 1967 and three in a 1971 mission.

The Apollo lunar landing missions involved complex feats of navigation and control, as well as many stages of propulsion. The basic flight plan included Earth orbit, transfer from Earth to lunar orbit, descent to the surface of the Moon, return to lunar orbit, and return to Earth. The lunar descent was made by the two-stage lunar module (LM). The service module (SM), crewed by one of the three astronauts, was left in orbit. After the surface exploration, the ascent stage of the lunar module rejoined the service module in lunar orbit. The service module provided propulsion to return the command module to the Earth. The command module alone made the final reentry with the three astronauts.

Equally sophisticated feats of astronautics were performed by some of the crewless missions. There were literally hundreds of relatively simple Earth orbital missions by both the United States and the Soviet Union. They included military reconnaissance satellites, communications satellites, and satellites for weather observation, recording radiation intensities, and biological investigations. *See* COMMUNICATIONS SATELLITE; MILITARY SATELLITES; SATELLITE NAVIGATION SYSTEMS; SCIENTIFIC AND APPLICATIONS SATELLITES.

The United States alone has launched successful crewless probes to every planet but Pluto. It is now possible to manipulate many of these probes by radio command from Earth—to correct their courses, start, stop, and aim their cameras, and control a great variety of other on-board equipment. *See* SPACE NAVIGATION AND GUIDANCE; SPACE PROBE.

Two notable accomplishments were the soft landing of crewless spacecraft on Mars and swingbys of Jupiter and Saturn by the Voyager probes. In both cases, spectacular photographs were relayed back to Earth, and a great deal of other data were obtained. The photographs of the moons of Jupiter were particularly breathtaking.

After termination of the Apollo program the only crewed spaceflights were orbital missions carried out by the Soviet Union. These missions set new records for human endurance in space and established the feasibility of voyages lasting at least a year.

The crewed space program of the United States resumed with completion of the space shuttle, a partially reusable craft having a winged orbital stage. A four-shuttle fleet made numerous flights with great technological success. In January 1986, on the twenty-fifth shuttle mission, however, the shuttle *Challenger* exploded shortly after liftoff, killing all seven astronauts on board. The accident was attributed to a leak in a joint in one of the solid-propellant booster rockets. The accident set the United States space program back at least 2 years. Further flights were suspended until the causes could be ascertained and appropriate remedies applied. Shuttle flights resumed in September 1988. *See* SPACE SHUTTLE.

In July 1988, Russia launched a crewless probe to the planet Mars. That country has been accumulating thousands of crew hours in space, presumably in preparation for a crewed mission to Mars. The United States has preferred to concentrate on construction of a rather large crewed space station. China has become the third spacefaring nation with a brief orbital crewed mission in 2003 and a successful five-day mission in 2005.

There has been an increased interest in commercial activities in space worldwide. The European space community has a strictly commercial launch activity in French Guiana, where Ariane vehicles are regularly launched with commercial payloads. *See* SPACE TECHNOLOGY.

Robert C. Truax

Bibliography. H. S. F. Cooper, *Before Lift-off*, 1987; J. D. Rummel, V. M. Ivanov, and J. Rummel (eds.), *Space and Its Exploration*, 1993; H. L. Shipman, *Humans in Space: Twenty-First Century Frontiers*, 1989.

Astronomical atlases

Books, CD-ROMs, or computer-accessible sets showing stars, constellations, or other astronomical phenomena and their locations in the sky. Astronomical atlases provide two-dimensional views of the distributions of objects.

The first sets of astronomical sky positions were mere lists of stars visible to the unaided eye and were not drawn to make an atlas. Hipparchus's star catalog (circa 127 B.C.) divided the stars into brightness classes, from which the current magnitude scale descends. Hipparchus's work is known mostly from Ptolemy's *Almagest*, written about A.D. 150 but available from translations about 1000 years after that. Ptolemy's catalog included 1028 stars. It is said that Ptolemy plotted his stars onto a globe; no trace of that globe is known, though other star globes may descend from it. *See* MAGNITUDE (ASTRONOMY).

Early star maps. Comets were pictured in the *Nuremberg Chronicle* (1493), but the same woodcuts were used multiple times, often in different orientations. So a question arises as to when representations of the sky become scientifically accurate.

Albrecht Dürer, in Nuremberg, made woodcuts dated 1515 that show the Ptolemaic stars and constellations, one map for each hemisphere. The first real star atlas was *De le stelle fisse* by Alessandro Piccolomini (1540), published in Venice. Piccolomini showed the different constellations, though crudely. He showed all but one of Ptolemy's constellations, and the Southern Cross as well. No constellation figures accompany the star charts. He showed four levels of brightness and marked several stars in each constellation with letters. His book was written in

Italian, a spoken language, rather than scientific and scholarly Latin. *See* CONSTELLATION.

Copernicus's historic Sun-centered diagram of 1543 was schematic, and diagramed only a circle, the outermost, for the "sphere of fixed and immobile stars." When Thomas Digges, in 1576, added an appendix to the book *Prognostication Everlasting* (1556) of his father, Leonard Digges, he provided the first discussion and first diagram of the Copernican system in English. His diagram shows stars extending outward at varying distances, something he discusses in the text. But the stars are not accurately located.

The star positions could be plotted with increasing accuracy as measuring accuracy improved. Tycho Brahe (1572) discussed his supernova, and Johannes Kepler (1604) discussed his own supernova with star diagrams included. *See* SUPERNOVA.

Golden age of star atlases. The first of the major beautiful star atlases was the *Uranometria* of Johannes Bayer, published in Augsburg in 1603. Bayer's 51 charts (**Fig. 1**), engraved by Alexander Mair, show the individual constellations, and additional northern and southern hemispheres show the overall arrangement of constellations. Bayer's atlas has grids to mark the star positions.

Bayer's lettering scheme is still used today for the stars that were so lettered. Bayer usually assigned the Greek letters in approximate order of brightness in a given constellation, with similar brightnesses assigned in order of position, and with special objects like the Big Dipper assigned Greek letters in order around the asterism. Bayer also used some lowercase Roman letters when he ran out of Greek letters for a given constellation. He used Tycho's observations for northern hemisphere objects and observations of Pieter Dirckszoon Keyzer, a Dutch navigator, for southern hemisphere objects.

Johannes Hevelius used his own observations to make an even more handsome catalog, *Firmamentum Sobiescianum sive Uranographia Joh. Hevelii*, published in Gdansk in 1687 (**Fig. 2**). He used Edmond Halley's observations of southern stars. Hevelius drew the constellations as they would be seen from outside the celestial sphere looking in (God's view), so they appear backward from the way they were in Bayer's charts or, indeed, to human eyes. He introduced 11 new constellations, 7 of which survive. Hevelius's constellation figures are sharply and incisively drawn, basically one to a page.

John Flamsteed, the first Astronomer Royal, made his own set of detailed observations of the northern stars and used them to create his *Atlas Coelestis*, published under the supervision of his friends in London in 1729 after his death. Most pages show a few constellations together. The French globe maker Fortin revised the book in 1776, dividing Hydra into two parts, and the now 27-plate atlas appeared in several French editions.

Johann Doppelmayr, a Nuremberg professor, published his *Atlas Coelestis* in 1742. It contains a variety of illustrations of astronomical events and

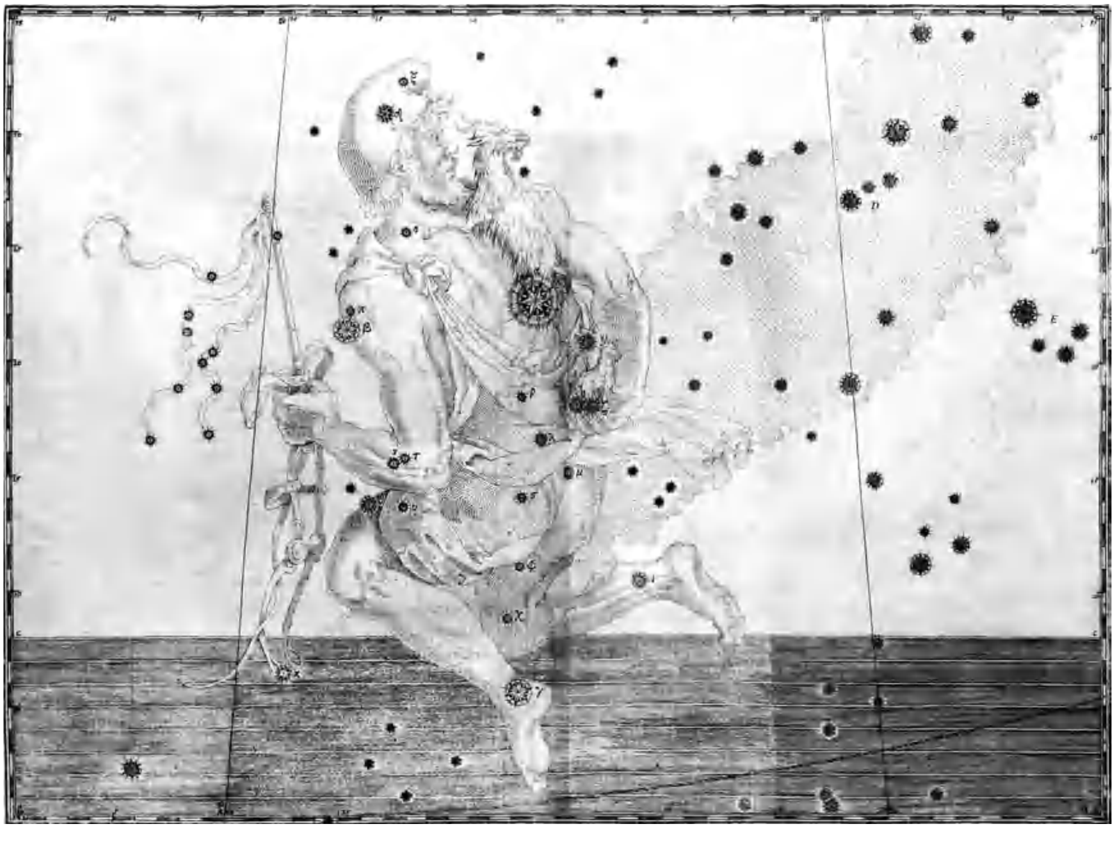


Fig. 1. Auriga, from Johann Bayer's star atlas, *Uranometria* (1603). (*Jay M. Pasachoff collection*)

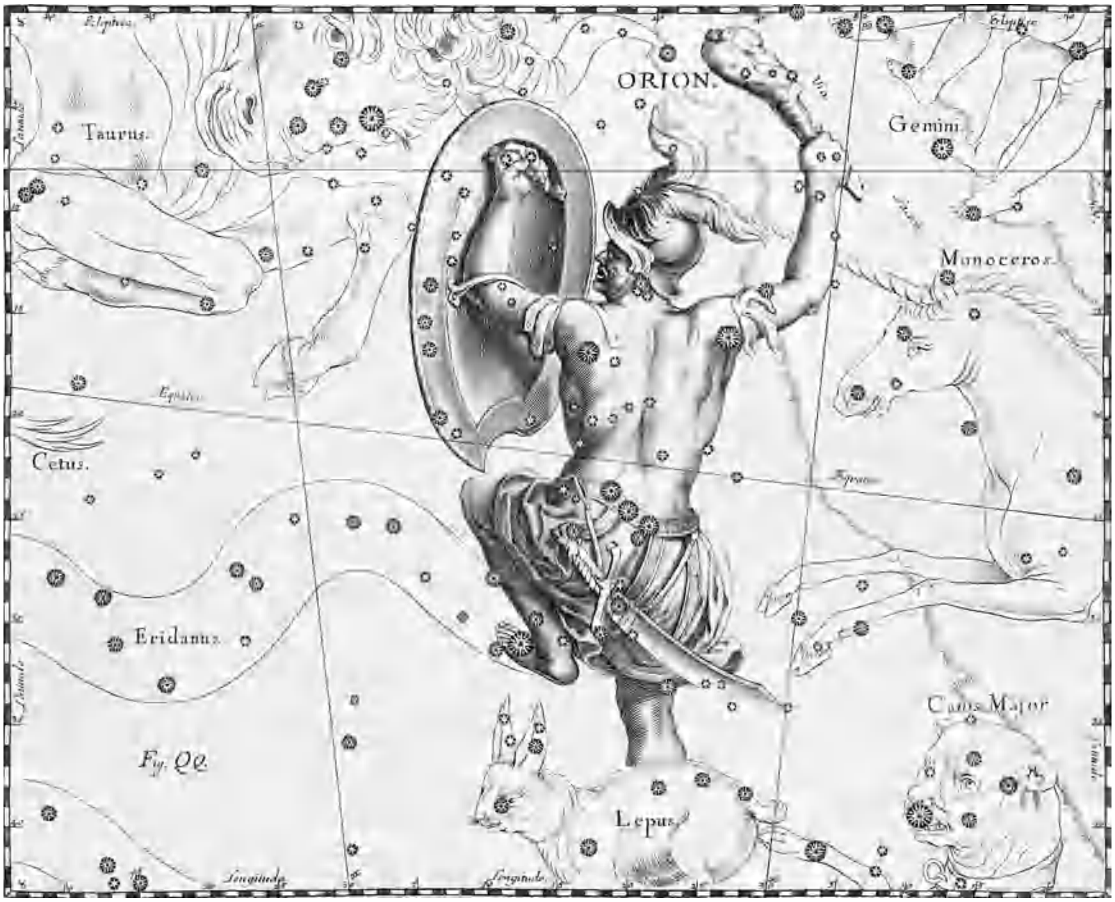


Fig. 2. Orion from Johannes Hevelius's star atlas (1687). The constellation is drawn as it would be seen from outside the celestial sphere and is therefore backward from the way it appears from Earth. (*Jay M. Pasachoff collection*)

configurations, including eclipses and transits, and its star atlas is a minor part of the book, with many constellations shown together.

John Bevis, discoverer of the Crab Nebula in 1731, produced a star atlas based on Bayer's (as is apparent from comparing the constellation figures), using also the star catalogs of Halley and Hevelius. But his apparent partner, the instrument maker John Neale, went bankrupt from the cost of the plates, and the atlas, *Uranographia Britannica*, was never published. Some two dozen sets of plates, mostly in the form of bound atlases, are known (**Fig. 3**). Some of them were sold at auction in 1785 and then to the public as *Atlas celeste* in 1786. In his plates, Bevis depicted not only the main constellations on a page but also the constellations around it in a lighter line strength. Using Hevelius's observations, he wound up with 79 constellations. His atlas had over 3550 stars.

The era of great star atlases ended with the overwhelming *Uranographia* of Johan Elert Bode, published in Berlin in 1801. The book contains 18 maps showing 99 constellations (**Fig. 4**) plus maps of the northern and southern hemispheres. This huge atlas, unwieldy to handle, contained over 17,000 stars, over 10 times more than Bayer's. It also included nonstellar objects that had been cataloged by William Herschel.

Nineteenth-century progress. The cataloging and mapping of nonstellar objects by Charles Messier in 1770 and, more extensively, by William Herschel, subsequent to his discovery of Uranus, led to a shift in emphasis. Herschel's catalog of stars led to the *General Catalogue* and eventually to the *New General Catalogue* and its two additional *Index Catalogues*, which provide many of the notations still in use today. *See* ASTRONOMICAL CATALOGS; MESSIER CATALOG.

Crude popular atlases appeared, such as that by Elijah Burritt in the 1830s in the United States. F. W. A. Argelander's *Atlas des nordlichen gestirnten Himmels*, published in Bonn in 1863, shows over 300,000 stars. The B.D. numbers from the accompanying star catalog, the *Bonner Durchmusterung*, are still in common use. A supplement of 1887 added stars below the equator; it was revised in 1951. The third edition of the northern atlas was not ready until 1954.

Carte du Ciel. In the 1880s, an international consortium set up the *Carte du Ciel*, a project to use 17 identical refracting telescopes around the world to chart the entire sky photographically. It is now held that this project, by tying up European telescopes for decades on slow-going and fruitless work, led to the assumption of the lead in astrophysics by United States astronomers as the twentieth century arrived.

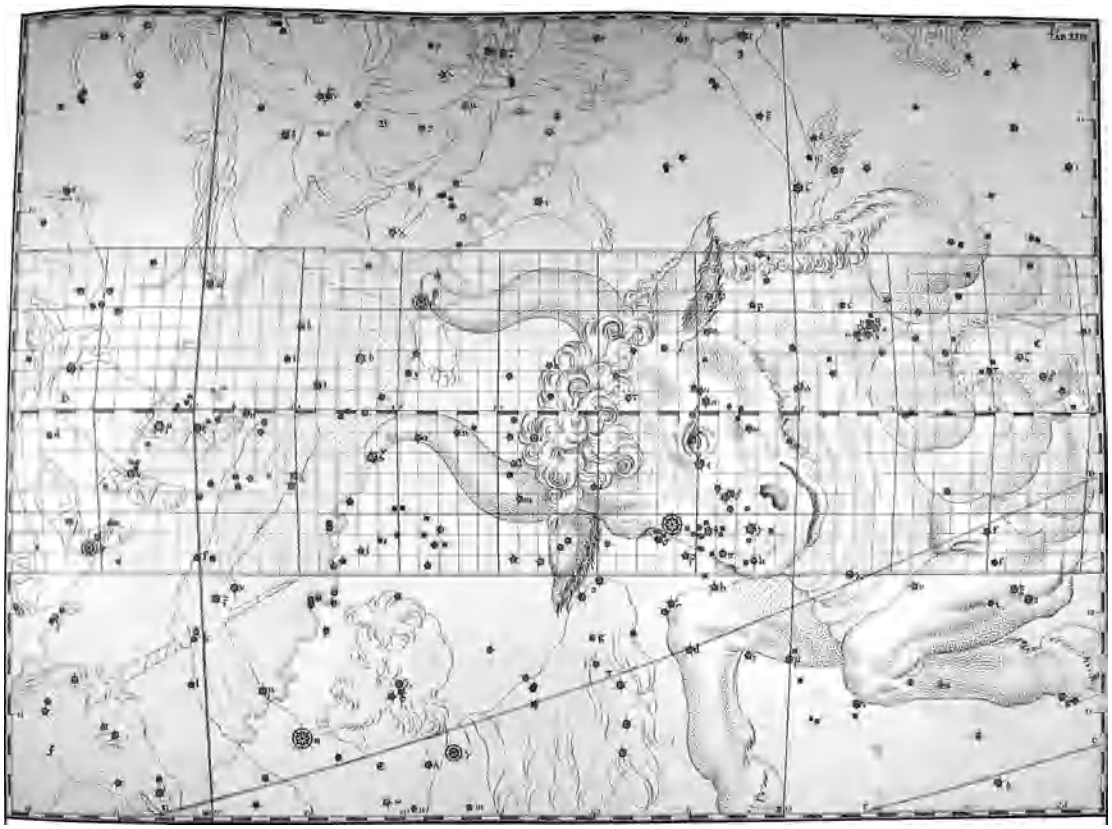


Fig. 3. Taurus and surrounding constellations from John Bevis's unpublished star atlas. (*Jay M. Pasachoff collection*)

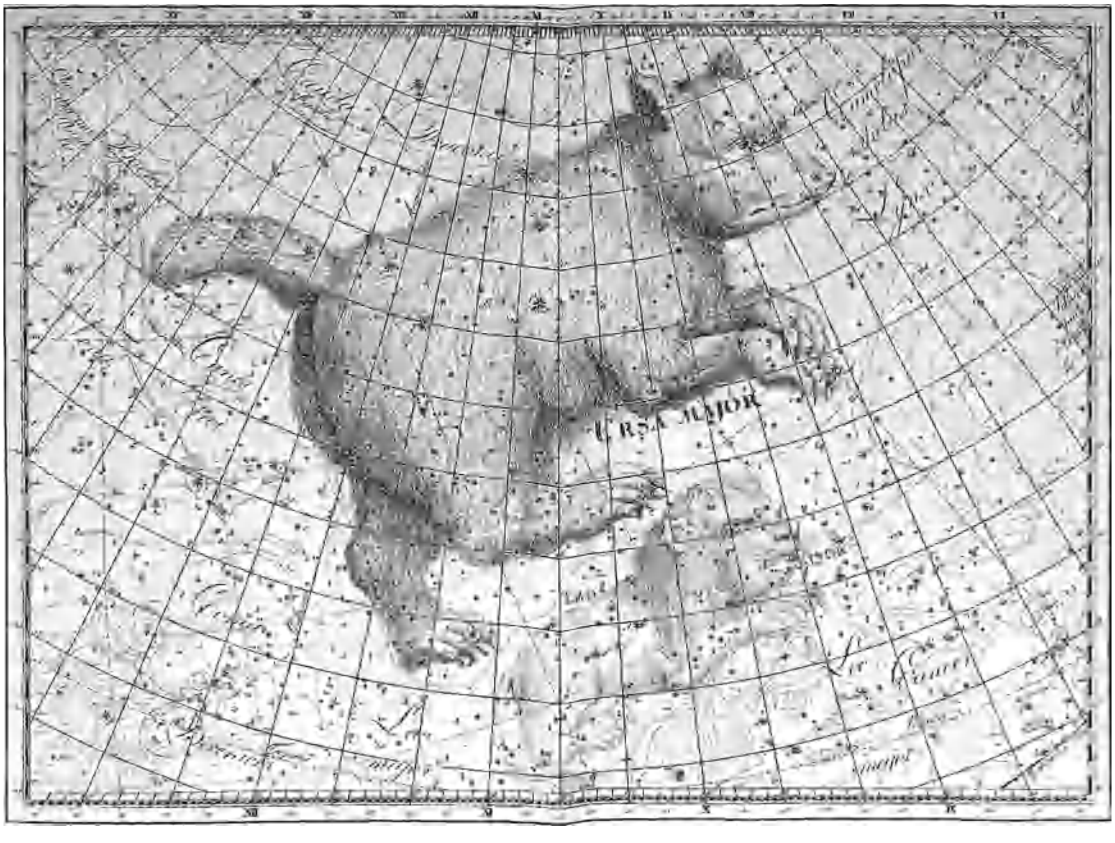


Fig. 4. Ursa Major from Johan Elert Bode's star atlas, *Uranographia* (1801). (*Jay M. Pasachoff collection*)

Photographic atlases. In 1927, E. E. Barnard produced *A Photographic Atlas of Selected Regions of the Milky Way*, edited by E. B. Frost and M. R. Calvert after Barnard's death. Volume 1 contains 51 prints, and volume 2 provides accompanying charts with identifications.

The invention in the 1930s of the Schmidt camera, capable of making wide-field photographs by using fourth-order optics, allowed the whole sky to be mapped. The National Geographic Society–Palomar Observatory Sky Survey began in 1949 to take plates using the 48-in. (1.2-m) Schmidt camera at the Palomar Observatory in California. By 1956, the whole sky above a certain southern latitude, -33°, was photographed in two colors, showing stars down to about 20th magnitude, about a million times fainter than stars that can be seen with the unaided eye. The glass plates, over 900 pairs, were printed onto transparencies and onto paper. The Ohio State University Radio Observatory provided transparent overlays showing and identifying stars, nebulae, galaxies, and other objects.

A correcting plate that was good in different colors was later installed on the Palomar Schmidt camera, which was renamed the Oschin Schmidt camera. A second-generation survey added an infrared color and used finer-grained plates, winding up with 894 sets of three plates.

Both Palomar Observatory Sky Surveys have been digitized. They are now available online, and the original Palomar Observatory Sky Survey can be purchased inexpensively on CD-ROMs.

The United Kingdom Schmidt Telescope at the Anglo-Australian Observatory in Siding Spring, Australia, and a Schmidt telescope at the European Southern Observatory's site at La Silla in Chile were used to photograph the southern sky in two visible colors and one infrared color, completing the sky coverage begun at Palomar. *See* SCHMIDT CAMERA.

Popular atlases. Many atlases showing reasonable numbers of stars for amateur observers appeared in the twentieth century. The most popular included the *Skalnate Pleso Atlas* of Antonin Becvar, compiled at Tatranska Lomnica, now in Slovakia. *Norton's Star Atlas*, with a few dozen plates showing the dome of the sky, has been through many editions.

Wil Tirion of the Netherlands is a leading celestial cartographer whose *Sky Atlas 2000.0* led atlases into the 21st century. He has since changed his method of drawing from handwork to computers and has provided atlases for many sources in a variety of projections. *Sky Atlas 2000.0*'s second edition, by Tirion and Roger W. Sinnott, was published in 1998.

In his 1963 *Peterson Field Guide to the Stars and Planets*, Donald H. Menzel used photographic star charts. Starting in 1983, a reduced-scale version of Tirion's atlas started appearing in the *Peterson Field Guide to the Stars and Planets*, then under the names of Menzel and Jay M. Pasachoff, revised in 1992 under the names of Pasachoff and Menzel and in 2000 by Pasachoff.

Contemporary amateur astronomers and others rely on *Sky Catalogue 2000.0* by Alan Hirshfeld, R. W. Sinnott, and François Ochsenbein (1991, 1985) and *NGC 2000.0*, edited by Sinnott. Sinnott and Michael A. C. Perryman did the *Millennium Star Atlas* (1997), using the accurate positions measured by the *Hipparcos* satellite of the European Space Agency. Tirion, Barry Rappaport, and George Lovi published *Uranometria 2000.0*, covering the whole sky, in 1986 with stars to magnitude 9.5 and an essay on historic atlases. Tirion, Rappaport, and Will Remaklus produced *Uranometria 2000.0 Deep Sky Atlas*, with volume 1 dealing with the northern hemisphere and volume 2 with the southern hemisphere. The volumes show over 280,000 stars on 220 star charts down to magnitude 9.75, plus over 25,000 galaxies, and use *Hipparcos* data. *See* PARALLAX (ASTRONOMY).

Galaxies. The deep-sky atlases mentioned above chart the apparent two-dimensional positions of galaxies. In his *Hubble Atlas of Galaxies*, (1961, 1984), Allan Sandage illustrated the various galaxy types but not their positions. *The Carnegie Atlas of Galaxies* (1994) by Sandage and John Bedke extended the work. Hans Vehrenberg, a distinguished amateur, covered the whole sky photographically in his *Atlas of Deep Sky Splendors*, showing the orientation of the objects.

A true mapping of space through the three-dimensional positions of galaxy clusters was published by Brent Tully of the University of Hawaii. In the early twenty-first century, two projects have mapped hundreds of thousands of galaxies and extended three-dimensional knowledge to greater distances than ever before on a uniform scale. They are the 2dF project carried out at Siding Spring, Australia, and the Sloan Digital Sky Survey. They are available online and in various technical publications. *See* GALAXY, EXTERNAL.

Solar-system atlases. Galileo made seven wash drawings of the lunar surface in 1609, and his *Sidereus Nuncius* published seven other engravings of the lunar surface in 1610. Hevelius's *Selenographia*, published in Gdansk in 1647, mapped the lunar surface in detail. A full set of lunar photographs from ground-based observations is the *Photographic Lunar Atlas* were produced by Gerard P. Kuiper in 1960. The National Geographic Society produced an airbrushed set of lunar maps; a reduced-scale version is available in Pasachoff's *Peterson Field Guide to the Stars and Planets*. In 1999, Michael Light produced *Full Moon*, in which original NASA *Apollo* photographs of the lunar surface were reprocessed. *See* MOON.

Ron Greeley and R. Batson have produced the *NASA Atlas of the Solar System* (1997) and a *Compact NASA Atlas of the Solar System* (2001). Gridded charts of various solar-system objects are produced on a regular basis by the U.S. Geological Survey's Flagstaff office of Astrogeology.

Twenty-first-century atlases. Atlases are now available online. The *Hubble Space Telescope's Guide Star Catalogue* provides positions for over 16 million objects, and can be called up on a computer in atlas form to make charts. *See* SPACE TELESCOPE, HUBBLE.

Programs are available for home computers that allow users to chart any given area of the sky on any scale, including deep-sky objects and solar-system objects. Capabilities continue to expand.

Jay M. Pasachoff

Bibliography. R. Greeley and R. Batson, *Compact NASA Atlas of the Solar System*, Cambridge University Press, 2001; M. Light, *Full Moon*, Knopf, 1999; J. M. Pasachoff, *Peterson Field Guide to the Stars and Planets*, 4th ed., Houghton Mifflin, 2000; A. Sandage and J. Bedke, *The Carnegie Atlas of Galaxies*, 2 vols., Carnegie Institution of Washington, 1994; R. W. Sinnott and Michael A. C. Perryman, *Millennium Star Atlas*, Sky Publishing Corp., 1997; W. Tirion, B. Rappaport, and W. Remaklus, *Uranometria 2000.0 Deep Sky Atlas*, 2 vols., Willmann-Bell, 2001; W. Tirion and R. W. Sinnott, *Sky Atlas 2000.0*, 2d ed., Sky Publishing Corp., 1998.

Astronomical catalogs

Lists or enumerations of astronomical data relevant to astronomy, navigation, geodesy, and space science applications. Astronomical catalogs vary a great deal in form and content depending upon the type of the data and the objects to which they are referring. The essential data may be positions and motions of the celestial objects, magnitudes, spectra, radial velocities of stars, or energy fluxes. The celestial objects may be stars, galaxies and other galactic and extragalactic objects, or solar system bodies, which are generally ordered by increasing right ascensions and identified by a catalog number. Catalogs are either derived directly from observations or are compiled from different sources.

Access to the catalogs. Astronomical data centers have been established by international exchange agreements to develop and provide, in machine-readable form and on-line, catalogs that were previously published in book form. The Astronomical Data Center (ADC) at the National Aeronautics and Space Administration's Goddard Space Flight Center discontinued its services in 2002. The Centre de Données Astronomiques de Strasbourg (CDS), in France, currently provides the most complete library of published astronomical catalogs available on line with standard documentation and format and a relational database management system. This service is also available in Canada, the United States, Japan, India, the United Kingdom, and China. Other data centers provide access to data and catalogs related to specific projects or astronomical objects. Such data centers are now developed in the general frame of the Virtual Observatory, by which astronomers will have World Wide Web access to any kind of data in common standardized formats. *See* WORLD WIDE WEB.

Reference systems and fundamental catalogs. The positions and motions listed for celestial objects should refer to a primary reference system which is realized by a reference frame. A considerable change was the transition from a reference system realized by stellar positions and proper motions to the International Celestial Reference System (ICRS), based on observed extragalactic radio sources. The ICRS and the International Celestial Reference Frame (ICRF) that realizes the ICRS have been adopted by the International Astronomical Union starting from 1998.

The ICRS is centered at the solar system barycenter and is nonrotating with respect to the ensemble of distant extragalactic objects. It has no intrinsic orientation but was aligned close to the mean equator and dynamical equinox of J2000.0 (the date January 1, 2000, at 12 hours Terrestrial Time) for continuity with previous fundamental reference systems. *See* ASTRONOMICAL COORDINATE SYSTEMS; DYNAMICAL TIME; EQUATOR; EQUINOX.

The ICRF is the catalog of adopted positions (right ascensions and declinations) of 608 extragalactic radio sources, mostly quasars, observed with very long baseline interferometry (VLBI), which is currently the most accurate realization of the ICRS. Of these objects, 212 are defining sources, with position uncertainties of the order of 0.5 milliarcsecond (mas), that establish the orientation of the ICRS axes. Extensions of the ICRF were released in 1999 and 2004 with 59 and 109 additional sources, respectively. *See* ASTRONOMICAL COORDINATE SYSTEMS; QUASAR; RADIO TELESCOPE.

The ICRS is realized at optical wavelengths by a subset of about 100,000 stars in the *Hipparcos Catalogue*. This is referred to as the *Hipparcos* Celestial Reference Frame (HCRF) which realizes the *Hipparcos Catalog* Reference System (HCRS). *See* ASTROMETRY; CELESTIAL REFERENCE SYSTEM.

Pre-Hipparcos catalogs of stellar positions. The *Fifth Fundamental Catalogue* (*FK5*), published in 1988 by the Astronomisches Rechen-Institut in Heidelberg, Germany, was the International Astronomical Union (IAU) realization of the celestial reference system up to 1998. It contains the positions at J2000.0 of 1535 stars, mostly brighter than ninth magnitude, and an additional 3117 stars, down to a magnitude of about 9.5. *See* MAGNITUDE (ASTRONOMY).

Moderately bright stars (seventh to ninth magnitude), selected on the basis of one star per square degree of the sky, have been related to the *FK5* by meridian circle observations in order to form a system of sufficient density to serve as position references for photographic observations. Typical of catalogs of such reference stars is the *International Reference System* (*IRS*) catalog (1991), which was the world standard reference system in the pre-*Hipparcos* era. *See* ASTRONOMICAL TRANSIT INSTRUMENT.

A photographic survey of the northern celestial hemisphere resulted in the *AGK3* (1975), the third in a series of catalogs published by the German Astronomical Society. A revised version of this catalog has been published by the Astronomisches Rechen-Institut as the *Position and Proper Motion* (*PPM*) *Catalog* (1991) of 181,731 stars brighter than the 9th magnitude, north of -2.5° declination. A similar photographic catalog for the southern celestial hemisphere is the *Second Cape Photographic Catalog* (*CPC2*) [1993] of 250,000 stars to a limiting magnitude of 10.5.

An important photographic catalog covering the entire sky to a limiting magnitude of 13 was the result of an international undertaking involving 19 observatories, with each assigned zones of declination and observing with nearly identical telescopes. The catalog, known as the *Carte du Ciel* (*CdC*) or *Astrographic Catalogue* (*AC*), finally completed for all zones in 1964, provides star positions in the form of rectangular coordinates, as measured on the plates. A reduction of the *AC* was made at U.S. Naval Observatory (USNO) in Washington, D.C. using its *Astrographic Catalogue Reference Stars* (*ACRS*) with positions and proper motions of about 320,000 stars.

The *Smithsonian Astrophysical Observatory Star Catalogue* (*SAOC*) was compiled from selected visual and photographic catalogs, initially for the determination of artificial satellite positions. It contains positions and proper motions of 259,000 stars over the entire sky. It was published in book form (1966) and updated in machine-readable form with a reduction to the *FK5* system.

The Guide Star Catalog (*GSC*), constructed for the pointing of the *Hubble Space Telescope*, provides positions and magnitudes for nearly 19 million objects in the 6th to 15th magnitude range, of which more than 15 million are classified as stars. The *USNO-A1.0 Catalog* contains *GSC*-based positions of 488,006,860 sources which were detected by the U.S. Naval Observatory's Flagstaff Station Precision Measuring Machine (PMM) in the framework of several photographic sky surveys. *See* HUBBLE SPACE TELESCOPE.

Hipparcos and Tycho Catalogues. The *Hipparcos* and *Tycho Catalogues* (1997) have both resulted from instruments on board the European Space Agency's *Hipparcos* satellite, operational from late 1989 to 1993. They include a large quantity of astrometric and photometric data, as well as annexes featuring variability and double- and multiple-star data. *See* PARALLAX (ASTRONOMY).

The *Hipparcos Catalogue* of 118,218 stars is complete to visual magnitude $V = 7.3$ and includes stars as faint as visual magnitude 12. It was linked to the ICRF through radio stars. Accuracies are 1 to 3 mas for positions at epoch 1991.25 and 1 to 2 mas/year for proper motions. *Hipparcos* also contains broadband visual photometric data including variability information in the 1–2 millimagnitude range for the brighter stars.

The *Tycho Catalogue* (also named *Tycho-1 Catalog*) of about 1 million stars is an extension of the *Hipparcos Catalogue* that is more complete at magnitudes 8–11 but less accurate by a factor of 30. It contains two-color photometry for most of its stars.

Post-Hipparcos star catalogs. New versions of star catalogs have been formed by referring them to the HCRS, such as the *AC 2000* and the *USNO-A2.0* catalog. The *USNO B1.0 Catalog* (2003) expands this series in providing 1,042,618,261 stars or galaxies and in the inclusion of proper motions. It is expected to be complete down to magnitude 21 with accuracies of 200 mas for positions at J2000, 0.3 mag for brightnesses in up to five colors, and 85% for distinguishing stars from nonstellar objects.

Several catalogs have aimed to combine Tycho observations with century-old positions, such as the *Tycho Reference Catalogue* (*TRC*), the *ACT Reference Catalog*, and the *All-Sky Compiled Catalogue* (*ASCC-2.5*). They have been superseded by the *Tycho-2 Catalog* of 2,539,913 brightest stars, which combines a sophisticated reduction of the Tycho stars and 144 ground-based star catalogs. Accuracies are from 10 to 100 mas for positions and from 1 to 3 mas/year for proper motions.

The *USNO CCD Astrographic Catalog* (*UCAC*) has been undertaken for global sky coverage down to 16th magnitude at red wavelengths. Preliminary releases of this catalog are *UCAC1* in the southern hemisphere and *UCAC2* covering 86% of the sky. The latter is a compiled catalog of 48 million stars which are primarily in the 8.0–16.0 magnitude range, including all catalogs used for *Tycho-2*, plus the recent epoch ground-based observations of the *UCAC* project. Accuracies are 20 to 70 mas for positions and 1 to 7 mas/year for proper motions.

A revision of the *Guide Star Catalog, GSC II*, is intended to provide positions, proper motions, magnitudes, and classifications for all objects derived from the Digitized Sky Survey. A preliminary version (*GSC2.2*) includes 435,457,355 objects with magnitude limits 18.5 and 19.5 in the two relevant photographic wavelengths, respectively.

Catalogs of nonstellar objects. J. L. E. Dreyer's *New General Catalogue of Nebulae and Starclusters* (1888) contains 7840 objects originally classified as nonstellar, and was supplemented by two Index Catalogs in with an additional 5386 objects. The NGC and IC numbers assigned in these catalogs remain the most commonly used designations. A *Master List of Nonstellar Astronomical Objects* (1980) provides approximately 185,000 listings from 270 catalogs, with multiple listings of objects appearing in several catalogs. *See* GALAXY, EXTERNAL; NEBULA; STAR CLUSTERS.

With the expansion of astronomical research beyond the visible part of the electromagnetic spectrum to the ultraviolet, x-ray, and gamma-ray regions at short wavelengths and to the infrared and radio regions at long wavelengths, and with modern very large sky surveys, the number of nonstellar objects has increased dramatically. An *Index of Extragalactic Radio Source Catalogs* (1977) provides precise references to the large number of individual catalogs. In this category, the third version of the *Reference Catalog of Bright Galaxies* (1991) contains 23,022 objects, the *Cambridge Observatory Discovery Catalogs of Quasars*, named *3C* and *4C* (1959 and 1965), are reference catalogs; and the *Catalog of Quasars and Active Galactic Nuclei*, (11th edition, 2003) contains 48,921 entries with position, redshift, photometry, and flux densities. *See* RADIO ASTRONOMY; SLOAN DIGITAL SKY SURVEY.

Catalogs and databases of several million extragalactic objects with various facilities (including cross-identification) are available at the NASA/IPAC Extragalactic Database (NED) and the HYPERLEDA database maintained at Lyon Observatory (France), with mirror sites in several countries.

The number of known infrared sources has rapidly increased, especially since the launch in 1982 of the *Infrared Astronomical Observatory Satellite* (*IRAS*), whose survey has provided the *IRAS Faint Source Catalog* and the *IRAS Catalogue of Point Sources*, which contains some 250,000 infrared point sources. Astrometric positions of stellar sources from the *IRAS Point Source Catalog* have been published by the U.S. Naval Observatory in the *Catalog of Positions of Infrared Stellar Sources* (2001), containing 37,700 entries with *Hipparcos* and *Tycho-2* astrometry and photometry.

The Two Micron All Sky Survey (2MASS) project has imaged the entire sky in three near-infrared bands, producing a point-source catalog (2000, 2003) of 470 million objects, mostly stars, and an extended-source catalog of 1.6 million objects, with accuracy in position of about 70 mas. The Deep Astronomical Survey of the Southern Sky (DENIS), in two near-infrared bands and one optical, was conducted by a European consortium and produced (2003) a catalog of 195,204,157 point sources.

Ultraviolet astronomical observations of stellar and nonstellar objects have been obtained from rocket vehicles, spacecraft, and satellites. Catalogs of these photometric and spectroscopic observations are available at CDS. *See* ULTRAVIOLET ASTRONOMY.

Early discoveries of x-ray sources were from rocket observations, but a rapid increase in known sources occurred after the launch of the *UHURU* satellite in 1970 and other satellites which followed. Catalogs and databases of these observations are available from NASA High Energy Astrophysics Science Archive Center (HEASARC) or CDS. The *ROSAT All-Sky Bright Source* and *ROSAT All-Sky Faint Source Catalogs* are among the most frequently used catalogs in this category. *See* CHANDRA X-RAY OBSERVATORY.

Gamma rays have been detected from quasars, galaxies, and highly evolved stars. Catalogs of observations from spacecraft and satellites, such as the *International Gamma-Ray Astrophysics Laboratory* (*INTEGRAL*), are available from HEASARC. *See* GAMMA-RAY ASTRONOMY.

Catalogs of astrophysical data. Among the numerous catalogs in this category is the monumental *Henry Draper Catalog* (*HD*) of spectral classification, which with its extension includes data for 275,000 stars, with an updated machine-readable version (1989). The introduction of luminosity classes in spectral classification, such as the prevailing Morgan-Keenan-Kellman (MK) system (1943), led to classification of some 33,000 stars in this system, compiled into a catalog by the Mount Stromlo Observatory in Australia (1981). This catalog also contains the magnitudes and colors of stars on the U, B, V photometric system. The Asiago Database on Photometric Systems (ADPS) project (2000) aims to list and investigate all existing photometric systems, and provides a link to a regularly updated *General Catalogue of Photometric Data* with an associated master index. *See* ASTRONOMICAL SPECTROSCOPY; SPECTRAL TYPE.

Catalogs of stellar radial velocities are listed in two bibliographies produced by the Kitt Peak National Observatory, covering data upto 1970, and the *General Catalog of Mean Radial Velocities* (2000) has been compiled with data from 1970 to 1990.

Catalogs of special stellar objects. There are many catalogs in this category. Among the most frequently used are the *Combined General Catalog of Variable Stars* (*GCVS*); the *Washington Double Star Catalog* (*WDS*) of the U.S. Naval observatory (2000), which is the official database for about 100,000 double and multiple stars for the IAU; the *Bright Star Catalog* of all stars brighter than magnitude 6.5; and the *Yale General Catalog of Trigonometric Stellar Parallaxes* (4th edition, 1994), containing more than 15,000 entries with a precision of 4 mas. *See* BINARY STAR; STAR; VARIABLE STAR. Nicole Capitaine

Bibliography. K. Aa. Strand, *Basic Astronomical Data*, 1963; H. Eichhorn, *Astronomy of Star Positions*, Ungar, 1974; K. J. Johnston et al. (eds.), *Towards Models and Constants for Sub-Microarcsecond Astrometry*, International Astronomical Union Colloquium 180, 2000 ; J. Kovalevsky and P. K. Seidelmann, *Fundamentals of Astrometry*, Cambridge University Press, 2004; H. G. Walter and O. J. Sovers, *Astrometry of Fundamental Catalogs: The Evolution from Optical to Radio Reference Frames*, Springer, 2000.

Astronomical coordinate systems

Schemes for locating astronomical objects in space. To an observer on the Earth's surface, the stars of the night sky appear to be placed upon a spherical shell of infinite radius with the observer at the center. Celestial objects appear to move or to maintain constant locations with respect to the stars, and at any given time their position on this imaginary sphere, called the celestial sphere, can be specified by two angles, called celestial coordinates, whose values depend upon what coordinate system is used. *See* CELESTIAL SPHERE.

Each coordinate system is defined by a fundamental plane and a principal axis. For example, on the Earth's surface, longitude and latitude coordinates are used to determine positions. In this system, which is analogous to astronomical coordinate systems, the fundamental plane is that of the Earth's Equator, and the principal axis is defined by a line running from the Earth's center to a point on the Equator at the longitude of Greenwich, England. Longitude is measured east or west of Greenwich. Angular positions along the fundamental plane, like longitude, can be expressed in terms of either degrees or hours of time. One hour equals 15°. Angular measurements north or south of the fundamental plane, like

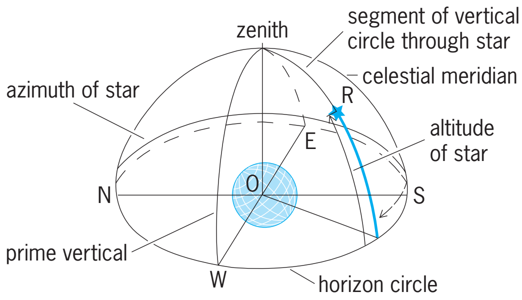


Fig. 1. Horizon system of astronomical coordinates.

latitude, are measured from 0 to 90° at the poles. In astronomical coordinate systems, the degrees north are denoted with a plus sign (+), and degrees south are with a minus sign (−). *See* EQUATOR; LATITUDE AND LONGITUDE.

Horizon system. The boundary between the hemisphere of the sky that is visible and the hemisphere which is hidden from view by the Earth is called the horizon (**Fig. 1**). The observer is located at the center of the system (*O*), the pole directly overhead is termed the zenith, and the opposite pole, the nadir. These pole directions are aligned with a plumb line, which is determined by the observer's local gravity. The fundamental horizon plane is 90° from the poles, and for astronomical applications the principal axis is most often taken to pass through the north point. Great circles that pass through the zenith and nadir are termed vertical circles; the one passing through the east and west points is termed the prime vertical, and that passing through the north and south points is called the celestial meridian. *See* HORIZON; ZENITH.

The longitudinal coordinate of a celestial object is termed its azimuth and is most often measured eastward from the north point to the object's vertical circle; and the latitudinal coordinate, termed its altitude, is measured along the object's vertical circle, north or south from the horizon plane to the object. For example, the position of the star *R* in Fig. 1 would have an approximate azimuth and altitude of 210° and $+60^\circ$. The object's zenith distance, measured along the object's vertical circle, is the angle measured from the zenith point to the object.

Because the horizon coordinate system is fixed with respect to the observer on the Earth's surface, it must rotate with the Earth. Hence the azimuth and altitude of a celestial object are constantly changing with time. For many astronomical applications, such as the listing of stellar and planetary positions, a nonrotating, coordinate system is required. The most common of these nonrotating, or inertial, coordinate systems is the equatorial system.

Equatorial system. The fundamental plane of the equatorial coordinate system can be visualized by imagining that the Earth's equatorial plane is extended to intersect the celestial sphere. An alternate fundamental plane, the ecliptic plane, is the extension of the Earth's mean orbital plane onto the celestial sphere (**Fig. 2**). These planes intersect at two points, called equinoxes, with the angle between them ϵ being termed the obliquity of the ecliptic. This angle is about 23.4°. *See* ECLIPTIC; EQUINOX.

Due to the Earth's motion about the Sun, observers on Earth see an apparent motion of the Sun along the ecliptic plane. The point where the Sun's annual apparent motion takes it northward across the equatorial plane is called the vernal equinox Υ, and the line between the Earth's center and this point defines the principal axis for both the equatorial and ecliptic coordinate systems. The apparent passage of the Sun through the vernal equinox, on about March 21, marks the beginning of spring in the Northern Hemisphere. Because of disturbing effects of the Sun and Moon on the Earth's figure, the Earth's rotation axis precesses, causing the celestial pole to describe an approximate circular motion about the ecliptic pole once every 26,000 years. This causes the location of the vernal equinox to drift westward along the ecliptic about 50 arc-seconds each year. Hence for an inertial coordinate system, where the principal axis is not moving, an epoch must be specified at which time the coordinate system is held fixed. In practice, the beginning of the year 2000 is most often used as an epoch. *See* EARTH ROTATION AND ORBITAL MOTION; PRECESSION OF EQUINOXES.

The north and south celestial poles represent the extension of the Earth's North and South poles onto the celestial sphere. For a celestial object (for example, object *X* in Fig. 2), the longitudinal coordinate is termed the right ascension α and is measured eastward along the celestial equator from the vernal equinox Υ to the great circle passing through the object and the north and south celestial poles. The latitudinal coordinate, called the declination δ, is then measured along the object's great circle, north or south from the Equator to the object.

Ecliptic system. The ecliptic coordinate system is often used when representing the orbital motions of the planets, asteroids, and comets. The fundamental plane is that of the ecliptic, and as in the equatorial system, the principal axis is the line extending from the Earth's center to the vernal equinox. The position of a celestial object is defined, in the ecliptic system, by the ecliptic longitude and latitude. The longitude

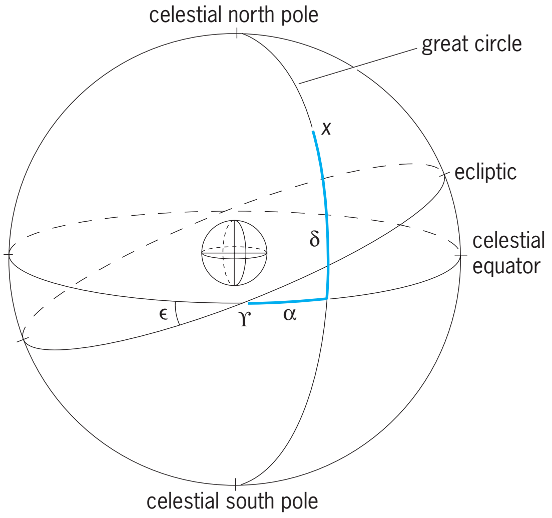


Fig. 2. Equatorial system of astronomical coordinates.

is measured eastward along the ecliptic plane from the vernal equinox to the great circle that passes through the object as well as the north and south ecliptic poles. The ecliptic latitude is then measured along this great circle either north or south from the ecliptic plane to the object.

Galactic system. Astronomers working with stars and other objects within the Milky Way Galaxy often find it convenient to use the galactic disk as the fundamental plane of their coordinate system, and the line extending from the galactic center to the Sun's location as the principal axis. With the Sun as the origin, galactic longitude is measured from the principal axis along the galactic equator eastward to the great circle that passes through the object and the north and south galactic poles. Galactic latitude is measured along the object's great circle, either north or south from the galactic plane to the object. In equatorial coordinates, the position of the north galactic pole, at the 2000 epoch, is approximately $\alpha = 12.9$ h (192.9°) and $\delta = +27.1^\circ$. *See* MILKY WAY GALAXY.

Donald K. Yeomans

Bibliography. S. Bell and C. Hohenkerk, Correspondence: Controversy in *The Astronomical Almanac 2006, The Observatory*, 125:271–272, 2005; P. Duffett-Smith, *Practical Astronomy with Your Calculator*, 3d ed., Cambridge University Press, 1988; C. A. Murray, *Vectorial Astronomy*, Adam Hilger, 1983; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, University Science Books, 1992; L. G. Taff, *Computational Spherical Astronomy*, 1981, rev. ed., Krieger, 1991.

Astronomical imaging

The production of a permanent two-dimensional record of a scene of interest to astronomy. In direct imaging, the scene is a section of the sky optically imaged with a telescope. In other types of imaging, the scene is the output of an instrument, often a spectroscope. Although astronomical images can be produced using light from any part of the electromagnetic spectrum, this article mainly concerns images made in the optical and near-infrared regions (wavelengths of 320–1000 nanometers). The overriding concerns in astronomical imaging are resolution, wavelength sensitivity, and above all, detective quantum efficiency.

Astronomers can use three techniques to construct an image. The first, the subject of this article, employs a two-dimensional, multielement panoramic detector such as a photographic plate or the retina of the human eye to detect the elements of the scene. The other two techniques build an image over time, either by scanning the scene with a single-element detector or by interferometry (interpreting the changing interference pattern of light waves detected by multiple telescopes as the spacing of the telescopes changes). *See* INTERFEROMETRY.

Most panoramic detectors belong to a class called photon detectors, which make use of an individual photon's ability to alter the quantum-mechanical state of one or more electrons in the detector. The other detector classes, thermal detectors and wave detectors, are difficult to fabricate as panoramic devices.

Arrays. At optical wavelengths as well as in the very near infrared, astronomers usually select the charge-coupled device (CCD) as the best available panoramic detector. CCDs are arrays—they consist of a grid of spatially discrete photon detectors, rather than the almost continuous distribution of sensitive emulsion found on a photographic plate. *See* CHARGE-COUPLED DEVICE.

At other wavelengths, astronomers also use arrays of photon detectors. At shorter wavelengths (x-rays and ultraviolet light), these are sometimes modified CCDs, but at wavelengths longer than 1.0 micrometers (infrared to submillimeter) the arrays are other devices. The following discussion concentrates on CCDs, but most of the general principles discussed apply to imaging with any detector array. *See* INFRARED ASTRONOMY.

MOS capacitor. Each detecting element in a CCD array is a metal-oxide-semiconductor (MOS) capacitor, a good example of a photon detector. The basic structure of this very simple solid-state device is a three-layer sandwich (**Fig. 1**). As illustrated, the left-hand layer is a block of *p*-type silicon semiconductor whose left-hand face is connected to electrical ground. The middle of the sandwich is a thin layer of insulator (usually silicon dioxide). The right-hand layer is a thin coating of metal (usually highly doped polysilicon), which is held at a positive voltage. *See* SEMICONDUCTOR; TRANSISTOR.

The conductivity of the *p*-type silicon is due to positively charged holes (the majority carriers), which are repelled toward ground by the positive voltage of the metal layer, producing a depletion region devoid of charge carriers. Illuminated through the transparent electrode, the device is sensitive to light because a photon absorbed in the depletion region will ionize a silicon atom, creating an electron-hole pair. The electron will move to the right, toward the positive electrode (the potential well), but cannot cross the nonconducting oxide layer. For every photon absorbed by the silicon, another electron is stored at the semiconductor-oxide boundary. At the end of an exposure, the charge on the capacitor is read out and converted to a digital signal.

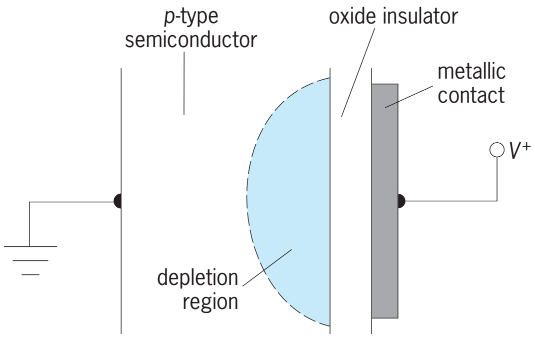


Fig. 1. Cross-sectional view of the physical structure of an MOS capacitor. Positive voltage (usually a few volts) applied to the metal layer creates a depletion region in the semiconductor.

Characterization of Array Detectors

From roughly 1880 to 1980, photography served as the primary method of astronomical imaging. What motivated professional astronomers to move to CCDs and other arrays? More generally, what makes one detector better than another? All the following characteristics are important.

Efficiency. Astronomers construct telescopes to gather as many photons as possible, and it seems perverse if a detector does not use a large fraction of these expensive photons to produce its signal. The quantum efficiency (QE) is a common measure of detector efficiency. It is usually defined as the fraction of photons incident on the detector that actually contribute to the signal. In a perfect detector, every incident photon would be absorbed in a fashion that contributes equally to the signal, and the detector would have a QE of 1.00.

Noise. Although efficiency in a detector is important, the real value of any measurement depends crucially on its uncertainty. The uncertainty in the output signal produced by a detector is often called the noise, and the signal-to-noise ratio (SNR) is a widely used indicator of the quality of a measurement. However, a perfect detector does not produce a signal with zero noise and infinite signal-to-noise ratio. *See* ELECTRICAL NOISE; SIGNAL-TO-NOISE RATIO.

There is an uncertainly inherent in measuring the strength of any incident stream of photons. For a photon-counting device, this uncertainty arises from the Poisson statistics of random photon arrivals, and is given by Eq. (1), where N is the number of photons

$$\sigma = \sqrt{N} \tag{1}$$

actually counted and σ is the uncertainty in N. A perfect detector, with QE = 1, counts all incident photons and will therefore produce a signal-to-noise ratio given by Eq. (4).

$$\mathrm{SNR}_{\mathrm{perfect}} = \frac{N_{\mathrm{out}}}{\sigma_{\mathrm{out}}} = \frac{N_{\mathrm{in}}}{\sigma_{\mathrm{in}}} = \sqrt{N_{\mathrm{out}}} = \sqrt{N_{\mathrm{in}}} \tag{2}$$

Real detectors will have a smaller signal-to-noise ratio, because they either count fewer photons or add output noise. The detective quantum efficiency (DQE) describes this departure of a real detector from perfection and, better than the QE alone, measures the quality of a detector in a particular situation. For a perfect detector, DQE = QE = 1. For any detector, DQE $\leq$ QE.

In detecting faint sources, a typical photographic emulsion, hypersensitized to increase its QE, might have QE = 0.04 and, because of noise due to emulsion granularity, DQE = 0.005. Modern astronomical CCDs degrade the information in the input hardly at all, and their DQEs are close to their QEs—in the range 0.20–0.95. Thus, both efficiency and noise considerations strongly select the CCD over photography. *See* PHOTOGRAPHY.

Spectral response and discrimination. The QE of a detector is generally a function of the wavelength of the input photons. Some excellent detectors are useless at some wavelengths. For example, silicon devices, including CCDs, cannot respond to photons with wavelengths greater than 1.1 μm. One can imagine an ideal detector that measures both the intensity and the wavelength of each incoming photon. Designs of such detectors have been proposed and might well replace CCDs if proven equivalent in DQE and other properties.

In professional astronomy, color images (which illustrate rough spectral distributions) are almost always assembled indirectly by combining three separate monochrome frames with different spectral responses.

Linearity. In an ideal detector, the output signal is directly proportional to the input illumination. Departures from strict linearity are common in modern infrared-sensitive devices and are a serious problem in photography. CCDs are highly linear devices except at very high levels of input, where individual array elements will saturate. Further increases in input will not move the output signal above the saturation level. *See* SATURATION.

Response time. The minimum time interval over which the detector can measure changes in illumination is an important parameter. Readout procedures for large astronomical CCDs, for example, can limit their response time to several seconds or more, and measuring rapid brightness changes is an application where astronomers replace CCDs with specialized devices.

Size, pixel number, and array resolution. Clearly, the physical size of each detector of an array will determine how closely spaced its elements, or pixels (for picture elements) can be. Large arrays are more easily manufactured for some types of detectors (such as MOS capacitors) than for others (such as bolometers and wave detectors). There is an obvious advantage in field-of-view for detectors with a large number of pixels. For any array recording telescopic images, it is best if the pixel size and telescope resolution match. Astronomical imaging is improved if the resolving power of the telescope is improved, either by placing it in space or by employing adaptive optics. *See* ADAPTIVE OPTICS.

At the beginning of the CCD era, photographic plates had a clear advantage in pixel number: For a very moderate cost, a photographic plate had a very large area (tens of centimeters on a side), and thus, in effect, contained up to 10^9 pixels (1000 megapixels). Mosaics of CCD arrays, although quite expensive, are beginning to approach the pixel count of medium-sized photographic plates.

Data handling. An important characteristic of modern solid-state detectors is the ease with which their output can be stored as a digital image accessible by a computer. This is critically important in the usual situation in which an astronomer wishes to analyze the image quantitatively. It is nearly essential for images transmitted by space-based telescopes.

General Operation and Properties of CCDs

Researchers at Bell Laboratories produced the first CCDs in the early 1970s as a kind of computer memory. A few astronomers quickly recognized the

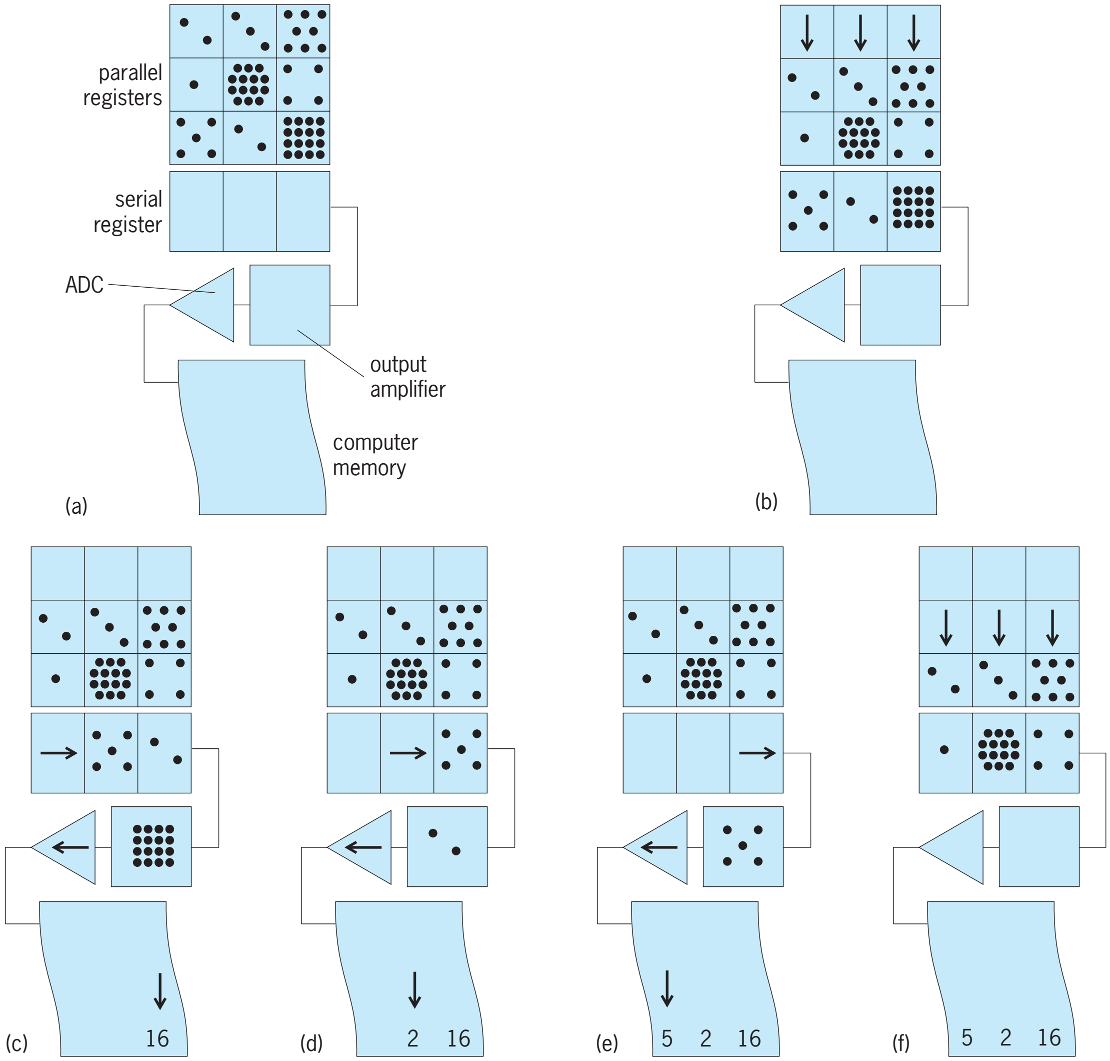


Fig. 2. **CCD components and readout. (*a*) Accumulated photoelectrons in a 3 × 3 array of capacitors. Reading the array: (*b*) The bottom row shifts into the serial register; all remaining rows shift down by one pixel in the parallel register. (*c–e*) Serial register reads one column at a time. (*f*) Step *b* is repeated: all the rows shift down one pixel.**

device's potential as a panoramic light detector, and by 1976 had recorded the first CCD images of celestial objects. The CCD array has become a standard component in military, industrial, and consumer applications that include scanners, copiers, and still and video cameras. Such widespread applicability has supported research and development costs. The resulting rapid evolution of the scientific CCD, initially spurred by the design for detectors used in the *Galileo* mission to Jupiter, has revolutionized the practice of optical observational astronomy.

To understand the basic ideas behind the array operation, one may imagine a matrix of MOS capacitors placed behind a shutter in the focal plane of a telescope. Photons strike the pixels of the array, and each accumulates photoelectrons at a rate that is proportional to the intensity of light incident on the pixel. At the end of the exposure, the array contains an electronic record of the image.

The defining attribute of the CCD is the method whereby the electron pattern stored in the array is converted into a useful form (numbers in a computer). One may consider, for example, an array of detector pixels three columns wide and three rows tall (**Fig.** 2*a*). Each column of pixels is called a parallel register. There is one additional row of capacitors (not exposed to light), called the serial register.

Reading the array first requires a shift of pixel content down the columns of the parallel registers by one pixel (Fig. 2*b*), that is, "coupling" the charge, providing the name "charge-coupled device." Electrons originally stored in row 3, for example, shift to the serial register.

A second operation now reads the newly filled serial register by shifting its contents to the right by one pixel (Fig. 2*c*). The electrons in the rightmost pixel shift into a new structure: an output amplifier that ultimately converts the charge to a voltage, and an analog-to-digital converter (ADC) that converts the voltage into the number stored in computer memory. The CCD continues this shift-and-read of the serial register, one pixel at a time (Fig. 2*d* and 2*e*), until the entire serial register has been read. The

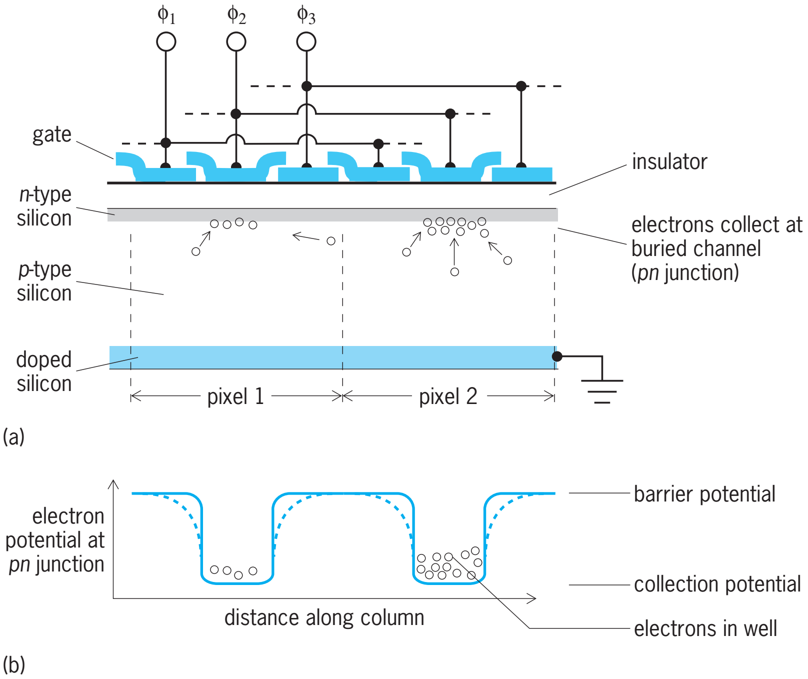


Fig. 3. Gate structure in a three-phase buried-channel CCD. (*a*) Two pixels, in cross section. (*b*) Electron potential *pn* junction. Collection and barrier potentials on the gates isolate the pixels from each other during an exposure. Overlapping gates produce a gradient in the barrier region (broken curve in lower figure) that enhances collection.

device then repeats parallel shifts (Fig. 2*f*) [followed by serial shifts and reads] until the entire array has been read to memory. *See* ANALOG-TO-DIGITAL CONVERTER.

To persuade the electrons stored in one capacitor to move to the neighboring capacitor, the single electrode of a pixel is usually replaced with two or more electrodes whose voltages are manipulated. In a three-phase CCD, for example, each pixel contains three electrodes or gates (**Fig. 3**).

Readout time and read noise. To maximize the DQE, the amplifier and analog-to-digital converter of a scientific CCD introduce very little noise, usually less than a few electrons per pixel. The noise these circuits adds depends on how quickly they do their job—the faster, the noisier. A large, slow-scan CCD for astronomy may require many tens of seconds to read a single image. This is an important difference between astronomical CCDs and commercial devices like digital cameras and webcams.

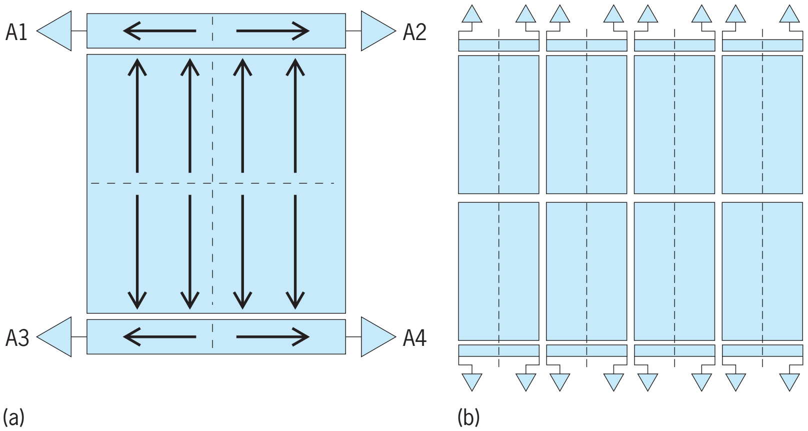


Fig. 4. Large CCD strategies. (*a*) Large monolithic detector with multiple serial registers and amplifiers (four, in this case). Read time is reduced by a factor equal to the number of amplifiers, and total charge transfer efficiency (CTE) is improved. (*b*) Mosaic of eight arrays butted to form a single large-area detector.

Lengthy readout times increase response time and waste telescope time. To speed read time, some chips use multiple amplifiers on a single array. Each of four amplifiers, for example, can simultaneously read one-quarter of an array (**Fig.** 4*a*).

An additional strategy is to build a mosaic of several very closely spaced but electrically independent CCDs, an approach that also circumvents the practical difficulties in manufacturing monolithic CCDs larger than 10–20 megapixels. A relatively simple combination of offset multiple exposures will fill in those parts of an image masked by the gaps between mosaicked CCDs. Kitt Peak National Observatory, for example, placed the eight-element Mosaic Imager in service in 1998 (Fig. 4*b*). As of 2005, several large observatories support mosaic arrays of up to 40 CCDs with total sizes of 200–400 megapixels.

Dark current, cooling, and vacuum enclosures. Thermal agitation can generate electron-hole pairs in the depletion zone of a MOS capacitor. The resulting steady flow of thermal electrons into the CCD potential wells is called dark current. At room temperature, the dark current can easily fill the potential wells and saturate a scientific CCD in a few seconds. Dark current adds some number of electrons to whatever number of photoelectrons are produced in a pixel and also introduces an associated uncertainty (noise). This uncertainty can never be removed, so the signal-to-noise ratio is always degraded by dark current.

Astronomers almost always operate CCD cameras at very low temperatures to reduce dark current. Cameras on spacecraft, if properly shaded from sunlight, can operate at low temperatures without other provisions for cooling. On the ground, a common cooling method places the CCD in thermal contact with a bath of liquid nitrogen (LN$_2$), a chemically inert substance that boils at 77 K (−196°C or −321°F). For liquid nitrogen–cooled CCDs, a typical operating temperature is around −100°C (−148°F). This technique requires sealing the detector in a small vacuum chamber as well as a carefully designed system for handling liquid nitrogen.

At the cost of somewhat higher dark current, relatively inexpensive and easy-to-use solid-state thermoelectric (Peltier junction) coolers can maintain a detector in the −20 to −50°C (−4 to −58°F) range, where the dark current of specially engineered CCDs (the MPP-CCD) is acceptably low for many astronomical applications. *See* PELTIER EFFECT; THERMOELECTRICITY.

Some non-CCD arrays used in the infrared require cooling with expensive liquid helium, which boils at 4.2 K (−269°C or −452°F). A relatively complicated apparatus, the closed-cycle helium refrigerator, can cool detectors to the 10 to 60 K (−263 to −213°C or −442 to −352°F) range without consuming large amounts of liquid helium.

Charge-transfer efficiency. In a CCD readout, the fraction of electrons in a pixel that are successfully moved to the next pixel is called the charge-transfer efficiency. One can avoid errors in the image only if the charge-transfer efficiency is very close to one. Multi-megapixel arrays require charge-transfer efficiency values approaching 0.999999 (known as "six nines"). To avoid decreases in charge-transfer efficiency, astronomical CCDs cannot operate at temperatures below about −100°C (−148°F) and must be designed so that all electrons collect in a region safely removed from the semiconductor–insulator interface. Manufacturers can produce this so-called buried channel by adding a thin layer of *n*-type semiconductor between the *p*-type material and the insulator (Fig. 3).

Surface issues. If a CCD is illuminated from the front side, the insulator and (especially) the gates will absorb and reflect photons at all wavelengths, but particularly in the blue and ultraviolet.

Backthinning. QE improves at all wavelengths, but especially at short wavelengths, if the semiconductor layer is made very thin, so that the buried channel can be illuminated from the back. The delicate technology for fabricating the very fragile backthinned CCD is costly to set up and difficult to perfect, so a thin, back-illuminated CCD is substantially more expensive than the thick, front-illuminated version of the same device.

Frontside options. Several strategies can improve the short-wavelength QE of a front-side illuminated CCD. Gate materials like indium tin oxide are more transparent than polysilicon, but harder to fabricate. Some designs leave gaps between the electrodes, and direct incoming light through them with microlenses. Another strategy coats the front surface with a thin layer of phosphor. A typical phosphor might glow with green light when excited by incoming ultraviolet light, thus producing a detector response to the ultraviolet. Except for cost, all these strategies have so far proven inferior to backthinning.

Antireflection coatings. An appropriate antireflection coating can minimize the reflection losses from any surface. The coating is most effective for light whose wavelength is four times its thickness, so a CCD designer must choose the coating with the intended use of the detector in mind. Usually CCDs are coated either to enhance the short-wavelength response or to minimize reflectivity near the device's QE maximum.

Processing Array Data

No matter what source is being investigated, certain calibration steps, called preprocessing, are necessary to remove the instrumental signature from array data. Further processing is then carried out to improve image quality.

Preprocessing. The calibration includes subtractive corrections and steps to improve linearity and flat-field response.

Subtractive corrections. If a detector is not exposed to a signal from the telescope but simply sits in the dark for a time, it can nevertheless produce a positive signal. This results from two effects: the signal bias, or zero level, which is always present to ensure amplifier linearity, and the accumulated thermal dark current, which grows with exposure time. The calibration step here is to take several exposures with the shutter closed, compute the bias and dark rate, and subtract the appropriate levels from every image produced by the array.

Linearity. A nonlinear detector requires a special calibration step to convert its response into one directly proportional to the input illumination.

Flat-field response. Identical light signals generally do not produce identical responses in every pixel of a detector array. This defect can arise because of structural quantum-efficiency differences intrinsic to the array or because of imperfections in the optical system. The calibration procedure is to observe images of a perfectly uniform source (for example, a projector screen or the twilight sky) and then divide every data frame by this flat field image.

Processing. After preprocessing, subsequent treatment of images depends on the information being sought. Background removal is often important. Removal of a smooth background (due to sky glow or light pollution) is usually a simple subtractive operation. Cosmic rays strike pixels randomly and generate electrons in them, and these are most easily removed if multiple exposures of the same scene can be combined (astronomical light sources repeat positions, but cosmic rays do not). Combining multiple exposures can also fill in detector defects like insensitive pixels or the gaps in a mosaic. Since the SNR generally increases with the number of electrons or photons counted, combining multiple exposures increases data quality.

Frederick R. Chromey

Bibliography. R. Berry and J. Burnell, *The Handbook of Astronomical Image Processing*, Willmann-Bell, Richmond, 2000; S. B. Howell, *Handbook of CCD Astronomy*, Cambridge University Press, 2000; C. Kitchin, *Telescopes and Techniques*, 2d ed., Springer, London, 2003; I. McLean, *Electronic Imaging in Astronomy*, Wiley, 1997.

Astronomical photography

The application of the photographic process to astronomy, including monochrome photography and color photography.

Monochrome Photography

Monochrome photography was one of the premiere tools of astronomical research during most of the twentieth century. It offered two major advantages: the integration of signal (through time exposure) allowed the accumulation of photons from very faint objects which could not otherwise be seen; and the storage of information in an efficient and permanent form allowed protracted, in-depth study of astronomical objects away from the telescope. Monochrome astronomical photography became a standard technique that was applied to direct imaging, spectroscopy, photometry, polarimetry, and astrometry. *See* ASTROMETRY; ASTRONOMICAL SPECTROSCOPY; PHOTOMETRY; POLARIMETRY.

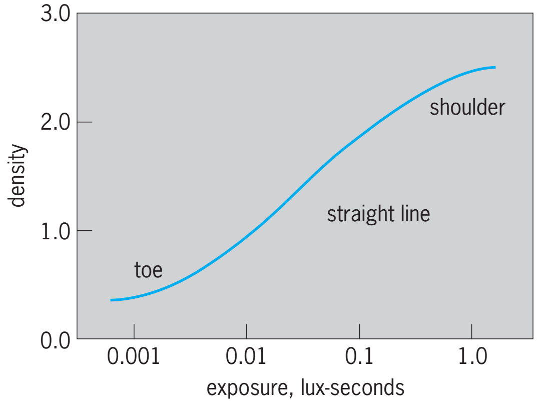


Fig. 1. Typical characteristic curve describing the relationship between density and exposure for a hypothetical monochrome emulsion.

Nonetheless, because astronomers were continually searching for increased sensitivity and wavelength coverage, the aggressive attempts to improve photographic emulsions had, by the 1970s, begun to reach the limits of what was reasonably attainable. At the same time a new type of electronic imager, the charge-coupled device (CCD), developed for military applications, began to become available to the scientific community. Space-based imaging applications of the charge-coupled device, particularly the development of the TI 800 × 800-px chip for use in the *Galileo* spacecraft, stimulated expanded use of this mode of imaging at ground-based observatories. By the end of the twentieth century, charge-coupled devices had become so pervasive in astronomical observatories that it was clear that monochrome astronomical photography was, in many respects, a technique of the past. Rarely used now by professional astronomers, the techniques of high-quality monochrome astrophotography are primarily preserved by amateur astronomers. Some advances are still being achieved, primarily through computer-processing of digitized photographic images. *See* CHARGE-COUPLED DEVICES.

Performance. Although the acquisition of astronomical photographs has declined dramatically, large photographic archives will continue to serve as important resources. An understanding of the properties of photographic emulsions will remain necessary to fully exploit these archives.

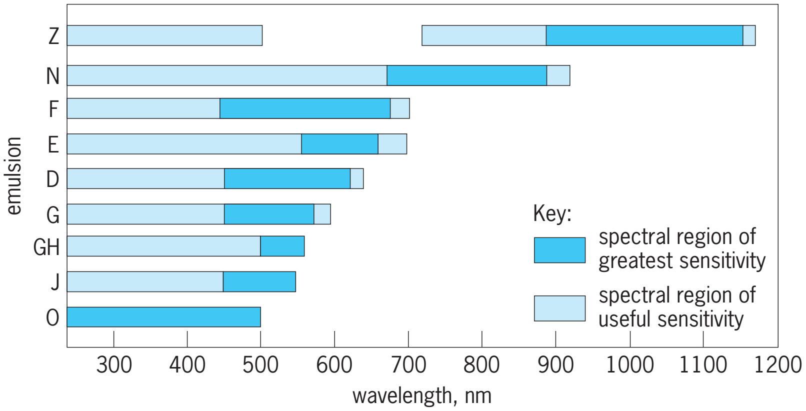


Fig. 2. Spectral sensitization for a variety of monochrome astronomical emulsions. (*After Eastman Kodak Co., Scientific Imaging with Kodak Films and Plates, Publ. P-315, 1987*)

Characteristic curve. One important property of a photographic emulsion is represented by the characteristic curve, also known as the Hurter-Driffield (H-D) curve (**Fig. 1**). This curve is variously plotted as the transmittance, opacity, or density of the processed emulsion versus the exposure, defined to be the product of light intensity and exposure time. Unlike the now-favored electronic detectors, which exhibit a linear analog to the characteristic curve, the H-D curve of an emulsion has a distinct nonlinearity, which complicates the use of photographic data. One result is that on the toe and shoulder of the curve, that is, for both very faint and very bright objects, the monochrome photograph provides little photometric distinction.

Nonetheless, the characteristic curve is very useful in selecting the most suitable emulsion for a particular application, and in optimizing the exposure time and processing necessary for making the most useful image of the specific objects of interest. The straight-line portion of the characteristic curve defines the range of exposures that may most reliably be used for determination of brightnesses of stars. The slope of the curve gives an index defining the level of contrast of the photograph.

Wavelength sensitivity. Photographic emulsions are also characterized by their sensitivity as a function of wavelength (**Fig. 2**). O emulsions are blue sensitive, J are useful into the green, E and F have very broad responses into the red, and N is usable into the near-infrared. The most important professionally used astronomical emulsions, which are still available, are the IIIa-J, IIIa-F, I-N and IV-N (progressively more sensitive to the red and near-infrared portions of the electromagnetic spectrum). These plates are produced in large sizes and were most recently used professionally in all-sky-survey imaging. The general wavelength sensitivity of a photographic emulsion can be matched to the specific scientific problem at hand; fine tuning of the spectral window that is imaged is accomplished by use of optical filters in conjunction with the emulsion. The characteristic curves of these emulsions vary widely. For example, near-infrared emulsions have much lower sensitivity (and hence longer required exposure) than more blue-sensitive emulsions.

Image structure. The image-structure characteristics of emulsions are critical to the application of the photograph. The most important of these characteristics is the granularity, or graininess, of the photograph. The size of the clumps of silver halide grains which form the image in the emulsion have a strong effect on the resolving power of the photograph, that is, the fineness of the detail which can be measured. Granularity is indicated by the prefix in the emulsion designator, types 103, II, III, IV, V, and 649 being progressively finer grained. In general, lower granularity is obtained at the expense of longer exposure times to achieve similar densities in the processed photographs.

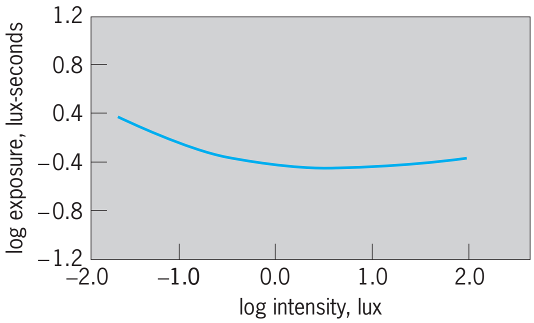


Fig. 3. Result of reciprocity failure for a hypothetical monochrome emulsion. The graph shows the amount of exposure required to produce a constant density at different levels of illumination for a single photographic emulsion type. Loss of sensitivity at very high or very low levels of illumination is reflected in the greater exposures required at these extreme levels.

Reciprocity failure. In the low-light-level environment of astronomical imaging, exposure times can be quite long, from minutes to hours. In theory, the law of reciprocity for chemical reactions implies that the total exposure is equal to the intensity times the exposure time. In practice, for most photographic emulsions, this law fails for long exposure times (**Fig. 3**). Reciprocity failure, whereby increased exposure time does not produce a commensurate increase in total exposure, has been partially mitigated by special processing of certain astronomical emulsions. These emulsions are distinguished by the letter a in the type designator, for example, IIa-O. Two primary factors in the reciprocity characteristics of photographic emulsions are the temperature of the emulsion during exposure, and the gaseous environment in the camera as the emulsion is exposed.

Hypersensitization. Several techniques have been used to improve the performance of astronomical emulsions. Generally described as hypersensitization, they include cooling the emulsion during exposure (to mitigate reciprocity failure), baking the plate prior to exposure, removing oxygen and water from the emulsion prior to use, treating the plates with hydrogen gas prior to exposure, and preflashing the emulsion with a low level of uniform light. In addition, there are processing steps which can improve the performance of the emulsion, such as the extension of development times beyond what is normally specified.

Applications. The great majority of monochrome astrophotography is now done by amateur astronomers, though much of it is quite professional in nature and can provide interesting images with resolutions and image scales not presently attainable with charge-coupled-device images. Targets for this type of imaging are frequently large, extended objects, such as nearby galaxies, galactic star fields, and various types of extended nebulae, as well as comets with extended tails. Large, systematic photographic programs have disappeared from astronomy, and the remaining photography is done primarily on an ad-hoc basis.

Photographic emulsions are now most commonly used by amateur astronomers and, for this application, are found primarily on flexible film substrates, and only rarely on glass plates. The emulsion types which are most popular for astronomical imaging are Kodak Technical Pan Film (Type 2415), and Kodak T-MAX 400 Professional Film. These emulsions are widely used in a standard 35-mm format, although they are also available in larger-size sheets on very stable Estar substrates. Even among amateur astronomers, the process of gas-hypering film to increase photosensitivity is fairly common.

Computers are widely used to provide enhancements to monochrome photographic images. Films are typically scanned at high resolution to provide digital image files for subsequent processing. Software tools of many types are used, ranging from commercial applications packages designed for photoprocessing for desktop publishing to special-purpose astronomical packages. Multiple photographic images of the same field can be added in order to increase signal-to-noise ratio and dynamic range, and images are frequently contrast-enhanced, or digitally filtered to improve sharpness. Very precisely controlled color images are often produced from monochrome photography using multiple images of a given field obtained through glass band-pass filters, and digital image processing techniques are used to produce very large, high-resolution mosaics of astronomical images by precisely abutting adjacent images.

Both existing and new photographic archives will be given additional power as techniques of digitization of large databases improve, and as digital image processing technology continues to evolve. Many important photographic archives have already been digitized and made available in electronic formats, and it is expected that many more will be digitized in coming years. *See* IMAGE PROCESSING. Eric R. Craine

Color Photography

The language of astronomy abounds with references to color, and the concept of color is implicit in many of the measurements that astronomers make. Color index, for example, is a quantity related to the temperature of a star, while the redshift of a galaxy is used to indicate its recessional velocity. More directly, stars may be described as red giants or white dwarfs or even blue stragglers. *See* BLUE STRAGGLER STAR; COLOR INDEX; REDSHIFT; STAR.

These names reflect the underlying importance of color in astrophysics and cosmology, and though the colors involved are subtle and difficult to distinguish by the eye in its dark-adapted state, special photographic techniques can be used to display them. A realistic representation of the true colors of celestial bodies can reveal new relationships in familiar objects and add an important third dimension to the morphology and brightness information of the more usual monochrome representations.

Color films. As discussed above, special photographic materials are necessary to accommodate the unusual requirements of photography in astronomy. Not only is the amount of incoming radiation to be

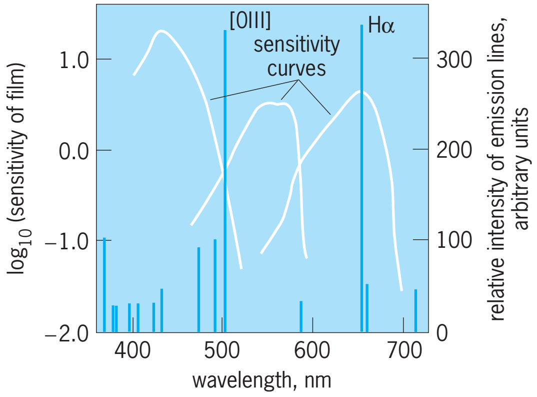


Fig. 4. Comparison of the spectral sensitivity curves of a color film with the emission-line spectrum of the Orion Nebula.

detected extremely small, but it is accompanied by unwanted light from the night sky (the airglow). The materials must therefore combine extreme sensitivity at long exposures with high contrast. The ability to detect faint objects is ultimately more dependent on the contrast and resolution of the photographic material than on the light grasp of the telescope or available observing time. On the other hand, color films are designed for general use at levels of illumination where high contrast and low-light-level efficiency are unimportant. In addition, these films are intended primarily to reproduce the broadband colors of everyday life, and for this the rather uneven spectral response of their individual layers is unimportant (**Fig. 4**). Unfortunately, gaseous nebulae emit most of their visible radiation in the form of monochromatic emission lines from the ionized elements present. In the visible region, the strong green forbidden lines of doubly ionized oxygen [O III] near 500 nanometers and the rich red of a hydrogen recombination line H*α* at 656 nm often predominate. The green line falls at a minimum in the blue-green sensitivity of a typical color film and is not well recorded; however, the red line coincides with the peak sensitivity of the red-sensitive layer. Thus, color films always show gaseous nebulae as red, largely irrespective of the contribution from the green oxygen line, whereas yellow (red + green) would be a more realistic representation in many cases. *See* NEBULA.

A further problem is the effect of long exposures on the relative sensitivity of the three layers, which are differently affected by low-intensity reciprocity failure. Changes in both sensitivity and contrast of the layers are found, and exposures which are long enough to be astronomically useful often produce severe color-balance distortion. However, in the absence of materials prepared specifically for color astrophotography, attempts have been made to minimize the problems of ordinary color films and exploit their advantages of low cost, ease of use, and ready availability. Both reversal and negative films have been used. Most workers appear to favor the higher contrast and brighter colors of reversal films, but the greater latitude and color-correction possibilities of negative materials are considerable advantages, especially since exposure times are difficult to assess. The long-exposure properties of both types of film can be improved by user-applied preexposure treatments.

Cooled cameras. As discussed above, low-intensity reciprocity failure of both color and monochrome films is reduced if the long exposure is made at a low temperature. Most experiments have been made with cameras designed for fairly small formats and cooled to about −103°F (−75°C) with solid carbon dioxide. Care must be taken to avoid the condensation of water vapor on the film during exposure and to ensure that the cooling is uniform across the frame. This becomes increasingly difficult with the large formats used in professional astronomy, and is impossible in some types of telescopes. Exposure at low temperature reduces the effects of low-intensity reciprocity failure on both the speed and color balance of all types of films.

Hypersensitization of color film. Some of the techniques which are used for spectroscopic plates may also be applied to color films. Baking both in nitrogen and in forming gas, a 2–4% hydrogen-in-nitrogen mixture, is useful. Forming gas is available in special kits specifically for hypersensitizing small quantities of films for astrophotography. Films are baked for several hours in a flow of the gas at 150°F (65°C) just prior to exposure and then (preferably) exposed in a nitrogen atmosphere. Substantial long-exposure speed gains are reported. However, the process affects each of the three sensitive layers differently, and some shift in color balance may be experienced.

These user-applied processes reduce some of the disadvantages of color films for astrophotography, and push development may be used in addition to the above, to increase both speed and contrast. However, the basic problem of uneven spectral response remains. As a result, color films can reproduce only realistic colors of the brighter, continuous-spectrum objects, such as planets, stars, and galaxies. Faint objects and emission nebulae are not well recorded.

Indirect color photography. An alternative approach is to use the oldest system of color photography, the three-color separation technique. This process derives directly from James Clerk Maxwell's 1861 demonstration of additive color imaging. In this system, three exposures are made with combinations of photographic emulsions and filters chosen to record the red, green, and blue parts of the spectrum on separate plates or film. In professional astronomy, these passbands are achieved with Eastman Kodak's fine-grain IIIa-J and IIIa-F (blue-green and red-sensitive) emulsions with glass filters to record the appropriate broadband colors. Kodak's Technical Pan film also works well after hypersensitization, but for the faintest objects under truly dark skies exposure times can be 90 min per color, even with an *f*/2.5 telescope such as the United Kingdom Schmidt. *See* SCHMIDT CAMERA.

Filter-emulsion passbands are chosen to ensure adequate overlap between adjacent colors so that hues intermediate between red, green, and blue are well recorded. With care in selection of these parameters, coverage of the visible region is much more uniform than is possible with conventional color film. This overlap ensures that narrow emission lines across the visible spectrum are well recorded, which is not the case with color film (Fig. 4).

The same principles are used with digital detectors such as charge-coupled devices, which are much more sensitive than hypersensitized silver-based photographic materials. As yet, they lack the essentially unlimited area and small pixels that give conventional materials their "photographic" characteristics that translate into a distinctive esthetic quality. Charge-coupled-device images are usually combined into three-color images with a computer, using commercially available software, and similar methods can be used to combine digitized versions of photographic red, green, and blue exposures.

Recombining color separations. Two methods are possible to recover the color information in three-color separations. The subtractive process involves imagewise combinations of yellow, magenta, and cyan dyes or pigments and is now rarely practiced outside professional printing applications, though it was once popular as the dye transfer process used for the finest color reproduction. It was also the basis of the Technicolor process. Dyes or pigments used subtractively also form the the color systems used in all commercial films and printing processes. *See* PRINTING.

Much more flexible in the astronomical application is the additive process, which allows several levels of image manipulation before the monochromes are combined. Additive color photography involves mixing colored light, rather than colored compounds, and its most common manifestation is the color television or computer screen image, which a magnifier shows to be made up of blue, green, and red dots or strips. When used photographically, monochrome positive copies are made by contact copying the three original separation negatives onto a suitable film material. *See* ELECTRONIC DISPLAY; PICTURE TUBE.

At this stage, a wide range of image-manipulation techniques can be applied to enhance small or faint features and to adjust the contrast of the original images. The positives are subsequently enlarged sequentially in register onto a positive-working color material (such as Cibachrome or Ektachrome) or onto a negative stock for subsequent printing by normal substractive methods. The positives are enlarged through the equivalent of the taking filters, the red positive through a red filter, and the green through a green filter, and so forth. This process is easier to use and more flexible than any of the subtractive systems and can yield color prints of both scientific and esthetic merit.

Color balance. The eye is a poor discriminator of color at low light levels, and it is not possible to check the color balance of astronomical photographs against the original scene or against color memory to verify color fidelity. Color pictures produced on the 154-in. (4-m) Anglo-Australian telescope are balanced by comparison with a neutral gray step-wedge image projected onto the plates during the telescope exposure. The light source of the projector is filtered to a color temperature of 5500 K (9400°F; equivalent to sunlight). Thus, sunlike stars appear white on pictures from this observatory, and stars hotter or cooler than the Sun appear blue and yellow respectively. Likewise, the balance of color in an emission-line nebula is adjusted with reference to the same gray step wedge on the assumption that the spectrum has been recorded equivalently in all three passbands. *See* PHOTOGRAPHY. David F. Malin

Bibliography. Eastman Kodak Co., *Scientific Imaging with Kodak Films and Plates*, Publ. P–315, 1987; T. Hallas and D. Hallas, Astrophotography with a twist, *Sky Telesc.*, 96(5):130–137, November 1998; I. McLean, *Electronic Imaging in Astronomy*, Wiley, 1997; D. F. Malin, An Integrated system of astronomical photography, Parts II and III, in Eastman Kodak, *TechBits Mag.*, Issue 1, pp. 1–10, and Issue 3, pp. 3–11, 1990; D. F. Malin, *A View of the Universe*, 1993; D. F. Malin and P. G. Murdin, *The Colours of the Stars*, 1984; G. Walker, *Astronomical Observations*, Cambridge University Press, 1987.

Astronomical spectroscopy

The use of spectroscopy (the analysis of light as a function of wavelength) as a tool for obtaining observational data on the chemical compositions, physical conditions, and radial velocities of astronomical objects. Astronomical applications of optical spectroscopy from ground-based observatories cover the electromagnetic spectrum from the near-ultraviolet (wavelengths around 0.3 micrometer) through the visible (0.4–0.7 μm) and into the near-infrared (2 μm). Space-based observatories, such as the Hubble Space Telescope, the *Far Ultraviolet Spectroscopic Explorer*, and the *Spitzer Space Telescope*, extend spectroscopic observations from the far-ultraviolet (0.1 μm) to the far-infrared (200 μm). Work at shorter wavelengths (x-ray and gamma-ray spectroscopy) and longer wavelengths (submillimeter and radio wavelengths) requires techniques other than those discussed here. *See* GAMMA-RAY ASTRONOMY; HUBBLE SPACE TELESCOPE; INFRARED ASTRONOMY; RADIO ASTRONOMY; SPITZER SPACE TELESCOPE; SUBMILLIMETER ASTRONOMY; ULTRAVIOLET ASTRONOMY; X-RAY ASTRONOMY.

Usually a spectrograph is fitted to a reflecting telescope, which serves as a light collector. The image of the celestial body being studied is focused on the spectrograph slit, which limits the region under study (thus improving the spectral resolution) and reducing the contribution by the night sky. The diverging light beam then passes from the slit to a collimator (either a lens or mirror). This produces parallel light, which is then dispersed by a diffraction grating or prism. The dispersed light enters a camera, which focuses the spectrum onto a detector, either a charge-coupled device (CCD) in the case of an optical spectrograph, or an electronic array sensitive to infrared light. *See* CHARGE-COUPLED DEVICES; DIFFRACTION GRATING; SPECTROGRAPH; SPECTROSCOPY; TELESCOPE.

Resolving power. The resolving power of a spectrograph is usually denoted by R, and is equal to the wavelength of light, λ, divided by the spectral resolution, $\Delta\lambda$; that is $R = \lambda/\Delta\lambda$. The size of the spectral resolution element is set by the width of the slit, the grating, the focal length of the camera, and the detector resolution. Typically $R \approx 200$ for low-dispersion survey work, $R \approx 5000$ for "classical" spectroscopy, and $R \approx 50{,}000$–$100{,}000$ for the highest-resolution work. *See* RESOLVING POWER (OPTICS).

Multiobject spectroscopy. It is often desirable to obtain spectroscopy of many of the objects within a telescope's field of view in a single exposure. A variety of methods are available to accomplish such surveys.

Slitless spectroscopy. It is possible to take spectra of all of the brighter objects within the field of view by not using a spectrograph at all, but by combining a low-dispersing element directly with the telescope. For instance, an objective prism may be placed in front of the telescope, which is often a Schmidt camera. Slitless spectroscopy has been used for large stellar surveys such as the Henry Draper Catalogue and, more recently, the Case Michigan Survey. Alternatively, a grism or grens (a hybrid between a prism and a grating) may be placed in the converging beam near the focal plane. Such spectroscopy is usually of low resolving power ($R \approx 100$–500) and is limited to the objects brighter than the sky background, as slitless spectroscopy has no way of attenuating the amount of sky seen by the detector. However, the use of large-format arrays of charge-coupled devices which can minimize the effects of the sky background opens new possibilities in searches for faint objects. *See* ASTRONOMICAL CATALOGS; SCHMIDT CAMERA.

Slitlet masks. In this technique, a picture is usually taken of a region containing several astronomical objects of interest; the exact locations of these objects are determined, and small slits (slitlets) are then milled in the corresponding locations in a metal plate. This plate is substituted for the slit in a conventional spectrograph. Care must be taken that the spectra do not overlap, which greatly limits the number of objects that can be observed at any one time to 10–100. Such systems have the advantage of good subtraction of the night sky, and very high efficiency. This technique has enjoyed a revival, with multiobject slitlet masks in use on the world's largest telescopes [for example, the DEIMOS spectrograph on the Keck 10-m (386-in.) telescope, GMOS spectrographs on the Gemini 8-m (315-in.) telescopes, and the IMACS spectrograph on the 6.5-m (256-in.) Magellan (clay) telescope].

Fiber-fed spectroscopy. Rather than milling slitlets in a plate, holes may be drilled, which are then plugged with optical fibers. (Such an arrangement is often referred to as a plugboard.) The light is then transported via the fibers to a spectrograph mounted on an optical bench in a laboratorylike environment adjacent to the telescope. Alternatively, robotics may be used to position fibers in the focal plane; the fibers are then anchored to a metal plate via magnets. At the spectrograph, the fibers are arrayed in a line and act as the spectrograph slit. Although significant light losses may occur in this scheme (typically factors of 2–3), hundreds of objects can be observed simultaneously, leading to very effective use of the telescope. The faintest object that can be usefully observed is dependent upon how well the relative transmissions of the various fibers can be calibrated, since some fibers will be sampling the night-sky spectra alone, while others will contain the spectra of the night sky plus a faint astronomical object of interest. *See* FIBER-OPTICS IMAGING; OPTICAL FIBERS.

Echelle spectroscopy. Normal spectrographs employ diffraction gratings that are intended to be used in low orders ($n = 1$, 2, or 3), with colored glass filters used to prevent overlap of adjacent orders. Echelle spectrographs differ from conventional systems in that they employ gratings intended to be used in very high orders ($n > 10$), resulting in very high resolving power. Normally these orders would fall on top of one another, rendering the data useless. An echelle uses a second dispersal element, usually another grating but sometimes a prism, at right angles to the first, in order to separate the successive spectral strips from each other. A large range of wavelengths can be obtained in the format of nearly parallel segments, well suited for charge-coupled devices.

Coudé spectroscopy. In the past, high resolving power was achieved primarily by using a large, stationary spectrograph located off the polar axis of an equatorial telescope. Light would be brought through the hollow axis of the telescope by a series of flat mirrors (the coudé train). Coudé spectrographs are very stable, as they are located off the telescope, and the optics and gratings can be quite large as they do not need to be attached to the telescope. Echelles and fiber-fed bench-mounted spectrographs have largely replaced the coudé spectrograph; most modern telescopes are now built with altitude-azimuth mounts rather than equatorial mounts, leading to the further decline of coudé systems.

Integral field spectroscopy. In this application, a close-knit bundle of optical fibers is placed in the focal plane and is used to observe an extended astronomical object, such as a gaseous nebula or a galaxy. The light is transmitted via the fibers to a bench-mounted spectrograph. Although the fibers are in a linear array at the spectrograph, their locations in the focal plane are known, and sophisticated data reduction techniques allow the astronomer to reconstruct a spectral "image" of the object.

Infrared spectroscopy. Infrared spectroscopy is sensitive to light at wavelengths longer than about 1 μm through about 200 μm. At ground-based telescopes, observations are possible in the near-infrared (1–5 μm) and in the mid-infrared (5–30 μm). Far-infrared (30–200 μm) spectroscopy is limited to space-based observatories (such as the *Infrared*

Astronomical Satellite, Infrared Space Observatory, and *Spitzer Space Telescope*) because these wavelengths are absorbed by the Earth's atmosphere.

Infrared spectroscopy is a necessary tool for the study of objects embedded in dusty regions, for example the nuclei of certain galaxies, dust-enshrouded massive stars, and regions of star formation in molecular clouds local to the Milky Way Galaxy. Because dust very effectively scatters shorter-wavelength light, there are some objects which are visible only in infrared light. Furthermore, the Earth's atmosphere is more stable in the infrared, as the distorting twinkling of starlight is less at infrared wavelengths, resulting in sharper images of extended objects and binary star systems, useful for angularly resolved spectroscopy of such objects. Most molecules produce rich absorption spectra in the infrared. Cool objects like very young stellar embryos recently formed in molecular cloud cores, very low mass stars, brown dwarfs, and giant planets not only are brighter in infrared light but also manifest more molecular lines than atomic lines and thus are more suitably studied in the infrared regime. *See* MOLECULAR CLOUD; TWINKLING STARS.

Infrared spectroscopy is a relatively young field; early studies using Fourier transform spectrometers (discussed below) began in the 1970s but were limited to the most prominent targets, principally the Sun and the brightest stars. The development of sensitive infrared array spectrometers in the late 1980s and early 1990s paved the way for a plethora of discoveries, from the ultraluminous infrared galaxies, to the largest known redshifts in the universe, to the existence and nature of the brown dwarfs, failed stars intermediate in mass between stars and giant planets. Current generations of infrared spectrometers in use and under construction incorporate traditional features of visible light spectrometers, such as echelle gratings, integral field units, and multiobject capability. *See* BROWN DWARF; GALAXY, EXTERNAL; INFRARED SPECTROSCOPY; REDSHIFT.

Fourier transform spectroscopy. Fourier transform spectroscopy, used particularly in the near-infrared, employs a concept entirely different from the spectrographs described above. Instead of being dispersed in a spectrograph, the light of a wide band of wavelengths is passed through a Michelson interferometer with variable spacing of its two apertures. The resulting interferogram, which is an electronic record of the interference signal produced by the interferometer as the separation of the apertures is varied, is converted into a record of intensity versus wavelength by a computer, and is of extremely high spectral resolution. *See* INTERFEROMETRY.

Spectrophotometry. Often it is useful to know the actual flux of a source in physical units arriving at the Earth as a function of wavelength. In order to achieve this, care must be taken to account for atmospheric light losses. The data can then be calibrated by observing spectrophotometric standard stars; these are stars whose spectral energy distribution is known via comparison to laboratory sources. Such calibration removes the wavelength dependence in the response of the telescope, spectrograph, and detector, and hence allows a direct comparison of the spectrum with model calculations, providing information complementary to that determined from analysis of spectral lines.

Spectropolarimetry. Measurements of polarization in continuum and line spectra often enable the recognition of nonthermal processes (such as synchrotron radiation) in the source. In many instances, they make it possible to assess the importance of light scattering by electrons or small solid grains. *See* POLARIMETRY.

Precision radial velocities. Measurements of the wavelength displacement, due to the Doppler effect, between the spectral lines in a celestial object and a laboratory source give the line-of-sight velocity of the source with respect to the observer. With conventional comparison sources (discharge tubes) and standard spectrographs, the precision of radial velocities is generally of order 1–3 km/s (0.6–1.8 mi/s). However, improvements have pushed this to 5 m/s^{-1} (20 ft/s) or better by using fiber-fed bench-mounted spectrographs, or by using absorption cells of gas (such as iodine) through which the starlight passes to provide an excellent internal reference system. The use of precision radial velocities, combined with dedicated observational programs, has resulted in the detection of planets around nearby stars, evidenced by small, periodic radial velocity variations in their parent stars. *See* DOPPLER EFFECT; PLANET.

Applications. The application of astronomical spectroscopy extends from solar system objects (the Sun, planets, and comets) to Milky Way objects (stars, including binary stars, ordinary novae, and cataclysmic variables; and gaseous nebulae, such as supernova remnants, H II regions, and planetary nebulae) and to distant galaxies and quasars.

Emission-line spectroscopy. Rarefied gas seen against a cooler background will produce an emission-line spectrum. In gaseous nebulae, such as H II regions and planetary nebulae, gas is ionized by the extreme ultraviolet emission of hot stars. As the electrons recombine with atomic nuclei, emission lines are

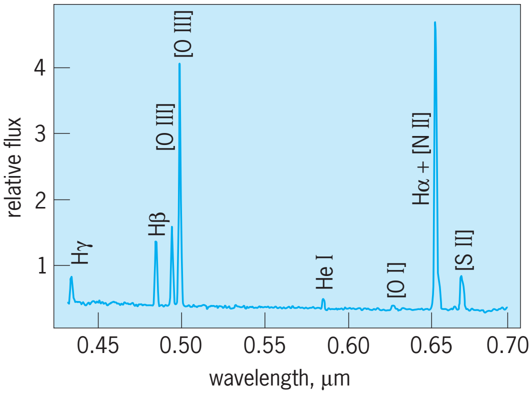


Fig. 1. Spectrum of the largest H II region in the irregular galaxy NGC 1156. The continuum comes from the underlying stellar population. Forbidden lines of neutral and doubly ionized oxygen (O I), (O III), singly ionized nitrogen (N II), and singly ionized sulfur (S II) are visible.

produced. Spectroscopy of the gas reveals physical conditions (density, electron temperatures) as well as chemical composition. The strongest optical nebular lines are the hydrogen Balmer lines and forbidden lines of doubly ionized oxygen ("forbidden" in that these lines would not be seen at higher densities). Forbidden lines of other elements are also usually present, in addition to neutral helium lines (**Fig. 1**). *See* NEBULA; PLANETARY NEBULA.

Some stars also show emission lines in their spectra, indicative of hot gas seen in extended envelopes or disks around these stars. Examples include Be stars (B-type stars whose spectra show emission lines) and Wolf-Rayet stars. *See* WOLF-RAYET STAR.

Absorption-line spectroscopy. More commonly, gas is seen projected against a hotter source, as is the case in a stellar atmosphere or in gas found in interstellar space seen against stars or galaxies. In this situation an absorption-line spectrum is produced.

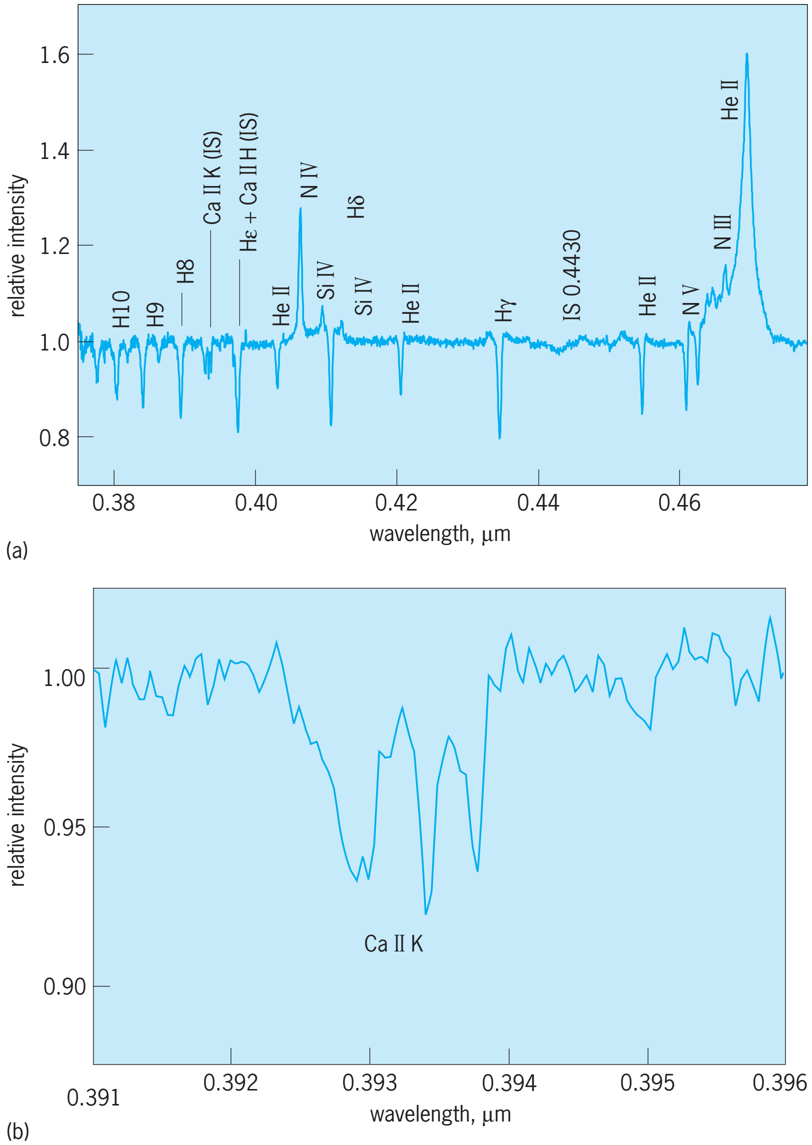


Fig. 2. Spectra of the star Sk-69° 22, one of the hottest, most luminous, and massive stars known, in the Large Magellanic Cloud. (*a*) Part of the optical spectrum, 0.375–0.478 μm. IS indicates absorption lines from interstellar gas. (*b*) Detailed spectrum of the K line of singly ionized calcium (Ca II), 0.391–0.396 μm.

The Sun was the first astronomical object whose spectrum was analyzed; Joseph Fraunhofer discovered hundreds of dark bands superposed on an otherwise continuous range of colors in 1815. Subsequent analysis of the dark-line Fraunhofer spectrum allowed astronomers to deduce the solar chemical composition, the temperature, density values, and their variation with depth in the solar photosphere. Later discovery of the splitting of spectral lines due to the Zeeman effect demonstrated the existence of magnetic fields on the Sun. The attenuated outer envelopes of the Sun, the chromosphere and corona, require special instrumentation, as they are very faint compared to the bright surface of the Sun. *See* CHROMOSPHERE; SOLAR CORONA; SOLAR MAGNETIC FIELD; SUN; ZEEMAN EFFECT.

Differences in the spectral appearance of stars are due primarily to differences in their effective temperature (the temperature at the surface), and only secondarily to surface gravity (pressure). Differences in chemical compositions are usually a tertiary effect, except for stars with significant chemical peculiarities. The spectra of most stars can be classified in a two-parameter spectral class and luminosity system. Strengths and appearance of different spectra lines supply classification criteria. Classification indices are usually temperature-sensitive lines; luminosity criteria are usually density-sensitive lines. Stars of high luminosity have lower atmospheric densities. The spectral sequence is described by letters O, B, A, F, G, K, M, and L, ranging from hot, about 50,000 K (90,000°F), to cool, 2000 K (3500°F), with decimal subdivisions. Luminosity classes range from I (very luminous) to V (main-sequence dwarf stars). The Sun's spectral class is G2 V, which means it is closer to G0 than to K0 and is a dwarf star. *See* HERTZSPRUNG-RUSSELL DIAGRAM; SPECTRAL TYPE; STAR.

The properties of stellar spectra may be illustrated by the spectrum of the star Sk-69° 22 (**Fig. 2*a***), one of the hottest, most luminous, and massive stars known, with a spectral type of O3 I. This star is located in the Large Magellanic Cloud, the nearest galaxy to the Milky Way Galaxy. Absorption lines of hydrogen (H) and singly ionized helium (He II) dominate. The lack of neutral helium (He I), and the presence of four-times ionized nitrogen (N V), require that the star's temperature be at least 50,000 K (90,000°F). The presence of emission lines of doubly and triply ionized nitrogen and triply ionized silicon (N III, N IV, Si IV), and the very strong, broad emission line of He II at 0.4686 μm, indicate that this star is of extremely high luminosity; these emission lines are formed in the expanding outer envelope of the star.

Certain stars, particularly white dwarfs, have intense magnetic fields whose strength can be measured by the splitting of spectral lines similar to that done for the Sun, although without the possibility of spatial resolution over the disk of the star. In fact, understanding of the importance of magnetic fields in astrophysics derives largely from spectroscopic data. *See* STELLAR MAGNETIC FIELD; WHITE DWARF STAR.

The presence of matter in interstellar space manifests itself in sharp, strong absorption lines. Despite the extremely low density of gas found between stars, the distances to some stars are so great that the effective column depth can be appreciable. Among the strongest interstellar lines in the optical are the H and K resonance lines of singly ionized calcium (Ca II). Broad, diffuse interstellar bands are also seen, the strongest of which in the optical occurs at 0.443 μm. Both of these can be seen in the spectrum of Sk-69° 22, and detailed inspection of its Ca II K line (Fig. 2*b*) shows multiple components, with radial velocities corresponding to gas in both the Milky Way Galaxy and the Large Magellanic Cloud. Extremely luminous, distant quasars make it possible to probe the conditions of the interstellar medium when the universe was quite young, since their light passes through the interstellar media of a number of intervening galaxies. *See* INTERSTELLAR MATTER; QUASAR.

Spectroscopy can be obtained of the resolved, luminous members of the very nearest galaxies; for all but this handful, the spectrum of a galaxy is a composite of its typically 10^{12} stellar members. Such integrated spectroscopy can be analyzed by population synthesis, in which a library of stars is mixed in appropriate proportions until something similar to the observed spectrum is seen. Such studies can place useful limits on the relative numbers of various types of stars, but whether the mix is a unique match to the integrated spectrum is often open to question without additional information.

Reflection spectroscopy. The spectra of solar system objects (other than the Sun) are primarily reflected sunlight. Spectra of planets with atmospheres also show lines and bands of molecules which make it possible to assess atmospheric temperatures and densities in favorable circumstances. For instance, the presence of carbon dioxide (CO_2) in the atmosphere of Venus was established in 1932 from spectroscopic observations. The spectra of comets show a reflection spectrum of the Sun upon which are superposed bright bands of CN, molecular carbon, and other molecules. They also show intense fluorescence effects. *See* COMET; VENUS.

Continuous energy distributions. In addition to the spectral lines, the energy distribution of the continuum—the light between the spectral lines—often provides information about the physical conditions which is highly complementary to the information that can be gleaned from the spectral lines themselves. Measurement of the intrinsic continuum shape requires use of spectrophotometry, and the ability to correct for any reddening due to interstellar material. The continuous energy distribution will sometimes reveal limitations in stellar atmosphere models that would otherwise cause analysis of the spectral lines to yield misleading results. In addition, the continuous energy distribution, particularly over a large wavelength interval, provides information that is necessary for population synthesis to produce unique results.

Philip Massey; Lisa Prato; Lawrence H. Aller

Bibliography. L. H. Aller, *The Atmospheres of the Sun and Stars*, 2d ed., Ronald Press, New York, 1963; P. M. Gray (ed.), *Fiber Optics in Astronomy II*, Astronomical Society of the Pacific, San Francisco, 1993; C. R. Kitchin, *Optical Astronomical Spectroscopy*, Institute of Physics Publishing, Bristol, 1995; A. D. Thackeray, *Astronomical Spectroscopy*, Macmillan, New York, 1961.

Astronomical transit instrument

A telescope adapted to the observation of the passage, or transit, of an astronomical object across the meridian of the observer. The astronomical transit instrument is the classic instrument of positional astronomy, which is the study of the positions and motions of astronomical objects. (The specific categories of astronomy concerned with these investigations are astrometry and celestial mechanics.) The chief variants of the classic design include the vertical circle, the horizontal transit circle, the broken or prism transit, the photographic zenith tube and, most commonly, the meridian or transit circle. *See* ASTROMETRY; CELESTIAL MECHANICS.

The astronomical transit instrument was first developed by the Danish astronomer Ole Roemer in 1689. The modern transit instrument has an objective (lens) with a diameter of 6–10 in. (15–25 cm) and a focal length of 72–90 in. (180–230 cm). The instrument consists of a telescope mounted on a single fixed horizontal axis of rotation. The horizontal axis has a central hollow cube (or sometimes a sphere) and two conical semiaxes ending in cylindrical pivots. The objective and imaging halves of the telescope are also fastened to the cube of the instrument, perpendicular to the horizontal axis. Rotation of the instrument in its bearings, or wyes, permits the optical axis to sweep only in the plane of the meridian. Because it is constrained to a single plane, it can be solidly mounted on massive piers, creating a stability not present in other telescope designs. An accurate clock is the essential ancillary scale by which the transits of the astronomical objects are observed.

The astronomical transit instrument takes advantage of a special case of the astronomical triangle, which is composed of arcs of great circles on the celestial sphere. Its vertices are, respectively, the north celestial pole, the zenith point of the observer, and the celestial object under observation. The angle at the north celestial pole represents the hour angle of the object; hence, when the object is on the meridian, the hour angle is zero, and the triangle degenerates to a single arc, a segment of the meridian. At that instant the local sidereal time equals the right ascension of the celestial object. *See* ASTRONOMICAL COORDINATE SYSTEMS; CELESTIAL SPHERE; SIDEREAL TIME.

Meridian or transit circle. For many years, the major observatories of the world had astronomical transit instruments called meridian or transit circles. These instruments are similar to the transit instrument described above except that they are also capable of measuring the distance of the object along the meridian to obtain its declination. The addition of a large, accurately calibrated circle attached to the horizontal axis allows large angles to be measured, and serves as the scale for measuring declination. The circle originally was read with a micrometer and later photographed to be measured with a special measuring engine. Most circles are now read with some form of charge-coupled-device (CCD) camera. An accurate (atomic) clock is used as the scale for determining the right ascension of the object. *See* ATOMIC CLOCK.

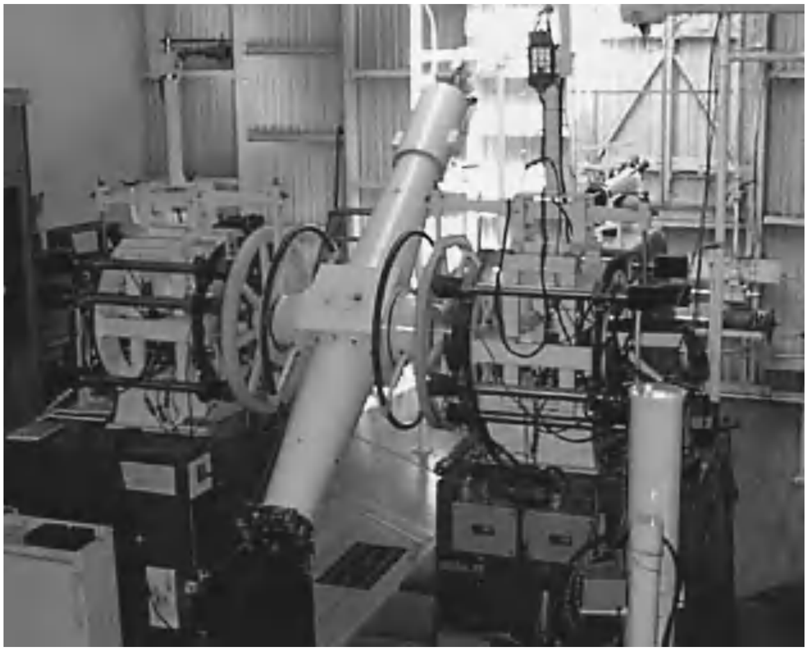


Fig. 1. Six-inch (15-cm) transit circle, U.S. Naval Observatory. (*Official U. S. Naval Observatory photograph*)

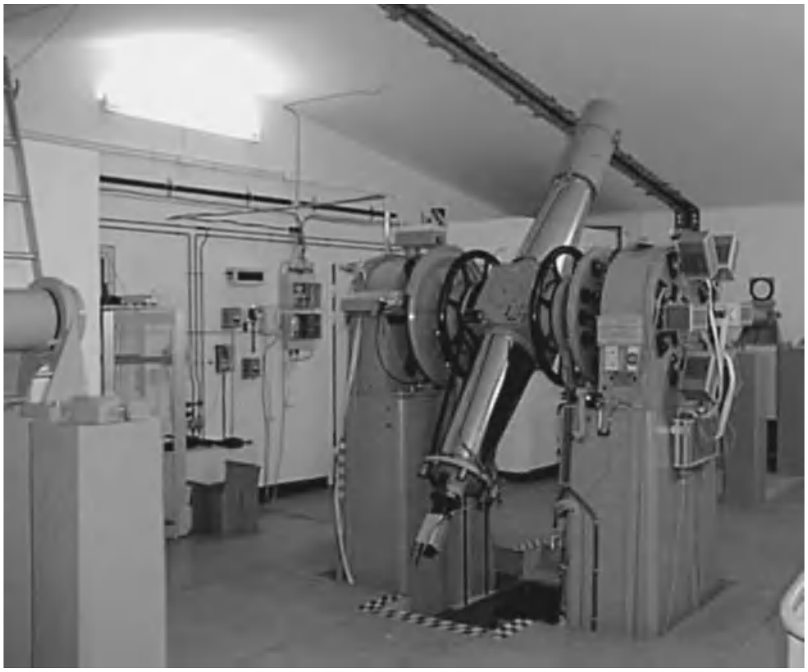


Fig. 2. Carlsberg Meridian Circle, La Palma, Canary Islands. (*University of Copenhagen*)

Traditionally, a very high precision micrometer at the eyepiece end of the instrument contained a movable wire, or pair of wires, and a stationary grid of vertical wires in the x and y directions for use in registering the transit and the declination. The micrometer and eyepiece have now been replaced with electronic detection systems, most commonly charge-coupled devices. While such systems have improved the accuracy and efficiency of the observations, they cannot be used far from the Equator due to the curved paths that the stars follow across the charge-coupled devices, and it is difficult to observe objects that are not point sources, such as the Sun and planets. *See* CHARGE-COUPLED DEVICES.

The 6-in. (15-cm) transit circle of the U.S. Naval Observatory was designed and constructed at the end of the nineteenth century, but it was improved continuously with the latest technological developments (**Fig. 1**). It was the first transit circle to have an electronic circle and the first to have the micrometer data read directly into an electronic computer. An atomic clock provided the computer with the signals to synchronize the right ascension readings from the micrometer. The divided glass circle was scanned with six charge-coupled-device cameras. The probable error of a single set of circle readings was 0.07″. However, the declination of a star may contain uncertainties of 0.25–0.50″ due to errors from other sources, such as atmospheric refraction, mechanical flexure of the instrument, and residual errors in the divided circle. The probable error of a single observation of the right ascension data of an equatorial star was about 0.012 s. When combined with other measurements to form a catalog, the positions have a probable error of around 0.05 second of arc. The introductions to the fundamental catalogs produced with this instrument contain complete descriptions of the equipment, methods, and results.

The 6-in. transit circle was used to observe not only the brighter stars (as faint as ninth magnitude) but also the Sun, Moon, planets, and several of the brighter asteroids. After 100 years of observations, this instrument was taken out of service.

Corrections. It is extremely difficult to adjust the instrument to the point of perfection, where a central charge-coupled-device column will trace the true meridian as the instrument is rotated on its pivots; therefore, corrections must be determined and applied to the observational data. The three principal instrumental errors that require correction are azimuth, collimation, and level. The azimuth correction is the horizontal angle between the axis of rotation and the true east-west direction and is obtained by observing circumpolar stars. The collimation correction is the angle between the line from the optical center of the telescope objective to the center of the charge-coupled device and the plane perpendicular to the horizontal axis of rotation, and is measured using small horizontal telscopes called collimators. The level correction is the angle that the axis of rotation makes with the plane of the horizon. In addition, a nadir value is obtained that is the zero point of the declination scale. The level and nadir values are obtained by observing a reflection of the focal plane in a basin of mercury.

These instrumental errors vary hourly, daily, and seasonally and, in many cases, are reproducible over these time periods. The main contributors to these errors are the variation in temperature within the pavilion and around the opening in the roof and air strata close to the building caused by wind patterns. Using computers, the values of the temperature, pressure, and humidity can be sampled, and many of these errors can be modeled. The instrument is carefully designed with modern optics and metallurgy to attempt to minimize these effects.

Another error, the clock correction, represents the error between the true sidereal time that the star should transit the local meridian and the time of transit recorded by the local sidereal clock. At one time the clocks had error rates that were significant enough to warrant a correction; however, today they have sufficient accuracy that this correction is considered as a constant for each night's work. This correction is evaluated through observations on bright stars whose positions are already well known.

Applications. The classical astronomical transit instrument had three interrelated uses: (1) from a known position on the Earth, observations of transits of stars led to the determination of their right ascension and declination with respect to the astronomical coordinate system; (2) with a knowledge of the positions of the stars observed and of Greenwich time, the longitude of the observer could be computed from observations of the time of transit of a star; and (3) the determination of corrections to the clock were made by the observation of stars of known position with an instrument situated at a known longitude. As the accuracy of clocks improved (and, specifically, with the development of atomic clocks), the third item fell out of use. To produce a fundamental star catalog, it was necessary to observe solar system objects at the same time as the stars. *See* LATITUDE AND LONGITUDE.

In 1997 the European Space Agency published the results of observations made with the *Hipparcos* satellite. The accuracy of the 100,000 star positions far exceeds anything obtainable with transit circles. In addition, modern astrographs with charge-coupled-device cameras can measure far more star positions than the transit circle and with greater accuracy. The development of mathematical techniques involving overlapping charge-coupled-device frames globally interconnected using powerful computers extends the narrow astrograph field to wide angles. Nevertheless, transit circles still carry on specialized observing programs. The Carlsburg Meridian Telescope at La Palma in the Canary Islands is an example of a modern transit circle used to provide additional accurate star positions to improve the proper motions of previously measured stars and for the measurement of solar system bodies, especially asteroids (**Fig. 2**). It is completely automatic, being operated remotely via the Internet. Use of a charge-coupled device as the detector generally limits these instruments to areas near the Equator. *See* ASTEROID.

A transit circle with collimators and azimuth marks is capable of determining positions of stars over wide angles in a fundamental reference system determined by making concurrent observations of solar system bodies. With the International Celestial Reference Frame (ICRF), based on radio observations of quasars and realized in the optical wavelengths by the *Hipparcos Catalogue*, replacing the fundamental reference stars, as realized by the FK4 and FK5, the transit circle is no longer the determining instrument for the stellar reference frame, and is mostly used now for small, specialized programs. *See* ASTRONOMICAL CATALOGS; CELESTIAL REFERENCE SYSTEMS.

F. Stephen Gauss

Bibliography. H. Eichhorn, *Astronomy of Star Positions*, Dover, 1974; R. Green, *Spherical Astronomy*, Cambridge University Press, 1985; H. C. King, *The History of the Telescope*, 1955, reprint, Dover, 2003; G. P. Kuiper and B. M. Middlehurst (eds.), *Telescopes: Stars and Stellar Systems*, vol. 1, University of Chicago Press, 1960; T. Page and L. W. Page, *Telescopes*, Macmillan, 1966; P. van de Kamp, *Principles of Astrometry*, W. H. Freeman, 1967.

Astronomical unit

The basic unit of length in the solar system. The astronomical unit (AU) is also used to a limited extent for interstellar distances through the definition of the parsec (1 pc = 206,265 AU). It is nearly equal to the mean distance a between the center of mass of the Sun and the center of mass of the Earth-Moon system ($a = 1.00\,000\,23$ AU), and for that reason it is often convenient to think of it as the mean distance between the Sun and Earth. *See* PARSEC.

The most accurate determination of the length of the astronomical unit in physical units, such as meters, is obtained from phase-modulated continuous-wave (CW) radio signals beamed to other planets. The round-trip travel times of the signals are determined by cross-correlating the returned signal from the planet with the transmitted signal, and as a result, planetary distances are measured directly.

Prior to the inception of planetary radar technology around 1961, determinations of the astronomical unit relied on parallax effects in the positions of planets and asteroids as viewed from distant points on the surface of the Earth, or on spectroscopic Doppler shifts in the light of stars near the plane of the ecliptic. These techniques were limited in accuracy to about 43,000 mi (70,000 km). On the other hand, early radar determinations were accurate to 300 mi (500 km), and by accumulating several years of planetary distance measurements, it was possible to refine this accuracy to 0.6 mi (1 km), the limit of passive radar bounce technology. Continuous-wave signals returned from the National Aeronautics and Space Administration (NASA) landers on the surface of Mars

-s

resulted in a determination of the astronomical unit to an accuracy of 20 m (65 ft). *See* DOPPLER EFFECT; EARTH ROTATION AND ORBITAL MOTION; PARALLAX (ASTRONOMY).

With an adopted value of 299,792,458 m/s (186,282.397051 mi/s) for the speed of electromagnetic propagation in vacuum, the value of the astronomical unit from NASA's Viking and Pathfinder missions is 149,597,870,692 m (92,955,807.268 mi).

John D. Anderson

Bibliography. D. W. Kurtz and G. E. Bromage (eds.), *Transits of Venus: New Views of the Solar System and Galaxy*, IAU Colloquium 196, 2005; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, 1992.

Astronomy

The study of the universe and the objects in it through scientific investigation. Since much of contemporary astronomy uses the laws and methods of physics, the terms "astronomy" and "astrophysics" are usually used interchangeably. However, modern astronomy also uses techniques from many other scientific disciplines, including chemistry, geology, and biology, for which the terms astrochemistry, planetary science, and astrobiology are increasingly used.

The goal of astronomical research is to understand celestial objects and the nature and evolution of the universe by using whatever techniques are appropriate. Astronomers' breadth of knowledge of different types of astronomical objects often gives insights into understanding any given object or physical process.

Scope. Since the advent of spectroscopy in the late nineteenth century, allowing the remote sensing and analysis of the composition and motion of distant objects, the distinction between astronomy and astrophysics has largely evaporated. The view that astronomers study objects in the sky and astrophysicists seek to explain them has been superseded. Some branches of astronomy that are not strictly included in astrophysics involve the study of the motions of the planets and satellites as well as of spacecraft launched from Earth (celestial mechanics) and the measurement of positions and motions of stars (astrometry). However, the extreme astrometric accuracy of the *Hipparcos* space mission in the 1990s gave data that were quickly applied to fundamental distance determinations that are used in assessing the expansion of the universe and other cosmological parameters. So even the traditional fields of astronomy are now so intertwined with astrophysics that no strict distinction is possible. *See* ASTROMETRY; CELESTIAL MECHANICS.

The use of geological knowledge and methods in analyzing close-up observations from spacecraft of planets and their satellites and of comets and asteroids closely links the disciplines of astronomy and planetary science. Indeed, the discovery of planets around distant stars holds for even closer relations in the future. Methods of studying molecules in interstellar clouds involve chemical knowledge. Planetary science and astrochemistry come together with astronomy in the search for life outside the solar system, part of the search for extraterrestrial intelligence (SETI). The National Aeronautics and Space Administration (NASA), the United States space agency, has placed a priority on astrobiology, including the investigation of Mars and bringing samples back to Earth. *See* ASTEROID; ASTROBIOLOGY; COMET; EXTRATERRESTRIAL INTELLIGENCE; INTERSTELLAR MATTER; PLANET; SATELLITE (ASTRONOMY); SOLAR SYSTEM.

Technologies. Astronomers often lead in employing new technologies, pushing them to the limit in exploring extremely faint signals in various parts of the electromagnetic spectrum. Nearly all astronomical research is now heavily dependent on computers, which control telescopes, help analyze data, make theoretical calculations, display results, provide Internet links for e-mail and the World Wide Web, and allow efficiencies through word processing, image processing, and statistical analysis programs. Even the shapes of telescope mirrors of some of the newest and largest instruments are controlled through computer feedback loops. *See* DIGITAL COMPUTER; IMAGE PROCESSING.

Astronomical imagery is now dominated by light-sensitive silicon chips known as charge-coupled devices (CCDs). In these devices, electric charge results when a pixel is struck by light, and this charge is then transferred to adjacent pixels in the process of being read out. CCDs are approximately 100 times more sensitive than film, and are used at all professional observatories as well as by increasing numbers of students and amateur astronomers. Thus an image or spectrum that formerly took over an hour to record on film can now be recorded in about a minute, with the added benefit over film that the law of reciprocity holds for the data; that is, the strength of the electronic signal produced is directly and linearly proportional to the intensity of the incoming signal. CCDs are sensitive throughout the optical and near-infrared portions of the electromagnetic spectrum. Other electronic devices are used in other parts of the spectrum to record incoming signals. *See* ASTRONOMICAL IMAGING; CHARGE-COUPLED DEVICES.

Fiber optics are used for a variety of astronomical purposes, including the taking of hundreds of galaxy images simultaneously from the field of view of a telescope and bringing the light to a spectrograph that can produce simultaneous spectra of all the objects. This multiplexing effect can increase the efficiency of sky surveys many times. *See* FIBER-OPTICS IMAGING.

The technology of adaptive optics, in which the shape of a mirror is changed slightly at a high rate (often faster than 1 Hz) to compensate for the blurring of astronomical images caused by the Earth's atmosphere, and for slower active optics, was developed largely in secret for military purposes. It has now been released to the public and is being increasingly pursued to eliminate the twinkling of stars. *See* ADAPTIVE OPTICS.

Optical telescopes. The opening of the 5-m (200-in.) Hale telescope at the Palomar Observatory on Palomar Mountain, California, in 1948 marked the beginning of a great period of development in optical astronomy. The light-gathering power of this telescope allowed cosmological study that extended most of the way to the beginning of time in the universe. It was joined in the task by several 4-m-class (160-in.) telescopes and by one less successful larger telescope. Apparently, the maximum size of such traditional telescopes had been reached.

In the 1990s, new techniques of telescope making allowed the completion of several telescopes in the 10-m (400-in.) class, twice the diameter and thus four times the collecting area of the Hale telescope. The first to be completed were the twin telescopes at the W. M. Keck Observatory on Mauna Kea in Hawaii. At this high site (about 4200 m or 13,800 ft in altitude) with excellent steady air (giving steady images known as good seeing), low amounts of water vapor overhead leading to good infrared transparency, and dark and usually clear skies are located several of the largest telescopes in the world. In 1999 the 8-m (320-in.) Subaru ("Pleiades") telescope of the Japanese National Observatory and the 8-m Gemini North telescope of an international consortium (with 50% U.S. participation) were established there. In the Southern Hemisphere, in Chile, better placed to observe various southern objects including the center of the Milky Way Galaxy, the four 8-m telescopes of the European Southern Observatory's Very Large Telescope (VLT) began operation. These telescopes, often operated independently, are also operated together as interferometers, as are the Keck telescopes, giving increased resolution. These techniques are rapidly improving. The Gemini South 8-m telescope also operates in Chile. An even larger telescope, though with limited pointing ability in the sky, is operated in Texas for spectroscopic use, and a similar spectroscopic telescope has been built in South Africa.

Still larger telescopes are in the conceptual design stage, though without definitive funding or construction plans. They include the 30-m (100-ft) California Extremely Large Telescope (CELT) and the Overwhelmingly Large Telescope (OWL).

The existing large telescopes have proven useful in taking spectra of the optical counterparts of gamma-ray bursts, proving that they are very far away; and in analyzing the distances to faraway galaxies and in measuring the redshifts of their spectra, leading to the current cosmological models of the expansion of the universe and the conclusion that the rate of expansion is accelerating. *See* COSMOLOGY; GAMMA-RAY BURSTS; HUBBLE CONSTANT; TELESCOPE.

Telescopes in space. The 1990s saw the thorough use of the vantage points of space for astronomical observation, exemplified by NASA's series of Great Observatories. In 1991 the *Compton Gamma-Ray Observatory* was launched, and in the following years mapped about one gamma-ray burst per day in addition to many other objects and events. The *Hubble Space Telescope* was launched in 1990 to study the ultraviolet and visible parts of the spectrum. Its repair in 1993, with secondary mirrors compensating for a focusing problem with the main mirror, brought it to full working order, and a 1996 upgrade included an improved two-dimensional spectrograph and infrared capability. The *Chandra X-Ray Observatory*, launched in 1999, provides high-resolution x-ray images, and is the same size and scope as Hubble. It studies various types of celestial objects and processes, such as black holes of stellar and galactic sizes. The *Spitzer Space Telescope*, launched in 2003, formerly the *Space Infrared Telescope Facility*, is the fourth of this series of Great Observatories. The European Space Agency's XMM-Newton x-ray telescope and INTEGRAL gamma-ray telescope are also important, large-scale spacecraft. *See* ASTROPHYSICS, HIGH-ENERGY; BLACK HOLE; GAMMA-RAY ASTRONOMY; HUBBLE SPACE TELESCOPE; INFRARED ASTRONOMY; X-RAY ASTRONOMY; X-RAY TELESCOPE.

Smaller spacecraft have also made valuable contributions. The *Cosmic Background Explorer* (*COBE*) in the early 1990s mapped the cosmic background radiation and provided important clues to the origin of the universe and its large-scale structure. It has been superseded by NASA's *Wilkinson Microwave Anisotropy Probe*. NASA's Explorer series continues with a variety of mid-sized spacecraft. The European Space Agency's *Solar and Heliospheric Observatory* (*SOHO*) and *Infrared Space Observatory* (*ISO*) have also sent back valuable observations. *See* COSMIC BACKGROUND RADIATION; SUN.

Telescopes across the spectrum. The atmosphere blocks most of the electromagnetic spectrum from reaching the Earth's surface, leaving windows of transparency mostly in the optical and radio parts of the spectrum. Radio astronomers have made the most of their window of transparency with such telescopes as the 100-m (328-ft) fully steerable telescope outside Bonn, Germany; the 330-m (1083-ft) Arecibo dish in Puerto Rico, in which the secondary mirror can be moved over a spherical bowl covered with mesh to give some limited tracking ability; and the precise, 100-m × 110-m (300-ft × 360-ft) Robert C. Byrd Green Bank Telescope in West Virginia, opened in 2000, for example, to allow studies of interstellar molecules. Radio astronomers pioneered the use of interferometry, taking advantage of the long wavelengths at which they observe. The Very Large Array (VLA) of radio telescopes of the National Radio Astronomy Observatory extends about 25 km (15 mi) over a plain in New Mexico, synthesizing the aperture to provide the resolution (though not the collecting area) of a single telescope of that size. The Very Long Baseline Array uses a dedicated set of telescopes on various sites across the globe to synthesize a telescope the size of the Earth. In India, the Giant Metrewave Radio Telescope near Pune has 30 antennas each 45 m (150 ft) across; larger than the VLA, it is more sensitive at long wavelengths but cannot image at the shorter wavelengths at which the VLA often observes. *See* RADIO ASTRONOMY; RADIO TELESCOPE.

The ozone layer and other constituents of the atmosphere block the shortest wavelengths from penetrating to the Earth's surface, so observations of gamma rays, x-rays, and most of the ultraviolet region require telescopes in space. Similarly, water vapor blocks most of the infrared, allowing only a few windows of transparency, and observations from balloons and, to go higher, from spacecraft are necessary to cover the whole infrared spectrum. *See* OZONE; ULTRAVIOLET ASTRONOMY.

Spectroscopy. Much of astronomy involves breaking down the incoming celestial radiation into its component wavelengths, a process known as spectroscopy. The brighter the object, the finer the spectroscopic resolution in wavelength that can be obtained. Spectroscopic studies can reveal the temperature of an object, the identity and proportions of its chemical elements, and the velocities of its constituents toward and away from the Earth. Light from the Sun and other objects is sometimes polarized, and studies of such polarization can give information about the magnetic fields present or about scattering processes. *See* ASTRONOMICAL SPECTROSCOPY; POLARIMETRY.

Non-electromagnetic-radiation telescopes. Though most of what astronomers study is electromagnetic radiation in its various forms, some particles do arrive at the Earth. The expansive definition of a telescope includes anything used in astronomy to observe the sky. Several neutrino telescopes have been used to detect neutrinos from the Sun and, in one instance, from a supernova. The original neutrino telescope was a large tank of cleaning fluid based in a mine far underground. Subsequent neutrino telescopes are tanks of gallium and even larger tanks of purified water, in which interactions with neutrinos have detectable consequences. These neutrino telescopes have pinpointed the "solar neutrino problem" as arising from problems in understanding the fundamental physics of elementary particles rather than from uncertainties or errors in the astronomical measurements. *See* NEUTRINO; NEUTRINO ASTRONOMY; SOLAR NEUTRINOS.

Cosmic rays are particles from the Sun and from more distant objects in space. These primary cosmic rays interact with particles in the Earth's atmosphere to make secondary cosmic rays, and the pace of observation of these secondary cosmic rays as well as the few primary cosmic rays that reach the Earth is increasing. *See* COSMIC RAYS.

Gravitational waves are a consequence of Einstein's general theory of relativity, and their existence was verified indirectly by careful study of the period of a binary pulsar. A pair of interferometers was built on Earth to attempt direct detection of such gravitational waves, which should result from such distant events as the merger of two neutron stars. This Laser Interferometer Gravitational-wave Observatory (LIGO) is in operation in Louisiana and in Washington but is in constant modification to improve its sensitivity, which is currently not at the level necessary to expect discoveries. Similar facilities are in operation and in construction at other locations around the world. *See* GRAVITATIONAL RADIATION.

Theory and computation. Theoretical calculations of the nature of astronomical objects or processes are known as theoretical astrophysics. Theoretical astrophysicists analyze the evolution of stars, analyze the solar wind of expanding plasma around the Sun, trace the explosion phase of supernovae, calculate the distribution of sizes of structures formed in the early eons of the universe, or analyze the formation of the light elements in the first minutes after the big bang. Cosmological calculations use Einstein's general theory of relativity. The discovery of the dark energy that makes up about 70% the content of the universe has led to a wide-ranging set of new cosmological ideas. *See* BIG BANG THEORY; PLASMA (PHYSICS); RELATIVITY; SOLAR WIND; STAR; STELLAR EVOLUTION; SUPERNOVA.

The availability of supercomputers, powerful and fast computers capable of handling large amounts of data, has led to three-dimensional simulations of, for example, the effects of varying amounts of cold dark matter (such as undiscovered types of subatomic particles) and hot dark matter (such as neutrinos) in the formation of large-scale structure in the early universe. Such dark matter, making up approximately 30% the contents of the universe, apparently exceeds the amount of matter detectable through more traditional methods of observation. From a variety of observations, including those of distant supernovae and those of the cosmic background radiation from the *Wilkinson Microwave Anisotropy Probe*, it has been found that the bulk of the contents of the universe, about 70%, are of dark energy. This dark energy may be in the form of the cosmological constant, energy transformable into mass in an amount calculable with Einstein's equation. *See* SIMULATION; SUPERCOMPUTER; UNIVERSE.

Models of the oscillations detectable on the Sun's surface through long-time-series observations made on the ground from the Global Oscillation Network Group (GONG) or from space with instruments on the *Solar and Heliospheric Observatory* (*SOHO*) are used to improve understanding of the solar interior, a process known as helioseismology. *See* HELIOSEISMOLOGY.

Experimental astrophysics. Laboratory astrophysics involves the measurement of basic parameters that are used in calculations of physical or chemical processes relevant to astronomy. Though the Sun and stars as well as interstellar space are excellent laboratories for studying processes at high temperature or in vacuums unobtainable on Earth, sometimes fundamental parameters such as cross sections of atomic and molecular collisional excitation and ionization, as well as nuclear parameters, can be measured in laboratories on Earth. *See* ATOMIC STRUCTURE AND SPECTRA; MOLECULAR STRUCTURE AND SPECTRA.

Historical astronomy. Studying astronomy's history provides data for comparison with the present. For example, the paths of solar eclipses from hundreds or thousands of years ago can be analyzed to study the rate of rotation of the Earth and the relation between astronomical timekeeping and timekeeping by atomic methods, leading to an improved understanding of processes inside the Earth. *See* ARCHEOASTRONOMY; DAY; EARTH ROTATION AND ORBITAL MOTION; ECLIPSE; TIME; YEAR. Jay M. Pasachoff

Bibliography. G. R. Burbidge et al. (eds.), *Annual Review of Astronomy and Astrophysics*, Annual Reviews; M. Cassé (transl. by S. Lyle), *Stellar Alchemy: The Celestial Origin of Atoms*, Cambridge University Press, 2003; K. Croswell, *The Universe at Midnight*, Free Press, 2001; M. Harwit, *Astrophysical Concepts*, 3d ed., Springer, 1998; M. A. Hoskin, *The Cambridge Concise History of Astronomy*, Cambridge University Press, 1999; D. Leverington, *A History of Astronomy: From 1890 to the Present*, Springer, 1995; J. M. Pasachoff, *Astronomy: From the Earth to the Universe*, 6th ed., Brooks/Cole, 2002; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millennium*, 2d ed., Brooks/Cole, 2004, 3d ed., 2007.

Astrophysics, high-energy

The study of the universe as revealed by high-energy, invisible forms of light: x-rays and gamma rays. These radiations are produced in the cosmos when gas is heated to millions of degrees Kelvin or electrons have been accelerated to near the speed of light by violent and extreme conditions. Exploding stars, neutron stars, black holes, and galaxy clusters, the most massive objects in the universe, are among the objects studied.

Instrumentation. The high energies of x-rays and gamma rays have two important consequences for astronomical research. First, these forms of light are absorbed by the atmosphere, so telescopes to detect them must be placed on spacecraft above the atmosphere. Second, the telescopes must be constructed differently. Gamma rays have such high energy that they cannot be focused by traditional techniques, although indirect methods can give a rough estimate of their direction. *See* SATELLITE ASTRONOMY.

X-rays will reflect off mirrors, but only if they strike at grazing angles, like a stone skipping across a pond. For this reason, x-ray mirrors have to be carefully shaped and aligned nearly parallel to the incoming x-rays. These barrel-shaped mirrors are nested one inside the other to increase the collection area, and therefore the sensitivity, of the telescope.

The *Chandra X-ray Observatory*, launched by the National Aeronautics and Space Administration (NASA) in July 1999, is the premier focusing x-ray telescope. It is an assembly of four pairs of mirrors. Chandra's mirrors are the smoothest mirrors ever constructed. The largest of the mirrors is almost 4 feet (1.2 m) in diameter and 3 ft (0.9 m) long. *See* X-RAY TELESCOPE.

The European Space Agency's *XMM*, a powerful telescope launched in December 1999, has 58 mirrors. These mirrors are not as smooth as *Chandra*'s mirrors, so *XMM* cannot make images of the same crispness, but it can detect fainter sources and measure the energies of x-rays very accurately.

NASA's *Rossi X-ray Timing Explorer* (*RXTE*), launched in December 1995, has the ability to study changes in the intensity of x-rays produced in the violent environment around neutron stars and black holes on time scales ranging from microseconds to months. Two other NASA missions involving international cooperation, the *High Energy Transient Explorer* (*HETE-2*), launched in 2001, and the *Swift* satellite, launched in 2004, are dedicated to the exploration of short-lived bursts of x-rays and gamma rays.

NASA's *Compton Gamma-Ray Observatory*, which inaugurated a new era in gamma-ray astronomy with its unprecedented sensitivity and coverage of a large range of gamma-ray energies, was deorbited and reentered the Earth's atmosphere on June 4, 2000. The *Compton Observatory* has been succeeded by the European Space Agency's *International Gamma-Ray Astrophysics Laboratory* (*INTEGRAL*), launched in October 2002, an observatory with a complement of gamma-ray, x-ray, and optical telescopes. Several new telescopes are designed to observe flashes of visible light created high in Earth's atmosphere when extremely high energy (teraelectronvolt or TeV) gamma rays hit the upper atmosphere. For example, the *High Energy Spectroscopic System* (*HESS*), an array of four 12-m (39-ft) mirrors on a mountain in central Namibia, has made a rough image of TeV gamma rays from the remnant of a supernova explosion. *See* CERENKOV RADIATION; GAMMA-RAY ASTRONOMY.

Supernovae. When a massive star (ten or more times as massive as the Sun) has used up the nuclear fuel that makes it shine, the pressure drops in the central core of the star. Gravity crushes the matter in the core to higher and higher densities. Temperatures soar to billions of degrees Kelvin. The intense heat generated in the collapse produces a cataclysmic rebound that sends high-speed debris flying outward at speeds in excess of 5000 mi/s (8000 km/s). A thermonuclear shock wave races through the now expanding stellar debris, fusing lighter elements into heavier ones and producing a brilliant visual outburst with the brightness of several hundred million suns.

A massive star explodes about once every 50 years in the Milky Way Galaxy. The shell of matter thrown off by the supernova creates a magnetized bubble of multimillion-degree gas and high-energy particles called a supernova remnant. The hot gas expands and produces x-rays for thousands of years (**Fig. 1**). Gamma rays from radioactive elements have also been detected from supernova remnants by gamma-ray telescopes such as those that were on the *Compton Gamma-Ray Observatory*.

The study of remnants of exploded stars, or supernovae, is essential for understanding the origin of life on Earth. The cloud of gas and dust that collapsed to form the Sun, Earth, and other planets was composed mostly of hydrogen and helium, with a small amount of heavier elements such as carbon, nitrogen, oxygen, and iron. The only place where these and other

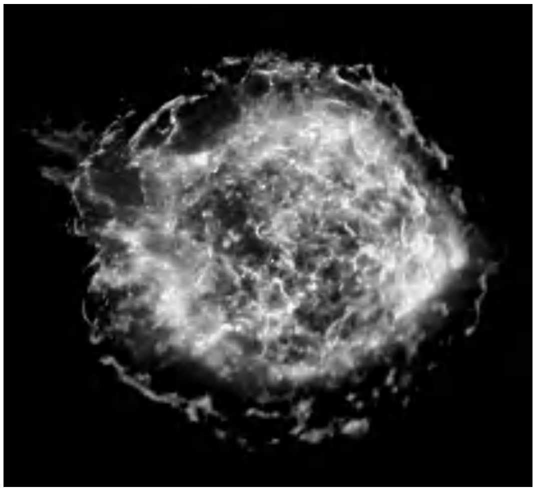


Fig. 1. ***Chandra X-ray Observatory*** **image of Cassiopeia A (Cas A), the 320-year-old remnant of a massive star that exploded. The image shows an expanding shell of hot gas produced by the explosion. This gaseous shell is about 10 light-years in diameter and has a temperature of about 5×10^7 Kelvin. (*NASA/Chandra X-ray Observatory Center/Goddard Space Flight Center/U. Hwang et al.*)**

heavy elements necessary for life are made is deep in the interior of a massive star. There they remain until a supernova explosion, spreads them throughout space. *See* NUCLEOSYNTHESIS; SUPERNOVA.

Neutron stars. When a massive star explodes, most of it is flung into space, but the core of the star is compressed to form a rapidly rotating dense ball of neutrons that is about 12 mi (20 km) in diameter. The collapse and rapid rotation of the neutron star cause it to become highly magnetized. A magnetized, rapidly rotating neutron star can produce electric voltages of 10^{16} V.

Neutron star gravity, which is more than 10^{11} times stronger than gravity on Earth, is overwhelmed by the electric field, and particles are pulled off the neutron star and accelerated to speeds near the speed of light. An intense shower of electrons and antimatter electrons, or positrons, is produced by these particles. The pulsed emission from the Crab Nebula, observed at all wavelengths from radio through gamma rays, is thought to be caused by this process (**Fig. 2**). *See* CRAB NEBULA; PULSAR.

As particles stream out from the pulsar and spiral around magnetic field lines, they produce a distinctive kind of radiation known as synchrotron radiation. The Crab Nebula's bell-shaped appearance in the x-ray image is due to synchrotron radiation from a huge magnetized bubble of high-energy electrons that is several light-years in diameter. Dozens of these so-called pulsar wind nebulae have been discovered by the *Chandra X-ray Observatory*. *See* SYNCHROTRON RADIATION.

The rotation-powered activity of neutron stars such as the Crab pulsar can last only a few thousand years. However, if the neutron star has a nearby companion star, its x-ray intensity may increase again in a million years or so. When the companion star enters the red giant stage of its life, it will increase greatly in size and gas will flow from the giant star onto the neutron star. The gas will be heated to tens of millions of degrees Kelvin as it falls onto the surface of the neutron star, and will glow brightly in x-rays.

The Milky Way Galaxy contains several hundred of these neutron star x-ray binaries. Depending on the details of the rate at which the matter falls onto the neutron star, and how the magnetic field of the neutron star guides the inflow, the star will be observed to pulse, flicker, or flare up violently in x-rays. X-ray binaries provide a unique opportunity to study neutron stars. A similar process allows the study of even stranger objects, black holes. *See* BINARY STAR; NEUTRON STAR.

Black holes and quasars. When some very massive stars collapse, they will form black holes. A black hole does not have a surface in the usual sense of the word. There is simply a region in space around a black hole beyond which nothing can be seen, because nothing can escape from inside this region. This region is called the event horizon.

Anything that passes beyond the event horizon is doomed to be crushed as it descends ever deeper into the gravitational well of the black hole. Neither visible light, nor x-rays, nor any other form of electromagnetic radiation given off by the particle can escape.

A black hole cannot be seen directly. One way to find one is by observing the energy released by matter that is falling toward the black hole. As gas and dust particles swirl toward a black hole, they speed up and form a flattened disk. Friction caused by collisions between the particles heats them to

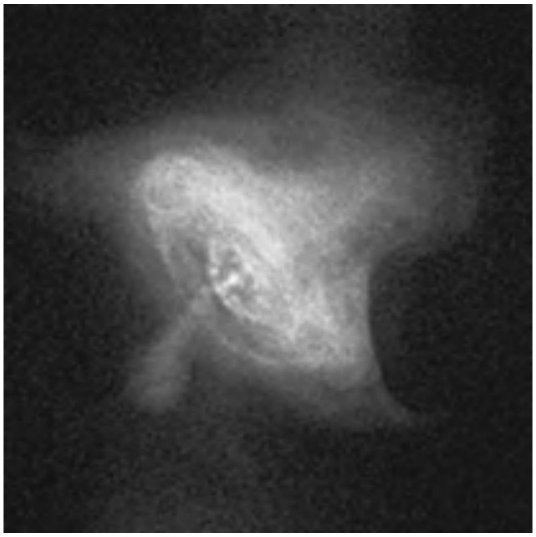


Fig. 2. ***Chandra X-ray Observatory*** **image of the Crab Nebula, a supernova remnant and pulsar in the constellation Taurus. The image shows the central pulsar, a rapidly spinning neutron star, or pulsar that emits pulses of radiation 30 times a second, surrounded by tilted rings of high-energy particles that appear to have been flung outward over a distance of more than a light-year from the pulsar. (*NASA/Chandra X-ray Observatory Center/ Smithsonian Astrophysical Observatory*)**

extreme temperatures. Just before the particles pass beyond the event horizon, they produce x-rays and gamma rays as their temperatures approach 10^8 Kelvin.

One of the best methods for finding a black hole has been to study x-ray binary systems. Although most of these systems consist of a normal star in a close orbit around a neutron star, there are about 20 cases in which the observations indicate that the mass of the invisible companion star is so great—more than three times the mass of the Sun—that it must be a black hole.

Neutron stars or black holes may be the explanation for one of the most important unsolved problems in astrophysics, gamma-ray bursts. As determined by the BATSE detector that was on the *Compton Gamma-Ray Observatory* and the Italian-Dutch satellite *Beppo-Sax*, these mysterious outbursts are observed to occur about once per day. In about a minute, gamma-ray bursts release as much energy as the Sun will give off at all wavelengths in its lifetime of 10^{10} years. There appear to be at least two types of gamma-ray bursts. Short bursts are likely produced by the merger of two neutron stars, or a neutron star and a black hole. Long bursts represent the explosions of extremely massive stars, greater than 50 times the mass of the Sun. According to the theory, a massive black hole forms in the center of the star just before the explosion. As matter in the center of the star pours into the black hole, a titanic explosion occurs, ejecting matter outward at nearly the speed of light. *See* GAMMA-RAY BURSTS.

Black holes grow when matter falls into them. A black hole in the center of a galaxy where stars are densely packed may grow to the mass of 10^9 suns. Energy released from large clouds of gas as they fall into these supermassive black holes can be stupendous. This is the accepted explanation for quasars, sources in which the power output at the center of a galaxy can be a thousand times greater than an entire galaxy of 10^{11} stars. *See* QUASAR.

One of the most intriguing features of supermassive black holes is that they do not suck up all the matter that falls within their sphere of influence. Some of the matter falls inexorably toward the black hole, and some explodes away from the black hole in high-energy jets that move at near the speed of light (**Fig. 3**). These jets produce radio, optical, x-ray, and gamma radiation. The matter swirling around the black hole must somehow be producing enormous electric and magnetic fields that accelerate electrons to extremely high energies. Exactly how this happens is unknown and is a major focus of research. *See* BLACK HOLE.

Galaxy clusters and dark matter. More than half of all galaxies in the universe are members of groups of galaxies or larger collections of galaxies, called clusters. X-ray observations have shown that most clusters of galaxies are filled with vast clouds of multimillion-degree gas. The mass of this gas, which was heated when it collapsed from a much larger size, is greater than all the stars in all the galaxies in a cluster of a thousand galaxies. Galaxy clusters are the largest and most massive gravitationally bound objects in the universe.

The x-ray-producing hot gas found in a typical cluster of galaxies presents a great mystery. Over time this extremely hot gas should escape the cluster, since the galaxies and gas do not provide enough gravity to hold it in. Yet the gas remains in clusters of all ages. Scientists have concluded that some unobserved form of matter, called dark matter, is providing the gravity needed to hold the hot gas in the cluster. An enormous amount of dark matter is needed—about three to ten times as much matter as that observed in the gas and galaxies. This means that most of the matter in the universe may be dark matter.

The candidate that best reproduces the observations is called cold dark matter—hypothetical subatomic particles that produce no light and can at present be detected only through gravity. Detailed measurements of the size and temperature of the hot gas clouds in galaxy clusters with x-ray telescopes could help solve the dark matter mystery. *See* DARK MATTER; GALAXY, EXTERNAL.

While an explanation for the source of dark matter is still lacking, an effect that is even more enigmatic has been discovered. Astronomers have observed that the visible light from type 1a supernovae, which act as standard candles, is fainter than expected in distant galaxies. The best explanation is that they are more distant than originally thought, which implies that the expansion of the universe must be accelerating. *Chandra's* measurements of the dark-matter content of clusters of galaxies have verified this astounding result by an independent method.

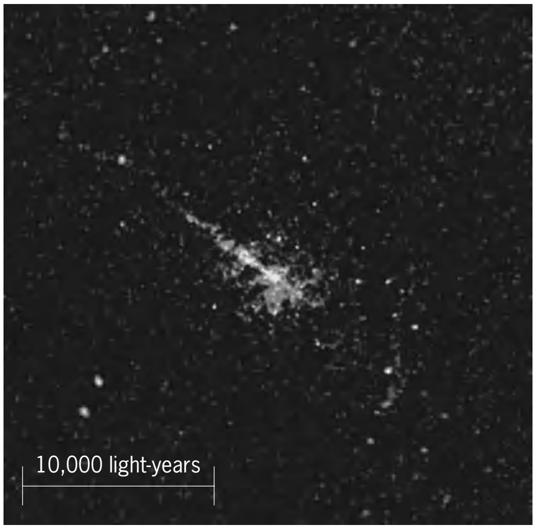


Fig. 3. ***Chandra X-ray Observatory*** **image of NGC 5128, a radio galaxy in the constellation Centaurus, 10^7 light-years from Earth. The image shows a bright source in the nucleus of the galaxy, which is thought to be due to a supermassive black hole. The jet extending to the upper left, far outside the galaxy, is caused by explosive activity around the black hole. (*NASA/Chandra X-ray Observatory Center/ Smithsonian Astrophysical Observatory*)**

Cosmic acceleration can be explained if the space between galaxies is filled with a mysterious dark energy that has the property that, as the universe expands, more dark energy is created. The existence of dark energy requires either a modification of Einstein's theory of general relativity or a major revision of some other area of fundamental physics. Assuming that dark energy is responsible for the acceleration, combining the *Chandra* results with observations of type 1a supernovae and the cosmic microwave background radiation indicates that dark energy makes up about 75% of the energy density of the universe, dark matter about 21%, and visible matter about 4%. *See* COSMIC BACKGROUND RADIATION; COSMOLOGY; DARK ENERGY; RELATIVITY; UNIVERSE; X-RAY ASTRONOMY.

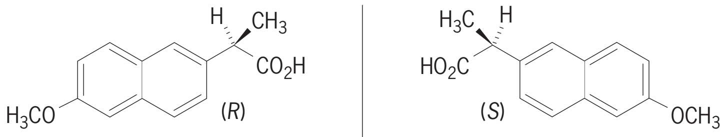
Wallace Tucker

Bibliography. M. Begelman and M. Rees, *Gravity's Fatal Attraction*, Scientific American Library, 1996; D. Berry, *Smithsonian Intimate Guide to the Cosmos*, Smithsonian/Madison Press, 2004; P. Charles and F. Seward, *Exploring the X-ray Universe*, Cambridge University Press, 1995; J. Pasachoff, *A Field Guide to the Stars and Planets*, 4th ed., Houghton Mifflin, 2002; G. Schilling, *Evolving Cosmos*, Cambridge University Press, 2004; E. Schlegel, *The Restless Universe*, Oxford University Press, 2002; J. Silk, *On the Shores of the Unknown: A Short History of the Universe*, Cambridge University Press, 2005; K. Thorne, *Black Holes and Time Warps*, W. W. Norton, 1994; W. Tucker and K. Tucker, *Revealing the Universe: The Making of the Chandra X-ray Observatory*, Harvard University Press, 2001; K. Weaver, *The Violent Universe: Joyrides through the X-ray Cosmos*, Johns Hopkins University Press, 2005.

Asymmetric synthesis

A chemical reaction or series of reactions leading to the predominant or exclusive formation of a single enantiomer, that is, one of an isomeric pair of chemical compounds whose molecules are nonsuperimposable mirror images. Enantiomers are said to be chiral. Many organic compounds of interest for pharmaceuticals, flavorings, and pesticides are chiral. A specific target of an asymmetric synthesis is a product that contains only the desired enantiomer rather than a mixture of enantiomers of chiral molecules. *See* MOLECULAR ISOMERISM.

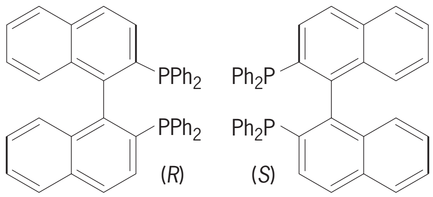
One of the most common structural features of many chiral molecules is the presence of a chiral center, a carbon atom in a molecule that has four different atoms or groups attached to it. For example, two enantiomers, (**1**) and (**2**), of naproxen exist and have the same constituents, but owing to the relative orientations of their substituents in space, these structures are not superimposable. In depicting three-dimensional structures, drawings use a wedge-shaped bond to the methyl group to indicate that it is above the plane of the naphthyl group and a dashed-line bond to indicate that it is below the plane. There are rules of nomenclature that unambiguously label the three-dimensional aspects of a center by a chirality descriptor, such that (**1**) is (*R*)-naproxen and (**2**) is (*S*)-naproxen. If a mirror plane is present in a molecule, it is not chiral and its mirror image will be identical with the original.



(**1**) (**2**)

mirror

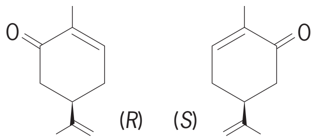
Although a chiral center is the most common source of chirality, one may also have chirality associated with an axis that can give rise to enantiomers. 2,2′-Bis(diphenylphosphino)-1,1′-binaphthyl (BINAP) is a common ligand used in transition-metal catalysis for synthesis of one enantiomer of a product, and BINAP can be obtained as the pure (*R*) or (*S*) enantiomer, (**3**) and (**4**).



(**3**) (**4**)

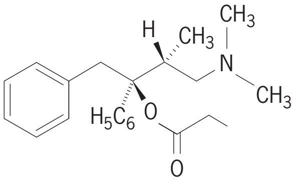
See STEREOCHEMISTRY.

Chirality in molecules is an important factor in determining their biological activity. One of the enantiomers of a chiral substance is often more effective than the other for applications in pharmacology, agriculture (pesticides or herbicides), or food chemistry (flavoring agents). Most complex molecules in a living organism are chiral and therefore interact with enantiomers of a given substance in different ways. This is true with regard to perception of odors where the enantiomers of carvone (**5** and **6**) have distinctively different fragrances: the (*R*) enantiomer smells like spearmint and the (*S*) enantiomer like caraway seeds. Pharmacological effects can also be pronounced; for example, (2*S*, 3*R*)-propoxyphene (**7**) is an analgesic, whereas (2*R*, 3*S*)-propoxyphene (**8**) is a cough suppressant. In other cases, one enantiomer can be effective and the other toxic. In the past, drugs were often sold as racemates (an equal mixture of both enantiomers) when the least effective enantiomer was not a risk or was converted to the active enantiomer in the body. *See* RACEMIZATION.

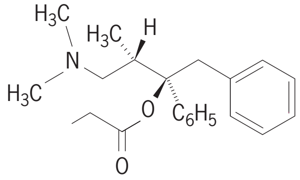


(**5**) (**6**)

There has been a movement toward preparation of pure enantiomers for pharmaceuticals owing to the

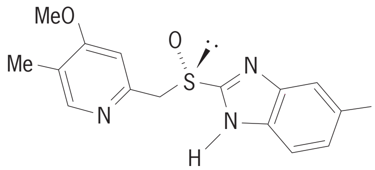


(**7**)



(**8**)

potential side effects of the ineffective enantiomer, which, if not converted to the active enantiomer, at least would present a metabolic burden on an individual. Another advantage of providing a single enantiomer drug is that, as viewed by the patent office, the pure enantiomer is a different compound than the racemate. Hence, the racemic version of omeprazole, a \$6 billion-a-year drug, is marketed in the enantiomerically pure form as esomeprazole (**9**). The “es” corresponds to the (*S*) configuration of the sulfoxide. This marketing tactic for preserving patent rights is known as a chiral switch.



(**9**)

Distinguishing enantiomers. The chirality descriptors (*R*) and (*S*) provide a method of labeling the three-dimensional aspects of a chiral center, and for a molecule with one center would describe the absolute configuration of that molecule. Although these enantiomeric molecules may behave differently in a biological system or with other chiral molecules, almost all of their physical properties are identical. One exception is their interaction with polarized light. Plane-polarized light passing through a solution containing a single enantiomer will rotate the plane of polarization clockwise (dextrorotatory or +) or counterclockwise (levorotatory or −), for a given wavelength of light. This provides an experimental parameter by which a particular enantiomer can be identified. The relationship between the direction of rotation of light and the absolute configuration is complex, and this physical property cannot often be used reliably to determine absolute configuration. For compounds containing heavy atoms, x-ray scattering can provide a method for determining absolute configuration in a crystal structure, and this provides a method for correlating observed optical rotation with configuration. When both are known, the nomenclature often signifies the physical property, as well as the absolute configuration, for example, (*R*)-(−)-carvone (**5**). By knowing the degree of rotation of a standard concentration (1 g/L) of a pure enantiomer, one can calculate a specific rotation, $[\alpha]_D$. Hence the amounts of each enantiomer in a sample that is not enantiomerically pure can often be determined by measuring its $[\alpha]_D$. The $[\alpha]_D$ of (**5**) is -61°, whereas the $[\alpha]_D$ of (**6**) is $+61^\circ$, and an equal mixture of both, a racemate, would have an $[\alpha]_D$ of 0°. *See* OPTICAL ACTIVITY.

A sample that contains an unequal amount of enantiomers is said to be chiral nonracemic, and the enantiomeric excess (ee) is given by Eq. (1), where *R* and

$$\mathrm{ee} = \frac{|R - S|}{R + S} \times 100 \qquad (1)$$

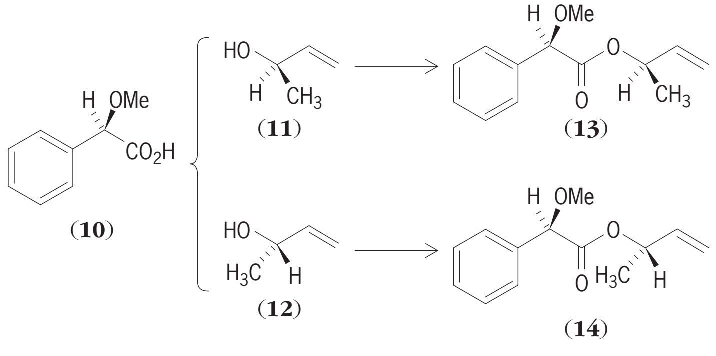
S correspond to the concentrations of the respective enantiomers.

This definition in terms of percentages of enantiomers is analogous to the effect on $[\alpha]_D$ of the presence of one enantiomer rotating light in the opposite direction to the other in mixtures. The term ee is now preferred over the older term of optical purity as the rotation may not be proportional to concentration owing to effects such as dimerization. Modern methods, such as gas chromatography and nuclear magnetic resonance (NMR) spectroscopy, use interactions with other chiral molecules to determine the percentages of enantiomers.

In the absence of the influence of nonracemic substances, symmetrically constituted reagents would be expected to yield racemic mixtures. The aim of an asymmetric synthesis is to selectively prepare enantiomerically pure products through the use of a nonracemic material, particularly avoiding the necessity of carrying out physical separations.

Strategies. If one prepares a racemate, the enantiomers can be separated in a number of ways.

Diastereomer separation. A classic method involves reaction of the mixture of enantiomers with an enantiomerically pure reagent and then separating the diastereomeric products. Diastereomers are no longer related as mirror images and thus have different chemical and physical properties. For example, an enantiomerically pure acid (**10**) would react with a racemic chiral alcohol, (**11**) and (**12**), to yield diastereomeric esters [reaction (2)].



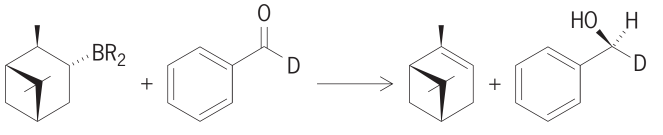
(2)

Since diastereomers (**13**) and (**14**) are formed, they can be distinguished by physical methods such as NMR or retention times on a gas chromatograph. This approach would provide an analysis of the enantiomeric purity of a nonracemic alcohol without resorting to the use of optical rotation.

In many cases, diastereomers can be separated by various physical methods, such as crystallization, and subsequent removal of the original chiral fragment would provide a method for isolation or resolution of a pure isomer. Thus, hydrolysis of pure (**13**) would yield (**11**) of 100% ee.

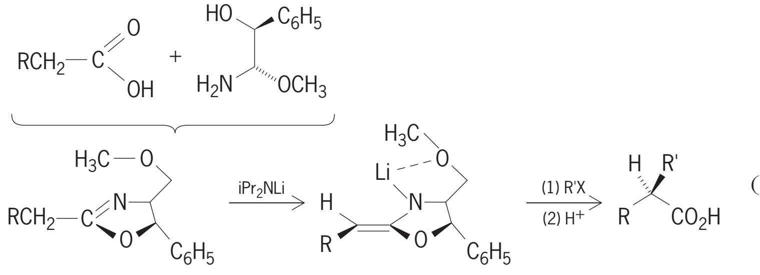
This type of separation of enantiomers does not fit the aim of an asymmetric synthesis, where physical separations are supposed to be avoided in principle.

Chiral reagent. If a chiral reagent, substituent in the molecule, or other entity influences the formation of new chiral center, one has the essence of an asymmetric synthesis. A common approach is the use of a chiral reagent. An example is the asymmetric reduction of a deuterated aldehyde with alpine-borane (**15**) which allows recovery of α-pinene (**17**) from which the reagent was derived and provides the chiral alcohol in ∼100% ee [reaction (3)]. Note that (**18**) would be nearly impossible to resolve from a racemate by a conventional resolution via diastereomers.

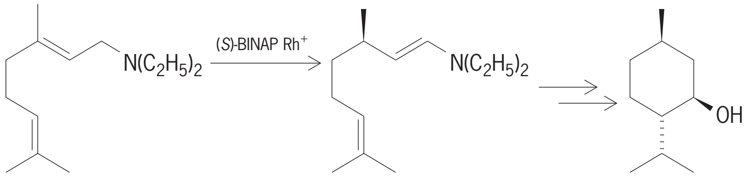


(**15**) (**16**) (**17**) (**18**) (3)

Another reliable strategy is the temporary incorporation of an enantiomerically pure chiral auxiliary to direct the stereochemistry of a reaction and subsequent cleavage of it from the product. For example, if one wishes to enantioselectively substitute an α-position on a carboxylic acid, derivatization with an amino alcohol with a directing group can provide the necessary stereocontrol of the reaction [reaction (4)].



(**19**) (**20**) (**21**) (4)

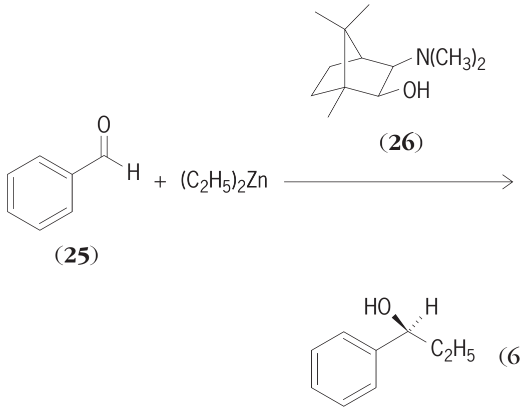


(**22**) (**23**) (**24**) (5)

Chiral catalyst. These strategies require a 1:1 use of chiral reagent or auxiliary per chiral molecule produced. Ideally, one would prefer to have the chirality multiplied in some way that large numbers of enantiopure product molecules could be formed for each enantiopure agent that was used. Asymmetric catalysis provides a way to do this. An example is the Takasago process that is used commercially for the preparation of (−)-menthol (**24**). Cationic rhodium complexes prepared from BINAP are very efficient catalysts for the enantioselective isomerization of allylic amines to enamines [reaction (5)].

Thus, the chirality within the catalyst is transferred in the isomerization of the achiral geranylamine (**22**) to the enamine (**23**). A small amount of catalyst can isomerize a large number of molecules and create chiral product in high enantiomeric yield, and this provides a method of multiplying the chirality that exists in the catalyst. Asymmetric catalysis is an important component of modern approaches to asymmetric synthesis. *See* CATALYSIS.

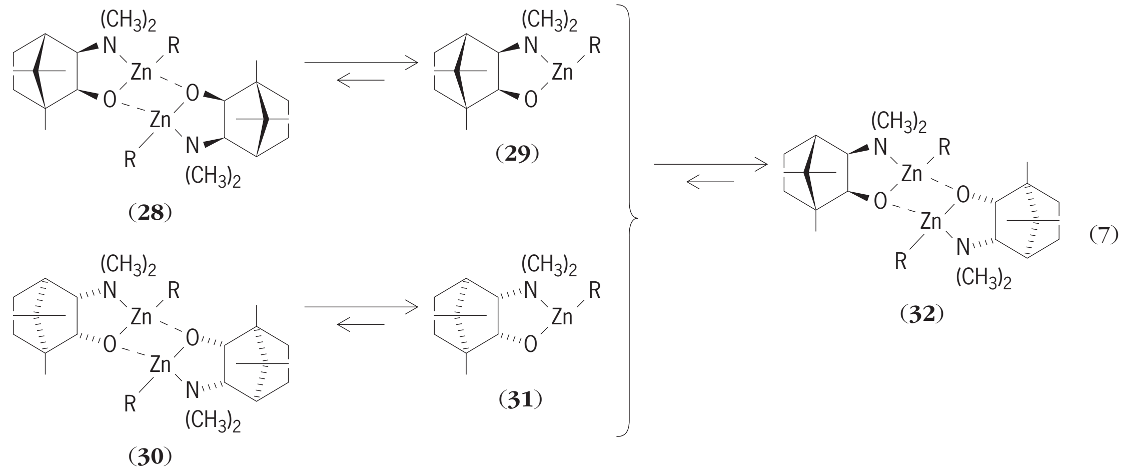
Chiral amplification. Intuitively one would expect that an enantiomerically pure catalyst would be required to yield products in high enantiomeric purity; however, this is not always the case. The rate of addition of diethylzinc to benzaldehyde is dramatically enhanced by the addition of an amino alcohol. If the amino alcohol is chiral and nonracemic, an enantioselective addition can take place. When enantiopure DAIB [3-exo-(dimethylamino)isoborneol] (**26**) is used, a product with 98% ee is obtained [reaction (6)].



(**25**) (**26**) (**27**) (6)

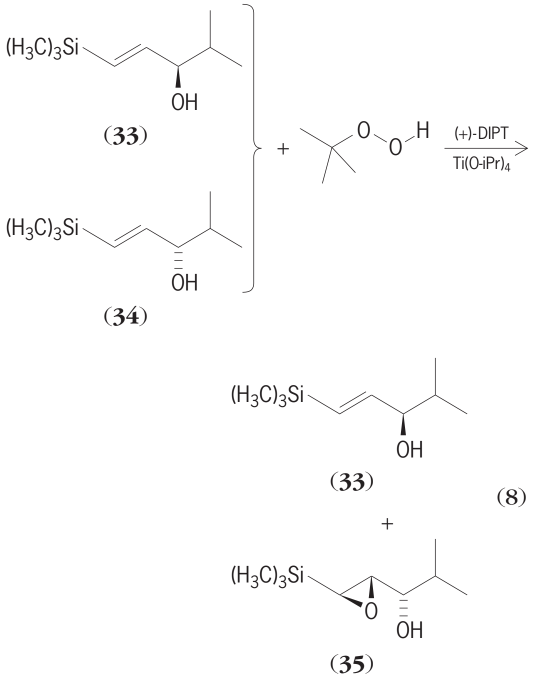
A fascinating feature of this reaction is that when DAIB has an enantiomeric purity of only 21% ee, the product (**27**) is obtained in 96% ee. The origin of this effect can be found in the relative stability of dimers containing either two molecules of the same enantiomer of DAIB or different enantiomers [reaction (7)].

Monomers (**29**) and (**31**) are considered to be the active species. In this case, the homo dimers (**28**) and (**30**) are highly dissociated, whereas the hetero dimer (**32**) remains mostly intact. Thus, virtually all of monomer (**31**) is tied up in the unreactive hetero dimer, and this leaves the excess (**29**) as the predominant species effecting the reaction.



(7)

Kinetic resolution. Another approach to obtaining enantiomers in high ee from a racemate involves the enantioselective conversion of one of the enantiomers into a different compound. This can be achieved with enzyme hydrolysis of esters, but one of the most commonly used methods for deracemization employs selective epoxidation of allylic alcohols using tartrate esters [reaction (8)].

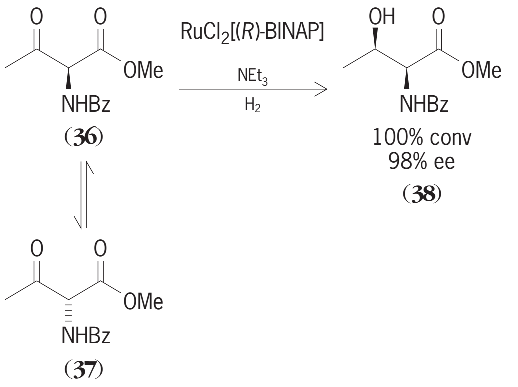


(8)

This approach not only provides an enantioselective synthesis of an epoxide (**35**) from (**34**) in 99% ee, but also leaves (**33**) behind in >99% ee. A disadvantage of this approach is that half of the starting material remains unreacted if the goal was to obtain pure epoxide.

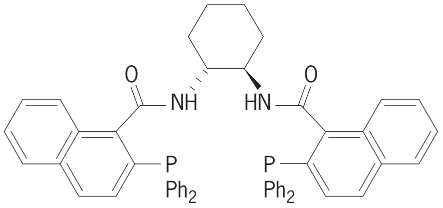
Dynamic kinetic resolution. The problem associated with 50% conversion of a racemic starting material can be overcome if a method for racemization of the starting material is available. Under certain conditions, racemization of the substrate may occur on a time scale that is fast compared to the reaction. This can be observed in ketone hydrogenations [reaction (9)]. In this case, (**36**) hydrogenates more rapidly than (**37**) with a ruthenium catalyst using the (*R*)- BINAP ligand.

Treating palladium with (**39**) provides a system that effectively racemizes the intermediates formed with diene monoepoxides, but also catalyzes enantioselective attack by pthalimide (**40**).

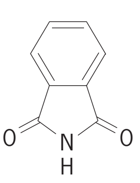


100% conv
98% ee

(9)

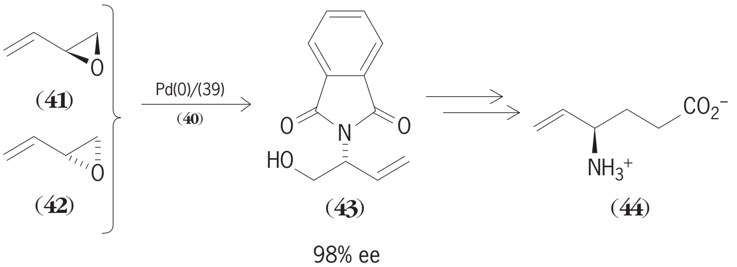


(**39**)



(**40**)

This procedure has been used by B. M. Trost to convert the racemic epoxide (**41**)/(**42**) to a precursor of the drug (*S*)-vigabatrin (**44**) [reaction (10)]. Since this process does not involve prior racemization of the substrate, it has been designated as a dynamic kinetic asymmetric transformation.



(10)

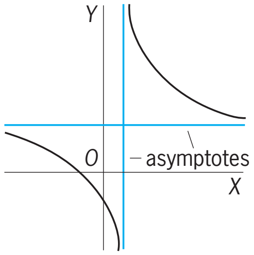
98% ee
99% yield

J. W. Faller

Bibliography. C. Bolm and J. A. Gladysz, Enantioselective catalysis, *Chem. Rev.*, 103:2761–3400, 2003; E. Eliel, S. Wilen, and L. Mander, *Stereochemistry of Organic Compounds*, Wiley, New York, 1994; R. E. Gawley and J. Aube, *Principles of Asymmetric Synthesis*, Elsevier, Oxford, 1996; R. Noyori, *Asymmetric Catalysis in Organic Synthesis*, Wiley, New York, 1994; J. Seyden-Penne, *Chiral Auxiliaries and Ligands in Asymmetric Synthesis*, Wiley, New York, 1995.

Asymptote

A line that is a limit of lines tangent to a curve as the contact points of those tangents approach infinity along the curve. Thus, an asymptote of a curve is an ordinary line (that is, not the "line at infinity") that is tangent to a curve at the points in which the curve intersects the line at infinity (see **illus.**). Analytically, a line $y = mx + b$ is an algebraic curve $f(x, y) = 0$, provided the constants m and b are such that the coefficients of the two terms of highest degree of $f(x, mx + b)$ are zero. For example, to find m and b so that $y = mx + b$ is an asymptote of $x^3 + y^3 - 3xy = 0$, elimination of y yields



Asymptotes of a hyperbola.

$$(1 + m^3)x^3 + 3x^2(bm^2 - m) + 3x(mb^2 - b) + b^3 = 0$$

Equating the two leading coefficients to zero gives $m = b = 1$. Thus $y + x + 1 = 0$ is an asymptote. *See* ANALYTIC GEOMETRY; HYPERBOLA.

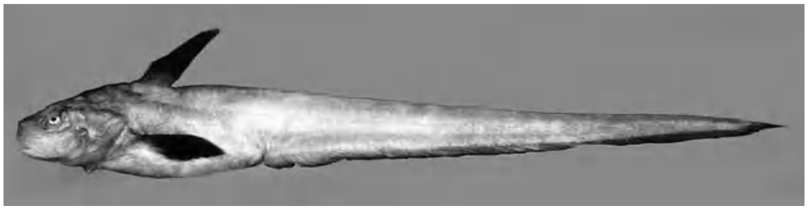
Leonard M. Blumenthal

Ateleopodiformes

An order of actinopterygian (ray-finned) fishes consisting of only one family, Ateleopodidae, known as jellynose fishes. Ateleopodiformes along with the order Stomiiformes constitute a teleost superorder called the Stenopterygii. Jellynose fishes are very distinctive in having a large head and bulbous snout; a long anal fin that unites with a much reduced caudal fin; a dorsal fin placed forward with 3 to 13 rays; a pelvic fin in a jugular position with a single ray in adults; and a largely cartilaginous skeleton (see **illustration**). Individuals may reach a length of 2 m (6.6 ft) at a depth range of 140–1281 m (called the bathydemersal environment). The family, consisting of 4 genera and 12 species, is represented in the eastern Atlantic, Caribbean Sea, eastern Pacific off Central America, and the Indo-West Pacific. *See* ACTINOPTERYGII; TELEOSTEI.

Herbert Boschung

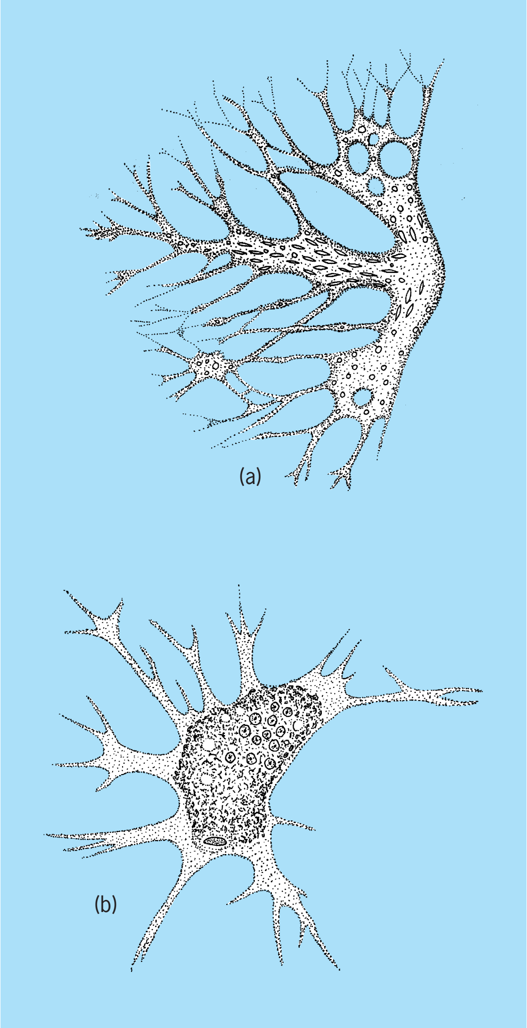
Bibliography. J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994; M. M. Smith and P. C. Heemstra (eds.), *Smith's Sea Fishes*, Macmillian South Africa, Johannesburg, 1986.



***Itimaia plicatellus*, a jellynose fish found in deep water in the east-central Pacific Ocean. (*Photo © John E. Randall*)**

Athalamida

An order of Granuloreticulosia in which the naked amebas form branched, threadlike interconnected pseudopodia (reticulopodia). Species are



Athalamids. (*a*) *Biomyxa vagans*. (*b*) *Arachnula impatiens*. (*After R. P. Hall, Protozoology, Prentice-Hall, 1953*)

known from fresh, salt, and brackish water. General characteristics are difficult to select, and the order may lack the taxonomic stability to survive accumulation of more nearly adequate information. Heterogeneity may extend even to a genus; described species of *Biomyza* differ appreciably in morphology and exhibit, for example, uninucleate and multinucleate condition in different species (although these may represent young and mature stages in life cycles). The granular cytoplasm in *Biomyza* (**illus.** *a*) shows no clear separation into ectoplasm and endoplasm. In addition to *Biomyza*, the genera *Arachnula* (illus. *b*), *Gymnophrys*, and *Pontomyxa* have been assigned to this order. *See* GRANULORETICULOSIA; PROTOZOA; RHIZOPODEA; SARCODINA; SARCOMASTIGOPHORA.

Richard P. Hall

Atheriniformes

An order of actinopterygian (ray-finned) fishes that comprise, with the Beloniformes and Cyprinodontiformes, the series Artherinomorpha. Atherinimorphs are more or less intermediate between typical soft-rayed and spiny-rayed fishes, but are not thought to be the ancestry of the latter. The order Atheriniformes, consisting of about 285 species in six families, are relatively small fishes, usually with an elongate body; opercular and preopercular margins lacking spines or serrations; lateral line absent or nearly so; usually two dorsal fins, with the first, if present, having weak spines; an anal fin larger than the soft dorsal fin and usually preceded by one spine; and usually abdominal pelvic fins, but subthoracic or absent in some (see family Phallostethidae below).

Habitats and distribution. Atheriniformes are represented by marine, fresh-water, and brackish water species. The marine species occupy continental and island shores of the tropical and temperate seas of the world. The fresh-water and brackish water species are found in northern and eastern Australia, New Guinea, the Philippines, eastern Indonesia, Southeast Asia, Malaya Peninsula, Madagascar, Africa, and North and South America.

The families Atherinidae and Atherionidae (Old World silversides) and the similar Atherinopsidae (New World silversides) share, in addition to the characters of the order, a slender, usually moderately compressed body; small terminal mouth; two widely separated dorsal fins; pectoral fins high on body and usually abdominal; scales large, usually cycloid, sometimes ctenoid; no lateral line; and broad silvery lateral band (black in preserved specimens). Atherinopsidae is the principal family of the western hemisphere. Fresh waters of Mexico have a rich silverside fauna: 39 species, primarily in two genera, *Atherinella* and *Chirostoma*. By contrast, all of the United States has only three species, which are limited to fresh water. The most widespread is the brook silverside (*Labidesthes sicculus*), ranging from the St. Lawrence River system south throughout much of the Mississippi Basin to the Gulf States. All in all, the family Atherinopsidae is represented in North America (including all of Mexico) by 11 genera and 56 species of which only 15 are limited to the marine environment, 10 from the Pacific, and 5 from the Atlantic. Two species of Old World silversides (Atherinidae) extend their range to the western North Atlantic, the hardhead silverside and reef silverside.

Another large family of atheriniforms is the Melanotaeniidae (rainbowfishes and blue eyes) consisting of 68 species from the fresh waters of northern and eastern Australia, New Guinea, and parts of eastern Indonesia. The body of the rainbow fishes is compressed; the dorsal fins are narrowly separated; the lateral line system is absent or weakly developed; the entire length of innermost pelvic fin ray is attached to the abdomen by a membrane; and, unlike most atheriniforms, they are colorful. The blue eyes are morphologically similar but differ in some features unique in the order; for example, the infraorbital series is reduced to a single bone (the lacrimal), and the mesethmoid is absent.

Reproduction and development. Reproduction in the brook silverside, as with some other atheriniforms, includes internal fertilization but without gross modification of fins and other skeletal structures to form clasping and copulatory organs; however, such organs are highly developed in the Phallostethidae, a family of 19 species of small fishes in Southeast Asia from the Philippines to Thailand. In this case, a unique copulatory structure of males called the pariapium is used to hold the female during mating and aids in transferring sperm packets. Derived from the pelvic skeleton, the pariapium is a complicated muscular and bony organ in the throat region that contains ducts from the urogenital system, as well as the terminal part of the intestine, all of which have migrated forward to below the pectoral fins in both sexes. The pelvic skeleton is absent in females.

Although fertilization is internal in both the brook silverside and the phallostethids, development is oviparous, with the female laying fertilized eggs. This reproductive strategy does not have all the advantages of viviparous development, as seen in some of the atheriniforms' near relatives in the orders Beloniformes and Cyprinodontiformes; however, fertilized eggs may be held and deposited at a time conducive to successful development, thus enhancing the reproductive potential of the species, especially one like the brook silverside, whose life cycle includes only one reproductive season. *See* ACTINOPTERYGII; CYPRINODONTIFORMES; TELEOSTEI. Herbert Boschung

Bibliography. B. S. Dyer, Atherinopsidae (Neotropical silversides), pp. 515–525 in R. E. Reis et al. (eds.), *Checklist of the Freshwater Fishes of South and Central America*, EDIPUCRS, Porto Alegre, Brazil, 2003; B. S. Dyer and B. Chernoff, Phylogenetic relationships among atheriniform fishes (Teleostei: Atherinomorpha), *Zool. J. Linn. Soc.*, 117:1–69, 1996; R. J. Lavenberg and M. Chernoff, Atherinidae: Pejerreyes, pp. 889–901 in W. Fischer et al. (eds.), *Guia FAO para Identification de Especies para lo*

Fines de la Pesca, Pacifico Centro-Oriental, 3 vols., FAO, Rome, 1995; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994; L. R. Parenti, A phylogenetic revision of the pallostethid fishes (Atherinomorpha, Phallostethidae), *Proc. Calif. Acad. Sci.*, 46(11):243–377, 1989; L. R. Parenti, Relationships of atherinomorph fishes (Teleostei), *Bull. Mar. Sci.*, 52(1):170–196, 1993; D. E. Rosen, The relationships and taxonomic position of the halfbeaks, killifishes, silversides, and their relatives, *Bull. Amer. Mus. Nat. Hist.*, vol. 127, no. 5, 1964.

Atlantic Ocean

The large body of seawater separating the continents of North and South America in the west from Europe and Africa in the east and extending south from the Arctic Ocean to the continent of Antarctica. The Atlantic is the second-largest ocean water body, and in area it covers nearly one-fifth of the Earth's surface. It receives the fresh-water runoff from a continental drainage area approximately four times larger than that draining into either the Pacific or Indian oceans. The two major divisions, North and South Atlantic oceans, have the Equator as the common boundary. The North Atlantic, because of projecting land areas and island arcs, has numerous subdivisions. These include three large mediterranean-type seas, the Mediterranean Sea, the Gulf of Mexico plus Caribbean Sea, and the Arctic Ocean; two small mediterranean-type seas, the Baltic Sea and Hudson Bay; and four marginal seas, the North Sea, English Channel, Irish Sea, and Gulf of St. Lawrence. *See* ARCTIC OCEAN; BALTIC SEA; EQUATOR; GULF OF MEXICO.

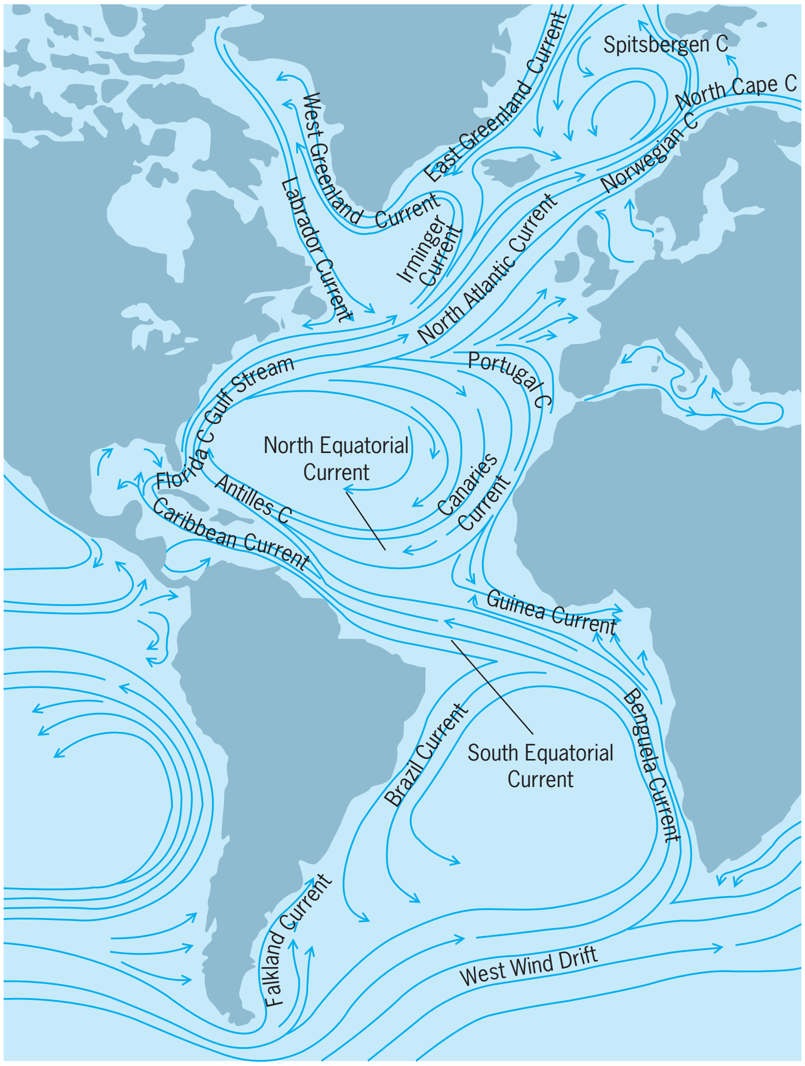


Fig. 1. Surface currents of the Atlantic Ocean. (*After J. Bartholomew, Advanced Atlas of Modern Geography, 3d ed., McGraw-Hill, 1987*)

Surface currents. Under the direction of M. F. Maury, the U.S. Naval Oceanographic Office began compiling wind and current measurements in the mid-1800s from logs of U.S. Navy ships as well as domestic and foreign merchant ships. The data were used to develop the first pilot charts and sailing directions. These historical ship drifts and pilot charts have provided most of what is known about the large-scale surface velocity patterns in the ocean. Modern measurements consisting of satellite-tracked surface drifters, moored current meters, and shipboard profiles have filled in many details of surface currents, including some of their variations in time. *See* INSTRUMENTED BUOYS.

Surface currents in the Atlantic Ocean are largely driven by the prevailing surface winds and flow in much the same direction as the winds (**Fig. 1**). Both the winds and currents circulate in large anticyclonic gyres centered at midlatitudes in the subtropical North and South Atlantic oceans. The currents tend to be much swifter on the western side of the gyres (Gulf Stream, Brazil Current) than on the east, because of the shape of the Earth and its rotation (Coriolis acceleration). A net northward flux of upper-layer water moves through the Atlantic, crossing the Equator near the western boundary. A compensatory southward flow of colder, denser water occurs at depth. The thermohaline circulation, which results in a net flux of heat from the South Atlantic to the North Atlantic, is driven by water mass density differences caused by geographical variations of temperature and salinity. *See* SEAWATER.

North Atlantic. The greater part of the water transported by the North Equatorial Current enters the Caribbean as the Caribbean Current and exits the Gulf of Mexico as the Florida Current, which continues northward along the North American coast as the Gulf Stream. Roughly half of the water flowing in the Caribbean Current and Florida Current enters from the North Equatorial Current; the other half originates in the South Atlantic and flows northward in the South Equatorial, North Brazil, and Guyana currents. *See* GULF STREAM.

South of the Grand Banks, the Gulf Stream divides into several branches. The North Atlantic Current flows across the Mid-Atlantic Ridge and forms several offshoots of relatively warm, saline waters that continue flowing in a northeastward direction. One of these, the Irminger Current, reaches Iceland.

Another flows across the Greenland-Scotland Ridge between the Faeroes and the Shetland Islands and thence along the Norwegian coast as the Norwegian Current. Part of this flow continues into the Barents Sea as the North Cape Current, and part of it reaches the Arctic Ocean as the West Spitsbergen Current. The transport of cold water southward in the East Greenland and Labrador currents partially compensates for the transport of relatively warmer water northward into the Arctic. The Irminger, West Greenland, and Labrador currents circulate around a cyclonic gyre centered in the Labrador Basin. Other branches of the North Atlantic Current flow southward as the Portugal and Canaries currents. These join the North Equatorial Current to complete the circuit of the North Atlantic.

Between the North and South Equatorial currents is the North Equatorial Countercurrent, which flows eastward across the Atlantic between 5°N and 10°N and into the Guinea Current. The Countercurrent is strongest during the second half of the year and disappears in the western Atlantic during the first half. The seasonal variation of the equatorial currents is largely due to the seasonal migration of the trade winds north and south across the Equator. Another eastward countercurrent that stretches across the Atlantic, the Equatorial Undercurrent, is located just below the surface layer centered on the Equator. This current is generated by the trade winds piling up water in the western Atlantic. *See* MARITIME METEOROLOGY; TROPICAL METEOROLOGY.

South Atlantic. The currents in the South Atlantic are, in many respects, the counterparts of those in the North Atlantic, for example the Brazil Current and the Gulf Stream, the Benguela and Canaries currents, and the Falkland and the Labrador currents. The Circumpolar Current or West Wind Drift flows eastward through the South Atlantic. Between the Circumpolar Current and Antarctica is located the Weddell Sea cyclonic gyre.

Ocean eddies. The Gulf Stream and other ocean currents are variable in position and time. Much of this variability is due to ocean eddies that are typically a few hundred kilometers in extent (**Fig. 2**). The eddies appear to be generated by instabilities of the swiftest currents and have velocities similar to them. Eddies are thought to be important to the dynamics of ocean circulation, and can transport water properties over long distances and times. Some of the longer-lived eddies survive several years, and change position by a thousand kilometers. In charts which show the general circulation, the eddies have been ignored (Fig. 1).

Surface temperature, salinity, and density. The sea surface temperature pattern is determined by geographical and seasonal variations of heating and cooling plus the advection of water by ocean currents and mixing (**Fig. 3**). Surface water is warmest in the equatorial band, where temperatures of 82°F (28°C) are found. Warm tropical water is carried northward in the Gulf Stream, where during winter large amounts of heat are released to the atmosphere. The northward heat flux in the Atlantic by ocean currents is thought to be about equal to that carried by the atmosphere.

Surface salinity is determined by geographical and seasonal variations in precipitation and evaporation and also advection and mixing. Highest salinities (greater than 37.0‰) are found in the subtropical gyres, where evaporation exceeds precipitation. Surface salinity is low (less than 35‰) in the doldrums, where heavy precipitation occurs. Salinity values below 30‰ occur where currents, such as the East Greenland, West Greenland, and Labrador, transport melting ice. Offshore of the major rivers such as the Amazon, which enters the Atlantic near the Equator, the surface salinity can be reduced considerably below usual oceanic values.

The warm and salty surface water of the subtropical gyres is less dense than the much colder and fresher water observed in the far north and south. The densest surface water in the open Atlantic is found during late winter in the Labrador Basin and Weddell Sea, sites of deep- and bottom-water formation.

Deep circulation. Subsurface water masses of the Atlantic originate from (1) water forced down from the surface in regions of convergence, (2) wintertime cooling and convection, and (3) inflows from other oceans and seas. The surface water in certain areas in the far north and south takes on a particularly high density in winter under the influence of climatic conditions. These water masses sink to a depth where their density is similar to that of the surrounding water, and then spread out at that level. At the same time the newly formed water mixes with the surrounding waters, modifying its characteristics. Overflows into the open North Atlantic from the Norwegian and Greenland seas and the Mediterranean Sea descend and mix with surrounding water until they, too, reach neutral buoyancy and spread out from their source regions as distinct layers. In addition, water from the Pacific enters the South Atlantic in the Circumpolar Current, and water from the Indian Ocean enters the South Atlantic around the southern tip of Africa as filaments and eddies of the Agulhas Current. In this way a multistoried stratification arises.

Of the Atlantic water masses, the densest is the very cold Weddell Sea Water, which is formed near the coast of Antarctica and extends northward along the sea floor to around 10°S (**Fig. 4**). Above this is Circumpolar Water, which enters the South Atlantic through the Drake Passage and which is split into lower and upper parts by higher-salinity water flowing southward from the North Atlantic at depths of 6600–9900 ft (2000–3000 m). The lower Circumpolar Water crosses the Equator in the west and flows northward in the western North Atlantic as far as 40°N. It also enters the eastern Atlantic through the Romanche Fracture Zone near the Equator. In the north, cold dense water flows into the North Atlantic from the Norwegian and Greenland seas over the ridge connecting Greenland and Scotland to form the abyssal water of the northern North Atlantic. Above this layer Labrador Sea Water is

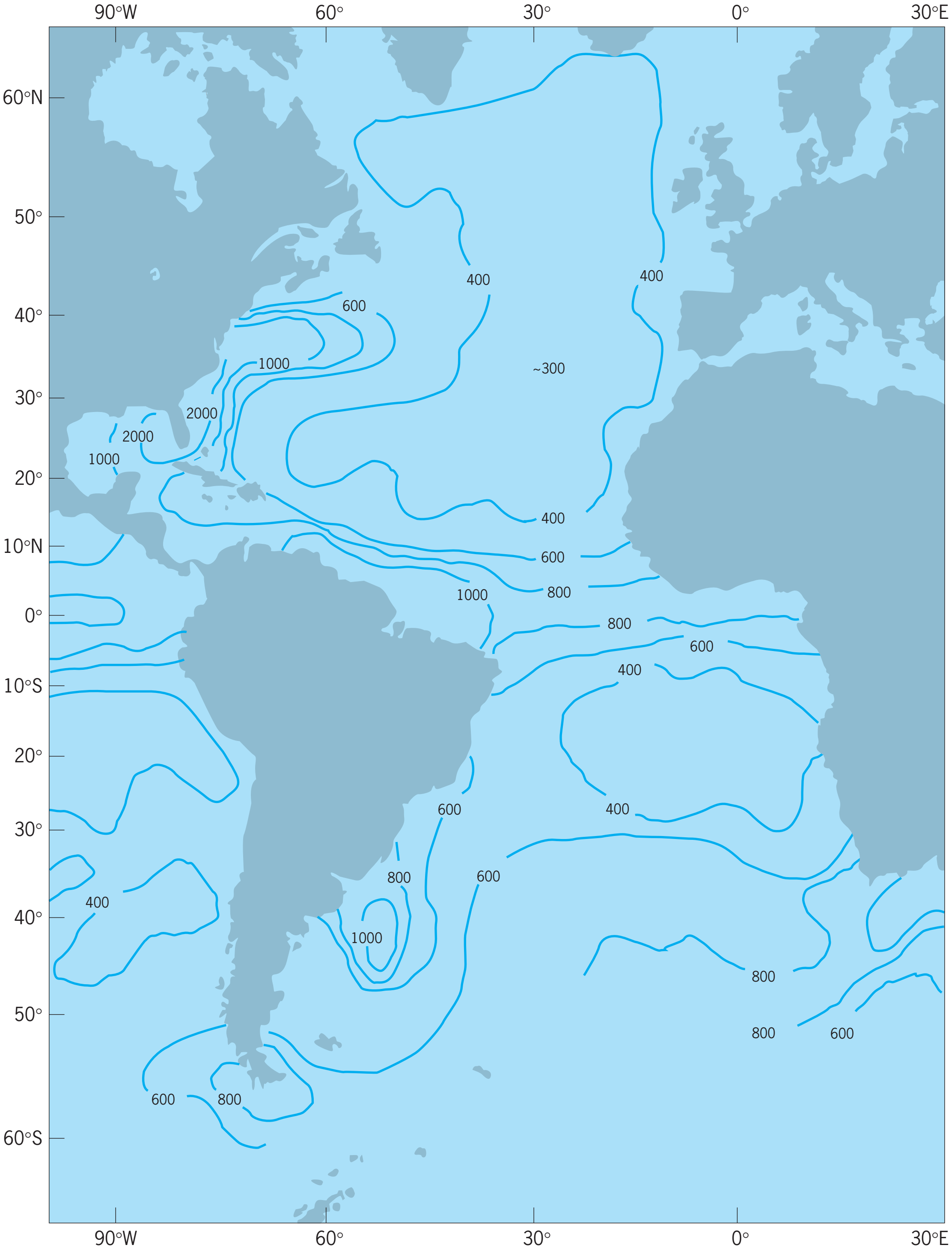

90°W
60°
30°
0°
30°E
60°N
50°
40°
30°
20°
10°N
0°
10°S
20°
30°
40°
50°
60°S
400
400
600
1000
~300
2000
2000
1000
400
600
800
1000
800
600
400
400
600
800
600
1000
400
800
800
600
600
800

Fig. 2. Eddy kinetic energy per unit mass (cm^2/s^2) for the Atlantic Ocean based on historical ship drifts. Contour lines represent cm^2/s^2. 1 cm^2 = 0.155 in.^{2} (*After K. Wyrtki, L. Magaard, and J. Hager, Eddy energy in the oceans, J. Geophys. Res., 81: 2641–2646, 1976*)

formed by deep convection in winter to roughly 5000 ft (1500 m).

Both Labrador Sea Water and the Norwegian Greenland Overflow Water are more saline at their sources than water originating in the South Atlantic, and their salinities are increased somewhat through mixing with the salty water of the Mediterranean Sea, which flows out of the Strait of Gibraltar. Mediterranean Water forms a huge lens of salty water which spreads westward across the Atlantic at depths of 3300–9900 ft (1000–3000 m) between 20° and 40°N with highest values in the eastern Atlantic near 3300 ft (1000 m). The combination of the Overflow Water, Labrador Sea Water, and Mediterranean Water forms North Atlantic Deep Water. This deep water mass flows southward along the western boundary

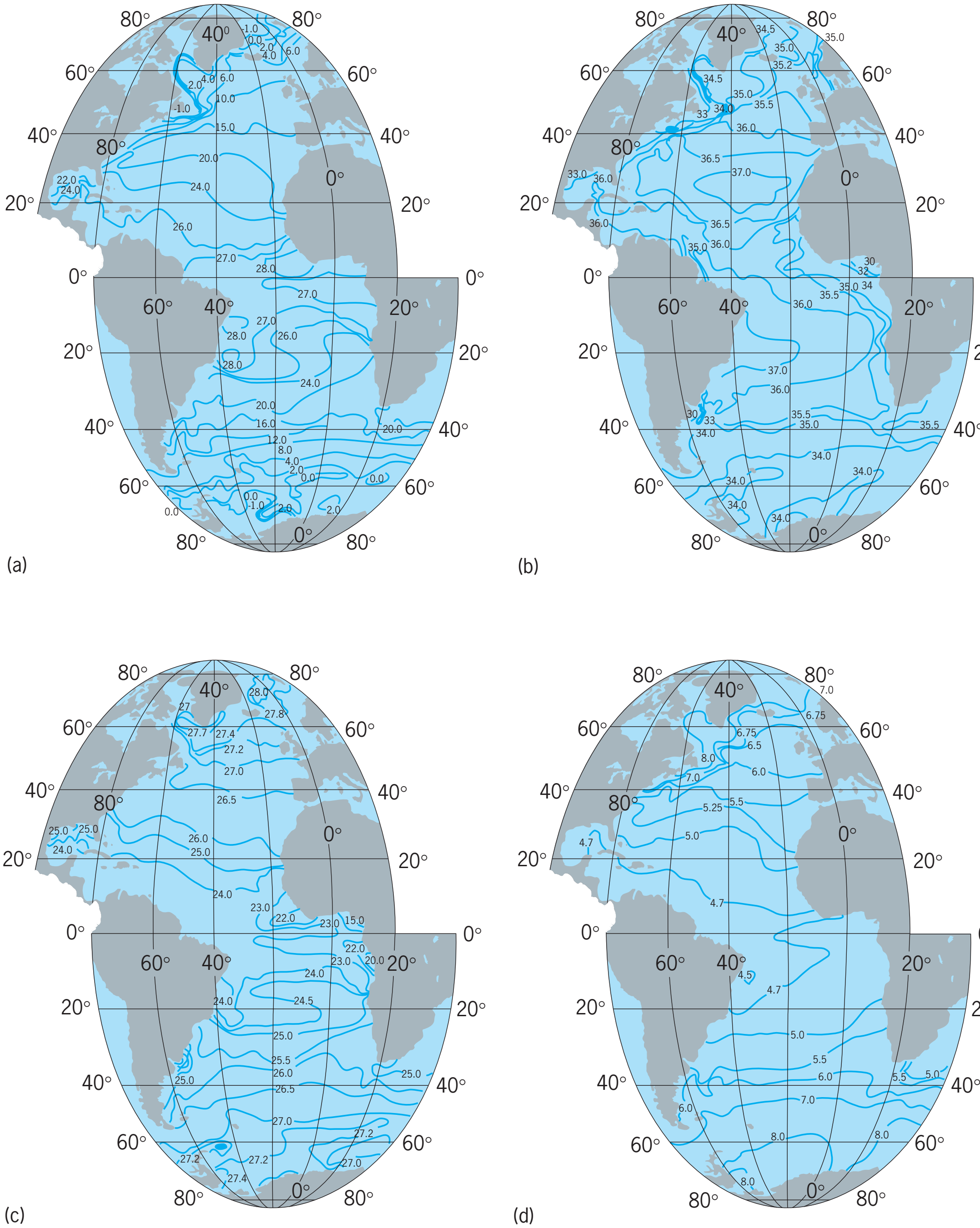

(a)
(b)
(c)
(d)
80°
60°
40°
20°
0°
-1.0
0.0
2.0
4.0
6.0
10.0
15.0
20.0
22.0
24.0
26.0
27.0
28.0
16.0
12.0
8.0
34.5
35.0
35.2
35.5
36.0
36.5
37.0
33.0
34.0
30
32
34
27.8
27.7
27.4
27.2
26.5
25.0
25.5
23.0
22.0
24.5
7.0
6.75
6.5
8.0
6.0
5.5
5.25
5.0
4.7
4.5

Fig. 3. Sea surface patterns. (*a*) Temperature; contour lines represent °C. (*b*) Salinity; contour lines represent a ratio based on relative conductivity. (*c*) Density; contour lines represent (kg/m^3) − 1000. (*d*) Oxygen; contour lines represent cm^3/liter. (*After J. L. Reid, Prog. Oceanog., 33:1–92, 1994*)

of the Atlantic, crosses the Equator, and extends into the Circumpolar Current, splitting it into upper and lower parts.

Above the deep water in the South Atlantic is the upper Circumpolar Water, and above this is the Intermediate Water, which is low in salinity and high in oxygen. The Intermediate Water spreads northward into the North Atlantic at depths of around 2600 ft (800 m) to at least 20°N, and even farther in the Gulf Stream. A warm upper layer overlies the Intermediate Water between the oceanic polar fronts. Temperatures in these waters are greater than 46°F (8°C).

Compared with that of the Indian and Pacific oceans, the deep circulation in the Atlantic is very vigorous, a result of deep water formation and strong thermohaline flows. The fastest speeds in the deep ocean reaching 50 cm/s (around 1 knot) are (1) underneath the major western boundary currents such as the Gulf Stream and Brazil Current; (2) along the western boundary, where the deep and abyssal meridional flows are concentrated; and (3) at the

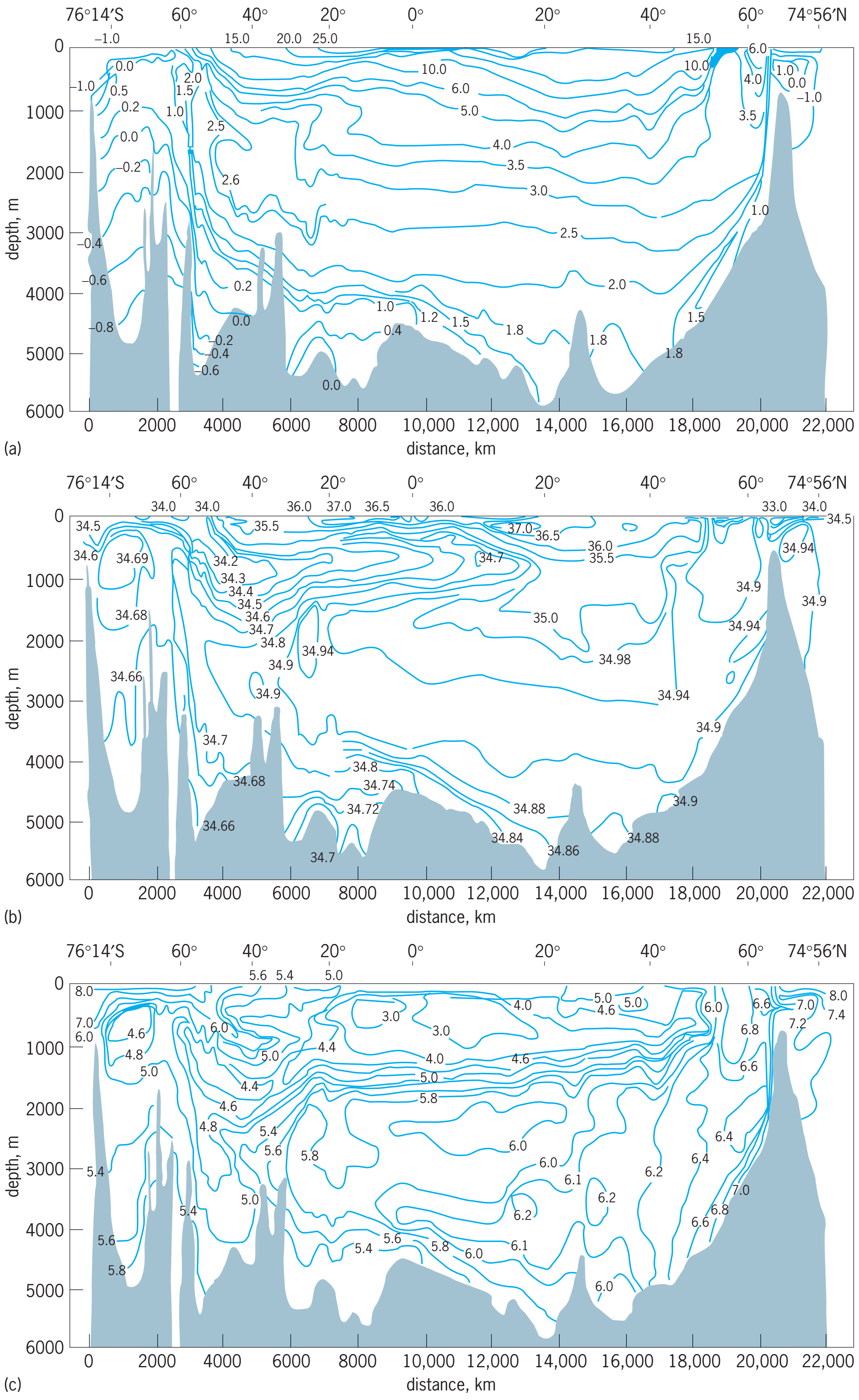


Fig. 4. North-south vertical sections of (*a*) potential temperature (°C) (*b*) salinity; and (*c*) oxygen (ml/liter) extending from the Weddell Sea at 76°S northward through the deeper part of the western basins and the Denmark Strait into the Norwegian-Greenland Sea at 75°N (*after J. L. Reid, On the total geostrophic circulation of the North Atlantic Ocean: Flow patterns, tracers, and transports, Prog. Oceanog., 33:1–92, 1994*).

sites of overflows, where dense water cascades down to its depth of neutral buoyancy much like an undersea waterfall. A particularly vigorous and complicated circulation scheme exists where the southward-flowing deep western boundary current encounters the deep extension of the Gulf Stream. Deep western boundary currents appear to have local energetic recirculations flowing counter to and offshore of the boundary currents. *See* OCEAN CIRCULATION. Philip L. Richardson

Bottom topography. Work on the transoceanic telegraph cables begun in 1860 stimulated development of deep-sea soundings. Since 1922 use of the echo sounder has made it possible to obtain a great number of depth soundings. The mean depth of the Atlantic Ocean is 12,690 ft (3868 m), and its volume is 76,000,000 mi^3 (318,000,000 km^3). *See* ECHO SOUNDER; MARINE GEOLOGY.

Broad shelves with depths less than 660 ft (200 m) are found in the region of the North Sea and the British Isles, on the Grand Banks of Newfoundland, and off the coasts of northeastern South America and Patagonia. The Mid-Atlantic Ridge, which extends from the Arctic Ocean to 55°S, is less than 9900 ft (3000 m) beneath the surface and is characterized by a pronounced relief. It separates the east and west Atlantic troughs, both of which have relatively uniform relief. The east and west troughs are connected in the vicinity of the Equator by the Romanche Deep, the only deep submarine passage through the Mid-Atlantic Ridge, with a depth of 25,500 ft (7728 m). *See* MID-OCEANIC RIDGE.

Three marked east-west ridges—the Greenland-Scotland Ridge in the North Atlantic and the Walvis and Rio Grande ridges in the South Atlantic—and several less conspicuous east-west rises separate the two Atlantic troughs into a series of basins including the West European, Canary, and Angola in the eastern Atlantic and the North American, Brazilian, and Argentine basins in the western Atlantic. Only isolated seamounts (such as the Great Meteor, Altair, and Atlantis seamounts) rise from the floor of the deep basins. Greatest depths occur in the narrow trenches along island arcs: 30,240 ft (9219 m) in the Puerto Rico Trench and 27,110 ft (8264 m) in the South Sandwich Trench. *See* DEEP-SEA TRENCH; SEAMOUNT AND GUYOT.

Islands. The islands of Jan Mayen, Iceland, St. Paul, Ascension, St. Helena, Tristan da Cunha, Gough, and Bouvet are parts of the Mid-Atlantic Ridge and are of purely volcanic origin. Other islands of volcanic origin but which lie outside the Mid-Atlantic Ridge are the Faeroes, Madeira, Fernando Poo, Príncipe, São Tomé, Annobón, Fernando Noronha, Trinidad, and the South Sandwich Islands. The Azores, Canary Islands, Cape Verde Islands, and Lesser Antilles are of predominantly volcanic origin. The Bermudas are the northernmost coral reefs. They rise from an old submarine volcanic cone. All the other islands in the Atlantic Ocean are continental in character, such as Spitsbergen, Bear Islands, the British Isles, Greater Antilles, Falkland Islands, and South Georgia. *See* ARCTIC AND SUBARCTIC ISLANDS; OCEANIC ISLANDS; WEST INDIES. Gunter O. Dietrich

Bibliography. G. Bearman, *Ocean Circulation*, 1989; K. O. Emery and E. Uchupi, *The Geology of the Atlantic Ocean*, 1984; S. G. Gorshkov, *World Ocean Atlas*, vol. 2: *Atlantic and Indian Ocean*, 1979; J. L. Reid, On the total geostrophic circulation of the South Atlantic Ocean: Flow patterns, tracers, and transports. *Prog. Oceanog.*, 23:149–244, 1989; J. L. Reid, On the total geostrophic circulation of the North Atlantic Ocean: Flow patterns, tracers, and transports, *Prog. Oceanog.*, 33:1–92, 1994; W. J. Schmitz, Jr., and M. S. McCartney, On the North Atlantic circulation, *Rev. Geophys.*, 31(1):29–49, 1993; B. A. Warren and C. Wunsch (eds.), *Evolution of Physical Oceanography: Scientific Surveys in Honor of Henry Stommel*, 1981.

Atmosphere

A gaseous layer that envelops the Earth and most other planets in the solar system. Earth, Venus, Mars, Jupiter, Saturn, Uranus, Neptune, and Titan (Saturn's largest satellite) are all known to possess substantial atmospheres that are held by the force of gravity. The structure and properties of the various atmospheres are determined by the interplay of physical and chemical processes. Structural features of Earth's atmosphere detailed below can often be identified in the atmospheres of other planetary bodies. *See* PLANETARY PHYSICS.

Composition. The composition of the Earth's atmosphere is primarily nitrogen (N_2), oxygen (O_2), and argon (Ar) [see **table**]. The concentration of water vapor (H_2O) is highly variable, especially near the surface, where volume fractions can vary from nearly 0% to as high as 4% in the tropics. There are many minor constituents or trace gases, such as neon (Ne), helium (He), krypton (Kr), and xenon (Xe), that are inert, and active species, such as carbon dioxide (CO_2), methane (CH_4), hydrogen (H_2), nitrous oxide (NO), carbon monoxide (CO), ozone (O_3), and sulfur dioxide (SO_2), that play an important role in radiative and biological processes.

In addition to the gaseous component, the atmosphere suspends many solid and liquid particles. Aerosols are particulates usually less than 5 micrometers in diameter that are created by gas-to-particle reactions or are lifted from the surface by the wind. A portion of these aerosols can become centers of condensation or deposition in the growth of water and ice clouds. Cloud droplets and ice crystals are made primarily of water with some trace amounts of particles and dissolved gases. Their diameters range from a few micrometers to about 100 μm. Water or ice particles larger than about 100 μm begin to fall because of gravity and may result in precipitation at the surface. *See* AEROSOL; CLOUD PHYSICS; PRECIPITATION (METEROLOGY).

One of the remarkable properties of the Earth's atmosphere is the large amount of molecular oxygen in the presence of gases, such as nitrogen, methane, water vapor, hydrogen, and others, that are capable of being oxidized. The atmosphere is in a highly oxidizing state that is far from chemical equilibrium.

Composition of the atmosphere*

Molecule	Fraction volume near surface	Vertical distribution
Major constituents		
N_2	7.8084×10^{-1}	Mixed in homosphere; photochemical dissociation high in thermosphere
O_2	2.0946×10^{-1}	Mixed in homosphere; photochemically dissociated in thermosphere, with some dissociation in mesosphere and stratosphere
Ar	9.34×10^{-3}	Mixed in homosphere with diffusive separation increasing above
Important radiative constituents		
CO_2	3.5×10^{-4}	Mixed in homosphere; photochemical dissociation in thermosphere
H_2O	Highly variable	Forms clouds in troposphere; trace amounts in stratosphere; photochemical dissociation above mesosphere
O_3	Variable	Small amounts, 10^{-8}, in troposphere; important layer, 10^{-6} to 10^{-5}, in stratosphere; dissociated above
Other constituents		
Ne	1.82×10^{-5}	
He	5.24×10^{-6}	Mixed in homosphere with diffusive separation increasing above
Kr	1.14×10^{-6}	
CH_4	1.15×10^{-6}	Mixed in troposphere; dissociated in upper stratosphere and above
H_2	5×10^{-7}	Mixed in homosphere; product of H_2O photochemical reactions in lower thermosphere, and dissociated above
NO	$\sim 10^{-8}$	Photochemically produced in stratosphere and mesosphere

*Other gases, for example, CO, N_2O, NO_2, and many by-products of atmospheric pollution also exist in small amounts.

This is in sharp contrast to the atmospheres of Venus and Mars, the planets closest to the Earth, which are composed almost entirely of the more oxidized state, carbon dioxide. The chemical disequilibrium on the Earth is maintained by a continuous source of reactive gases derived from biological processes. Life plays a vital role in maintaining the present atmospheric composition. *See* ATMOSPHERIC CHEMISTRY; MARS; VENUS.

Vertical structure. The total mass of the Earth's atmosphere is about 5.8×10^{15} tons (5.3×10^{15} metric tons). The vertical distribution of gaseous mass is maintained by a balance between the downward force of gravity and the upward pressure-gradient force. The balance is known as the hydrostatic balance or the barometric law. Hence, the declining atmospheric pressure that is measured while ascending in the atmosphere is a result of gravity. The globally averaged pressure at mean sea level is 1013.25 millibars (101,325 pascals).

Below about 60 mi (100 km) in altitude, the atmosphere's composition of major constituents is very uniform. This region is known as the homosphere to distinguish it from the heterosphere above 60 mi (100 km), where the relative amounts of the major constituents change with height. In the homosphere, there are sufficient atmospheric motions and a short enough molecular free path to maintain uniformity in composition. Above the boundary between the homosphere and the heterosphere, known as the homopause or turbopause, the mean free path of the individual molecules becomes long enough that gravity is able to partially separate the lighter molecules from the heavier ones. The mean free path is the average distance that a particle will travel before encountering a collision. Hence, the average molecular weight of the heterosphere decreases with height as the lighter atoms dominate the composition.

Radiative transfer. The vertical structure of the atmosphere (**Fig. 1**) is in large part determined by the transfer properties of the solar and terrestrial radiation streams. The energy of the smallest unit of radiation, the photon, is directly proportional to its frequency. The type of interaction that occurs between photons and the atmosphere depends on the energy of the photons. *See* PHOTON.

The most energetic of the photons are x-rays and extreme ultraviolet radiation of the electromagnetic spectrum, which are capable of dissociating and ionizing the gaseous molecules. The less energetic near-ultraviolet photons are able to excite molecules and atoms into higher electronic levels. As a result, most of the ultraviolet and x-ray radiation is attenuated by the upper atmosphere. A cloudless atmosphere is relatively transparent to visible light, where most of the solar energy resides. At the opposite end of the spectrum, toward the lower frequencies of radiation, is the infrared which is capable of inducing various vibrational and rotational motions in triatomic and polyatomic molecules.

In order to maintain an energy balance, the Earth must emit to space about the same amount of radiation as it absorbs from the Sun (**Fig. 2**). The terrestrial radiation occurs in the infrared part of the spectrum and is strongly affected by water vapor, clouds, carbon dioxide, and ozone and other trace gases. The ability of these gases to absorb and emit in the infrared allows them to effectively trap some of the outgoing radiation that is emitted by the surface, creating the so-called greenhouse effect which is responsible for making the average global surface temperature about 59°F (15°C). The Earth's carbon dioxide concentration is increasing as a result of fossil-fuel burning and deforestation. Since carbon dioxide absorbs and emits infrared radiation, the greenhouse effect is being enhanced, and this is responsible for much of the currently observed global warming which may become a major environmental problem in this century. *See* GREENHOUSE EFFECT; INSOLATION; TERRESTRIAL RADIATION.

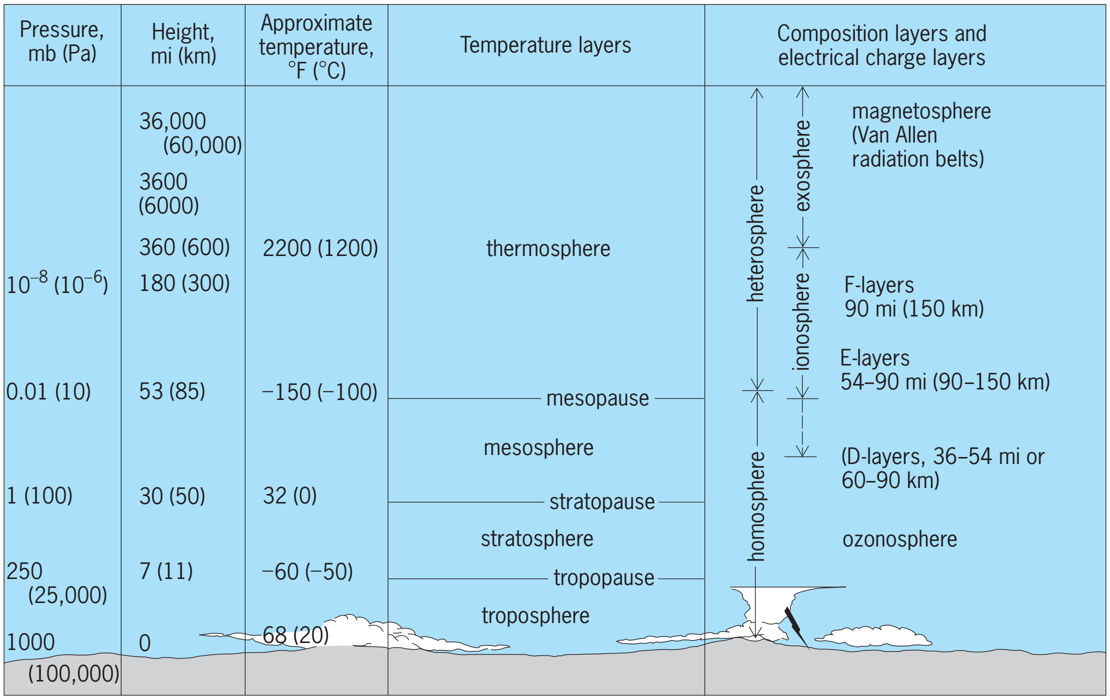

Pressure, mb (Pa)
Height, mi (km)
Approximate temperature, °F (°C)
Temperature layers
Composition layers and electrical charge layers
36,000 (60,000)
3600 (6000)
360 (600)
2200 (1200)
thermosphere
10^−8 (10^−6)
180 (300)
0.01 (10)
53 (85)
−150 (−100)
mesopause
mesosphere
1 (100)
30 (50)
32 (0)
stratopause
stratosphere
250 (25,000)
7 (11)
−60 (−50)
tropopause
troposphere
1000 (100,000)
0
68 (20)
heterosphere
homosphere
exosphere
ionosphere
magnetosphere (Van Allen radiation belts)
F-layers 90 mi (150 km)
E-layers 54–90 mi (90–150 km)
(D-layers, 36–54 mi or 60–90 km)
ozonosphere

Fig. 1. Layers of the atmosphere, shown in terms of pressure, height, temperature, and compositional properties. (*After M. Neiburger, J. G. Edinger, and W. D. Bonner, Understanding Our Atmospheric Environment, W. H. Freeman, 1973*)

Troposphere. The atmospheric layer that extends from the surface to about 7 mi (11 km) is called the troposphere. The tropopause, which is the top of the troposphere, has an average altitude that varies from about 11 mi (18 km) near the Equator to about 5 mi (8 km) near the Poles. The actual tropopause height varies considerably on time scales from a few days to an entire year. *See* TROPOPAUSE; TROPOSPHERE.

The troposphere contains about 80% of the atmospheric mass and exhibits most of the day-to-day weather fluctuations that are observed from the ground. Temperatures generally decrease with increasing altitude at an average lapse rate of about 17°F/mi (6°C/km), although this rate varies considerably, depending on time and location (**Fig. 3**).

On average, about half of the total solar radiative flux incident at the top of the atmosphere is absorbed by the Earth's surface. This raises the surface temperature sufficiently to induce atmospheric circulations in the troposphere that redistribute the heat

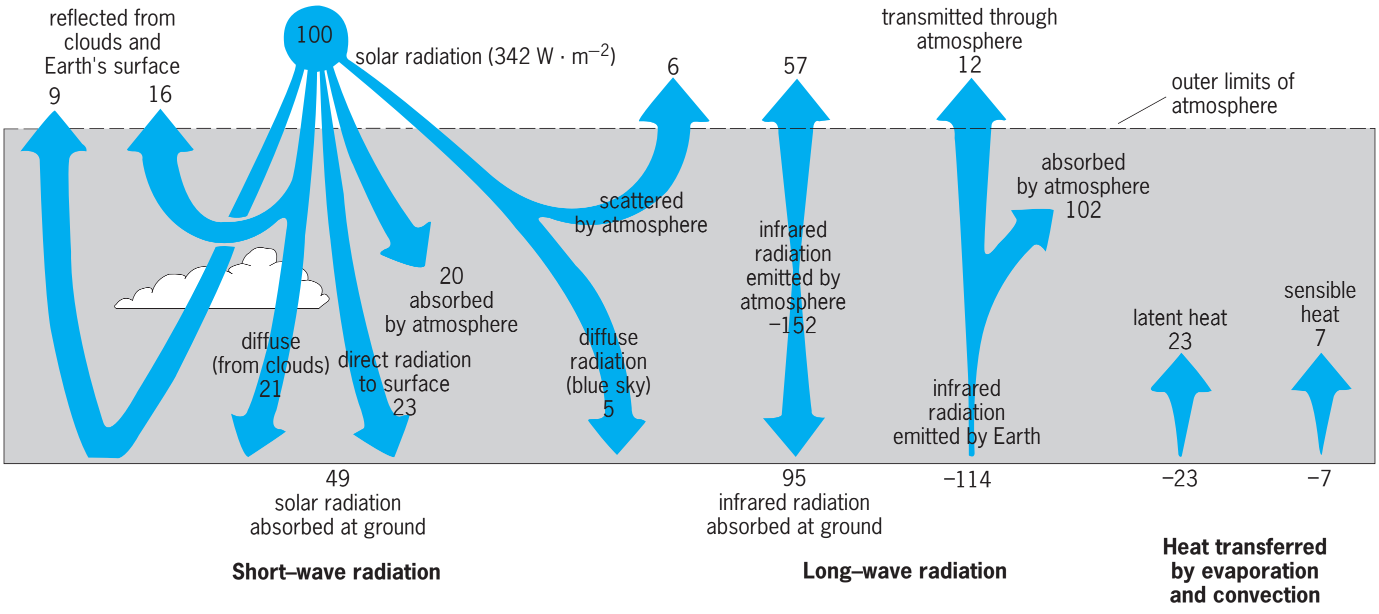

reflected from clouds and Earth's surface
9
16
100
solar radiation (342 W · m^−2)
6
57
transmitted through atmosphere
12
outer limits of atmosphere
absorbed by atmosphere
102
scattered by atmosphere
infrared radiation emitted by atmosphere
−152
20 absorbed by atmosphere
sensible heat
7
latent heat
23
diffuse (from clouds)
21
direct radiation to surface
23
diffuse radiation (blue sky)
5
infrared radiation emitted by Earth
49 solar radiation absorbed at ground
95 infrared radiation absorbed at ground
−114
−23
−7
Short–wave radiation
Long–wave radiation
Heat transferred by evaporation and convection

Fig. 2. Annually and globally averaged radiation energy budget of the Earth's atmosphere expressed as percentages of the total incoming solar radiation; the numerical values sum to a net balance for the surface, atmosphere, and space separately. (*After J. T. Kiehl and K. E. Trenberth, Earth's annual global mean energy budget, Bull. Amer. Meteorol. Soc., 78:197, 1997*)

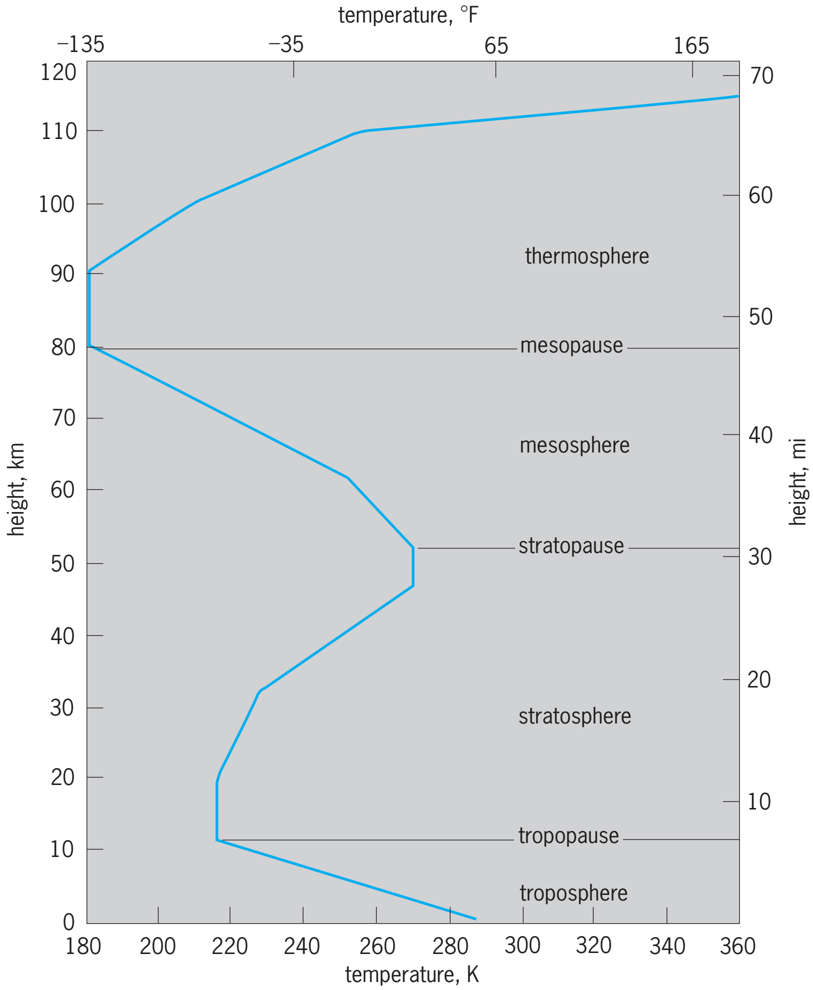


Fig. 3. Temperature of the United States standard atmosphere plotted as a function of height and showing the various thermal layers. The standard atmosphere is intended to be a global and annual representative average. (*After M. Neiburger, J. G. Edinger, and W. D. Bonner, Understanding Our Atmospheric Environment, W. H. Freeman, 1973*)

both vertically and horizontally throughout the troposphere. As an air parcel that is heated from contact with the surface is forced to rise, it becomes surrounded by lower pressure. The air expands and must cool in order to conserve total energy, resulting in the observed decrease of temperature with height.

In order to transfer heat from regions with a radiational surplus, such as the tropics, to regions with a radiational deficit, such as the high latitudes, the atmosphere goes into motion and produces the general circulation relative to the Earth's surface. The large-scale wind patterns are dominated by a balance between the horizontal pressure-gradient force and the Coriolis force. The horizontal pressure gradients form when uneven solar heating creates horizontal temperature gradients. The Coriolis force is a fictitious force that arises because the rotating Earth is an accelerating body. The Coriolis force acts at right angles to the wind and is proportional to the wind speed. Winds that start moving from high to low pressure become deflected to the right in the Northern Hemisphere and to the left in the Southern Hemisphere, so that eventually the winds blow parallel to lines of constant pressure. This is called geostrophic balance and is responsible for the westerlies and the jet streams that are found in the mid and high latitudes and the easterlies in the low latitudes or tropics. *See* ATMOSPHERIC GENERAL CIRCULATION; CORIOLIS ACCELERATION; GEOSTROPHIC WIND; JET STREAM; WIND.

Superimposed on the background westerly and easterly winds are numerous types of disturbances or waves that give the weather its complex character and dominate the transport of energy and momentum in the troposphere. Examples of such disturbances include (1) the Hadley cell, a thermally direct average circulation that extends from latitude 30°S to 30°N and transports heat and momentum away from the Equator; (2) the barotropic disturbance, a wave instability that forms in the presence of a very strong horizontal wind shear; (3) the baroclinic disturbance, a wave instability that grows in the presence of a strong north-south temperature gradient and is responsible for the extratropical cyclones which we observe as low-pressure systems with rain or snow in the midlatitudes; (4) tropical cyclones such as hurricanes and typhoons, disturbances that convert the potential energy of the moist and warm sea surface air into kinetic energy; and (5) thunderstorms, deep convective disturbances that rapidly overturn localized regions of the atmosphere in which relatively cold, dry air overlays warm, moist air. *See* ATMOSPHERIC ELECTRICITY; DYNAMIC INSTABILITY; HURRICANE; THUNDERSTORM.

The winds not only transport heat in a form that can be felt directly, namely sensible heat, but also transport latent heat, which is a form of heat that is carried by the water vapor. Latent refers to the fact that the heat which is absorbed when water is evaporated into the atmosphere is released only when the water condenses to form clouds and precipitation in the atmosphere. The global hydrological cycle, which includes the clouds and precipitation, is a major contributor to the global redistribution of heat. *See* HYDROMETEOROLOGY.

The amount of water vapor present in a parcel of air can be measured by the partial vapor pressure that is exerted by those water molecules. Since the maximum water vapor pressure (saturation vapor pressure) decreases exponentially with decreasing temperature, most of the water vapor condenses out to form clouds as a moist air parcel ascends through the troposphere. It is sufficiently cold at the tropopause (−110 to −40°F or −80 to −40°C) that very little water is able to escape above it, an effect known as the cold trap. As a result of the cold trap, the air is very dry and usually cloud-free above the tropopause.

Stratosphere. This is the atmospheric layer that extends from the tropopause up to the stratopause at about 30 mi (50 km) above the surface. It is characterized by a nearly isothermal layer in the first 6 mi (10 km) overlaid by a layer in which the temperature increases with height to a maximum of about 32°F (0°C) at the stratopause. *See* STRATOSPHERE.

The reversal in the temperature lapse rate is a result of direct absorption of solar radiation, mainly by ozone and oxygen at the ultraviolet frequencies. Ozone (O_3), is a poisonous and reactive gas. Most of the ultraviolet light is attenuated in the stratosphere by ozone and oxygen; hence the surface of the Earth is protected from receiving doses that are harmful to life. In spite of the great importance of stratospheric ozone, it is merely a trace gas with peak concentrations of about 10–15 parts per million by volume at an altitude of about 20 mi (33 km) in the tropics. Ozone concentrations are also influenced by stratospheric winds and hence are highly variable in space and time. Ozone photochemistry is highly complex and involves chemical cycles from the hydrogen, nitrogen, and chlorine groups. It is the chlorine chemistry that has received particular attention in light of the declining ozone concentrations over the Antarctic during spring in the Southern Hemisphere, creating the so-called ozone hole. The chlorofluorocarbons emitted by certain industrial activities have contributed to this decline. *See* PHOTOCHEMISTRY.

The reversal of the temperature lapse rate makes the stratosphere vertically stable. This stability limits the amount of vertical mixing and results in molecular residence times of many months to years. Another consequence of a stable stratosphere is that it acts as a lid on the troposphere, confining the strong vertical overturning and the surface-based weather phenomena of the troposphere. *See* WEATHER.

Occasionally a strong volcano injects large amounts of aerosols, commonly sulfate particles, into the stratosphere that may reside there for several years before they are removed by slow descent. While in the stratosphere, the sulfate particles can perturb the radiative balance by increasing the stratospheric temperature and cooling the troposphere. Significant events of this type were the El Chichón eruption in 1982 and the Pinatubo eruption in 1991. *See* VOLCANO.

Mesosphere. This is the atmospheric layer extending from the stratopause up to the mesopause at an altitude of about 53 mi (85 km). The mesosphere is characterized by temperatures decreasing with height at a rate of about 12°F/mi (4°C/km). Although the mesosphere has less vertical stability than the stratosphere, it is still more stable than the troposphere and does not experience rapid overturning. The coldest temperatures of the entire atmosphere are encountered at the mesopause, with values as low as −150°F (−100°C).

The temperature lapse rate found in the mesosphere is a result of the gradual weakening with height of the direct absorption of solar radiation by ozone. The radiative infrared cooling to space by the carbon dioxide molecules is responsible for the low temperatures near the mesopause. On rare occasions, the temperatures may be low enough for noctilucent clouds to form during the high-latitude summer. *See* MESOSPHERE; NOCTILUCENT CLOUDS.

Thermosphere. This part of the atmosphere is found above the mesopause. The thermosphere is characterized by rising temperatures with height up to an altitude of about 190 mi (300 km) and then is nearly isothermal above that. Although there is no clear upper limit to the thermosphere, it is convenient to consider it extending several thousand kilometers. Embedded within the thermosphere is the ionosphere, comprising those atmospheric layers in which the ionized molecules and atoms are dominant.

Molecular species dominate in the lower thermosphere, while atomic species are dominant above 190 mi (300 km). The distribution of the constituents is controlled by diffusive equilibrium in which the concentration of each constituent decreases exponentially with height according to its molecular weight. Hence the concentration of the heavier constituents such as nitrogen, oxygen, and carbon dioxide will decrease with height faster than the lighter constituents such as helium and hydrogen. At an altitude of 560 mi (900 km), helium becomes the dominant constituent, while hydrogen dominates above 1900 mi (3000 km).

Diffusion becomes the dominant mode of transport in the thermosphere, because the low air densities allow the mean free path of the molecules and atoms to become very large, about 3 ft at 60 mi (1 m at 100 km), and it increases rapidly with height. The thermosphere is exposed to the complete spectrum of radiation that is emitted by the Sun. This includes the extreme ultraviolet (wavelengths less than about 103 nanometers) and some x-ray radiation that is capable of dissociating and ionizing molecules. Although the amount of energy contained in these high-frequency (short-wavelength) radiation bands is small compared to the total solar flux in the visible band, they still dominate the radiative heating in the thermosphere. *See* SOLAR RADIATION.

Of the total absorbed solar energy in the thermosphere, about 35% goes into heating the neutral particles, raising the kinetic temperature, and about 45% is reradiated out of the thermosphere as ultraviolet airglow. Airglow occurs when molecules and atoms reemit temporarily absorbed single photons of solar radiation. The remainder of the incoming absorbed solar radiation (about 20%) is stored as chemical energy of oxygen atoms formed when an oxygen molecule is dissociated. This energy is released as heat when oxygen atoms recombine below 60 mi (90 km) in altitude. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; AIRGLOW.

In order to maintain thermal balance, the thermosphere must lose the heat it receives from the absorption of solar radiation. The emission of thermal, infrared radiation is weak, and so balance is achieved by thermal conduction of heat from the high temperatures of the upper thermosphere to the coldest temperatures at the mesopause. At the mesopause, there is sufficient carbon dioxide and ozone pressure to allow heat to be emitted radiatively to space. The reason that thermal conduction is such a good heat-transfer mechanism in the thermosphere, and not in the lower atmospheric layers, is the large mean free path of the atoms and molecules at these heights. *See* HEAT RADIATION.

The upper thermosphere, above 190 mi (300 km), is a very active region both thermally and dynamically; it exhibits large responses to changes in the solar radiative forcing and auroral particles. Temperatures can vary from about 600°F (300°C) at night to 2200°F (1200°C) during the day. A global circulation occurs in the thermosphere, with winds between 110 and 450 mi/h (50 and 200 m/s) blowing from the day to night side. This prevents the diurnal temperature change from becoming even larger. Also, auroral storms can induce a high- to low-latitude circulation in the thermosphere. *See* AURORA.

Ionosphere. This can be defined operationally as that part of the atmosphere that is sufficiently ionized to affect the propagation of radio waves. In the ionosphere, the dominant negative ion is the electron, and the main positive ions include O^+, NO^+, and ${O_2}^+$.

The ionosphere is classified into four subregions. The D region extends from 40 to 60 mi (60 to 90 km) and contains complex ionic chemistry; most of the ionization is caused by ultraviolet ionization of NO and by galactic cosmic rays. This region is responsible for the daytime absorption of radio waves, which prevents distant propagation of certain frequencies. The E region extends from 60 to 90 mi (90 to 150 km) and is caused primarily by the x-rays from the Sun. The F1 region from 90 to 125 mi (150 to 200 km) is caused by the extreme ultraviolet radiation from the Sun and disappears at night. The F2 region includes all the ionized particles above 125 mi (200 km), with the peak ion concentrations occurring near 190 mi (300 km). *See* COSMIC RAYS; IONOSPHERE.

Exosphere. The term exosphere is used to refer to the atmosphere above 300 mi (500 km), where the probability of interatomic collisions is so low that some of the atoms traveling upward with sufficient velocity can escape the Earth's gravitational field. The dominant escaping atom is hydrogen since it is the lightest constituent. Calculations of the thermal escape of hydrogen (also known as the Jeans escape) yield a value of about 3×10^8 atoms $\cdot$ $cm^{-2} \cdot s^{-1}$. This is a very small amount since at this rate less than 0.5% of the oceans would disappear over the current age of the Earth.

The main source of the escaping hydrogen is water vapor that becomes dissociated by ultraviolet radiation above the tropopause. The importance of the cold water trap at the tropopause is evident, as this limits the availability of water for dissociation and provides a limit on the hydrogen escape. It is speculated that Venus may have lost all its water because such a cold trap did not limit the hydrogen escape.

Magnetosphere. This is the region surrounding the Earth where the movement of ionized gases is dominated by the geomagnetic field. The lower boundary of the magnetosphere, which occurs at an altitude of nearly 75 mi (120 km), can be roughly defined as the height where there are enough neutral atoms that the ion-neutral particle collisions dominate the ion motion. The dynamics of the magnetosphere is dictated in part by its interaction with the plasma of ionized gases that blows away from the Sun, the solar wind. The solar wind interacts with the Earth's magnetic field and severely deforms it, producing a magnetosphere around the Earth. It extends about 40,000 mi (60,000 km) toward the Sun, but extends beyond the orbit of the Moon away from the Sun. *See* MAGNETOSPHERE; SOLAR WIND; VAN ALLEN RADIATION.

Vertical energy transport. The division of the Earth's atmosphere into layers is based primarily on the thermal and chemical properties of each layer. Even though physical properties are found that distinguish the layers from each other, it is important to realize that interactions of mass and energy do occur between the layers.

The most obvious energy transfer is the outgoing terrestrial infrared radiation that escapes from the Earth. However, atmospheric dynamicists have also been studying the upward transfer of energy and momentum by the vertical propagation of waves. As a wave propagates upward, it encounters lower air densities, and the wave amplitude grows to conserve the total energy flux. Eventually the wave amplitude becomes so large that the wave is said to break, not unlike ocean waves breaking on the shoreline, and it deposits its energy and momentum at that height. The dissipation that occurs when a wave breaks is achieved by molecular conduction, viscosity, and ion drag.

It is believed that wave breaking may play an important role in the heating and circulation patterns of certain upper atmospheric regions. The two main wave types are atmospheric tides and gravity waves. Atmospheric tides are initiated in the stratosphere when ozone is heated, and can produce a vertically propagating wave. Gravity waves are oscillations produced by the stable buoyancy force; they also grow in amplitude as they move upward. *See* AERONOMY; ATMOSPHERIC TIDES; METEOROLOGY; UPPER-ATMOSPHERE DYNAMICS.

Glen B. Lesins

Bibliography. R. M. Goody, *Principles of Atmospheric Physics and Chemistry*, 1995; J. T. Houghton, *The Physics of Atmospheres*, 3d ed., 2002; J. M. Wallace and P. V. Hobbs, *Atmospheric Science: An Introductory Survey*, 1977; R. P. Wayne, *Chemistry of Atmospheres*, 3d ed., 2000.

Atmosphere, evolution of

Variation with time of the chemical composition and total weight of the Earth's atmosphere. The atmosphere is a most tenuous envelope; its mass is less than one-millionth that of the solid Earth; its density even at sea level is less than one-thousandth that of rocks, and virtually all of the atmosphere is below a height only one-hundredth of an earth radius above the surface of the Earth. But the atmosphere is taken so much for granted that one tends to be surprised at the thought that it has a history, that its chemical composition and total weight have varied through time. On reflection, it would, however, be odd to find that the atmosphere has not changed during the long years of the Earth's existence, and that its weight and composition have not responded to the complicated series of events that have left such clear marks on the Earth's crust.

The study of atmospheric evolution is difficult, because there seems to be no way of obtaining reliable samples of the atmosphere that are more than about 400,000 years old. But the quest is not altogether hopeless. The origin of life, the continued existence of animals since at least 600 million years ago (Ma), the marks which interaction of the atmosphere with surface rocks have left on ancient sediments, and the nature of volcanic gases through geologic time all give clues about the chemical evolution of the atmosphere. While it is not possible to solve all the puzzles of the history of the Earth's atmosphere with these clues, the broad pattern of atmospheric evolution is emerging; and a more complete understanding is within reach.

It may be wise to start by defining the problem rather precisely. Today the atmosphere contains a small number of major components and a very large number of minor and trace components. Each component exerts a pressure, which is essentially constant for some components at sea level and which is variable in time and space for other components (**Tables 1** and **2**). The problem of the chemical evolution of the Earth's atmosphere can be considered solved when an accurate plot can be made of the variation with time of the pressure of these components and others that may have existed in the atmosphere in the past. This is obviously a large undertaking, but it is made somewhat less so by the mutual exclusion of certain gases as major components of the atmosphere. At present, for instance, gases such as ammonia (NH_3), methane (CH_4), and hydrogen (H_2) can exist in the atmosphere only as trace components, because they are unstable in the presence of large quantities of oxygen. Conversely, oxygen could not have been a major component of an atmosphere in which ammonia, methane, and hydrogen were abundant. *See* ATMOSPHERIC CHEMISTRY.

TABLE 1. Essentially nonvariable constituents of today's air*

Constituent	Content, %
N_2	78.084
O_2	20.946
CO_2	0.036
A	0.934
Ne	18.18×10^{-6}
He^4	5.24×10^{-6}
He^3	6.55×10^{-12}
Kr	1.14×10^{-6}
Xe	0.087×10^{-6}
H_2	0.5×10^{-6}
N_2O	0.5×10^{-6}

*The classification of H_2 and N_2O as nonvariable constituents is uncertain.

TABLE 2. Variable constituents of today's dry air

Constituent	Content
O_3	0 to 0.07 ppm (summer) 0 to 0.02 ppm (winter)
SO_2	0 to 1 ppm
NO_2	0 to 0.02 ppm
CH_4	1 to 2 ppm
CH_2O	Uncertain
I_2	0 to 10^{-10} g · cm^{-3}
NaCl	Order of 10^{-10} g · cm^{-3}
NH_3	0 to trace
CO	0 to trace (0.8 cm atm)

Origin of free oxygen. Free oxygen is somewhat of an anomaly on the Earth. Most rocks that are more than a few feet below the Earth's surface are out of equilibrium with free oxygen and are oxidized in contact with the atmosphere. This is seen in the development of red hydrous ferric oxide minerals in soil zones above many rock types in temperate and tropical areas and in the presence of the red to reddish-brown sediments that are so common in the western United States.

The origin of oxygen in the Earth's atmosphere has been a source of continuing controversy for many decades. Of the two theories that have been dominant, the first proposes that atmospheric oxygen has been produced through geologic time by green plant photosynthesis, during which carbon is effectively separated from oxygen in carbon dioxide. In a very rough manner this can be written as reaction (1).

$$6CO_2 + 6H_2O \xrightarrow{\text{photo-synthesis}} \underbrace{C_6H_{12}O_6}_{\text{Carbohydrate}} + 6O_2 \tag{1}$$

This reaction runs in the opposite direction during the decay of plant material and the breathing of animals. It can be shown that nearly all of the oxygen produced by photosynthesis is consumed by respiration, but the small amount that is not lost in this manner could easily account for the present, rather large quantity of atmospheric oxygen.

The second theory proposes that ultraviolet light from the Sun decomposes water molecules in the upper atmosphere. Most of the water molecules recombine, but there is a finite possibility that some hydrogen atoms will escape from the Earth's atmosphere before recombination has taken place. Oxygen atoms, being 16 times as heavy as hydrogen atoms, escape much more slowly or not at all. The decomposition of water vapor followed by hydrogen escape is therefore a distinctly plausible manner of generating free atmospheric oxygen. The escape rate of hydrogen from the atmosphere is almost certainly rapid. The critical factor in determining escape rates is the temperature of the upper atmosphere, and this is now well known from rocket measurements. But a strongly limiting factor for oxygen production by this mechanism is the formation of ozone (O_3) in the stratosphere. Ozone absorbs ultraviolet light very readily, and forms a screen that prevents ultraviolet light from reaching the lower levels of the atmosphere, where water vapor is abundant. Most of the hydrogen that manages to reach the upper atmosphere today is a constituent not of water vapor but of biologically produced methane, and the rate of hydrogen escape from the Earth's atmosphere is controlled largely by the rate of upward transport of this compound. Nevertheless, the hydrogen loss rate is currently a very small fraction of the oxygen production rate by photosynthesis. The oxygen content of the present atmosphere is therefore only slightly influenced by hydrogen loss from the upper atmosphere, and depends almost exclusively on the operation of the feedback system that links the rate of oxygen production during photosynthesis to the rate of oxygen use by weathering and the decay of organic matter. *See* PHOTOSYNTHESIS.

Biologic evidence. The proposition that the present high concentration of oxygen in the atmosphere is largely due to oxygen production during photosynthesis implies that oxygen was much less abundant prior to the existence of photosynthetic organisms. This view is in harmony with the requirement that free oxygen was absent from the atmosphere during the development of life. *See* PREBIOTIC ORGANIC SYNTHESIS.

It is likely that life evolved very early in Earth history. The oldest known unmetamorphosed sedimentary rocks, those found near North Pole, Australia, are about 3.5 billion years old and contain microscopic bits of carbon that are probably biologic in origin. They also have stromatolitic structures, which may be the work of cyanobacteria. There is little doubt, then, that life began more than 3.5 billion years ago (Ga), and there is reasonably compelling evidence that life began more than 3.8 Ga. It is not yet clear when organisms that were able to produce free oxygen developed, but evidence from the isotopic composition of carbon in sedimentary rocks suggests that this occurred before 2.7 Ga. *See* STROMATOLITE.

The biologic evidence for oxygen at levels approaching those of the present is still scant until the close of the Precambrian Era, some 543 Ma. The sequence of animals that developed during the latest part of the Precambrian Era and the beginning of the Phanerozoic Era strikingly parallels the zoning of animals in contemporary settings of progressively greater oxygen content. This correspondence suggests, but does not prove, that evolutionary events were related to changes in the level of atmospheric oxygen. *See* PRECAMBRIAN.

It has been proposed that the invasion of the land by plants and animals about 400 Ma occurred when the oxygen pressure had risen to about one-tenth of its present value. At this oxygen pressure the intensity of ultraviolet radiation at the Earth's surface would have been so low that it no longer presented a health hazard. Although this is no more than an interesting speculation, the persistence of animals requiring oxygen in rather large quantities indicates that oxygen levels in the atmosphere have probably never been lower than two-tenths of the present value during the past 400 million years. There may have been times when the oxygen pressure was greater than at present.

Most, if not all, coal deposits contain charcoal. This fact implies that during the past 350 Ma the atmosphere has contained more than about 15% oxygen, the minimum oxygen level required to sustain fires. Many coal deposits also contain the remains of very large trees. If the oxygen content of the atmosphere had been more than 30–35%, forest fires would probably have been so frequent and so intense that trees would not have been able to grow to such great heights. The evidence from coal deposits therefore indicates that the oxygen content of the atmosphere has been between 15 and 35% during the past 350 Ma.

Evidence from sediments. Today oxidation of rocks at the Earth-atmosphere interface is pervasive. If oxygen was essentially absent from the atmosphere in times past, one might expect to see relatively less oxidation in the minerals of ancient sediments, and to see the formation of new minerals in these sediments, which are less oxidized than their modern counterparts. Iron, manganese, uranium, and sulfur are among the elements which today respond most readily to oxidation at the Earth–atmosphere interface. Uraninite (UO_2), for instance, reacts rapidly with atmospheric oxygen to form a variety of higher oxides and hydrous oxides. A concerted search for uraninite in black sands during World War II was unsuccessful. And yet, uraninite, which has apparently survived weathering and transport, occurs as ores in the sedimentary rocks of the Dominion Reef series and the Witwatersrand series of South Africa (**Figs. 1** and **2**), as well as at Blind River, Canada, and at Serra

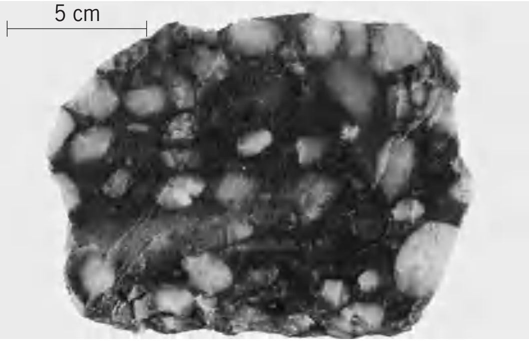


Fig. 1. Polished slab of gold ore, Modder Deep, South Africa. The white pebbles are quartz, SiO_2, and the major metallic constituent is pyrite, FeS_2. Most of the gold ore also contains uraninite, UO_2, which is apparently detrital.

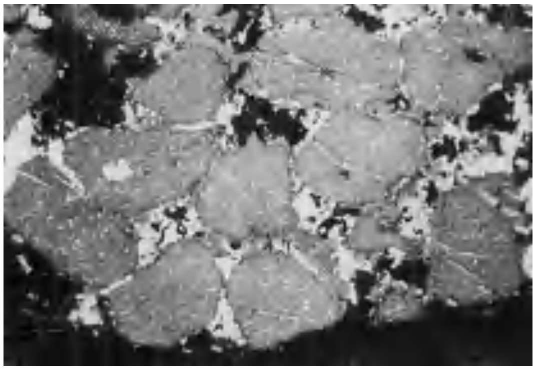


Fig. 2. Photomicrograph (×280) of a polished section of uranium ore, West Rand Cons Mine, South Africa. Gray grains are uraninite, UO_2, which are probably detrital. White material in cracks and between grains of uraninite is brannerite, an oxide of uranium, titanium, and calcium. (*From Ramdohr, Abhandl. Deut. Akad. Wiss. Berlin, 1958*)

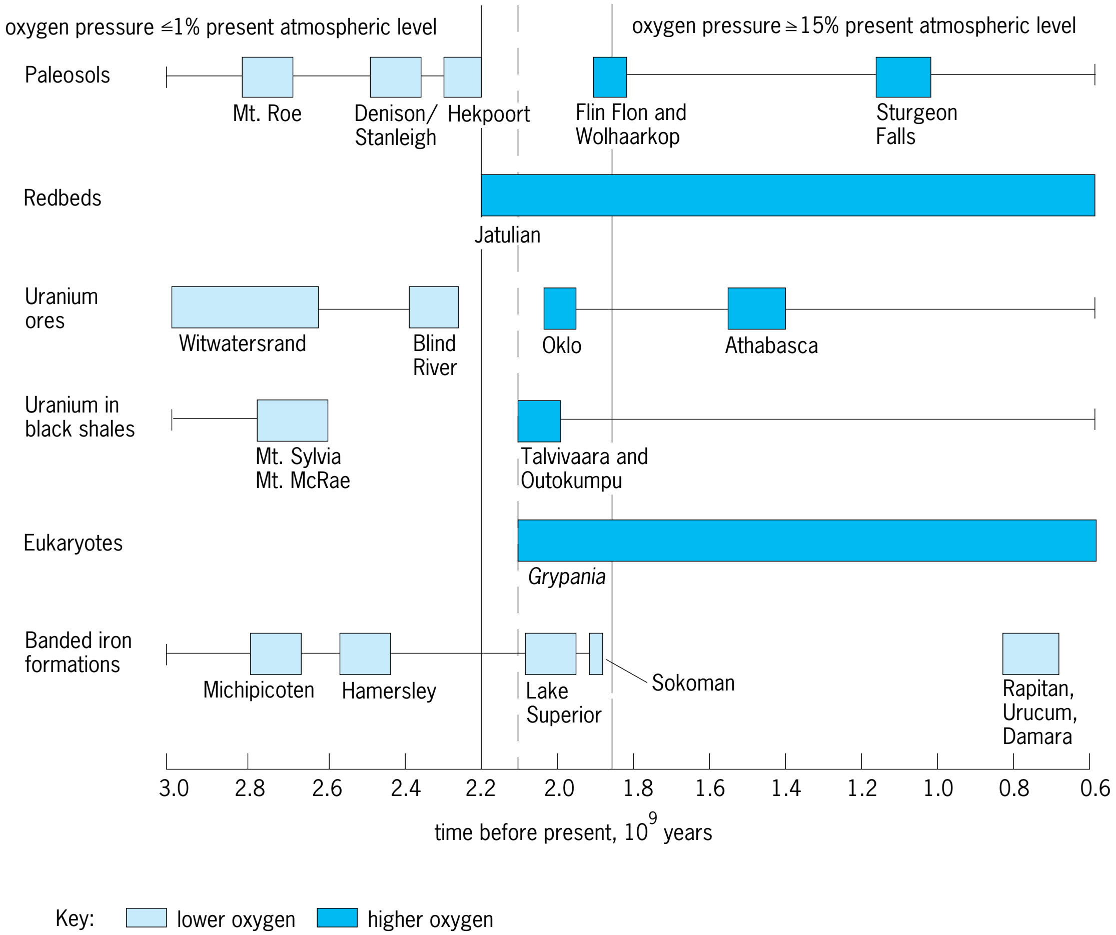


Fig. 3. **Summary of data bearing on the O_2 content of the Precambrian atmosphere. (*After H. D. Holland, Early Proterozoic atmospheric change, in S. Bengston, ed., Early Life on Earth, Columbia University Press, 1994; also R. Rye and H. D. Holland, Amer. J. Sci., 298:621–672, 1998*)**

de Jacobina, Brazil. The sediments in these areas are all more than 2.3 Ga. Nearly all geologists who have closely studied the ore deposits have agreed that much of the uraninite is a residue of weathering. This view has been corroborated by measurements of the age of uraninite grains in the deposits. They appear to be older than the sediments in which they occur. If the grains are indeed detrital, the rate of oxidation of uraninite must have been much slower prior to 2.3 Ga than it is today. Careful studies of the rate of oxidation of uraninite have shown that the oxygen pressure must have been less than about 4×10^{-3} atm (4×10^{2} pascals) to permit the survival of uraninite during weathering, transport, and deposition of these ores.

Between 2.3 and 2.0 Ga the nature of uranium ore deposits changed from detrital to hydrothermal. The formation of hydrothermal uranium deposits requires a significant amount of atmospheric oxygen. A very significant rise in the oxygen level between 2.3 and 2.0 Ga is therefore indicated. Australian sandstones deposited between 2.75 and 3.25 Ga also contain other detrital minerals that are easily oxidized and that are absent from younger sandstones; this confirms the evidence from uranium minerals for a low-O_2 atmosphere prior to 2.7 Ga. Major differences between the oxidation state of iron in soils formed prior to 2.3 Ga and in those formed more recently than 1.9 Ga (**Fig. 3**) point to the same conclusion, as does the first appearance of highly oxidized redbeds about 2.2 Ga. A major positive excursion in the isotopic composition of carbon in carbonate rocks deposited between 2.25 and 2.06 Ga also suggests that oxygen levels climbed dramatically during this time interval (**Fig. 4**). In defining the changes in the isotopic composition, delta (δ) terms are used. The delta term for carbon is defined in Eq. (2), showing the

$$\delta^{13}\mathrm{C} = \left[\frac{\left(\frac{^{13}\mathrm{C}}{^{12}\mathrm{C}}\right)_{\mathrm{sample}} - \left(\frac{^{13}\mathrm{C}}{^{12}\mathrm{C}}\right)_{\mathrm{standard}}}{\left(\frac{^{13}\mathrm{C}}{^{12}\mathrm{C}}\right)_{\mathrm{standard}}}\right] \times 1000‰ \quad (2)$$

ratio of ^{13}C to ^{12}C in a sample and in a standard, which by convention is a Peedee belemnite. For sulfur, the delta term is defined in Eq. (3).

$$\delta^{34}\mathrm{S} = \left[\frac{\left(\frac{^{34}\mathrm{S}}{^{32}\mathrm{S}}\right)_{\mathrm{sample}} - \left(\frac{^{34}\mathrm{S}}{^{32}\mathrm{S}}\right)_{\mathrm{standard}}}{\left(\frac{^{34}\mathrm{S}}{^{32}\mathrm{S}}\right)_{\mathrm{standard}}}\right] \times 1000‰ \quad (3)$$

Evidence from gaseous emissions. A third line of evidence is available to tell something about the oxidation state of the atmosphere in a nonbiotic state. W. W. Rubey showed that many of the volatile

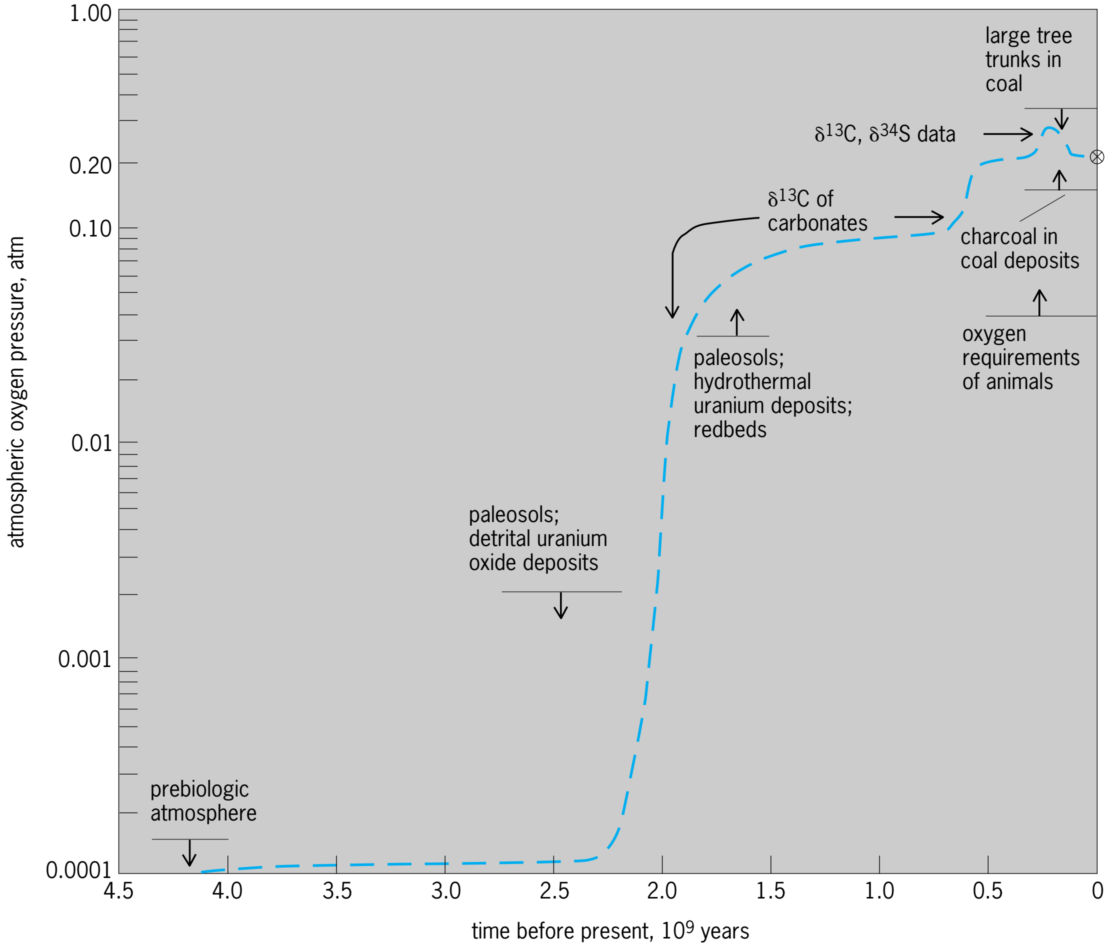


Fig. 4. Summary of the evidence for the evolution of atmospheric oxygen since the Earth was formed. 1 atm = 10^5 Pa.

constituents of the atmosphere and oceans cannot have been derived from the weathering of igneous rocks. He developed a very plausible argument for the concept that these "excess volatiles" have boiled out of the interior of the Earth during the course of its long history. Volcanoes give eloquent evidence for such boiling out today (**Fig. 5**), and it seems likely that a portion of the discharge of at least some hot springs has a deep source. In a nonbiotic state the chemistry of the atmosphere would be controlled largely by the chemistry of such emanations and by their interaction with surface rocks. Thus, if something were known about the oxidation of volcanic gases in the past, it might be possible to predict, at least within broad limits, the oxidation state of an atmosphere unaffected by biologic processes.

Fig. 5. Eruption of Mount Vesuvius in 1943. The gases emitted by volcanoes throughout geologic time have played a major role in the chemical evolution of the atmosphere of the Earth.

At present, volcanic gases consist mainly of water vapor, carbon dioxide, sulfur dioxide, hydrogen, carbon monoxide, and nitrogen (**Table 3**). Free oxygen is almost completely absent: The oxygen pressure

TABLE 3. Composition of typical Hawaiian volcanic gases

Gas	Vol %
H_2O	79.3
CO_2	11.6
SO_2	6.5
N_2	1.3
H_2	0.6
CO	0.4
Cl_2	0.05
Ar	0.04

in these gases as they emerge is about 10^{-7} atm (10^{-2} Pa). If a gas mixture is defined as being neutral from an oxidation-reduction point of view when it contains neither free hydrogen nor free oxygen, then these gases are slightly on the reduced side, since they contain a small amount of free hydrogen (H_2) and carbon monoxide (CO). Today these reduced gases react rapidly with atmospheric oxygen. In the absence of atmospheric oxygen, a variety of different reactions might take place. All of these tend to produce an essentially neutral atmosphere which would be similar to the present atmosphere, with the difference that oxygen would be virtually absent.

The oxidation state of volcanic gases today is controlled in large part by the oxidation state of the associated lavas. This in turn is reflected in the ratio of ferrous iron (Fe^{2+}) to ferric iron (Fe^{3+}) in lavas, a ratio which tends to be preserved on cooling. Studies of this ratio in basalts have shown that the Fe^{2+}/Fe^{3+} ratio has probably not changed significantly during at least the past 2 Ga. It is fairly certain, then, that the oxidation state of volcanic gases has not varied greatly during the second half of Earth history. There are good theoretical reasons for believing that this same state of affairs prevailed at least back to 3 Ga, but it is possible that shortly after its birth the Earth vented volcanic gases which were more highly reducing. The cause for this difference would have been the presence of metallic iron in the upper part of the Earth's mantle. Today this iron is largely concentrated in the Earth's core, far below the part of the mantle where lavas are generated. *See* EARTH INTERIOR.

Rise of oxygen pressure. It may be well to summarize the evidence regarding the history of the pressure of oxygen in the Earth's atmosphere before considering the pressure of the other constituents through geologic time. The largest area of ignorance surrounds the events prior to the accumulation of the earliest known sediments, some 3.8 Ga. It is possible that the atmosphere was quite reducing as a consequence of the injection of highly reduced volcanic gases into the atmosphere during the first 100 Ma of Earth history.

Shortly after 3 Ga the atmosphere probably contained less than 4×10^{-3} atm (4×10^{2} Pa) oxygen. Green plant photosynthesis was already under way, but the rate of burial of organic carbon was insufficient to support more than a small oxygen pressure in the atmosphere. This condition probably prevailed until about 2.2 Ga. The first widespread redbed sequences were deposited at about that time, and no uranium deposits of the Witwatersrand–Blind River type younger than 2.3 Ga have been discovered. Both observations are consistent with a major increase in atmospheric oxygen between 2.3 and 2.0 Ga.

Another rise may well have taken place at the end of the Precambrian and may have been important for the development of animal life. Finally, the colonization of the land about 400 Ma may have caused a further rise in the oxygen pressure. Since that time the oxygen pressure has probably not fluctuated greatly (Fig. 4).

Rare gases other than helium. Of the atmospheric constituents other than oxygen, the rare gases, nitrogen, and carbon dioxide are probably the most interesting. Studies have shown that the Earth is very depleted in rare gases in comparison with the Sun. Even xenon, a heavy element which certainly cannot escape from the Earth's atmosphere today at anything but a geologically insignificant rate, appears to be only about one-millionth as abundant in comparison with silicon as in the Sun. From these studies it was concluded that the Earth was essentially devoid of an atmosphere after it had reached its present size and gravitational field, although it is possible that an original gaseous envelope containing the missing rare gases was swept away by strong magnetic fields which penetrated throughout the solar system, or during the violent Hayashi phase early in the history of the Sun, or during an impact with a Mars-size object. Whichever view turns out to be correct, the Earth seems to have been left nearly devoid of an atmospheric blanket at some time not long after the end of the accumulation phase.

Rubey's argument for the degassing of the Earth's interior is strongly supported by the rather anomalous abundance and isotopic composition of atmospheric argon. This gas is much more abundant than the other rare gases (Table 1), and consists almost entirely of the argon isotope of mass 40, which in the normal course of nuclear events should not be much more abundant than the argon isotopes of mass 36 and mass 38. The anomaly is readily explained by the degassing of argon-40 from the Earth after its production from the radioactive decay of potassium-40. Although the potassium content of the Earth is not well known, there is little doubt that the decay of the available potassium could account for the present abundance of atmospheric argon-40. It seems likely that degassing was most intense during the early history of the Earth, but the discovery of excess helium-3 in the Pacific Ocean indicates that the release of primordial gases is continuing from mid-ocean ridges. The atmospheric pressure of neon, argon, krypton, and xenon has probably increased gradually during the course of Earth history. *See* ARGON; INERT GASES; RADIOISOTOPE.

Other gases. The abundance of helium in the atmosphere is anomalously low. The quantity of helium which has been generated in the solid Earth by the decay of uranium and thorium series nuclides and has entered the atmosphere with other gases is much larger than the quantity of helium now present in the atmosphere. Helium has clearly escaped from the atmosphere into interplanetary space. *See* HELIUM.

Nitrogen is a good deal more reactive than the rare gases. It is cycled through living organisms at a geologically rapid rate but seems to reside mainly in the atmosphere. From a thermodynamic point of view, nitrate should be quite abundant in the oceans today. Its concentration in seawater is kept to very low levels by organisms for whom nitrate is an important and frequently limiting nutrient. *See* NITROGEN.

Ammonia is thermodynamically unstable in the presence of free oxygen; a major fraction of the nitrogen that is now in the atmosphere would be converted to ammonia only if the hydrogen pressure were in excess of 10^{-3} atm. Such a hydrogen pressure could have existed only early in the Earth's history.

The value of the hydrogen pressure, and even the very existence of a highly reducing early atmosphere, is presently very uncertain. R. Brett pointed out that the relatively large nickel content of olivine in the upper mantle is inconsistent with the presence of large amounts of metallic iron in the upper mantle early in Earth history. If this is correct, volcanic gases may never have been much more reducing than at present.

Methane could have been a major constituent of an early reducing atmosphere. It is likely that its pressure was already less than 10^{-3} atm (10^2 Pa) during the deposition of the sediments of the Isua area in West Greenland 3.8 Ga. Methanogens, that is, organisms that generate methane, may have been a significant part of the early biosphere, and may have been responsible for the presence of a significant amount of methane in the early atmosphere. However, methane must have been a very minor component ever since atmospheric oxygen became more than a trace constituent of the atmosphere. The very small concentration of CH_4 since then has almost certainly been controlled by the balance between its rate of biologic production and the rate of its photochemical decomposition in the atmosphere. *See* METHANE.

It seems likely that the carbon dioxide pressure has been reasonably close to the value at which hydrous magnesium aluminum silicate minerals are in equilibrium with kaolinite, quartz , calcite, and dolomite [reaction (4)]. All of these minerals are abundant in

$$\underset{\text{Chamosite}}{Mg_5Al_2Si_3O_{10}(OH)_8} + \underset{\text{Calcite}}{5CaCO_3} + 5CO_2 \rightarrow \underset{\text{Dolomite}}{5CaMg(CO_3)_2} + \underset{\text{Kaolinite}}{Al_2Si_2O_5(OH)_4} + \underset{\text{Quartz}}{SiO_2} + 2H_2O \tag{4}$$

sediments and sedimentary rocks, and it is likely that such reactions have helped to maintain the CO_2 pressure within two orders of magnitude of its present value for much of geologic time. There has been considerable debate concerning the CO_2 content of the atmosphere during the Archean, that is, more than 2.5 Ga. Some geochemists have proposed that the CO_2 pressure was in excess of 10^{-1} atm (10^4 Pa) and possibly as high as 10 atm (10^6 Pa). However, evidence from the composition of paleosols indicates that the CO_2 pressure has probably been less than 3×10^{-2} atm (3×10^3 Pa) since 2.75 Ga.

The analyses of small air bubbles trapped in the ice sheets of Greenland and Antarctica have had a profound impact on thinking about atmospheric evolution and the origin of ice ages. The CO_2 content of the atmosphere has varied by about 50% during the past 160,000 years (**Fig. 6**). Levels of CO_2 during the last interglacial period around 125,000 years ago were about 270 parts per million (ppm) [2.7×10^{-4} atm or 27 Pa]. The CO_2 content in the atmosphere dropped gradually to about 200 ppm during the last ice age but rebounded rapidly to about 270 ppm at the end of the last ice age about 10,000 years ago. Thereafter it varied little until the beginning of the industrial revolution. Since then fossil fuel burning and, to a lesser extent, deforestation have pushed the CO_2 content of the atmosphere up to about 370 ppm (3.70×10^{-4} atm or 37 Pa). The CO_2 pressure is apt to rise to roughly twice that value by the end of the twenty-first century, unless economics and the fear of adverse climate changes produce a major shift away from the burning of fossil fuels as the dominant source of the world's energy.

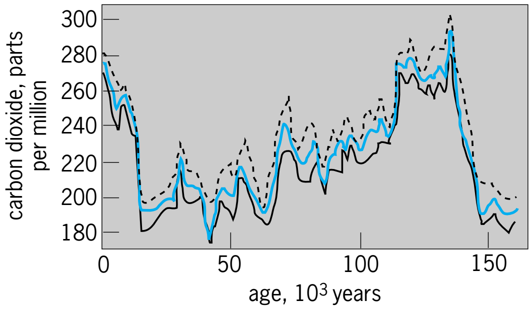


Fig. 6. Carbon dioxide concentration in air bubbles in the Vostok ice core, Vostok Station, Antarctica. (*After C. Lorius, Antarctic ice core: CO_2 and climatic change over the last climatic cycle, EOS, 69(26):681, 683–684, June 28, 1988*)

Today CO_2 is second in abundance to water in gases issuing from volcanoes. Yet the ratio of CO_2 to water in the atmosphere plus the oceans is minuscule. Carbon dioxide has clearly been scavenged very thoroughly from the atmosphere-hydrosphere system. Its resting place is easily discovered in the elemental carbon of sedimentary rocks and in the carbonate rocks that have been deposited throughout the entire range of Earth history that is accessible through the study of sedimentary rocks (**Fig. 7**). The two dominant carbonate minerals, calcite ($CaCO_3$) and dolomite [$CaMg(CO_3)_2$], are the major components of limestones and dolomites. The removal of CO_2 from the atmosphere, its reaction with surface rocks, and its ultimate burial are processes that are chemically, mineralogically, and biologically complex.

Three stages of evolution. The above discussion seems to lead quite naturally to a threefold division of atmospheric history (**Table 4**). During the first stage, very shortly after the accretion of the Earth, the atmosphere may have been quite reducing. Reduced volatiles issuing from volcanoes probably would have given rise to an atmosphere consisting of methane with minor quantities of hydrogen, nitrogen, and ammonia. After metallic iron was removed from the upper mantle, the oxidation state of volcanic gases probably approached its present value, methane was replaced by carbon dioxide, and ammonia was converted to nitrogen. In the atmosphere during this second stage, nitrogen was its dominant component,

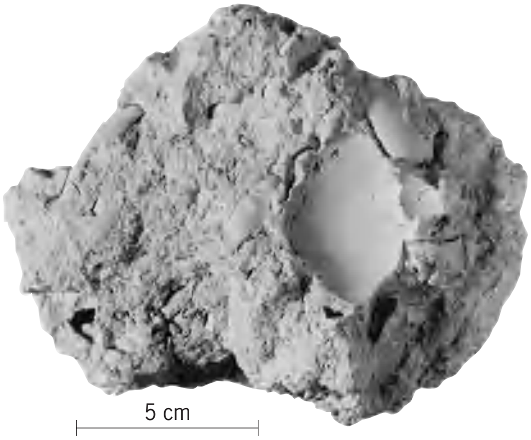


Fig. 7. Sample of Tertiary limestone from Florida. Nearly all carbon dioxide injected into the atmosphere has been removed in limestones and dolomites. During the past 6×10^4 years much removal has taken place through the action of organisms with carbonate shells. Prior to that time, removal by inorganic precipitation and by stromatolites was probably dominant.

and carbon dioxide and argon were its most important minor constituents.

The third stage opened when the rate of oxygen production became sufficiently great that oxygen became more than a trace component of the atmosphere. The transition from stage 2 to stage 3 probably occurred between 2.3 and 2.0 Ga. Since the opening of stage 3, the oxygen pressure has climbed to its present value by a path which was probably simple but could have been complex in detail. During this period the nitrogen, neon, argon, krypton, and xenon pressures have also gradually climbed to their present value, while the helium pressure remained reasonably constant, suspended between the rate of input and the rate of escape. The CO_2 pressure has probably been less than 100 times and more than half the present value during stage 3. There may have been a general decrease toward its present value, but fluctuations around a generally decreasing trend have probably been significant.

TABLE 4. Summary of data on the probable chemical composition of the atmosphere during stages 1, 2, and 3

Components	Stage 1	Stage 2	Stage 3
Major components: $P > 10^{-2}$ atm (10^3 Pa)	CH_4 H_2 (?)	N_2	N_2 O_2
Minor components: 10^{-2} atm $> P >$ 10^{-4} atm (10 Pa)	H_2 (?) H_2O N_2 H_2S NH_3 Ar	H_2O CO_2 A O_2 (?)	Ar H_2O CO_2
Trace components: 10^{-4} atm $> P >$ 10^{-6} atm (10^{-1} Pa)	He	Ne He CH_4 NH_3 (?) SO_2 (?) H_2S (?)	Ne He CH_4 Kr

Society is actively modifying the composition of the atmosphere. The major atmospheric constituents are essentially immune to human impacts, but the CO_2 content of the atmosphere has increased significantly, as has that of methane, and there have been major changes in the concentration of many trace gases. Some of these gases are entirely new to the atmosphere. Their concentration is extremely small, but the effects of some, like the chlorofluorocarbons, have been disproportionately large.

Heinrich D. Holland

Bibliography. S. Bengtson (ed.), *Early Life on Earth*, 1994; H. D. Holland, *The Chemical Evolution of the Atmosphere and Oceans*, 1984; F. S. Rowland and I. S. A. Isaksen (eds.), *The Changing Atmosphere*, 1988; S. H. Schneider and P. J. Boston (eds.), *Scientists on Gaia*, 1991; J. W. Schopf (ed.), *Earth's Earliest Biosphere*, 1992; J. W. Schopf and C. Klein (eds.), *The Proterozoic Biosphere*, 1992.

Atmospheric acoustics

The science of sound in the atmosphere. Atmospheric acoustics is concerned with sound outdoors rather than indoors. The atmosphere has a structure that varies in both space and time, and these variations have significant effects on a propagating sound wave. In addition, when sound propagates close to the ground, the type of ground surface has a strong effect.

Development. Atmospheric acoustics originated with the rebirth of science in the seventeenth century. Early work focused on measuring the speed of sound. In 1636, Marin Mersenne measured this speed by timing the interval between the flash and sound of a gun blast. He obtained the inaccurate value of 448 m/s (1470 ft/s). A contemporary, Pierre Gassendi, noted that the speed of sound was independent of its intensity, since the speed was the same whether the sound was made by a large cannon or a musket. The next century saw the first precise measurements of the speed of sound. In 1738 measurements under the direction of the Academy of Paris yielded a value of 332 m/s (1089 ft/s) when corrected to 0°C (32°F). At about the same time, G. L. Bianconi showed that the speed of sound definitely increased with increasing air temperature.

Sound travels relative to air with the speed given by Eq. (1), where T is absolute temperature in kelvins

$$c = (\gamma R_0 T/M)^{1/2} \qquad (1)$$

(temperature in degrees Celsius + 273.15), R_0 is the universal gas constant, γ is the specific heat ratio, and M is the average molecular weight. For dry air of normal composition, $M = 29.0$ and Eq. (1) gives $c = 331$ m/s (1086 ft/s) at 0°C (32°F). Thus, the value found in 1738 was within 0.3% of the best modern value.

The sustained development of atmospheric acoustics did not begin until the last half of the nineteenth century. The effects of atmospheric profiles on the refraction of sound received considerable attention. In the early twentieth century, there were significant advances in understanding molecular absorption processes in the atmosphere. Starting in the 1960s, concern over increasing environmental noise led to increased understanding of the effects of the ground on sound propagation outdoors. At about the same time, the invention of the echosonde opened a new area of research on the acoustic sounding of the atmosphere.

Atmospheric sound attenuation. As sound propagates in the atmosphere, several interacting mechanism attenuate and change the spectral or temporal characteristics of the sound received at a distance from the source. The attenuation means that sound propagating through the atmosphere decreases in level with increasing distance between source and receiver. The total attenuation, in decibels, can be approximated as the sum of three nominally independent terms, as given in Eq. (2), where A_{div} is the

$$A_{\text{total}} = A_{\text{div}} + A_{\text{air}} + A_{\text{env}} \qquad (2)$$

attenuation due to geometrical divergence, A_{air} is the attenuation due to air absorption, and A_{env} is the attenuation due to all other effects and includes the effects of the ground, refraction by a nonhomogeneous atmosphere, and scattering effects due to turbulence.

Sound energy spreads out as it propagates away from its source due to geometrical divergence. At distances that are large compared with the effective size of the sound source, the sound level decreases at the rate of 6 dB for every doubling of distance. The phenomenon of geometrical divergence, and the corresponding decrease in sound level with increasing distance from the source, is the same for all acoustic frequencies. In contrast, the attenuation due to the other two terms in Eq. (2) depends on frequency and therefore changes the spectral characteristics of the sound.

Air absorption. Dissipation of acoustic energy in the atmosphere is caused by viscosity, thermal conduction, and molecular relaxation. The last arises because fluctuations in apparent molecular vibrational temperatures lag in phase the fluctuations in translational temperatures. The vibrational temperatures of significance are those characterizing the relative populations of oxygen (O_2) and nitrogen (N_2) molecules. Since collisions with water molecules are much more likely to induce vibrational state changes than are collisions with other oxygen and nitrogen molecules, the sound attenuation varies markedly with absolute humidity. *See* CONDUCTION (HEAT); MOLECULAR STRUCTURE AND SPECTRA; VISCOSITY.

The total attenuation due to air absorption increases rapidly with frequency (**Fig. 1**). For example, at a pressure of 1 atm, temperature of 20°C (68°F), and relative humidity of 70%, the total attenuation is about 1 dB for every 100 m (328 ft) at 2 kHz, but is close to 100 dB for every 100 m at 20 kHz. For this reason, applications in atmospheric acoustics are restricted to sound frequencies below a few thousand hertz if the propagation distance exceeds a few hundred meters. *See* SOUND ABSORPTION.

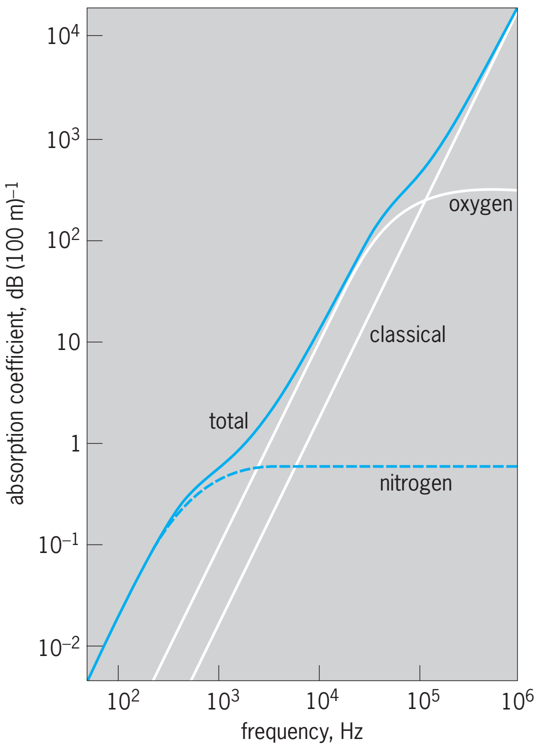


Fig. 1. Frequency dependence of attenuation coefficient in units of decibels per 100 m (dB/328 ft), temperature of 20°C (68°F), and a relative humidity of 70%. Total attenuation from air absorption and contributions from classical absorption (that due to viscosity and thermal conduction) and molecular absorption of oxygen and nitrogen are shown. (*After J. E. Piercy, T. F. W. Embleton, and L. C. Sutherland, Review of noise propagation in the atmosphere, J. Acous. Soc. Amer., 61:1403–1418, 1977*)

Effects of the ground. When the sound source and receiver are above a large flat ground surface in a homogeneous atmosphere, sound reaches the receiver via two paths (**Fig. 2**). There is the direct path from source to receiver and the path reflected from the ground surface. Most naturally occurring ground surfaces are porous to some degree, and their acoustical property can be represented by an acoustic impedance. The acoustic impedance of the ground is in turn associated with a reflection

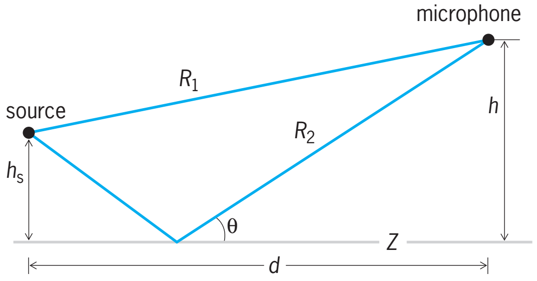


Fig. 2. Diagram of reflection of sound from a flat ground surface with impedance *Z*, showing direct (R_1) and reflected (R_2) paths from source to receiver, source height h_s, microphone height *h*, distance of propagation *d*, and angle of incidence or grazing angle θ.

coefficient that is typically less than unity. In simple terms, the sound field reflected from the ground surface suffers a reduction in amplitude and a phase change.

If the incident waves are plane, the reflection coefficient R_p on a plane ground surface is given in its simplest from by Eq. (3), where θ is the angle of

$$R_p = \frac{\sin\theta - (\rho c/Z)}{\sin\theta + (\rho c/Z)} \tag{3}$$

incidence or grazing angle and $\rho c/Z$ is the ratio of the characteristic impedance of the air (ρc, where ρ is the air density) to the acoustic impedance (Z) of the ground surface. At grazing incidence (θ approaches 0), R_P always approaches −1. Therefore, in the case of plane waves, the reflected sound cancels the direct sound at grazing incidence. *See* ACOUSTIC IMPEDANCE.

Real sound sources outdoors can be considered point sources at distances that are large compared with the effective size of the source. Wavefronts are no longer plane but spread spherically away from the source. A good approximation for the total sound pressure from a point source above a plane ground surface is given by Eq. (4), where p_d is the sound

$$p_t = p_d + R_p p_r + p_g \tag{4}$$

field along the direct path, p_r is the sound field along the reflected path, and p_g is a correction term that allows for the curvature of the incident sound field and, under certain conditions, the possible existence of a surface wave.

When the source and receiver are both relatively near the ground and are a large distance apart, the direct and reflected fields become nearly equal and the grazing angle tends to zero. The direct and reflected sound fields then cancel each other because R_P approaches −1, and any sound reaching the receiver is given by the third term, p_g. The consequences of these phenomena are illustrated by the spectra in **Fig. 3**. The sound pressure levels are calculated for a source height of 1.8 m (6 ft), a microphone height of 1.5 m (5 ft), and distances between 125 m (410 ft) and 4 km (2.4 mi). The levels are corrected for geometrical spreading and molecular absorption, and therefore show the effects of the ground only. The large dips around 500 Hz are not true interference dips due to comb filtering from path-length differences, but are the result of the third term in Eq. (3). The position of the dip does not change with source-receiver separation. Therefore, spatial averaging does not remove the decreased sound pressure levels at these frequencies. This has important consequences for environmental noise since 500 Hz is a common dominant frequency of many noise sources (ground transportation noise, for example). Naturally occurring ground provides a significant amount of natural shielding in the case of propagation at grazing incidence.

Refraction of sound. Straight ray paths are rarely achieved outdoors. In the atmosphere, both the wind and temperature vary with height above the ground. The velocity of sound relative to the ground is a function of wind velocity and temperature; hence it also varies with height, causing sound waves to propagate along curved paths. There are two distinct cases.

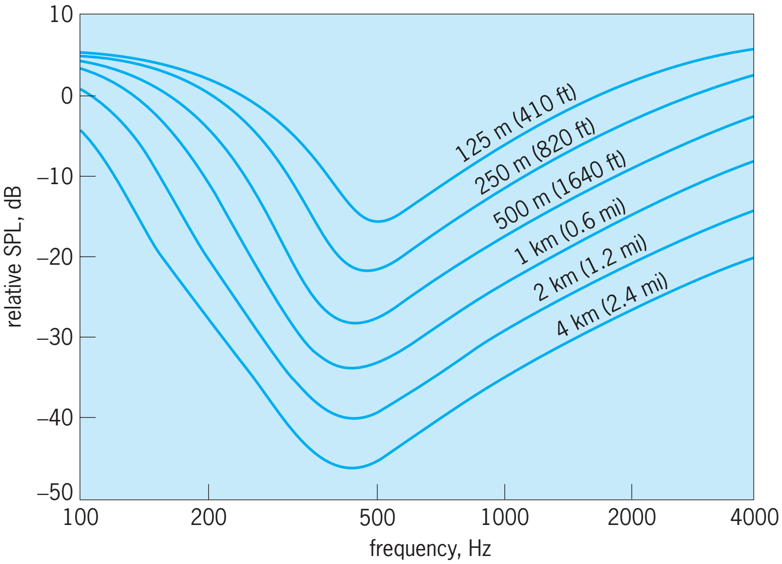


Fig. 3. Relative sound pressure levels (SPL) for propagation from a source over mown grass for source height of 1.8 m (6 ft), microphone height of 1.5 m (5 ft), and distances of propagation indicated.

The speed of the wind decreases with decreasing height above the ground because of drag on the moving air at the surface. Therefore, the speed of sound relative to the ground increases with height during downwind propagation, and ray paths curve downward (**Fig. 4***a*). For propagation upwind, the sound speed decreases with height, and ray paths curve upward (Fig. 4*b*). In the case of upward refraction, a shadow boundary forms near the ground beyond which no direct sound can penetrate. Some acoustic energy penetrates into a shadow zone via creeping waves that propagate along the ground and that continually shed diffracted rays into the shadow zones. The dominant feature of shadow-zone reception is the marked decrease in a sound's higher-frequency content. The presence of shadow zones explains

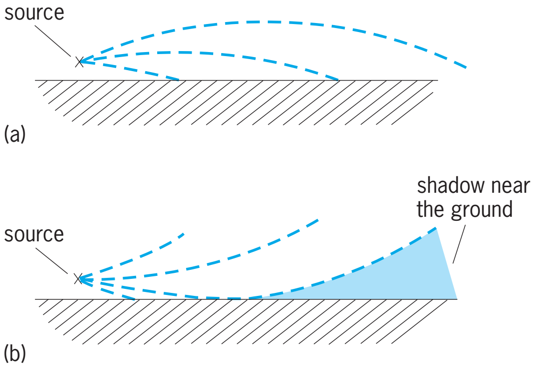


Fig. 4. Curved ray paths. (*a*) Refraction downward, during temperature inversion or downwind propagation.
(*b*) Refraction upward, during temperature lapse or upwind propagation.

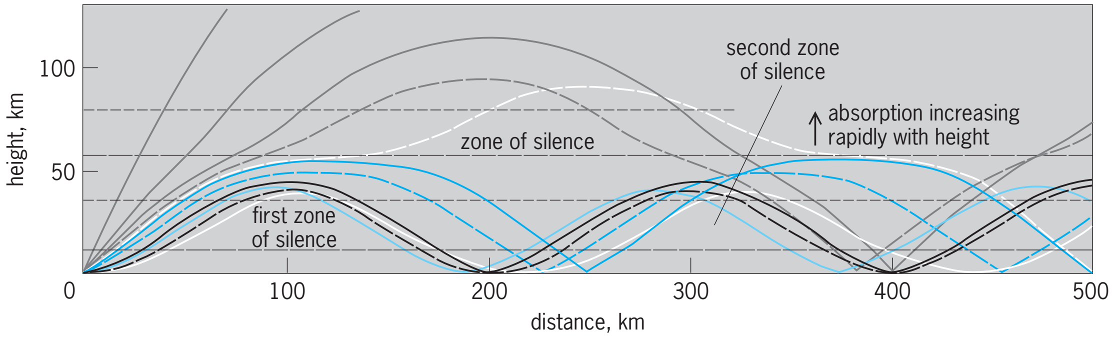


Fig. 5. Representative long-range ray paths of sound radiating from a source in the atmosphere and traveling from west to east in the Northern Hemisphere in winter. 1 km = 0.6 mi. (*After T. F. Malone, ed., Compendium of Meteorology, American Meteorological Society, 1951*)

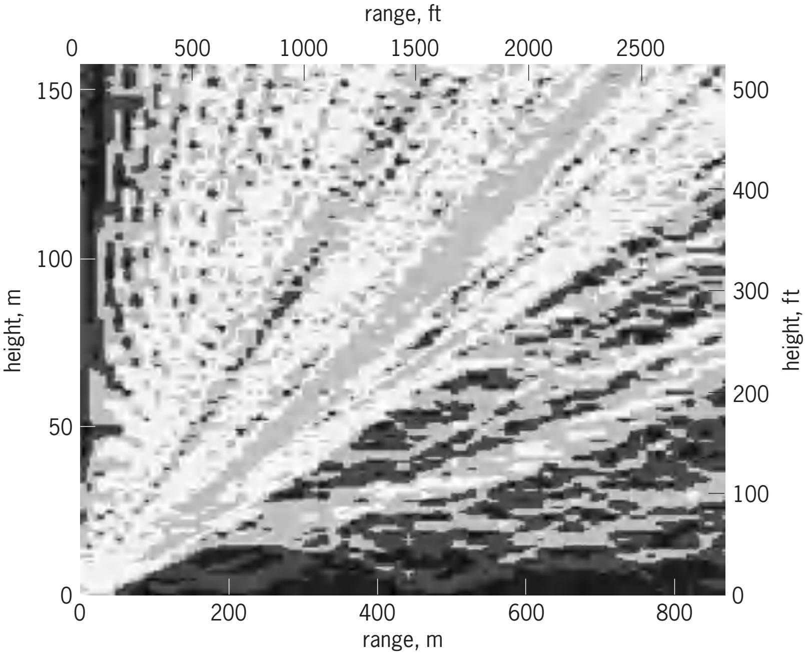


Fig. 6. Contour plot of sound levels in the presence of upward refraction in a turbulent atmosphere. The source is very close to the ground at zero range. The white areas are regions of higher sound levels.

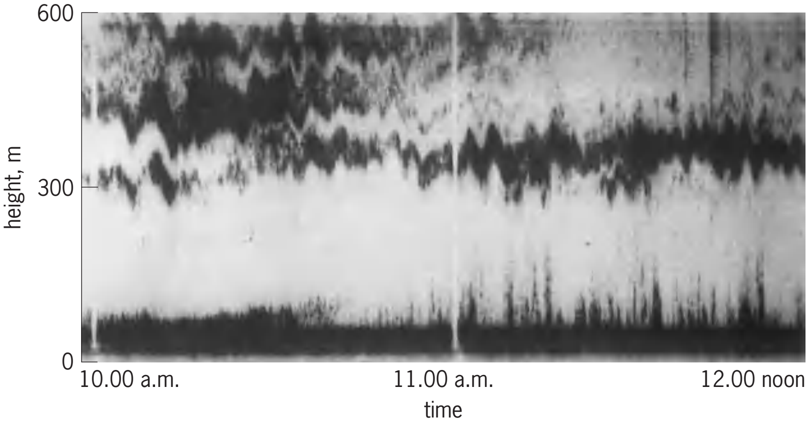


Fig. 7. Facsimile record of acoustic echo sounding of atmosphere at the South Pole, February 9, 1975. 300 m = 984 ft, 600 m = 1969 ft. Dark regions above 300 m altitude correspond to temperature inhomogeneities associated with the passage of convective plumes over the site. (*After E. H. Brown and F. F. Hall, Jr., Advances in atmospheric acoustics, Rev. Geophys. Space Phys., 16:47–110, 1978*)

why sound is generally less audible upwind of a source.

Refraction by temperature profiles is analogous. During the day, solar radiation heats the Earth's surface, resulting in warmer air near the ground. This condition is called a temperature lapse and is most pronounced on sunny days. A temperature lapse is the common daytime condition during most of the year, and also causes ray paths to curve upward. After sunset there is often radiation cooling of the ground, which produces cooler air near the surface. In summer under clear skies, such temperature inversions begin to form about 2 hours after sunset. Within the temperature inversion, the temperature increases with height, and ray paths curve downward.

The effects of refraction by temperature and wind are additive and produce rather complex sound speed profiles in the atmosphere (**Fig. 5**). Rays can be trapped in a sound channel where the effective speed of sound has a minimum. Audible sound is often received at distances of a few hundred kilometers from large explosions, even though the sound may be inaudible at closer distances. *See* REFRACTION OF WAVES.

Effects of turbulence. The atmosphere is also an unsteady medium with random variations in temperature, wind velocity, pressure, and density. In practice, only the temperature and wind velocity variations significantly contribute to the effects of turbulence on a sound wave. Turbulence in the atmosphere causes the effective sound speed to fluctuate from point to point, so a nominally smooth wave front develops ripples. One result is that the direction of a received ray may fluctuate with time in random manner. Consequently, the amplitude and phase of the sound at a distant point will fluctuate with time. The acoustical fluctuations are in some respects analogous to more familiar optical phenomena such as the twinkling of light from a star. The acoustical fluctuations are clearly audible in the noise from a large aircraft flying overhead. *See* TWINKLING STARS.

Turbulence in the atmosphere also scatters sound from its original direction. A collimated beam will spread in width so that the amplitude along the axis will appear to undergo an additional attenuation. Another consequence of sound scattering by turbulence is the degradation of refractive acoustic shadow zones (**Fig. 6**). *See* TURBULENT FLOW.

Echosonde. Atmospheric acoustics has found extensive application in the study of meteorological disturbances in the lower atmosphere. The echosonde, invented by L.G. McAllister in the late 1960s, is the forerunner and also the prototype of current acoustic sounding instrumentation. A transducer pointing upward transmits a short acoustic pulse and then receives weak echoes over a relatively long period of time. Echoes received at time Δt after transmission have been reflected from inhomogeneities at height $c\Delta t/2$.

A facsimile display (**Fig. 7**) is produced by a side-by-side superposition of the echo histories for a long succession of such sounding experiments; this is interpreted as a picture of the time evolution of the height profile of the atmosphere's ability to backscatter acoustic waves. Darker regions on the display imply that, at the corresponding times and heights, the backscattering was strong. Meteorological interpretations of such records requires experience and an understanding of fundamental atmospheric processes. Refined systems in conjunction with the theory of sound propagation through a turbulent atmosphere enable a quantitative determination of temperature and wind-velocity structures above the ground and hence a characterization of the turbulence. *See* SOUND. Gilles A. Daigle

Bibliography. M. J. Crocker (ed.), *Encyclopedia of Acoustics*, Wiley-Interscience, 1997; H. H. Hubbard (ed.), *Aeroacoustics of Flight Vehicles: Theory and Practice*, Acoustical Society of America, 1994; A. D. Pierce, *Acoustics: Introduction to Its Physical Principles and Applications*, 1981.

Atmospheric chemistry

A scientific discipline concerned with the chemical composition of the Earth's atmosphere. Topics include the emission, transport, and deposition of atmospheric chemical species; the rates and mechanisms of chemical reactions taking place in the atmosphere; and the effects of atmospheric species on human health, the biosphere, and climate.

Atmosphere. Over 99.9% of the total mass of the atmosphere is contained in the lowest 50 km (30 mi) above the Earth's surface. This region is divided into the troposphere and the stratosphere. The troposphere extends from the surface to 8–18 km (5–11 mi) altitude, depending on latitude and season, and is characterized by a rapid decrease of temperature with altitude due to solar heating of the surface; it contains 85% of total atmospheric mass. The stratosphere extends from the top of the troposphere (the tropopause) to 50 km (30 mi; the stratopause), and is characterized by an increase of temperature with altitude due to absorption of solar ultraviolet radiation by ozone (O_3). Buoyancy facilitates vertical motions in the troposphere but suppresses these motions in the stratosphere. As a result, vertical mixing of the troposphere takes only about 1 month, while vertical mixing of the stratosphere takes about 5 years. *See* STRATOSPHERE; TROPOSPHERE.

Global transport in the atmosphere is driven primarily by the thermal contrast between high and low latitudes and by the effect of the Coriolis force. East-west transport of species around the globe requires only a few weeks; north-south transport is slower. Exchange of air between the Northern and Southern hemispheres takes about 1 year. *See* ATMOSPHERIC GENERAL CIRCULATION.

A useful quantity in atmospheric chemistry is the atmospheric lifetime, defined as the mean time that a molecule resides in the atmosphere before it is removed by chemical reaction or deposition. The atmospheric lifetime measures the time scale on which changes in the production or loss rates of a species may be expected to translate into changes in the species concentration. The atmospheric lifetime can also be compared to the time scales for transport to infer the spatial variability of a species in the atmosphere; species with lifetimes longer than a decade tend to be uniformly mixed, while species with shorter lifetimes may have significant gradients reflecting the distributions of their sources and sinks.

Principal constituents. The principal constituents of dry air are nitrogen (N_2; 78% by volume), oxygen (O_2; 21%), and argon (Ar; 1%). The atmospheric concentrations of N_2 and Ar are largely determined by the total amounts of N and Ar released from the Earth's interior since the origin of the Earth. The atmospheric concentration of O_2 is regulated by a slow atmosphere-lithosphere cycle involving principally the conversion of O_2 to carbon dioxide (CO_2) by oxidation of organic carbon in sedimentary rocks (weathering), and the photosynthetic conversion of CO_2 to O_2 by marine organisms which precipitate to the bottom of the ocean to form new sediment. This cycle leads to an atmospheric lifetime for O_2 of about 4 million years. It is thought that O_2 has remained at near-present levels in the atmosphere for the past 100 million years, reflecting a balance between sediment formation and weathering. An additional fast cycle between O_2 and CO_2 takes place by oxidation of organic carbon in the biosphere, balanced by photosynthesis; but this cycle does not regulate atmospheric O_2 levels, because it is limited by the supply of organic carbon. The total organic carbon in the biosphere, soil, and ocean reservoirs could convert at most 1% of atmospheric O_2 to CO_2. *See* ATMOSPHERE, EVOLUTION OF; BIOSPHERE; LITHOSPHERE; PHOTOSYNTHESIS.

Water vapor concentrations in the atmosphere range from 3% by volume in wet tropical areas to a few parts per million by volume (ppmv) in the stratosphere. Water vapor, with a mean atmospheric lifetime of 10 days, is supplied to the troposphere by evaporation from the Earth's surface, and it is removed by precipitation. Because of this short lifetime, water vapor concentrations decrease rapidly with altitude, and little water vapor enters the stratosphere. Oxidation of methane represents a major source of water vapor in the stratosphere, comparable to the source contributed by transport from the troposphere.

Carbon dioxide. The most abundant carbon species in the atmosphere is CO_2. It is produced by oxidation of organic carbon in the biosphere and in sediments. The atmospheric concentration of CO_2 has risen from 280 ppmv in preindustrial times (as determined from ice core measurements) to 355 ppmv in 1993. There is concern that this rise may cause significant warming of the Earth's surface because of the ability of CO_2 to absorb infrared radiation emitted by the Earth (that is, the greenhouse effect). *See* GREENHOUSE EFFECT.

There is a natural cycle of carbon through geochemical reservoirs (**Fig. 1**). The total amount of carbon present in the atmosphere is small compared to that present in the other geochemical reservoirs, and therefore it is controlled by exchange with these reservoirs. Equilibration of carbon between the atmosphere, biosphere, soil, and surface ocean reservoirs takes place on a time scale of decades. The carbon is eventually transported to the deep ocean, mainly by sinking of cold surface water at high latitudes. Carbon then resides in the deep ocean for approximately 700 years before upwelling to the surface. A small fraction of the carbon in the deep ocean is incorporated into sediments, where it has a lifetime of 400 million years against uplift to the surface and weathering.

Human activity over the past 100 years has disrupted the natural carbon cycle by increasing the fluxes of carbon from the sediments to the atmosphere (fossil fuel combustion) and from the biosphere to the atmosphere (land-use changes, in particular tropical deforestation). The rate of increase of atmospheric CO_2 during the 1980s was 1.8 ppmv/year, that is, 4.0×10^{12} kg (8.8×10^{12} lb) C/year. Estimated sources from fossil fuel combustion and tropical deforestation during the decade were 6.0×10^{12} kg (13×10^{12} lb) C/year and 1.6×10^{12} kg (3.5×10^{12} lb) C/year, respectively, and the estimated sink of CO_2 from net uptake by the ocean was 2.0×10^{12} kg (4.4×10^{12} lb) C/year. Mass balance requires an additional CO_2 sink of 1.6×10^{12} kg (3.5×10^{12} lb) C/year, which could possibly be contributed by accumulation of carbon in the biosphere and in soils at northern midlatitudes. The capacity of the ocean, biosphere, and soil reservoirs to take up additional carbon is a crucial issue for assessing the response of atmospheric CO_2 to changes in emissions. *See* CARBON DIOXIDE.

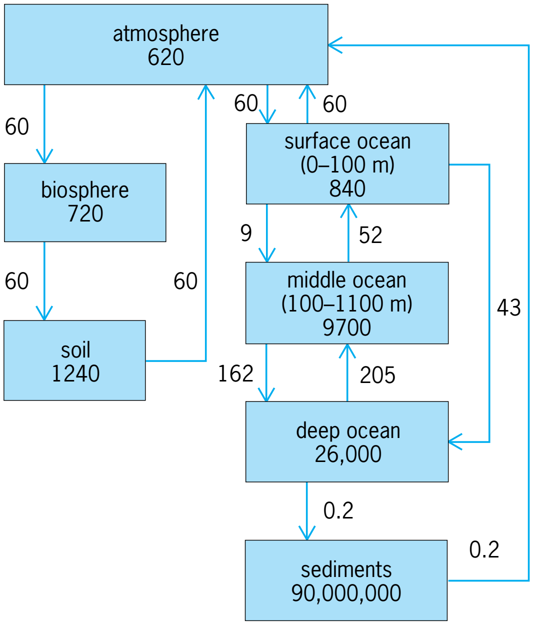


Fig. 1. Natural carbon (C) cycle. The total mass of carbon in each geochemical reservoir is in units of 10^{12} kg C; the carbon fluxes between reservoirs are in units of 10^{12} kg C/year. 1 kg = 2.2 lb. 1 m = 3.3 ft.

Other carbon species. Methane is the second most abundant carbon species in the atmosphere and an important greenhouse gas. It is emitted by anaerobic decay of biological carbon (for example, in wetlands, landfills, and stomachs of ruminants), by exploitation of natural gas and coal, and by combustion. It has a mean lifetime of 12 years against atmospheric oxidation by the hydroxyl (OH) radical, its principal sink. The concentration of methane in the atmosphere has risen from 0.8 ppmv in preindustrial times to 1.7 ppmv in 1993; this rise reflects an increase of sources over the past century, and could also reflect in part a decrease of OH concentrations. *See* METHANE.

Many hydrocarbons other than methane are emitted to the atmosphere from vegetation, soils, combustion, and industrial activities. The emission of isoprene [$H_2C{=}C(CH_3){-}CH{=}CH_2$] from deciduous vegetation is particularly significant. Nonmethane hydrocarbons have generally short lifetimes against oxidation by OH (a few hours for isoprene), so that their atmospheric concentrations are low. They are most important in atmospheric chemistry as sinks for OH and as precursors of tropospheric ozone, organic nitrates, and organic aerosols.

Carbon monoxide (CO) is emitted to the atmosphere by combustion, and it is also produced within the atmosphere by oxidation of methane and other hydrocarbons. It is removed from the atmosphere by oxidation by OH, with a mean lifetime of 2 months. Carbon monoxide is the principal sink of OH and hence plays a major role in regulating the oxidizing power of the atmosphere. Typical concentrations of CO in the troposphere range from about 50 parts per billion by volume (ppbv) in clean regions of the Southern Hemisphere to more than 200 ppbv over the polluted continents.

Nitrogen oxides. Nitrous oxide (N_2O) is of environmental importance as a greenhouse gas and as the stratospheric precursor for the radicals NO and NO_2. The principal sources of N_2O to the atmosphere are microbial processes in soils and the oceans; the main sinks are photolysis and oxidation in the stratosphere, resulting in an atmospheric lifetime for N_2O of about 130 years. The atmospheric concentration of N_2O has risen from 290 ppbv to 310 ppbv over the past century, a trend related at least in part to the growing use of nitrogen-based fertilizer.

The radical species NO and NO_2, referred to collectively as NO_x, play a major role in regulating the concentration of ozone both in the stratosphere and in the troposphere. They are produced in the stratosphere by oxidation of N_2O, and are emitted to the troposphere by combustion, lightning, and microbial processes in soils. The lifetime of NO_x against oxidation to nitric acid is about 1 day; in the troposphere, nitric acid is removed principally by deposition, while in the stratosphere it cycles back to NO_x. Reaction of NO_x with organic radicals in hydrocarbon-rich regions of the troposphere produces organic nitrates, in particular peroxyacetylnitrate [$CH_3C(O)OONO_2$], which may be transported over long distances and eventually decompose to provide a diffuse reservoir source of NO_x. Tropospheric concentrations of NO_x range from over 1 ppbv in polluted regions with large combustion sources to 0.01–0.1 ppbv in remote regions.

The deposition of nitric acid is an important source of nitrogen nutrient to terrestrial ecosystems (in marine ecosystems, upwelling from the deep ocean dominates atmospheric deposition as a source of nitrogen). It has been argued that emission of NO_x from fossil fuel combustion could fertilize the biosphere and thus facilitate the removal of CO_2 from the atmosphere. In ecosystems where nitrogen is not a limiting nutrient, however, deposition of nitric acid causes acidification and a general decrease of vegetative productivity. *See* ECOSYSTEM; NITROGEN CYCLE; UPWELLING.

Tropospheric ozone and OH. About 90% of total atmospheric ozone resides in the stratosphere, where it is produced by photolysis of O_2. The ultraviolet photons ($\lambda < 240$ nm) needed to photolyze O_2 are totally absorbed by ozone and O_2 as solar radiation travels through the stratosphere. As a result, ozone concentrations in the troposphere (0.01–0.1 ppmv) are much lower than in the stratosphere (1–10 ppmv). *See* PHOTOLYSIS.

Tropospheric ozone plays a central role in atmospheric chemistry by providing the primary source of the strong oxidant OH. It is also an important greenhouse gas. In surface air, ozone is of great concern because of its toxicity to humans and vegetation. Ozone is supplied to the troposphere by slow transport from the stratosphere, and it is also produced within the troposphere by a chain reaction involving oxidation of CO and hydrocarbons by OH in the presence of NO_x. The chain is initiated by photolysis of ozone to produce singlet oxygen $O(^1D)$, as in reaction (1). The principal fate of $O(^1D)$ is stabilization to the ground-level triplet state $O(^3P)$ by collision with an inert molecule M [either N_2 or O_2; reaction (2)]. The $O(^3P)$ then reacts with O_2 to reform ozone [reaction (3)]. However, a small percentage of the $O(^1D)$ atoms reacts with water vapor, producing OH [reaction (4)].

$$O_3 + h\nu(\lambda < 320\,\text{nm}) \longrightarrow O_2 + O(^1D) \quad (1)$$

$$O(^1D) + M \longrightarrow O(^3P) + M \quad (2)$$

$$O(^3P) + O_2 + M \longrightarrow O_3 + M \quad (3)$$

$$O(^1D) + H_2O \longrightarrow 2OH \quad (4)$$

Once OH is produced, the chain propagates in sequence by reaction of hydrocarbons and CO with OH to yield peroxy radicals; reaction of the peroxy radicals with NO to produce NO_2 and recycle OH; and photolysis of NO_2 to produce O_3 while recycling NO. In the simplest case of the oxidation of CO, the propagation steps are as shown in reactions (5)–(8). The net reaction is given by (9). The chain is terminated by scavenging of radicals, for example reaction (10).

$$CO + OH \longrightarrow CO_2 + H \quad (5)$$

$$H + O_2 + M \longrightarrow HO_2 + M \quad (6)$$

$$HO_2 + NO \longrightarrow OH + NO_2 \quad (7)$$

$$NO_2 + h\nu(\lambda < 400\,\text{nm}) \longrightarrow NO + O(^3P) \quad (8)$$

$$O(^3P) + O_2 + M \longrightarrow O_3 + M \quad (3)$$

$$CO + 2O_2 \xrightarrow{h\nu} CO_2 + O_3 \quad (9)$$

$$HO_2 + HO_2 \longrightarrow H_2O_2 + O_2 \quad (10)$$

See CHAIN REACTION (CHEMISTRY); FREE RADICAL; TRIPLET STATE.

Ozone production by the above mechanism is particularly rapid in urban areas, where emissions of NO_x and of reactive hydrocarbons are high, resulting in unhealthful ozone levels. Regional-scale episodes of high ozone extending spatially over 10^5–10^6 km^2 (10^5 mi^2) are often observed in the United States during summer, and are thought to be primarily caused by interaction of NO_x emitted from fossil fuel combustion with isoprene emitted from vegetation. On a global scale, the above mechanism is estimated to represent a source of tropospheric ozone at least comparable in magnitude to input from the stratosphere, and perhaps much larger.

The strong oxidant OH produced by reactions (4) and (7) is the principal agent for removal of many environmentally important atmospheric species, including CO, methane, methyl bromide, hydrochlorofluorocarbons (HCFCs), NO_x, and sulfur species. Understanding the factors regulating OH concentrations is therefore a crucial issue in atmospheric chemistry. The lifetime of OH against oxidation of CO and hydrocarbons, its principal sinks, is only about 1 second; because of this short lifetime, OH concentrations are governed locally by a chemical steady state that is dependent both on solar ultraviolet radiation and on the concentrations of ozone, water vapor, NO_x, CO, and hydrocarbons. Typical daytime OH concentrations in the troposphere are of order 10^6 radicals cm^{-3}. Concentrations of OH may have declined over the past century because of anthropogenic emissions of CO and hydrocarbons, but this point is a matter of controversy. Any such decline would have been compensated at least in part by anthropogenic emissions of NO_x, resulting in increased ozone and corresponding OH production by reactions (4) and (7). There is concern that future

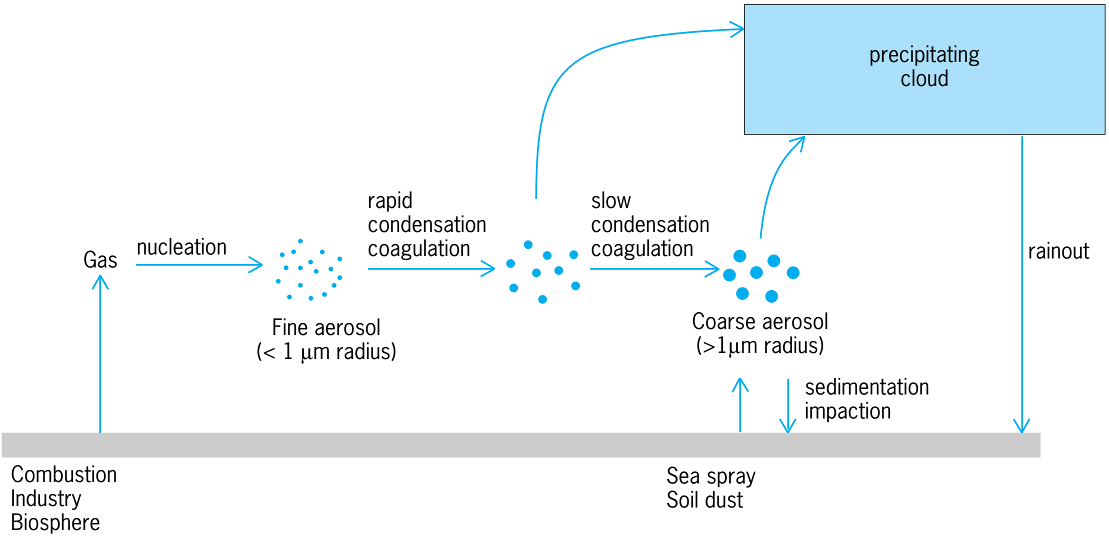

precipitating cloud
Gas
nucleation
Fine aerosol (< 1 μm radius)
rapid condensation coagulation
slow condensation coagulation
Coarse aerosol (>1μm radius)
rainout
sedimentation impaction
Combustion
Industry
Biosphere
Sea spray
Soil dust

Fig. 2. Processes that determine aerosol concentrations.

changes in anthropogenic emissions of CO, hydrocarbons, and NO_x may have the potential for causing large changes in the concentration of OH and hence in the oxidizing power of the atmosphere. *See* ULTRAVIOLET RADIATION.

Atmospheric aerosols. The term aerosol refers to an ensemble of liquid and solid particles suspended in air. Aerosol number concentrations in the troposphere range from 10 to 10^6 particles cm^{-3}, with most of the aerosol mass present in particles with a radius of 0.1–10 micrometer. Aerosols are environmentally important in a number of ways. They cool the Earth by scattering solar radiation, limit visibility, serve as condensation nuclei for cloud formation, and provide media for certain chemical reactions that would otherwise not occur in the atmosphere. At the high concentrations sometimes present in urban air, aerosols may represent a significant public health risk. *See* ALBEDO.

There is a group of processes that determine aerosol concentrations (**Fig. 2**). Aerosols emitted mechanically to the atmosphere (for example, sea spray and soil dust) are called primary; they are relatively large ($>1\ \mu m$ radius) and deposit within a period of a few days by sedimentation or impaction on surfaces. Aerosols produced by condensation of gases (for example, sulfate and soot) are called secondary; they grow rapidly by condensation and coagulation to 0.1–1 μm radius, but further growth is slow, in part because the coagulation rate is limited by particle inertia. Secondary aerosols are removed from the atmosphere principally by rainout, with a mean lifetime in the troposphere of about 10 days. Aerosol lifetimes in the stratosphere are long, 1–2 years, because of the lack of precipitation. Volcanic aerosols injected into the stratosphere can thus effectively scatter solar radiation and have a sustained cooling effect on the Earth's climate.

Sulfuric acid produced in the atmosphere by oxidation of sulfur dioxide (SO_2) is a major component of the secondary aerosol and an important contributor to acid deposition. Sources of SO_2 to the atmosphere include emission from combustion, smelters, and volcanoes, and oxidation of oceanic dimethylsulfide [$(CH_3)_2S$] emitted by phytoplankton. It is estimated that about 75% of total sulfur emission to the atmosphere is anthropogenic. The increase of planetary albedo due to anthropogenic sulfate aerosol over the last century could possibly have caused a significant cooling of the Earth, offsetting the concurrent warming caused by rising concentrations of greenhouse gases. *See* AEROSOL; AIR POLLUTION.

Models. The general approach for computing the concentration *c* of a species in an atmospheric chemistry model is to solve the mass conservation equation for that species, including terms from transport, emission (*E*), deposition (*D*), chemical production (*P*), and chemical loss (*L*). The coordinate system of the model may be fixed with respect to the Earth (eulerian framework) or tied to moving parcels of air (lagrangian framework). In the eulerian framework, the mass conservation equation is written, for example, as Eq. (11), where *N* is the air density (both *c* and *N* are in units of moles per unit volume).

$$\frac{\partial c}{\partial t} = -\nabla \cdot (\mathbf{U}c) + \nabla \cdot \mathbf{K}N\nabla\left(\frac{c}{N}\right) + E - D + P - L \quad (11)$$

The first term on the right-hand-side represents transport by the mean circulation (**U** is the wind vector), and the second term is a simple parametrization for turbulent mixing derived by analogy with molecular diffusion (**K** is a turbulent diffusivity tensor). The *P* and *L* terms are usually functions of the concentrations of other species, so that closure may require additional mass conservation equations (one for each of the chemically interacting species). Models for atmospheric aerosols are further complicated by the need to account for a multiplicity of phases and for interactions between phases (for example, gas-particle mass transfer and aerosol coagulation). *See* ATMOSPHERE; FLUID-FLOW PRINCIPLES. Daniel J. Jacob

Bibliography. B. J. Finlayson-Pitts and J. N. Pitts, Jr., *Chemistry of the Upper and Lower Atmosphere: Theory, Experiments, and Applications*, 1999; T. E. Graedel and P. J. Crutzen, *Atmospheric Change: An Earth's System Perspective*, 1993; J. Houghton, *Global Warming: The Complete Briefing*, 3d ed., 2004; D. Jacob, *Introduction to Atmospheric Chemistry*, 1999; J. H. Seinfeld and N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 1997; P. Warneck, *Chemistry of the Natural Atmosphere*, 2d ed., 1999.

Atmospheric electricity

Electrical processes that take place in the lower atmosphere. This activity is ubiquitous and is of two kinds: the intense local electrification accompanying storms, and the much weaker fair-weather electrical activity occurring over the entire globe that is produced by the thousand or so electrified storms continuously in progress. The relative importance of the various mechanisms that cause storms to accumulate electrically charged particles is unknown, and the extent to which atmospheric electricity affects meteorological processes has yet to be determined. It is possible that atmospheric electricity may be important in forming precipitation and in providing energy to severe tornadoes. *See* PRECIPITATION (METEOROLOGY); TORNADO.

Disturbed-weather phenomena. The usual height of thunderstorms is about 6 mi (10 km); however, they can be as low as 2.5 mi (4 km) or as high as 13 mi (20 km). These electrified storms are characterized by heavy precipitation and strong updrafts and downdrafts with speeds as high as 65 mi/h (30 m/s).

As the result of various electrification processes, regions of charged water and ice particles accumulate in the cloud associated with the thunderstorm. In over 90% of such clouds, positive charge is in the upper part and negative charge is in the lower part, thus forming a positive, vertical dipole. When the electrified regions have dimensions of miles or kilometers and contain charges of tens of coulombs, electric fields of several hundred kilovolts per meter are produced, sufficient in intensity to cause dielectric breakdown in the form of lightning. A majority of these flashes, known as intracloud discharges, exchange charge between the upper positively charged and the lower negatively charged regions within the cloud. A minority, cloud-to-ground discharges, carry charge from the lower charged region, which is usually negative, to the ground. An average thunderstorm generates a charging current of roughly an ampere and potential differences of a hundred million volts that cause lightning flashes to occur about every 10 s, each transporting about 10 coulombs of charge.

Electrical conductivity. The atmosphere is an electrical conductor because it contains ions. These small electrically charged clusters of molecules are produced by radioactivity, solar radiation, and cosmic rays. Ions transport charge because they move under the influence of an electric field. The upper level of the atmosphere, above about 24 mi (40 km), known as the ionosphere or electrosphere, is a good conductor because of the presence of electrons and ions produced by solar and extraterrestrial ionizing radiation. In this region the air is so conductive that the time required for a charge to leak away, the electrical relaxation time, may be as little as 10^{-7} s. At lower levels, where the atmosphere is more dense and there is much less ionizing radiation, it is a much poorer conductor. At sea level, where ions are produced at the low rate of about 10 ion pairs $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$ by cosmic rays and terrestrial radioactivity, the atmosphere has very low conductivity. Here the electrical relaxation time can be 10^3 s or longer. If suspended cloud or aerosol particles are present, the conductivity can be further reduced, because many of the ions lose their mobility by becoming attached to these much more massive particles. In clouds, fog, or smoke the electrical relaxation time can sometimes exceed 10^4 s.

The air over thunderstorms is sufficiently conductive for a current to flow from the positive cloud top, carrying positive charge to the ionosphere. Measurements over the tops of thunderclouds show conduction currents with values ranging from one-tenth to several amperes. The frequency with which a cloud gives lightning is found to be roughly proportional to its conduction current, suggesting a relationship between this current and the electrification process that causes lightning.

The cloud is neither a sink nor a source of charge. Therefore, the current of positive charge flowing from the top of the cloud to the atmosphere is balanced by an equal transport of charge from the earth to the lower part of the cloud. This takes place through various processes, such as cloud-to-ground lightning, point discharge, falling precipitation, and conduction. Although the details are not clear, the transfer of charge within the cloud occurs as the result of intracloud lightning, the downward motion of negatively charged cloud and precipitation particles, and the upward motion of positively charged cloud particles.

The combined effect of the approximately one thousand thunderstorms in progress at any given time in the Earth's atmosphere provides an electric current of about a kiloampere that continuously deposits negative charge on the surface of the Earth and conveys an equal positive charge into the atmosphere by conduction through the ionosphere. *See* ATMOSPHERE; IONOSPHERE; THUNDERSTORM.

Fair-weather field. The currents from thunderstorms maintain a negative charge of about 1 million coulombs on the Earth and an equal net positive charge in the atmosphere, raising the ionosphere to a negative potential of about 200,000–300,000 V with respect to the Earth. Under the influence of this potential difference a vertical conduction current of about 1 microampere/km^2 carries positive charge to the Earth, giving rise to a fair-weather electric field which, because it is inversely proportional to atmospheric conductivity, decreases exponentially with height. At sea level the field intensity is about 100 V/m and at 10 km 10 V/m, indicating 90% of the positive space charge in the atmosphere lies below 10 km (6 mi). The fair-weather field at the Earth's surface shows fluctuations over minutes or hours as a result of space charge carried by air motions and local variations in atmospheric conductivity. However, in undisturbed locations far at sea or over the polar regions, the field is less variable and a diurnal variation of about 20% can be recognized that is independent of position or local time. The variation is at a minimum a few hours after midnight and at a maximum in the late afternoon Greenwich Mean Time (GMT). This corresponds to the similar diurnal variation of thunderstorm electrical activity over the globe, which is at a minimum when the Sun is shining on the Pacific Ocean, and at a maximum when it is shining on the Earth's large landmasses.

Electricity as global index. Measuring lightning activity over large areas is accomplished by means of sferics detection, satellite observations, lightning detection networks, and evaluation of Schumann resonances. (Electromagnetic waves with frequencies of several to tens of hertz circle the Earth in the waveguide formed by the layer of air between the Earth and the ionosphere. Those waves with resonances of 8, 14, and 20 Hz are known as Schumann resonances.) The magnitude of the conduction currents flowing between thunderclouds and the upper atmosphere is closely related to the frequency of lightning. Therefore, measurements of the potential difference between the Earth and the ionosphere yield information concerning the rate that lightning is taking place over the entire globe. Because the frequency of lightning occurrence is a measure of convective activity, which increases rapidly with temperature, atmospheric electrical measurements are being used to detect and monitor global warming. *See* LIGHTNING; SFERICS.

Electrification mechanisms. Two processes give rise to the electrified cloud and precipitation particles that cause lightning. Such charged particles are produced within the cloud by charge-separation processes that take place when particles collide or break up. They are also introduced into the cloud when charge is brought to the cloud from its environment by conduction currents, convection, or lightning. To form the large regions of charged particles and the extensive electric fields that cause lightning, charged particles must be moved over large distances in opposition to electrical forces. This movement requires approximately 100 MW of power, which is supplied by falling, charged, precipitation particles and by convection that carries the charged particles in strong updrafts and downdrafts. It will not be possible to have a satisfactory understanding of the relative importance of these two charge transport mechanisms until far more is known about air motions within the cloud and details of the nature and the location of the various electrified cloud and precipitation particles. *See* CLOUD PHYSICS.

Effects of thunderstorm electricity. Electrical activity represents no more than a few percent of the total energy of a storm. It is probably for this reason that most meteorologists have assumed that atmospheric electricity plays a negligible role in weather phenomena. There is evidence, however, to suggest that electricity may be more important than formerly supposed. For example, a close correlation has been observed between the formation and dissipation of a very intense tornado and the lightning activity in the parent storm. Laboratory investigations of the effect of electric fields on the coalescence of water drops show that the electrical state of the cloud can influence the subsequent production of rain. Evidence that electricity may be important in natural clouds is afforded by radar observations that show heavy rain sometimes forms after less than a minute in the region of the cloud in which lightning occurs. Because the formation and fall of precipitation is known to produce a large effect on the dynamics of clouds and storm systems, thunderstorm electricity may, in an indirect way, produce significant effects on meteorological processes.

If the development of electric fields can affect the behavior of a cloud, it follows that variables that influence electrification will also influence the weather. Examples of such variables are extraterrestrial ionizing radiation, radioactivity, and aerosols from volcanoes or other natural and anthropogenic sources. Human activities can affect atmospheric electrical processes in several ways. Tall towers, skyscrapers, and airplanes can trigger lightning that otherwise would not have occurred. Techniques for research have been developed that initiate lightning under a thundercloud at a desired place and time by firing a rocket that trails a grounded wire. *See* WEATHER MODIFICATION.

Possible role of fair-weather electricity. There is evidence that fair-weather electricity, which is caused by thunderstorms, may in turn play a role in causing them to become electrified. Experiments show that a cloud growing from air artificially electrified with negative, instead of the usual natural, positive space charge develops into a thundercloud of inverted polarity. This finding suggests, in accord with various influence theories, that the electrification process in thunderclouds may be initiated, and its polarity determined, by the small fair-weather charges present in the atmosphere. Evidence that fair-weather atmospheric electrical variables may affect thunderstorm electrification was provided by observations of a thundercloud formed by a fire deliberately set for scientific studies (prescribed forest fire): the thundercloud grew from air that contained negative space charge, and it gave flashes that were exclusively of positive polarity.

Anomalies. With the advance of technology, observations have accumulated that reveal facets of atmospheric electricity that have yet to be explained. Pictures taken with sensitive video cameras from the ground, from airplanes, and from the space shuttle show various luminous phenomena that proceed out of the tops of active thunderclouds high into the stratosphere or ionosphere. These rare phenomena, which have been seen but not photographed before, sometimes resemble a lightning channel and at other times an aurora or a glow discharge. Observations from lightning-detection networks show that, although most clouds produce almost exclusively negative flashes to ground, a small percentage of otherwise similar storms produce flashes of the opposite polarity. Space exploration discloses that intense electrical activity is present in the very different atmospheres of other planets. Lightning has now been reported on Venus, Jupiter, and Saturn. *See* PLANETARY PHYSICS; WEATHER. Bernard Vonnegut

Bibliography. J. A. Chalmers, *Atmospheric Electricity,* 1957; Geophysics Study Committee, Commission on Physical Sciences, Mathematics, and Resources, *The Earth's Electrical Environment,* 1986; C. Magono, *Thunderstorms,* 1980; R. Reiter, *Phenomena in Atmospheric and Environmental Electricity,* 1992.

Atmospheric entry

The motion of a body traveling from space through the atmosphere surrounding a planet. Entry bodies can be natural bodies, such as meteors or comets, or vehicles, such as ballistic missiles or the space shuttle orbiter. Entry begins when the body reaches the sensible atmosphere (defined as 400,000 ft or 122 km altitude for Earth).

The primary forces acting on an entry body are aerodynamic lift and drag, gravity, propulsion, and centrifugal acceleration. Of particular concern to the designer of entry vehicles is the control of the trajectory to minimize the effects of heating on the thermal protection system and aerodynamic loading on the vehicle structure. From Newton's law of motion, the forces on the vehicle determine the resulting trajectory as the body traverses the atmosphere. As shown in **Fig. 1**, the aerodynamic drag opposes the velocity vector. The aerodynamic lift is perpendicular to the velocity vector but not necessarily in the plane of the velocity and gravity vectors. The centrifugal force opposes the gravity vector. Any propulsive force is generally used to retard the velocity, prior to entry into the sensible atmosphere. *See* CENTRIFUGAL FORCE.

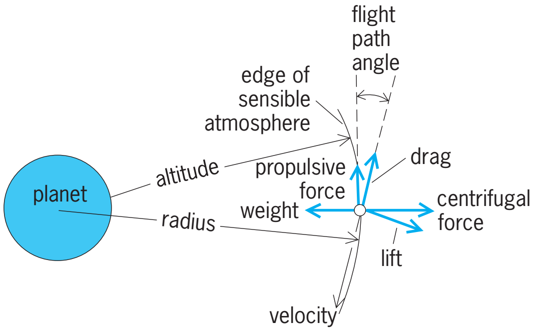


Fig. 1. Entry free-body diagram, showing forces on a vehicle entering the sensible atmosphere.

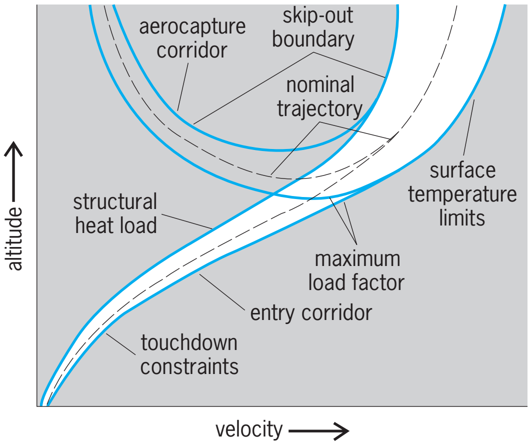


Fig. 2. Guidance corridor. Constraints on the trajectory, which determine the boundaries of the corridor, are indicated.

Aerocapture vehicles use the atmosphere to deplete energy prior to orbit capture in order to significantly reduce the amount of propellant required and thus reduce the mass requirements. Aerocapture is important for planetary exploration and will probably be used for the human exploration of the planets because of the large mass requirements of these missions.

Trajectory control. For controllable vehicles, the concept of trajectory control refers to the management of the kinetic and potential energy so as to maneuver from the initial entry conditions to the desired final conditions with due regard to system constraints and dispersions. One manner of accomplishing this is to establish a guidance corridor (**Fig. 2**). Initially, the flight path is steep enough to prevent skipping out of the atmosphere but shallow enough to keep the maximum temperature and structural load factor within limits. Later in flight, the total heat load into the structure, which increases with time of flight, becomes a constraint. Finally, touchdown or orbit capture conditions dictate orientation and velocity boundaries. The on-board guidance system attempts to follow the nominal trajectory, near the middle of the corridor, with the width of the corridor determined by dispersions. Dispersions are uncertainties, unknowns, errors, or unplanned changes in the guidance factors determining the control of the vehicle. Atmospheric dispersions on Earth are dominated by density deviations of up to 50% at higher altitudes and by surface winds associated with storms and weather fronts at lower altitudes. Other error sources include aerodynamic uncertainties and navigation uncertainties which are of particular concern for planetary aerocapture. *See* SPACE NAVIGATION AND GUIDANCE.

Aerodynamic characteristics. The aerodynamic forces arise from the effect of the airstream on the vehicle. The aerodynamic characteristics are primarily a function of the body shape, air speed, and orientation with respect to the velocity vector. The most dominant aerodynamic force is the vehicle drag, which provides the deceleration. High-drag bodies are characterized by large, blunt reference profiles.

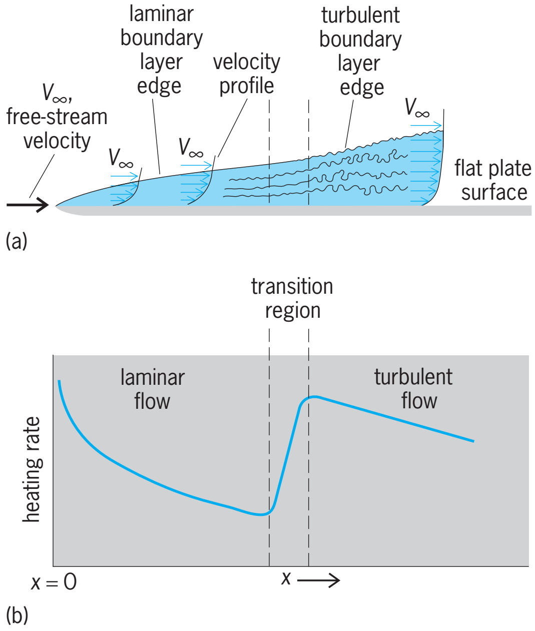


Fig. 3. Laminar and turbulent velocity profiles and heating-rate distribution on a flat plate. (*a*) Boundary layer development on a flat plate showing velocity profiles and streamlines in the boundary layer. (*b*) Surface heating rate as a function of distance *x* from flat-plate leading edge.

The lift force is perpendicular to the drag force, works perpendicular to the velocity vector, and is the primary force vector for trajectory control. Lift is modulated by changing the angle of attack or, more effectively, by rotating the lift vector about the velocity vector.

The ratio of lift to drag (*L*/*D*) determines the amount of trajectory control available. For the space shuttle this ratio is 1.1, while for the Apollo entry capsule this ratio was 0.3. The higher *L*/*D* of the shuttle gives larger ranging capability, more precise trajectory control, and point-landing capability. Depending on the navigation accuracy, aerocapture vehicles are being considered with *L*/*D* values of 0.3 to 1.5. *See* AERODYNAMIC FORCE.

Heating. Vehicles entering the atmosphere experience heat transferred from the hot air surrounding the spacecraft to the colder wall of the spacecraft. The transfer of heat or energy is accomplished by conduction, radiation, and convection. Conduction and radiation depend on a temperature difference between the fluid (air) and the vehicle wall, and convection is associated with the motion of the fluid. The temperature difference provides the potential for energy transfer or heat transfer.

A vehicle traveling at supersonic or hypersonic velocities deflects the air and forms a shock wave. The air between the detached bow shock and the vehicle is heated to very high temperatures by molecular collisions converting the kinetic energy to thermal energy. Approximately 97–98% of this energy flows past the vehicle into the free stream. The remaining 2% has to be managed by the thermal protection system on the spacecraft. The dissipation of heat depends on the vehicle's configuration and orientation to the free stream and the altitude and velocity at which the vehicle is traveling. *See* AEROTHERMODYNAMICS; HYPERSONIC FLIGHT; SHOCK WAVE; SUPERSONIC FLIGHT.

When the air flows along the surface, particles near the surface slow down through the action of viscous forces. At some distance from the surface, the velocity of the particles approaches the velocity of the undisturbed free stream. The region between the surface and some arbitrary percentage of the free-stream velocity (90–99%) is called the boundary layer, which is thin relative to the linear dimension of the vehicle. The flow within this boundary layer can be either laminar, which is smooth and orderly, or turbulent, which is irregular and characterized by eddies or fluctuating motion in the flow (**Fig. 3**). Turbulent flow has faster-moving particles and higher rates of heat transfer to the surface than laminar flow. Thus, it is desirable for entry vehicles to maintain laminar flow as long as possible to minimize the surface temperature. For aerocapture vehicles, flight near the upper corridor boundary helps to maintain laminar flow. *See* BOUNDARY LAYER FLOW; FLUID FLOW.

Heat protection systems. The advent of the space age with its progression from simple ballistic configurations, such as Mercury spacecraft, to the complex high-lift-to-drag orbiter of the Space Shuttle Transportation System has dictated development of new materials for use in the thermal protection system. Ablator materials, which were used on Mercury, Gemini, and Apollo spacecraft, accommodated the convective heating through absorption, vaporization, and the resultant char layer that reradiated the heat to the atmosphere. Ablators, however, are not reusable. *See* NOSE CONE.

The reuse requirement for the shuttle orbiter necessitated development of new concepts of thermal protection. One concept is an external low-density insulator that can withstand high temperatures for multiple orbiter entries for at least 100 flights. The insulator is a tile fabricated from high-purity silica fibers reinforced with silica binder. Two temperature regimes utilize this insulator. One, designated HRSI (high-temperature reusable surface insulator), consists of 6 × 6 in. (15 × 15 cm) tiles applied to the lower surface, where temperatures vary from 1200 to 2300°F (922 to 1533 K). The other, designated LRSI (low-temperature reusable surface insulator), consists of 8 × 8 in. (20 × 20 cm) tiles on the side fuselage and upper wing, where temperatures vary from 700 to 1200°F (644 to 922 K). The materials are the same but have different coatings. These coatings are both waterproof borosilicate glass, but the HRSI coating contains black pigment for high emittance at high temperatures and the LRSI tiles have a white coating for low solar absorptance. *See* HIGH-TEMPERATURE MATERIALS.

Another insulator, FRSI (flexible reusable surface insulation), consists of 3 × 4 ft (0.9 × 1.2 m) sheets of Nomex felt and is bonded to areas on the upper

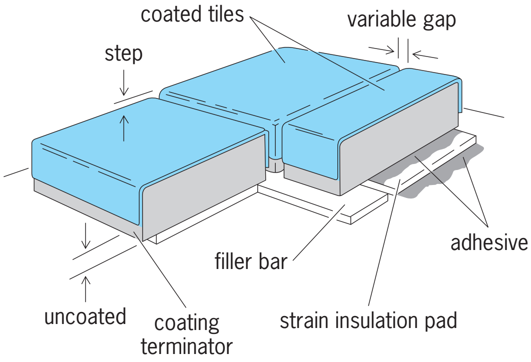


Fig. 4. Thermal-protection-system tile configuration on the shuttle orbiter. (*After R. L. Dotts, D. M. Curry, and D. J. Tillian, The shuttle orbiter thermal protection system materials, design and flight performance overview, 13th Intersociety Conference on Environmental Systems, San Francisco, July 11–15, 1983, SAE Tech. Pap. Ser., 831118, 1983*)

surface where temperatures are less than 700°F (644 K). Nomex felt is also used as the strain isolation pad, between the tiles and the structure. The tile configuration is shown in **Fig. 4**. The filler bar is also coated nomex felt, and the adhesive is room-temperature vulcanizing silicone.

The temperature on the wing leading edges and on the nose of the shuttle orbiter was predicted to exceed 2300°F (1533 K). A reinforced carbon-carbon system was developed to withstand up to 3000°F (1922 K). Reinforced carbon-carbon is a laminate of woven graphite cloth with a carbon binder and a silicon carbide coating which prevents oxidation. The locations of the various types of thermal protection on the surface of the orbiter are shown in **Fig. 5**. *See* COMPOSITE MATERIAL.

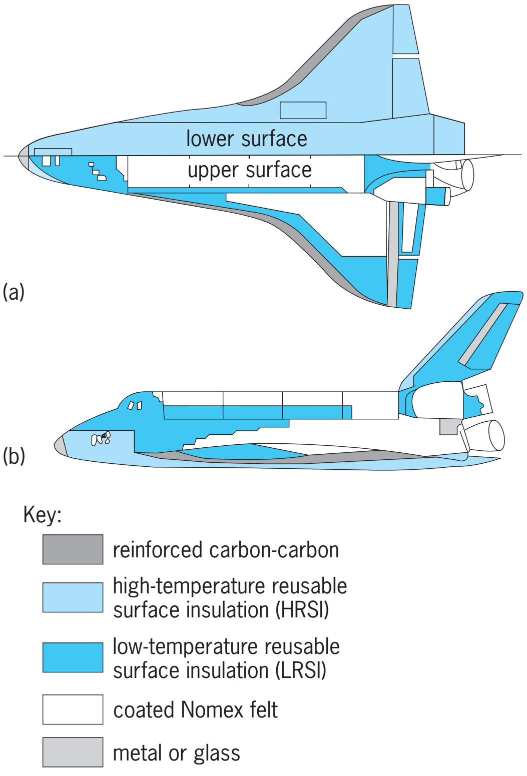

lower surface
upper surface
(a)
(b)
Key:
reinforced carbon-carbon
high-temperature reusable surface insulation (HRSI)
low-temperature reusable surface insulation (LRSI)
coated Nomex felt
metal or glass

Fig. 5. Location of various types of thermal protection on the surface of the shuttle orbiter. (*a*) Top view of upper surface and bottom view of lower surface. (*b*) Side view. (*After R. L. Dotts, D. M. Curry, and D. J. Tillian, The shuttle orbiter thermal protection system materials design and flight performance overview, 13th Intersociety Conference on Environmental Systems, San Francisco, July 11–15, 1983, SAE Tech. Pap. Ser., 831118, 1983*)

These insulating systems are installed as smooth surfaces so that tile-to-tile displacements and gap widths between the tiles are minimized. Stringent requirements of thermal protection system installation are imposed on the manufacturer to assure laminar boundary-layer flow during the significant heating portion of the entry trajectory.

Most of the heat protection systems that have been studied for planetary aerocapture vehicles are similar to the entry systems described above. However, the entry velocity for some of the return missions is too high for the use of tiles, and ablative heat shields will probably be required. *See* BALLISTIC MISSILE; MISSILE; SPACE PROBE; SPACE SHUTTLE; SPACECRAFT STRUCTURE. Richard L. Barton; Dorothy B. Lee; Joe D. Gamble

Bibliography. A. J. Chapman, *Fundamentals of Heat Transfer*, 1987; M. J. Griffin and J. R. French, *Space Vehicle Design*, 2d ed., 2004; W. Hankey, *Re-Entry Aerodynamics*, 1988; F. J. Regan and S. M. Anandakrishnan, *Dynamics of Atmospheric Entry*, 1993.

Atmospheric general circulation

The statistical description of atmospheric motions over the Earth, their role in transporting energy, and the transformations among different forms of energy. Through their influence on the pressure distributions that drive the winds, spatial variations of heating and cooling generate air circulations, but these are continually dissipated by friction. While large day-to-day and seasonal changes occur, the mean circulation during a given season tends to be much the same from year to year. Thus, in the long run and for the global atmosphere as a whole, the generation of motions nearly balances the dissipation. The same is true of the long-term balance between solar radiation absorbed and infrared radiation emitted by the Earth–atmosphere system, as evidenced by its relatively constant temperature. Both air and ocean currents, which are mainly driven by the winds, transport heat. Hence the atmospheric and oceanic general circulations form cooperative systems. *See* MARITIME METEOROLOGY; OCEAN CIRCULATION.

Owing to the more direct incidence of solar radiation in low latitudes and to reflection from clouds, snow, and ice, which are more extensive at high latitudes, the solar radiation absorbed by the Earth–atmosphere system is about three times as great in the equatorial belt as at the poles, on the annual average. Infrared emission is, however, only about 20% greater at low than at high latitudes. Thus in low latitudes (between about 35°N and 35°S) the Earth–atmosphere system is, on the average, heated and in higher latitudes cooled by radiation. The Earth's surface absorbs more radiative heat than it emits, whereas the reverse is true for the atmosphere. Therefore, heat must be transferred generally poleward and upward through processes other than radiation. At the Earth–atmosphere interface, this transfer occurs in the form of turbulent flux of sensible heat and through evapotranspiration (flux of latent heat). In the atmosphere the latent heat is released in connection with condensation of water vapor. *See* CLIMATOLOGY; HEAT BALANCE, TERRESTRIAL ATMOSPHERIC.

Considering the atmosphere alone, the heat gain by condensation and the heat transfer from the Earth's surface exceed the net radiative heat loss in low latitudes. The reverse is true in higher latitudes. The meridional transfer of energy, necessary to balance these heat gains and losses, is accomplished by air currents. These take the form of organized circulations, whose dominant features are notably different in the tropical belt (roughly the half of the Earth between latitudes 30°N and 30°S) and in extratropical latitudes. *See* METEOROLOGY; STORM.

Principal circulations. Characteristic circulations over the Northern Hemisphere are sketched in **Fig. 1**. In the upper troposphere, there are two principal jet-stream systems: the subtropical jet (STJ) near latitude 30°, and the polar-front jet (PFJ), with large-amplitude long waves and superimposed shorter waves associated with cyclone-scale disturbances. The long waves on the polar-front jet move slowly eastward, and the shorter waves move rapidly. At the Earth's surface, northeast and southeast trade winds of the two hemispheres meet at the intertropical convergence zone (ITCZ), in the vicinity of which extensive lines and large clusters of convective clouds are concentrated. Westward-moving waves and vortices form near the intertropical convergence zone and, in summer, within the trades. Heat released by condensation in convective clouds of the intertropical convergence zone, and the mass of air conveyed upward in them, drive meridional circulations (right of Fig. 1), whose upper-level poleward branches generate the subtropical jet stream at their poleward boundaries. *See* TROPICAL METEOROLOGY.

In extratropical latitudes, the circulation is dominated by cyclones and anticyclones. Cyclones develop mainly on the polar front, where the temperature contrast between polar and tropical air masses is concentrated, in association with upper-level waves on the polar-front jet stream. In winter, cold outbreaks of polar air from the east coasts of continents over the warmer oceans result in intense transfer of heat and water vapor into the atmosphere. Outbreaks penetrating the tropics also represent a sporadic exchange in which polar air becomes transformed into tropical air. Tropical airstreams,

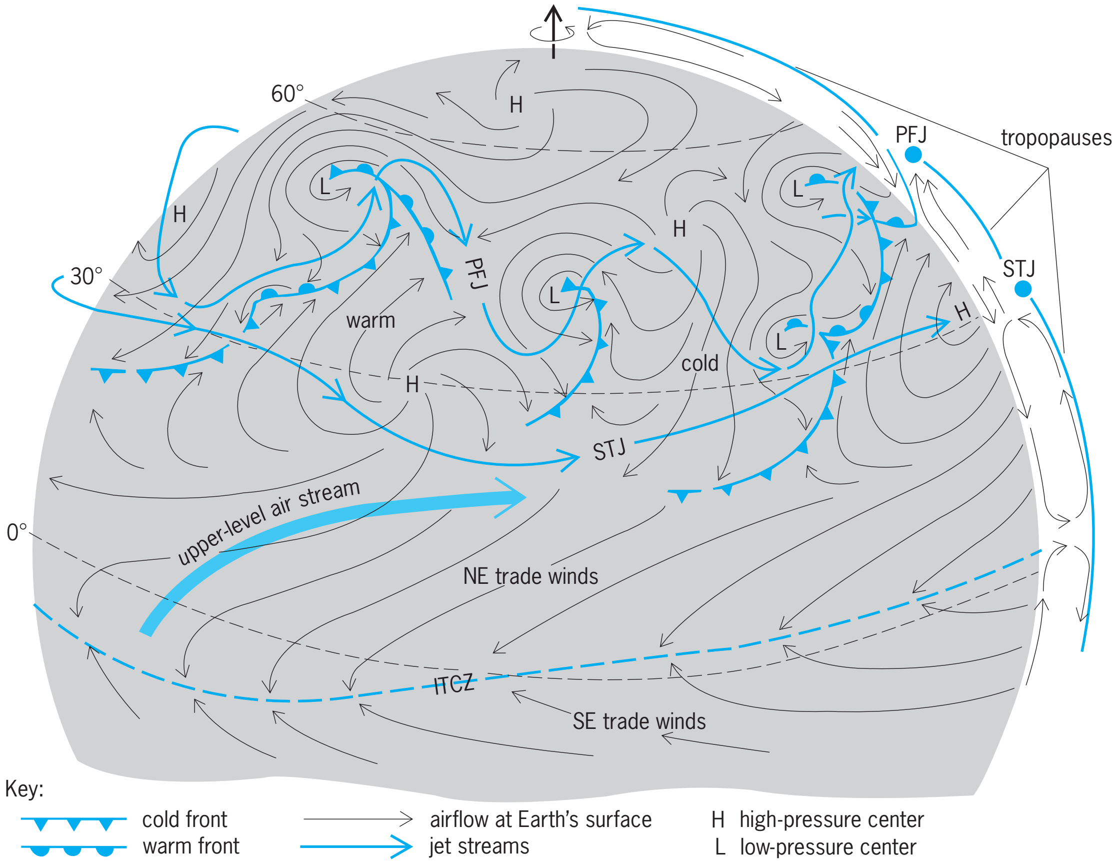


Fig. 1. Schematic circulations over the Northern Hemisphere in winter. The intertropical convergence zone (ITCZ) lies entirely north of the Equator in the summer. Eastward acceleration in the upper-level tropical airstream is due to Earth rotation and generates the subtropical jet stream (STJ). The vertical section (right) shows the dominant meridional circulation in the tropics and shows airstreams relative to the polar front in middle latitudes.

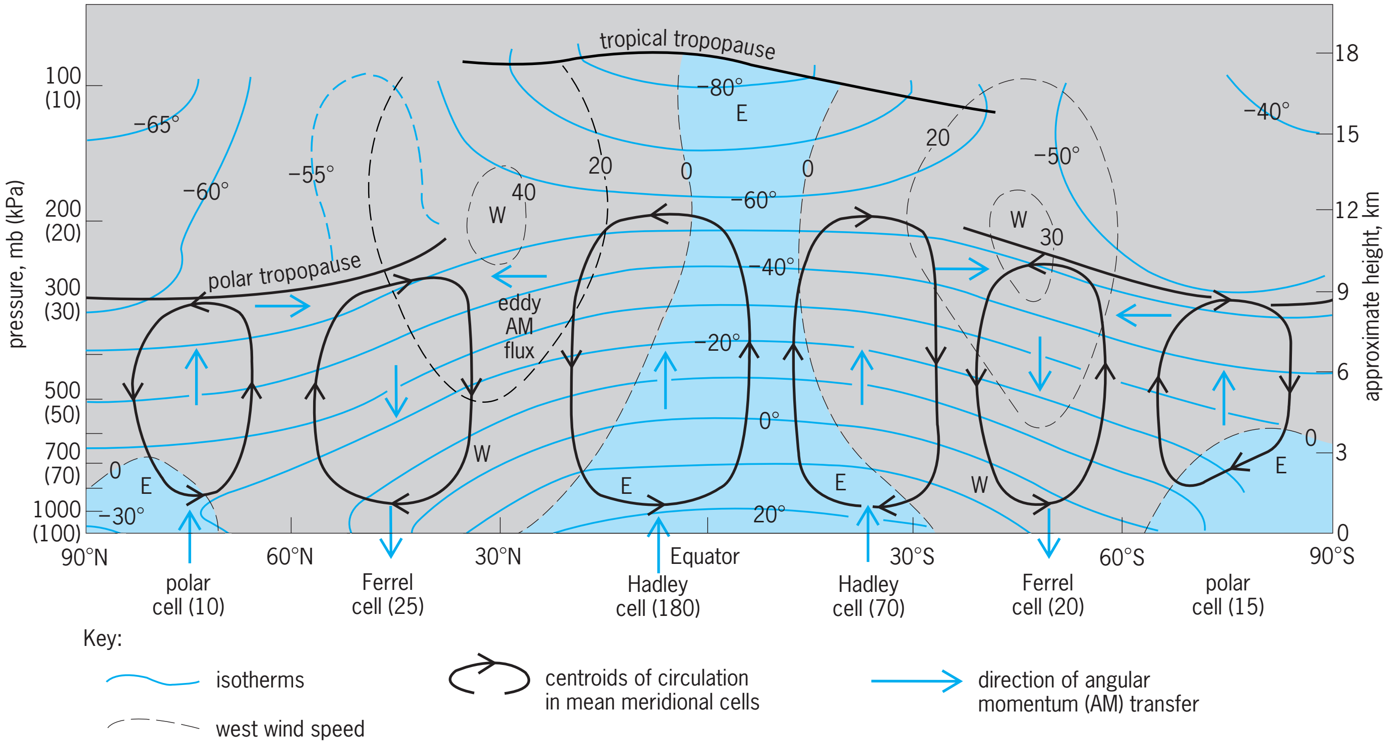

tropical tropopause
polar tropopause
pressure, mb (kPa)
100 (10)
200 (20)
300 (30)
500 (50)
700 (70)
1000 (100)
approximate height, km
18
15
12
9
6
3
0
−65°
−60°
−55°
−80°
E
−60°
−40°
−20°
0°
20°
−30°
−50°
−40°
40
W
20
0
0
20
W
30
eddy AM flux
W
E
E
E
W
0
E
0
E
90°N
60°N
30°N
Equator
30°S
60°S
90°S
polar cell (10)
Ferrel cell (25)
Hadley cell (180)
Hadley cell (70)
Ferrel cell (20)
polar cell (15)
Key:
isotherms
west wind speed
centroids of circulation in mean meridional cells
direction of angular momentum (AM) transfer

Fig. 2. Pole-to-pole section (450-fold vertical exaggeration) in January showing zonally averaged temperature in °C and wind speed in m/s; W and E denote westerlies and easterlies, respectively. Mean meridional cells are named and their intensities are given in terms of mass flow in megatons per second.°F = (°C × 1.8) + 32. 1 m/s = 2 knots. 1 km = 0.6 mi.

poleward on the west sides of the subtropical highs, then supply heat and water vapor to the extratropical disturbances. *See* CYCLONE; FRONT.

The characteristic flow in cyclones takes the form of slantwise descending motions on their west sides and ascent to their east in which extensive clouds and precipitation form. Heat that is released in condensation drives the ascending branch, and the descending branch consists of polar air that has been cooled by radiation in higher latitudes. When viewed relative to the meandering polar-front zone (right of Fig. 1), the combined sinking of cold air and ascent of warm air represents a conversion of potential energy into kinetic energy. This process maintains the polar jet stream. The branches of the circulation transfer heat both upward, to balance the radiative heat loss by the atmosphere, and poleward, to balance the radiative heat deficit in high latitudes.

Mean temperature and wind. A principal object of general circulation studies has been to explain the zonal mean (long-term average around latitude zones) structure of the atmosphere, shown for January in **Fig. 2**. The zonal (west-east) wind component is almost everywhere dominant and is in quasi-geostrophic balance with the mean meridional pressure gradient. The pressure gradient changes with height in accordance with the distribution of air density, which at a given pressure is inverse to temperature. Hence the distribution of zonal wind is related to that of temperature, as expressed by Eq. (1), where

$$\frac{\partial \bar{u}}{\partial z} = -\frac{g}{(2\Omega \sin\phi)\bar{T}}\frac{\partial \bar{T}}{\partial y} \qquad (1)$$

u denotes zonal wind component; z, height above sea level; g, acceleration of gravity; Ω, angular velocity of the Earth; ϕ, latitude; T, Kelvin temperature; and y, distance northward. Overbars denote values averaged over longitude and time. *See* GEOSTROPHIC WIND.

Only in the lowest kilometer or so, where surface friction disturbs the geostrophic balance, and in the vicinity of the Equator is the mean meridional (south-north) component comparable to the zonal wind. Because of the nature of the atmosphere as a shallow layer, the mean vertical wind component is weak. Whereas the magnitude of the mean zonal wind varies 100 mi/h (45 m/s), and the mean meridional wind up to 6.5 mi/h (3 m/s), the mean vertical wind nowhere exceeds 0.02 mi/h (1 cm/s). The vertical component cannot be observed directly, but can be calculated from the distribution of horizontal motions.

In the troposphere and lower stratosphere, the zonal circulation is similar in winter and summer, with easterlies in low latitudes and westerlies in higher latitudes, except in small regions of low-level easterlies around the poles. The strongest west winds are, in the winter hemispheres, observed near latitude 30° at about 7 mi (12 km). In summer, the west-wind maxima are weaker and located farther poleward. *See* STRATOSPHERE; TROPOSPHERE.

In the troposphere, the zonal wind increases upward according to Eq. (1) and with a general poleward decrease in temperature. In the lower stratosphere over most of the globe, this temperature gradient is reversed, and the wind decreases with height. Above about 12 mi (20 km), separate wind systems exist, with prevailing easterly winds in summer and westerlies in winter that attain speeds up to 130–175 mi/h (60–80 m/s) near 40 mi (60 km) height in high latitudes. *See* JET STREAM.

The much weaker meridional circulation consists of six separate cells. Their general locations and nomenclatures are shown in Fig. 2, along with the approximate circulations in terms of mass flux. For each cell, only central streamlines are shown, but these represent flows that are several kilometers deep in the horizontal branches, while each vertical branch represents gentle ascending or descending motions over latitude belts some thousands of kilometers wide. The tropical Hadley cells, best developed in the winter hemisphere, are mainly responsible for maintaining the westerly winds as described below.

Angular momentum balance. The motion of the atmosphere plus the eastward speed of a point on the Earth's surface represent the total motion relative to the Earth's axis. The angular momentum (M) of a unit mass of air is given by Eq. (2), where a represents

$$M = (u + \Omega a \cos\phi) a \cos\phi \qquad (2)$$

the radius of the Earth and the other terms are as defined for Eq. (1). Considering the mean value in a zonal ring, this quantity is conserved unless the ring is subjected to a torque.

The surface easterlies in low latitudes and westerlies in middle latitudes (Fig. 2) exert the principal torques upon the atmosphere, due to frictional drags that are opposite to the direction of the surface winds.

Since the torques would tend to diminish the westerlies and easterlies, and this is not observed to occur over the long run, it follows that angular momentum has to be transferred from the belts of surface easterlies to the zones of westerlies. Calculations show that this meridional flux occurs mainly in the upper troposphere. Hence angular momentum has to be brought upward from the surface layer in low latitudes, transferred poleward, and ultimately brought down to the Earth in the belt of westerlies (Fig. 2).

The vertical transfers are mainly accomplished by the mean meridional circulations, involving the second term on the right-hand side of Eq. (2), which is much larger than the first. Considering the $\cos^2\phi$ factor in this term, there is a net upward transfer by the Hadley cells since their ascending branches are at lower latitudes than the descending branches. Similarly, the opposite Ferrel circulations bring angular momentum downward. With conservation of total angular momentum in the poleward upper branches of the Hadley cells, u in Eq. (2) increases at the expense of the Ω term. This process accounts for the generation of the strongest upper-level westerly winds in subtropical latitudes. From this source, eddies (waves in which poleward flow is associated with a stronger west-wind component than in equatorward flow) transfer angular momentum poleward to sustain the west winds at higher latitudes. *See* ANGULAR MOMENTUM.

Heat energy balance. Maintenance of an average temperature distribution such as in Fig. 2 depends upon a balance between the effects of heat sources and sinks (for example, radiation) and of heat transport by air motion. The sources and sinks are largely a function of latitude and elevation, such that meridional and vertical fluxes of heat energy are required. Two methods by which these fluxes are calculated from different kinds of observations give results that offer a check upon one another.

In the first method estimates are made of the rate of change of any property X per unit area of the Earth's surface due to sources and sinks. To maintain an unchanged condition, their integrated value, over the area north of a latitude ϕ, must equal the northward flux F_ϕ of the property across the latitude circle. This requirement is expressed by Eq. (3), in which t denotes time.

$$F_\phi = -2\pi a^2 \int_\phi^{90°N} \frac{d\overline{X}}{dt} \cos\phi \, d\phi \qquad (3)$$

By employing this method, the energy balance may be calculated by substituting the quantities summarized in **Table 1** into Eq. (3). The listed heat sources comprise R_a, net (absorbed minus emitted) radiation of the atmosphere; R_e, net radiation at the Earth's surface; Q_s, flux of sensible heat from the surface to the atmosphere; LE, flux of latent heat from the surface (E denoting rate of evapotranspiration and L, heat of vaporization); and LP, release of latent heat in the atmosphere as estimated from the observed rate of precipitation P.

The second method is to compute the fluxes directly from aerological observations (made by balloon-borne radiosondes). If x denotes a given property per unit mass of air, the flux is given by Eq. (4), where v is the meridional wind component

$$F_\phi = \frac{2\pi a \cos\phi}{g} \int_0^{p_0} (\overline{x}\,\overline{v} + \overline{x'v'}) \, dp \qquad (4)$$

(positive northward). The integration, with pressure p as a vertical coordinate (related to height z as in Fig. 2), is extended from the bottom ($p = p_0$) to the top of the atmosphere ($p = 0$). Here $\overline{x}$ and $\overline{v}$ denote values averaged over time and longitude, and x' and v' are deviations from these mean values at a given pressure surface. A corresponding expression can be written for the vertical fluxes.

The forms of heat energy x are listed in **Table 2**. The atmospheric energy comprises the sum $c_pT + gz$, the two quantities being interchangeable during

TABLE 1. Sources of atmospheric properties

Different sources	$d\overline{X}/dt$
Atmospheric heat	$\overline{R}_a + \overline{Q}_s + L\overline{P}$
Latent heat	$L(\overline{E} - \overline{P})$
Heat of Earth's surface	$\overline{R}_e - \overline{Q}_s - L\overline{E}$
Heat of atmosphere and Earth	$\overline{R}_a + \overline{R}_e$

TABLE 2. Atmospheric properties used for flux computations

Property	x (per unit mass)*
Atmospheric energy	
Sensible heat	C_pT
Potential energy	gz
Total atmospheric energy	$C_pT + gz$
Latent heat	Lq

*Here c_p is specific heat of air; q is specific humidity, or mass of water vapor per unit mass of air.

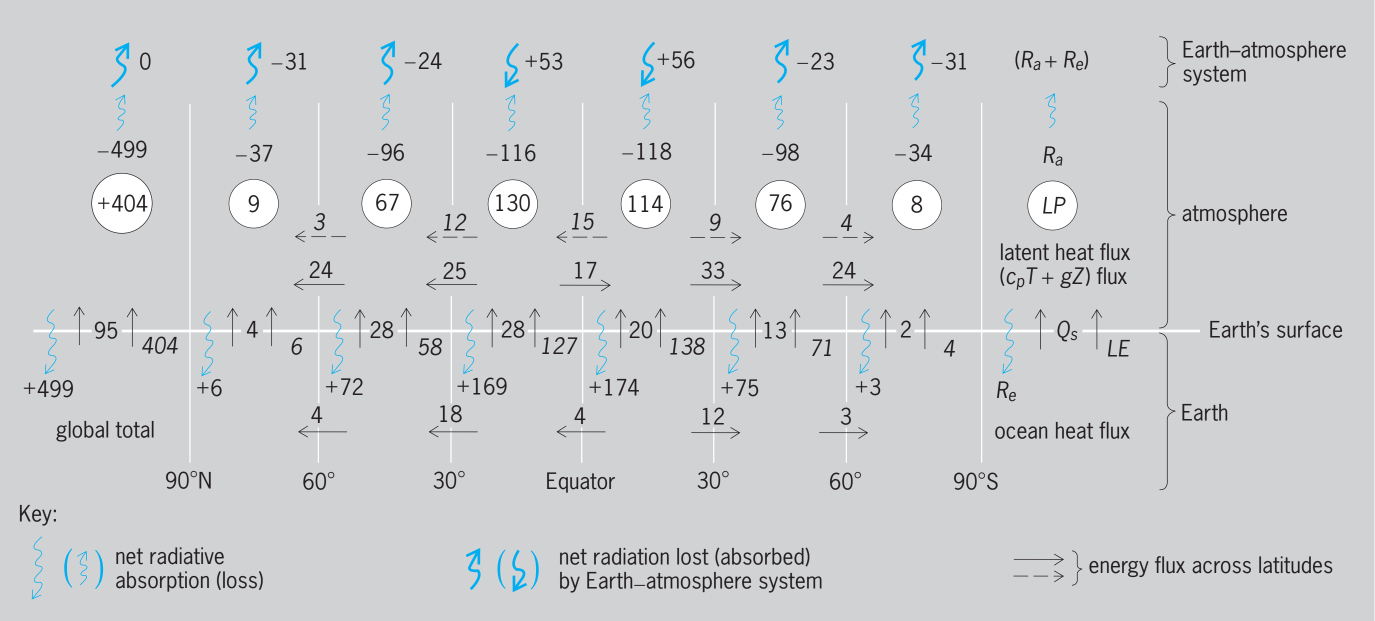
vertical air movement; decompression of rising air diminishes its sensible heat content by an amount equal to the increase of potential energy associated with change of elevation. Almost everywhere $c_pT + gz$ increases upward, and Lq generally but not always decreases upward.

The first term of Eq. (4) represents the mean meridional circulations in Fig. 2, which dominate in the tropical belt. Water vapor conveyed equatorward by the trade winds is condensed in tall convective clouds near the intertropical convergence zone. As a result of latent heat release, the atmospheric energy ($c_pT + gz$) is augmented, so that the upper-level flow of a Hadley cell carries more heat poleward than does the equatorward low-level flow. The second term represents eddy fluxes, dominant in subtropical and higher latitudes, associated with cyclones and anticyclones (Fig. 1) whose poleward flows are both warmest and richest in water vapor.

Various components of the energy balance are summarized, by 30° belts of latitude, in **Fig. 3**. Of the net radiation absorbed at the Earth's surface over the whole globe, 81% is expended in evaporation. Correspondingly, 81% of the net radiative loss by the atmosphere is compensated by release of latent heat when water vapor condenses and falls out as rain, snow, or hail, and 19% by transfer of sensible heat from the Earth. In the tropical belt 30°N–30°S, the Earth–atmosphere system gains heat by radiation; the excess is exported to higher latitudes as atmospheric heat and latent heat, and by ocean currents. Considering the tropical-latitude and temperate-latitude belts of the two hemispheres, significant differences in the apportionments of evaporation and sensible heat transfer from the Earth to the atmosphere, and of the meridional transports of energy in the various forms, arise from the greater dominance of continents in the Northern Hemisphere. On the annual average, water-vapor-laden trade winds (the lower branches of Hadley cells in Fig. 2) converge at about 5°N, where the greatest precipitation is observed. Minima of precipitation occur in the belts near 30°N, where mean descending motions occur. Secondary maxima of precipitation, in latitudes 40–50°, are associated with frequent extratropical cyclones. *See* HYDROMETEOROLOGY.

The frequency and intensity of cyclones, and the contrasts between their cold and warm air masses, are much greater in winter than in summer in the Northern Hemisphere; these variations are much less pronounced in the oceanic Southern Hemisphere. Thus there is a fourfold greater poleward transport of sensible heat in middle latitudes of the Northern Hemisphere in winter than in summer, contrasted with only about a 30% seasonal variation in the Southern Hemisphere. In the tropics, large seasonal changes in the intensities of the Hadley circulation, together with a migration of the rain belt of the intertropical convergence zone (Fig. 1), are most pronounced in the monsoon regions of Asia-Australia and Africa. *See* MONSOON METEOROLOGY.

Between late spring and early autumn a given hemisphere receives solar radiation far in excess



0
−31
−24
+53
+56
−23
−31
(Ra + Re)
Earth–atmosphere system
−499
−37
−96
−116
−118
−98
−34
Ra
+404
9
67
130
114
76
8
LP
atmosphere
3
12
15
9
4
latent heat flux
24
25
17
33
24
(cpT + gZ) flux
95
404
4
6
28
58
28
127
20
138
13
71
2
4
Qs
LE
Earth's surface
+499
+6
+72
+169
+174
+75
+3
Re
global total
4
18
4
12
3
ocean heat flux
Earth
90°N
60°
30°
Equator
30°
60°
90°S
Key:
net radiative absorption (loss)
net radiation lost (absorbed) by Earth–atmosphere system
energy flux across latitudes

Fig. 3. Annual heat balance for the Earth as a whole and for each 30° latitude belt. Units are 10^{14} W. All values are keyed to the column at right. Italic numbers apply to water vapor (latent heat) flux.

of the amount it loses by infrared radiation. The excess heat is stored mainly in the oceans during the warm seasons and given up to the atmosphere as sensible and latent heat during the cooler seasons. Thus the oceans serve as an energy reservoir that tempers the seasonal changes of atmospheric temperature over them and over neighboring land areas invaded by marine air masses. *See* MARITIME METEOROLOGY.

C. W. Newton

Bibliography. J. R. Holton, *An Introduction to Dynamic Meteorology,* 3d ed., 1992; J. P. Peixóto and A. H. Oort, Physics of climate, *Rev. Mod. Phys*., 56:365–429, 1984; S. Petterssen, *Introduction to Meteorology*, 3d ed., 1969.

Atmospheric tides

Those oscillations in any or all atmospheric fields whose periods are integral fractions of either lunar or solar days. Oscillations with a period of a day are called diurnal, with a period of a half day semidiurnal, and with a period of a third of a day terdiurnal. A common nomenclature does not exist for smaller fractions of a day. The sum of all tidal variations is referred to as the daily variation. As a practical matter, the subject of atmospheric tides is generally restricted to oscillations on a global spatial scale (thus excluding sea breezes). Moreover, the bulk of attention is devoted to migrating tides, which are those tidal oscillations that depend only on local time.

Atmospheric tides tend to be rather small in the troposphere, although the tidal oscillations in rainfall are surprisingly large. Their importance stems from two primary factors: (1) Tidal oscillations tend to increase in amplitude with height and become major components of the total meteorology above about 50 km (30 mi). (2) The subject has played a prominent role in the intellectual history of meteorology, and it still provides a remarkable example of scientific methodology in an observational science. In connection with the second item, it should be noted that tides are unique among meteorological systems in that they have perfectly known periods and relatively well known sources of forcing (though the determination of these sources constitutes a major part of the history of the subject).

Observational description. The determination of an oscillation by means of data requires at least two measurements per period. Since most meteorological upper air soundings are taken only twice each day, such data can be used only to marginally determine diurnal oscillations. Occasionally, stations obtain soundings four times per day, which in turn permits determinations of semidiurnal oscillations. Rain gages assign rainfall to specific hours, and averages over many years allow the determination of the daily variation of rainfall. Surface pressure is monitored at a great many stations with effectively (from the point of view of tidal analyses) continuous time resolution. Therefore, surface pressure has traditionally been the field most thoroughly analyzed for tides.

To a fair degree of approximation, the distributions of solar semidiurnal (S_2) and diurnal (S_1) tides in surface pressure (P_s) are given by Eqs. (1) and (2), where

$$S_2(P_s) = a \sin^3\theta \sin(2t + 158^\circ) + bP_2(\theta)\sin(2t_u + 118^\circ) \quad (1)$$

$$S_1(P_s) = c \sin^3\theta \sin(t + 12^\circ) \quad (2)$$

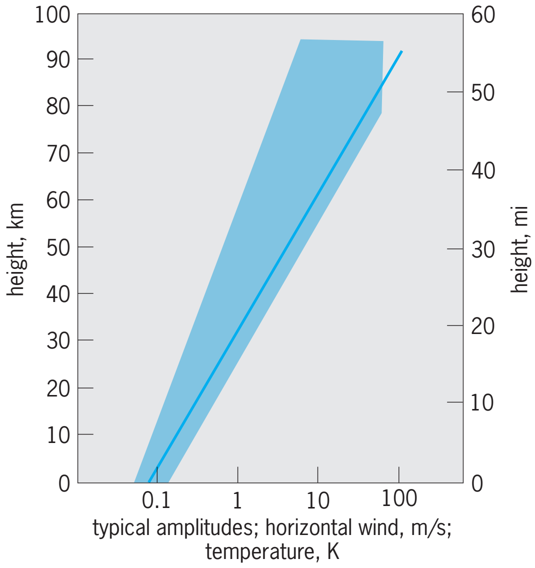
$a = 1.16$ millibar (116 pascals), $b = 0.085$ millibar (8.5 Pa), $c = 0.593$ millibar (59.3 Pa), $\theta =$ colatitude, $t =$ local time $t_u =$ Greenwich (Universal) Time, and $P_2\theta \equiv {}^1/_2\cos^2\theta - 1)$.

The first term on the right-hand side of Eq. (1) represents a migrating semidiurnal tide with maximum surface pressure near 10 A.M. and 10 P.M. local time, while the second term represents a standing oscillation in Universal Time. The diurnal tide is primarily migrating with maximum pressure occurring near 5 A.M. local time. The lunar semidiurnal tide in surface pressure is similar in distribution to the migrating part of the solar semidiurnal tide but only about one-twentieth its strength; maximum lunar semidiurnal surface pressure typically occurs about 1 lunar hour and 13 lunar hours after lunar transit. Clearly, the solar semidiurnal tide dominates the surface pressure. The solar diurnal component is not only smaller but also far more irregular.

Rainfall is commonly observed to have a daily variation. The diurnal component, though often quite large, has a very irregular phase; on the other hand, the solar semidiurnal component is surprisingly uniform, amounting to about 10–20% of the mean daily rainfall in the tropics with maximum rainfall at about 4 A.M. and 4 P.M. local time. Maximum semidiurnal rainfall appears to occur somewhat later in middle latitudes. *See* PRECIPITATION (METEOROLOGY).

Data become more sparse when attempts are made to analyze tides above the surface. Analyses of radiosonde wind measurements have shown that solar semidiurnal oscillations in horizontal wind are primarily in the form of migrating tides. Diurnal oscillations, on the other hand, are significantly affected by regional, nonmigrating components up to at least 20 km (12 mi). Above this height, the diurnal oscillations also tend to be dominated by migrating components. There is a tendency for the diurnal oscillations to have more phase variation with height, especially at low latitudes. The **illustration** shows the approximate variation of tidal amplitudes with height. As a rough rule of thumb, oscillations in temperature tend to have magnitudes in kelvins comparable to the amplitudes in wind in meters per second. The illustration shows the general increase in amplitude with height. There is also no longer a clear dominance of the semidiurnal oscillations over the diurnal oscillations once the upper-level fields are considered. The amplitude increase with height renders the detection of tidal oscillations at greater altitudes somewhat easier since the tides are becoming a larger (and sometimes dominant) feature of the total fields. *See* OSCILLATION.

Meteorological rocket soundings have permitted extension of the knowledge about tides to about a 60-km (36-mi) altitude. Below an altitude of 50 km



Variation of tidal amplitudes with height. The shaded area shows the general range of solar diurnal and semidiurnal amplitudes (on a logarithmic scale) as a function of height. The line corresponds to amplitude varying inversely with the square root of the basic pressure. In classical tidal theory, amplitudes scale by this factor, which leads to energy density being independent of height.

(30 mi), semidiurnal horizontal wind oscillations tend to be smaller than 1–2 m/s; however, diurnal oscillations are appreciably stronger, reaching 3–4 m/s near 40 km (24 mi) and rising to about 6–10 m/s above 50 km (30 mi). Above 50 km (30 mi) the semidiurnal oscillations also increase to about 6 m/s. Diurnal oscillations equatorward of 30° latitude tend to be associated with rapid vertical phase variations; poleward of 30° these phase variations tend to be absent. Semidiurnal oscillations tend to have no phase variations with height, except for a 180° phase shift someplace between 30 and 50 km (18 and 30 mi). Data above 60 km (36 mi) are obtained from larger rockets and satellite observations, as well as from a variety of radar techniques. Amplitudes are generally found to still increase with height up to about 80 or 90 km (48 or 54 mi). At 80 km (48 mi), amplitudes of as much as 40–50 m/s are found. At this height, diurnal oscillations tend to be dominant equatorward of 30°, while semidiurnal oscillations tend to be dominant poleward of 30°. The diurnal tidal contributions to the temperature come close to producing static instability, and may be an important source of turbulence in the upper atmosphere. Semidiurnal tides begin to show evidence of more phase variation with height. There tends to be a minimum in diurnal amplitudes near about 105 km (65 mi); within the thermosphere, tidal amplitudes are typically large, increasing asymptotically to constant values above 200 km (120 mi). These values are commonly in excess of 100 m/s and 100 K. *See* CONVECTIVE INSTABILITY; DYNAMIC INSTABILITY; METEOROLOGICAL ROCKET.

Historical considerations. The behavior of tides in the surface pressure was already well known by the nineteenth century. Ocean tides were known to be dominated by the lunar semidiurnal component. This was consistent with their being forced primarily by the tidal component of the gravitational potential. Here, lunar forcing is stronger than solar forcing (because the radius of the Earth is a much larger fraction of the lunar distance than the solar distance, and because gravitational forcing is primarily semidiurnal). The fact that atmospheric tides are primarily solar rather than lunar suggests a thermal origin. Lord Kelvin, however, noted that thermal forcing would be primarily diurnal rather than semidiurnal. It was thus puzzling that the surface pressure tide was primarily semidiurnal. He put forth the hypothesis that the atmosphere might be resonant at 12 h, and this might explain the preferential semidiurnal response.

The search for resonance was the primary focus of theoretical research on atmospheric tides through the 1950s. It was recognized that the resonance properties of the atmosphere were dependent on the thermal structure of the atmosphere. The high point in the search occurred in 1937, when C. L. Pekeris discovered that the thermal structure needed for solar semidiurnal resonance was almost identical to the independently inferred structure based on observations. The history of this inquiry had followed almost perfectly the paradigm for research in an observational science. Observations of tides had led to a hypothesis whose development led to independent implications (namely, the thermal structure of the atmosphere), which could be confirmed or refuted by additional measurements. Success, however, does not rigorously prove the original hypothesis; indeed, in the years following World War II, both observational and theoretical advances led to the rejection of the resonance hypothesis. Specific thermal sources of forcing (namely, insolation absorption by both water vapor and ozone) have proven adequate to account for observed tidal fields below about 100 km (60 mi) without any appeal to resonance. The agreement is by no means perfect, and it now appears that the latent heating associated with the daily variation in rainfall may constitute an additional important source of forcing.

Theory. The classical theory of atmospheric tides is based on a linearization of the equations of motion for a rotating spherical inviscid atmosphere about a state of rest. The resulting equations allow simple harmonic dependences on time and longitude, and are separable in their dependence on height and latitude. Latitude dependence is described by Laplace's tidal equation, which, for a given frequency and wave number, can be solved for an infinite number of eigenfunctions (known as Hough functions) and their associated eigenvalues (known as equivalent depths). Height dependence is described by the so-called vertical structure equation. This inhomogeneous equation includes (in either the equation or its boundary conditions) the gravitational and thermal forcing for each Hough mode, which is the complete solution for all physical fields (that is, horizontal and vertical wind, temperature, pressure, and density) including their vertical structure associated with a given Hough function. The nature of the vertical response is determined by both the thermal structure of the atmosphere and the mode's equivalent depth. As a rule, the vertical structure equation yields only the trivial zero solution when there is no forcing. However, for a few equivalent depths it is sometimes possible to get nontrivial solutions in the absence of forcing; these are then said to be resonant equivalent depths. *See* EIGENFUNCTION.

The nature of the Hough functions and equivalent depths is significantly determined by the ratio of the oscillation frequency to the quantity 2Ω (the Coriolis parameter at the poles where Ω is the Earth's rotation rate). When this quantity is greater than 1, there are no latitudes where the frequency equals the local Coriolis parameter ($2\Omega \sin \phi$, where ϕ is the latitude). The Hough functions in these cases are closely related to spherical harmonics, and the equivalent depths are positive. Positive equivalent depths are generally associated with wavelike propagation in the vertical, though sufficiently large equivalent depths lead to vertical trapping (that is, exponential as opposed to wavelike behavior). The situation is markedly different when this ratio is less than 1. Then, two latitudes exist where the frequency equals the local Coriolis parameter. Beyond these latitudes the Coriolis parameter exceeds the frequency, and this, it turns out, inhibits vertical propagation. The way that this manifests itself for Laplace's tidal equation is that the eigensolutions divide into two classes: one with amplitudes concentrated poleward of these critical latitudes and having negative equivalent depths (which are always associated with trapping in the vertical), and the other with amplitudes concentrated equatorward of these critical latitudes and having positive equivalent depths. With reference to atmospheric tides, the critical latitudes for semidiurnal modes are essentially the poles, and all modes have positive equivalent depths. However, for diurnal modes the critical latitudes occur at 30°, and the eigenfunctions are divided into the two categories described above. As it turns out, the gravest diurnal mode, into which the bulk of the thermal forcing goes, has a negative equivalent depth. The so-called graveness of a mode is determined by the number of zeros that the associated Hough function has; the gravest mode is that mode for which the associated Hough function has no zeros except at the poles. Thus the effect of upper-level thermal forcing is unable to effectively penetrate to the surface. This is a primary reason for the dominance of the semidiurnal oscillation at the ground. The confinement of positive equivalent-depth diurnal modes to the region equatorward of 30° also accounts for the relative dominance of diurnal oscillations in this region in the neighborhood of 80 km (48 mi) altitude. *See* CORIOLIS ACCELERATION.

While the classical theory of atmospheric tides is adequate for many purposes, recent years have seen a substantial development of theory well beyond the classical theory to include the effects of mean winds, viscosity, and thermal conductivity. Very little progress has been made in explaining the tides in rainfall. *See* ATMOSPHERE; EARTH TIDES; TIDE.

Richard S. Lindzen

Bibliography. S. Chapman and R. S. Lindzen, *Atmospheric Tides*, 1970; Lord Kelvin (W. Thomson), On the thermodynamic acceleration of the earth's rotation, *Proc. Roy. Soc. Edinburgh*, 11:396–405, 1882; R. S. Lindzen, Atmospheric tides, *Annu. Rev. Earth Planet. Sci.*, 7:199–225, 1979; R. S. Lindzen, *Dynamics in Atmosphere Physics*, 1990; H. Volland, *Atmospheric Tidal and Planetary Waves*, 1988.

Atmospheric waves, upper synoptic

Horizontal wavelike oscillations in the pattern of wind flow aloft in the upper troposphere, usually with reference to the stronger portion of the westerly current in midlatitudes. The flow is anticyclonically curved in the vicinity of a ridge line in the wave pattern, and is cyclonically curved in the vicinity of a trough line. *See* TROPOSPHERE.

Any given hemispheric upper flow pattern may be represented by the superposition of sinusoidal waves of various lengths in the general westerly flow. Analysis of a typical pattern discloses the presence of prominent long waves, of which there are usually three or four around the hemisphere, and of distinctly evident short waves, of about half the length of the long waves.

Typically, each short-wave trough and ridge is associated with a particular cyclone and anticyclone, respectively, in the lower troposphere. The development and intensification of one of these circulations depends in a specific instance upon the details of this association, such as the relative positions and intensities of the upper trough and the low-level cyclone. These circulations produce the rapid day-to-day weather changes which are characteristic of the climate of the mid-latitudes.

The long waves aloft do not generally correspond to a single feature of the circulation pattern at low levels. They are relatively stable, slowly moving features which tend to guide the more rapid motion of the individual short waves and of their concomitant low-level cyclones and anticyclones. Thus, by virtue of their position and amplitude the long waves can exert an indirect influence on the character of the weather over a given region for a period of the order of weeks.

The motion, but not the amplification, of long waves and short waves can be predicted with considerable success by application of the principle of conservation of absolute vorticity to the flow pattern at midtropospheric elevations.

A blocking situation is one in which waves do not progress through the latitude belt of the westerlies. Blocking situations are frequently accompanied by extreme meteorological events; precipitation and cool temperatures persist near upper-level cyclones, and dry, warm weather persists near upper-level anticyclones. A blocking pattern usually consists of a ridge (anticyclone) over a trough (cyclone), a high-amplitude ridge, or flow shaped like an uppercase Greek omega (Ω). Because of the preference for blocking off the west coasts of Europe and North America, it appears that topography must play an important role in blocking. *See* ATMOSPHERE; JET STREAM; STORM; VORTEX; WEATHER FORECASTING AND PREDICTION; WIND.

Frederick Sanders; Howard B. Bluestein

Bibliography. S. Colucci, An observational study of mid-tropospheric blocking during recent winter seasons, *The General Circulation: Theory, Modeling, and Observations,* Advanced Study Program of NCAR, Boulder, 397–405, 1978; R. Dole, The objective representation of blocking patterns, *The General Circulation: Theory, Modeling, and Observations,* Advanced Study Program of NCAR, Boulder, pp. 406–426, 1978.

Atoll

An annular coral reef, with or without small islets, that surrounds a lagoon without projecting land area.

Physiography. Most atolls are isolated reefs that rise from the deep sea. They vary considerably in size and shape. Small rings, some without islets, may be less than a mile in diameter, but many atolls have a diameter of 20 mi (32 km) or more and bear numerous islets. The largest atoll, Kwajalein in the Marshall Islands, has an irregular shape, is about 75 mi (120 km) long, 15–20 mi (24–36 km) wide, and covers 840 mi^2 (2180 km^2).

The reefs of the atoll ring are flat, pavementlike areas, large parts of which, particularly along the seaward margin, may be exposed at times of low tide (**Fig. 1**). The reefs vary in width from narrow ribbons to broad bulging areas more than a mile across. On windward reefs exposed to constant trade wind surf, the most prominent feature is commonly a low, cuestalike wave-resistant ridge whose steeper side faces the sea. The ridge, as much as 3 ft (1 m) above lowest tides, is composed of pink or rose-colored calcareous algae. To seaward of the ridge the gentle submarine slope is cut by a series of grooves. These may be slightly sinuous, in places branching as they extend seaward from the reef front. They are separated by flattened spurs or buttresses that appear to have developed mainly by coral and algal growth. The grooves and buttresses form the "toothed edge" that is so conspicuous when an atoll is viewed from the air. Beyond the toothed edge may be a shallow terrace, then a steep (35°) submarine slope that flattens progressively at deeper levels until it reaches the ocean floor. Some grooves of the toothed margin extend back through the marginal algal ridge for short distances as open surge channels; in places of exceptional development, deeper interconnecting channels may extend for several hundred feet under the reef flat, roofed over by algal growth and exposed here and there in open pools, or ending as a series of blowholes that spout with every wave. These structures taken together form a most effective baffle that robs the incoming waves of much of their destructive power, as they bring a constant supply of refreshing seawater with oxygen, food, and nutrient salts to wide expanses of reef. On leeward reefs, protected from the constant surf that characterizes windward reefs, the algal margin is less developed. Coral growth may be profuse to the reef edge which drops off sharply, in places vertically, for several hundred feet.

Inside the margin of the reef may be several zones that are rich in corals or other reef builders, but large parts of the reef flat are composed of cemented reef rock, bare or thinly veneered with loose sand or coarser reef debris. When the reef flats of Eniwetok Atoll were drilled, a solid plate of hard limestone was found below that extends to depths of 10–15 ft (3–4.5 m) and contains corals in growth position. Growth of organisms and deposition of sediments on the reef are controlled by the prevailing winds and currents and other ecologic factors. The sediments, including cemented rock, that make up the reef surface and its islets are composed entirely of carbonates secreted by shallow-water organisms.

The reefs of most atolls are broken, commonly to leeward, by one or more passes. Some of these may be as deep as the lagoon and give ready access to that body of water. Within atoll groups, lagoon depths show little relationship to size, except that very small lagoons are typically shallower than larger ones in the same group. Most atoll lagoons in the Caroline and Marshall islands range in depth from 100 to 300 ft (30 to 90 m), but the lagoon of Christmas Atoll, a very large atoll in the Line Islands, is only 23 ft (7 m) deep. Hundreds of coral knolls rise from the lagoon floor of many atolls. More than 2000 were charted in Eniwetok lagoon. Other atolls may show

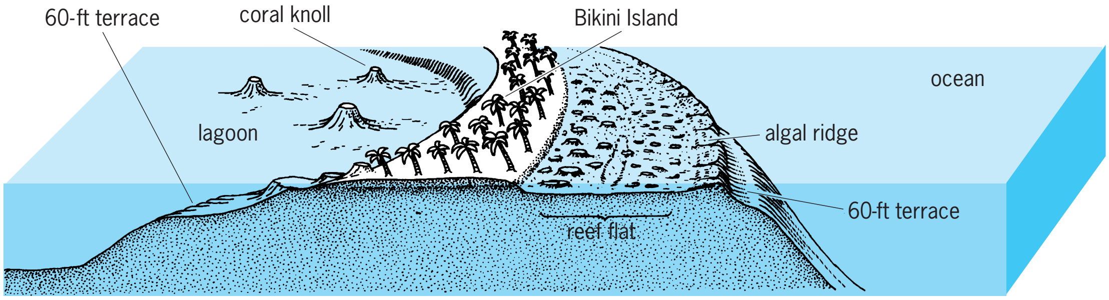


Fig. 1. Section showing features of rim of Bikini Atoll.

very few coral knolls, and some relatively shallow lagoons (Canton in the Phoenix Islands, Mataiva in the Tuamotus, Christmas Atoll) are divided into compartments by an intersecting network of linear coral ridges.

On many atolls islets are spaced at irregular intervals along the reef. Most of these are small and rarely rise as much as 20 ft (6 m) above the reef flat. They are composed of reef detritus (coral heads and sand) piled up by waves and winds. The sediments are unconsolidated except for layers of beach rock (cemented conglomerate and sandstone) at intertidal levels. Islets located on reefs that are subjected periodically to violent storms may exhibit one or a series of ridges or ramparts built of blocks of reef rock, coral heads, and finer material. These structures apparently can be built, moved, or destroyed by the waves of a single major storm. Low sand dunes may occur on the borders of islands inside sand beaches. Inland from the beaches and ramparts, sand and gravel flats occur and large areas are covered with coconut palms, breadfruit trees, and *Pandanus* (screw pines). More isolated islets may contain groves of native *Pisonia*, inside a beach ridge covered with *Tournefortia*, *Cordia*, and other native trees.

Many atolls exhibit terraces around their margins and inside their lagoons. Shallow drilling on Eniwetok islets has shown that the reef rock to a depth of 30 ft (9 m) or more below sea level was deposited beginning about 8000 years ago during the Holocene sea-level rise; the reef limestone immediately below that level and presumably that of the terraces was more than 130,000 years old and grew during the previous interglacial sea-level rise. *See* HOLOCENE.

Distribution. Existing atolls, like other types of coral reefs, require strong light and warm waters and are limited in the existing seas to tropical and near-tropical latitudes. A large percentage of the world's more than 400 atolls are contained in an area known as the former Darwin Rise that covers much of the central and western Pacific. Atolls are also numerous in parts of the Indian Ocean and a number are found, mostly on continental shelves, in the Caribbean area. *See* INDIAN OCEAN; PACIFIC OCEAN; TROPICAL METEOROLOGY.

The Darwin Rise also contains more than 150 flat-topped seamounts or guyots, which were atoll-like structures that existed at or nearly at sea level during Cretaceous time but are now drowned several thousand feet deep. Some, possibly many, atolls such as Bikini and Eniwetok have grown on a foundation of Cretaceous guyots during the Cenozoic. Similar atoll-like structures existed in shallow Paleozoic seas. These ancient atolls were built by a variety of organisms; the types of corals and algae that are primarily responsible for the present reefs were not then in existence. Many Paleozoic atolls became reservoirs of petroleum, and are being studied and mapped in detail by deep drilling. *See* PALEOZOIC; SEAMOUNT AND GUYOT.

Origin and development. Most atolls that rise from the deep sea have probably been built on submerged volcanoes. The reef caps of four such atolls (Eniwetok, Midway, Mururoa, and Fangataufa) have been drilled and the existence of a foundation of volcanic flows confirmed. In these atolls it was found that the volcanoes had been at least partly truncated by erosion prior to the initiation of reef growth. Many of the existing atolls probably originated as near-surface reefs in Tertiary time. The oldest limestones beneath Eniwetok are late Eocene, although the underlying basalt is Late Cretaceous (76.5 million years old), and Midway's oldest limestones are early Miocene. The bases of Mururoa and Fangataufa are late Miocene, and other atolls may prove to be appreciably younger. *See* EOCENE; MIOCENE; TERTIARY.

Wherever extended three-dimensional studies have been carried out on atolls—Funafuti in the Ellice Islands (Tuvalu), Kita-Daito-Jima, Bikini and Eniwetok in the Marshall Islands, Midway in Hawaii, or Mururoa and Fangataufa in the Tuamotus—they have proved the atolls to be complex structures; surface similarities may mask striking differences in underground structure. Wherever tested by the drill, the thicknesses of reef caps of atolls greatly exceed the depth limits of the reef builders, and these caps appear to have developed during long periods of subsidence.

The structure of Eniwetok Atoll in the Marshall Islands has been revealed by drilling (**Fig. 2**). The limestone cap, nearly a mile in thickness, rests on the truncated summit of a basaltic volcano that rises 2 mi (3 km) above the ocean floor. In one of the drillholes the lower limestones are late Eocene and formed as forereef deposits on an outer slope. Correlative limestones in a second hole were formed in shallow water. The 4000-ft (1200-m) section of limestone that forms the cap of Eniwetok records an overall subsidence of that amount over a period exceeding 40 million years. Subsidence of this magnitude even exceeds the amounts so brilliantly postulated by Charles Darwin. The sinking has not been steady or continuous but was interrupted by periods of relative emergence, as during periods of glaciation. The emergent stages are recorded in zones of

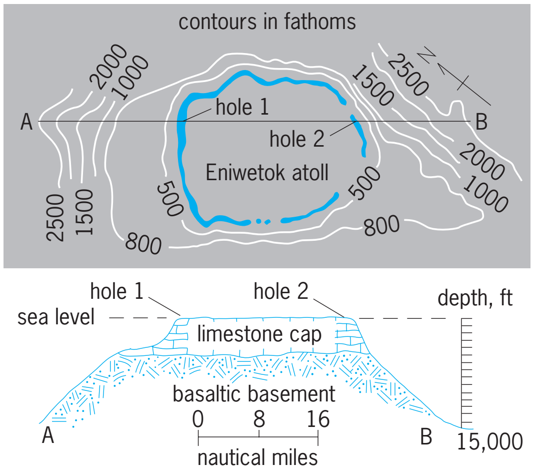


Fig. 2. Structure of Eniwetok Atoll in Marshall Islands. Lower limestones in hole 1 are late Eocene. Hole 2 reveals correlative limestones. 1 ft = 0.3 m; 1 naut mi = 1.85 km; 1 fathom = 1.83 m.

recrystallized and partially dolomitized sediments overlain by less altered limestone. Such sequences, termed solution unconformities, were formed when the top of the atoll stood above the sea for appreciable time. During such periods land snails lived on the island and land plants left a record of spores and pollen; Eniwetok and perhaps many other atolls functioned as stepping stones in the distribution of shallow-water marine life in the Pacific. *See* OCEANIC ISLANDS; REEF; SEQUENCE STRATIGRAPHY.

Harry S. Ladd; Joshua I. Tracey, Jr.

Bibliography. A. Guilcher, *Coral Reef Geomorphology*, 1988; N. M. Hill (ed.), *The Sea*, vol. 3, 1963, reprint 1981; H. S. Ladd et al., Drilling on Midway Atoll, Hawaii, *Science*, 156(3778):1088–1094, 1967; H. S. Ladd and S. O. Schlanger, *Drilling Operations on Eniwetok Atoll*, USGS Prof. Pap. 260-Y, 1960; H. J. Wiens, *Atoll Environment and Ecology*, 1962.

Atom

A constituent of matter consisting of z negatively charged electrons bound predominantly by the Coulomb force to a tiny, positively charged nucleus consisting of Z protons and $(A - Z)$ neutrons. Z is the atomic number, and A is the mass or nucleon number. The atomic mass unit is $u = 1.661 \times 10^{-27}$ kg. *See* ATOMIC MASS UNIT; ATOMIC NUMBER; COULOMB'S LAW; ELECTRON; MASS NUMBER; NEUTRON; PROTON.

Atoms are the smallest objects that have a chemical identity. Atoms are thus the basic building blocks of matter. Electrically neutral atoms ($z = Z$) with the range $Z = 1$ (hydrogen) to $Z = 92$ (uranium) make up the periodic table of the elements naturally occurring on Earth. (An element is the name of an atom having a given atomic number.) *See* ELEMENT (CHEMISTRY); PERIODIC TABLE.

Isotopes of a given element have different values of A but nearly identical chemical properties, which are fixed by the value of Z. Certain isotopes are not stable; they decay by various processes called radioactivity. Atoms with Z greater than 92 are all radioactive but may be synthesized, either naturally in stellar explosions or in the laboratory using accelerator techniques. *See* ISOTOPE; NUCLEAR STRUCTURE; RADIOACTIVITY; TRANSURANIUM ELEMENTS.

Atoms with $Z - z$ ranging from 1 to $Z - 1$ are called positive ions. Those having $z - Z = 1$ are called negative ions; none has been found with $z - Z$ greater than 1. *See* ION; NEGATIVE ION.

A Rydberg atom has one electron in a highly excited orbit. Its radius grows and binding energy decreases with the square and the inverse square, respectively, of the principal quantum number n. A neutral Rydberg atom with $n = 100$ has a huge diameter, approximately 10^{-6} m or 4×10^{-5} in. (near that of a bacterium). Its miniscule binding energy, approximately 1.36 millielectronvolts, explains its fragility in collisions with other particles. Occurring naturally in the interstellar medium, Rydberg atoms with n-values up to several hundred emit spectral lines that have been detected with radio telescopes. They have also been produced and studied in the laboratory. *See* RYDBERG ATOM.

An exotic atom has a different charged particle substituted for an electron, its nucleus, or both. For example, positronium consists of an electron and its antiparticle, a positron. *See* ATOMIC STRUCTURE AND SPECTRA; HADRONIC ATOM; MUONIUM; PIONIUM; POSITRONIUM.

Peter M. Koch

Atom cluster

Clusters are aggregates of atoms (or molecules) containing between three and a few thousand atoms that have properties intermediate between those of the isolated monomer (atom or molecule) and the bulk or solid-state material. The study of such species has been an increasingly active research field since about 1980. This activity is due to the fundamental interest in studying a completely new area that can bridge the gap between atomic and solid-state physics and also shows many analogies to nuclear physics. However, the research is also done for its potential technological interest in areas such as catalysis, photography, and epitaxy. A characteristic of clusters which is responsible for many of their interesting properties is the large number of atoms at the surface compared to those in the cluster interior. For many kinds of atomic clusters, all atoms are at the surface for sizes of up to 12 atoms. As the clusters grow further in size, the relative number of atoms at the surface scales as approximately $4N^{-1/3}$, where N is the total number of atoms. Even in a cluster as big as 10^5 atoms, almost 10% of the atoms are at the surface. Clusters can be placed in the following categories:

1. Microclusters have from 3 to 10–13 atoms. Concepts and methods of molecular physics are applicable.

2. Small clusters have from 10–13 to about 100 atoms. Many different geometrical isomers exist for a given cluster size with almost the same energies. Molecular concepts lose their applicability.

3. Large clusters have from 100 to 1000 atoms. A gradual transition is observed to the properties of the solid state.

4. Small particles or nanocrystals have at least 1000 atoms. These bodies display some of the properties of the solid state.

Geometry. Traditionally, solid-state physics describes a crystal as an infinitely extending periodic ordering of atoms with translational symmetry. Body-centered cubic (bcc), face-centered cubic (fcc), and hexagonal close packing (hcp) are the most common arrangements. These orderings are not normally found among clusters. The packing of the atoms in clusters can be investigated by electron diffraction on the cluster beam. The most favored geometry for rare-gas (neon, argon, and krypton) clusters of up to a few thousand atoms (**Fig. 1**) is icosahedral. The fivefold rotational axis of these structures, which first appears in clusters of seven atoms, is forbidden for the symmetry of the standard lattices of solid-state physics. As confirmed by computer simulation

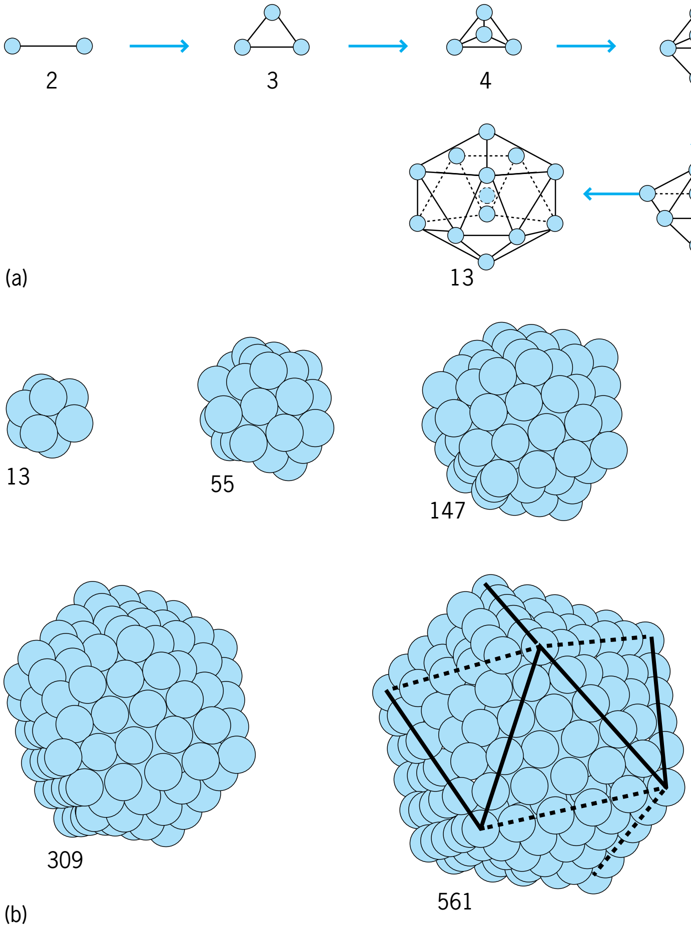


Fig. 1. Growth sequence of neutral rare-gas clusters. Number of atoms in cluster, *N*, is indicated for each cluster. (*a*) Microclusters. (*b*) Small and large clusters, all of which are icosahedra with fivefold symmetry. (*After H. Haberland, ed., Clusters of Atoms and Molecules, vol. 1, Springer-Verlag, 1994*)

methods, the icosahedral geometry allows the atoms to be aggregated into a more compact structure with a smaller potential energy than the structures allowed in solid-state physics, such as the face-centered cubic structure of bulk rare-gas solids. *See* CRYSTAL STRUCTURE.

However, the preferred cluster geometry depends critically on the bonding between the monomers in the clusters. For example, ionic clusters such as those of sodium chloride [$(NaCl)_N$] very rapidly assume the cubic form of the bulk crystal lattice, and for metallic clusters it is the electronic structure rather than the geometric structure which is most important. A summary of the different kinds of clusters, classified in terms of their chemical bonding, is given in the **table**. Mercury clusters (Hg_N) are particularly intriguing since they show van der Waals, covalent, or metallic bonding, depending on the cluster size. *See* CHEMICAL BONDING.

Production. A cluster apparatus typically consists of two separately pumped vacuum chambers which are connected by a small conical aperture, the skimmer. The central part of the cluster beam, formed in the first vacuum chamber, passes through the skimmer and is investigated in some way, ionized, and detected in a mass spectrometer in the second chamber.

There are two main types of sources for producing free cluster beams. In a gas-aggregation source, the atoms or molecules are vaporized into a cold, flowing rare-gas atmosphere. They cool by undergoing many collisions with the rare-gas atoms and aggregate into clusters. In a jet-expansion source, a gas is expanded under high pressure through a small hole into a vacuum. An extreme cooling of the rotational and vibrational degrees of freedom as well as the relative motion occurs during this adiabatic expansion, leading to cluster formation.

Classification of clusters according to their chemical bonding

Kind of cluster	Example	Average binding energy per atom, eV	Elements for which cluster type is found
Metallic clusters: half-filled band of delocalized electrons	(alkali metal)$_N$, Al_N, Cu_N, Fe_N, Pt_N, W_N, Hg_N with $N > 200$	0.5±3	Elements of the lower-left corner of the periodic table
Covalent clusters: directed bonding by electron pairs through *sp*-hybridization	C_N, Si_N, Hg_N, $80 \geq N \geq 30$	1±4 (Hg ≈ 0.5)	B, C, Si, Ge
Ionic clusters: bonding due to Coulomb force between ions	$(NaCl)_N$, $(CaBr_2)_N$	2±4	Metals from the left side of the periodic table with electronegative elements from the right side
Hydrogen-bonded clusters: strong dipole-dipole attraction	$(HF)_N$, $(H_2O)_N$	0.15±0.5	Molecules with closed electronic shells that contain H and strong electronegative elements (F, O, N)
Molecular clusters: like van der Waals with an additional weak covalent contribution	$(I_2)_N$, $(S_6)_N$, (organic molecules)$_N$	0.3±1	Organic molecules, some other closed-shell molecules
Van der Waals clusters: induced dipole interaction between atoms and molecules with closed electronic shells	(rare gas)$_N$, $(H_2)_N$, $(CO_2)_N$, Hg_N with $N < 10$	0.01±0.3	Rare gases, closed-shell atoms and molecules

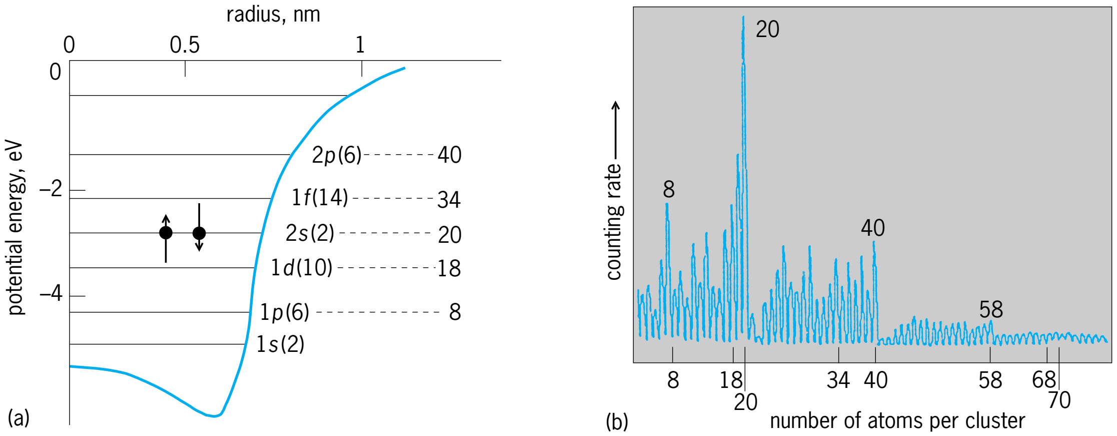


Fig. 2. Jellium model. (*a*) Potential energy of the electrons as a function of the cluster radius for a spherical sodium cluster with 20 atoms (Na_{20}). Electron energy levels are also shown, with their radial and angular-momentum quantum numbers and the degeneracies of the levels (in parentheses). The total number of electrons in a given shell is indicated on the right. Only the two electrons in the highest occupied orbital are indicated. (*b*) Mass spectrum of sodium (Na_N) clusters, showing high intensities (magic numbers) for clusters with completely filled electronic shells. (*After W. A. de Heer et al., Electronic shell structure and metal clusters, Sol. State Phys., 40:93–181, 1987*)

One crucial requirement for both these sources is to obtain a sufficient density of monomers to produce a beam with a high enough intensity of clusters for further investigation.The common way of doing this for materials with a high melting point is to use a laser-vaporization or ion-sputtering source. An intense ion or laser beam is directed onto the surface of the target material and removes material in the form of neutral or charged particles, which then cluster together. The number of clusters, their size, and their temperature can be controlled by combining this method with those of the gas-aggregation or jet-expansion source.

Properties. In most situations, the valence electrons of the atoms making up the clusters can be regarded as being delocalized, that is, not attached to any particular atom but with a certain probability of being found anywhere within the cluster. The simplest and most widely used model to describe the delocalized electrons in metallic clusters is that of a free-electron gas, known as the jellium model. The positive charge is regarded as being smeared out over the entire volume of the cluster while the valence electrons are free to move within this homogeneously distributed, positively charged background. The calculated potential for the electrons in a spherical jellium approximation typically looks like the example in **Fig. 2*a***. Here, the inner part of the effective potential resembles the bottom of a wine bottle. The electronic energy levels are grouped together to form shells. The jellium potential is very similar to the Woods-Saxon potential, used to describe the interaction between protons and neutrons in nuclear physics, and the same classification of energy levels is found, wherein the energy levels are characterized by the radial quantum number (1, 2, 3, . . .) and the angular-momentum quantum number (*s*, *p*, *d*, . . .). *See* NUCLEAR STRUCTURE.

The predicted shell closures can be seen in mass spectra of alkali-metal clusters. If the neutral clusters produced from a given cluster source are hot, they can fragment on leaving the source. The intensity of relatively stable clusters will thus grow at the expense of their less stable neighbors, producing so-called magic numbers in the mass distribution. The neutral mass distribution is then probed by ionizing the clusters with a photon energy that is just slightly higher than the ionization potential of the clusters to avoid further fragmentation after ionization. This operation was used to obtain the mass spectrum of sodium clusters in Fig. 2*b*, which gives a beautiful confirmation of the electronic shell model. *See* MAGIC NUMBERS.

The delocalized electrons can be excited by photon absorption or electron collisions into a collective motion with respect to the positively charged nuclei, which is known as a plasmon resonance. This is very analogous to the giant resonances, that is, the collective motion of neutrons and protons, found in atomic nuclei. The plasmon frequency for metal clusters depends on the cluster size and material but is generally in the visible range of the spectrum. It is the plasmon excitation of metal clusters or nanoparticles in glass that is responsible for the strong colors seen, for example, in medieval stained glass windows. The optical absorption of metal clusters is compared to that of atoms and the bulk metal in **Fig. 3**. *See* GIANT NUCLEAR RESONANCES; PLASMON.

Fullerenes. Fullerenes are clusters of carbon atoms, C_{2N}, with N greater than 11, which have unique hollow structures. They were discovered with a laser vaporization source combined with an adiabatic gas expansion. It is now possible to produce and isolate macroscopic amounts of some of these clusters, in particular C_{60} and C_{70}. The cluster C_{60} is especially interesting because of its highly symmetrical truncated icosahedral geometry (just like a soccer ball) and a range of fascinating properties, such as a quasi-three-dimensional aromatic chemistry and superconductivity when doped with alkali atoms. The availability of macroscopic amounts of a mass-selected, neutral atomic cluster with

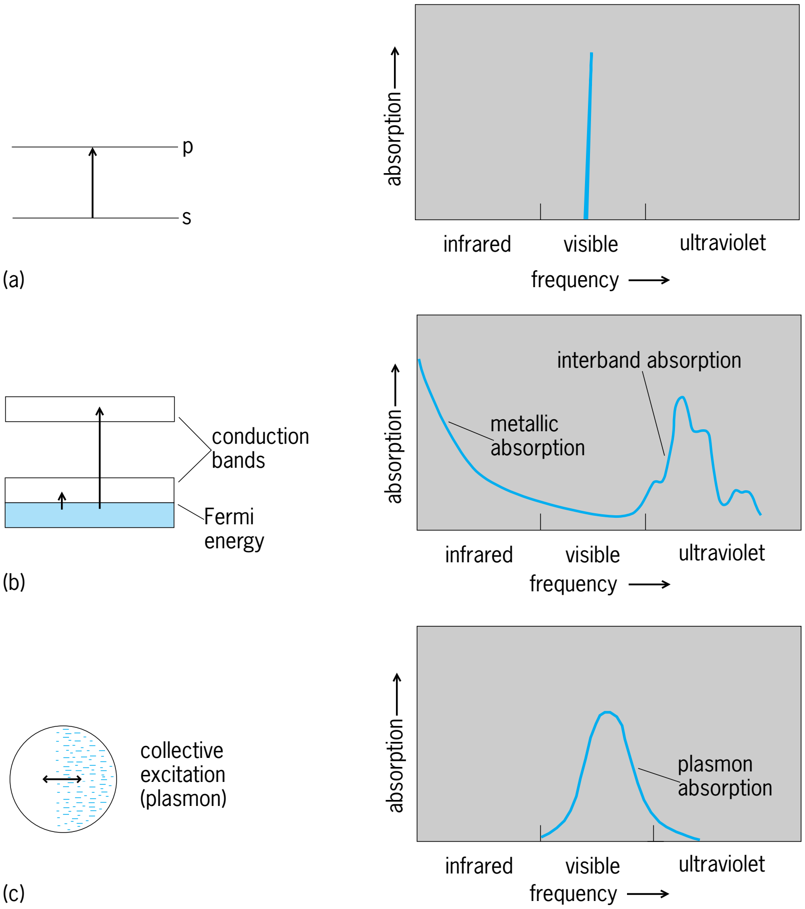


Fig. 3. Absorption of light by an alkali metal in various states of aggregation. Absorption spectra are shown on the right and schematic structures on the left. (*a*) Metal atom. (*b*) Metal crystal. (*c*) Metal cluster.

well-defined geometrical and electronic properties has opened up a new range of experiments and theoretical treatments to probe the dynamics of a relatively simple atomic cluster system with a large but finite number of degrees of freedom. *See* FULLERENE.

Probing with laser pulses. A recent development in the study of atomic clusters is the use of laser pulses with durations of less than 100 femtoseconds (10^{-13} s) to probe the time scale for energy coupling between the different degrees of freedom in the cluster in real time. This work, which is presently in its infancy, will provide very detailed information on the dynamics of such complex systems and will certainly provide a stringent test of theoretical models. *See* OPTICAL PULSES.

Clusters on surfaces. The overall emphasis in the field is gradually shifting from gas-phase studies to the study of clusters deposited on surfaces. The aim is to produce new materials with tailor-made electrical, optical, or catalytic properties. Cluster-beam epitaxy, where large metallic clusters are deposited at high kinetic energy (greater than 10 eV per atom) onto substrates, is already producing superior-quality mirror coatings for use in high-power laser applications. Eleanor Campbell

Bibliography. M. S. Dresselhaus, G. Dresselhaus, and P. C. Eklund, *Science of Fullerenes and Carbon Nanotubes*, 1996; H. Haberland (ed.), *Clusters of Atoms and Molecules*, 2 vols., 1994; U. Kreibig and M. Vollmer, *Optical Properties of Metal Clusters*, 1995.

Atom economy

The maximum incorporation of the starting materials into the final product for any chemical reaction.

A broad imperative for synthetic chemists today is the need to develop innovative and sophisticated synthetic methods to keep pace with the continually changing variety of molecular targets. In 1991, Barry M. Trost presented a set of coherent guiding principles to evaluate the efficiency of a specific chemical process, which has subsequently altered the way many chemists design and plan their syntheses. He proposed that within synthetic efficiency there are two broad categories: selectivity and atom economy. Selectivity is defined as chemo-, regio-, and stereoselectivity, while atom economy seeks to maximize the incorporation of the starting materials into the final product. An additional corollary is that if maximum incorporation cannot be achieved, then ideally the side-product quantities should be minute and environmentally innocuous.

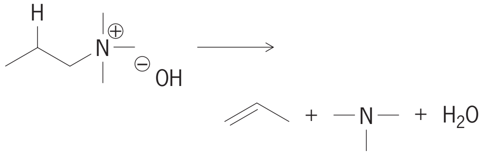
In the past, attaining the highest yield and product selectivity were the governing factors in chemical synthesis. As a result, multiple reagents often were used in stoichiometric quantities that were not incorporated into the target molecule, resulting in significant side products and waste. Today, the underlying standard is synthetic efficiency—the ability to maximize the incorporation of the starting materials into the final product and to minimize by-products.

The reaction yield and the atom-economy yield are calculated by different means. The reaction yield [Eq. (1)] is concerned only with the quantity of desired product isolated, relative to the theoretical quantity of the product, and does not provide information in terms of the efficiency of the transfer of molecular weight from the reactants into the desired product, which is the concern of atom economy. Atom economy takes into account all the reagents used and the unwanted side products, along with the desired product [Eq. (2)].

$$\text{Reaction yield} = \frac{\text{quantity of product isolated}}{\text{theoretical quantity of product}} \times 100\% \qquad (1)$$

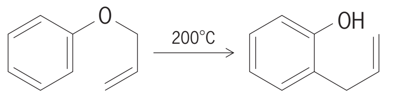
$$\text{Atom economy} = \frac{\text{molecular weight of products}}{\text{molecular weight of all products}} \times 100\% \qquad (2)$$

Atom-uneconomic reactions. Examples of atom-uneconomic reactions are the Wittig reaction (formation of triphenylphosphine oxide), Grignard reaction, and substitution and elimination reactions (3). *See* GRIGNARD REACTION.



(3)

Atomic-economic reactions. Some atomic-economic reactions are the Diels-Alder reaction, rearrangements [reaction (4)], and concerted reactions. Not all rearrangements are equal in that some require large amounts of acids to push the reaction forward (for example, the Beckmann rearrangement), and so those acids are considered to be ancillary reagents and factor into the economy of the reaction. *See* DIELS-ALDER REACTION.



(4)

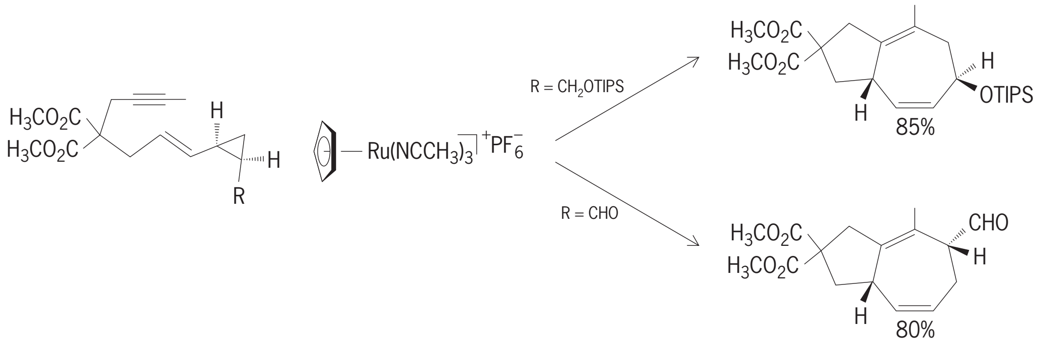
Currently, many chemists are working on increasing the number of reactions in the "toolbox" by developing ideal and innovative methodologies that combine both selectivity and atom economy. Transition metals have proven to be excellent catalysts for the stereoselective transformation of many organic molecules, and since they are used catalytically, they are ideal for the development of atomic-economic methodologies. Reaction scheme (5) is an example of an transition-metal-catalyzed, atom-economic reaction.

The ruthenium catalyst complex is easily accessible, and its generality makes it the catalyst of choice. With this catalyst, the reactions are usually complete in 30 minutes at ambient temperature, and based upon the substituent there is regioselective control. In addition, this reaction is 100% atom-economic. *See* ASYMMETRIC SYNTHESIS; CATALYSIS; STEREOCHEMISTRY.

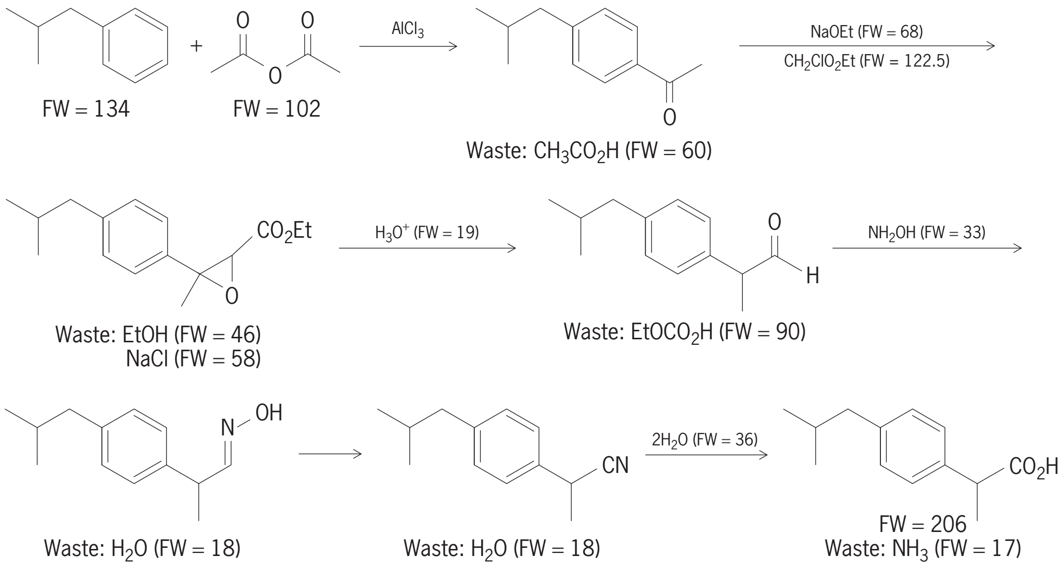
Synthesis of ibuprofen. A practical application of atom economy in industry is the synthesis of ibuprofen. The traditional synthesis of ibuprofen was developed in the 1960s. It was a six-step synthesis that used stoichiometric amounts of reagents and generated large quantities of waste by-products that needed further treatment [reaction scheme (6)]. There are numerous waste by-products, and the overall atom economy of the synthesis is 40%, or conversely, there is an overall waste of 60%.

Reaction scheme (7) shows a more atom-economic synthesis of ibuprofen, having an atom economy percentage of 77% (or 99% if the acetic acid generated in the first step is recovered).

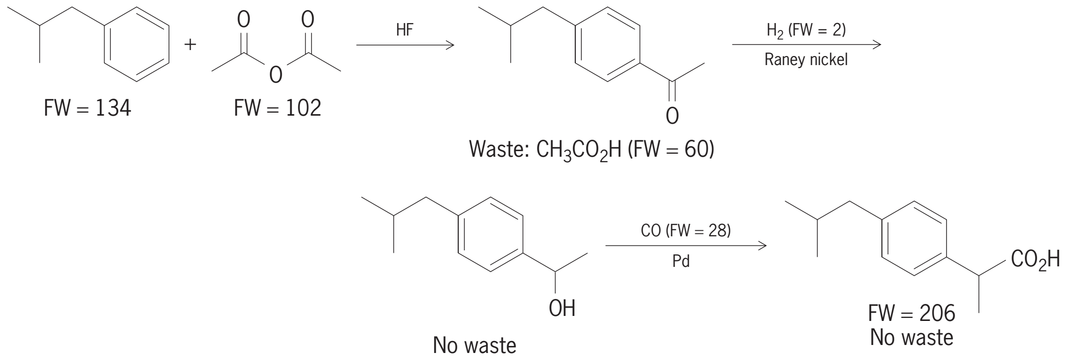
For both reaction schemes (6) and (7), the starting materials and products are the same for the first step, with the only difference being the catalyst. In reaction scheme (6), $AlCl_3$ is not truly a catalyst but an auxiliary reagent that is needed in stoichiometric


H3CO2C
H3CO2C
H
H
R
Ru(NCCH3)3 +PF6−
R = CH2OTIPS
R = CHO
H
OTIPS
85%
CHO
H
80%

(5)


FW = 134
O
O
FW = 102
AlCl3
NaOEt (FW = 68)
CH2ClO2Et (FW = 122.5)
Waste: CH3CO2H (FW = 60)
CO2Et
H3O+ (FW = 19)
O
H
NH2OH (FW = 33)
Waste: EtOH (FW = 46)
NaCl (FW = 58)
Waste: EtOCO2H (FW = 90)
N
OH
CN
2H2O (FW = 36)
CO2H
FW = 206
Waste: H2O (FW = 18)
Waste: H2O (FW = 18)
Waste: NH3 (FW = 17)

(6)



(7)

amounts and produces aluminum trichloride hydrate as a waste by-product. Hydrogen fluoride used in reaction shceme 7 is a catalyst, which is recovered and repeatedly reused. In addition, the Raney nickel and palladium catalysts are also recovered and reused. As a result, almost no waste is generated.

Since reaction scheme (7) requires three steps [relative to six for reaction scheme (6)], there is the additional elimination of waste by not having three extra steps. This results in the ability to produce larger quantities of ibuprofen in less time and with less capital expenditure. Therefore, besides having an atom-economic and environmentally friendly green synthesis, the product has a larger profit capability. *See* ORGANIC SYNTHESIS.

Charlene C. K. Keh; Chao-Jun Li

Bibliography. P. T. Anastas and J. C. Warner, *Green Chemistry: Theory and Practice*; Oxford, New York, 1998; B. M. Trost, The atom economy—a search for synthetic efficiency, *Science*, 254:1471–1477, 1991; B. M. Trost, On inventing reactions for atom economy, *Acc. Chem. Res.*, 35:695–705, 2002.

Atom laser

An atom laser is a device that generates an intense coherent beam of atoms through a stimulated process. It does for atoms what an optical laser does for light. The atom laser emits coherent matter waves, whereas the optical laser emits coherent electromagnetic waves. Coherence means, for instance, that atom laser beams can interfere with each other. *See* COHERENCE.

Properties. Laser light is created by stimulated emission of photons, a light amplification process. Similarly, an atom laser beam is created by stimulated amplification of matter waves. The conservation of the number of atoms is not in conflict with matter-wave amplification: The atom laser takes atoms out of a reservoir and transforms them into a coherent matter wave similar to the optical laser, which converts energy into coherent electromagnetic radiation (but, in contrast, the number of photons need not be conserved). *See* LASER.

The condition of high intensity means that there are many particles per mode or quantum state. A thermal atomic beam has a population per mode of only 10^{-12} compared to much greater than 1 for an atom laser. The realization of an atom laser therefore required the development of methods to greatly enhance the mode occupation. This enhancement is accomplished by cooling the reservoir of atoms to microkelvin temperatures or below. *See* LASER COOLING.

In the case of an ideal atom laser, the output beam should be monochromatic and directional, and have a well-defined phase and intensity. For atoms, being monochromatic means that their velocity spread is extremely small. Such beams propagate with minimum spreading, and can be focused by atom lenses to a small spot size. The minimum spreading and the minimum spot size are limited by Heisenberg's uncertainty relation in the same way as the propagation of a single-mode optical laser beam is diffraction limited. The analogy between light and matter waves is exploited in the field of atom optics. *See* ATOM OPTICS.

The different nature of atoms and photons implies different properties of light and atom beams. Unlike light, an atomic beam cannot travel far through air. It scatters off air molecules in less than a micrometer. Vacuum is thus required for all atom laser experiments. Also, slow atoms are strongly affected by gravity. Furthermore, a dense atom beam will show spreading in excess of the Heisenberg uncertainty limit because of the interactions between the atoms.

Elements. A laser requires a cavity (resonator), an active medium, and an output coupler (see **table**).

Cavity. Various analogs of laser cavities for atoms have been realized. The most important ones are

Analogies between an atom laser and the optical laser

Atom laser*	Optical laser
Atoms	Photons
Matter waves	Electromagnetic waves
Atom trap	Laser cavity
Atoms in the Bose condensate	Photons in the lasing mode
Thermal atoms	Gain medium
Evaporative cooling	Excitation of the gain medium
Stimulated scattering of atoms	Stimulated emission of photons
Critical temperature for Bose-Einstein condensation	Laser threshold

*Based on evaporative cooling.

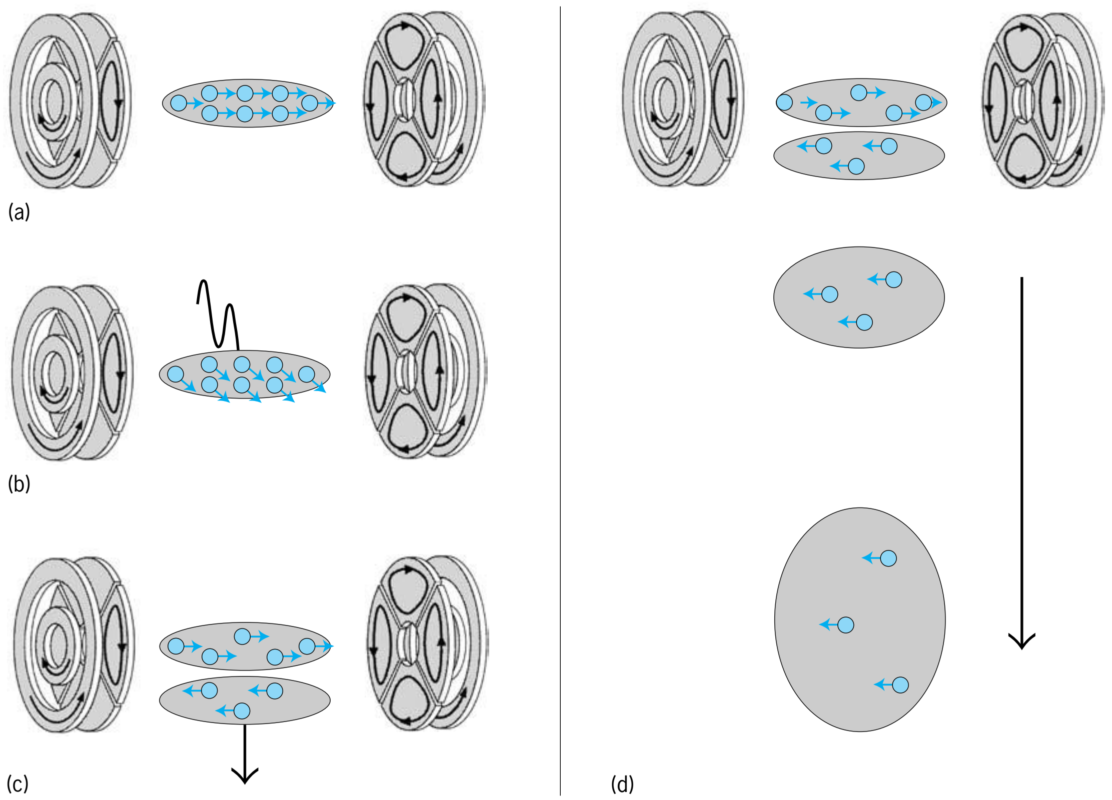


Fig. 1. Principle of the radio-frequency output coupler of an atom laser. (*a*) Bose-Einstein condensate confined in an atom trap. (*b*) Tilting of the spins of the atoms by a short radio-frequency pulse. (*c*) Splitting of the cloud into a trapped cloud and an out-coupled cloud. (*d*) Repetition of steps *a* and *c* creates a sequence of output pulses.

magnetic traps (which use the force of an inhomogeneous magnetic field on the atomic magnetic dipole moment) and optical dipole traps (which use the force exerted on atoms by focused laser beams). Confinement of atoms between two atom mirrors has been suggested and is analogous to a Fabry-Perot cavity for light. Even a single-mirror cavity is possible, where atoms perform multiple bounces off a mirror in the vertical direction and return because of gravity (an atomic "trampoline"). *See* INTERFEROMETRY; PARTICLE TRAP.

Active medium. The active medium is a reservoir of atoms which are transferred to one state of the confining potential, which is the analog of the lasing mode. The reservoir can be atoms confined in other quantum states of the atom cavity or an ultraslow atomic beam. The atoms are transferred to the lasing mode either by collisions or by optical pumping. The transfer of atoms is efficient only for an ultracold sample, which is prepared by laser cooling or evaporative cooling. This cooling ensures that the atoms in the reservoir occupy only a certain range of quantum states which can be efficiently coupled to the lasing mode.

Output coupler. The output coupler extracts atoms from the cavity, thus generating a pulsed or continuous beam of coherent atoms. A simple way to accomplish this step is to switch off the atom trap and release the atoms. This method is analogous to cavity dumping for an optical laser, and extracts all the stored atoms into a single pulse. A more controlled way to extract the atoms requires a coupling mechanism between confined quantum states and propagating modes.

Such a beam splitter for atoms can be realized by applying the Stern-Gerlach effect to atoms in a magnetic trap (**Fig. 1**). Initially, all the atoms have their electron spin parallel to the magnetic field, say spin up (Fig. 1*a*), and in this state they are confined in the trap. A short radio-frequency pulse rotates (tilts) the spin of the atoms by a variable angle (Fig. 1*b*). Quantum-mechanically, a tilted spin is a superposition of spin up and spin down. Since the spin-down component experiences a repulsive magnetic force, the cloud of atoms is split into a trapped cloud and an out-coupled cloud (Fig. 1*c*). By using a series of radio-frequency pulses, a sequence of coherent atom pulses can be formed (Fig. 1*d*). These pulses are accelerated downward by gravity and spread out. *See* QUANTUM MECHANICS.

Figure 2 shows such a sequence of coherent pulses. In this case, sodium atoms are coupled out from a magnetic trap by radio-frequency pulses every 5 ms. The atom pulses are observed by illuminating them with resonant laser light and imaging their shadows, which are caused by absorption of the light. Each pulse contains 10^5–10^6 sodium atoms.

Other output coupling schemes have been suggested, including optical transitions which eject atoms from the cavity because of the recoil of the absorbed photon, and tunneling through thin barriers of light.

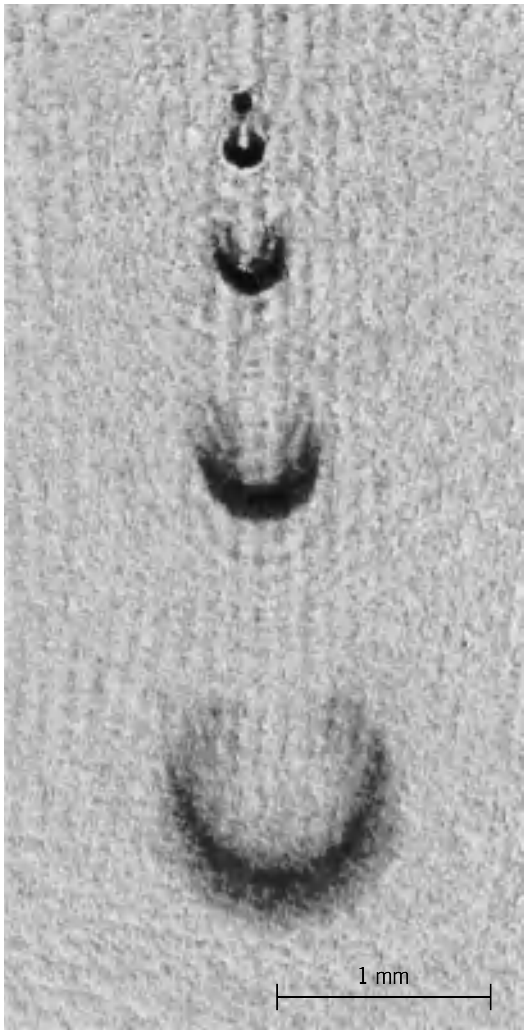


Fig. 2. Pulsed atom laser in operation, with pulses of coherent sodium atoms coupled out from a Bose-Einstein condensate that is confined in a magnetic trap.

Gain process. An atom laser is possible only for bosonic atoms. The accumulation of atoms in a single quantum state is based on Bose-Einstein statistics. Two different mechanisms have been discussed which may provide gain in an atom laser: elastic collisions and spontaneous emission of photons. *See* BOSE-EINSTEIN STATISTICS.

The case of elastic collisions is closely related to Bose-Einstein condensation. When a gas of bosonic particles is cooled down, it forms a Bose-Einstein condensate characterized by a macroscopic occupation of the ground state of the system. This process happens suddenly at the Bose-Einstein condensation transition temperature. The atoms in a Bose-Einstein condensate are coherent to first and higher order. An atom laser based on Bose-Einstein condensation operates in thermal equilibrium. Atom lasing is achieved simply by cooling down the gas. Below a certain temperature, nature maximizes entropy by generating a Bose-Einstein condensate. For photons, the situation is very different: Cooling a black-body cavity reduces its energy density in proportion to the fourth power of its absolute temperature (Stefan-Boltzmann law). Thus, at very low temperatures the cavity is empty. This is how entropy is maximized when the number of particles is not conserved. *See* BOSE-EINSTEIN STATISTICS.

It is instructive to look more closely at the stimulated amplification process that takes place when a Bose-Einstein condensate forms. In a normal gas, atoms scatter among a myriad of possible quantum states. But when the critical temperature for Bose-Einstein condensation is reached, they scatter predominantly into the lowest energy state of the system. This abrupt process is closely analogous to the threshold for operating an optical laser. The presence of a Bose-Einstein condensate causes stimulated scattering into the ground state. More precisely, the presence of a condensate with N_0 atoms enhances the probability that an atom will be scattered into the condensate by a factor of $N_0 + 1$.

In an atom laser, the analog to excitation of the active medium can be accomplished by evaporative cooling. The evaporation process creates a cloud which is not in thermal equilibrium and relaxes toward lower temperatures. This results in growth of the condensate. After equilibration, the gain process halts and the condensate fraction remains constant until further cooling is applied. In thermal equilibrium, there is still stimulated scattering of atoms into the condensate. However, this process is in dynamic equilibrium with collisions that knock atoms out of the condensate.

An atom laser was realized by extracting a beam of atoms from a Bose-Einstein condensate and explicitly demonstrating its coherence. The proof of the coherence was obtained by observing a high-contrast interference pattern when two Bose-Einstein condensates overlapped. The two condensates were created by cooling a gas of sodium atoms in a double-well potential. After the condensates were released from the trap, they accelerated downward, spread out ballistically, and eventually overlapped and interfered. The interference pattern could be directly photographed. It had a period of 15 micrometers, a gigantic length for matter waves. (Room-temperature atoms have a matter wavelength of 0.05 nm, 300,000 times smaller.) *See* INTERFERENCE OF WAVES.

An atom laser based on Bose-Einstein condensation is a special case of macroscopic occupation of a quantum state. In this case, the atoms accumulate in the ground state and are in thermal equilibrium. More generally, atom lasers can operate in higher-order modes and also as a driven system which is not in thermal equilibrium. (This is the situation in an optical laser.) The lasing mode is distinguished by preferential population of atoms or minimum loss. It has been suggested that this condition can be realized by optical pumping. In this case, atoms in the reservoir are optically excited. When they decay by spontaneous emission, they can reach final momentum states that differ from the initial momentum by the photon recoil. If one state within this range has a macroscopic population, then the rate of spontaneous emission into this final state is enhanced, and there is an amplification process similar to the one described above for elastic collisions. The case of optically excited atoms shows very clearly the symmetry between the optical laser and the atom laser: The rate of emission to a final state, which is specified by a particular state of the atom inside the cavity and a particular mode of the photon field, is proportional to $(N + 1)(n + 1)$, where N is the number of atoms in this level of the cavity, and n is the number of photons in this mode. The first factor in this expression is the bosonic stimulation by atoms that is responsible for the amplification process in the atom laser, and the second describes the amplification process in the optical laser.

Potential applications. Although a basic atom laser has now been demonstrated, major improvements are necessary before it can be used for applications, especially in terms of increased output power and reduced overall complexity. The atom laser provides ultimate control over the position and motion of atoms at the quantum level, and might find use where such precise control is necessary, for example, for precision measurements of fundamental constants, tests of fundamental symmetries, atom optics (in particular, atom interferometry and atom holography), and precise deposition of atoms on surfaces. Since the matter wavelength of atoms can be extremely short (it decreases in inverse proportion to the atomic velocity), the ultimate limit to the spatial resolution is not the matter wavelength but the size of the atom. *See* FUNDAMENTAL CONSTANTS; NANOTECHNOLOGY; SYMMETRY LAWS (PHYSICS).

Wolfgang Ketterle

Bibliography. M. R. Andrews et al., Observation of interference between two Bose condensates, *Science*, 275:637–641, 1997; B. Goss Levi, Bose condensates are coherent inside and outside an atom trap, *Phys. Today*, 50(3):17–18, March 1997; D. Kleppner, A beginner's guide to the atom laser, *Phys. Today*, 50(8):11–13, August 1997; H.-J. Miesner et al., Bosonic stimulation in the formation of Bose-Einstein condensate, *Science*, 279:1005–1007, 1998; G. Taubes, First atom laser shoots pulses of coherent matter, *Science*, 275:617–618, 1997.

Atom optics

The use of laser light and nanofabricated structures to manipulate the motion of atoms in the same manner that rudimentary optical elements control light. The term refers to both an outlook in which atoms in atomic beams are thought of and manipulated like photons in light beams, and a collection of demonstrated techniques for doing such manipulation (see **table**). Two types of atom optics elements have existed for some time: slits and holes used to collimate molecular beams (the analog of the pinhole camera), and focusing lenses for atoms and molecules (for example, hexapole magnets and quadrupole electrostatic lenses). However, in the 1980s the collection of optical elements for atoms expanded dramatically because of the use of near-resonant laser light and fabricated structures to make several types of mirrors as well as diffraction gratings. The diffraction gratings are particularly interesting because they exploit and demonstrate the (de Broglie) wave nature of atoms in a clear fashion. *See* LASER.

Diffraction gratings. Diffraction gratings for atoms have been made by using either a standing wave of light or a slotted membrane. The standing light wave makes a phase grating (that is, it advances or retards alternate sections of the incident wavefront but does not absorb any of the atom wave), so that the transmitted intensity is high. This approach requires the complexity of a single-mode laser, and introduces the complication that the light acts differently on the various hyperfine states of the atom. The slotted membrane, however, absorbs (or backscatters) atoms which strike the grating bars but does not significantly alter the phase of the transmitted atoms; it is therefore an amplitude grating. It works for any atom or molecule, regardless of internal quantum state, but with total transmission limited to about 40% by the opacity of the grating bars and requisite support structure.

The effect of a slotted membrane type of diffraction grating on atoms and molecules is demonstrated by an experiment in which a beam composed of both sodium atoms and a smaller number of sodium dimers (Na_2), moving at the same speed, is diffracted from a nanofabricated transmission grating (**Fig. 1**). Because the molecules have twice the mass of the atoms and are traveling at the same speed, they have twice the momentum, p. The de Broglie wave that describes their motion has a wavelength which is equal to Planck's constant divided by their momentum, and is therefore one-half of the atoms' de Broglie wavelength. As a result, the spacing of the peaks in the diffraction pattern is half as large. The diffraction pattern therefore has small peaks, attributed to the less populous molecules, in between adjacent larger ones.

If the atoms are removed from the beam by resonant laser light, leaving only the molecules, the smaller peaks are, indeed, unaffected, as is expected if they arise only from molecules in the combined beam. A similar diffraction experiment with a supersonic helium beam provided definitive evidence for the existence of the helium dimer, the most weakly bound molecule known. *See* INTERMOLECULAR FORCES.

The above diffraction pattern may be regarded as a demonstration that atoms and molecules behave

Comparison of optical elements for photons and atoms

Photons	Atoms
Lens	Electromagnetic fields (static, light) Zone plate
Mirror	Crystal surface Liquid helium Evanescent light
Phase grating	Standing light wave
Amplitude grating	Nanofabricated bars
Beam splitter	Only diffraction gratings (no achromatic elements)

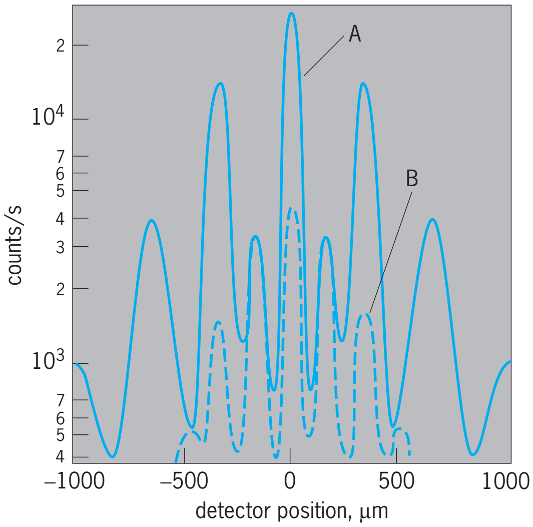


Fig. 1. Diffraction patterns from a transmission grating [period = 100 nm] (curve A) of a beam composed of both sodium atoms (Na) and sodium dimers (Na_2), and (curve B) of the same beam after the atoms have been removed with a deflecting laser beam, leaving only Na_2 molecules.

like waves. This demonstration differs fundamentally from similar demonstrations with neutrons and electrons in that atoms and molecules are composite particles; a sodium molecule is about 0.6 nanometer in diameter. In the experiment described above, both the de Broglie wavelength, 0.016 nm, and the coherence length, 0.12 nm, of these waves (the distance over which they remain in phase) were smaller than the size of the object which they describe. *See* DIFFRACTION GRATING.

In a diffraction grating, the constant spacing of the bars produces a constant angle of diffraction so that all the diffracted waves travel in parallel away from the grating. If the spacing is systematically decreased away from the center of the grating, the angle of diffraction is greater for rays that strike the grating farther from the center, and these rays intersect the undeflected rays, which pass through the center some distance downstream of the grating. For the Fresnel zone plate, in which the grating spacing varies inversely with distance from the center, all rays intersect at the same downstream point; that is, they converge to a focus. Such devices have been demonstrated for atoms in both a spherical variety that focuses initially parallel atoms to a point and a cylindrical type that focuses parallel atoms to a line. *See* DIFFRACTION.

Atom interferometers. Atom interferometers have been demonstrated through several different experimental routes, involving both microscopic fabricated structures and laser beams. These interferometers are the first examples of optical systems composed of the elements of atom optics like those discussed above. Atom interferometers, like optical interferometers, are well suited for application to a wide range of fundamental and applied scientific problems. Scientific experiments with atom interferometers divide naturally into three major categories: measurements of atomic and molecular properties, fundamental tests and demonstrations, and inertial effects.

Atomic polarizability measurements. An atomic property that has been precisely determined with an atom interferometer is the polarizability of sodium, a quantity that parametrizes its interaction with electric fields, low-frequency radiation, and nearby atoms. In the interferometer used to measure this, a diffraction grating was used to split the atom wave into two components whose centers were separated by 55 micrometers before a second grating diffracted them back toward each other so they recombined to form interference fringes at the surface of a third grating (**Fig. 2**). This approach allowed the insertion of a stretched metal foil between the atom waves in the two sides of the interferometer, so that a uniform electric could be applied to the atom wave on one side of the interferometer.

Because an atom is polarized by an electric field, this field acts as a slight depression in the potential energy, increasing the momentum of the atom as it passes through the electric field. The result is a shorter de Broglie wavelength and therefore an increased phase accumulation relative to the wave that passes on the side of the metal foil with no field. The measured phase shift increased quadratically with the applied field, permitting a precision determination of the polarizability of the ground state of sodium (**Fig. 3**). *See* DIPOLE MOMENT.

This experiment also illustrates the paradox that results from an attempt to imagine that the atom has a classical position even when no measurement is made. The polarizability depends on the size of the atom, and the phase shift is proportional to the product of the polarizability and the time spent in the field. The observed phase shift of the fringes is consistent with the entire atom being present on the side with the electric field for the entire time of passage through the interaction region. However, the observation of the interference fringes is possible only if the atom wave splits roughly evenly between the two sides of the interferometer; indeed, the fringes

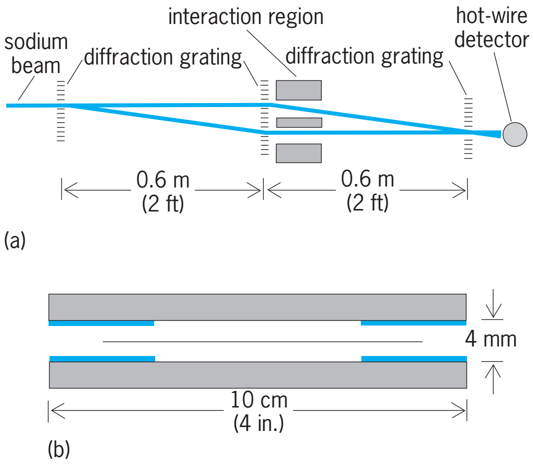


Fig. 2. Separated-beam atom interferometer. (*a*) Diagram of the apparatus. Diffraction gratings have a period of 200 nm. (*b*) Detail of the interaction region.

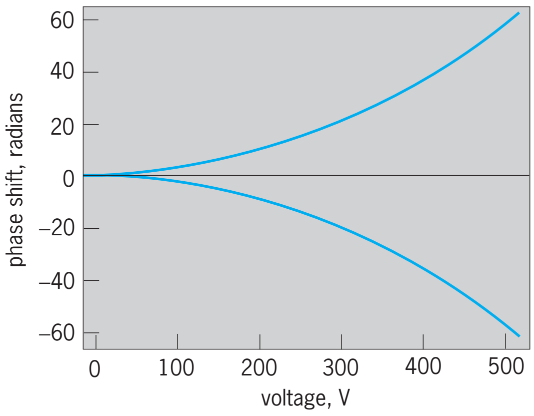


Fig. 3. Phase shift of the interference pattern of a separated-beam atom interferometer as a function of voltage applied to the left or right side of the interaction region. In either case, the magnitude of the phase shift shows a quadratic increase with voltage.

will disappear if a barrier prevents the atoms from traveling on either side of the interferometer. The quantum-mechanical description of the above situation is to add together the wave amplitudes of both possibilities (that is, the atom goes on one side and the atom goes on the other side) to predict the atom intensity observed at the detector. Along with this result comes the impossibility of localizing the atom, even in the imagination, without an intervening measurement to determine which side it traveled on.

Probing atomic interactions. Experiments have been conducted in which the atom wave on one side of the metal foil passes through a gas not present on the other side. The most useful perspective is that of conventional optics: The passage of a wave through a medium is described in terms of an index of refraction whose real part is proportional to the phase shift of the wave and whose imaginary part is proportional to the absorption of the wave. These factors shift the position and decrease the amplitude of the atom interference pattern, respectively. For sodium waves passing through helium and neon, the phase shift is found to be a much more sensitive probe of the interaction between sodium and the target gas than is the absorption. *See* ABSORPTION; REFRACTION OF WAVES; SCATTERING OF ELECTROMAGNETIC RADIATION.

Measurement of inertial effects. The sensitivity of atom interferometers to inertial effects results from the fact that there is no force on the atoms between the gratings, so that they travel along straight lines if viewed from an inertial (unaccelerating) coordinate system with respect to which the interferometer accelerates or rotates. When the atoms recombine to form interference fringes, they appear displaced because of the movement of the interferometer during the atoms' transit time. Since the de Broglie wavelength of the atoms is very small, atom interferometers are extremely sensitive to accelerations and rotations. The sensitivity of atom interferometers to rotation, which was demonstrated in the first interferometer to use light gratings, is of the order of 10^{10} greater than the phase shift in a light interferometer with the same geometry (a factor of $\sim 10^5$ from the longer transit time, and $\sim 10^5$ from the shortness of the de Broglie waves relative to light waves). Atom interferometer sensitivity to gravitation has been demonstrated at a level of 3 parts in 10^8, within an order of magnitude of the accuracy of the best absolute gravimeters. *See* ATOMIC STRUCTURE AND SPECTRA; FRAME OF REFERENCE; GRAVITY METER; INTERFERENCE OF WAVES; INTERFEROMETRY; MOLECULAR BEAMS; OPTICS; QUANTUM MECHANICS.

David E. Pritchard

Bibliography. C. S. Adams, M. Siegel, and J. Mlynek, Atom optics, *Phys. Rep.*, 240:143–210, 1994; F. Flam, Making waves with interfering atoms, *Science*, 252:921-922, 1991; R. Pool, Catching the atom wave, *Science*, 268:1129–1130, 1995; Special issue on atom optics and atom interferometry, *Appl. Phys. B*, vol. 54, 1992.

Atomic beams

Unidirectional streams of neutral atoms passing through a vacuum. These atoms are virtually free from the influence of neighboring atoms but may be subjected to electric and magnetic fields so that their properties may be studied. The technique of atomic beams is identical to that of molecular beams. For historical reasons the latter term is most generally used to describe the method as applied to either atoms or molecules.

The method of atomic beams yields extremely accurate spectroscopic data about the energy levels of atoms, and hence detailed information about the interaction of electrons in the atom with each other and with the atomic nucleus, as well as information about the interaction of all components of the atom with external fields. *See* MOLECULAR BEAMS.

Polykarp Kusch

Atomic bomb

A device for suddenly producing an explosive neutron chain reaction in a fissile material such as uranium-235 (^{235}U) or plutonium-239 (^{239}Pu). In a wider sense, any explosive device that derives its energy from nuclear reactions, including not only the foregoing fission weapon but also a fusion weapon, which gets its energy largely from fusion reactions of heavy hydrogen isotopes, and a fission-fusion weapon, which derives its energy from both fission and fusion. Because an atomic bomb derives its energy from nuclear reactions, it is properly called a nuclear explosive or nuclear weapon. *See* HYDROGEN BOMB; NUCLEAR FISSION; NUCLEAR FUSION; NUCLEAR REACTION; PLUTONIUM; URANIUM.

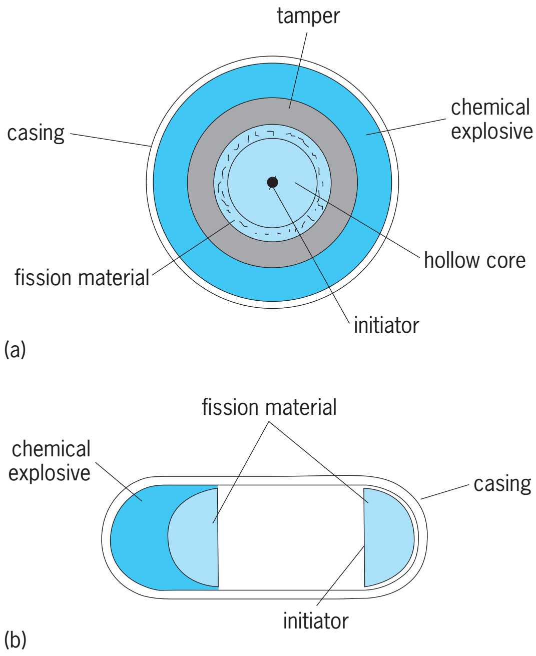
Materials. Of the two principal fissile materials, the cheaper but less potent ^{235}U is present in natural uranium usually in the proportion of 1 part to 139 parts of ^{238}U and is separated from it by various enrichment processes. Weapons-grade plutonium is manufactured from ^{238}U in a special military production reactor that has enough excess neutrons for the reaction below.

$$^{238}\text{U} + n \rightarrow {}^{239}\text{U (23-min half-life)} \rightarrow {}^{239}\text{Np (2.3-day half-life)} \rightarrow {}^{239}\text{Pu}$$

See ISOTOPE SEPARATION; NUCLEAR REACTOR.

Weapon cores are made of very high fractions of fissile materials: highly enriched 93% uranium-235 or weapon-grade 94% plutonium-239. In most power reactors, fuel is only 3–4% uranium-235, and the plutonium-239 by-product is less than 80% isotopically pure. Although low-quality fissile materials, such as reactor-grade plutonium, can be used by expert weapon designers to make a fission explosive, such devices have not been placed in the arsenals of states known or suspected of making nuclear weapons.

Conceptual design. A fission bomb before ignition consists of a mass of fissile material and surrounding tamper—beryllium oxide or other reflector of neutrons intended ultimately to improve the neutron multiplication factor k—arranged in a geometry so favorable to neutron leakage that k is less than 1. These materials are suddenly compressed into a geometry where k substantially exceeds 1. This is done with chemical explosives that either implode a spherical subcritical mass of fission material (**illus.** *a*) or else drive two subcritical sections together in a gun-barrel type of arrangement (illus. *b*). At the same time, enough neutrons are artificially introduced to start an explosively divergent (expanding) chain reaction. *See* CHAIN REACTION (PHYSICS).



Conceptual diagrams illustrating the main features of (*a*) implosion and (*b*) gun-barrel type of fission weapons. (*After A. De Volpi et al., Born Secret: The H-Bomb, the "Progressive" Case and National Security, Pergamon, 1981*)

Fission-explosive devices intended for military application are highly sophisticated combinations of pure materials, precise design, and reliable electronics. The high-explosive lenses are shaped charges that generate and focus a compressive implosion wave. The fissile material or composite shell is levitated in a hollow core in order to allow the blast wave to impart maximum momentum to the imploding tamper-reflector. A chain-reaction initiator introduces at the optimal moment a burst of external neutrons. Either a forcibly mixed radioactive neutron source or a more useful miniature electronic neutron generator can be used for neutron injection. The latter allows controlled variability of explosive yields.

More neutrons are released in fission than are lost because of absorption and leakage, which means that a chain reaction can be maintained. With the right arrangement, the extra neutrons expelled will cause fissions in nearby fissile atoms, and the rate of fissioning grows exponentially. The rate of reaction ultimately tapers off. In a sufficiently efficient device, the reaction subsides as fission depletes the fuel (a total mass of 2.2 lb or 1 kg of heavy nuclei undergoes fission for each 17 kilotons or 7.1×10^{13} joules of energy release); but in a less efficient device, k is reduced below unity primarily through the explosive disassembly of the compact geometry. Temperatures are attained of the astrophysical order of 10^8 K, at which electrons have energies of the order 10 keV, and the energy peak of the blackbody (heat) radiation compares with that of the electronic kinetic energy. *See* HEAT RADIATION.

For optimum yield, the injection of neutrons must occur at precisely the right time as the fissile material approaches its critical configuration, within 10^{-3} s or less, depending on the type of weapon and its fissile material. In particular, if there are too many neutrons around too soon, the bomb may be triggered prematurely (predetonate) and "fizzle" (yield only a fraction of its potential). The implosion technique was developed to counter the predetonation tendency of plutonium, which has a high inherent neutron background. The poorer the quality of the plutonium, the worse the predetonation problem, and the more painstaking the weapon design becomes.

Yield. The explosive energy (yield) of a nuclear weapon is usually expressed in kilotons or megatons. A kiloton is the amount of energy liberated in the explosion of 1000 tons of TNT (10^{12} calories or 4.18×10^{12} J), and a megaton is a thousand times as large. The fission bombs that destroyed Hiroshima (gun-barrel type) and Nagasaki (implosion type) had estimated explosive yields of 13 and 22 kilotons, respectively. Fractional kiloton yields can be obtained (tactical nuclear weapons). Fission weapons have been tested up to approximately 500 kilotons, overlapping the yield of multistage fusion explosives (strategic nuclear weapons). Fission explosives are central to all nuclear warheads, including thermonuclear weapons, and some are known to be small enough to fit within a 127-mm (5-in.) artillery shell.

Acquisition of materials and components remain the major obstacle for state-supported production of military devices. Less knowledgeable individuals could conceivably assemble crude devices by using diverted materials, but highly lethal fission explosives, with yields of 1 kiloton or so, are extremely impractical with lower- (reactor-) grade materials. However, serious hazards could be generated through high-explosive dispersal of the radioactive materials. Meticulous control and safeguarding of separated fissile materials is essential, a process which is aided by the self-indicating radioactive nature of fissile materials.

Explosion effects. The nuclear explosive energy is communicated by mechanical shock and radiative transport to the surrounding water, earth, or air, ionizing it out to a radius which, in the case of explosions in air, is known as the fireball radius (150 yd or 140 m in about 1 s after a 20-kiloton nuclear explosion). Energy goes out from such a fireball into the surrounding relatively transparent air, in not very different orders of magnitude in the form of a shock wave and in the form of heat radiation that may continue for a number of seconds. *See* NUCLEAR EXPLOSION; RADIOACTIVE FALLOUT.

A. De Volpi

Bibliography. A. C. Brown and C. B. McDonald, *The Secret History of the Atomic Bomb*, 1977; A. DeVolpi, Fissile materials and nuclear weapons proliferation, *Annu. Rev. Nuc. Particle Sci.*, 36:83–114, 1986.

Atomic clock

A device that uses an internal resonance frequency of atoms (or molecules) to measure the passage of time. The terms atomic clock and atomic frequency standard are often used interchangeably. A frequency standard generates pulses at regular intervals. A frequency standard can be made into a clock by the addition of an electronic counter, which records the number of pulses. *See* DIGITAL COUNTER.

Basic principles. Most methods of timekeeping rely on counting some periodic event, such as the rotation of the Earth, the motion of a pendulum in a grandfather clock, or the vibrations of a quartz crystal in a watch. An atomic clock relies on counting periodic events determined by the difference of two different energy states of an atom. According to quantum mechanics, the internal energy of an atom can assume only certain discrete values. A transition between two energy states with energies E_1 and E_2 may be accompanied by the absorption or emission of a photon (particle of electromagnetic radiation). The frequency ν of this radiation is given by the equation below, where h is Planck's constant. A basic

$$h\nu = |E_2 - E_1|$$

advantage of atomic clocks is that the frequency-determining elements, atoms of a particular isotope, are the same everywhere. Thus, atomic clocks constructed and operated independently will measure the same time interval, that is, the length of time between two events. In order for the two clocks to agree on the time, they must be synchronized at some earlier time. *See* ATOMIC STRUCTURE AND SPECTRA; ENERGY LEVEL (QUANTUM MECHANICS); QUANTUM MECHANICS.

An atomic frequency standard can be either active or passive. An active standard uses as a reference the electromagnetic radiation emitted by atoms as they decay from a higher energy state to a lower energy state. An example is a self-oscillating maser. A passive standard attempts to match the frequency of an electronic oscillator or laser to the resonant frequency of the atoms by means of a feedback circuit. The cesium atomic beam and the rubidium gas cell are examples of passive standards. Either kind of standard requires some kind of frequency synthesis to produce an output near a convenient frequency, such as 5 MHz, that is proportional to the atomic resonance frequency. *See* FEEDBACK CIRCUIT; LASER; MASER; OSCILLATOR.

Accuracy and stability. Two different gages of the quality of a clock are accuracy and stability. The accuracy of a frequency standard is defined in terms of the deviation of its frequency from an ideal standard. In practice, it might be defined in terms of the frequency differences measured between independently constructed and operated standards of the same type. Improving the accuracy depends on understanding and controlling all the parameters that might cause the frequency to shift. The stability of a frequency standard is defined in terms of the constancy of its average frequency from one interval of time to the next. For many frequency standards, the stability initially improves with increasing measurement time but eventually gets worse. That is, a more precise measurement of the frequency can be made by averaging together successive measurements, until some imperfection in the apparatus causes the frequency to change. The stability increases with increased Q (resonance frequency divided by the width of the resonance) and with increased measurement signal-to-noise ratio. *See* Q (ELECTRICITY); SIGNAL-TO-NOISE RATIO.

Common types. The three most commonly used types of atomic clock are the cesium atomic beam, the hydrogen maser, and the rubidium gas cell. The cesium clock has high accuracy and good long-term stability. The hydrogen maser has the best stability for periods of up to a few hours. The rubidium cell is the least expensive and most compact and also has good short-term stability.

Cesium atomic-beam clock. This clock (**Fig. 1**) uses a 9193-MHz transition between two hyperfine energy states of the cesium-133 atom. Both the atomic nucleus and the outermost electron have magnetic moments; that is, they are like small magnets, with a north and a south pole. The two hyperfine energy states differ in the relative orientations of these magnetic moments. The cesium atoms travel in a collimated beam through an evacuated region. Atoms in the different hyperfine states are deflected into different trajectories by a nonuniform magnetic field. Atoms in one of the two states are made to pass

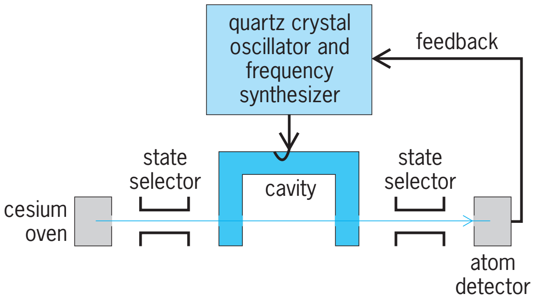


Fig. 1. **Cesium atomic-beam clock. (*After G. Kamas, ed., Time and Frequency Users' Manual, NBS Tech. Note 695, 1977*)**

through a microwave cavity, where they are exposed to radiation near their resonance frequency. The resonant radiation may cause the atom to make a transition from one state to the other; if that happens, the atom is deflected by a second, nonuniform magnetic field onto a detector. *See* ELECTRON SPIN; HYPERFINE STRUCTURE; MAGNETIC MOMENT; MOLECULAR BEAMS; NUCLEAR MOMENTS.

The *Q* of the resonance is over 10^8 for some laboratory standards and somewhat less for the smaller standards that are commercially available. Cesium atomic beams, including variants such as optically pumped atomic beams and atomic fountains, are the most accurate of all atomic clocks. The best models have an error of less than 1 part in 10^{15}, or about 1 s in 3×10^7 years. For this reason, cesium has become the basis of the international definition of the second: the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine states of the ground state of the cesium-133 atom. The cesium clock is especially well suited for applications such as timekeeping, where absolute accuracy without recalibration is necessary. Measurements from many cesium clocks throughout the world are averaged together to define an international time scale that is uniform to parts in 10^{14}, or about 1 microsecond in a year. *See* ATOMIC TIME; DYNAMICAL TIME; PHYSICAL MEASUREMENT.

Hydrogen maser. This instrument (**Fig. 2**) is based on the hyperfine transition of atomic hydrogen, which has a frequency of 1420 MHz. Atoms in the higher hyperfine energy state are selected by a focusing magnetic field, so that they enter an evacuated storage bulb inside a microwave cavity. The atoms bounce off the poly(tetrafluoroethylene)-coated walls for about 1 s before they are induced to make a transition to the lower hyperfine state by a process called stimulated emission. The stimulated emission from many atoms creates a self-sustaining microwave oscillation.

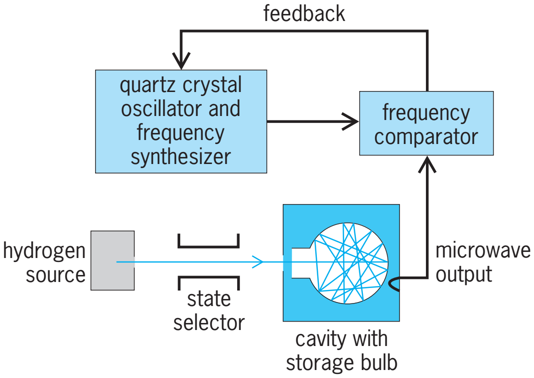


Fig. 2. **Hydrogen maser. (*After G. Kamas, ed., Time and Frequency Users' Manual, NBS Tech. Note 695, 1977*)**

The resonance *Q* is about 10^9. The best hydrogen masers have a stability of about 1 part in 10^{15} for averaging periods of 10^4 s. Over longer periods of time, the frequency drifts, primarily because of changes of the cavity tuning. Collisions with the walls cause the frequency to be shifted by about 1 part in 10^{11} relative to that of a free atom, but the magnitude of the shift varies from one device to another. This shift limits the accuracy of the hydrogen maser to about 1 part in 10^{12}.

The hydrogen maser can also be operated as a passive device, with improved long-term stability, due to the addition of automatic cavity tuning. The short-term stability is worse than that for an active maser.

Rubidium gas cell. This device (**Fig. 3**) is based on the 6835-MHz hyperfine transition of rubidium-87. The rubidium atoms are contained in a glass cell together with a buffer gas, such as argon, that prevents them from migrating to the cell walls. A method called optical pumping is used to prepare the atoms in one hyperfine state. Filtered light from a rubidium resonance lamp is absorbed by atoms in one of the two hyperfine states, causing them to be excited to a higher state, from which they quickly decay to the other hyperfine state. If the atoms are then subjected to microwave radiation at the hyperfine transition frequency, they are induced to make transitions back to the other hyperfine state. They can then absorb light again from the lamp; this results in a detectable decrease in the light transmitted through the cell.

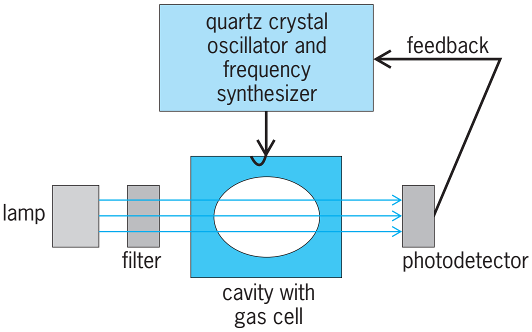


Fig. 3. **Rubidium gas cell. (*After G. Kamas, ed., Time and Frequency Users' Manual, NBS Tech. Note 695, 1977*)**

The *Q* is only about 10^7, but the short-term stability is quite good, reaching 1 part in 10^{13} for averaging times of 1 day. After longer periods, changes in the buffer gas pressure and the lamp cause the frequency to drift. The accuracy is not better than 1 part in 10^{10}. Rubidium standards are used in applications that do not require the accuracy of a cesium standard.

Laboratory types. Many other kinds of atomic clocks, such as thallium atomic beams and ammonia and rubidium masers, have been demonstrated in the laboratory. The first atomic clock, constructed at the National Bureau of Standards in 1949, was based on a 24-GHz transition in the ammonia molecule.

Some laboratories have tried to improve the cesium atomic-beam clock by replacing the magnetic state selection with laser optical pumping and fluorescence detection. Improved performance is expected because of increased signal-to-noise ratio and a more uniform magnetic field. One such standard, called NIST-7, was in operation at the U.S. National Institute of Standards and Technology and was formerly the primary frequency standard for the United States. Other laboratories have studied atomic-beam standards based on magnesium, calcium, or methane, which have frequencies higher than that of cesium.

Atomic frequency standards can also be based on optical transitions. One of the best-developed optical frequency standards is the 3.39-micrometer (88-THz) helium-neon laser, stabilized to a transition in the methane molecule. Research is being carried out on the use of laser-cooled, neutral atoms, such as calcium or strontium, as optical frequency standards.

Frequency synthesis chains have been built to link the optical frequency to radio frequencies. A new device for optical frequency synthesis, the femtosecond-laser frequency comb, has largely replaced the earlier, more cumbersome methods based on the frequency doubling and mixing of many separate lasers. In this device, a femtosecond mode-locked laser generates an optical spectrum made up of many equally spaced components. The frequency of a particular component is given by $f(m) = f_0 + mf_{\text{rep}}$, where m is an integer. The pulse repetition rate f_{rep} of the mode-locked laser can easily be measured by detecting the light with a photodiode, while the offset frequency f_0 can be determined by a method called self-referencing. The frequency of another laser can be determined from the beatnote with the nearest component of the spectrum. This provides a method to transfer the stability of an optical frequency to the frequency of the pulse repetition rate, which is in the radio-frequency region. This makes possible an optical clock, given a stable optical frequency.

Ion traps, which confine ions in a vacuum by electric and magnetic fields (**Fig. 4**), have been studied for use in both microwave and optical atomic clocks. They provide a benign environment for the ions while still allowing a long measurement time. Microwave clocks based on buffer-gas-cooled, optically pumped mercury-199 or ytterbium-171 ions have been built and show good stability. Values of Q as high as 1.5×10^{13} have been observed on the ytterbium-171 hyperfine transition. Other trapped ion standards make use of laser cooling to reduce frequency errors due to Doppler shifts. Laser cooling is a method by which resonant light pressure is used to damp the motion of atoms. A microwave clock based on laser-cooled mercury-199 ions has demonstrated an accuracy close to that of the best cesium clocks. It is possible to observe extremely high values of Q on optical transitions of trapped ions. An optical transition has been observed in a single, trapped mercury ion with a Q of 1.6×10^{14}. Similar results have been obtained with other ions, such as strontium-88 and ytterbium-171. An optical frequency standard based on such an ion might be capable of an accuracy of 1 part in 10^{18}. *See* LASER COOLING; PARTICLE TRAP.

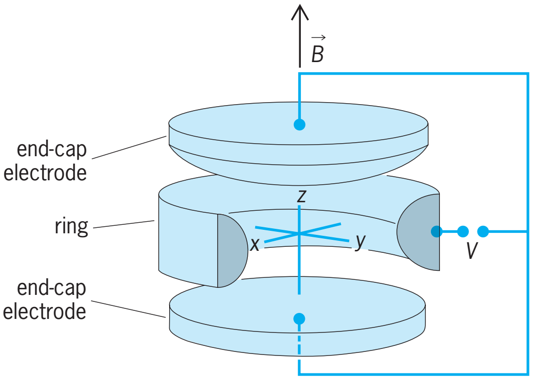


Fig. 4. Electrodes used to create the confining electric potential for a Penning ion trap or a Paul ion trap. An electric potential *V*, which is static for a Penning trap and oscillating for a Paul trap, is applied between the ring electrode and the end-cap electrodes. The Penning trap requires a uniform magnetic field *B*. (*After F. J. Rogers and H. E. Dewitt, eds., Strongly Coupled Plasma Physics, Plenum, 1987*)

The atomic fountain (**Fig. 5**) is another method of using laser-cooled atoms in a frequency standard. In this method, neutral atoms, such as cesium, are laser cooled and state selected and then launched upward on a ballistic trajectory. The atoms pass through a microwave cavity, as in the atomic-beam method. They are moving slowly enough that they fall back down through the cavity a second time, after which they are detected by laser-induced fluorescence. The general principles are the same as in an atomic-beam standard, but greater precision is possible because the flight time of an atom through the apparatus is longer. Cesium atomic fountain clocks have now replaced cesium atomic beams at national measurement laboratories, such as the National Institute of Standards and Technology in Boulder, Colorado. The best ones have accuracies better than 1 part in 10^{15}.

The coherent population trapping (CPT) clock is a new type of microwave atomic clock, which is more interesting for its compact size than for its accuracy and stability. As in the rubidium cell clock, a hyperfine resonance in gas-phase atoms is excited. Rather than being excited directly by microwave radiation at the hyperfine transition frequency, the resonance is excited by two laser light fields whose frequency difference is equal to the hyperfine transition frequency. In practice, the two fields are generated by modulation of the frequency of a single laser. Such a clock can be made very compact, since there is no need for a microwave cavity to enhance a microwave field. CPT clocks based on the cesium hyperfine transition are currently being developed. In some of them, the cesium gas cell measures only a few millimeters on a side, and the entire device, exclusive of the power

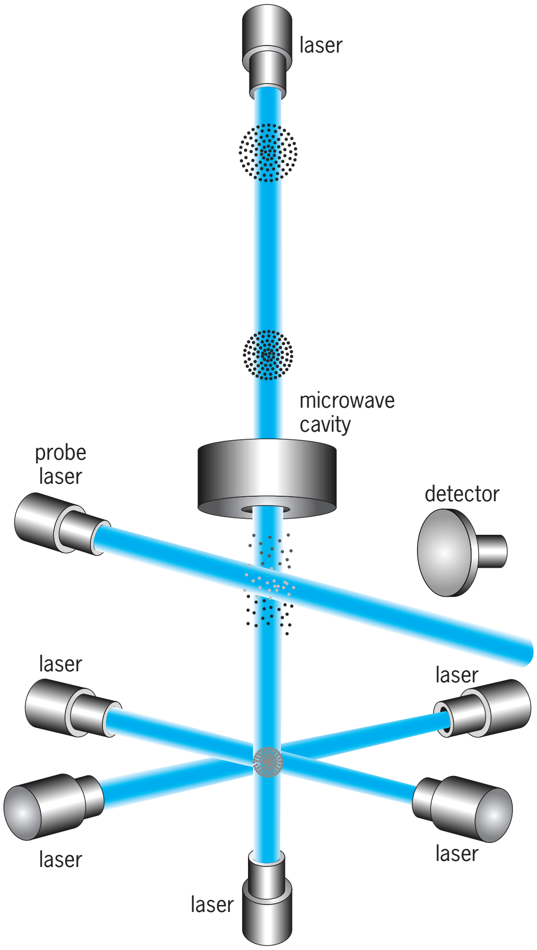


Fig. 5. Cesium atomic fountain clock. A sample of cesium atoms is trapped and laser-cooled at the intersection of six laser beams. The atoms are then launched upward, passing through the microwave cavity, then falling back through the cavity. After passing through the cavity the second time, the atoms are illuminated by the probe laser, and the induced fluorescence is detected. (*National Institute of Standards and Technology*)

supply, might fit into a volume of 1 cm^3.

Applications. Atomic clocks are used in applications for which less expensive alternatives, such as quartz oscillators, do not provide adequate performance. The use of atomic clocks in maintaining a uniform international time scale has already been mentioned; other applications are described below. *See* QUARTZ CLOCK.

Navigation. The Global Positioning System is a satellite-based system that enables a user with a suitable radio receiver to determine position within about 10 m (33 ft). The satellites send out accurately timed radio pulses, from which the user's receiver can calculate its location and time. The satellites and the ground stations, but not the users, need atomic clocks (usually cesium clocks). *See* SATELLITE NAVIGATION SYSTEMS.

Communications. Various digital communications systems require precise synchronization of transmitters and receivers in a network. Some systems use time-division multiplexing, in which many channels of information are sent over the same line by sequentially allotting a small time slot to each channel. Timing is very critical when there are several sources of information with their own clocks. The primary timing is provided by cesium clocks. *See* ELECTRICAL COMMUNICATIONS; MULTIPLEXING AND MULTIPLE ACCESS; PULSE MODULATION.

Radio astronomy. Very long baseline interferometry is a technique that allows two or more widely separated radio telescopes to achieve very high angular resolution by correlation of their signals. The system has the resolution that a single telescope would have if its aperture were equal to the distance between the telescopes. This can be thousands of miles. The stable timing needed to correlate the signals is provided by hydrogen masers. *See* RADIO ASTRONOMY; RADIO TELESCOPE.

Space exploration. Navigation of space probes by Doppler tracking requires very stable local oscillators, derived from atomic frequency standards. Doppler tracking relies on determining the velocity of the spacecraft by measuring the frequency shift of a signal after it has been echoed to the Earth by a transponder on the spacecraft. Stable local oscillators are also needed for studies of planetary atmospheres and rings by fluctuations of the radio signals transmitted through them. *See* SPACE NAVIGATION AND GUIDANCE.

Fundamental science. According to A. Einstein's special and general theories of relativity, a moving clock runs slower than a stationary one, and a clock on the Earth's surface runs slower than one far from the Earth. These predictions were verified to high accuracy by an experiment in which a hydrogen maser was launched in a rocket to an altitude of 10,000 km (6000 mi). *See* CLOCK PARADOX; RELATIVITY; TIME.

Wayne M. Itano

Bibliography. J. C. Bergquist, S. Jefferts, and D. J. Wineland, Time measurement at the Millennium, *Phys. Today*, 54(3):37–42, 2001; W. M. Itano and N. F. Ramsey, Accurate measurement of time, *Sci. Amer.*, 269(1):56–65, July 1993; J. Jespersen and J. Fitz-Randolph, *From Sundials to Atomic Clocks*, National Institute of Standards and Technology, Monogr. 155, U.S. Government Printing Office, Washington, D.C., 1999; T. Jones, *Spliting the Second: The Story of Atomic Time*, Institute of Physics Publishing, Bristol and Philadelphia, 2000; F. G. Major, *The Quantum Beat: The Physical Principles of Atomic Clocks*, Springer, New York, 1998; N. F. Ramsey, Precise measurement of time, *Amer. Sci.*, 76(1):42–49, 1988.

Atomic Fermi gas

A gas of atoms, generally ultracold, in which the individual atoms possess half-integer spin (that is, are fermionic). Since the 1980s, researchers have developed powerful new techniques for cooling atoms to ultralow temperatures. Among the most significant achievements made possible by these technical developments was the creation in 1995 of the long-sought Bose-Einstein condensate. In a

Bose-Einstein condensate, atoms are cooled to such a low temperature that they collect in the quantum-mechanical ground state of their confinement volume. Bose-Einstein condensation is possible only in the class of particles known as bosons. Particles of the other class, fermions, are forbidden to occupy the same quantum state and are thus prevented from directly condensing.

Progress in studying ultracold fermions has lagged behind that of bosons, not because of lack of interest, but because they are experimentally more difficult to cool. In fact, fermions are especially significant because the fundamental building blocks of matter, protons, neutrons, and electrons, are all fermions. Most of the field of condensed matter physics is concerned with the behavior of fermions, especially electrons. Of particular interest is the phenomenon of Cooper pairing, where fermions, such as the electrons in a superconductor, or the atoms in helium-3 (^{3}He), form correlated pairs when cooled to sufficiently low temperature. Recent experimental progress in cooling atomic Fermi gases has enabled the realization of Cooper pairing in the gas phase. Because of the inherent simplicity and tunability of the atomic interactions in an ultracold gas, this work may help clarify many troubling issues in condensed-matter physics, including the mechanisms at work in high- temperature superconductors.

Behavior of bosons and fermions. We have known since the work of Satyendranath Bose, Albert Einstein, Enrico Fermi, and Paul Dirac in the 1920s that there are two fundamental types of particles. Bosons are particles with integer spin angular momentum ($0, \hbar, 2\hbar, \ldots$) and fermions are those with half integer ($^1/_2\hbar, ^3/_2\hbar, \ldots$). Since protons, neutrons, and electrons are each spin-$^1/_2$ particles, atoms made with an odd number of these constituents are themselves composite fermions. Examples are lithium-6 (^{6}Li) and potassium-40 (^{40}K), which happen to be the only stable fermionic isotopes of the alkali-metal elements. The differences between bosons and fermions are significant only when they become quantum-degenerate, that is, when the size of the atomic quantum-mechanical de Broglie wavelength, Λ, which increases as temperature decreases, is comparable to the average separation between atoms. When this condition is met, bosons undergo the Bose-Einstein condensation phase transition. Identical fermions, on the other hand, are forbidden by the Pauli exclusion principle from occupying the same quantum level (**Fig. 1**). As Cooper pairs are composite bosons, superconductivity and fermionic superfluidity are thought to be related to Bose-Einstein condensation of pairs. *See* BOSE-EINSTEIN CONDENSATION; EXCLUSION PRINCIPLE; QUANTUM MECHANICS; QUANTUM STATISTICS.

Cooling methods. The methods for cooling Fermi gases and for confining them in atom traps are similar to those used for bosons. The primary methods are laser cooling, atom trapping, and evaporative cooling. Laser cooling utilizes radiation pressure to slow atoms. Since temperature is related to the kinetic energy of the atoms, reducing their speed is equivalent to cooling. Laser cooling is typically limited to temperatures in the range of 100 millionths of a degree above absolute zero, or 100 microkelvin. Atom traps use laser radiation pressure or magnetic fields to confine atoms without walls. Since the atom temperatures are so low, isolation from physical boundaries is absolutely necessary to prevent heating.

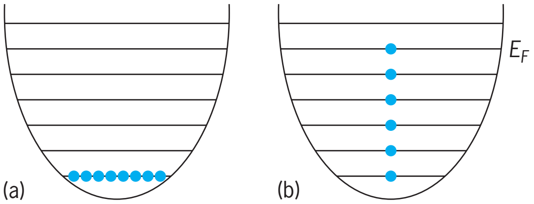


Fig. 1. Bosons and fermions at absolute temperature $T = 0$. Atoms in a magnetic trap feel a harmonic restoring potential, as indicated by the parabolic wells. The equally spaced lines in each well represent the quantized energy levels of the harmonic potential. (*a*) Bosons. At $T = 0$, they collectively occupy the ground level of the system. (*b*) Fermions. These particles must obey the Pauli exclusion principal, which forbids more than one identical fermion per level. At $T = 0$, they completely fill the energy-level ladder up to the Fermi energy E_F. The filling of energy levels in this manner is analogous to the way electrons fill the lowest available atomic orbitals in the periodic table of the elements.

Even lower temperatures, of the order of 100 nanokelvin (100 billionths of a degree above absolute zero), are needed to achieve quantum degeneracy in low-density atomic gases. These temperatures, as low as any produced by any method or found anywhere in the universe, are attainable using evaporative cooling. Evaporative cooling exploits the fact that there is always a distribution of energies in a gas, including some very energetic atoms in the "tail" of the distribution. By systematically removing only the hottest atoms from the trap, a large fraction of energy can be removed, while minimizing the number of atoms lost.

Effective evaporative cooling, however, requires that the atoms in the gas undergo elastic thermalizing collisions to continuously replenish the tail of hot atoms. The main experimental difficulty encountered with fermions is that the Pauli exclusion principle prevents identical fermions from interacting, so they are unable to thermalize. Several ways around this difficulty have been developed. First, one can make use of the different projections of the nuclear and atomic spin of the atom to make subsets of atoms that are distinguishable from other subsets. For example, one can make a mixture of ^{40}K atoms in two different nuclear spin states, so that the nuclear spins of some of the atoms point in a different direction than the others. Atoms in the same spin state will not interact, but two atoms in different spin states will. Unfortunately, this method cannot be universally used, because some spin states cannot be magnetically trapped, and because collisions between some spin states can result in the creation of atoms in a third state, with a release of excess energy that sends the atoms flying from the trap.

A more general way to cool fermions is to "sympathetically" cool them with bosons. The bosons are evaporatively cooled as usual, while the fermions are cooled by interaction with the bosons. In this way, ^{6}Li has been cooled to temperatures as low as 20 nanokelvin using either lithium-7 (^{7}Li) or sodium-23 (^{23}Na) atoms as the "refrigerant." After the fermions have been cooled, the bosons can be selectively removed from the trap by driving transitions to untrapped quantum states using microwave or optical pulses. The gas is probed by shining a laser beam through the gas of atoms and imaging the shadow cast by the atoms onto a charge-coupled-device (CCD) camera. *See* LASER COOLING; PARTICLE TRAP.

Fermi pressure. In 1926, Fermi realized that the energy of a gas of fermions is finite at zero absolute temperature (Fig. 1), and consequently it exerts a pressure. This "Fermi pressure" is a purely quantum-mechanical effect, as the pressure p from the classical ideal gas law, $p = nkT$, vanishes as the temperature, T, approaches 0. (In this expression, n is the number density of the gas, and k is the Boltzmann constant.) The implications of this result were soon realized by S. Chandrasekhar: under certain conditions electron Fermi pressure could prevent the gravitational collapse of relatively cold, aging stars known as white dwarfs. This prediction was verified by observation, and the same stabilization method was later found to apply to neutron stars. The effect of Fermi pressure has also been demonstrated for a trapped gas of ^{6}Li atoms (**Fig. 2**). The figure shows a gas containing both bosons (^{7}Li) and fermions (^{6}Li), at three different temperatures. Although both types of atoms are mixed together in the trap at the same time, they can be individually imaged. The images are separated horizontally in the figure to show the difference in behavior between bosons and fermions. Figure 3*a* corresponds to the highest temperature, $T \sim T_F$, where T_F is the Fermi temperature (Fig. 1). At this relatively high temperature, there is little difference in the spatial size of the fermions compared with the bosons. However, as the temperature is reduced to $\sim$0.5 T_F in Fig. 3*b*, and finally to $\sim$0.25 T_F in Fig. 3*c*, it becomes clear that the fermions occupy more volume than do the bosons. This difference is a manifestation of Fermi pressure, albeit on a much different size and energy scale than for stars. *See* FERMI-DIRAC STATISTICS; NEUTRON STAR; WHITE DWARF STAR.

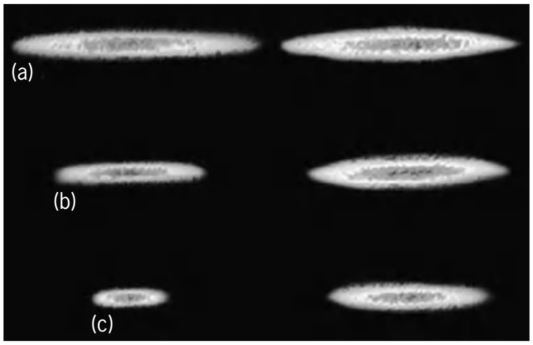


Fig. 2. **Fermi pressure in a gas of trapped atoms. The images on the left are of the bosons (^{7}Li), while those on the right are of the fermions (^{6}Li) taken at progressively lower temperatures. Each pair of images corresponds to a particular temperature. The ^{6}Li and ^{7}Li atoms are simultaneously trapped in the same volume, but the images are separated for clarity. (*a*) Temperature $T \cong 1.5\ T_c \cong T_F$, where T_c is the critical temperature for Bose-Einstein condensation, and T_F is the Fermi temperature. (*b*) $T \cong 0.5\ T_F$. (*c*) $T \cong T_c \cong 0.25\ T_F$. (*From A. G. Truscott et al., Observation of Fermi pressure in a gas of trapped atoms, Science, 291:2570–2572, 2001*)**

Cooper pairing in an atomic gas. One of the most remarkable achievements in physics in recent years was the realization of Cooper pairing in an ultracold atomic Fermi gas. The pairing of fermions is a phase transition, just as Bose-Einstein condensation is for bosons. In fact, these phase transitions are closely related because a pair of fermions is itself a composite boson. The realization of Cooper pairing required that the atoms be cooled to even lower temperatures than have been demonstrated thus far. Furthermore, the atoms must attract one another. Since identical ultracold fermions cannot interact, as was the case for evaporative cooling, the experimental schemes use a mixture of fermions in two different spin states. A powerful technique for producing strong, tunable interactions has been developed using a Feshbach resonance, in which a magnetic field tunes the energy of a bound state of two atoms, a diatomic molecule, to be equal to the energy of a pair of free atoms. The resulting collisional resonance can be used to create strong or weak interactions that are either repulsive or attractive depending on the value of the magnetic field. Because of the versatility it affords, this interaction "knob" is one of the most compelling aspects of quantum gas research. *See* RESONANCE (QUANTUM MECHANICS).

Traditional superconductors and superfluid ^{3}He form pairs in the weakly interacting limit, where the size of the Cooper pairs is much larger than the average distance between fermions. This situation is very well described by the BCS (Bardeen-Cooper-Schrieffer) theory of superconductivity. On the other hand, the interactions in high-temperature superconductors are believed to be much stronger,

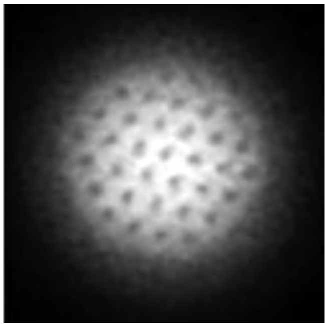


Fig. 3. **Vortex array in a paired gas of ^{6}Li atoms. Vortices with quantized angular momentum form when the gas is stirred. They appear as dark regions in the figure because centrifugal forces push atoms away from the vortex core. Interactions between atoms cause the vortices to form a regular lattice array. Such an array of quantized vortices is very similar to those created in Bose-Einstein condensates of ultracold bosonic atoms and is a hallmark of superfluidity. (*From M. W. Zwierlein et al., Vortices and superfluidity in a strongly interacting Fermi gas, Nature, 435:1047–1051, 2005*)**

and hence the pairs are spatially smaller. Feshbach resonances have enabled the investigation of pairing in atomic Fermi gases over a range of interaction strength, but the experiments thus far have explored mainly the strongly interacting regime where the pairs are almost molecular in character. In this regime, there is yet no complete theory of Cooper pairing, and so the experiments provide new insight into this physical realm. The evidence for Cooper pairing obtained in recent years includes direct measurements of the correlation between atom pairs, measurements of the heat capacity, and the observation of quantized vortices (**Fig. 3**), a hallmark of superfluidity and one of the consequences of Cooper pairing. *See* LIQUID HELIUM; QUANTIZED VORTICES; SUPERCONDUCTIVITY; SUPERFLUIDITY.

Randall Hulet

Bibliography. B. DeMarco and D. S. Jin, Onset of Fermi degeneracy in a trapped atomic gas, *Science*, 285:1703–1706, 1999; D. Jin, A Fermi gas of atoms, *Phys. World*, 15(4):27–32, April, 2002; H. J. Metcalf and P. van der Straten, *Laser Cooling and Trapping*, Springer, 1999; C. J. Pethick and H. Smith, *Bose-Einstein Condensation in Dilute Gases*, Cambridge, 2002; A. G. Truscott et al., Observation of Fermi pressure in a gas of trapped atoms, *Science*, 291:2570–2572, 2001; M. W. Zwierlein et al., Vortices and superfluidity in a strongly interacting Fermi gas, *Nature*, 435:1047–1051, 2005.

Atomic force microscopy

A form of microscopy in which a sharp tip is scanned over the surface of a sample while sensing the interaction force between the tip and the sample. Due to its ability to observe and manipulate surfaces in aqueous solution, atomic force microscopy (AFM) has revolutionized the way in which microscopists explore biological structures, from single molecules to living cells.

Principle and technique. The basic idea behind AFM is to use so-called near-field physical interactions (that is, repulsive or attractive forces) acting between a sharp tip and the specimen to generate three-dimensional images of the surface without dependence on an incident beam, as in optical and electron microscopies (**Fig. 1**). The sample is mounted on a piezoelectric scanner that allows high-resolution, three-dimensional positioning. The force interacting between tip and specimen is measured by the bending, or "deflection," of a soft cantilever spring, typically made of silicon or silicon nitride. The cantilever deflection is detected by a laser beam, usually focused on the free end of the cantilever, and reflected into a photodiode. *See* INTERMOLECULAR FORCES; PIEZOELECTRICITY; SURFACE PHYSICS.

Operational modes. AFM can be operated in several different modes: contact, tapping, noncontact, and force spectroscopy.

Contact mode. In contact mode, the tip and sample are placed in contact, and the tip is simply dragged across the surface, resulting in a topographic image of the surface. The scanning is usually done under feedback control, in which the sample is moved toward or away from the cantilever during the scan so that the bending of the cantilever, normal to the surface, remains constant in order to maintain constant force. The up and down motion of the sample is, therefore, a record of the sample topography.

Tapping mode. In tapping mode (also known as intermittent mode), the cantilever is oscillated near its resonant frequency using a piezoelectric actuator. Moving the oscillating tip until it lightly touches the surface reduces the oscillation amplitude. The reduction in oscillation amplitude now becomes the feedback control signal, which can be used to measure the surface topography. Due to its intermittent sample contact, tapping mode AFM can significantly reduce lateral forces during imaging, which may be advantageous for imaging soft biological samples.

Noncontact mode. AFM may also be operated in the noncontact mode, in which the tip is oscillated at a distance from the surface where attractive van der Waals forces dominate.

Force spectroscopy. Besides recording images, AFM can measure minute forces within or between biomolecules with piconewton (pN; 10^{-12} N) sensitivity. In this operating mode, known as force spectroscopy, the cantilever deflection is recorded as a function of the vertical displacement of the piezoelectric scanner; that is, deflection is recorded as the sample is pushed toward the tip and retracted. The resulting force versus separation distance curves provide a wealth of information on biophysical properties and biomolecular interactions and permit the manipulation of single biomolecules.

Nanoscale visualization of biological structures. Since the late 1980s, AFM has proved very useful in imaging the structure of biomolecules, lipid membranes, membrane proteins, and cells. Remarkably, images can be recorded in real-time under physiological conditions, with unprecedented signal-to-noise

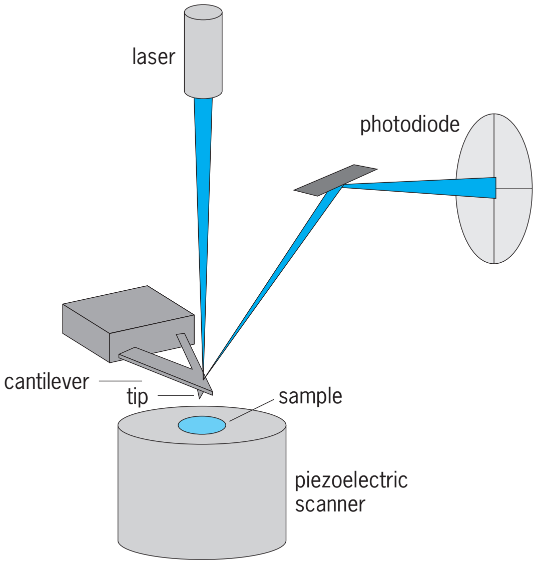
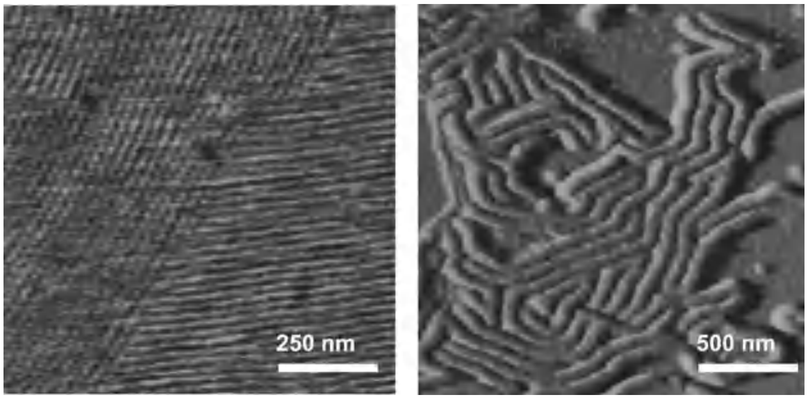


Fig. 1. Principle of AFM.

-s



(a) (b)

Fig. 2. Application of AFM imaging to lipid membranes. (*a*) Nanostructures with well-defined bendings and a ∼30 nm periodicity, formed at the surface of a mixed dipalmitoylphosphatidylcholine/surfactin bilayer supported on mica. (*b*) Another type of nanostructure with domains protruding ∼5 nm above the surface of a dipalmitoylphosphatidylcholine/dioleoylphosphatidic acid bilayer imaged in the presence of the simian immunodeficiency virus fusion peptide.

ratio and resolution, thereby complementing the range of instruments available to investigate structural properties. The power of this approach for probing lipid bilayers is illustrated in **Fig. 2**. The AFM images, recorded directly in buffer solution, reveal that remarkable nanostructures are formed at the bilayer surface in the presence of lipopeptide or fusion peptide. This example illustrates how AFM is enabling biophysical researchers to investigate membrane structure, properties, and interactions (such as molecular packing, structural defects, nanodomains, interaction with external agents, fusion and adhesion events). These nanoscale studies are also expected to have an important impact on nanobiotechnology for the design of biosensor and biomimetic surfaces (synthetic surfaces that mimic biological surfaces). *See* BIOELECTRONICS; BIOSENSOR; NANOSTRUCTURE.

The stability and structural regularity of two-dimensional arrays of membrane proteins make them well suited for high-resolution imaging. In the past decade, progress in AFM instrumentation and methodology has allowed various membrane proteins to be explored with subnanometer resolution, such as the hexagonally packed intermediate (HPI) protein layer of *Deinococcus radiodurans*, the purple membrane of the archaean *Halobacterium*, and porin (protein channel) crystals of *Escherichia coli*. Remarkably, conformational changes of individual membrane proteins can be detected and related to function. For instance, voltage and pH-dependent conformational changes were observed in bacterial porins and correlated with the closure of the channel entrance, suggesting that this is a mechanism to protect the cells from drastic changes in the environment. *See* CELL MEMBRANES; CELL PERMEABILITY.

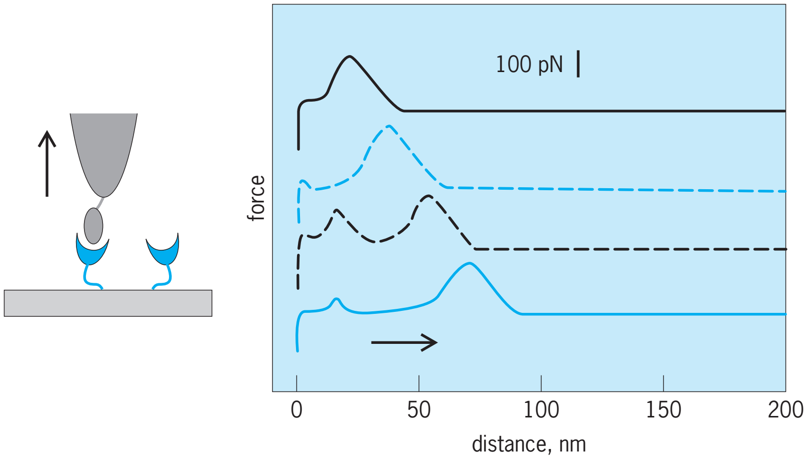


Fig. 3. Application of AFM force spectroscopy to the detection of single molecular recognition events. Force–distance curves were recorded between an AFM tip and a substrate functionalized with oligoglucose and concanavalin A, respectively. The ∼100 pN unbinding forces reflect the interaction between single lectin and carbohydrate molecules.

Another exciting feature of the instrument is the possibility to monitor dynamic processes using time-lapse imaging; that is, for the first time microscopists can observe single biomolecules at work, such as ribonucleic acid (RNA) polymerase moving along deoxyribonucleic acid (DNA) during the transcription process, and can follow dynamic events in living cells, such as cell growth and division. *See* DEOXYRIBONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA).

There are several technical challenges to address in the future, including reducing tip–specimen interaction forces for high-resolution imaging of soft samples such as cells and improving time resolution to probe dynamic processes such as molecular conformational changes.

Detecting and manipulating single biomolecules. The piconewton force sensitivity of AFM force spectroscopy enables it to be used to manipulate single biomolecules and to measure their inter- and intramolecular interactions in relation to function. In particular, the AFM tip can be used to pull on flexible macromolecules (such as DNA, proteins, and polysaccharides) to learn about their elastic behavior as well as their conformational and folding/unfolding properties. While pioneering experiments were performed on pure molecules, such as the polysaccharide dextran, recent studies have also focused on membranes and cells. For instance, force curves recorded via AFM on the bacterial HPI layer showed multiple adhesion peaks that were attributed to the sequential unzipping of entire bacterial pores. Remarkably, the resulting molecular defects could be directly visualized in the manipulated specimen using high-resolution imaging. Another example of such a nanodissection experiment is the extraction of individual bacteriorhodopsins from purple membranes.

Measuring molecular recognition forces. AFM force spectroscopy is also well suited to measure molecular recognition forces between receptors and ligands, which play important roles in the life sciences and form the basis of many biotechnological applications. By attaching specific biomolecules on the AFM tips, researchers can measure the binding forces between complementary molecules, such as biotin–avidin, antibody–antigen, and DNA strands. The measured forces are usually in the range of 50–200 pN, depending on the nature of the molecules and the pulling rate. An example of such an experiment is shown in **Fig. 3**. To measure lectin–carbohydrate interactions, a solid substrate was modified with the lectin concanavalin A and complementary oligoglucose molecules were attached to an AFM tip. Force–distance curves recorded between the modified tip and substrate showed unbinding forces of ∼100 pN that were attributed to the specific interaction between individual lectin and carbohydrate molecules.

Besides providing fundamental insights into the properties and functions of biomolecules, these AFM-based analyses have important promise in medicine for the rapid, highly sensitive detection of biomolecules such as antigens and toxins. Moreover, biotips can also be used to detect specific molecules on cell surfaces and to map their distribution, opening the door to new applications in cell biology and microbiology for elucidating the molecular bases of cell-surface processes such as cell adhesion. *See* CELL (BIOLOGY); CELL ADHESION; MOLECULAR RECOGNITION.

Yves F. Dufrêne

Bibliography. G. Binnig, C. F. Quate, and C. Gerber, Atomic force microscope, *Phys. Rev. Lett.*, 56:930–933, 1986; R. J. Colton et al., *Procedures in Scanning Probe Microscopies*. Wiley, Chichester, 1998; B. P. Jena and J. K. H. Hörber, *Atomic Force Microscopy in Cell Biology: Methods in Cell Biology*, vol. 68, Academic Press, San Diego, 2002; V. J. Morris, A. R. Kirby, and A. P. Gunning, *Atomic Force Microscopy for Biologists*, Imperial College Press, London, 1999; Z. Shao et al., Biological atomic force microscopy: What is achieved and what is needed, *Adv. Phys.*, 45:1–86, 1996.

Atomic mass

The mass of an atom or molecule on a scale where the mass of a carbon-12 (^{12}C) atom is exactly 12.0 atomic mass units. The mass of any atom is approximately equal to the total number of its protons and neutrons multiplied by the atomic mass unit, $u = 1.6605387 \times 10^{-24}$ gram. (Electrons are much lighter, about 0.0005486 u.) No atom differs from this simple formula by more than 1%, and stable atoms heavier than helium all lie within 0.3%. *See* ATOMIC MASS UNIT.

This simplicity of nature led to the confirmation of the atomic hypothesis—the idea that all matter is composed of atoms, which are identical and chemically indivisible for each chemical element. In 1802, G. E. Fischer noticed that the weights of acids needed to neutralize various bases could be described systematically by assigning relative weights to each of the acids and bases. A few years later, John Dalton proposed an atomic theory in which elements were made up of atoms that combine in simple ways to form molecules.

In reality, nature is more complicated, and the great regularity of atomic masses more revealing. Two fundamental ideas about atomic structure come out of this regularity: that the atomic nucleus is composed of charged protons and uncharged neutrons, and that these particles have approximately equal mass. The number of protons in an atom is called its atomic number, and equals the number of electrons in the neutral atom. The electrons, in turn, determine the chemical properties of the atom. Adding a neutron or two does not change the chemistry (or the name) of an atom, but does give it an atomic mass which is 1 u larger for each added neutron. Such atoms are called isotopes of the element, and their existence was first revealed by careful study of radio-active elements. Most naturally occurring elements are mixtures of isotopes, although a single isotope frequently predominates. Since the proportion of the various isotopes is usually about the same everywhere on Earth, an average atomic mass of an element can be defined, and is called the atomic weight. Atomic weights are routinely used in chemistry in order to determine how much of one chemical will react with a given weight of another. *See* ATOMIC STRUCTURE AND SPECTRA; ISOTOPE; RELATIVE ATOMIC MASS.

In contrast to atomic weights, which can be defined only approximately, atomic masses are exact constants of nature. All atoms of a given isotope are truly identical; they cannot be distinguished by any method. This is known to be true because the quantum mechanics treats identical objects in special ways, and makes predictions that depend on this assumption. One such prediction, the exclusion principle, is the reason that the chemical behavior of atoms with different numbers of electrons is so different. Other consequences, such as the absence of certain features of the optical spectra of molecular gases, and the existence of Bose-Einstein condensates, have also been observed, confirming the truly indistinguishable nature of different atoms of the same isotope. *See* EXCLUSION PRINCIPLE; MOLECULAR STRUCTURE AND SPECTRA; QUANTUM MECHANICS; QUANTUM STATISTICS.

Measurement of atomic masses. In 1913, J. J. Thomson made the first true atomic mass measurements by deflecting a beam of neon ions (charged atoms) with a magnetic field. His resolution was just sufficient to separate neon-20 from neon-22. However, his basic method, measuring the curvature of a beam of ions (charged atoms) in a magnetic field, formed the basis of mass spectrometry for many years. F. W. Aston made great improvements to ion beam, mass spectrometers during the 1920s and 1930s. The common isotopes were identified, and improvements in resolution brought mass accuracies below a thousandth of a mass unit. This accuracy made it possible to verify the equivalence of mass and energy, expressed in the formula $E = mc^2$, where c is the speed of light. In this case, m is the mass difference between reactants and products, and E is the energy liberated in the associated nuclear reaction. For example, when two atoms of deuterium (mass 2.0141 u) fuse to form a helium atom (mass 4.0026 u), 0.0256 u is converted to energy. When multiplied by c^2, this is a tremendous amount of energy, several million times the chemical energy that would be released if the deuterium reacted with oxygen to form heavy water. Thus a kilogram of nuclear fuel can liberate as much energy as several hundred tons of coal. Atomic mass measurements now provide measurements of nuclear reaction energies far more accurately than they can be predicted from nuclear theory. *See* MASS SPECTROSCOPE; NUCLEAR FUSION; NUCLEAR REACTION.

Single-ion mass measurements. Starting about 1990, a new method of atomic mass measurement was developed with accuracies of about 1 part in 10^{10}, a hundred or more times more accurate than had been possible before. The new technique involves isolating a single ion in a Penning trap, and measuring the frequency at which it orbits in a steady magnetic field produced by a superconducting magnet. The ion is set in motion in a tiny orbit, and its motion is detected with sensitive superconducting electronics. This cyclotron frequency is given by qB/m, where q is the charge of the ion, B is the magnetic field strength, and m is the ion mass. Comparing the cyclotron frequency of the unknown ion with the frequency of a carbon-containing atom gives its mass in u, since 1 u is defined as 1/12 the mass of a carbon-12 atom. *See* PARTICLE ACCELERATOR; PARTICLE TRAP.

The tremendous improvement in accuracy brought on by single-ion measurements not only has improved traditional applications of mass spectrometry such as the calibration of the gamma-ray spectrum, but also has opened the door to new applications in fundamental physics and metrology including evaluation of fundamental constants and setting a lower limit on the neutrino mass, and may ultimately lead to the replacement of the kilogram mass standard. *See* FUNDAMENTAL CONSTANTS; NEUTRINO.

Determination of α. Perhaps the most important dimensionless number in physics is the fine structure constant, $\alpha = 1/137.036$, which underlies much of atomic, nuclear, and elementary particle physics. In addition to determining the spacing of fine structure lines in atomic spectra, α sets the strength of various interactions, including the electromagnetic, and appears in numerous other important ratios. The square of the fine structure constant equals (with several small corrections) twice the ratio of the highest frequency of ultraviolet light emitted by a hydrogen atom, called the Rydberg, to the frequency of a photon whose energy equals the rest energy of an electron, mc^2. The hydrogen ultraviolet frequency has been measured to about 1 part in 10^{11}. The frequency associated with the electron mass is determined indirectly from Penning trap measurements of its atomic mass. The ratio of an atomic mass to its corresponding frequency equals $1000/c^2$ times the molar Planck constant, N_Ah, where N_A is Avogadro's number and h is Planck's constant. Several measurements of N_Ah have been under taken, all requiring accurate atomic mass measurements. Together these may provide the best determination of α. *See* ELECTRON SPIN; FINE STRUCTURE (SPECTRAL LINES); FUNDAMENTAL INTERACTIONS; RYDBERG CONSTANT.

Redefinition of kilogram. The mass of an atom can now be measured much more accurately than the mass of the platinum-iridium cylinder that defines the kilogram. This artifact has been placed in a vault at the International Bureau of Weights and Measures in Sèvres, France. Adsorbed gases and damage from cleaning limit the accuracy of this standard to 1 part in 10^9 or worse. A difficult compromise is made between using the standard and preserving it, with the result that it was compared with other standards only three times in the twentieth century. These shortcomings make it very attractive to define the kilogram in terms of atomic masses, since atoms are widely available and all identical. This would be done by defining the Avogadro number N_A to be exactly $6.02214199 \times 10^{23}$, for example. This number of carbon atoms would then have a mass of 12 grams by definition. Such a definition can be useful only if standard masses can be made in which the number of atoms is known. This can most likely be accomplished using silicon, since the technology exists for making large ultrapure crystals of silicon and since the atomic mass of silicon-28 has been measured accurately. The spacing between atoms in silicon can be accurately measured using x-ray channeling and laser beams of known wavelength. An atomic definition of the kilogram would complete the process of making all the fundamental metrological standards (length, time, voltage, resistance, and so forth) depend on reproducible natural constants rather than the preservation and stability of any particular artifact. *See* AVOGADRO NUMBER; PHYSICAL MEASUREMENT.

Fred L. Palmer; David E. Pritchard

Bibliography. G. Audi et al., The Nubase evaluation of nuclear and decay properties, *Nucl. Phys.*, A624:1–122, 1997; F. DiFilippo et al., Accurate masses for fundamental metrology, *Phys. Rev. Lett.*, 73:1481–1484, 1994.

Atomic mass unit

A unit of mass equal to exactly $^1/_{12}$ of the mass of an atom of carbon-12 (^{12}C), the predominant isotope of carbon. The unit, also known as the dalton, is often abbreviated amu, and is designated by the symbol u. The relative atomic mass of a chemical element is the average mass of its atoms expressed in atomic mass units. *See* RELATIVE ATOMIC MASS.

Before 1961, two versions of the atomic mass unit were in use. The unit used by physicists was defined as $^1/_{16}$ of the mass of an atom of oxygen-16 (^{16}O), the predominant isotope of oxygen. The unit used by chemists was defined as $^1/_{16}$ of the average mass of the atoms in naturally occurring oxygen, a mixture of the isotopes ^{16}O, ^{17}O, and ^{18}O. In 1961, by international agreement, the standard based on ^{12}C superseded both these older units. It is related to them by: 1 amu (international) $\cong$ 1.000318 amu (physical) $\cong$ 1.000043 amu (chemical). *See* ATOMIC MASS.

Jonathan F. Weil

Atomic nucleus

The central region of an atom. Atoms are composed of negatively charged electrons, positively charged protons, and electrically neutral neutrons. The protons and neutrons (collectively known as nucleons) are located in a small central region known as the nucleus. The electrons move in orbits which are large in comparison with the dimensions of the nucleus itself. Protons and neutrons possess approximately equal masses, each roughly 1840 times that of an electron. The number of nucleons in a nucleus is given by the mass number A and the number of protons by the atomic number Z. Nuclear radii r are given approximately by $r = 1.2 \times 10^{-15}\ \mathrm{m}\ A^{1/3}$. *See* NUCLEAR STRUCTURE. Henry E. Duckworth

Atomic number

The number of elementary positive charges (protons) contained within the nucleus of an atom. It is denoted by the letter Z. Correspondingly, it is also the number of planetary electrons in the neutral atom.

Relation to x-ray wavelengths. The concept of atomic number emerged from the work of G. Moseley, done in 1913–1914. Using a crystal of potassium ferrocyanide in the Bragg x-ray spectrometer, he measured the wavelengths of the most energetic rays (K and L lines) produced by using the elements calcium to zinc as targets in an x-ray tube. The square root of the frequency, $\sqrt{\nu}$, of these x-rays increased by a constant amount in passing from one target to the next. These data, when extended, gave a linear plot of atomic number versus $\sqrt{\nu}$ for all elements studied, using 13 as the atomic number for aluminum and 79 for that of gold. Further, when extrapolated, this plot predicted the atomic number of hydrogen to be 1 and that of uranium to be 92. *See* X-RAY SPECTROMETRY.

Moseley's atomic numbers were quickly recognized as providing an accurate sequence of the elements, which the chemical atomic weights had sometimes failed to do. For example, the elemental pairs argon-potassium, cobalt-nickel, tellurium-iron, which had appeared in the periodic table in the reverse order of their atomic weights, were shown in their correct order by their atomic numbers. Additionally, the atomic number sequence indicated the positions of elements that had not yet been discovered. Two examples were $Z = 43$ (technetium) and $Z = 61$ (promethium); both elements are unstable and were not correctly identified until they had been produced in nuclear reactions.

Atomic and nuclear properties. The atomic number not only identifies the chemical properties of an element but facilitates the description of other aspects of atoms and nuclei. Thus, atoms with the same atomic number are isotopes and belong to the same element, while nuclear reactions may alter the atomic number: in alpha decay, $Z \rightarrow Z - 2$; in beta emission, $Z \rightarrow Z + 1$; and in positron emission or K capture, $Z \rightarrow Z - 1$. *See* ISOTOPE; RADIOACTIVITY.

When specifically written, the atomic number is placed as a subscript preceding the symbol of the element, while the mass number (A) precedes as a superscript, for example, $^{27}_{13}\mathrm{Al}$, $^{238}_{92}\mathrm{U}$. *See* ATOMIC STRUCTURE AND SPECTRA; ELEMENT (CHEMISTRY); MASS NUMBER. Henry E. Duckworth

Atomic physics

The study of the structure of the atom, its dynamical properties, including energy states, and its interactions with particles and fields. These are almost completely determined by the laws of quantum mechanics, with very refined corrections required by quantum electrodynamics. Despite the enormous complexity of most atomic systems, in which each electron interacts with both the nucleus and all the other orbiting electrons, the wavelike nature of particles, combined with the Pauli exclusion principle, results in an amazingly orderly array of atomic properties. These are systematized by the Mendeleev periodic table. In addition to their classification by chemical activity and atomic weight, the various elements of this table are characterized by a wide variety of observable properties. These include electron affinity, polarizability, angular momentum, multiple electric moments, and magnetism. *See* ATOMIC MASS; PERIODIC TABLE; QUANTUM ELECTRODYNAMICS; QUANTUM MECHANICS.

Each atomic element, normally found in its ground state (that is, with its electron configuration corresponding to the lowest state of total energy), can also exist in an infinite number of excited states. These are also ordered in accordance with relatively simple hierarchies determined by the laws of quantum mechanics. The most characteristic signature of these various excited states is the radiation emitted or absorbed when the atom undergoes a transition from one state to another. The systemization and classification of atomic energy levels (spectroscopy) has played a central role in developing an understanding of atomic structure. Control of populations of ensembles of excited atoms has led to the laser, which is itself now used to obtain even more refined information concerning atomic structure than has hitherto been possible. *See* LASER; SPECTROSCOPY.

Atomic radiation represents one of nature's most refined probes for the study of a wide range of natural phenomena, such as the effects of fluctuations in empty space, of the anomalous magnetic moment of the electron, and even subtler high-order corrections to atomic energy levels caused by quantum electrodynamics. The isolated atom is one of the most reliable systems known in nature in terms of the reproducibility of its energy levels, and is used as a primary frequency standard for time calibration. *See* ATOMIC CLOCK; FUNDAMENTAL CONSTANTS; PARITY (QUANTUM MECHANICS).

The problem of the mutual interaction of atoms or fragments (for example, electrons and ions) is still more complicated than that of the isolated atom, but such interactions are important since they govern a wide variety of practical phenomena, including atmospheric physics, laser physics, plasma generation for controlled thermonuclear plasmas, materials research (including the influence of radiation on matter), and chemical reactions and molecular formation. This area is the domain of atomic collisions. *See* SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

Finally, the study of inner shell structure and interactions of heavier atoms at very high energies has an active field of research. This is the domain of high-energy atomic physics which has led to a better understanding of the structure and the dynamics of the inner, strongly bound atomic electrons. *See* ATOM; ATOMIC STRUCTURE AND SPECTRA; ELECTRON; NUCLEAR PHYSICS. Benjamin Bederson

Atomic spectrometry

A branch of chemical analysis that seeks to determine the composition of a sample in terms of which chemical elements are present and their quantities or concentrations. Unlike other methods of elemental analysis, however, the sample is decomposed into its constituent atoms which are then probed spectroscopically.

In the ultimate atomic spectrometric procedure, a sample would be completely disintegrated into its constituent atoms. The atoms would then be sorted element by element and counted one by one. Advances in physics and chemistry have indicated that this atom-counting approach is feasible; however, the required equipment is complex and costly. Therefore, routine atomic spectrometric measurements are simpler but less elegant or sensitive.

In routine atomic spectrometry, a device called the atom source or atom cell is responsible for producing atoms from the sample; there are many different kinds of atom sources. After atomization of the sample, any of several techniques can determine which atoms are present and in what amounts, but the most common are atomic absorption, atomic emission, atomic fluorescence (the least used of these four alternatives), and mass spectrometry.

Most atomic spectrometric measurements (all those just mentioned except mass spectrometry) exploit the narrow-line spectra characteristic of gas-phase atoms. Because the atom source yields atomic species in the vapor phase, chemical bonds are disrupted, so valence electronic transitions are unperturbed by bonding effects. As a result, transitions among atomic energy levels yield narrow spectral lines, with spectral bandwidths commonly in the 1–5-picometer wavelength range. Moreover, because each atom possesses its unique set of energy levels, these narrow-band transitions can be measured individually, with little mutual interference. Thus, sodium, potassium, and scandium can all be monitored simultaneously and with minimal spectral influence on each other. This lack of spectral overlap remains one of the most attractive features of atomic spectrometry. *See* LINE SPECTRUM; SPECTRUM.

Atomic absorption spectrometry (AAS). In atomic absorption spectrometry, light from a primary source is directed through the atom cell, where a fraction of the light is absorbed by atoms from the sample. The amount of radiation that remains can then be monitored on the far side of the cell. The concentration of atoms in the path of the light beam can be determined by Beer's law, which can be expressed as the equation below, where P_0 is the light intensity

$$\log \frac{P_0}{P} = kC$$

incident on the atom cell, P is the amount of light which remains unabsorbed, C is the concentration of atoms in the cell, and k is the calibration constant, which is determined by means standard samples having known concentrations. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION.

The two most common kinds of atom cells employed in atomic absorption spectrometry are chemical flames and electrical furnaces. Chemical flames are usually simple to use, but furnaces offer higher sensitivity. The simplicity of the flame-based system results from the convenient means by which samples can be introduced. Ordinarily, a sample solution (prepared by dissolving the sample in suitable acids, solvents, or mixtures) is merely sprayed into the flame by means of a pneumatic nebulizer. To ensure that the greatest possible fraction of sample material reaching the flame is actually atomized, a spray chamber is interposed between the sprayer (nebulizer) and flame. In the spray chamber, large droplets in the aerosol settle out or are shattered, so only the finest mist reaches the flame. In the flame, the aerosol droplets are evaporated, the resulting solute (sample) particles are dried, and the dried particles are vaporized.

Unfortunately, even with this procedure the sample particles that reach the flame are seldom atomized completely. Moreover, the presence of particular sample concomitants can influence the degree of atomization of others. Such incomplete atomization often leads to effects known as matrix interferences. To overcome such interferences, so-called releasing agents are usually added to sample and standard (known-concentration) solutions so that the degree of atomization of both is as similar as possible.

Electrically heated carbon furnaces can reach temperatures comparable to those of chemical flames (2800 K or 4580°F). As a result, the furnaces also are effective but not 100% efficient at sample atomization. Thus matrix interferences are also common in furnace-based atomic absorption spectrometry.

The principal attraction of furnaces is that, unlike flames, they can generate sample atoms in a relatively still environment. Consequently, the atom cloud from even small quantities of a sample can be very concentrated, and sensitivity is very high for atomic absorption spectrometry. While the flame-based technique is capable of measuring sample concentrations down to 100 nanograms/ml (parts per billion), furnace atomic absorption spectrometry offers detection limits on the order of 0.1 ng/ml, a thousandfold gain.

Another attribute of furnace atomization is the ability to handle microliter quantities of sample solution. While flames must usually be fed continuously with sample solutions at flow rates of 1–5 ml/min, furnaces operate satisfactorily with total sample volumes of 10–50 microliters. Because the total volume of required sample solution is small, furnaces can measure extremely low quantities of sample mass. Typical detection limits for many elements by furnace atomic absorption spectrometry are 10^{-13} g.

The most common primary light source employed in atomic absorption spectrometry is the hollow-cathode lamp. Conveniently, the hollow-cathode lamp emits an extremely narrow line spectrum of one, two, or three elements of interest. As a result, the atomic absorption spectrometry measurement is automatically tuned to the particular spectral lines of interest. This attribute simplifies the instrumentation and measurement procedure substantially. However, because an effective hollow-cathode lamp can be made to emit the narrow-line spectra of only one to three chemical elements, determining additional elements requires several lamps. Although several schemes for the automatic switching of lamps have been proposed and are commercially available, the inconvenience of lamp exchanging has kept atomic absorption spectrometry from becoming a truly simultaneous multielement technique. Instead, the method is usually employed for quantitative analysis of specific elements in a sample.

Although spectral interferences among elements are rare in atomic absorption spectrometry, a problem known as nonspecific absorption is fairly common. Regardless of whether a sample is atomized in a flame or furnace, it might generate molecular fragments or smoke. Understandably, furnaces cause greater problems in this regard than do flames. These molecular fragments and smoke can attenuate the light from the hollow-cathode lamp in a manner which is indistinguishable from atomic absorption. Consequently, the presence of nonspecific absorption artificially inflates the apparent concentration of the element in a sample. To overcome this problem, a number of background-correction methods are available on commercial instruments. They all operate on the principle that light scattering by smoke and absorption by molecular fragments both occur over a relatively broad spectral range, compared with atomic absorption. Comparing the broad-band absorption of an atomized sample with the narrow-band absorption of the light from a hollow-cathode lamp permits the contribution from atomic absorption spectrometry to be unraveled from that of nonspecific absorption.

Atomic emission spectrometry (AES). In this type of spectrometry, atomic species are measured by their emission spectra. For such spectra to be produced, the atoms must first be excited by thermal or nonthermal means. Therefore, the atom sources employed in atomic emission spectrometry are hotter or more energetic than those commonly used in atomic absorption spectrometry. Although several such sources are in common use, the dominant one is the inductively coupled plasma. From the simplest standpoint, the inductively coupled plasma is a flowing stream of hot, partially ionized (positively charged) argon. Power is coupled into the plasma by means of an induction coil. At a typical operating frequency of 27–40 MHz and an input power of 1–1.5 kW, the inductively coupled plasma achieves temperatures between 6000 and 8000 K (10,000 and 14,000°F). Such an elevated temperature dissociates sample components even more effectively than in a chemical flame, so the inductively coupled plasma does not suffer the serious matrix effects that are commonly encountered in flame atomic absorption spectrometry. Furthermore, atoms produced in the inductively coupled plasma are efficiently excited at the plasma's high temperature. Detection limits are accordingly low and are on the order of 0.1–1 ng/ml for most elements. Importantly, emission occurs from all elements simultaneously in the inductively coupled plasma, so that either rapid sequential or truly simultaneous multielement analysis is possible.

There are two common modes for observing emission spectra from an inductively coupled plasma. The less expensive and more flexible approach employs a so-called slew-scan spectrometer. With this instrument, spectral lines are accessed in rapid sequence, so that a number of chemical elements can be measured rapidly, one after the other. Moreover, because each viewed elemental spectral line can be scanned completely, it is possible to subtract spectral emission background independently for each element. Spectral interferences are therefore reduced somewhat.

The alternative approach is to view all spectral lines simultaneously, either with a number of individual photodetectors keyed to particular spectral lines or with a truly multichannel electronic detector driven by a computer. Although this alternative approach is somewhat less flexible than the slew-scan system, it enables samples to be analyzed more rapidly and permits transient atom signals (as from a furnace-based atomizer) to be recorded.

Because the inductively coupled plasma produces fewer severe interelement interferences, exhibits better detection limits, and permits analysis of a greater number of samples per unit time, it has assumed much of the workload that was formerly assigned to atomic absorption spectrometry. Nonetheless, inductively coupled plasma atomic emission is not without its shortcomings. For example, emission spectra from individual elements in the inductively coupled plasma are far more complex than are atomic absorption spectra. Also, the spectral background from the inductively coupled plasma consists not only of emission lines of argon (the plasma support gas) but also of a continuum (extremely broad-band) background in addition to emission bands from molecular fragments introduced with the sample or by entrainment of atmospheric gases. Therefore, spectral interferences in inductively coupled plasma atomic emission are far more serious than are those in atomic absorption spectrometry.

Commercial instrument manufacturers have sought to overcome as many of these difficulties as possible. Most instruments using inductively coupled plasmas are computer-controlled, and many include programs with menus that permit the judicious selection of those spectral lines that are most likely to be free of spectral interference. Also, most instruments are equipped with spectral scanning features, whether they be of the slew-scan or direct-reading type. These features permit the background beneath each spectral line to be measured and the net atomic signal deduced.

Another limitation of inductively coupled plasma atomic emission measurements is sensitivity. Although the technique is capable of measuring lower concentrations than flame atomic absorption spectrometry, it does not yet possess the sensitivity that is achieved with furnace atomic absorption spectrometry. Moreover, except for specialized designs, most inductively coupled plasma units are not capable of measuring microliter quantities of sample solution. Finally, most instruments using inductively coupled plasmas are still somewhat more complex to operate than comparably priced units designed for atomic absorption spectromerty. *See* EMISSION SPECTROCHEMICAL ANALYSIS.

Mass spectrometry (MS). Elemental mass spectrometry has been practiced for many years in the form of spark-source mass spectrometry and, more recently, glow-discharge-lamp mass spectrometry. However, a hybrid technique that combines the inductively coupled plasma with a mass spectrometer (ICP-MS) has assumed a prominent place.

At the high temperatures present in an inductively coupled plasma, many atomic species occur in an ionic form. These ions can be readily extracted into a mass spectrometer via an interface that separates the atmospheric-pressure plasma and the mass spectrometer. In a typical configuration, a tapered, water-cooled metallic cone is inserted into the inductively coupled plasma. A small orifice (typically 0.04 in. or 1 mm in diameter) in the cone then allows the passage of ions from the plasma into a reduced-pressure zone held ordinarily at pressures near 1 torr (133 pascals). The ion beam generated in this differentially pumped region is then passed to one or more successively lower-pressure zones until it is extracted into a mass spectrometer, where ions are separated according to their mass. The mass spectrometer region is usually held at pressures near 10^{-5} torr (1.33×10^{-3} Pa).

In commercial instrumentation combining an inductively coupled plasma and mass spectrometry, a quadrupole mass spectrometer is employed. In this type of system, ions of different elements are separated sequentially according to mass. The separated ions then pass one after the other into a detector where they can be measured electronically. Differences in atomic mass make it possible to identify elements; the intensity of their signals then allows determination of their concentrations.

The advantages of the combination of inductively coupled plasma and mass spectrometry are substantial. The system is capable of some of the best detection limits in atomic spectrometry, typically 10^{-3} to 10^{-2} ng/ml for most elements. Also, virtually all elements in the periodic table can be determined during a single scan. The method is also capable of providing isotopic information, unavailable by any other atomic spectrometric method for such a broad range of elements. Finally, elemental mass spectra are relatively simple; there are very few naturally occurring isotopes for each element and also virtually no mono-isotopic elements that suffer severe spectral overlap with isotopes of other elements.

Use of this hybrid technique does present some disadvantages. A substantial number of diatomic and cluster ions are often observed in spectra produced by the combination of inductively coupled plasma and mass spectrometry. They originate from a combination of sample-borne species, from argon, from atmospheric gases, and from solvent vapor. These polyatomic ions often overlap with masses of desired elements. It has been found that the incidence of interfering ions can be reduced through use of aqueous or nitric acid solutions, but some problems remain.

From a practical standpoint, widespread use of the inductively coupled plasma combined with mass spectrometry is limited largely by the cost of instrumentation, which is between two and five times that of competing atomic absorption spectrometry and inductively coupled plasma atomic emission instruments. Also, few standard procedures have been developed, and a rather intensive effort would be needed for the development of methods for the determination of most new samples. Nonetheless, the extremely high sensitivity that is provided by the combination of inductively coupled plasma and mass spectrometry with its isotopic selectivity and relatively easily interpreted spectra suggests that it is likely to achieve increased prominence in the future. *See* MASS SPECTROMETRY.

Gary M. Hieftje

Bibliography. P. W. M. J. Boumans (ed.), *Inductively Coupled Plasma Emission Spectroscopy*, pt. I: *Methodology, Instrumentation, and Performance*, pt. II: *Applications and Fundamentals*, 1987; J. D. Ingle and S. R. Crouch, *Spectrochemical Analysis*, 1988; A. Montaser and D. W. Golightly, *Inductively Coupled Plasmas in Analytical Atomic Spectrometry*, 2d ed., 1992; D. G. Peters, J. M. Hayes, and G. M. Hieftje, *Chemical Separations and Measurements*, 1974; W. J. Price, *Spectrochemical Analysis by Atomic Absorption*, 1985.

Atomic structure and spectra

The idea that matter is subdivided into discrete building blocks called atoms, which are not divisible any further, dates back to the Greek philosopher Democritus. His teachings of the fifth century B.C. are commonly accepted as the earliest authenticated ones concerning what has come to be called atomism by students of Greek philosophy. The weaving of the philosophical thread of atomism into the analytical fabric of physics began in the late eighteenth and the nineteenth centuries. Robert Boyle is generally credited with introducing the concept of chemical elements, the irreducible units which are now recognized as individual atoms of a given element. In the early nineteenth century John Dalton developed his atomic theory, which postulated that matter consists of indivisible atoms as the irreducible units of Boyle's elements, that each atom of a given element has identical attributes, that differences among elements are due to fundamental differences among their constituent atoms, that chemical reactions proceed by simple rearrangement of indestructible atoms, and that chemical compounds consist of molecules which are reasonably stable aggregates of such indestructible atoms. *See* CHEMISTRY.

Electromagnetic nature of atoms. The work of J. J. Thomson in 1897 clearly demonstrated that atoms are electromagnetically constituted and that from them can be extracted fundamental material units bearing electric charge that are now called electrons. These pointlike charges have a mass of 9.109×10^{-31} kg and a negative electric charge of magnitude 1.602×10^{-19} coulomb. The electrons of an atom account for a negligible fraction of its mass. By virtue of overall electrical neutrality of every atom, the mass must therefore reside in a compensating, positively charged atomic component of equal charge magnitude but vastly greater mass. *See* ELECTRON.

Thomson's work was followed by the demonstration by Ernest Rutherford in 1911 that nearly all the mass and all of the positive electric charge of an atom are concentrated in a small nuclear core approximately 10,000 times smaller in extent than an atomic diameter. Niels Bohr in 1913 and others carried out some remarkably successful attempts to build solar system models of atoms containing planetary pointlike electrons orbiting around a positive core through mutual electrical attraction (though only certain "quantized" orbits were "permitted"). These models were ultimately superseded by nonparticulate, matter-wave quantum theories of both electrons and atomic nuclei. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

The modern picture of condensed matter (such as solid crystals) consists of an aggregate of atoms or molecules which respond to each other's proximity through attractive electrical interactions at separation distances of the order of 1 atomic diameter (approximately 10^{-10} m) and repulsive electrical interactions at much smaller distances. These interactions are mediated by the electrons, which are in some sense shared and exchanged by all atoms of a particular sample, and serve as an interatomic glue that binds the mutually repulsive, heavy, positively charged atomic cores together. *See* SOLID-STATE PHYSICS.

Planetary atomic models. Fundamental to any planetary atomic model is a description of the forces which give rise to the attraction and repulsion of the constituents of an atom. Coulomb's law describes the fundamental interaction which, to a good approximation, is still the basis of modern theories of atomic structure and spectra: the force exerted by one charge on another is repulsive for charges of like sign and attractive for charges of unlike sign, is proportional to the magnitude of each electric charge, and diminishes as the square of the distance separating the charges. *See* COULOMB'S LAW.

Also fundamental to any planetary model is Thomson's discovery that electrically neutral atoms in some sense contain individual electrons whose charge is restricted to a unique quantized value. Moreover, Thomson's work suggested that nuclear charges are precisely equally but oppositely quantized to offset the sum of the constituent atomic electron charges. Boyle's elements, of which over a hundred have been discovered, may then be individually labeled by the number of quantized positive-charge units Z (the atomic number) residing within the atomic nucleus (unit charge $= +1.602 \times 10^{-19}$ coulomb). Each nucleus is surrounded by a complement of Z electrons (with charges of -1.602×10^{-19} C each) to produce overall charge neutrality. Molecules containing two, three, . . . , atoms can then be thought of as binary, ternary, . . . , planetary systems consisting of heavy central bodies of atomic numbers Z_1, Z_2, . . . , sharing a supply of $Z_1 + Z_2 + \cdots$ electrons having the freedom to circulate throughout the aggregate and bond it together as would an electronic glue.

Atomic sizes. All atoms, whatever their atomic number Z, have roughly the same diameter (about 10^{-10} m), and molecular sizes as well as macroscopic solid sample sizes tend to be the sum of the aggregate atomic sizes. It is easy to qualitatively, though only partially, account for this circumstance by using Coulomb's law. The innermost electrons of atoms orbit at small radii because of intense electrical attractions prevailing at small distances. Because electrons are more than 2000 times lighter than most nuclei and therefore move less sluggishly, the rapid orbital motion of inner electrons tends to clothe the slow-moving positive nuclear charge in a negative-charge cloud, which viewed from outside the cloud masks this positive nuclear charge to an ever-increasing extent. Thus, as intermediate and outer electrons are added, each experiences a diminishing attraction. The Zth, and last, electron sees $+Z$ electronic charge units within the nucleus compensated by about $(Z - 1)$ electronic charges in a surrounding cloud, so that in first approximation the Zth electron orbits as it would about a bare proton having $Z = 1$, at about the same radius. Crudely speaking, an atom of any atomic number Z thus has a size similar to that of the hydrogen atom. When the molecular aggregates of atoms of concern to chemistry are formed, they generally do so through extensive sharing or exchanging of a small number (often just one) of the outermost electrons of each atom. Hence, the interatomic spacing in both molecules and solids tends to be of the order of one to a very few atomic diameters.

Chemical reactions. In a microscopic planetary atomic model, chemical reactions in which rearrangements result in new molecular aggregates of atoms can be viewed as electrical phenomena, mediated by the electrical interactions of outer electrons. The reactions proceed by virtue of changing positions and shapes of the orbits of the binding electrons, either through their internal electrical interactions or through their being electrically "bumped" from outside (as by collisions with nearby molecules).

Difficulties with the models. Before continuing with a detailed account of the successes of detailed planetary atomic models, it is advisable to anticipate some severe difficulties associated with them. Some of these difficulties played a prominent role in the development of quantum theory, and others present as yet unsolved and profound challenges to classical as well as modern quantum theories of physics.

A classical planetary theory fails to account for several atomic enigmas. First, it is known that electrons can be made to execute fast circular orbits in large accelerating machines, of which modern synchrotrons are excellent examples. As they are then in accelerated motion, they radiate light energy, as predicted by Maxwell's theory, at a frequency equal to that of their orbital circular motions. Thus, a classical planetary theory fails to explain how atoms can be stable and why atomic electrons do not continuously radiate light, falling into the nucleus as they do so. Such a theory also does not account for the observation that light emitted by atomic electrons appears to contain only quantized frequencies or wavelengths. Furthermore, a classical planetary theory would lead one to expect that all atomic electrons of an atom would orbit very close to their parent nuclei at very small radii, instead of distributing themselves in shells of increasing orbital radius and thickness, which seem able to accommodate only small, fixed numbers of electrons per shell.

Any theory of atomic structure must deal with the question of whether the atom is mostly empty space, as a planetary system picture suggests, or whether the entire atomic volume is filled with electronic charges in smeared out, cloudlike fashion. Such a theory must also be concerned with whether electrons and nuclei are pointlike, structureless, and indivisible, or whether they can be subdivided further into smaller constituents, in analogy to the way in which Thomson was able to extract electrons as constituents of atoms. Questions which still have not received definitive answers concern how much energy is needed to construct electrons and nuclei, what their radii are, and why they do not fly apart under the explosive repulsion of half their charge distribution for the other half. A more advanced theory than the classical planetary model is also required to determine whether electrons really interact with each other and with nuclei instantaneously, as Coulomb's law would have it, or whether there is a finite interaction delay time. Finally, a fundamental question, which can be answered only by quantum theory, is concerned with whether electrons and nuclei behave as pointlike objects, as light does when it gives rise to sharp shadows, or whether they behave as extended waves which can exhibit diffraction (bending) phenomena, as when light or water waves bend past sharp edges and partially penetrate regions of shadow in undulatory intensity patterns.

Scattering experiments. A key experimental technique in answering many important questions, such as those just posed, is that of scattering, in which small, high-speed probe projectiles are used to interact with localized regions of more extended objects from which they rebound, or scatter. A study of the number of scattered projectiles, and their distributions in speed and angle, gives much information about the structure of target systems, and the internal distributions, speeds, and concentrations of constituent bodies or substructures. *See* SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

The scattering of x-rays by solids, for example, was used by C. Barkla to indicate that the number of electrons in an atom is approximately half the atomic weight. The mechanism of the production of the secondary (scattered) x-rays is indicated in **Fig. 1*a***. X-rays are electromagnetic waves of wavelength considerably smaller than the size of the atom. If an x-ray sets an atomic electron into vibration, there follows the emission of a wave of lesser amplitude which can be observed at directions outside the incident beam. *See* X-RAYS.

Rutherford's experiments on the scattering of alpha particles represented an important step in understanding atomic structure. Alpha particles are helium ($Z = 2$) nuclei emitted at high velocities by some radioactive materials. A beam of these particles was directed at thin foils of different metals, and the relative numbers scattered at various angles were observed. While most of the particles passed through the foil with small deflections, such as the lower particle in Fig. 1*b*, a considerable number of very large deflections occurred. The upper particle in Fig. 1*b* has undergone such a deflection. The precise results could be explained only if the positive electric charge of the atom is concentrated in the very small volume of the nucleus, which also contains almost all of the atom's mass. The diameter of the nucleus, found to depend on the atomic mass, was about 10^{-14} m for heavy nuclei. The nuclear charge was found to be the atomic number Z times the magnitude of the electron's charge. *See* ALPHA PARTICLES.

The results of the scattering experiments therefore established the model of the atom as consisting of a small, massive, positively charged nucleus surrounded by a cloud of electrons to provide

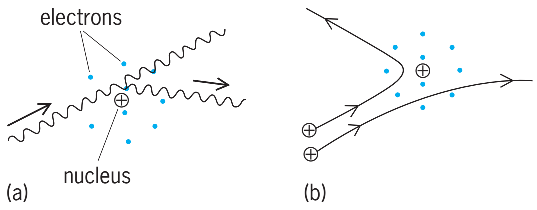


Fig. 1. Scattering by an atom (*a*) of x-rays, (*b*) of alpha particles.

electrical neutrality. As noted above, such an atom should quickly collapse. The first step toward an explanation as to why it did not, came from the Bohr picture of the atom.

Bohr Atom

The hydrogen atom is the simplest atom, and its spectrum (or pattern of light frequencies emitted) is also the simplest. The regularity of its spectrum had defied explanation until Bohr solved it with three postulates, these representing a model which is useful, but quite insufficient, for understanding the atom.

Postulate 1: The force that holds the electron to the nucleus is the Coulomb force between electrically charged bodies.

Postulate 2: Only certain stable, nonradiating orbits for the electron's motion are possible, those for which the angular momentum associated with the motion of an electron in its orbit is an integral multiple of $h/2\pi$ (Bohr's quantum condition on the orbital angular momentum). Each stable orbit represents a discrete energy state.

Postulate 3: Emission or absorption of light occurs when the electron makes a transition from one stable orbit to another, and the frequency ν of the light is such that the difference in the orbital energies equals $h\nu$ (A. Einstein's frequency condition for the photon, the quantum of light). The description of these apparently nearly instantaneous quantum jumps remains a subject of research, as discussed below.

Here the concept of angular momentum, a continuous measure of rotational motion in classical physics, has been asserted to have a discrete quantum behavior, so that its quantized size is related to Planck's constant h, a universal constant of nature. The orbital angular momentum of a point object of mass m and velocity v, in rotational motion about a central body, is defined as the product of the component of linear momentum mv (expressing the inertial motion of the body) tangential to the orbit times the distance to the center of rotation. *See* ANGULAR MOMENTUM.

Modern quantum mechanics has provided justification of Bohr's quantum condition on the orbital angular momentum. It has also shown that the concept of definite orbits cannot be retained except in the limiting case of very large orbits. In this limit, the frequency, intensity, and polarization can be accurately calculated by applying the classical laws of electrodynamics to the radiation from the orbiting electron. This fact illustrates Bohr's correspondence principle, according to which the quantum results must agree with the classical ones for large dimensions. The deviation from classical theory that occurs when the orbits are smaller than the limiting case is such that one may no longer picture an accurately defined orbit. Bohr's other hypotheses are still valid.

Quantization of hydrogen atom. According to Bohr's theory, the energies of the hydrogen atom are quantized (that is, can take on only certain discrete values). These energies can be calculated from the electron orbits permitted by the quantized orbital angular momentum. The orbit may be circular or elliptical, so only the circular orbit is considered here for simplicity. Let the electron, of mass m and electric charge $-e$, describe a circular orbit of radius r around a nucleus of charge $+e$ and of infinite mass. With the electron velocity v, the angular momentum is mvr, and the second postulatebecomes Eq. (1).

$$mvr = n\left(\frac{h}{2\pi}\right) \qquad (n = 1, 2, 3, \ldots) \qquad (1)$$

The integer n is called the principal quantum number. The centripetal force required to hold the electron in its orbit is the electrostatic force described by Coulomb's law, as shown in Eq. (2). Here ϵ_0 is

$$\frac{mv^2}{r} = \frac{e^2}{4\pi\epsilon_0 r^2} \qquad (2)$$

the permittivity of free space, a constant included in order to give the correct units to the statement of Coulomb's law in SI units. *See* QUANTUM NUMBERS.

The energy of an electron in an orbit consists of both kinetic and potential energies. For these circular orbits, the potential energy is twice as large as the kinetic energy and has a negative sign, where the potential energy is taken as zero when the electron and nucleus are at rest and separated by a very large distance. The total energy, which is the sum of kinetic and potential energies, is given in Eq. (3).

$$E = \frac{mv^2}{2} - mv^2 = -\frac{mv^2}{2} \qquad (3)$$

The negative sign means that the electron is bound to the nucleus and energy must be provided to separate them.

It is possible to eliminate v and r from these three equations. The result is that the possible energies of the nonradiating states of the atom are given by Eq. (4).

$$E = -\frac{me^4}{8\epsilon_0^2 h^2} \cdot \frac{1}{n^2} \qquad (4)$$

The same equation for the hydrogen atom's energy levels, except for some small but significant corrections, is obtained from the solution of the Schrödinger equation, as modified by W. Pauli, for the hydrogen atom. *See* SCHRÖDINGER'S WAVE EQUATION.

The frequencies of electromagnetic radiation or light emitted or absorbed in transitions are given by Eq. (5) where E' and E'' are the energies of the initial

$$\nu = \frac{E' - E''}{h} \qquad (5)$$

and final states of the atom. Spectroscopists usually express their measurements in wavelength λ or in wave number σ in order to obtain numbers of a convenient size. The frequency of a light wave can be thought of as the number of complete waves radiated per second of elapsed time. If each wave travels at a fixed velocity c (approximately 3×10^8 m/s in vacuum), then after t seconds, the distance traveled by the first wave in a train of waves is ct, the number

-s

of waves in the train is νt, and hence the length of each must be $ct/(\nu t) = c/\nu = \lambda$. The wave number, defined as the reciprocal of the wavelength, therefore equals ν/c. The quantization of energy $h\nu$ and of angular momentum in units of $h/(2\pi)$ does not materially alter this picture. The wave number of a transition is shown in Eq. (6). If $T = -E/(hc)$, then Eq. (7) results. Here T is called the spectral term.

$$\sigma = \frac{\nu}{c} = \frac{E'}{hc} - \frac{E''}{hc} \tag{6}$$

$$\sigma = T'' - T' \tag{7}$$

The allowed terms for hydrogen, from Eq. (4), are given by Eq. (8). The quantity R is the important

$$T = \frac{me^4}{8\epsilon_0^2 ch^3} \cdot \frac{1}{n^2} = \frac{R}{n^2} \tag{8}$$

Rydberg constant. Its value, which has been measured to a remarkable and rapidly improving accuracy, is related to the values of other well-known atomic constants, as in Eq. (8). *See* RYDBERG CONSTANT.

The effect of finite nuclear mass must be considered, since the nucleus does not actually remain at rest at the center of the atom. Instead, the electron and nucleus revolve about their common center of mass. This effect can be accurately accounted for and requires a small change in the value of the effective mass m in Eq. (8). The mass effect was first detected by comparing the spectrum of hydrogen with that of singly ionized helium, which is like hydrogen in having a single electron orbiting the nucleus. For this isoelectronic case, the factor Z^2 must be included in the numerator of Eqs. (2) and (8) to account for the greater nuclear charge. The mass effect was used by H. Urey to discover deuterium, one of three hydrogen isotopes, by the shift of lines in its spectrum because of the very small change in its Rydberg constant. *See* ISOELECTRONIC SEQUENCE; ISOTOPE SHIFT.

Elliptical orbits. In addition to the circular orbits already described, elliptical ones are also consistent with the requirement that the angular momentum be quantized. A. Sommerfeld showed that for each value of n there is a family of n permitted elliptical orbits, all having the same major axis but with different eccentricities. **Figure 2*a*** shows, for example, the Bohr-Sommerfeld orbits for $n = 3$. The orbits are labeled s, p, and d, indicating values of the azimuthal quantum number $l = 0$, 1, and 2. This number determines the shape of the orbit, since the ratio of the major to the minor axis is found to be $n/(l = 1)$. To a first approximation, the energies of all orbits of the same n are equal. In the case of the highly eccentric orbits, however, there is a slight lowering of the energy due to precession of the orbit (Fig. 2*b*). According to Einstein's theory of relativity, the mass increases somewhat in the inner part of the orbit, because of greater velocity. The velocity increase is greater as the eccentricity is greater, so the orbits of higher eccentricity have their energies lowered more. The quantity l is called the orbital angular momentum quantum number or the azimuthal quantum number. *See* RELATIVITY.

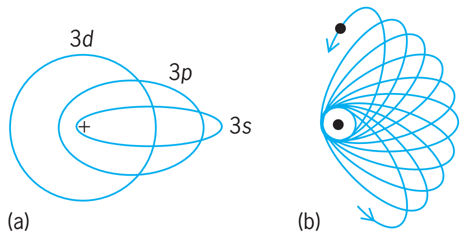


Fig. 2. Possible elliptical orbits, according to the Bohr-Sommerfeld theory. (*a*) The three permitted orbits for $n = 3$. (*b*) Precession of the 3*s* orbit caused by the relativistic variation of mass. (*After A. P. Arya, Fundamentals of Atomic Physics, Allyn and Bacon, 1971*)

A selection rule limits the possible changes of l that give rise to spectrum lines (transitions of fixed frequency or wavelength) of appreciable intensity. The rule is that l may increase or decrease only by one unit. This is usually written as $\Delta l = \pm 1$ for an allowed transition. Transitions for which selection rules are not satisfied are called forbidden; these tend to have quite low intensities. The quantum number n may change by any amount, although the probability of large changes in n is generally low. Selection rules for hydrogen may be derived from Bohr's correspondence principle. However, the selection rules, as well as the relation between n and l, arise much more naturally in quantum mechanics. *See* SELECTION RULES (PHYSICS).

Multielectron Atoms

In attempting to extend Bohr's model to atoms with more than one electron, it is logical to compare the experimentally observed terms of the alkali atoms, which contain only a single electron outside closed shells, with those of hydrogen. A definite similarity is found but with the striking difference that all terms with $l > 0$ are double. This fact was interpreted by S. A. Goudsmit and G. E. Uhlenbeck as due to the presence of an additional angular momentum of $^1/_2(h/2\pi)$ attributed to the electron spinning about its axis. The spin quantum number of the electron is $s = {}^1/_2$.

The relativistic quantum mechanics developed by P. A. M. Dirac provided the theoretical basis for this experimental observation. *See* ELECTRON SPIN.

Exclusion principle. Implicit in much of the following discussion is W. Pauli's exclusion principle, first enunciated in 1925, which when applied to atoms may be stated as follows: no more than one electron in a multielectron atom can possess precisely the same quantum numbers. In an independent, hydrogenic electron approximation to multielectron atoms, there are $2n^2$ possible independent choices of the principal (n), orbital (l), and magnetic (m_l, m_s) quantum numbers available for electrons belonging to a given n, and no more. Here m_l and m_s refer to the quantized projections of l and s along some chosen direction. The organization of atomic electrons into shells of increasing radius (the Bohr radius scales as n^2) follows from this principle, answering the question as to why all electrons of a heavy atom do not collapse into the most tightly bound orbits. *See* EXCLUSION PRINCIPLE.

Examples are: helium ($Z = 2$), two $n = 1$ electrons; neon ($Z = 10$), two $n = 1$ electrons, eight $n = 2$ electrons; argon ($Z = 18$), two $n = 1$ electrons, eight $n = 2$ electrons, eight $n = 3$ electrons. Actually, in elements of Z greater than 18, the $n = 3$ shell could in principle accommodate 10 more electrons but, for detailed reasons of binding energy economy rather than fundamental symmetry, contains full $3s$ and $3p$ shells for a total of eight $n = 3$ electrons, but often a partially empty $3d$ shell. The most chemically active elements, the alkalies, are those with just one outer orbital electron in an n state, one unit above that of a completely full shell or subshell. *See* ELECTRON CONFIGURATION.

Spin-orbit coupling. This is the name given to the energy of interaction of the electron's spin with its orbital angular momentum. The origin of this energy is magnetic.

A charge in motion through either "pure" electric or "pure" magnetic fields, that is, through fields perceived as "pure" in a static laboratory, actually experiences a combination of electric and magnetic fields, if viewed in the frame of reference of a moving observer with respect to whom the charge is momentarily at rest. For example, moving charges are well known to be deflected by magnetic fields. But in the rest frame of such a charge, there is no motion, and any acceleration of a charge must be due to the presence of a pure electric field from the point of view of an observer analyzing the motion in that reference frame. *See* RELATIVISTIC ELECTRODYNAMICS.

A spinning electron can crudely be pictured as a spinning ball of charge, imitating a circulating electric current (though Dirac electron theory assumes no finite electron radius—classical pictures fail). This circulating current gives rise to a magnetic field distribution very similar to that of a small bar magnet, with north and south magnetic poles symmetrically distributed along the spin axis above and below the spin equator. This representative bar magnet can interact with external magnetic fields, one source of which is the magnetic field experienced by an electron in its rest frame, owing to its orbital motion through the electric field established by the central nucleus of an atom. In multielectron atoms, there can be additional, though generally weaker, interactions arising from the magnetic interactions of each electron with its neighbors, as all are moving with respect to each other and all have spin. The strength of the bar magnet equivalent to each electron spin, and its direction in space are characterized by a quantity called the magnetic moment, which also is quantized essentially because the spin itself is quantized. Studies of the effect of an external magnetic field on the states of atoms show that the magnetic moment associated with the electron spin is equal in magnitude to a unit called the Bohr magneton. *See* MAGNETISM.

The energy of the interaction between the electron's magnetic moment and the magnetic field generated by its orbital motion is usually a small correction to the spectral term, and depends on the angle between the magnetic moment and the magnetic field or, equivalently, between the spin angular momentum vector and the orbital angular momentum vector (a vector perpendicular to the orbital plane whose magnitude is the size of the orbital angular momentum). Since quantum theory requires that the quantum number j of the electron's total angular momentum shall take values differing by integers, while l is always an integer, there are only two possible orientations for s relative to l: s must be either parallel or antiparallel to l. (This statement is convenient but not quite accurate. Actually, orbital angular momentum is a vector quantity represented by the quantum number l and of magnitude $\sqrt{l(l = 1)} \cdot (h/2\pi)$. There are similar relations for spin and total angular momentum. These statements all being true simultaneously, the spin vector cannot ever be exactly parallel to the orbital angular momentum vector. Only the quantum numbers themselves can be described as stated.) **Figure 3*a*** shows the relative orientations of these two vectors and of their resultant j for a p electron (one for which $l = 1$). The corresponding spectral term designations are shown adjacent to the vector diagrams, labeled with the customary spectroscopic notation, to be explained later.

For the case of a single electron outside the nucleus, the Dirac theory gives Eq. (9) for the spin-orbit correction to the spectral terms.

$$\Delta T = \frac{R\alpha^2 Z^4}{n^3} \times \frac{j(j+1) - l(l+1) - s(s+1)}{l(2l+1)(l+1)} \qquad (9)$$

Here $\alpha = e^2/(2\epsilon_0 hc) \cong 1/137$ is called the fine structure constant. The fine structure splitting predicted by Eq. (9)

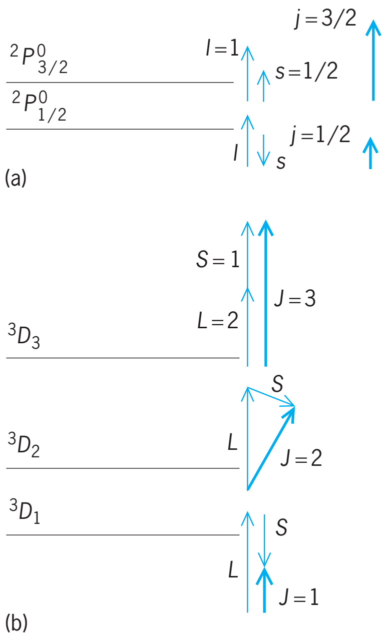


Fig. 3. Vector model for spectral terms arising from (*a*) a single *p* electron, and (*b*) two electrons, either two *p* electrons, or an *s* and a *d* electron.

is present in hydrogen, although its observation requires instruments of very high precision. A relativistic correction must be added.

In atoms having more than one electron, this fine structure becomes what is called the multiplet structure. The doublets in the alkali spectra, for example, are due to spin-orbit coupling; Eq. (9), with suitable modifications, can still be applied. These modifications may be attributed to penetration of the outer orbital electron within the closed shells of other electrons.

The various states of the atom are described by giving the quantum numbers n and l for each electron (the configuration) as well as the set of quantum numbers which depend on the manner in which the electrons interact with each other.

Coupling schemes. When more than one electron is present in the atom, there are various ways in which the spins and orbital angular momenta can interact. Each spin may couple to its own orbit, as in the one-electron case; other possibilities are orbit–other orbit, spin-spin, and so on. The most common interaction in the light atoms, called *LS* coupling or Russell-Saunders coupling, is described schematically in Eq. (10). This notation indicates that the l_i are

$$\{(l_1, l_2, l_3, \ldots)(s_1, s_2, s_3, \ldots)\} = \{L, S\} = J \tag{10}$$

coupled strongly together to form a resultant L, representing the total orbital angular momentum. The s_i are coupled strongly together to form a resultant S, the total spin angular momentum. The weakest coupling is that between L and S to form J, the total angular momentum of the electron system of the atom in this state. Suppose, for example, it is necessary to calculate the terms arising from a p $(l = 1)$ and a d $(l = 2)$ electron. The quantum numbers L, S, and J are never negative, so the only possible values of L are 1, 2, and 3. States with these values of L are designated P, D, and F. This notation, according to which $L = 0, 1, 2, 3, 4, 5$, etc. correspond to S, P, D, F, G, H, etc. terms, survives from the early empirical designation of series of lines as sharp, principal, diffuse, and so on.

The spin $s = {}^1/_2$ for each electron, always, so the total spin $S = 0$ or $S = 1$ in this case. Consider only the term with $S = 1$ and $L = 2$. The coupling of L and S gives $J = 1$, 2, and 3, as shown in Fig. 3*b*. These three values of J correspond to a triplet of different energy levels which lie rather close together, since the *LS* interaction is relatively weak. It is convenient to use the notation ^3D for this multiplet of levels, the individual levels being indicated with J values as a subscript, as $^3D_1$, $^3D_2$, and $^3D_3$. The superscript has the value $2S + 1$, a quantity called the multiplicity. There will also be a term with $S = 0$ and $L = 2$ and the single value of $J = 2$. This is a singlet level, written as ^1D or $^1D_2$. The entire group of terms arising from a p and d electron includes ^1P, ^3P, ^1D, ^3D, ^1F, ^3F, with all the values of J possible for each term.

The ^2P state shown in Fig. 3*a* is derived from a single electron which, since $L = l = 1$ and $S = s = {}^1/_2$, has J values of $^1/_2$ and $^1/_2$, forming a doublet. If there are three or more electrons, the number of possible terms becomes very large, but they are easily derived by a similar procedure. The resulting multiplicities are shown in the **table**. If two or more electrons are equivalent, that is, have the same n and l, the number of resulting terms is greatly reduced, because of the requirement of the Pauli exclusion principle that no two electrons in an atom may have all their quantum numbers alike. Two equivalent p electrons, for example, give only ^1S, ^3P, and ^1D terms, instead of the six terms ^1S, ^3S, ^1P, ^3P, ^1D, and ^3D possible if the electrons are nonequivalent. The exclusion principle applied to equivalent electrons explains the observation of filled shells in atoms, where $L = S = J = 0$.

Possible multiplicities with different numbers of electrons

Number of electrons:	1	2	3	4
Values of S:	$^1/_2$	1, 0	$^3/_2$, $^1/_2$	2, 1, 0
Multiplicities	Doublets	Singlets Triplets	Doublets Quartets	Singlets Triplets Quintets

Coupling of the *LS* type is generally applicable to the low-energy states of the lighter atoms. The next commonest type is called *jj* coupling, represented in Eq. (11). Each electron has its spin coupled to its

$$\{(l_1, s_1)(l_2, s_2)(l_3, s_3)\ldots\} = \{j_1, j_2, j_3, \ldots\} = J \tag{11}$$

own orbital angular momentum to form a j_i for that electron. The various j_i are then more weakly coupled together to give J. This type of coupling is seldom strictly observed. In the heavier atoms it is common to find a condition intermediate between *LS* and *jj* coupling; then either the *LS* or *jj* notation may be used to describe the levels, because the number of levels for a given electron configuration is independent of the coupling scheme.

Spectrum of Hydrogen

Figure 4*a* shows the terms R/n^2 in the spectrum of hydrogen resulting from the simple Bohr theory. These terms are obtained by dividing the numerical value of the Rydberg constant for hydrogen (109,678 cm^{-1}) by n^2, that is, by 1, 4, 9, 16, etc. The equivalent energies in electronvolts may then be found by using the conversion factor 1 eV = 8066 cm^{-1}. These energies, in contrast to the term values, are usually measured from zero at the lowest state, and increase for successive levels. They draw closer together until they converge at 13.598 eV. This corresponds to the orbit with $n = \infty$ and complete removal of the electron, that is, ionization of the atom. Above this ionization potential is a continuum of states representing the nucleus plus a free electron possessing a variable amount of kinetic energy.

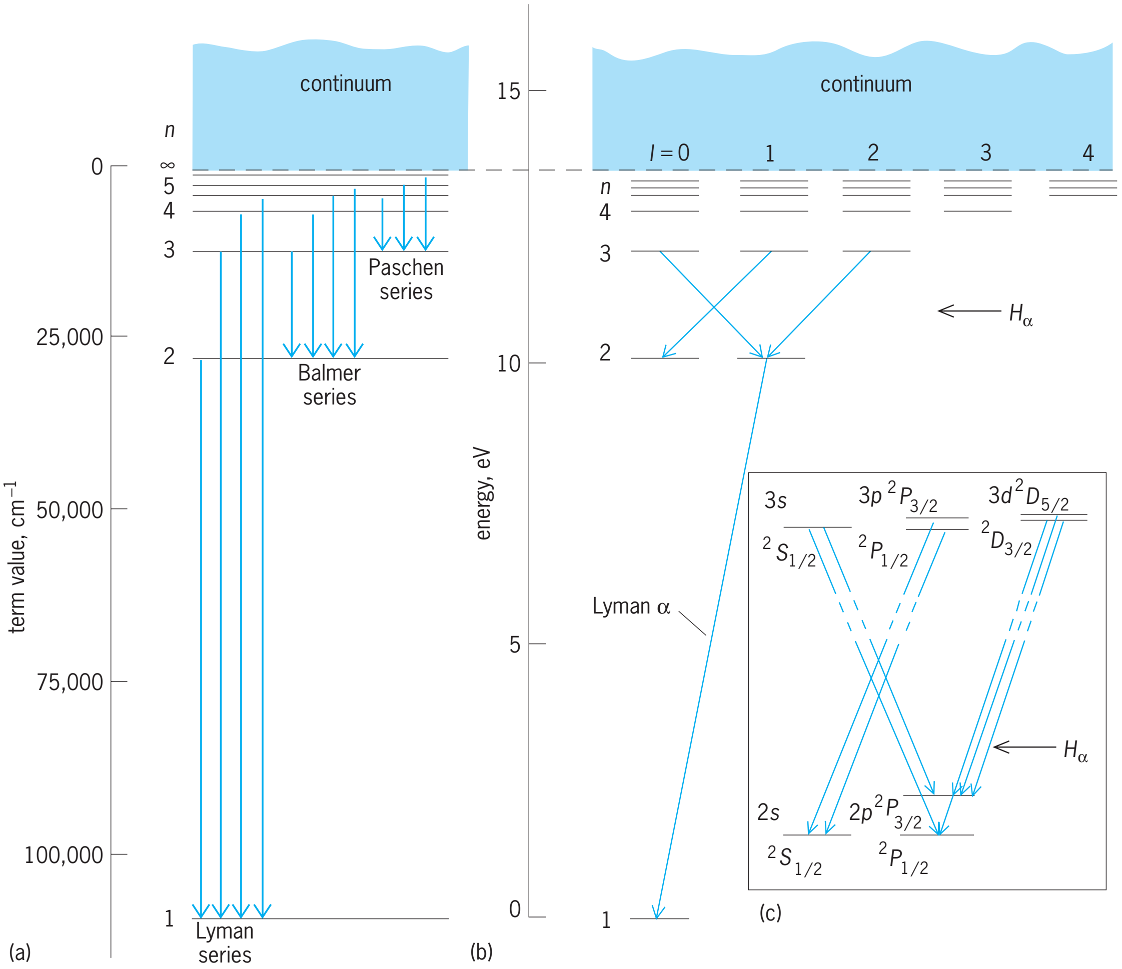


Fig. 4. Terms and transitions for the hydrogen atom. (*a*) Bohr theory. (*b*) Bohr-Sommerfeld theory. (*c*) Dirac theory.

The names of the hydrogen series are indicated in Fig. 4*a*. The spectral lines result from transitions between these various possible energy states. Each vertical arrow on the diagram connects two states and represents the emission of a particular line in the spectrum. The wave number of this line, according to Bohr's frequency condition, is equal to the difference in the term values for the lower and upper states and is therefore proportional to the length of the arrow. The wave numbers of all possible lines are given by Eq. (12), known as the Balmer formula,

$$\sigma = T'' - T' = R\left(\frac{1}{n''^2} - \frac{1}{n'^2}\right) \qquad (12)$$

where the double primes refer to the lower energy state (larger term value) and the single primes to the upper state. Any particular series is characterized by a constant value of n'' and variable n'. The only series with any lines in the visible region is the Balmer series, represented by Eq. (12) with $n'' = 2$. Its first line, that for $n' = 3$, is the bright red line at the wavelength 656.3 nanometers and is called H_α. Succeeding lines H_β, H_γ, and so on proceed toward the ultraviolet with decreasing spacing and intensity, eventually converging at 364.6 nm. Beyond this series limit there is a region of continuous spectrum. The other series, given by $n'' = 1, 3, 4$, etc., lie well outside the visible region. The Lyman series covers the wavelength range 121.5–91.2 nm in the vacuum ultraviolet, and the Paschen series 1875.1–820.6 nm in the infrared. Still other series lie at even longer wavelengths.

Since hydrogen is by far the most abundant element in the cosmos, its spectrum is extremely important from the astronomical standpoint. The Balmer series has been observed as far as H_{31} in the spectra of hot stars. The Lyman series appears as the strongest feature of the Sun's spectrum, as photographed by rockets and satellites orbiting above the Earth's atmosphere such as Skylab.

Nuclear Magnetism and Hyperfine Structure

Most atomic nuclei also possess spin, but rotate about 2000 times slower than electrons because their mass is on the order of 2000 or more times greater than that of electrons. Because of this, very weak nuclear magnetic fields, analogous to the electronic ones that produce fine structure in spectral lines, further split atomic energy levels. Consequently, spectral lines arising from them are split according to the relative orientations, and hence energies of interaction, of the nuclear magnetic moments with the electronic ones. The resulting pattern of energy levels and corresponding spectral-line components is referred to as hyperfine structure. *See* NUCLEAR MOMENTS.

The fine structure and hyperfine structure of the hydrogen terms are particularly important astronomically. In particular, the observation of a line of 21-cm wavelength, arising from a transition between the two hyperfine components of the $n = 1$ term, gave birth to the science of radio astronomy. *See* ASTRONOMICAL SPECTROSCOPY; RADIO ASTRONOMY.

Investigations with tunable lasers. The enormous capabilities of tunable lasers have allowed observations which were impossible previously. For example, high-resolution saturation spectroscopy, utilizing a saturating beam and a probe beam from the same laser, has been used to measure the hyperfine structure of the sodium resonance lines (called the D_1 and D_2 lines). Each line was found to have components extending over about 0.017 cm^{-1}, while the separation of the D_1 and D_2 lines is 17 cm^{-1}. The smallest separation resolved was less than 0.001 cm^{-1}, which was far less than the Doppler width of the lines. *See* DOPPLER EFFECT; HYPERFINE STRUCTURE; LASER SPECTROSCOPY.

Isotope shift. Nuclear properties also affect atomic spectra through the isotope shift. This is the result of the difference in nuclear masses of two isotopes, which results in a slight change in the Rydberg constant. There is also sometimes a distortion of the nucleus, which can be detected by ultrahigh precision spectroscopy. *See* MOLECULAR BEAMS; PARTICLE TRAP.

Lamb Shift and Quantum Electrodynamics

The Bohr-Sommerfeld theory, which permitted elliptical orbits with small relativistic shifts in energy, yielded a fine structure for H_α that did not agree with experiment. The selection rule $\Delta l = \pm 1$ permits three closely spaced transitions. Actually, seven components have been observed within an interval of 0.5 cm^{-1}. According to the Dirac theory, the spin-orbit and the relativistic correction to the energy can be combined in the single formula (13), so that levels with the same n and j coincide exactly, as shown in Fig. 4*c*. The splittings of the levels are greatly exaggerated in the figure. Selection rules, described later, limit the transitions to $\Delta j = 0$ and ± 1, so there are just seven permitted transitions for the Hα line. Unfortunately for the theory, two pairs of these transitions coincide, if states of the same j coincide, so only five lines of the seven observed would be accounted for.

$$\Delta T = \frac{R\alpha^2 Z^4}{n^3} \cdot \left(\frac{2}{2j+1} - \frac{3}{4n} \right) \qquad (13)$$

The final solution of this discrepancy came with the experimental discovery by W. Lamb, Jr., that the $2s^2S$ level is shifted upward by 0.033 cm^{-1}. The discovery was made by extremely sensitive microwave measurements, and it has subsequently been shown to be in accord with the general principles of quantum electrodynamics. The Lamb shift is present to a greater or lesser degree in all the hydrogen levels, and also in those of helium, but is largest for the levels of smallest l and n. Accurate optical measurements on the wavelength of the Lyman α line have given conclusive evidence of a shift of the $1s^2S$ level.

The knowledge of this 1*s* level shift has improved with great rapidity. By 2000, the transition frequency (energy of the transition divided by Planck's constant h) separating the 1*s* and 2*s* states of atomic hydrogen had been measured with an uncertainty of only 2 parts in 10^{14}, representing the most accurate measurement up to that time of any frequency or wavelength in the visible or ultraviolet regions of the light spectrum. This high accuracy is possible primarily because (1) the 2*s* state in hydrogen is so long-lived that the natural width of the transition is only about 1 Hz; and (2) by probing a cold atomic hydrogen beam with two simultaneous laser photons, each with half the transition energy, from opposite directions, the Doppler broadening arising from the motion of the target atoms in the laser field is made to cancel to first order. As further discussed below, such Doppler spreads are usually a major source of smearing out such measurements.

Though an extended discussion of the highly abstract field theory of quantum electrodynamics is inappropriate here, a simple picture of the lowest-order contribution to the Lamb shift in hydrogen can be given. No physical system is thought ever to be completely at rest; even at temperatures near absolute zero (about −273°C or −460°F), there is always a zero-point oscillatory fluctuation in any observable physical quantity. Thus, electromagnetic fields have very low-level, zero-point fluctuations, even in vacuum, at absolute zero. Electrons respond to these fluctuations with a zero-point motion: in effect, even a point charge is smeared out over a small volume whose mean square radius, though tiny, is not exactly zero. The smearing out of the charge distribution changes the interaction energy of a charged particle with any additional external field in which it may reside, including that of an atomic nucleus. As *s* electrons spend more time near the nucleus, where the nuclear-field Coulomb electric field is most intense, than do, say, electrons having more extended orbits, the change in interaction energy is greater for *s* electrons than, say, *p* electrons, and is such as to raise the energy of *s* electrons above the exact equality with that for *p* electrons of the same *n* arising in Dirac's theory of the electron.

The higher the charge number Z of the atomic nucleus, the more intense is the electric field experienced by atomic electrons, and hence the Lamb shift of the energy level of each atomic electron is larger. Experiments pioneered in the 1960s and later extended showed that few-electron atoms analogous to ordinary hydrogen, helium, and so forth could be prepared from atoms of higher Z by ionizing fast beams of higher-Z elements by using the thin-foil electron stripping techniques described below, and that Lamb shifts in these analogs could be measured. By 1986 the method had been extended as far as heliumlike uranium (U^{90+}), for which a one-electron Lamb shift of 70.4 ± 8.3 eV was measured

-S

compared with a theoretical value of 75.3 ± 0.4 eV. Once again the theory of quantum electrodynamics had exhibited its unchallenged status as the most exact and successful physical field theory ever created. *See* QUANTUM ELECTRODYNAMICS.

Doppler Spread

In most cases, a common problem called Doppler broadening of the spectral lines arises, which can cause overlapping of spectral lines and make analysis difficult. The broadening arises from motion of the emitted atom with respect to a spectrometer. Just as in the familiar case of an automobile horn sounding higher pitched when the auto is approaching an observer than when receding, so also is light blueshifted (to higher frequency) when approaching, and redshifted when receding from, some detection apparatus. The percentage shift in frequency or wavelength is approximately $(v/c)\cos\theta$, where v is the emitter speed, c the speed of light, and θ the angle between $\vec{v}$ and the line of sight of the observer. Because emitting atoms normally are formed with some spread in v and θ values, there will be a corresponding spread in observed frequencies or wavelengths. Several ingenious ways of isolating only those atoms nearly at rest with respect to spectrometric apparatus have been devised. The most powerful employ lasers and either involve the saturation spectroscopy mode mentioned above or use two laser photons which jointly drive a single atomic transition and are generated in lasers so arranged that the first-order Doppler shifts of the photons cancel each other.

Radiationless Transitions

It would be misleading to think that the most probable fate of excited atomic electrons consists of transitions to lower orbits, accompanied by photon emission. In fact, for at least the first third of the periodic table, the preferred decay mode of most excited atomic systems in most states of excitation and ionization is the electron emission process first observed by P. Auger in 1925 and named after him. For example, a singly charged neon ion lacking a 1*s* electron is more than 50 times as likely to decay by electron emission as by photon emission. In the process, an outer atomic electron descends to fill an inner vacancy, while another is ejected from the atom to conserve both total energy and momentum in the atom. The ejection usually arises because of the interelectron Coulomb repulsion. *See* AUGER EFFECT.

The vast preponderance of data concerning levels of excited atomic systems concerns optically allowed, single-electron, outermost-shell excitations in low states of ionization. Since much of the mass in nature is found in stars, and most of the elements therein rarely occupy such ionization-excitation states, it can be argued that presently available data provide a very unrepresentative description of the commonly occurring excited atomic systems in nature. When the mean lives of excited atomic systems are considered, the relative rarity of lifetime measurements on Auger electron-emitting states is even more striking. The experimentally inconvenient typical lifetime range (10^{-12} to 10^{-16} s) accounts for this lack.

Spectroscopy of Highly Ionized Atoms

Since the 1960s, an effective means of creating high ionization-excitation states, such as those that are common in stars and other high-temperature plasmas, has been provided by beam-foil spectroscopy. In this method, ions accelerated by a charged-particle accelerator pass through a thin carbon foil. The resulting beam of ions stripped of one or more electrons can be analyzed so the charge states and velocities are accurately known. Highly excited states are often produced by the interaction with the foil. The subsequent emission, detected with a spectrometer, allows the measurement of lifetimes of states which are difficult to produce in other sources. When, for example, electromagnetic atomic decays are observed, optical or x-ray spectrometers may be useful. Sometimes the intensity of the emitted radiation undergoes fluctuations called quantum beats, analogous to sound intensity heats of tuning forks of slightly different frequency, as can be seen in the photograph in **Fig. 5**. When, as is most frequently the case, Auger processes are dominant, less familiar but comparably effective electron spectrometers are preferred. *See* BEAM-FOIL SPECTROSCOPY.

Violent, high-velocity atomic collisions or electron-atom collisions are needed to reach the highest ionization-excitation states possible. Because highly ionized atoms are very important in stellar atmospheres, and in terrestrially produced highly ionized gases (plasmas) on which many potentially important schemes for thermonuclear fusion energy production are based, it is also important to overcome the Doppler spread problem in violent collisions. For sufficiently highly ionized, fast, heavy projectiles, lighter target atoms in a gaseous target

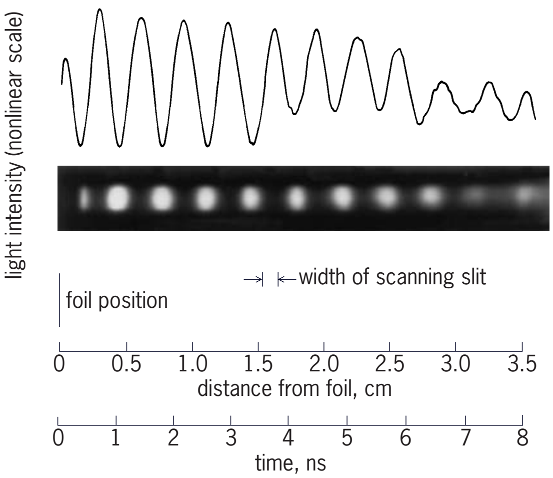


Fig. 5. Example of atomic process following excitation of a beam of ions in a thin solid target less than 1 μm in thickness. Quantum fluctuations in the light emitted by a 2-mm-diameter beam of such particles, traveling at a few percent of the speed of light, are shown as the atoms travel downstream from the target. (*After I. A. Sellin et al., Periodic intensity fluctuations of Balmer lines from single-foil excited fast hydrogen atoms, Phys. Rev., 184:56–63, 1969*)

can be ionized to very high states of ionization and excitation under single-collision conditions, while incurring relatively small recoil velocities. Sample spectra indicate that a highly charged projectile can remove as many as $Z - 1$ of the Z target electrons while exciting the last one remaining, all under single-collision-event conditions. Subsequent studies showed that the struck-atom recoil velocities are characterized by a velocity distribution no worse than a few times thermal, possibly permitting a new field of precision spectroscopy on highly ionized atoms to be developed.

In the 1980s and 1990s more direct methods of producing ions in high ionization-excitation states underwent rapid development. Ion sources have been developed that can produce beams of ions that have much greater intensities but suffer from higher kinetic energies (of the order of keV) than those which are readily available from ion sources based on the recoil ion production method (eV energies). With some exceptions, these ion sources usually produce lower ionization-excitation states than are found by using the very fastest ion beams or the beam-foil method; however, those methods are usually characterized by corresponding ion kinetic energies in the GeV range.

The two most prominent types of sources are electron cyclotron resonance (ECR) sources and cryogenic electron-beam ion sources (CRYEBIS). The former depend for their operation on ionization and excitation of ions by impact of electrons heated to high temperatures in an intense microwave discharge. The latter also exploit electron impact ionization and excitation, but accomplish this objective by first using electrostatic forces to trap ions in low ionization-excitation states inside intense electron beams, and then making use of additional electron impact excitation and ionization to achieve high ionization-excitation states. Periodic expulsion of ions from the trapping region after a suitable time delay is then used to form ion beams which are hence inherently pulsed in time, in contrast to beams from electron cyclotron resonance sources, which are normally extracted in a steady-state manner. *See* ION SOURCES.

In the 1990s a shorter, more compact version of the CRYEBIS, the electron beam ion trap (EBIT), underwent rapid development. Its operating principles are very similar, but it features close-up access to a short and very thin column of highly ionized ions excited by intense high-energy electron impact on atoms in the limited fields of view of specialized spectroscopic instruments mounted as close to the interaction region as possible. It is the favored instrument for increasingly precise x-ray spectroscopy on highly ionized atoms.

CRYEBIS sources are used as injectors for heavy-ion storage rings, and are designed to provide the much larger fluxes of high-charge-state ions needed to fulfill the potential of such rings. These machines recycle the supply of stored ions in the ring once every few microseconds through a given orbital point, enhancing their effective numbers, since each ion can be interrogated once per cycle by each of one or more local measuring instruments. Lifetime measurements can be carried out by monitoring decay times of the ions. The application of merged-electron-beam cooling in the ring (discussed further below) to greatly reduce the longitudinal and transverse velocities of stored ions is a further crucial attribute, permitting, for example, study of very low relative velocity interactions of merged electrons and ions inside the ring. *See* PARTICLE ACCELERATOR.

Relativistic Dirac Theory and Superheavy Elements

Dirac's theory of a single electron bound to a point nucleus implies a catastrophe for hydrogen atoms of high Z, if Z could be made on the order of $137 \cong 1/\alpha$ or larger [in which $\alpha = e^2/(2\epsilon_0 hc)$ is the fine-structure constant]. Dirac's expression for the electronic binding energy becomes unphysical at that point. Nuclear theorists have, however, contemplated the possible stable existence of superheavy elements, appreciably heavier than any observed heretofore, say, in the range $Z = 114$ to 126 or even larger. Though searches for such superheavy elements of either primordial or artificial origin have thus far proved unsuccessful, the advent of heavy-ion accelerators capable of bringing heavy particles to a distance of closest approach much smaller than the radius of the innermost shell in any atom, even a heavy one, raises the possibility of transient creation of a superheavy atom with a combined atomic number, $Z_{\text{comb}} = Z_1$ and Z_2, of about 180 or even greater (in which Z_1 and Z_2 are the individual atomic numbers of the colliding system).

The binding energy of a 1*s* electron in an atom increases rapidly with increasing atomic number. As already noted, in the usual linear version of the Dirac equation, a catastrophe occurs when the nuclear charge reaches $(1/\alpha) \cong 137$. When $Z_1 + Z_2$ is sufficiently large, the 1*s* binding energy can reach and exceed twice the rest energy of the electron $m_e c^2$, in which m_e is the electron mass and c is the speed of light. By varying the target or projectile, one can trace the path of the electron energies as they dive toward the negative-energy sea, or, in the case of the 1*s* electron, into the negative-energy sea, to give rise to bound states degenerate with the negative-energy continuum (**Fig. 6**). The concept of a negative-energy sea at $-m_e c^2$ was introduced by Dirac and used to predict the existence of a new particle, the positron, represented by a hole in the sea. The 1*s* level acquires a width because of the admixture of negative-energy continuum states, and the 3*s*-state width corresponds to a decaying state, provided there is nonzero final-state density. When a hole is present in the K shell, spontaneous positron production is predicted to take place. Production of an electron-positron pair is possible, since a final 1*s* ground state is available for the produced electron; the positron is predicted to escape with kinetic energy corresponding to overall energy balance. *See* ANTIMATTER; ELECTRON-POSITRON PAIR PRODUCTION.

Because of the effects of finite nuclear size in modifying the potential, $Z_1 + Z_2$ must reach some initial value Z_{CR} greater than 137 at which the "splash" into the continuum is made. According to nuclear-model-dependent estimates, the splash would occur at about 170. Some types of nonlinear additions to the Dirac equation may remove the diving into the negative-energy continuum, and these could be tested by observation of positron production as a function of $Z_1 + Z_2$. Other "limiting field" forms of nonlinear electrodynamics lead to larger values of Z_{CR}, and since the positron escape width turns out to be approximately proportional to the square of the difference between the effective quasi-nuclear charge and Z_{CR}, these nonlinearities could be tested too. Intensive searches for spontaneous positron production have been undertaken but have not yet met with success.

Such spontaneous positron production, if ever found, would represent an effect due to the potential energy of interaction of a combined nucleus of superheavy charge $Z_1 + Z_2$ with Dirac's negative energy sea. However, experiments, carried out chiefly in the 1990s, have observed copious production of dynamically induced positrons from the reservoir of kinetic energy brought into collisions of heavy projectile ions with atoms at speeds very close to the speed of light.

The physical picture here is that of a naked ion (no electrons remaining) passing by a target atom at such close range and so fast that the disturbance due to the passage of the ion's electrical charge closely resembles a very energetic transverse light pulse, whose spectrum of photon energies is entirely sufficient to liberate positrons from Dirac's negative energy sea via electron-positron pair production, in which both an electron and a positron are born in the Coulomb field of the struck atom. In this picture, free electron-positron pairs are formed by collisional ionization of electrons initially lying in Dirac's negative electron continuum. Sometimes excitation of a negative-continuum electron into an unoccupied bound state of either the projectile or target particle manifests itself as atomic capture of the electron produced in such pair production, reducing the electrical charge of the ion by one unit, while simultaneously producing a free positron. Eventually, in either process, the positron inevitably collides with some other electron and annihilates, converting its Einstein rest energy and that of a struck electron back into photons. Rates for both the production of free pairs and pair-produced electron capture (**Fig. 7**) are large—of the scale for ordinary atomic as opposed to elementary-particle-interaction reaction rates—and increase with the energy of the collision. *See* POSITRON; QUASIATOM.

Uncertainty Principle and Natural Width

A principle of universal applicability in quantum physics, and of special importance for atomic physics, is W. Heisenberg's 1927 uncertainty principle. In its time-energy formulation, it may be regarded as holding that the maximum precision with which the energy of any physical system (or the corresponding frequency ν from $E = h\nu$) may be determined is limited by the time interval Δt available for a measurement of the energy or frequency in question, and in no case can ΔE be less than approximately $h/(2\pi\,\Delta t)$. The immediate consequence is that the energy of an excited state of an atom

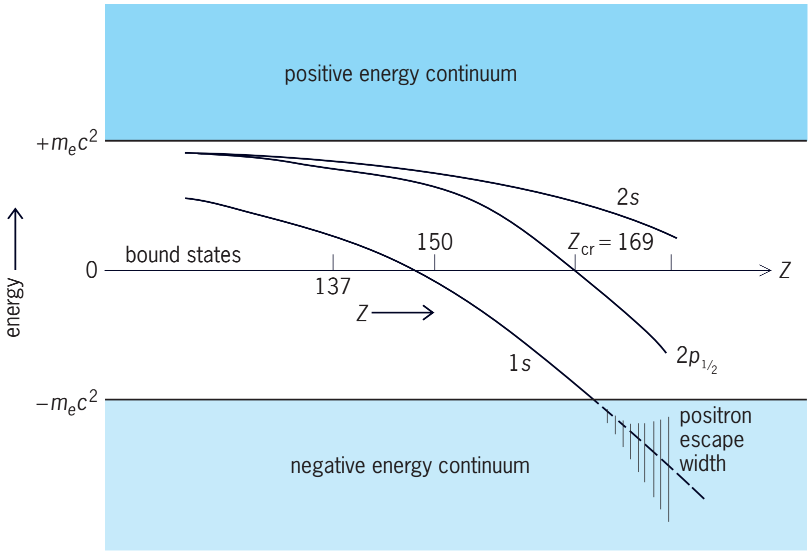


Fig. 6. **Behavior of the binding energies of the 1*s*, 2*p*$_{1/2}$, and 2*s* electrons with increasing atomic number *Z*. (*After J. H. Hamilton and I. A. Sellin, Heavy ions: Looking ahead, Phys. Today, 26(4):42–49, April 1973*)**

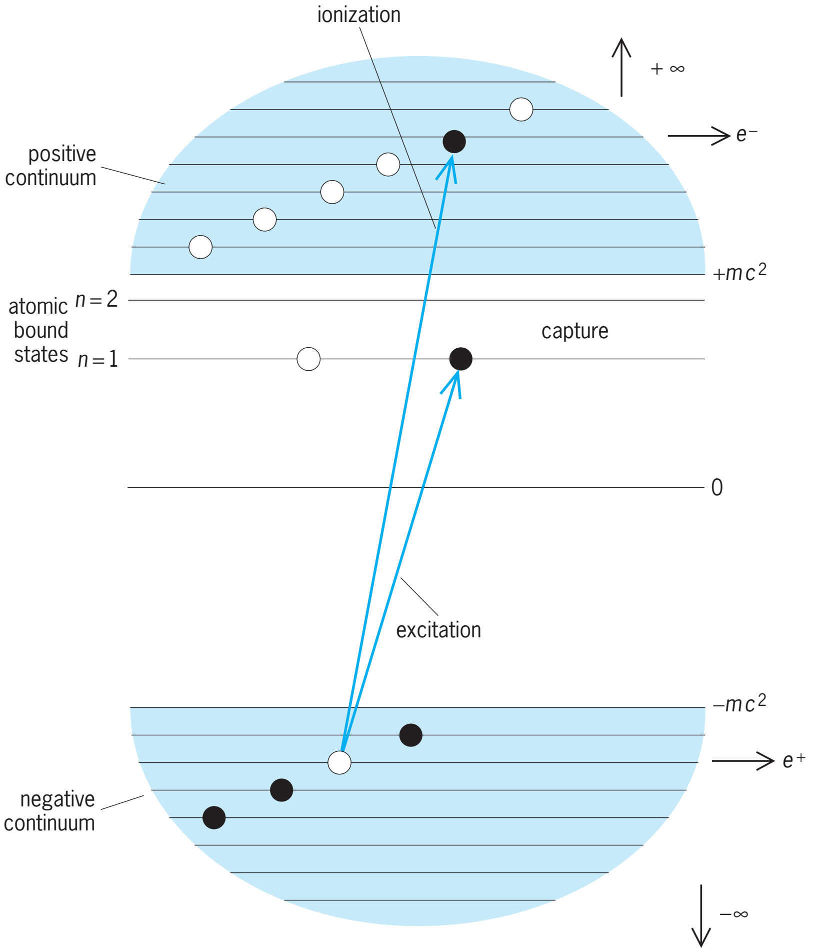


Fig. 7. **Free positron (e^+) and free (ionization) or bound (capture) electron (e^-) production, arising from pair-production creation of a hole (positron) in Dirac's negative-electron-energy sea. The energy mc^2 refers to the einsteinian rest-mass energies of electrons and positrons, of 511,000 eV each. (*After Charles Vane*)**

cannot be determined with arbitrary precision, as such states have a finite lifetime available for measurement before decay to a lower state by photon or electron emission. It follows that only the energies of particularly long-lived excited states can be determined with great precision. The energy uncertainty ΔE, or corresponding frequency interval $\Delta\nu$, is referred to as the natural width of the level. For excited atoms radiating visible light, typical values of $\Delta\nu$ are on the order of 10^9 Hz. (For a long time it was thought that the natural widths of levels participating in laser-cooling transitions set a lower limit to how low a temperature laser cooling can achieve. However, ingenious schemes have been developed that do not violate the uncertainty principle but skillfully evade it.) For inner electrons of heavy atoms, which give rise to x-ray emission, and also for Auger electron-emitting atoms, 10^{14} Hz is more typical. An energy interval $\Delta E = 1$ eV corresponds to about 2.4×10^{14} Hz. *See* UNCERTAINTY PRINCIPLE.

The time-energy uncertainty principle is perhaps best regarded as a manifestation of the basic wave behavior of electrons and other quantum objects. For example, piano tuners have been familiar with the time-frequency uncertainty principle for hundreds of years. Typically, they sound a vibrating tuning fork, or its modern electronic equivalent, of accepted standard frequency in unison with a piano note of the same nominal frequency, and listen for a beat note between the tuning fork and the struck tone. Beats, which are intensity maxima in the sound, occur at a frequency equal to the difference in frequency between the two sound sources. For example, for a fork frequency of 440 Hz and a string frequency of 443 Hz, three beat notes per second will be heard. The piano tuner strives to reduce the number of beats per unit time to zero, or nearly zero. To guarantee a frequency good to 440 ± 0.01 Hz would require waiting for about 100 s to be sure no beat note had occurred ($\Delta\nu\Delta t \gtrsim 1$). *See* BEAT.

Cooling and Stopping Atoms and Ions

Despite impressive progress in reducing Doppler shifts and Doppler spreads, these quantities remain factors that limit the highest obtainable spectroscopic resolutions. The 1980s and 1990s saw extremely rapid development of techniques for trapping neutral atoms and singly charged ions in a confined region of space, and then cooling them to much lower temperatures by the application of laser-light cooling techniques. Photons carry not only energy but also momentum; hence they can exert pressure on neutral atoms as well as charged ions. *See* LASER COOLING.

Schemes have been developed to exploit these light forces to confine neutral atoms in the absence of material walls, whereas various types of so-called bottle configurations of electromagnetic fields developed earlier remain the technique of choice for similarly confining ions. For neutral atoms and singly charged ions, whose energy levels (unlike those of most more highly charged ions) are accessible to tunable dye lasers, various ingenious methods have been invented to slow down and even nearly stop atoms and ions. More effective than cooling methods known from low-temperature physics, these methods often utilize the velocity-dependent light pressure from laser photons of nearly the same frequency as, but slightly less energetic than, the energy separation of two atomic energy levels to induce a transition between these levels. This procedure serves to cool atomic systems to submillikelvin temperatures. For example, an atom may absorb photons slightly below the precise resonance energy separation owing to the Doppler shift arising from its motion toward a laser source, and then reradiate fluorescent photons of slightly higher average energy in all directions. The atom thereby loses net directed momentum because of the redistribution in all directions and also loses net kinetic energy of motion because of the requirements of total energy conservation. The use of six laser beams in so-called optical molasses schemes permits atomic momentum (and hence kinetic energy) reduction in each of the six possible directions of motion. Other light forces on atoms, known as dipole gradient forces, have been used to deflect or focus beams of neutral atoms, exploiting the rapid spatial variation of electric fields inside laser beams. Depending on laser frequency, atoms are pulled toward or driven away from regions of high laser electric field because of the nature of the interaction of an inhomogeneous electric field with polarizable atoms.

This ingenious optical dipole trapping scheme was quickly improved upon by the invention of the magnetooptic trap. This device is larger and more robust. It combines optical forces provided by laser light with a weak magnetic field whose size goes through zero at the geometrical center of the trap and increases with distance from this center. The net result is a restoring force which confines sufficiently laser-cooled atoms near the center.

To universal surprise, initial measurements of the temperature of such trapped ions were only about 40×10^{-6} K above absolute zero. Until then, an anticipated value some six times greater had been regarded as a fundamental limit associated with the natural widths of the levels involved. Ingenious improvements have allowed cooling of ions to temperatures as low as 180×10^{-9} K.

For more highly ionized ions, annular storage rings are used in which radial confinement of fast ion beams (with speeds of approximately 10% or more of the speed of light) is provided by magnetic focusing, and longitudinal confinement automatically occurs because of the closed ionic orbits, similar to those in cyclotrons. Two cooling schemes are known to work on stored beams of charged particles, the so-called stochastic cooling method and the electron cooling method. In the former, deviations from mean stored particle energies are electronically detected, and electronic "kicks" that have been adjusted in time and direction are delivered to the stored particles to compensate these deviations. In electron cooling, which proves to be more effective for stored heavy ions of high charge, electron beams prepared with a narrow velocity distribution are merged with the stored ion beams. When the average speeds of the electrons and the ions are matched, the Coulomb interaction between the relatively cold (low-velocity-spread) electrons and the highly charged ions efficiently transfers energy from the warmer ions, thereby reducing the temperature of the stored ions. Electron cooling occurs quickly, and has been demonstrated to cool beams of fast protons sufficiently to reduce the ratio of the spread in the longitudinal momentum to the total momentum to better than 1 part per million. Several heavy-ion storage rings are presently in operation.

Successful Explanations and Unresolved Problems

Many of the atomic enigmas referred to above have been explained through experimental discoveries and theoretical inspiration. Examples include Rutherford scattering, the Bohr model, the invention of the quantum theory, the Einstein quantum relation between energy and frequency $E = h\nu$, the Pauli exclusion and Heisenberg uncertainty principles, and many more. A few more are discussed below, but only in part, as the subject of atomic structure and spectra is by no means immune from the internal contradictions and defects in understanding that remain to be resolved.

In 1927 C. Davisson and L. Germer demonstrated that electrons have wave properties similar to that of light, which allows them to be diffracted, to exhibit interference phenomena, and to exhibit essentially all the properties with which waves are endowed. It turns out that the wavelengths of electrons in atoms are frequently comparable to the radius of the Bohr orbit in which they travel. Hence, a picture of a localized, point electron executing a circular or elliptical orbit is a poor and misleading one. Rather, the electrons are diffuse, cloudlike objects which surround nuclei, often in the form of standing waves, like those on a string. Crudely speaking, the average acceleration of an electron in such a pure standing wave, more commonly called a stationary state in the quantum theory, has no time-dependent value or definite initial phase. Since, as in Maxwell's theory, only accelerating charges radiate, atomic electrons in stationary states do not immediately collapse into the nucleus. Their freedom from this catastrophe can be interpreted as a proof that electrons may not be viewed as pointlike planetary objects orbiting atomic nuclei. *See* ELECTRON DIFFRACTION.

Atoms are by no means indivisible, as Thomson showed. Neither are their nuclei, whose constituents can be ejected in violent nucleus–nucleus collision events carried out with large accelerators. Nuclear radii and shapes can, for example, be measured by high-energy electron scattering experiments, in which the electrons serve as probes of small size. *See* NUCLEAR STRUCTURE.

However, a good theory of electron structure still is lacking, although small lower limits have been established on its radius. Zero-point fluctuations in the vacuum smear their physical locations in any event. There is still no generally accepted explanation for why electrons do not explode under the tremendous Coulomb repulsion forces in an object of small size. Estimates of the amount of energy required to "assemble" an electron are very large indeed. Electron structure is an unsolved mystery, but so is the structure of most of the other elementary objects in nature, such as protons, neutrons, neutrinos, quarks, and mesons. There are hundreds of such objects, many of which have electromagnetic properties, but some of which are endowed with other force fields as well. Beyond electromagnetism, there are the strong and weak forces in the realm of elementary particle physics, and there are gravitational forces. Electrons, and the atoms and molecules whose structure they determine, are only the most common and important objects in terrestrial life. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS.

Though the action-at-a-distance concept built into a Coulomb's-law description of the interactions of electrons with themselves and with nuclei is convenient and surprisingly accurate, even that model of internal atomic interactions has its limits. It turns out that electrons and nuclei do not interact instantaneously, but only after a delay comparable to the time needed for electromagnetic waves to travel between the two. Electrons and nuclei do not respond to where they are with respect to each other, but only to where they were at some small but measurable time in the past.

Quantum jumps. Einstein is credited with introducing the idea of discontinuous transitions in the time evolution of microscopic quantum states, such as are involved in atomic transitions. Bohr's postulate 3, discussed earlier, implicitly assumed that electrons changing states and radiating photons do so in discontinuously short time intervals—examples of what are commonly referred to as quantum jumps. E. Schrödinger, the formulator of the Schrödinger equation, which well describes many features of the quantum behavior of electrons in atoms as well as that of numerous other microscopic quantum systems, hoped that the necessity of assuming discontinuous quantum jumps could be avoided. He hoped that refined calculations would show a smooth change between the wave function solutions which satisfy the Schrödinger equation for individual stationary Bohr orbits when transitions between orbits occur. The presently accepted status of quantum theory nevertheless remains that of a purely statistical theory: the numerical coefficients of terms which when added together constitute an atomic wave function which satisfies the Schrödinger equation for an atom in a so-called general, superposition state, determine only the average probability of finding an atom in a stationary state described by one of the terms. Finding an atom in a particular state means making a state determination measurement, such as detecting a photon whose frequency labels the final stationary state of an electron after a transition has occurred.

If by a quantum-mechanical system is meant an ensemble of a large number of identically prepared atoms, then these average values may indeed evolve

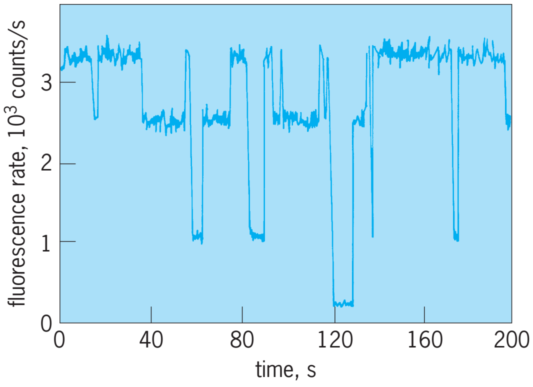


Fig. 8. Multiple quantum jumps occurring in the laser-excited 493-nm fluorescence of three Ba$^+$ ions. (*After W. Neuhauser and Th. Sauter, Observation of quantum jumps, Comments Atom. Mol. Phys., 21:83–95, 1988*)

continuously in time. However, if it is possible to prepare a very small number of isolated atoms in excited states in the same way again and again—in effect, to realize a time average, not an ensemble average over subsequent photon emissions—the question arises of whether discontinuous quantum jumps are truly observable.

Evidence that they are observable is seen in **Fig. 8**, which shows the intensity of light from large numbers of fluorescent photons which are emitted during rapid, repetitive laser excitation of a particular excited state in one or more of three Ba$^+$ ions contained in an electromagnetic ion trap. These ions most often relax back to the ionic ground states (states of no excitation) by the fluorescent photon emission whose intensity is observed. Occasionally, each of the ions relaxes instead to a state whose energy is intermediate, for which selection rules impede photon emission to the ground state—in effect, "shelving" the ion for a while in a state for which the laser is ineffective in repetitive reexcitation to the fluorescent state because its frequency does not match the frequency between the intermediate and upper levels. The apparently discontinuous switching to dark states is seen to reduce the fluorescent intensity to about two-thirds of the original level when one ion jumps to a dark state, to one-third of the level when two ions jump, and to nearly zero intensity when three ions jump. An additional remarkable feature of these observations is that the number of apparently simultaneous jumps of two or three ions at a time exceeds that expected on the basis of random coincidence of jumps by about two orders of magnitude. Cooperative interactions of even a very small number of atomic systems with the laser radiation field seem to have unexpectedly high probability.

Photon antibunching. Still other intimate connections between quantum jumps in atoms and the photons constituting the electromagnetic radiation field to which they give birth are illustrated by a phenomenon called photon antibunching. Fluorescent photon emission from a large ensemble of excited atoms in thermal equilibrium at some high temperature is said to be incoherent, meaning that individual atoms emit photons at random times; that is, the instant of photon emission of a given atom is not correlated with that of others in the ensemble. Under these circumstances the probability of a second photon being detected at some time interval Δt after detection of a first has a maximum at $\Delta t = 0$, a result called photon bunching since the photon emission tends to occur in bursts.

For one to a few ions stored in an electromagnetic trap, however, the smallest value of the probability can occur at $\Delta t = 0$, an effect called antibunching. That this result is plausible can be demonstrated by considering a trapped ion which returns to its ground state following fluorescent photon emission. If it is illuminated by a laser which reexcites it to the level from which fluorescence is again possible, the reexcitation is in general not instantaneous but occurs after some mean time delay determined by the laser intensity and the probability per unit time for resonant laser photon absorption, which is also a statistical process. The probability of two fluorescent photons being emitted in a time interval that is very short compared to that required on the average for reexcitation should therefore be very small, as the likelihood that the ion is still in its ground state should then be very high.

An illustration of antibunching for a single Mg$^+$ ion stored in an electromagnetic trap is shown in **Fig. 9**. The second-order correlation function (intensity correlation) $g^{(2)}$ of the $3^2S_{1/2} - 3^2P_{3/2}$ transition

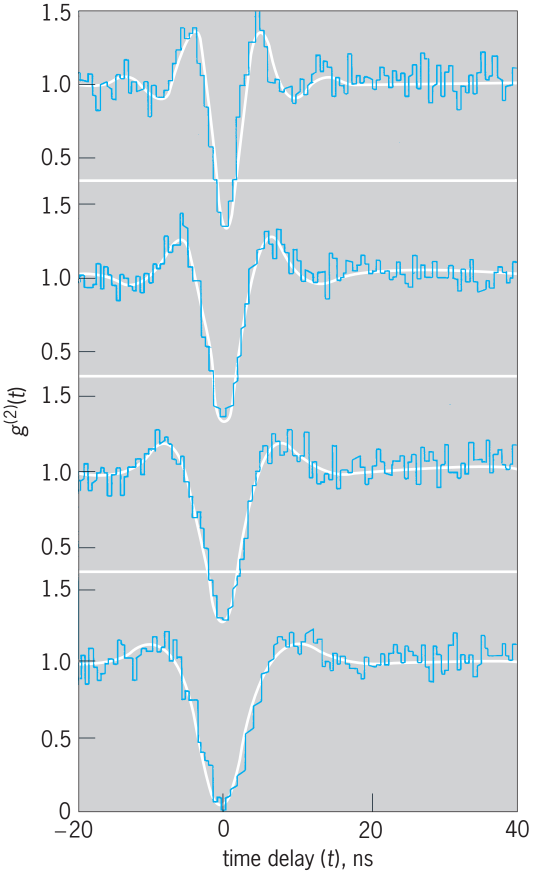


Fig. 9. Antibunching signal of a single Mg$^+$ ion for four different laser intensities. (*After F. Diedrich and H. Walther, Nonclassical radiation of a stored ion, Phys. Rev. Lett., 58:203–206, 1987*)

in Mg^+ is plotted versus the time delay t between two fluorescent photon emissions for four different laser intensities. The quantity $g^{(2)}(t)$ is proportional to the probability of a second photon being detected after a first emission with a time delay t. The large antibunching effect near zero time delay is strikingly demonstrated. Ivan A. Sellin

Bibliography. E. U. Condon and H. Odabasi, *Atomic Structure*, Cambridge University Press, 1980; R. D. Cowan, *The Theory of Atomic Structure and Spectra*, University of California Press, 1981; H. Friedrich, *Theoretical Atomic Physics*, 3d ed., Springer-Verlag, 2005; B. Levi, Work on atom trapping and cooling gets a warm reception in Stockholm, *Phys. Today*, 50(12):17–19, December 1997; B. Schwarzschild, Optical frequency measurement is getting a lot more precise, *Phys. Today*, 50(12):19–21, December 1997; L. Szasz, *The Electronic Structure of Atoms*, Wiley, 1992.

Atomic theory

The study of the structure and properties of atoms based on quantum mechanics and the Schrödinger equation. These tools make it possible, in principle, to predict most properties of atomic systems. A stationary state of an atom is governed by a time-independent wave function which depends on the position coordinates of all the particles within the atom. To obtain the wave function, the time-independent Schrödinger equation has to be solved. This is a second-order differential equation (1), where the first term in parenthesis is the so-called kinetic energy operator, the second term is the potential energy operator, and E is the energy of the state described by Ψ. The potential energy term contains the Coulomb interaction between all the particles in the atom, and in this way they are all coupled to each other. *See* DIFFERENTIAL EQUATION; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; SCHRÖDINGER'S WAVE EQUATION.

$$\begin{aligned} H\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}_N) &= \left(\sum_{i=1}^{N} \frac{-\hbar^2}{2m_i}\nabla_i^2 + V(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}_N)\right) \\ &\quad \times \Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}_N) \\ &= E\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}_N) \qquad (1) \end{aligned}$$

A many-particle system where the behavior of each particle at every instant depends on the positions of all the other particles cannot be solved directly. This is not a problem restricted to quantum mechanics. A classical system where the same problem arises is a solar system with several planets. In classical mechanics as well as in quantum mechanics, such a system has to be treated by approximate methods. *See* CELESTIAL MECHANICS.

Independent particle model. As a first approximation, it is customary to simplify the interaction between the particles. In the independent particle model the electrons are assumed to move independently of each other in the average field generated by the nucleus and the other electrons. In this case the potential energy operator in Eq. (1) will be a sum over one-particle operators and Eq. (1) can be rewritten as Eq. (2). The simplest wave function, Ψ, which will satisfy Eq. (2) is a product of one-particle orbitals, $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N = \varphi_1\varphi_2\cdots\varphi_N$. To fulfill the Pauli exclusion principle, the total wave function must, however, be written in a form such that it will vanish if two particles are occupying the same quantum state. This is achieved with an antisymmetrized wave function, that is, a function which, if two electrons are interchanged, changes sign but in all other respects remains unaltered. The antisymmetrized product wave function is usually called a Slater determinant. *See* EXCLUSION PRINCIPLE.

$$\begin{aligned} \left(\sum_{i=1}^{N} \frac{-\hbar^2}{2m_i}\nabla_i^2 + u(\mathbf{r}_i)\right) &\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}_N) \\ &= E\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}_N) \qquad (2) \end{aligned}$$

Hartree-Fock method. In the late 1920s, only a few years after the discovery of the Schrödinger equation, D. Hartree showed that the wave function to a good approximation could be written as a product of orbitals, and also developed a method to calculate the orbitals. Important contributions to the method were also made by V. Fock and J. C. Slater (thus, the Hartree-Fock method).

For a stationary atomic state the energy should be stable with respect to small variations of the wave function. If the expectation value of the energy is minimized under the condition that the wave function should be a single Slater determinant, the expression for the Hartree-Fock potential is found. The Hartree-Fock model thus gives the lowest-energy ground state within the assumption that the electrons move independently of each other in an average field from the nucleus and the other electrons.

To simplify the problem even further, it is common to add the requirement that the Hartree-Fock potential should be spherically symmetric. This leads to the central-field model and the so-called restricted Hartree-Fock method. For a closed-shell system the requirement for the restricted Hartree-Fock method is automatically fulfilled.

The Hartree-Fock potential is a one-particle potential, but it depends on the orbitals of all the electrons. A common approach is to start with a reasonable approximation of the orbitals, calculate the Hartree-Fock potential from these, and solve the differential equation. This gives a new set of orbitals and the procedure can be repeated until convergence occurs. The solution is then said to be self-consistent.

The Hartree-Fock method gives a qualitative understanding of many atomic properties. Generally it is, for example, able to predict the configurations occupied in the ground states of the elements. Electron binding energies are also given with reasonable accuracy. For example, the ionization energy of neon is measured to be 21.5 eV, while the result with the Hartree-Fock method is 23 eV.

Electron correlation. Correlation is commonly defined as the difference between the full many-body problem, Eq. (1), and the Hartree-Fock model. More specifically, the correlation energy is the difference between the experimental energy and the Hartree-Fock energy. There are several methods developed to account for electron correlation. Some important methods are the following.

In the configuration-interaction method, the total many-body wave function is expanded in a basis set, and the coefficients are determined by minimization of the expectation value of the energy. This method is widely used on both atoms and molecules. Basis sets consisting of analytic functions are often used. *See* QUANTUM CHEMISTRY.

The multiconfiguration Hartree-Fock method is a natural extension of the Hartree-Fock model. A number of configurations are chosen, and the mixing coefficients, as well as the radial part of the orbitals, are varied to minimize the expectation value of the energy.

In perturbation theory, the full hamiltonian operator, H, is divided up into two parts, $H = H_0 + V$, where H_0 can be, for example, the Hartree-Fock model (usually within the central field approximation), and then a perturbation expansion of V is carried out. The operator H_0 must be a reasonable approximation to H, otherwise the perturbation expansion may not converge. *See* PERTURBATION (QUANTUM MECHANICS).

All these methods require large-scale numerical calculations. The ability to perform accurate calculations increased dramatically during the 1980s and 1990s due to the development of fast computers.

One way to classify the different contributions to the correlation energy is to count the number of electrons involved in the different terms. The most important type of correlation is pair correlation. As an example, **Table 1** shows the importance of different correlation contributions to the ground state of the beryllium atom, a system with four electrons. The relative sizes of the contributions can be understood by comparison with a system of colliding classical particles. In such a system, collisions between two particles dominate and three-particle collisions are very rare. True four-particle collisions are also very rare, but two simultaneous two-particle collisions constitute a kind of effective four-particle effect which is more common.

Table 1 shows the contribution to the correlation energy only. The Hartree-Fock model accounts for 99% of the total binding energy of beryllium, and thus only around 1% of the binding energy is due to correlation. This success of the independent-particle model for the total energy is due to the large contributions from the inner electrons. These electrons interact primarily with the nucleus and only to a smaller extent with the other electrons, and thus are less influenced by correlation.

TABLE 1. Contributions to the total correlation energy for the ground state of beryllium

Number of particles involved	Contribution, %
Singles	0.7
Doubles	94.6
Triples	0.6
Quadruples	4.1

Strongly correlated systems. Although the Hartree-Fock model can qualitatively explain many atomic properties, there are systems and properties for which correlation is more important.

Many, but not all, atoms can bind an extra electron and form a negative ion. Since there is no long-range field outside a neutral atom, it is the detailed electron-electron interaction which makes binding possible. In general, the Hartree-Fock model makes quite poor quantitative predictions for the so-called electron affinity (the binding energy of the extra electron). Sometimes a calculation within the Hartree-Fock model will even predict that it is not possible to form a negative ion, but experiments as well as more refined correlated calculations show that it is indeed possible. Calcium is one such element; the electron affinity is around 0.02 eV, although no bound state is found in a Hartree-Fock calculation. *See* ELECTRON AFFINITY; NEGATIVE ION.

Excited states are often much more sensitive to electron correlation than the ground state. One well-known example comes from the beryllium atom. The ground state of beryllium can be denoted $(1s^2 2s^2)\ {}^1S_0$. In beryllium-like ions the first excited state of ${}^1S_0$ symmetry is the $(1s^2 2p^2)\ {}^1S_0$ state. In neutral beryllium, where the attractive interaction with the nucleus is relatively weaker, this state is not bound at all. The behavior can be explained only if electron correlation, which pushes it up above all the $(1s^2 2sns)\ {}^1S_0$ states (where n is the primary quantum number of the outer most electron), is accounted for. In fact, it is generally true that doubly excited states constitute strongly correlated systems. *See* ELECTRON CONFIGURATION.

In open-shell systems the electrons within the shell interact strongly with each other, and generally many states can be formed from one configuration. In an average level calculation the same Hartree-Fock orbitals are used for all these levels. This is a natural independent-particle starting point, but often large discrepancies are found when the result is compared with experiments. In the rare-earth elements (elements 57–70), the $4f$ shell is successively filled. For example, the Pr^{3+} ion is a system with two $4f$ electrons outside closed shells. Within this configuration it is possible to form 13 ${}^{2S+1}L_J$ states, and the independent-particle model does not even give their energies in the correct order. *See* RARE-EARTH ELEMENTS.

If the interest is not in calculating the total energy of a state but in understanding some other properties, effects beyond the central field model can be more important. Small changes in the wave function may have a large impact on some weak effect, although the total energy is insensitive. One example comes from the so-called hyperfine structure, that is, the interaction between the electronic and the nuclear magnetic moments. The most important contribution comes from the electronic wave function

-S

inside the nucleus. In closed shells there are as many electrons with spin up as spin down, and the hyperfine interactions from these electrons cancel each other. If there are open outer shells, the electrons in these shells will interact slightly differently with the closed-shell spin-up and spin-down electrons. In this way the closed shell will be polarized; that is, the overlap with the nucleus will differ for a spin-up and a spin-down electron. Since inner electrons have a large overlap with the nucleus, even a small imbalance gives rise to a substantial effect. *See* HYPERFINE STRUCTURE.

Relativistic effects. The Schrödinger equation is a nonrelativistic wave equation. In heavy elements the kinetic energy of the electrons becomes very large. For example, the speed of the innermost electron in lead is around 40% of the speed of light. For heavy systems it is therefore not sufficient to use a nonrelativistic theory. Instead, the calculations are based on the relativistic counterpart to the Schrödinger equation, the Dirac equation. The original Dirac equation describes an electron moving in an external potential, provided, for example, by the nucleus. To describe many-electron systems, the interaction between the electrons has also to be specified. Nonrelativistically it is given by the Coulomb interaction. In Dirac theory the electronic spin is an integrated part of the equation, and this gives the electron a magnetic moment; that is, the electrons can also interact magnetically with each other. In addition, the interaction cannot be instantaneous, but has to propagate with the speed of light.

It is possible to construct a Hartree-Fock model based on the Dirac equation, where the electron-electron interaction is given by the Coulomb interaction, a magnetic contribution, and a term which corrects for the finite speed (retardation) with which the interaction propagates. This method is commonly used for heavy elements. Correlation can also be calculated in a relativistic framework with this expression for the electron-electron interaction. Here special problems arise, however, from the positron solutions (the positron is the electron's antiparticle) to the Dirac equation. It is important to separate electron and positron solutions in calculations to obtain reliable results. Practical solutions to this problem were found and implemented in calculations only during the 1980s. *See* ANTIMATTER; RELATIVISTIC QUANTUM THEORY.

Radiative corrections. With the measurement of the so-called Lamb shift in hydrogen in the 1940s, it became clear that there are additional effects not accounted for by the Dirac equation. These effects, which are called radiative corrections, were soon afterward shown to arise when the electromagnetic field was quantized within the theory of quantum electrodynamics. This theory, which combines special relativity (the Dirac equation) with the quantization of the electromagnetic field, has been very successful. It gives in principle a complete description of how to handle many-particle systems. In practice, however, a full quantum electrodynamics treatment of an atom becomes very cumbersome, and only a few calculations on two- and three-electron systems exist. The radiative corrections are, however, important for inner-shell electrons, and their effect on the interaction between an electron and a nucleus is well understood. For many-body systems, calculations of radiative effects are usually done within some independent-particle model, and the result is added to a correlated relativistic calculation based on the Dirac equation. *See* QUANTUM ELECTRODYNAMICS.

TABLE 2. Ionization energy of the innermost electron in lead

Type of contribution	Energy, eV
Relativistic Hartree-Fock model, Coulomb interaction*	88,498
Magnetic and retardation effects	−322
Relativistic correlation	4
Radiative contributions	−167

*A nonrelativistic calculation would account for only around 90% of this result.

As an example of a heavy system where both relativistic and radiative effects are important, **Table 2** shows the different contributions to the energy required to remove the innermost electron in lead. This quantity is not measured directly, but after the electron has been removed another electron will fall down into the hole left behind and an x-ray photon will be emitted. The least energetic photon is measured to have an energy of 74,970.01 ± 0.17 eV. The calculated value, which includes the ionization energy of the innermost electron, is close to that: 74,971 eV. *See* ATOMIC STRUCTURE AND SPECTRA.

Eva Lindroth

Bibliography. R. D. Cowan, *The Theory of Atomic Structure and Spectra*, Los Alamos Series in Basic and Applied Sciences, University of California, 1981; C. F. Froese Fischer, T. Brage, and P. Jönsson, *Computational Atomic Structure: An MCHF Approach*, IOP, 1997; L. Lindgren and J. Morrison, *Atomic Many-Body Theory*, Series on Atoms and Plasma, 2d ed., Springer-Verlag, Berlin, 1986.

Atomic time

Time that is based on physics, specifically on a unit of duration that corresponds to a defined number of periods of radiation of a specific atomic transition of a chosen isotope. The fundamental unit of atomic time is the International System (Système International or SI) second, which is defined as the duration of 9,192,631,770 periods of radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom. *See* ATOMIC CLOCK.

International Atomic Time. International Atomic Time (Temps Atomique International, or TAI) is determined by the Bureau International des Poids et Mesures (BIPM) in Sèvres, France from commercial atomic time standards and primary frequency standards in many countries. The time from the different sources is compared by the Global Positioning System (GPS) and two-way satellite time transfer to determine the differences between the source time standards and a free running time scale at BIPM. The resulting TAI conforms as closely as possible to the definition of the SI second. TAI is steered to match the SI second; departures range 0.5–1.19 × 10^{-14}, with uncertainties of 0.2 × 10^{-14}.

TAI seeks to provide accuracy, long-term stability, and reliability. It was officially introduced in January 1972, but has been available since July 1955. TAI is the basis for a number of other time scales.

Terrestrial Time (TT). Terrestrial Time (TT) is the theoretical coordinate time whose mean rate is close to the mean rate of the proper time of an observer located on the geoid (the surface of constant gravity potential that extends the sea-level surface over the entire Earth). The unit of TT is the SI second. An accurate realization of TT is TT(TAI) = TAI + 32.184 seconds. *See* DYNAMICAL TIME.

Coordinated Universal Time (UTC). Since January 1, 1972, most, and now all, broadcast time services and civil times are based on the redefined Coordinated Universal Time (UTC), which is based on TAI and differs from TAI by an integer number of seconds. UT1 is based on the observationally determined rotational motion of the Earth, which conforms closely with the mean diurnal motion of the Sun. UTC is maintained within 0.9 s of UT1 by the introduction of 1-s steps (leap seconds) when necessary, usually at the end of December or June. The civil time around the world differs from UTC mostly by hour increments according to time zones. *See* EARTH ROTATION AND ORBITAL MOTION.

Satellite navigation. GPS and GLONASS (Global'naya Navigatsionnya Sputnikovaya Sistema) have atomic clocks on the satellites and use the accurate time and ephemerides of the satellites as a means of determining accurate positions on Earth. The satellite times are steered to match the rate of TAI and provide accurate sources of time. The GPS system is used for accurate time transfer and for the determination of the rotation and polar motion of the Earth. *See* SATELLITE NAVIGATION SYSTEMS.

Relativistic effects. The accuracy of atomic time requires that relativistic effects be taken into account. Hence, the definition of TAI is a coordinate time scale at what is known as the geocentric datum line, and having as its unit the SI second as obtained on the geoid in rotation. Then corrections to atomic clocks entering the formation of TAI must be made for gravitational potential and velocity and for the rotation of the Earth. Specifically, if there are two standard clocks at the same elevation on the Earth with respect to the geoid, the one closer to the Earth's equator will run slower than a clock closer to the pole, due to the greater rotational velocity near the equator. A standard clock at a lower elevation with respect to the geoid will run slower than a higher clock. *See* GEODESY; RELATIVITY.

Atomic clocks. There have been a large number of commercial cesium standards and a small number of primary cesium frequency standards available as the basis of atomic time. Since the 1990s, cesium fountains have been developed based on the principles of laser cooling of the cesium atoms. These cesium fountains (as of 2005, about five exist) have provided an order-of-magnitude improvement in accuracy. Optical frequency standards are now under investigation. These are based on optical atomic transitions, at optical rather than microwave frequencies, so they operate with a much smaller unit of time and can possibly achieve as much as five orders of magnitude improvement in accuracy. As of 2005, cesium atomic clocks achieve timing uncertainties of about 35 picoseconds per day, and optical atomic clocks may have uncertainties of 100 femtoseconds per day.

Due to the technology of atomic clocks, the SI second is the most accurately measured unit. Hence, the meter is now defined as the distance that light travels in 1/299,792,458 of a second in vacuum, rather than as a physical bar of metal. *See* LASER COOLING; LENGTH; PHYSICAL MEASUREMENT.

History. The civil time scales were defined based on the mean solar time until it was recognized that the rotation of the Earth was irregular. In the 1950s Ephemeris Time was introduced as a new time scale based on the apparent motion of the Sun. The ephemeris second was the new unit of time. However, it was difficult to determine Ephemeris Time from observations of the Sun, so observations of the Moon, which were more accurate, were used. Also, the definition of the ephemeris second was based on Simon Newcomb's theory of the motion of the Earth, which was fit to observations from the 1800s. So the ephemeris second matched the rate of rotation of the Earth from about 1850.

In the 1950s observations of the Moon were used to determine the atomic second with respect to the ephemeris second. The definition of the atomic second was given to agree with the ephemeris second. That is the definition of the SI second used today and the reason that atomic time does not match the current rate of rotation of the Earth.

P. Kenneth Seidelmann

Bibliography. S. A. Diddams et al. (2004), Standards of time and frequency at the outset of the 21st century, *Science*, 306:1318–1324, 2004; W. Markowitz et al., Frequency of cesium in terms of Ephemeris Time, *Phys. Rev. Lett.* 1:105–107, 1958; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, University Science Books, Mill Valley, California, 1992.

Atomization

The process whereby a bulk liquid is transformed into a multiplicity of small drops. This transformation, often called primary atomization, proceeds through the formation of disturbances on the surface of the bulk liquid, followed by their amplification due to energy and momentum transfer from the surrounding gas.

Spray formation processes are critical to the performance of a number of technologies and applications. These include combustion systems (gas turbine engines, internal combustion engines, incinerators, furnaces, rocket motors), agriculture (pesticide and fertilizer treatments), paints and coatings (furniture, automobiles), consumer products (cleaners, personal care products), fire suppression systems, spray cooling (materials processing, computer chip cooling), medicinals, (pharmaceutical), and spray drying (foods, drugs, materials processing). Current concerns include how to make smaller drops (especially for internal combustion engines), how to make larger drops (agricultural sprays), how to reduce the number of largest and smallest drops (paints and coatings, consumer products, medicinals, spray drying), how to distribute the liquid mass more uniformly throughout the spray, and how to increase the fraction of liquid that impacts a target (paints and coatings, spray cooling, fire suppression).

Spray devices (that is, atomizers) are often characterized by how disturbances form. The most general distinction is between systems where one or two fluids flow through the atomizer. The most common types of single-fluid atomizers (**Fig. 1**) are pressure (also called plain-orifice, hydraulic, or pneumatic), pressure-swirl, rotary, ultrasonic (sometimes termed whistle or acoustic), and electrostatic; other types include impinging jet, fan, and jet-swirl. Twin-fluid atomizers (**Fig. 2**) include internal-mix and external-mix versions, where these terms describe the location where atomizing fluid (almost always a gas) first contacts fluid to be sprayed (almost always a liquid).

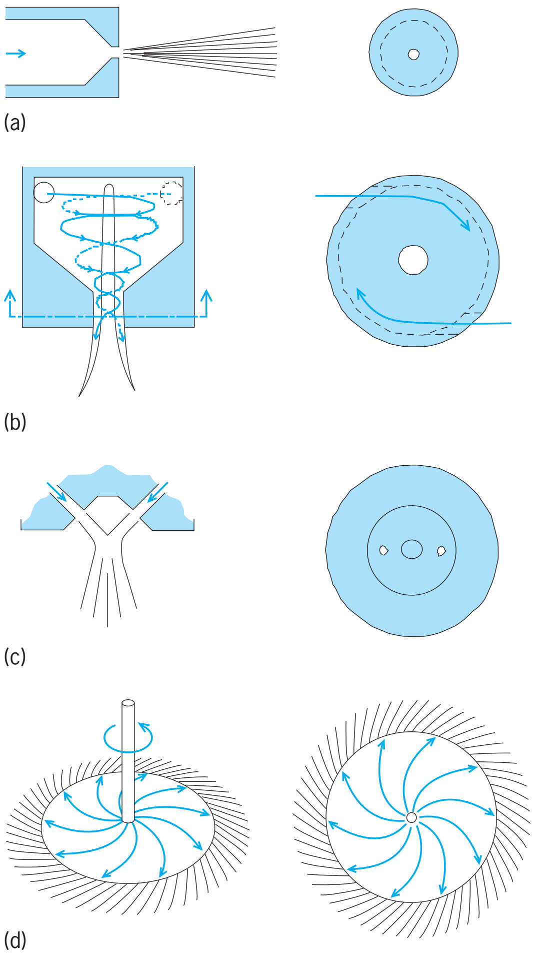


Fig. 1. Single-fluid atomizers. Diagrams on right are cross sections. (*a*) Pressure atomizer. (*b*) Pressure-swirl atomizer. (*c*) Impinging-jet atomizer. (*d*) Rotary atomizer.

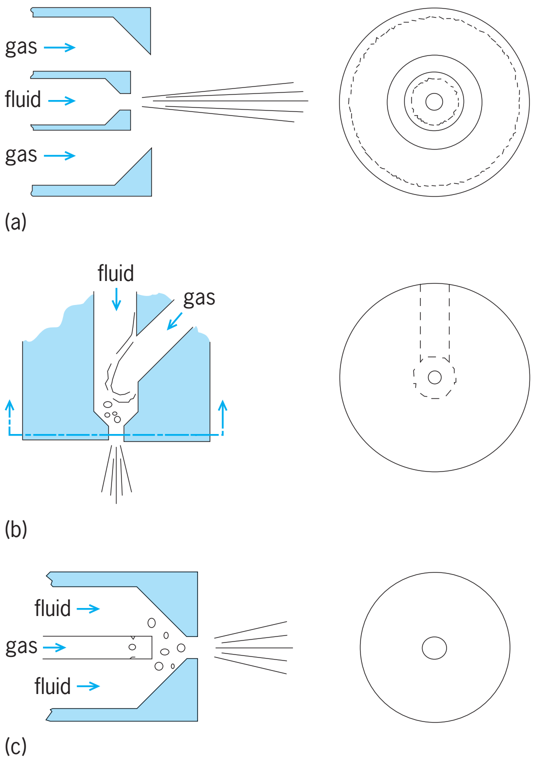


Fig. 2. Twin-fluid atomizers. Diagrams on right are cross sections. (*a*) External-mix atomizer. (*b*) Y-jet internal-mix atomizer. (*c*) Effervescent internal-mix atomizer.

While primary atomization is important, because of its role in determining mean drop size and the spectrum of drop sizes, subsequent processes also play key roles in spray behavior. They include further drop breakup (termed secondary atomization), drop transport to and impact on a target, drop evaporation (and perhaps combustion), plus drop collisions and coalescence. In addition, the spray interacts with its surroundings, being modified by the adjacent gas flow and modifying it in turn.

Basic processes. The spray formation process actually starts inside the atomizer itself. It is here that the fluid flow assumes its laminar or turbulent state. The latter is beneficial to forming small drops because precursors to fluid-mechanic instabilities are naturally present in turbulent flows. *See* LAMINAR FLOW; TURBULENT FLOW.

Primary atomization. Sheet or jet breakup resulting from fluid-mechanic instabilities is termed primary atomization. It depends on process operating conditions, fluid physical properties, and atomizer geometry.

Process operating conditions dictate the relative velocity between the fluid to be atomized and any surrounding gas. For single-fluid atomizers, fluid supply pressure determines the velocity at which fluid exits the atomizer into the (assumed) quiescent environment. For twin-fluid atomizers, the keys are the fluid and atomizing gas velocities. In all cases, it has been established that a larger relative velocity will accelerate the growth of instabilities and lead to smaller drops.

The fluid physical properties that most strongly influence spray formation are viscosity and surface tension. Density usually plays a smaller role; an increase in density usually leads to a small decrease in drop size.

An increase in shear (or dynamic) viscosity always leads to an increase in droplet size because viscosity directly opposes instability growth. If viscosity is high (say, above 100 centipoise or 0.1 pascal second), some types of atomizers (including pressure and pressure-swirl) can no longer form sprays without the use of excessively high fluid supply pressures. *See* VISCOSITY.

The process is more complex if the fluid is non-newtonian, especially if viscoelastic. The elastic contribution is especially efficient at damping out instabilities, thereby inhibiting breakup. Non-newtonian behavior is prevalent in paints and coatings, pharmaceutical and medicinal sprays, consumer product sprays, and adhesives. A practical result is that single-fluid atomizers cannot form sprays of reasonable quality from viscoelastic fluids. *See* NON-NEWTONIAN FLUID.

An increase in surface tension can enhance or attenuate instability growth. For primary atomization, increasing surface tension leads to smaller drops at low relative velocity. High relative velocity causes an increase in surface tension to lead to larger drops. When spray formation is dominated by secondary atomization, increasing surface tension always leads to larger drops. *See* SURFACE TENSION.

Atomizer geometry affects spray formation through the initial configuration of the fluid to be sprayed. Fluid exiting the atomizer is often spread into thin sheets or filaments to enhance the breakup process since smaller drops result from thinner sheets and smaller-diameter jets. Sheets can be flat, as in fan sprays, or annular, as in pressure-swirl sprays and prefilming twin-fluid sprays (where the liquid is spread into a thin sheet prior to its contact with the atomizing air). *See* JET FLOW.

Secondary atomization. If the relative velocity between an already formed drop and its surrounding gas is sufficiently high, the drop may undergo breakup, that is, secondary atomization. Its importance is evaluated by the drop Weber number, a dimensionless group representing the competition between aerodynamic distortion of the drop and restorative surface tension, given by Eq. (1). Here ρ_{gas} is the mass

$$\mathrm{We} = \rho_{\mathrm{gas}} V_{\mathrm{rel}}^2 d/\sigma \qquad (1)$$

density of the surrounding gas, V_{rel} is relative velocity between the drop and the surrounding gas, d is drop diameter, and σ is surface tension. Values of the Weber number greater than 12 indicate that secondary atomization may be present. Higher values indicate that drops break up via the bag, boundary-layer stripping, and catastrophic modes (**Fig. 3**). *See* DIMENSIONLESS GROUPS.

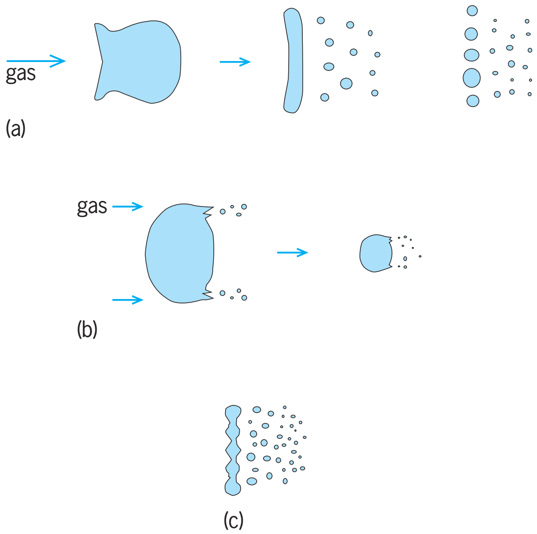


Fig. 3. Secondary atomization mechanisms. (*a*) Bag mode. (*b*) Boundary-layer stripping mode. (*c*) Catastrophic mode.

Drop transport. Droplet transport, or motion, is determined from the forces that action the drop using Newton's second law. Aerodynamic drag, F_{drag}, is usually the most important force. It is given by Eq. (2), where C_{drag} is the drag coefficient of the

$$F_{\mathrm{drag}} = (1/2)\rho_{\mathrm{gas}} C_{\mathrm{drag}} (\pi/4) d^2 V_{\mathrm{rel}}^2 \qquad (2)$$

drop. Since drops deform when moving, the drag coefficient depends on the drop's Reynolds number. Because droplet mass depends on diameter cubed while droplet drag force depends on diameter squared, large drops subject to aerodynamic drag retain their velocities relative to the surrounding gas much longer than smaller drops. *See* AERODYNAMIC FORCE; REYNOLDS NUMBER.

Additional forces of interest include gravity, when drops are moving slowly (for example, "drop-out" in consumer product sprays), and electrostatic, when significant electric fields or drop charging are present (automotive paint sprays, for instance). Electrostatic forces can be due to imposed external fields, or to fields resulting from the charged drops themselves. *See* ELECTROSTATICS.

Drop evaporation. Evaporation, along with primary and secondary atomization and coalescence, is a mechanism for changing drop size. Evaporation is dominated by energy transfer from the surrounding gas that goes into increasing the drop sensible and latent enthalpies. Both steady-state and transient (that is, drop heat-up) behavior can be important.

Analytical models are available to predict drop diameter as a function of time. The simplest assumes that all energy addition to the drop goes into latent enthalpy (that is, results directly in evaporation), the process is at steady-state, the drop is not affected by other drops, the process is spherically symmetric, the interior temperature is uniform throughout the drop, the drop is stationary, and the gas-phase conditions are not changed by evaporation. The d^2-law, Eq. (3), results, where d_0^2 is the square of the droplet

$$d^2 = d_0^2 - \lambda_{\text{evap}} t \qquad (3)$$

diameter when evaporation begins, t is the time since evaporation started, and λ_{evap} is the evaporation constant. More sophisticated models predict transient heat-up behavior and the influence of convective heat transfer when the droplet is in motion. Nonuniform temperatures and internal circulatory flows within the drop, as well as drop-to-drop interactions, have been considered. *See* ENTHALPY; EVAPORATION; PSYCHROMETRICS.

Drop collisions and coalescence. Drop collisions are rare in most parts of a spray. The exceptions are regions where the number of drops per unit volume is high, or where drops have widely differing velocities. Examples are immediately adjacent to the atomizer exit orifice, where the spray has not had time to spread, and near a target where drops are impinging on a surface and (perhaps) rebounding.

The importance of collisions can be assessed through a model that treats drops as rigid spheres. The probability of a collision is related to the square of the average diameter of the two drops, their relative velocity, and the number of drops per unit volume for both colliders.

Drop coalescence leads to an increase in drop diameter. However, not all droplet collisions result in coalescence. Colliding drops must have sufficient kinetic energy to overcome surface tension at impact, but not so much that they rebound. An increase in drop viscosity usually increases the likelihood of a collision resulting in coalescence.

Spray-surroundings interactions. A spray will interact with its surroundings as it moves away from the atomizer. This interaction changes spray cone angle, influences how far the spray penetrates into the surroundings, controls how much surrounding air is drawn into the spray, impacts mean drop velocity and the spread in droplet velocities, and varies the distribution of droplet mass throughout the spray cone.

There is a noticeable decrease in cone angle, or turning-in of the spray boundary. This results from the entrainment of surrounding gas which drives drops (especially the smallest ones) toward the spray centerline (**Fig. 4**).

The spray-surroundings interaction also limits how far a spray penetrates into the surrounding gas. Diesel engine, gas turbine, and consumer product sprays are applications where penetration is important. Droplet average velocity also depends on spray-surroundings interactions. It usually decreases with distance from the atomizer because the spray most often exits into a stagnant environment so that aerodynamic drag decelerates the droplets.

Spray-surroundings interactions also modify the distribution of fluid mass throughout the spray. For a hollow-cone spray (**Fig. 5**), such as that produced by a pressure-swirl atomizer, entrainment helps to homogenize the fluid mass distribution by "filling in" the spray core. Furthermore, the size dependence of droplet motion means that larger drops are more likely to penetrate toward the edges of the spray while smaller drops are more likely to be dragged toward the interior of the spray by inbound gas, so that mean drop size increases with distance from the spray axis.

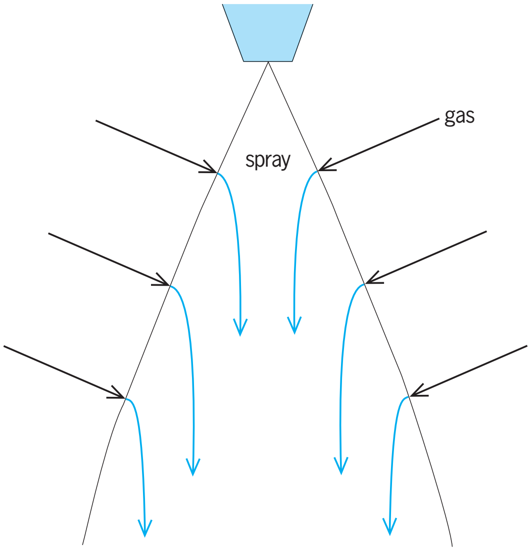


Fig. 4. Entrainment of surrounding air by a spray.

Single-fluid atomizers. Pressure and pressure-swirl atomizers require only a source of liquid at elevated pressure. Pressure atomizers (Fig. 1*a*) are found in diesel engines, aircraft engine afterburners, and some fire-suppression sprays. Their mean drop size performance is more sensitive to fluid viscosity than other types. In addition, fluid supply pressure variations can have such a large impact on atomizer performance that only a modest turn-down ratio (the ratio of the largest to the smallest liquid mass flow rates) is obtainable. Finally, pressure atomizers tend to produce sprays having narrow cone angles (half-angles of the order of 10°), so they are poor choices

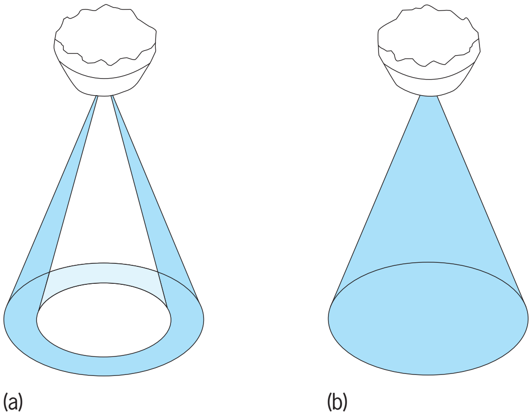


Fig. 5. Atomizer sprays. (*a*) Hollow-cone spray. (*b*) Solid-cone spray.

in applications where mixing the spray into its surroundings is important.

Pressure-swirl atomizers (Fig. 1*b*) produce sprays having larger, and controllable, cone angles. Values as high as 150° can be achieved. Modifications made to enhance the limited turn-down ratio characteristic of typical pressure-swirl atomizers have resulted in the dual-orifice (or duple), spill-return, and duplex designs. Pressure-swirl atomizers cannot produce reasonable quality sprays when the fluid viscosity climbs above about 100 cP (0.1 Pa · s).

Twin-fluid atomizers. The performance of twin-fluid atomizers tends to be less sensitive to fluid physical properties because they usually rely on secondary atomization. Secondary atomization depends somewhat on fluid surface tension, and to a lesser extent on viscosity and density.

Air-assist and air-blast atomizers are examples of external-mix units (Fig. 2*a*). Both can be used with highly viscous fluids. Their natural cone angles approach those of pressure atomizers, but can be extended using a mechanical deflection device called a pintle. Both types promote mixing between the fluid to be atomized and the surrounding gas, so they are desirable for combustion applications. An air-assist atomizer employs a smaller quantity of air than an air-blast unit, but at a higher velocity, and may require an external compressor for the atomizing gas supply.

Internal-mix designs include the Y-jet (Fig. 2*b*), effervescent (Fig. 2*c*), dissolved-gas, and flash atomizers. Effervescent atomizers can atomize highly viscous, and viscoelastic, fluids using a minimum amount of atomizing gas and large-diameter orifices. These energy-efficient devices form gas bubbles inside the atomizer body and often rely on primary atomization. Their natural cone angles are similar to those of pressure atomizers, but can be extended with mechanical deflection devices. Effervescent atomizers require an external supply of atomizing gas, but are relatively immune to clogging and plugging.

Dissolved-gas and flashing atomizers also form bubbles inside the atomizer body. They differ from effervescent types in their need for some combination of mass (dissolved gas) and energy transfer (flashing), in addition to bubble nucleation sites. This requirement can make them quite large. Moreover, it is necessary to dissolve gas into the fluid to be atomized, or mix in a high-vapor-pressure liquid, but an external compressor is unnecessary.

Measurement techniques. There are two principal methods for measuring drop sizes. Instruments based on forward light scattering (that is, diffraction) return information based on the number of drops per unit volume located along a line that crosses the spray. Instruments based on interference and refraction return information based on the rate of drops passing through a unit area located at a point in the spray. Diffraction instruments allow sprays to be characterized rapidly, but do not provide spatially precise information. Both techniques can interrogate transient sprays, and both provide the spray's mean drop size and the range of drop sizes produced. *See* DIFFRACTION; INTERFERENCE OF WAVES; REFRACTION OF WAVES; SCATTERING OF ELECTROMAGNETIC RADIATION.

Drop velocities are usually measured using interference-refraction instruments. The principles of operation are the same as those for laser Doppler velocimetry. Mean and root-mean-square (rms) velocity information is provided. *See* ANEMOMETER; FLOW MEASUREMENT.

Spray cone angles are usually measured via imaging. Charge-coupled-device (CCD) cameras, 35-mm cameras, movie cameras (both film and camcorders), and holography have been employed. Illumination sources include incandescent lamps, strobe lights, short-duration flashes, and arc lights. This same imaging apparatus can often be used to determine spray penetration.

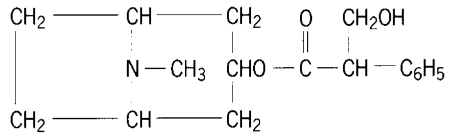
Mass distribution throughout the spray (patternation) can be measured using either intrusive mechanical devices or nonintrusive optical devices. *See* AEROSOL; PARTICULATES.

Paul E. Sojka

Bibliography. L. Bayvel and Z. Orzechowski, *Liquid Atomization*, Taylor and Francis, New York, 1993; A. H. Lefebvre, *Atomization and Sprays*, Hemisphere, New York, 1989; M. T. Lund, P. E. Sojka, and A. H. Lefebvre, Effervescent atomization at low mass flow rates, Part I: The influence of surface tension, *Atomiz. Sprays*, 3(1):77–89, 1993; S. D. Sovani, P. E. Sojka, and Y. R. Sivathanu, Prediction of drop size distributions from first principles: The influence of fluctuations in relative velocity and liquid physical properties, *Atomiz. Sprays*, 9(2):133–152, 1999.

Atropine

An alkaloid, $C_{17}H_{23}NO_3$, with the chemical structure below. The systematic chemical name is endo(±)-α-(hydroxymethyl)phenylacetic acid 8-methyl-8-azabicyclo[3.2.1]oct-3-yl ester, and in phamacy it is



sometimes known as *dl*-hyoscyamine. It occurs in minute amounts in the leaves of *Atropa belladonna, A. betica, Datura stramonium, D. innoxia, and D. sanguinea* as well as many related plants. It is chiefly manufactured by racemization of *l*-hyoscyamine, which is isolated from the leaves and stems of the henbane, *Hyoscyamus niger*. It melts at 114–116°C (237–241°F) and is poorly soluble in water. The nitrate and sulfate are used in medicine instead of the free base.

Atropine is used clinically as a mydriatic (pupil dilator). Dilation is produced by paralyzing the iris and ciliary muscles. Atropine is also administered in small doses before general anesthesia to lessen oral and air-passage secretions. Its ability to reduce these secretions is also utilized in several preparations commonly used for symptomatic relief of colds. *See* ALKALOID. Frank Wagner

Attention deficit hyperactivity disorder

A common psychiatric disorder of childhood characterized by attentional difficulties, impulsivity, and hyperactivity; known earlier as attention deficit disorder. A disorder or a syndrome comprises a group of signs and symptoms which tend to cluster together (symptoms refer to what the patient experiences; signs refer to what the physician observes). Other older names for this disorder include minimal brain dysfunction, minimal brain damage, hyperactivity, hyperkinesis, and hyperactive child syndrome. Over time, these names were modified due to their implications about etiology and core symptoms: minimal brain dysfunction seemed to imply that children with this disorder were brain-damaged, while hyperactivity and its synonyms named a feature seen in many but not all of these children. Recent studies indicate that attention deficit hyperactivity disorder tends to follow a chronic course, persisting in many cases not only into adolescence but into adulthood.

Clinical features. Some symptoms and signs may be present in all instances of the disorder, while others may occur in varying combinations or may not be present.

The three defining symptoms of attention deficit disorder are as follows: (1) *Attentional deficits.* The child is described as having a short attention span. Lacking "stick-to-itiveness," the child often fails to finish things he or she starts, does not seem to listen, and is easily distracted or disorganized. In more severe instances the child is unable to focus attention on anything, while in less severe cases attention can be focused on things of interest to the child. For example, the child may studiously absorb information about dinosaurs or tornadoes but ignore improper fractions. (2) *Impulsivity.* The child is often described as acting before thinking, shifting excessively and rapidly from one activity to another, or having difficulty waiting for a turn in games or group activities. (3) *Hyperactivity.* Many children with this disorder are hyperactive, and indeed, may have been noted to be so prior to birth. They are often described as always on the go or as driven by a motor. They may fidget, wiggle, move excessively, and have difficulty keeping still. This excessive activity is not noticeable when the children are playing. At such times, the child may be no more active than his or her peers; however, in the classroom or other quiet settings, the child cannot decrease his or her activity appropriately. Some affected children are active at a normal level or even sluggish (hence the inapplicability of the term "hyperactive" to a subset of individuals with attention deficit hyperactivity disorder); they have been studied less adequately than the hyperactive ones. On the basis of the predominating symptoms, children with attention deficit hyperactivity disorder are subcategorized as having hyperactive symptoms (hyperactive type), lacking hyperactivity (inattentive type), and having both inattention and hyperactivity or impulsivity (combined type).

Many children with attention deficit hyperactivity disorder frequently show an altered response to socialization. They are often described by their parents as obstinate, impervious, stubborn, or negativistic. With peers, many affected children are domineering or bullying, and thus may prefer to play with younger children. Another characteristic often seen in children with the disorder is emotional lability. Their moods change frequently and easily, sometimes spontaneously, and sometimes reactively so that the child is upset, sad, or angry.

Because of their behavioral difficulties, children with the disorder often have conflicts with parents, teachers, and peers. The child's behavior may engender friction with others. While this friction is sometimes used to explain the child's problems, more accurately it is a result of the child's disorder. Commonly, difficulties in discipline and inadequacies in schoolwork also lead to reproof and criticism. As a consequence, children with the disorder usually also have low self-esteem.

Associated disorders. Attention deficit hyperactivity disorder is frequently associated with other disorders, including disruptive behavior disorders, mood and anxiety disorders, and developmental disorders. The disruptive behavior disorders include oppositional defiant disorder and conduct disorder. Oppositional defiant disorder is characterized by stubbornness, defiance, and negativistic behavior. In conduct disorder, there is frequent recurrent violation of important rules at home or at school. There is often physical violence, lack of close relationships with others, and diminished or markedly reduced conscience. Specific learning disabilities represent a group of conditions (1) in which performance in psychometric testing in reading, spelling, mathematics, coordination, and other skills is below that which would be predicted on the basis of the child's intelligence, and (2) which is not explicable in terms of inadequate educational experience. In the case of reading and spelling, the disorder is often called dyslexia. When changes in mood are persistent and severe, the term "mood disorders" is applied; typical forms of mood disorders include depression and mania, sometimes known as bipolar disorder. Anxiety disorders refer to a group of conditions characterized by fearfulness and avoidance. Additional behavioral, developmental, and emotional difficulties are termed comorbidity. *See* AFFECTIVE DISORDERS.

Prevalence. It is estimated that 5–10% of children of elementary school age (roughly 6–19 years) manifest significant symptoms of attention deficit hyperactivity disorder. Moreover, it is estimated that the disorder may affect as many as 2–4% of adults. About twice as many boys as girls are affected with the disorder. The girls are much less likely than the boys to be aggressive and have serious behavioral difficulties, making the girls vulnerable to underidentification and undertreatment.

Causation. Formerly believed to be largely caused by brain damage, and more recently believed by some to be caused by food allergy, attention deficit hyperactivity disorder is now considered to be mainly hereditary. Supporting evidence is as follows: (1) The disorder runs in families. (2) In studies of adopted children with the disorder, psychiatric illnesses are found among biological parents rather than among adoptive parents. (3) Twin studies of the disorder document greater resemblance of monozygotic, or genetically identical, twins as compared to dizygotic twins, who share, on average, half of their varying genes. These twin studies suggest that attention deficit hyperactivity disorder is highly heritable. (4) Several investigators identified genes in the dopamine family as associated with this condition. Some researchers in biological psychiatry have suggested that many of the symptoms of the disorder are produced by an inherited metabolic abnormality in the brain. This abnormality is hypothesized to be underactivity in certain nerve tracts that is produced by deficient synthesis, release, or sensitivity to released neurotransmitters. (5) Not only do the parents (primarily the fathers) of affected children recount a history of the disorder in childhood, but there seems to be an increased frequency of other psychiatric disorders among relatives of children with the disorder. These other disorders include alcoholism, drug addiction, antisocial personality, and mood and anxiety disorders as well as learning and cognitive deficits.

Natural history. Formerly, it was believed that attention deficit hyperactivity disorder was outgrown during adolescence. Although some signs of the disorder such as excessive activity may diminish or disappear in some affected children, other signs such as attentional difficulties, impulsivity, and interpersonal problems may persist. It is now recognized that in about half of children with attention deficit hyperactivity disorder the disorder will persist into adulthood. Despite the fact that this disorder is common in adults, the lower rates of hyperactivity in adults may result in the condition being frequently overlooked.

Treatment. The treatment of the child or adult with this disorder involves three steps: evaluation, explanation of the problem to parents and child, and therapeutic intervention.

Evaluation requires a detailed history of the individual's psychological development and current functioning. When evaluating children, assessments commonly include information obtained from both parents and teachers and an interview with the child. Next, because the disorder is frequently associated with learning problems in school, it is desirable to obtain an individual intelligence test (such as the Wechsler Intelligence Scale for Children) as well as a test of academic achievement. Since attention deficit hyperactivity disorder is often associated with other psychiatric disorders, it is important to carefully evaluate the presence of these other conditions in the affected child, adolescent, or adult. If a diagnosis of attention deficit hyperactivity disorder is confirmed, the parents or family should be educated regarding the nature of the condition and associated conditions.

Medication and guidance are the mainstays of the treatment. Approximately 60–70% of children and adults manifest a therapeutic response to one of the major stimulant drugs, such as amphetamines and methylphenidate. A new generation of reformulated highly effective, once-daily, long-acting stimulants has emerged that has greatly facilitated the management of individuals with attention deficit hyperactivity disorder. Recently, a new nonstimulant medication (automoxetine) received FDA approval for the treatment of children and adults with attention deficit hyperactivity disorder. When effective, these medications increase attention, decrease impulsivity, and usually make the child more receptive to parental and educational requests and demands. Hyperactivity, when present, is usually diminished as well. These medications do not produce a "chemical straitjacket" but diminish the child's feeling of being driven. Although usually less effective (and, thus, second-line treatments), other medications can be helpful to individuals who cannot tolerate or do not respond to stimulants. The common mechanism of action for such medications is their impact upon the neurotransmitters dopamine and norepinephrine.

Because stimulants can be abused, concern about these medications has been expressed by the public. Yet, euphoria rarely occurs either in children or in adults with attention deficit hyperactivity disorder treated with adequate doses under medical supervision. Recent findings have emerged documenting that medication management of children with attention deficit hyperactivity disorder exerts a protective effect against the development of problems with drugs and alcohol in adolescence. Since medication is symptomatic rather than curative (that is, it only reduces the symptoms of the disorder), it must be used as long as the symptoms persist. Medication must be used in conjunction with concrete, specific, and flexible guidance techniques. Medication does not, however, improve specific developmental disorders (such as dyslexia) or coordination problems. Many children with attention deficit hyperactivity disorder alone or with associated disorders may require some form of remedial education since their behavioral problems have interfered with learning.

Joseph Biederman

Bibliography. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed., American Psychiatric Association, Washington, DC, 1994; J. Biederman et al., Gender differences in a sample of adults with attention deficit hyperactivity disorder, *Psych. Res.*, 53:13–29, 1994; J. Biederman, J. Newcorn, and S. Sprich, Comorbidity of attention deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders, *Amer. J. Psych.*, 148:564–577, 1991; S. V. Faraone et al., Attention deficit hyperactivity disorder in adults: An overview, *Biol. Psych.*, 48:9–20, 2000; S. V. Faraone and J. Biederman, The neurobiology of attention deficit hyperactivity disorder, in D. S. Charney, E. J. Nestler, and B. S. Bunney (eds.), *Neurobiology of Mental Illness*, pp. 788–801, Oxford,

-S

New York, 1999; L. Goldman et al., Diagnosis and treatment of attention-deficit/hyperactivity disorder in children and adolescents, *JAMA*, 279:1100–1107, 1998; T. Spencer, J. Biederman, and T. Wilens, Pharmacotherapy of ADHD: A life span perspective, in *American Psychiatric Press Review of Psychiatry*, pp. 87–128, American Psychiatric Association, Washington, D.C., 1997; T. Wilens, Does the medicating ADHD increase or decrease the risk for later substance abuse?, *Rev. Brasil. Psychiat.*, 25:127–128, 2003; T. Wilens et al., Does stimulant therapy of attention deficit hyperactivity disorder beget later substance abuse? A meta-analytic review of the literature, *Pediatrics*, 111:179–185, 2003.

Attenuation

The reduction in level of a transmitted quantity as a function of a parameter, usually distance. It is applied mainly to acoustic or electromagnetic waves and is expressed as the ratio of power densities. Various mechanisms can give rise to attenuation. Among the most important are geometrical attenuation, absorption, and scattering.

For unconfined radiation from a point source in free space, the power density (watts per square meter) decreases in proportion to the square of the distance, for purely geometrical reasons. The power densities, I_1 and I_2, at distances r_1 and r_2 from the source, are related by Eq. (1).

$$I_2 = I_1 \left(\frac{r_1}{r_2}\right)^2 \qquad (1)$$

See INVERSE-SQUARE LAW.

If the signal, in a parallel beam so that there is no geometrical attenuation, passes through a lossy medium, absorption reduces the power level, I, exponentially with distance, x, according to Eq. (2), where a is the attenuation coefficient.

$$I(x) = I(0)e^{-ax} \qquad (2)$$

See ABSORPTION; ABSORPTION OF ELECTROMAGNETIC RADIATION; SOUND ABSORPTION.

Scattering is said to occur if the power is not absorbed in the medium but scattered from inhomogeneities. *See* SCATTERING OF ELECTROMAGNETIC RADIATION.

More complicated situations occur with guided waves, such as acoustic waves in pipes or electromagnetic waves in transmission lines or waveguides, where absorption may take place and irregularities may cause reflection of some power. *See* TRANSMISSION LINES; WAVEGUIDE.

In electric circuits, constituent elements are often described as attenuators when they reduce the level of signals passing through them. *See* ATTENUATION (ELECTRICITY).

Attenuation is usually measured in terms of the logarithm of the power ratio, the units being the neper or the decibel. *See* DECIBEL; NEPER. A. E. Bailey

Attenuation (electricity)

The exponential decrease with distance in the amplitude of an electrical signal traveling along a very long uniform transmission line, due to conductor and dielectric losses. If the peak voltage at the sending end of the transmission line is denoted by V_0, the peak voltage at a distance x from the sending end is given by Eq. (1), where α is the attenuation constant of the line.

$$V_x = V_0 e^{-\alpha x} \qquad (1)$$

Characteristic impedance. Transmission-line theory shows that the input impedance of such a line has a purely resistive value Z_0 at high frequencies, where Z_0 is the characteristic impedance of the line. If the line is terminated at any point by a resistance of value Z_0, all of the power that reaches the termination will be dissipated and the line is said to be matched. If the line is terminated by any other impedance Z_t, part of the incident voltage will be reflected and the reflection coefficient Γ of the termination is defined as the ratio of the reflected voltage to the incident voltage. The value of Γ is shown by Eq. (2). Thus,

$$\Gamma = \frac{Z_t - Z_0}{Z_t + Z_0} \qquad (2)$$

when $Z_t = 0$, $\Gamma = -1$; when $Z_t = 3Z_0$, $\Gamma = 0.5$; and when Z_t approaches infinity, Γ approaches $+1$. *See* ELECTRICAL IMPEDANCE; IMPEDANCE MATCHING; REFLECTION AND TRANSMISSION COEFFICIENTS.

Neper and decibel. Electronic engineers usually express power ratios in a logarithmic manner, using either nepers or decibels (dB) as in Eqs. (3) and (4),

$$\text{Power ratio in nepers} = {}^1\!/_2 \log_e \frac{P_1}{P_2} \qquad (3)$$

$$\text{Power ratio in decibels} = 10 \log_{10} \frac{P_1}{P_2} \qquad (4)$$

where P_1 and P_2 are the power levels to be compared. The advantages of using the decibel or neper, instead of just P_1/P_2, are that very large power ratios can be expressed by conveniently small numbers, and the overall gain or loss of a cascade of matched networks can be found by addition and subtraction instead of multiplication. *See* DECIBEL; NEPER.

The well-known relationship between napierian and common logarithms yields Eq. (5).

$$\begin{aligned}&\text{Power ratio in decibels}\\ &\qquad = 8.686 \times \text{power ratio in nepers}\end{aligned} \qquad (5)$$

Insertion loss. When a generator with a reflection coefficient Γ_G is connected directly to a load with a reflection coefficient Γ_L, let the power dissipated in the load be denoted by P_3. Suppose now that a network, with two connectors called ports, is inserted between the generator and the load, and let this reduce the power dissipated in the load to P_4. Then, the insertion loss of this two-port network is given in decibels by Eq. (6).

$$L = 10 \log_{10} \frac{P_3}{P_4} \qquad (6)$$

Definition of attenuation. The attenuation of the two-port network is defined as the insertion loss when both the source and load are perfectly matched, that is, when $\Gamma_G = 0$ and $\Gamma_L = 0$. Insertion loss depends on the values of Γ_G and Γ_L as well as on the characteristics of the two-port network, whereas attenuation is a property only of the two-port network.

Returning now to the situation depicted by Eq. (1), the power entering the line is $(V_0)^2/2Z_0$, and the power at a distance x from the sending end is $(V_x)^2/2Z_0$; so, using Eq. (3), the attenuation in nepers over the distance x is seen to be given by Eq. (7).

$$A = {}^1\!/_2 \log_e \frac{(V_0)^2/2Z_0}{(V_x)^2/2Z_0} = {}^1\!/_2 \log_e(e^{2\alpha x}) = \alpha x \qquad (7)$$

Thus, α gives the attenuation of the line in nepers per unit length; attenuation in decibels per unit length is seen from Eq. (5) to be 8.686α.

Types of attenuators. Attenuators find numerous applications, typical examples being: in a signal generator, to vary the amplitude of the output signal; and in the input line to a television receiver that is very close to a television transmitter, so that overloading can be avoided. *See* SIGNAL GENERATOR; TELEVISION RECEIVER.

Attenuators for the dc (steady voltage) to very high frequency (VHF) range (frequencies from 0 to 300 MHz) often contain resistors arranged in T or π configurations (**Fig. 1***a*). The resistor values are chosen so that the required attenuation is obtained without changing the matching conditions. When a single T section is required to give an attenuation of A_t decibels and be matched to input and output lines of characteristic impedance Z_0, the design equations for the resistors R_1 and R_2 are Eqs. (8) and (9), where k is given by Eq. (10). Thus, when $Z_0 = 50\ \Omega$ and an

$$R_1 = Z_0 \frac{1-k}{1+k} \qquad (8)$$

$$R_2 = 2Z_0 \frac{k}{1-k^2} \qquad (9)$$

$$k = 10^{-A_t/20} \qquad (10)$$

attenuation of 10 dB is needed, $R_1 = 25.975\ \Omega$ and $R_2 = 35.136\ \Omega$.

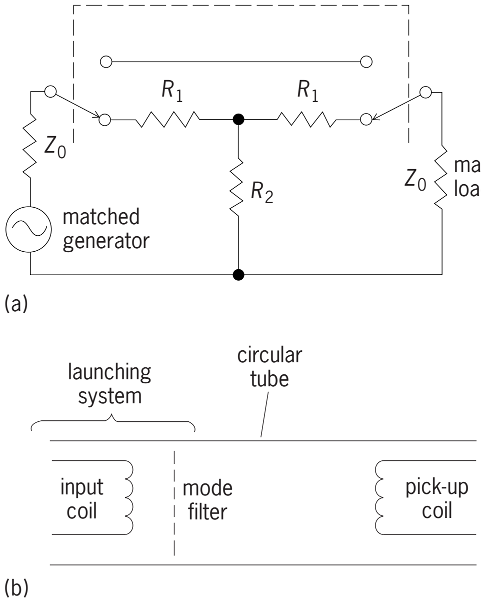


Fig. 1. Types of attenuators. (*a*) One-section T-type attenuator that can be switched in or out of a circuit between a matched generator and a matched load. (*b*) Piston attenuator.

Fixed coaxial attenuators can be designed to operate from dc to frequencies of more than 40 GHz. Various constructional techniques are used; for example, a thin resistive film can be deposited on the inner conductor, or a resistive film can be deposited on a substrate that is furnished with suitable launching, receiving, and earth electrodes.

Piston attenuators (sometimes called waveguide-beyond-cutoff attenuators) are used at both intermediate and microwave frequencies (Fig. 1*b*). The attenuation is varied by altering the separation between the two coils. The circular tube acts as a waveguide beyond cutoff, and the launching system is designed so that only one mode is excited in it. Assuming that the inner surface of the tube has perfect conductivity, an increase in the coil separation from x_1 to x_2 produces an attenuation change in decibels given by Eq. (11), where r is the radius of the cylinder, λ is

$$A_p = 8.686 \times 2\pi(x_2 - x_1)\left[\left(\frac{S_{\mathrm{nm}}}{2\pi r}\right)^2 - \frac{1}{\lambda^2}\right]^{1/2} \qquad (11)$$

the free-space wavelength of the applied signal, and S_{nm} is a constant which depends upon the mode that is excited. The TE_{11} mode is usually employed, and in this case $S_{\mathrm{nm}} = 1.84118$. A piston attenuator has a very high dynamic range and a linear decibel scale. When great care is taken in the design and manufacture, an accuracy as high as 0.0002 dB per 10 dB can be achieved over a 120-dB range. *See* WAVEGUIDE.

A variable waveguide attenuator can be produced by moving a lossy vane either sideways across the waveguide or into the waveguide through a longitudinal slot.

The rotary vane attenuator is a very popular instrument. At the input end, there is a rectangular-to-circular waveguide taper containing a fixed lossy vane perpendicular to the incident electric vector. The central section contains a lossy vane diametrically across a circular waveguide that can be rotated, and the output section is a mirror image of the input section. When the central vane is at an angle θ relative to the two fixed end vanes, the total attenuation in decibels is given by Eq. (12), where A_0 is the residual attenuation when all three vanes lie in the same plane.

$$A_{\mathrm{rva}} = 40 \log_{10}(\sec\theta) + A_0 \qquad (12)$$

Measurement of attenuation. Many different techniques for measuring attenuation have been devised. The power-ratio method is widely used. The simplest configuration requires only a stable well-matched filtered source and a well-matched low-drift power meter. Let the measured power be P_5 when the source is connected directly to the power meter and P_6 when the device under test is inserted between them. Then the attenuation of the device under test in decibels is given by Eq. (13). Insensitivity to

$$A_{\mathrm{dut}} = 10 \log_{10} \frac{P_5}{P_6} \qquad (13)$$

source power variations can be achieved by using two power meters (**Fig. 2**). A Kelvin-Varley voltage divider is adjusted to settings D_1 and D_2 that give nulls before and after inserting the device under test. Then the attenuation of the device under test is given by

$$A'_{\mathrm{dut}} = 10 \log_{10} \frac{D_1}{D_2} \qquad (14)$$

Eq. (14). If sufficient care is taken, this configuration will yield very accurate results over a 30-dB range.

Substitution methods of attenuation measurement are very popular, and the following techniques can be used: (1) radio-frequency (rf) substitution, for example, comparison with an accurate rotary vane attenuator; (2) intermediate-frequency substitution, for example, comparison with an i-f piston attenuator; (3) audio-frequency substitution, for example, comparison with an inductive voltage divider; and (4) dc substitution, for example, comparison with a Kelvin-Varley voltage divider. *See* INDUCTIVE VOLTAGE DIVIDER.

The device under test can be connected either in series or in parallel with the reference standard. For example, in a manually switched parallel i-f substitution system (**Fig. 3**), the device under test is inserted between matched isolators, and a continuous-wave signal at the required measurement frequency f is passed through it. This signal then has its frequency changed to, say, 30 MHz by a linear mixer that is driven by a local oscillator operating at a frequency of $f \pm 30$ MHz. A stable 30-MHz continuous-wave oscillator energizes the piston attenuator and, at each setting of the device under test, the piston is adjusted until the reading on the output voltmeter is the same in both positions of the switch. A system of this type has a dynamic range of about 100 dB.

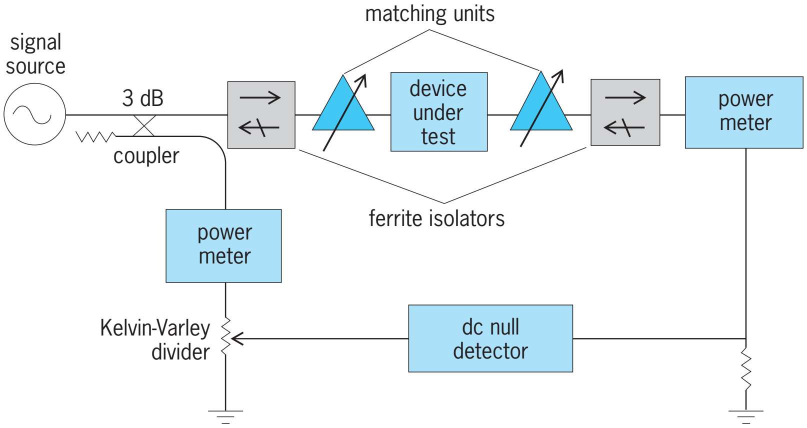


Fig. 2. **Dual-channel power ratio system for attenuation measurement.**

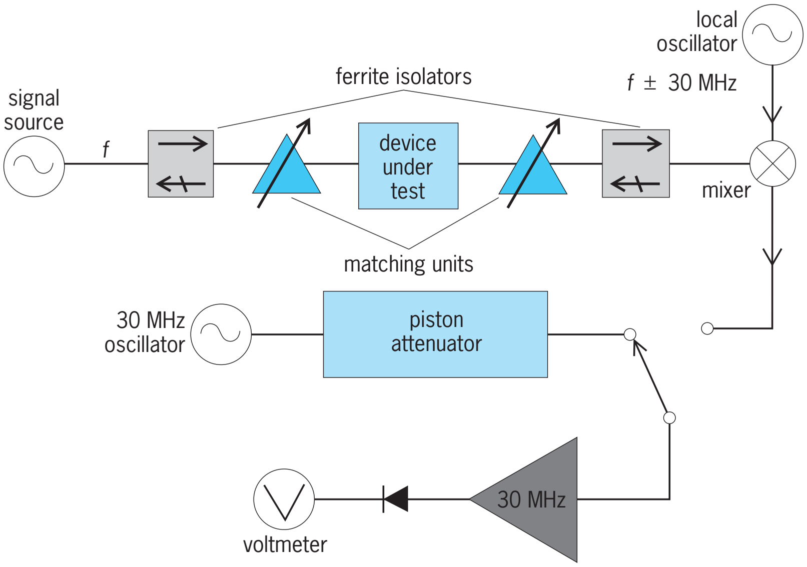


Fig. 3. **Parallel i-f substitution system for attenuation measurement.**

When extremely high dynamic range (up to 160 dB) is needed, this type of receiver is inadequate. However, the necessary extra sensitivity can be obtained by employing a dual-channel phase-sensitive detection system, which can extract signals that are deeply buried in noise. In one configuration (**Fig. 4**), a synthesized signal source and a synthesized local oscillator are operated 10 kHz apart, so that, after mixing, 10-kHz signals are produced in both channels. Isolation amplifiers with very high reverse attenuation are used to eliminate mixer-to-mixer leakage. One channel provides a clean reference signal for a lock-in analyzer, and the measurement process takes place in the other channel, which also feeds this analyzer. The lock-in analyzer contains in-phase and quadrature phase-sensitive detectors, whose outputs are $V \cos \phi$ and $V \sin \phi$, where V is the peak value of the 10-kHz signal emerging from the measurement channel, and ϕ is the phase difference between the signals in the two channels. These outputs are combined in quadrature to yield a direct voltage V_{out}, given by Eq. (15). Thus, V_{out} is directly related

$$V_{\mathrm{out}} = \{(V \cos \phi)^2 + (V \sin \phi)^2\}^{1/2} = V \qquad (15)$$

to the measurement signal V and is independent of the phase difference. Both a gage-block attenuator and an inductive voltage divider are placed in the measurement channel. When the device under test is inserted in this channel, appropriate amounts of attenuation are removed from both those devices to restrict the output voltage change to 20 dB (the maximum range over which the lock-in analyzer is highly linear). The attenuation in decibels through the device under test is given by Eq. (16), where $V_{\mathrm{out},1}$ and

$$A_{\mathrm{out}} = 20 \log_{10} \frac{V_{\mathrm{out},1}}{V_{\mathrm{out},2}} + A_{\mathrm{gba}} + A_{\mathrm{ivd}} \qquad (16)$$

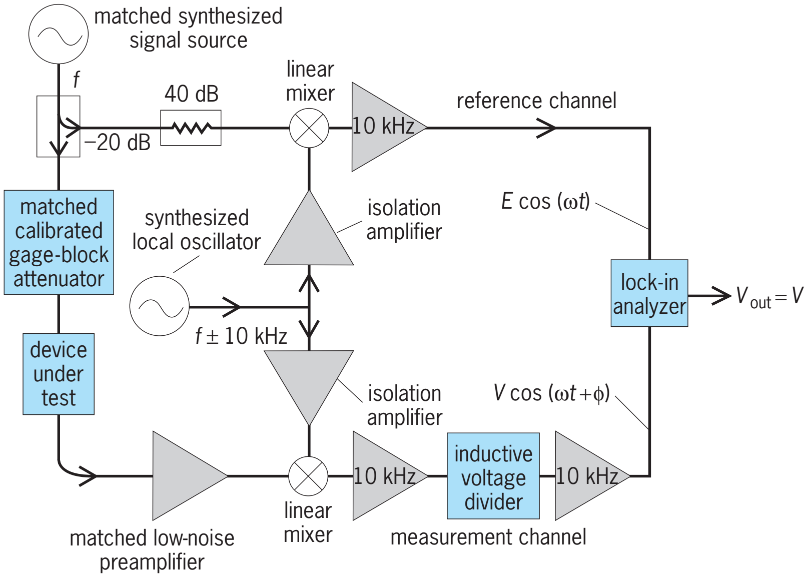


Fig. 4. Voltage-ratio-plus-gage-block system for attenuation measurements up to 160 dB. (*After G. J. Kilby, T. A. J. Smith, and F. L. Warner, The accurate measurement of high attenuation at radio frequencies, IEEE Trans. Instrum. Meas., 44:308–311, 1995*)

$V_{\mathrm{out},2}$ are the output voltages at the datum and calibration settings, respectively, and A_{gba} and A_{ivd} denote the attenuations in decibels removed from the gage-block attenuator and the inductive voltage divider when the device under test is inserted. The gage-block attenuator can be calibrated against the inductive voltage divider by using series substitution. To achieve accurate results up to 160 dB, careful attention must be paid to the elimination of leakage and circulating earth currents.

Low values of attenuation can be determined accurately by making reflection coefficient measurements on the device under test with a sliding short behind it. Several bridge techniques for measuring attenuation have been devised. Attenuation changes of less than 10^{-4} dB can be readily seen on a magic T bridge with a superheterodyne null detector. *See* MICROWAVE IMPEDANCE MEASUREMENT.

The attenuation in a waveguide can be found by making Q measurements on resonant sections of different lengths. *See* CAVITY RESONATOR; Q (ELECTRICITY).

The microwave reflection coefficient of a superconducting ring closed by a Josephson junction varies periodically as the current through an adjacent wire is increased. By using this effect (by connecting the device under test in series with the adjacent wire), discrete values of attenuation can be found very accurately by simply counting the number of zero crossovers and making use of a table of Bessel function roots. *See* JOSEPHSON EFFECT; SUPERCONDUCTIVITY.

When only moderate accuracy (on the order of ±0.5 dB) is required over a wide frequency range, a leveled swept source can be connected to the device under test, and the emerging signal can be fed to a diode detector that is followed by a logarithmic amplifier and oscilloscope. With this technique, resonances which can easily be missed using point-by-point methods are clearly revealed, and the effects of adjustments to the device under test can be seen immediately over a wide band. *See* AMPLITUDE-MODULATION DETECTOR.

Network analyzers became very popular after 1967. These instruments yield both the magnitude and phase angle of the transmission and reflection coefficients of the device under test over a wide frequency range. By using ingenious calibration and computer-correction techniques, high accuracy can be achieved. To obtain the phase information, a second channel is needed to provide a reference signal. In most network analyzers, mixers are employed and high-dynamic-range instruments of this type are now commercially available for use at frequencies up to 100 GHz. Considerable effort has been devoted since 1972 to six-port network analyzers in which the phase information is deduced solely from power measurements. *See* TRANSMISSION LINES.

Frank L. Warner

Bibliography. A. E. Bailey (ed.), *Microwave Measurements*, 2d ed., 1989; G. H. Bryant, *Principles of Microwave Measurements*, 1988; G. F. Engen, *Microwave Circuit Theory and Foundations of Microwave Metrology*, 1992; I. Kneppo, *Microwave Measurement by Comparison Methods*, 1988; S. R. Pennock and P. R. Shepherd, *Microwave Engineering with Wireless Applications*, Macmillan, London, 1998.

Audio amplifier

An electronic circuit for amplification of signals within or somewhat beyond the audio frequency range (generally regarded as 20 to 20,000 Hz). Audio amplifiers may function as voltage amplifiers (sometimes called preamplifiers), power amplifiers, or both. In the last case, they are often called integrated amplifiers. *See* POWER AMPLIFIER; VOLTAGE AMPLIFIER.

The function of integrated amplifiers (or of the combination of separate voltage amplifiers and power amplifiers used together) is to amplify a weak signal, such as from a microphone, phonograph pickup, tape player, radio tuner, or compact disc player, to a level capable of driving a loudspeaker or other type of transducer such as headphones at the desired sound level. Power amplifiers may have power ratings ranging from less than 1 W to several hundreds of watts. Stereo amplifiers consist of two identical, but electrically independent, amplifier circuits housed in a single chassis, often sharing a common power supply. Audio amplifiers are commonly constructed with solid-state devices (transistors and integrated circuits), although some amplifiers using vacuum tubes as the active, amplifying devices are still manufactured. *See* INTEGRATED CIRCUITS; TRANSISTOR; VACUUM TUBE.

Classification. Audio power amplifiers, or the power sections of integrated amplifiers, are classified in accordance with the manner in which the output stages conduct current. Class A amplifiers conduct a constant current regardless of whether an audio signal is applied to their input or not. They are therefore relatively inefficient. Class AB amplifiers conduct a moderate amount of current even in the absence of an input signal, but the current increases as signal levels increase. Most modern audio power amplifiers for high-fidelity applications employ this type of circuitry. Class B circuits conduct minute amounts of current in the absence of an input signal and are therefore the most efficient in terms of power dissipation versus power delivered to the loudspeaker or other connected load. Other amplification classes have been developed by specific manufacturers, such as class G and class H, each of which attempts to improve efficiency beyond that attainable with either class AB or class B circuits.

Performance characteristics. The ideal amplifier delivers an output signal that, aside from its higher power level, is identical in relative spectral content to the input signal. Normally, various forms of distortion are generated by the amplifier, such as harmonic distortion (multiples of the desired signal frequency), intermodulation distortion (spurious sum or difference frequencies created when multiple tones are applied to the amplifier simultaneously, as in the case of music or speech amplification), and transient intermodulation distortion (caused by rapid fluctuations of the input signal level). All forms of distortion are measured as percentages of the desired signal amplitude. Generally, distortion levels of under 1% or 0.5% are considered to be low enough for high-fidelity applications. *See* DISTORTION (ELECTRONIC CIRCUITS); FIDELITY.

Other parameters used to define an amplifier's characteristics include frequency response and signal-to-noise ratio (S/N). The frequency response is the range of frequencies that the amplifier can handle, usually quoted with a tolerance in decibels, for example: "frequency response: 20 to 20,000 Hz, ±0.5 dB." Signal-to-noise ratio, also quoted in decibels, is indicative of the amount of residual noise generated by the amplifier itself, as compared with the desired output signal level. Signal-to-noise levels greater than 70 dB are generally considered to be acceptable, although some amplifiers offer much better values. *See* ELECTRICAL NOISE; RESPONSE; SIGNAL-TO-NOISE RATIO.

Speaker impedance. Although not related to quality of sound reproduction, other important amplifier specifications include a listing of suitable loudspeaker impedances that may be safely connected to the amplifier. Loudspeakers generally have impedances of 4 or 8 ohms. Most audio amplifiers can drive either of these impedances, often delivering higher power when the lower-impedance loudspeakers are connected. If the loudspeaker impedance dips to too low a value, however, the amplifier may not be able to deliver the higher current required and may either fail or cause its protection circuits to temporarily disable the amplifier. While many amplifiers can power more than one loudspeaker per channel, wiring multiple loudspeakers in parallel across the output terminals of an audio amplifier lowers the net impedance into which the amplifier must operate. Thus, if two 8-ohm loudspeakers are connected in parallel, the net impedance across the amplifier terminals becomes 4 ohms. *See* AMPLIFIER; ELECTRICAL IMPEDANCE; LOUDSPEAKER.

Leonard Feldman

Bibliography. R. Angus, L. Feldman, and N. Eisenberg, *The New World of Audio: A Music Lover's Guide*, 1982; G. Ballou, *Handbook for Sound Engineers:*, 2d ed., 1991; M. Clifford, *Modern Audio Technology*, 1992.

Audiometry

The quantitative assessment of individual hearing, both normal and defective. In typical tests, pure tones are presented through headphones, though some tests use speech instead of tones. In bone conduction tests, tones are presented through a vibrator placed on the head. Audiometric tests may serve various purposes, such as investigation of auditory fatigue under noise conditions, human engineering study of hearing aids and communication devices, screening of individuals with defective hearing, and diagnosis and treatment of defective hearing. In all of these situations, individual hearing is measured relative to defined standards of normal hearing (**Fig. 1**).

Audiometer. The pure-tone audiometer is the instrument used most widely in individual hearing measurement. It is composed of an oscillator, an amplifier, and an attenuator to control sound intensity. For speech tests of hearing, called articulation tests, word lists are reproduced on records or tape recorders. Measurements of detectability or

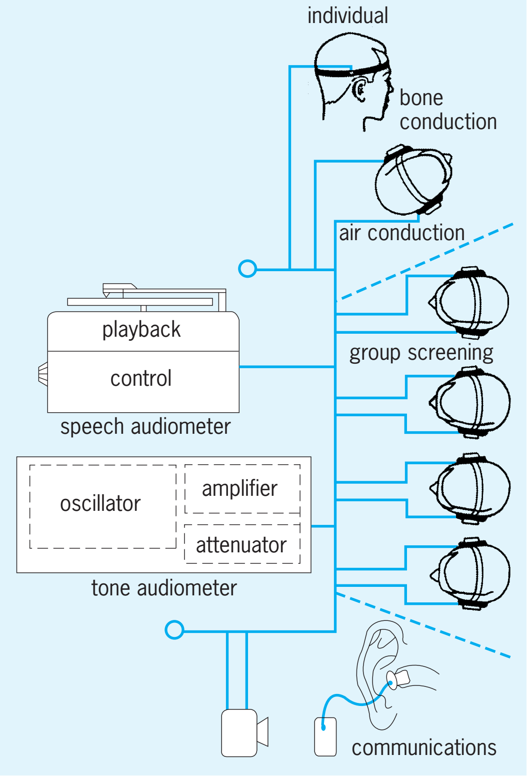


Fig. 1. Test equipment for tone and speech audiometry as applied to individual hearing measurement.

intelligibility can be made by adjusting the intensity of the test words. To make bone conduction tests, signals from the audiometer activate a vibrator located on the forehead or mastoid bone.

Reverberation and anechoic rooms. Scientific advance in audiometry demands careful control of all environmental sound. Two types of rooms especially constructed for research and measurement of hearing are shown in **Fig. 2**, the random diffusion, or reverberation, chamber and the anechoic room. In the reverberation chamber, sounds are randomly reflected from heavy nonparallel walls, floor, and ceiling surfaces. In the anechoic room, the fiber glass wedges absorb all but a small percentage of the sound. Such conditions make possible precise determination of the limits and variation of normal and defective hearing for all types and conditions of sound transmission.

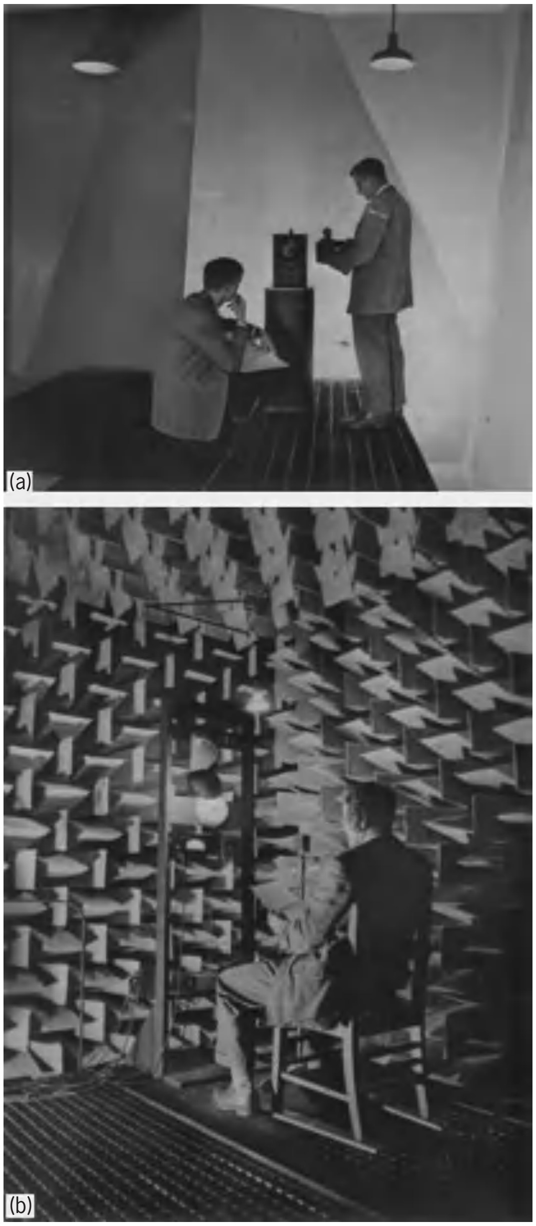


Fig. 2. Special environments for audiometry. (*a*) Random diffusion chamber for measuring hearing under controlled noise conditions. (*b*) Anechoic soundproofed room lined with fiber glass wedges. (*Courtesy of J. P. Egan*)

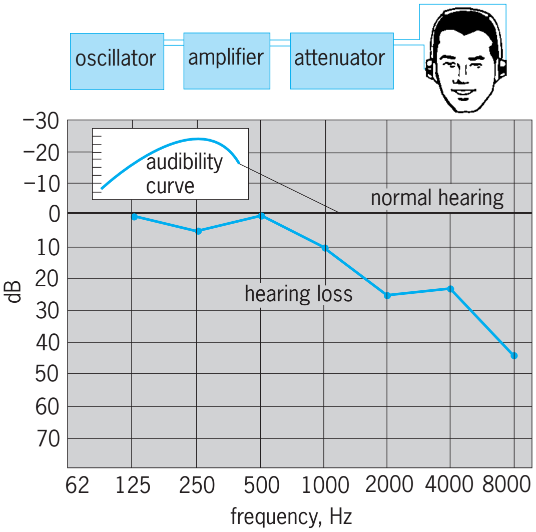


Fig. 3. Audiogram for determining pure-tone hearing loss at various frequency levels.

Audiogram. The measurement of hearing loss for pure tones is represented by the audiogram (**Fig. 3**). Sounds of different frequencies are presented separately to each ear of the individual, and the absolute threshold for each frequency is determined. The absolute threshold is the lowest intensity that can be detected by the individual who is being tested.

In clinical audiometry the status of hearing is expressed in terms of hearing loss at each of the different frequency levels. In the audiogram the normal audibility curve, representing absolute thresholds at all frequencies for the normal ear, is represented as a straight line of zero decibels. Amount of hearing loss is then designated as a decibel value below normal audibility, as defined by the American National Standards Institute (ANSI) or the International Organization for Standardization (ISO). The audiogram in Fig. 3 reveals a hearing loss for tones above 500 Hz. Automatic audiometers are now in use which enable one to plot an audiogram for oneself.

Articulation tests. Articulation tests are tests used to assess hearing and loss of hearing for speech. The threshold of intelligibility for speech is defined as the intensity level at which 50% of the words, nonsense syllables, or sentences used in the articulation test are correctly identified. The best of such articulation tests are phonetically balanced (PB) word tests, in which single-syllable words were selected so that the phonetic sounds naturally occurring in the English language would be proportionately represented in the test. The hearing loss for speech is determined by computing the difference in decibels between the individual intelligibility threshold and the normal threshold for that particular speech test. Discrimination loss for speech represents the difference between the maximum articulation score at a high intensity level (100 dB), expressed in percent of units identified, and a score of 100%. The measure of discrimination loss helps distinguish between conduction loss and nerve deafness.

Bone conduction tests. Bone conduction audiograms are compared with air conduction audiograms in order to analyze the nature of deafness. Loss in bone conduction hearing as well as air conduction hearing often indicates nerve deafness, as contrasted to middle-ear or conduction deafness. *See* EAR (VERTEBRATE).

Audiometric testing. Audiometric testing is adjusted to meet various needs of individuals. School audiometry makes possible rough determination of hearing loss in children, who can thereafter be tested with more precise techniques if some loss is found. Audiometry can be extended to assessment of hearing in very young children who cannot speak or follow directions. In such cases, certain reflexes, such as the psychogalvanic skin reflex, are conditioned to sound. Subsequently the presence or absence of the conditioned reflex can be used to indicate whether or not the child hears a particular sound. Young children can also be trained to press buttons to get a reward when sounds are present. Thereafter the occurrence of the learned response in relation to a test sound is used to determine the threshold of audibility.

Automatic audiometry. The main theoretical and technological issues of audiometry have focused on the design of psychometric and electrophysiological procedures, automation of testing, and application of computers to audiometer design.

Békésy audiometry. In 1974 G. von Békésy designed the first observer-controlled threshold-testing procedure. In Békésy audiometry, the observer indicates that he or she hears a sound by pressing a button. When the button is held down, the sound intensity is gradually reduced by a motor-driven attenuator until the observer releases the button, indicating that the sound is no longer heard. At this point, the sound intensity is gradually increased. As the observer makes these adjustments, the frequency of the test tone is advanced from 100 to 10,000 Hz and an audiogram is automatically charted. The audiogram in **Fig. 4** shows a marked hearing loss for tones above 3000 Hz.

Electrical response audiometry. Standard audiometric methods can be used to test children as young as 4 or even 3 years. The need to assess hearing in younger children and infants, and to be able to test some retarded and handicapped individuals, has prompted the development of automatic audiometric methods. Some of these methods measure reflexes and automatic reactions, such as the startle response and changes in heart rate.

More reliable and consistent than reflexes are various electrical measures of the auditory system's response to sounds. By placing electrodes at or near the eardrum, it is possible to measure the electrocochleogram, which is the auditory nerve's response to sounds. If electrodes are attached to the surface of the skull, it is possible to measure electrical potentials generated in the brainstem and cerebral cortex. Abnormalities in these electrical audiograms can indicate defective functioning in the ear or in the auditory nervous system.

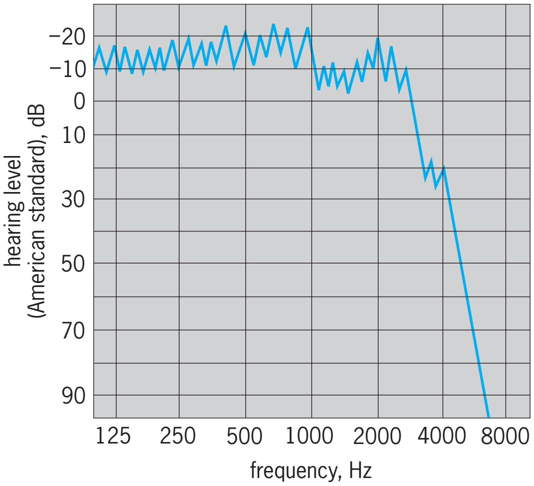


Fig. 4. Audiogram obtained by Békésy audiometric method of observer-controlled threshold testing.

Computer systems. Conventional data-processing instruments have been developed to program sequences of discrete sound stimuli that the observer must judge separately. These devices have been designed so that the observer should be limited to making only passive judgments, while the machine is programmed to do all the work. *See* COMPUTER.

Hybrid computer instruments have been constructed for the utilization of observer-controlled methods of audiometric measurement (**Fig. 5**). In

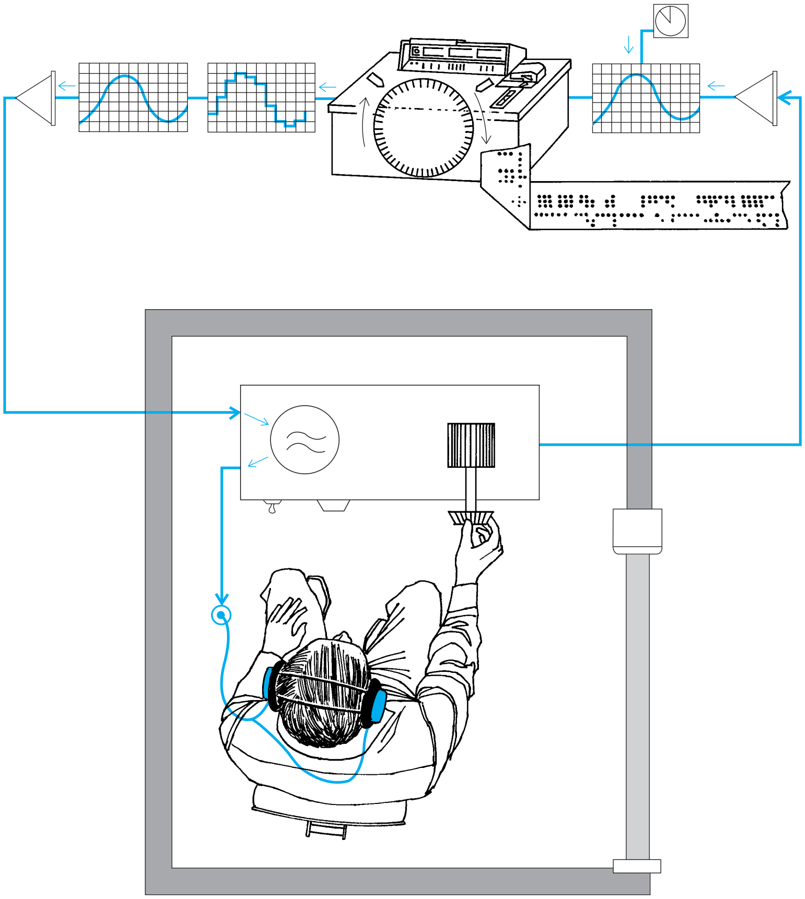


Fig. 5. Basic design of a hybrid computer system for dynamic audiometric testing.

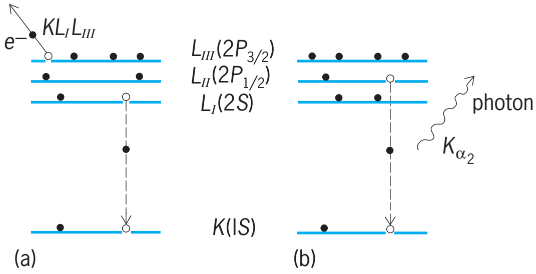
these systems the subject uses a hand-controlled adjustment to control either intensity or frequency of a variable oscillator system. Dynamic programming is used to generate predetermined changes in the frequency and intensity of the tones heard by the subject. The computer program can require the subject to compensate or negate changes in sound, to track positively the loudness of a second sound, or to generate sounds to a point of zero loudness. For the future such methods will permit the measurement of hearing that can be associated directly with impedance measures of ear reactivity. *See* ELECTRODERMAL RESPONSE; HEARING (HUMAN); HEARING IMPAIRMENT; LOUDNESS. Lawrence E. Marks

Bibliography. H. A. Beagley (ed.), *Audiology and Audiological Medicine*, 1981; H. Kaplan et al., *Audiometric Interpretation: A Manual of Basic Audiometry*, 2d ed., 1993; F. N. Martin, *Introduction to Audiology*, 4th ed., 1990.

Auger effect

One of the two principal processes for the relaxation of an inner-shell electron vacancy in an excited or ionized atom. The Auger effect is a two-electron process in which an electron makes a discrete transition from a less bound shell to the vacant, but more tightly bound, electron shell. The energy gained in this process is transferred, via the electrostatic interaction, to another bound electron which then escapes from the atom. This outgoing electron is referred to as an Auger electron and is labeled by letters corresponding to the atomic shells involved in the process. For example, a KL_IL_{III} Auger electron corresponds to a process in which an L_I electron makes a transition to the K shell and the energy is transferred to an L_I electron (**illus**. *a*). By the conservation of energy, the Auger electron kinetic energy E is given by $E = E(K) - E(L_I) - E(L_{III})$ where $E(K,L)$ is the binding energy of the various electron shells. Since the energy levels of atoms are discrete and well understood, the Auger energy is a signature of the emitting atom. *See* ELECTRON CONFIGURATION; ENERGY LEVEL (QUANTUM MECHANICS).

The other principal process for the filling of an inner-shell hole is a radiative one in which the transition energy is carried off by a photon (illus. *b*). Inner-shell vacancies in elements with large atomic number correspond to large transition energies and usually decay by such radiative processes; vacancies in elements with low atomic number or outer-shell vacancies with low transition energies decay primarily by Auger processes. *See* ATOMIC STRUCTURE AND SPECTRA.



Two principal processes for the filling of an inner-shell electron vacancy. (*a*) Auger emission; a KL_IL_{III} Auger process in which an L_I electron fills the K-shell vacancy with the emission of a KL_IL_{III} Auger electron from the L_{III} shell. (*b*) Photon emission; a radiative process in which an L_{II} electron fills the K-shell vacancy with the emission of a $K_{\alpha 2}$ photon.

Auger electron spectroscopy is an important tool in the analysis of the near-surface region of solids. Inner-shell vacancies in the atoms of solids are created by incident radiation, usually an energetic electron beam, and the energy spectrum of the outgoing Auger electrons is then used to determine the types of atoms present, and hence the elemental composition of the material. In solids, the outgoing Auger electron has a high probability of undergoing an inelastic or energy loss event which changes the Auger electron energy. The probability of an inelastic process is characterized by a mean free path, a thickness in which approximately 63% of the Auger electrons undergo an energy changing event. The mean free path can vary from approximately 0.5 nanometer to 100 nm, depending on the electron energy and the solid; thus Auger electron spectroscopy is useful for measurements of the elemental composition of a material in its first 0.5–100 nm. Elemental composition as a function of depth is determined by using Auger electron spectroscopy combined with some erosion process which slowly removes the surface layers; the Auger spectrum measured at various stages of the erosion process yields a depth profile of the elemental composition of the material. *See* ELECTRON SPECTROSCOPY; SURFACE PHYSICS. Leonard C. Feldman

Bibliography. P. Auger, The compound photoelectric effect, *J. Phys. Radium*, 1925; G. Cubiotti, G. Mundio, and K. Wandett (eds.), *Auger Spectroscopy and Electronic Structure*, 1989.

Auger electron spectroscopy

Auger electron spectroscopy (AES) is a widely used technique that detects the elements in the first atomic layers of a solid surface. Although many elements can be detected, hydrogen usually cannot be observed. Excellent spatial resolution can be achieved. Auger electron spectroscopy is important in many areas of science and technology, such as catalysis, electronics, lubrication, and new materials, and also understanding chemical bonding in the surface region. Auger spectra can be observed with gas-phase species.

Basic principles. In Auger electron spectroscopy an electron beam, usually 2–20 kV in energy, impinges on a surface, and a core-level electron in an atom is ejected. An electron from a higher level falls into the vacant core level. Two deexcitation processes are possible: An x-ray can be emitted, or a third electron is ejected from the atom (**Fig. 1**). This electron is an Auger electron; the effect is named after its discoverer, Pierre Auger. Auger electrons for surface analysis can be created from other sources, such as, x-rays, ion beams, and positrons. K capture (radioactive decay) is another source of Auger

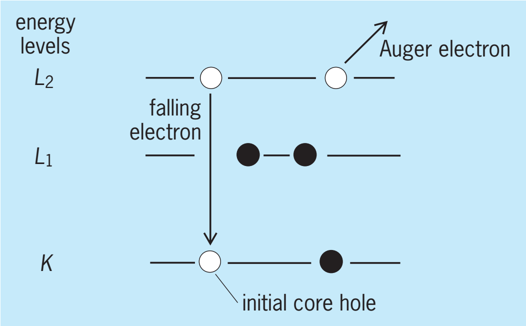


Fig. 1. **Schematic diagram of an Auger electron being ejected from a carbon atom for a KL_2L_2 transition. Round forms are electrons.**

electrons. The Auger yield (the ratio of Auger electrons to the number of core holes created) depends upon the element, the initial core level, and the excitation conditions. *See* AUGER EFFECT.

The surface sensitivity of Auger electron spectroscopy is based upon the fact that electrons with kinetic energy in the range of 30–2500 eV have an inelastic mean free path (IMFP) of about 0.5–3 nanometers for most materials. The inelastic mean free path is defined as the average distance that an electron will travel before it undergoes an inelastic collision with another electron or nucleus of the matrix constituent atoms. Very few Auger electrons created at depths greater than two to three times the IMFP will leave the bulk with their initial kinetic energy. Those electrons formed at greater depths undergo collisions with atoms and electrons in the bulk and lose part or all of the initial creation energy. In this energy range, all elements with an atomic number of 3 or more have detectable Auger electrons. Most elements can be detected to 1 atomic percent in the analysis volume, and relative atomic amounts usually can be determined.

X-ray notation is used to signify Auger transitions. For example, if a K ($1s$) electron in silicon is ejected and an L_3 ($2p_{3/2}$) electron falls into the K-level hole and another L_3 is ejected as the Auger electron, the transition is KL_3L_3. If a valence electron is involved, it often is denoted with a V. The energy relationship for Auger electrons is given by

$$E_{\mathrm{KE}}(j,k,l) = E_{\mathrm{KE}}(j) - E_{\mathrm{KE}}(k) - E_{\mathrm{KE}}(l) - U$$

where $E_{\mathrm{KE}}(j,k,l)$ is the energy of the Auger electron, $E_{\mathrm{KE}}(j)$ is the binding energy of the electron ejected during the initial ionization event, $E_{\mathrm{KE}}(k)$ is the binding energy of the electron that falls into the empty level, $E_{\mathrm{KE}}(l)$ is the binding energy of the ejected electron, and U is the hole-hole repulsion energy. U (about 1–10 eV) takes into account that there are two empty electron states in the excited atom. Handbooks containing the spectra of most elements are used for routine elemental identification. If a transition is within the same major elemental level, for example, $M_1M_{2,3}M_{2,3}$, it is called a Coster-Kronig transition. Chemical information sometimes can be obtained with Auger electron spectroscopy; for example, the $KL_{2,3}L_{2,3}$ silicon (Si) transition energies for Si and SiO_2 are 1616 and 1608 eV, respectively. The energy difference between L_2 and L_3 cannot be distinguished with Auger electron spectroscopy in this instance. Chemical information occasionally can be observed by line shape differences, for example, between carbides and graphite. With nonconducting samples an electron imbalance sometimes occurs on the surface that changes the observed transition energies. There are several ways to overcome this charging, but in extreme instances a spectrum cannot be observed.

A spectrum for copper is shown in **Fig. 2** in both integral and differential modes. In some instances, small peaks are easier to notice with differential spectra. Derivative spectra can be obtained instrumentally or mathematically from integral spectra. When an electron beam is used to produce Auger electrons, elemental maps can be made by analyzing only small energy regions over an area. This approach is called scanning Auger microscopy (SAM). These maps indicate spatial variations of elements over a selected area in the surface region. If the electron beam is moved in one direction, elemental line scans are produced. With the electron beam, secondary electrons are produced and images can be observed in the same way as with scanning electron microscopy. *See* ATOMIC STRUCTURE AND SPECTRA; ELECTRON MICROSCOPE.

Elemental distribution into the bulk can be obtained by sputtering, or “molecular sandblasting.” This is done with a beam of gas ions usually 0.5–5 kV in energy. At selected intervals, spectra are recorded for chosen elements. This technique is used to study ion-implanted or layered materials and thin oxide layers. In some instances the depth resolution can

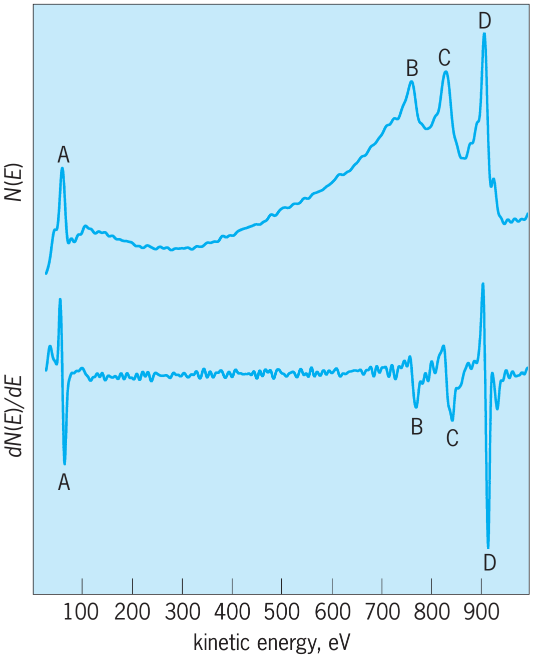


Fig. 2. **Integral (upper) and differential (lower) mode spectra of sputter-cleaned copper with the most intense peaks identified: A, $M_{2,3}M_{4,5}M_{4,5}$; B, $L_3M_{4,5}M_{4,5}$; C, $L_3M_{2,3}M_{4,5}$; D, $L_3M_{2,3}M_{2,3}$.**

be a few atomic layers with sample rotation during sputtering. Sputter profiles usually are not greater than a few micrometers.

Equipment. Auger electron spectroscopy spectra are obtained in a vacuum chamber at pressures below about 10^{-2} pascal (10^{-4} millibar), but most systems operate in the ultrahigh-vacuum (UHV) regime ($<10^{-6}$ Pa or 10^{-8} mb). These ultrahigh-vacuum chambers are constructed mostly with metal, and special handling of the materials in the system is required. In addition, specimens have to be handled with clean-room-type gloves to eliminate the transfer of oils and salts to the sample surface. Most systems have a sample introduction apparatus that permits the main chamber to remain at a low pressure when placing specimens in the analysis chamber. The samples are placed on a holder that permits their movement for analysis at specific locations. Many holders permit the sample temperature to be lowered to about 90 K (−183°C; −298°F) or raised to over 1000 K (727°C; 1340°F).

The electron beam is created with an electron gun that uses a heated filament of tungsten, a lanthanum boride (LaB_6) crystal, or a Schottky barrier filament (zirconium oxide, ZrO) coated on a single crystal face of tungsten. With the last source, the beam diameter can be below 20 nm. Current levels range from below 1 nanoamp to milliamps depending on the source. If x-rays are used to create Auger electrons, the ability to analyze small areas is lost, but chemical information can be obtained by measuring the energy difference between the x-ray photoelectrons and Auger electrons. The cylindrical mirror analyzer (CMA) or the concentric hemispherical analyzer (CHA) usually is employed to measure the energy of Auger electrons. The CMA has somewhat better throughput, but the CHA has much better energy resolution. Auger spectroscopy electrons are detected with channel or position-sensitive electron multipliers, and the signals are recorded with computers. Sputtering is done with an ion gun with current densities up to a few microamps per square centimeter. Usually an inert gas is employed at pressures of 10^{-2} Pa in the ionization region. Often, brief sputtering is used to remove atmospheric contaminants (usually containing carbon and oxygen) that are present with almost all materials that have been exposed to air. Noel H. Turner

Bibliography. D. Briggs and M. P. Seah (eds.), *Practical Surface Analysis*, 2d ed., 1990; C. L. Hedberg (ed.), *Handbook of Auger Electron Spectroscopy: A Book of Reference Data for Identification and Interpretation in Auger Electron Spectroscopy*, 3d ed., 1995.

Augite

A group of monoclinic calcic pyroxenes which have the general chemical formula $(Ca,Mg,Fe)(Mg,Fe)Si_2O_6$, in which calcium is the dominant cation in the first cation position. Monoclinic pyroxene with substantial iron or magnesium in place of calcium is called pigeonite, and has a different crystal structure from augite. Augite is generally considered to be a combination of the four end members, diopside ($CaMgSi_2O_6$), hedenbergite ($CaFe^{2+}Si_2O_6$), enstatite ($Mg_2Si_2O_6$), and ferrosilite ($Fe_2{}^{2+}Si_2O_6$), but it almost always has substantial aluminum and minor to substantial amounts of sodium, ferric iron, chromium, and titanium. These minor constituents enter as solid solutions of augite toward acmite ($NaFe^{3+}Si_2O_6$), jadeite ($NaAlSi_2O_6$), calcium Tschermak's molecule ($CaAl_2SiO_6$), ureyite ($NaCrSi_2O_6$), and $CaTiAl_2O_6$. The amount of substitution by any of these minor components depends upon the bulk composition of the rock as well as the conditions of formation. Four common varieties of augite are omphacite, which has about 50% substitution of jadeite; aegerine augite, which has substantial acmite substitution; fassaite, which has a considerable amount of calcium Tschermak's molecule in solution; and titanaugite, which is rich in titanium and has a characteristic purple color. *See* DIOPSIDE; ENSTATITE; PIGEONITE.

Augite in igneous and metamorphic rocks does not commonly occur as well-formed crystals; when it does, the crystals are dark green to black, short and prismatic. Augite crystallizes in a *C*-centered monoclinic cell, space group *C*2/*c*, and is commonly recognized by its prominent (110) pyroxene cleavages which form 87° angles. It displays typical pyroxene crystal chemistry—single chains of tetrahedral (SiO_4) groups alternating with strips containing two distinct types of octahedral sites accommodating Ca, Fe, Mg, Mn, Na, Cr, and Ti. The smaller Al cation can enter both tetrahedral and octahedral sites.

Augite occurs in both igneous and metamorphic rocks. It is nearly universal in basalts and gabbros, and occurs somewhat less frequently in less mafic igneous rocks. Alkali olivine basalts and other alkali-rich volcanic rocks contain augite which is commonly enriched in Na, Ti, Al, and Fe^{3+}. Magnesium-rich augite is a characteristic mineral in many ultramafic rocks and in rocks of the Earth's mantle. Augite and pigeonite are also rather common constituents of lunar basalts and basaltic meteorites. In metamorphic rocks, augite of a composition close to diopside occurs in marbles and in calc-silicate rocks. More compositionally complex augite occurs in high-grade metamorphosed amphibolites and granulites where it typically coexists with hornblende or orthopyroxene. The augite variety omphacite is found with magnesium-rich garnet in eclogite which is of basaltic composition, and apparently forms at very high pressures. *See* ECLOGITE; PYROXENE.

Robert J. Tracy

Bibliography. W. A. Deer et al., *Rock-Forming Minerals*, vol. 2: *Chain Silicates*, 1963; C. Klein, *Manual of Mineralogy*, 21st ed., 1993; J. J. Papike (ed.), *Pyroxenes and Amphiboles*, Mineral. Soc. Amer. Spec. Pap. 2, 1969.

Aulopiformes

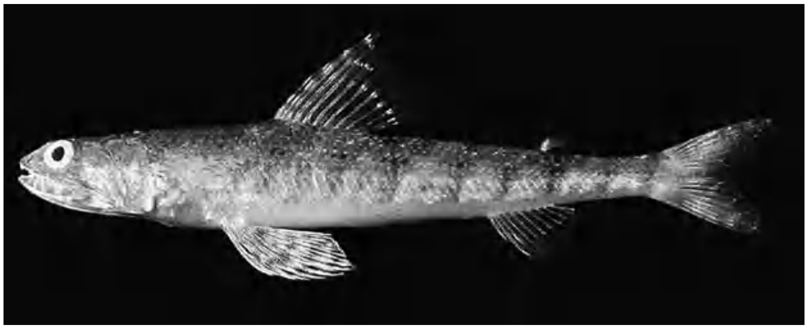
An order of about 219 species of teleost fishes ranked in four suborders. The order is distinguished by certain skeletal specializations of the gill arches that are unknown in other teleost fishes. They share characters with the order Myctophiformes (lanternfishes) to the extent that they are combined under the ordinal name Myctophiformes by some authors. In both orders the premaxilla excludes the maxilla from the gape of the mouth; the mouth is not protractile; an adipose fin is usually present; the caudal fin is forked; pelvic fins are usually abdominal; light-producing organs are often present; a swim bladder, if present, is physoclistous (not connected to the gut); and some families have hermaphroditic species with self-fertilization.

Morphological diversity. Descriptions of selected families follow to illustrate the diversity in morphology.

Giganturidae (telescope fishes). These fishes are mesopelagic in the Atlantic, Indian, and Pacific oceans). Telescope fishes are the most bizarre of the order. They have a large mouth, extending far behind the tubular eyes; a pectoral fin inserted above the gill opening; a forked caudal fin, with lower lobe greatly extended; and a greatly expandable stomach. The maximum length is 22 cm (9 in.). The family consists of one genus and two species.

Chlorophthalmidae (greeneyes). These fishes are benthic on the continental shelf and adjacent deep Atlantic, Indian, and Pacific oceans. They are moderately slender, rounded anteriorly, and compressed posteriorly; with large, lensless eyes that are directed dorsolaterally; a well-developed adipose fin; maxillary expanded post; a single supramaxilla. The tapetum of the eye is iridescent green; the top of the head is scaleless; and the caudal fin is forked, with lobes about equal length. The maximum length is about 25 cm (10 in.). The family consists of two genera and 20 species.

Synodontidae (lizardfishes). Members are benthic in the Atlantic, Indian, and Pacific oceans and, rarely, in brackish waters. The body is slender and cylindrical; head depressed; mouth large, with gape oblique and maxillary rudimentary; scales cycloid, and absent from top of head; adipose fin present, small to large; no luminous organs (see **illustration**). There are five genera and 55 species.


Orangemouth lizardfish (*Saurida flamma*). (*Photo © John E. Randall*)

Paralepididae (barracudinas). These fishes are found mid-depth to near surface in all oceans from the Arctic to the Antarctic. Scales are either present or absent; the anal fin base is long; there are moderately long pointed jaws; eyes are normal with a circular iris; an adipose fin is present; squamation varies from a scaled head, body, and vertical fins to completely naked, except for embedded scales in the lateral line; the dorsal fin is posteriorly placed; pelvic fins are abdominal; the anal fin is long; and luminous organs may be present or absent. The maximum length is 1 m (3.3 ft). There are 12 genera and 56 species.

Alepisauridae (lancetfishes). Members are found in the Atlantic, Indian, and Pacific oceans. The body is slender and subcylindrical; the scales and light organs are absent; the dorsal fin is high (as in sailfishes), extending from just behind the head to near the adipose fin; the pectoral fins are ventral, the pelvic fins are abdominal; the mouth is large, with well-developed teeth, and long palatine teeth. The maximum length is no less than 2 m (6.6 ft). There is one genus and two species.

Pelagic specializations. Ecologically there are two major groups of Aulopiforms, the benthic families and the pelagic to bathypelagic families. The latter group is confined to midwaters of deep oceans. These are predatory fishes ranging from only 12 cm (4.7 in.) in length to 2 m (6.6 ft). They have no swim bladder; no luminescent organs; a single row of small teeth on the premaxillaries, one or two rows of large teeth on the palatines (roof of mouth), and one to three rows on each dentary (mandible); and gillrakers in the form of spines or teeth. The skeleton of most is lightly ossified and the flesh may be flaccid. The lack of a swim bladder (hydrostatic organ) to maintain the fish's position in the water column is compensated to some extent by the light skeleton and the reduction or total loss of scales in many of the species. In total darkness species recognition, essential for reproduction in dioecious fishes, is usually enhanced by luminescent organs; however, this group of fishes lack the ability to produce light. For self-fertilizing hermaphroditic species, as some of these are, the lack of luminous organs is of little consequence. In this environment life is sparse. Many of these fishes are equipped with a large mouth, large knifelike teeth, and the ability to hover and make swift darts at their prey, assuring them success in obtaining a meal when the rare opportunity exists. In species lacking scales, the skin can stretch to accommodate very large prey. *See* OSTEICHTHYES.

Herbert Boschung

Bibliography. E. Bertelsen, G. Krefft, and N. B. Marshall, The fishes of the family Notosudidae, *Dana Rep.*, 86:1–114, 1976; K. E. Hartel and M. J. L. Stiassny, The identification of larval parasudis (Teleostei, Chlorophthalmidae): With notes on the anatomy and relationships of aulopiform fishes, *Breviora*, vol. 487, p. 23, 1986; R. K. Johnson, *Fishes of the Families Evermannellidae and Scopelarchidae: Systematics, Morphology, Interrelationships, and Zoogeography*, Fieldiana Zool. New Ser., vol. 12, p. 252, 1982; R. K. Johnson, Giganturidae: Development and relationships, pp. 199–201, in H. G. Moser et al. (eds.), *Ontogeny and Systematics of Fishes*, Spec. Publ. no. 1, Amer. Soc. Ichthy. Herp., 1984; R. K. Johnson and E. Bertelsen, The fishes of the family Giganturidae: Systematics, development, distribution, and aspects of biology, *Dana Rep.*, 91:1–45, 1991; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994; R. R. Rofen, Fishes of the Western North Atlantic: Families Paralepididae, Omosudidae, Anotopteridae, Evermannellidae, and Scopelarchidae, *Sears Found Mar. Res., Mem.* (Yale University), 1(5):205–481, 1966.

Aurora

An optical manifestation of a large-scale electrical discharge process which surrounds the Earth. The discharge is powered by the so-called solar wind–magnetosphere generator. The Sun continuously blows out its upper atmosphere, the corona, with a supersonic speed. This fully ionized and magnetized gas flow interacts with the Earth's magnetic field, resulting in a comet-shaped cavity (the magnetosphere) carved around the Earth, while the lines of force of the Earth's magnetic field and of the solar wind magnetic field interconnect. Electric power of as much as 10^{12} W is generated as the solar wind blows across the interconnected field lines near the comet-shaped boundary. A part of the electric current (carried mainly by electrons) thus generated flows between the magnetospheric boundary and an annular, ring-shaped region of the polar upper atmosphere along the lines of force of the Earth's magnetic field. *See* MAGNETOSPHERE.

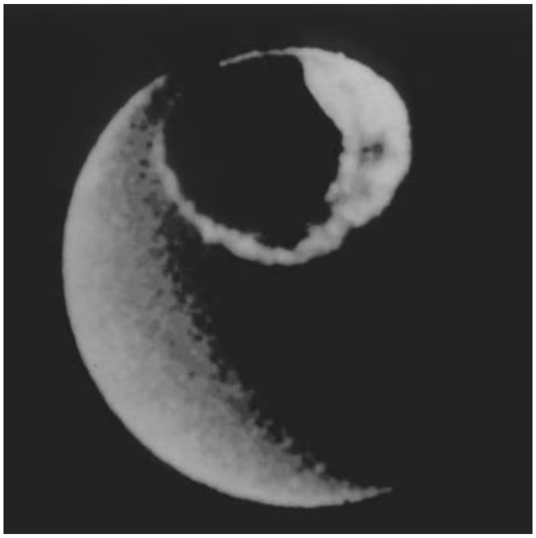


Fig. 1. Ring-shaped auroral glow observed from above the north polar region by the Dynamics *Explorer* satellite. (*L. A. Frank and J. D. Craven*, *University of Iowa*)

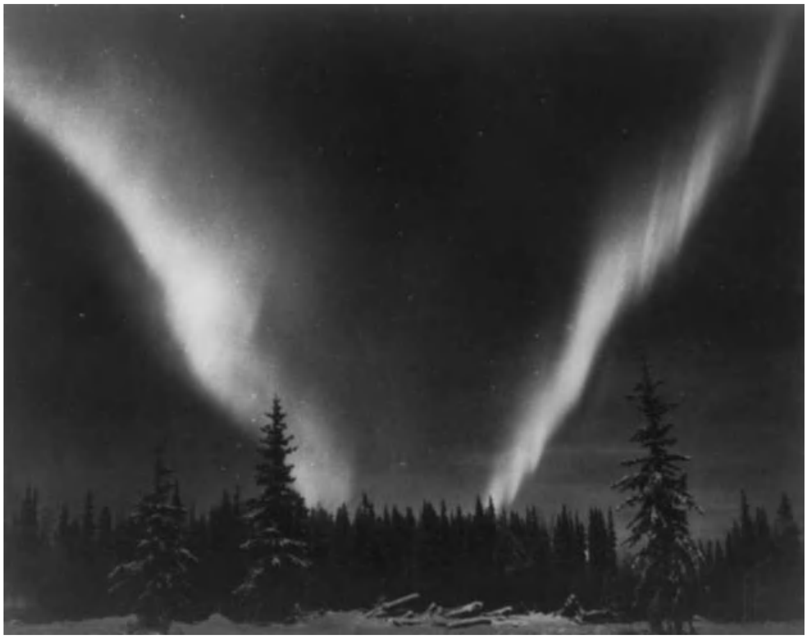


Fig. 2. Two curtain-shaped auroras stretching across the sky near Fairbanks, Alaska. (*Courtesy of Lee Snyder*)

As these electrons descend toward the Earth, they themselves develop an electrical potential drop of the order of a few kilovolts along the lines of force. As a result, the current-carrying electrons acquire energies of as much as a few kiloelectronvolts, sufficient to ionize and excite a few hundred atoms and molecules before they are stopped by the atmosphere at an altitude of about 60 mi (100 km).

Two ring-shaped glows, one in each hemisphere, are produced by upper atmospheric atoms and molecules which emit their own characteristic light after colliding with the current-carrying electrons. The most common light of the aurora (the greenish-white light) comes from excited oxygen atoms. Excited and ionized molecular nitrogen adds several band emissions. Imaging devices aboard satellites have successfully "photographed" both the northern and southern auroral rings (**Fig. 1**). Satellite-borne and rocket-borne detectors and radio probing devices, as well as advanced optical instrumentation, have increased considerably the understanding of the nature of the discharge electrons and the interaction between the upper atmosphere and auroral electrons. *See* SCIENTIFIC AND APPLICATIONS SATELLITES.

From a point on the ground, for example, Fairbanks, Alaska, only a small part of the ring-shaped glow can be observed. It is seen as a curtain-shaped glow, stretching from the eastern to the western horizon across the sky (**Fig. 2**). The bottom of the auroral curtain is sharply bounded and is located at about 60 mi (100 km) altitude. The upper boundary diffuses and extends to well above 180 mi (300 km). The auroral curtains undergo drastic changes from time to time, such as a sudden increase or rapid fluctuations in brightness, rapid poleward motions, or curling motions of various scales. When such activity occurs all along the ring-shaped region, it is called an auroral substorm. It lasts for a few hours and occurs a few times a day. The auroral discharge current during an auroral substorm can reach an intensity of several million amperes and generates intense magnetic disturbances, which cause polar magnetic substorms.

Magnetospheric physicists have determined the solar wind quantities which control the efficiency of the solar wind–magnetosphere generator and thus of auroral activity. They are the solar wind speed, and the intensity and the orientation (with respect to that of the Earth's magnetic field) of the solar wind magnetic field. Therefore, it has become possible to forecast auroral activity about one hour in advance by monitoring the solar wind at a distance of 200 earth radii upstream of the solar wind. The aurora becomes active during geomagnetic storms which occur often about 40 h after an intense solar flare. This is because the efficiency of the solar wind–magnetosphere generator becomes high and variable when the solar wind becomes gusty after a solar flare. During a great magnetic storm, the auroral ring expands from its usual latitude of about 67 to 50° or a little less. It is on such an occasion when the aurora can be seen widely across the continental United States. *See* ATMOSPHERE; ATOMIC STRUCTURE AND SPECTRA; GEOMAGNETISM; IONOSPHERE; PLASMA (PHYSICS); SOLAR WIND.

S.-I. Akasofu

Australia

An island continent situated in the Southern Hemisphere and extending from 10 to 44°S and from 113 to 153°E. Australia's total area (2,941,526 mi^2 or 7,618,517 km^2 for the Australian mainland and 26,383 mi^2 or 68,332 km^2 for the island of Tasmania) is somewhat less than that of the United States. Bounded on the west by the Indian Ocean and on the east by the Pacific Ocean, Australia straddles the Tropic of Capricorn. Originally part of the ancient supercontinent of Gondwanaland, which began to disintegrate 200 million years ago, Australia broke away from what is now Antarctica about 70 million years ago and began its long drift northward to its present position. Australia's long-term isolation has contributed to the uniqueness of its flora and fauna. Australia possesses about 650 species of birds and 400 species of reptiles, a large proportion of which are endemic. Most interesting are its 255 species of mammals, in particular the marsupials (lacking placenta), which include the kangaroo, wallaby, and koala; and the monotremes (egg-laying), which include the platypus and spiny anteater. These animals appear to represent stages along the evolutionary path toward fully developed placental mammals. *See* CONTINENTAL DRIFT; TROPIC OF CAPRICORN.

Australia's global position astride the southern tropic provides a climate that for much of the continent is characterized by clear skies, high temperatures, and relatively low rainfall (50% of its area has a median annual rainfall of less than 12 in. or 300 mm). The major cause of Australian aridity is the global subsidence of air at a latitude of about 30°S, warming as it sinks to produce belts of high pressure, clear skies, and low rainfall. Equatorward of the high-pressure belt, easterly trade winds blow back toward the Equator, while poleward generally westerly winds blow toward the higher-latitude low-pressure systems. While aridity dominates central Australia, coastal climates range from the seasonally wet monsoon tropics in the north to relatively moist temperate climates in the south (**Fig. 1**). *See* MONSOON METEOROLOGY; TROPICAL METEOROLOGY.

Australia is generally of remarkably low elevation and moderate relief. Seventy-five percent of the landmass lies between 600 and 1500 ft (180 and 450 m) in the form of a high plateau. A cross section from east to west shows first a narrow belt of coastal plain, then the steep escarpments of the eastern face of the Great Dividing Range, stretching 1200 mi (1900 km) from the north of Queensland to the south of Victoria (**Fig. 2***a*). The descent on the western slope of the Dividing Range is gradual until elevation in the inland basins is often below sea level. The land surface then rises gradually again across the great plateau until the low ranges of Western Australia fringing the plateau are reached, and beyond these lies another coastal plain. With the exception of the Gulf of Carpentaria and Cape York peninsula in the north and the Great Australian Bight in the south, there are few striking features in the configuration of the coast (Fig. 2*a*). Australia may conveniently be divided into three great physiographic regions, revealing characteristic geologic structure and major landform regions (Fig. 2*b*).

East Australian Highlands. A narrow plain extends north and south along the eastern coast, seldom more than 60–70 mi (100–110 km) in width, and occasionally only a few miles wide, but broader in the north than in the south. Flanking the plain are the series of ranges and tablelands making up the Great Dividing Range. In the south, one branch sweeps westward through Victoria, whereas the main branch continues due south, interrupted by the waters of Bass Strait, and terminates in Tasmania. Elevations are low, and the highest peak, Mount Kosciusko, is 7328 ft (2234 m). A few other peaks rise above 500 ft (1500 m) [**Table 1**]. This is a region of ancient mountains and of old and hard igneous and metamorphic rocks (dating mainly from the Paleozoic Era, 235–570 million years ago) that were raised in relatively recent geological times as a series of plateaus and low mountain blocks. The uplift introduced the rich deposits of gold, silver, lead, and copper that were a major stimulus to the early development of eastern Australia. The soils of the region are primarily shallow, except for much of Cape York and the area around Sydney. Active dissection by numerous short rivers has produced a broken surface of deep valleys and gorges on the eastern slopes, but few of these give access to the interior, for most terminate abruptly in rocky cliffs or turn and run parallel to the coast. Others run back to high plateaus or steep-sided, hazy-blue ranges. *See* PALEOZOIC.

TABLE 1. Australian mountains and summit elevations in feet (meters)

Australian Mainland	
Kosciusko (New South Wales)	7328 (2234)
Townsend (New South Wales)	7266 (2215)
Bogong (Victoria)	6508 (1984)
Feathertop (Victoria)	6306 (1922)
Bartle Frere (Queensland)	5438 (1658)
Tasmania	
Legge's Peak	5160 (1573)
Cradle Mountain	5069 (1545)

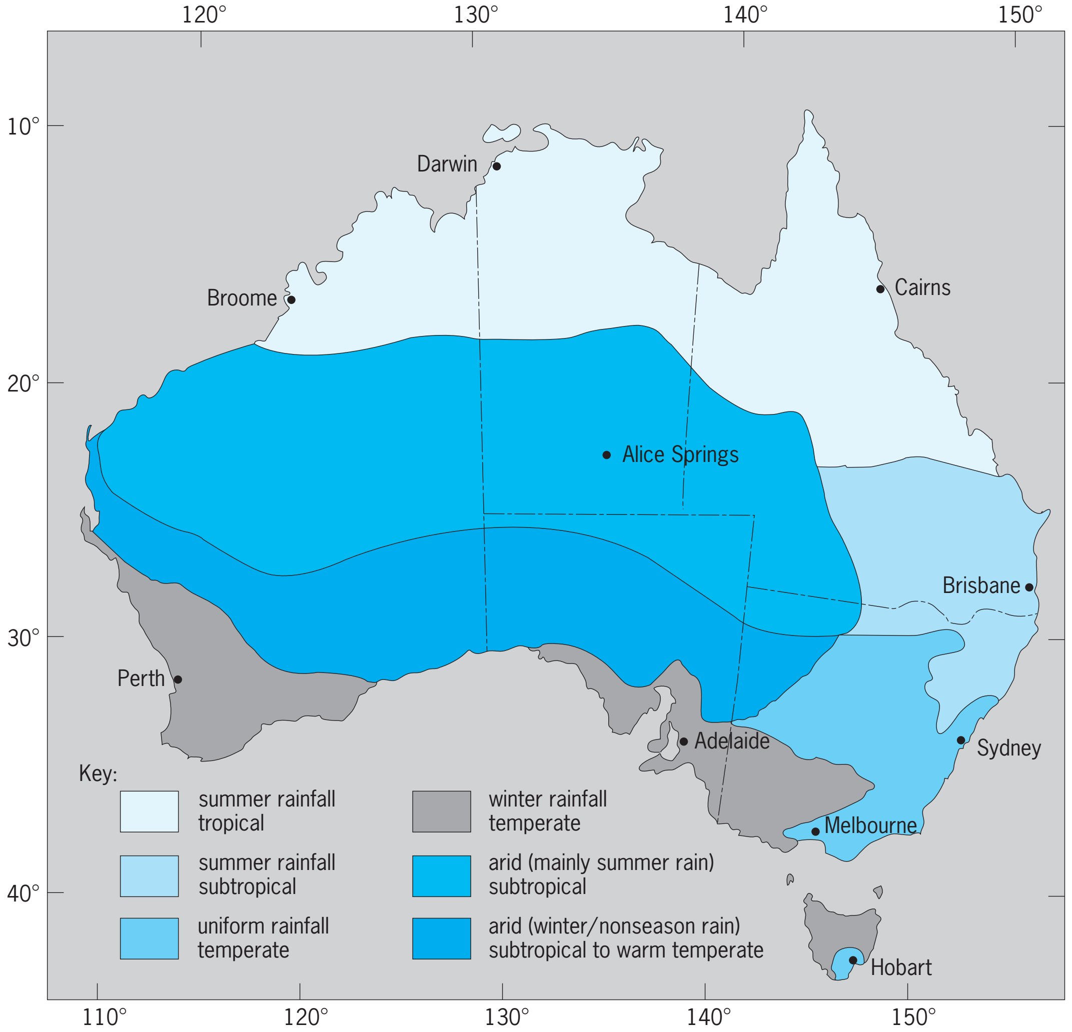

120°
130°
140°
150°
10°
20°
30°
40°
Darwin
Broome
Cairns
Alice Springs
Brisbane
Perth
Adelaide
Sydney
Melbourne
Hobart
Key:
summer rainfall tropical
summer rainfall subtropical
uniform rainfall temperate
winter rainfall temperate
arid (mainly summer rain) subtropical
arid (winter/nonseason rain) subtropical to warm temperate
110°
120°
130°
140°
150°

Fig. 1. Climate zones of Australia.

Located off the northeast coast and covering an area of 100,000 mi^2 (260,000 km^2) is the Great Barrier Reef, the largest living structure in the world. Extending 1300 mi (2000 km) from near the coast of Papua New Guinea to the Tropic of Capricorn, the reef actually comprises a chain of 2500 individual reef structures, many accumulating coral sand and debris to form small islands. Ribbon, or linear, reefs line the outer perimeter of the Great Barrier Reef at the continental shelf, at a distance ranging from 10 to 150 mi (16 to 240 km) offshore. There is concern about degradation of the central two-thirds of the reef as a result of the impact of tourism, agricultural runoff to the inshore waters, and infestations of the crown-of-thorns starfish which feed on growing coral. Partly in response to concerns about the health of the reef, it was declared a Marine Park in 1975 and was listed a UNESCO World Heritage Area in 1981. *See* REEF.

Climatically the region ranges from a tropical, primarily summer rainfall regime in the north (sometimes affected by the northwest monsoon) to a warm temperate, mainly uniform rainfall regime in the south (Fig. 1). This is the best-watered region in Australia, with annual rainfalls measuring generally above 25 in. (600 mm) and exceeding 250 in. (6000 mm) in some of the mountainous areas of the northeast. The generally sufficient rainfall is reflected in the extensive areas of gray-green *Eucalyptus* open forest and woodland (**Fig. 3**). Small pockets of tropical and temperate rainforest (closed forest) of more diverse floristic type occur along the east coast and in western Tasmania. Some of the river systems are of considerable size (**Table 2**). Large hydroelectric power plants have been developed in the East Australian Highlands and Tasmania. The elaborate Snowy River project, with a generating capacity of 3000 MW of electricity, also diverts water in tunnels under the mountains to the Murrumbidgee and Murray rivers, thus augmenting the supply of irrigation water in the drier interior of New South Wales and Victoria.

On the flanks of the East Australian Highlands are Australia's principal coal deposits, which are a major source of the nation's energy. Extensive black coal deposits were laid down in the Sydney and Bowen basins during the Permian Period (300–235 million years ago); and massive brown coal deposits of

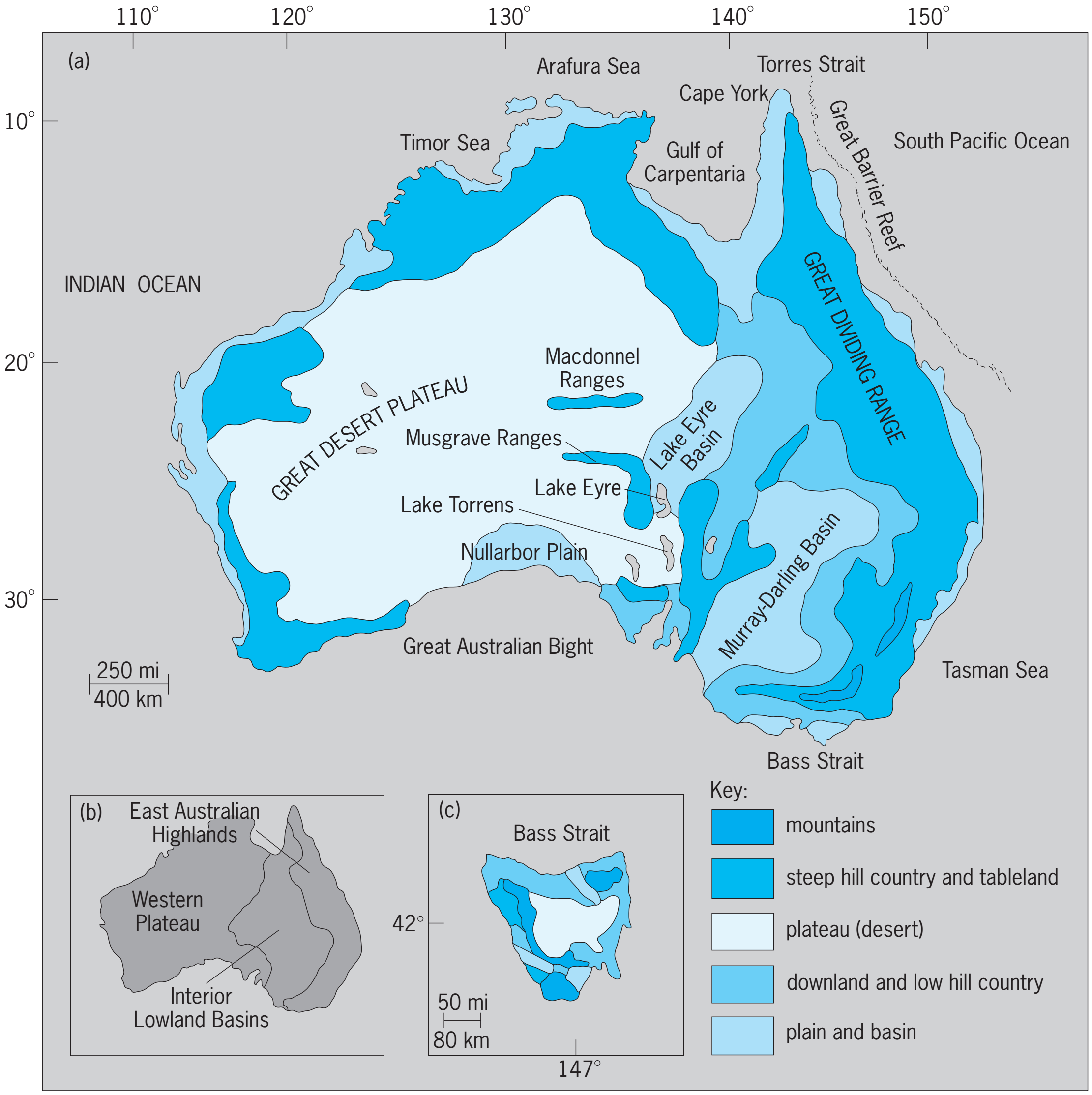

110°
120°
130°
140°
150°
(a)
10°
20°
30°
Arafura Sea
Torres Strait
Cape York
Timor Sea
Gulf of Carpentaria
Great Barrier Reef
South Pacific Ocean
INDIAN OCEAN
GREAT DIVIDING RANGE
Macdonnel Ranges
GREAT DESERT PLATEAU
Musgrave Ranges
Lake Eyre Basin
Lake Eyre
Lake Torrens
Nullarbor Plain
Murray-Darling Basin
Great Australian Bight
Tasman Sea
250 mi
400 km
Bass Strait
(b)
East Australian Highlands
Western Plateau
Interior Lowland Basins
(c)
Bass Strait
42°
50 mi
80 km
147°
Key:
mountains
steep hill country and tableland
plateau (desert)
downland and low hill country
plain and basin

Fig. 2. Land surface characteristics of mainland Australia and Tasmania. (*a*) Landforms of Australia. (*b*) Sketch map of predominant geologic structure and major landform regions of Australia. (c) Landforms of Tasmania.

Tertiary age (65–2 million years ago) lie in a great depression in the LaTrobe Valley of Victoria, south of the Great Divide and close to the coast of Victoria. Many important petroliferous basins lie adjacent to the major coal fields. The fields at Surat (Roma), flanking the divide in Queensland, and off the coast of Victoria in Bass Strait are among Australia's most productive deposits of petroleum and natural gas. *See* PERMIAN; PETROLEUM GEOLOGY; TERTIARY.

Tasmania. This small, relatively mountainous island lies 150 mi (240 km) southeast of the Australian mainland across Bass Strait. It is structurally similar to the East Australian Highlands and is a logical extension of that region. The dominant feature is the central plateau, falling from a general level of 3500 ft (1100 m) in the northwest toward the southeast. Its lake-studded surface is drained by the Derwent River (Table 2). To the west and south the plateau rises to 4000–5000 ft (1200–1500 m) in a range of rocky mountains, but to the north and east it is much lower

TABLE 2. Principal Australian rivers and lengths in miles (kilometers)

Australian Mainland	
Darling (Queensland and New South Wales)	1760 (2830)
Murray (Victoria, South Australia, and New South Wales)	1600 (2600)
Murrumbidgee (New South Wales)	1050 (1690)
Lachlan (New South Wales)	850 (1400)
Flinders (Queensland)	520 (840)
Diamantina (Queensland)	468 (753)
Murchison (Western Australia)	440 (710)
Burdekin (Queensland)	440 (710)
Bulloo (Queensland)	370 (600)
Mitchell (Queensland)	350 (560)
Tasmania	
South Esk	120 (190)
North Esk	45 (72)
Tamar (from confluence of North and South Esk)	40 (60)
Huon	105 (169)
Derwent	107 (172)
Gordon	90 (150)

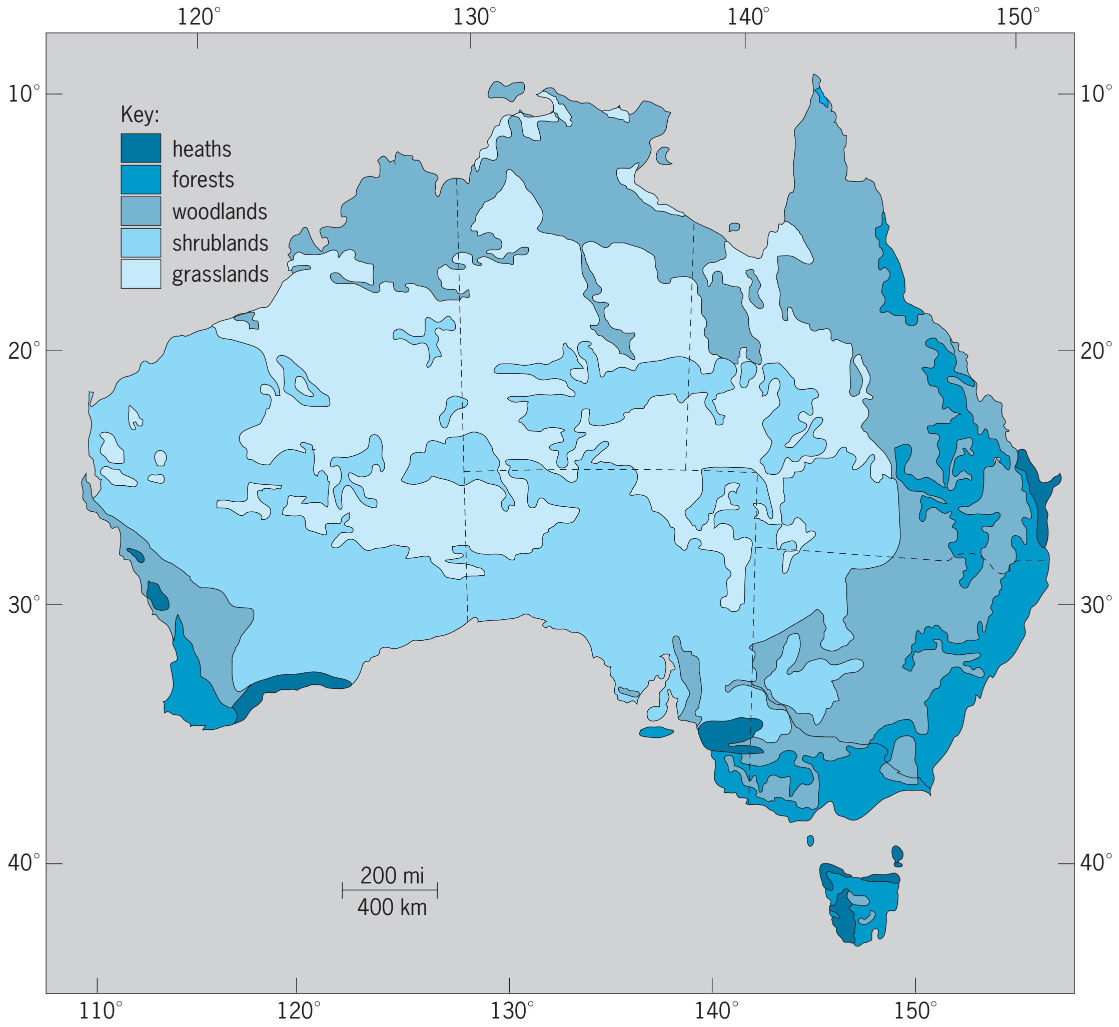


Fig. 3. **Basic vegetation distributions in Australia. (*After R. M. Moore ed., Australian Grasslands, Australian National University Press, 1970*)**

and terminates in bold scarps about 1000 ft (300 m) high which overlook the Tamar-Esk river lowlands. On the northeastern side of this valley lowland, an isolated block of rocky highlands includes the highest peak (5160 ft or 1570 m) in Tasmania (Table 1). In the southwest, the Gordon and Huon river valleys separate the mountains fringing the plateau from another outlying series of low mountains rising about 3500 ft (1100 m). The Tasmanian climate is temperate, with dominantly winter rainfall. A dense *Eucalyptus* forest and woodland covers most of the island except along the wetter west coast (rainfall exceeds 120 in. or 3000 mm in places), where dense beech forest predominates. Some coastal regions have extensive areas of heath vegetation. The rivers have short, rapid courses with relatively little seasonal variation in flow. Significant and exploitable mineral deposits of copper, lead, zinc, and tin occur, especially in northwest Tasmania.

Interior Lowland Basins. The inland comprises a region of mainly Mesozoic and Cenozoic (235 million years ago to the present) sedimentary rocks, surficial deposits, and deeply weathered mantle, occupying one-third of the continent between the western slope of the East Australian Highlands and the inner eastern margin of the ancient shield which forms the Western Plateau (Fig. 2*a* and *b*). Little land is over 500 ft (150 m), and some (in the region of Lake Eyre) is below sea level. *See* BASIN; CENOZOIC; MESOZOIC; SEDIMENTARY ROCKS.

Low rises separate three extensive drainage systems: One drains north to the Gulf of Carpentaria; the Lake Eyre Basin is one of inland drainage; and the third, the Murray-Darling Basin, is drained to the south. Except where the low Flinders, Barrier, and Mount Lofty ranges sharply break the surface in the south, the lowland is a dry and monotonous sandy-gravelly plain, crossed more often by broad, incised, dry river beds with occasional oxbow lakes (billabongs) than by rivers of running water. The rivers of the Murray-Darling Basin, draining the western slopes of the Great Dividing Range, have a marked seasonal variation in flow but never dry up in the lower reaches. South Australia's shallow lakes are more often dry expanses of encrusted white salt than bodies of water—the result of low rainfall and high

TABLE 3. Principal Australian lakes and areas in square miles (square kilometers)

Australian Mainland	
Eyre (South Australia)*	2970 (7690)
Torrens (South Australia)*	2230 (5780)
Mackay (Northern Territory and Western Australia)*	1990 (5100)
Gairdner (South Australia)*	1840 (2960)
Frome (South Australia)*	930 (2400)
Argyle (Western Australia)	430 (1100)
Corangamite (Victoria)	90 (230)
Tyrrell (Victoria)*	67 (170)
George (New South Wales)	60 (160)
Menindee (New South Wales)	60 (160)
Nerran (New South Wales)	55 (140)
Tasmania	
Great Lake	44 (110)
Sorell	19 (49)
St. Clair	15 (39)
Arthur	14 (36)
Echo	12 (31)

*Shallow salt pans, normally or seasonally dry.

evaporation (**Table 3**). In most parts of the region, water from deep artesian wells is available. *See* ARTESIAN SYSTEMS.

The climate of much of the region is arid, with annual rainfalls as low as 4 in. (100 mm) in the area around Lake Eyre. However, the coastal margins of the Gulf of Carpentaria exhibit a summer monsoonal climate, and in the far south of the region a temperate, moderate winter rainfall regime is present (Fig. 1).

Despite generally low rainfall, very little of the region is completely devoid of vegetation. In the south of the area, pockets of mallee heathland (dominated by low-growing *Eucalyptus*) predominate in the sand country of northwest Victoria and southeast South Australia. Much of the remainder of the Interior Lowland is covered in *Acacia* shrubland and grassland (Fig. 3), with patches of *Eucalyptus* savanna-woodland appearing in the north Gulf region and along the west coast of Cape York. *See* SAVANNA.

Minerals of importance in this region lie mainly along its margins, with significant deposits of gold, copper, lead, and zinc at the southeastern edge of the Barkly Tableland in the area of Mount Isa, and in the southern Barrier Range around Broken Hill. Extensive bauxite deposits occur on the west coast of Cape York, from where the raw material is transported to alumina refineries and aluminum smelters in Australia and overseas. Very significant natural gas fields dating from the Permian (300–235 million years ago) occur in the Cooper Basin in the Lake Eyre region. *See* BAUXITE; NATURAL GAS.

Western Plateau. This largest area occupies almost three-fifths of the continent; it is a great shield of ancient rocks, many of which date from the Archean Era (3600–2500 million years ago), standing 750–1500 ft (230–460 m) high. Sedimentary and low-grade metamorphic rocks dominate, but there are significant outcrops of high-grade metamorphic and igneous rocks. Some of the earliest evidence of life on Earth, microbe-sized fossils dating back 3465 million years, comes from the northwest of the Western Plateau. Much of the plateau is buried in deeply weathered desert sand, and only a few ridges of ancient mountains (such as the Macdonnel and Musgrave ranges) break the monotony of the surface. The sand is arranged in ridges as high as 50–60 ft (15–20 m), running remarkably parallel for great distances northwest to southeast. The ridges are more permanent features of the arid landscape than are the migratory dunes of other deserts. Local ephemeral streams form a net of inland drainage to scattered salt-lake basins. The limestone Nullarbor Plain, skirting the Great Australian Bight, is both treeless and riverless. Only in the southwestern corner of the continent (a temperate Mediterranean-like climate) and along the northwestern coast (a tropical monsoon climate) is rainfall sufficient to support a sclerophyll forest of *Eucalyptus* and a monsoon *Eucalyptus* woodland, respectively, each giving way inland to a succession of savanna, desert *Acacia* scrub, and dry grassland. Significant areas of heath occur in coastal regions of the southwest. In the north, coastal rivers are of considerable size, but they change from flooded torrents after seasonal summer rains to a succession of water holes in the winter dry season. In the same area, Lake Argyle, an engineered lake formed on the Ord River as part of an ambitious irrigation scheme, is Australia's largest fresh-water lake (Table 3). *See* ARCHEAN.

It is in this western region that many of the great postwar discoveries of minerals have been made that have helped transform the Australian economy. Chief among these is one of the world's greatest series of deposits of high-grade iron ore. To exploit them, settlements and networks of communications have been established in the hitherto unoccupied arid northwest of Western Australia, in the vicinity of the Hamersly and Opthalmia ranges; and ports have been constructed to facilitate the export of the ore to Japan and to transport ore to processing plants elsewhere in Australia. Significant gold deposits are located throughout the southern Western Plateau, and major oil and gas fields are located off the northwest coast in the Carnarvon Basin. Smaller gas and oil fields have been located west of Darwin, and in the Amadeus Basin in the south of the Northern Territory. Since the development of the Argyle deposit in the Kimberley region of the northern Western Plateau, Australia has become the world's largest diamond producer. *See* DIAMOND; ORE AND MINERAL DEPOSITS.

Nigel J. Tapper

Bibliography. Australian Surveying and Land Information Group, *Atlas of Australian Resources*, vol. 5: *Geology and Minerals*, 1988, vol. 6: *Vegetation*, 1990; D. N. Jeans (ed.), *Australia: A Geography*, vol. 1: *The Natural Environment*, 2d ed., 1986; D. Johnson, *The Geology of Australia*, 2005; State of the Environment Advisory Council, *Australia State of the Environment*, 1996; A. P. Sturman and N. J. Tapper, *Weather and Climate of Australia and New Zealand*, 1996.

Australopithecine

Any of the seven species that belong to the family Hominidae (comprising humans and their closest relatives). These species are not attributable to the genus *Homo*, but belong to at least three genera that existed between about 4.4 million years ago (Ma) and 1.2 Ma during the Pliocene and early Pleistocene epochs. All seven species are known only from Africa. Although some workers regard all of them as belonging to one genus, *Australopithecus*, it is clear that a second genus, *Ardipithecus*, should be recognized for the earliest known hominid fossils, and that the three "robust" species belong to a third genus, *Paranthropus*.

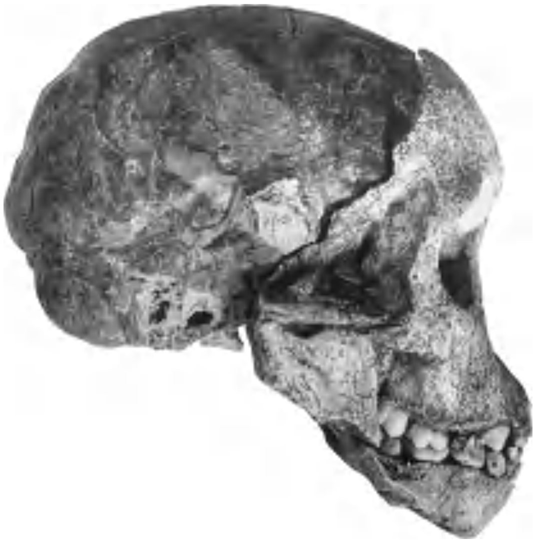


Fig. 1. Lateral view of the Taung skull of *Australopithecus africanus*, the type specimen of *Australopithecus* and the first early hominid specimen to be discovered in Africa. (*Courtesy of F. E. Grine*)

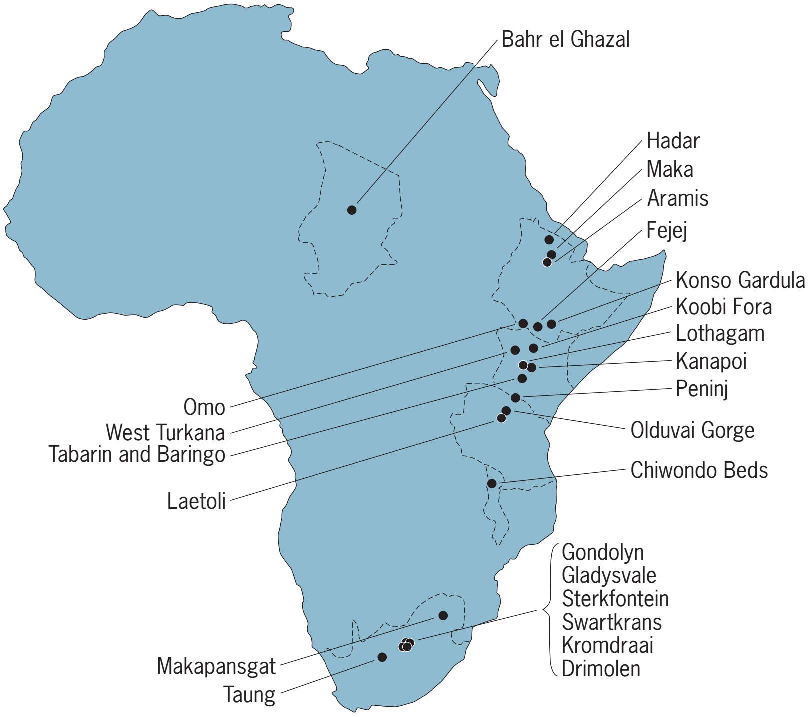


Fig. 2. Principal sites at which fossils of *Australopithecus* have been discovered in Chad, Ethiopia, Kenya, Tanzania, and South Africa.

Historical context. The name australopithecine comes from the taxon *Australopithecus* ("southern ape") *africanus*, which was coined by Raymond Dart for a fossil skull discovered in 1924 in Taung, South Africa (**Fig. 1**). The Taung skull had several distinctly hominid, or humanlike, features, but the claim that *A. africanus* was a human forebear was disputed by many of the leading paleoanthropologists of that time. The hominid status of *A. africanus* became widely accepted more than a decade later, largely because of work on *Australopithecus* fossils from Sterkfontein. Specimens of *A. africanus* are known also from the sites of Makapansgat and Gladysvale in southern Africa (**Fig. 2**). Faunal comparison with radiometrically dated sites in eastern Africa indicates that this species existed between about 3.0 and 2.3 Ma.

The name *Paranthropus* ("beside human") was coined in 1938 when fossils from Kromdraai, South Africa, were attributed to the taxon *P. robustus*. Fossils of this species are also known from the nearby sites of Swartkrans and Drimolen. These bones are dated, also by faunal comparisons with radiometrically dated sites in eastern Africa, to between about 1.8 and 1.5 Ma.

In 1959, a cranium with very large cheek teeth was discovered at Olduvai Gorge, Tanzania (Fig. 2). It represents a separate species of *Paranthropus*, *P. boisei*. Numerous *P. boisei* fossils have been discovered in Tanzania at Olduvai Gorge and Lake Natron, in southern Ethiopia along the Omo River (Shungura Formation), and especially in Kenya on the eastern side of Lake Turkana (Koobi Fora Formation). Fossils of *P. boisei* are dated to between 2.3 and 1.2 Ma.

In 1935, the fossil site of Laetoli was discovered south of Olduvai Gorge, but it was only during the 1970s that most of the hominid jaws and teeth were found there among volcanic ash layers that preserve spectacular hominid footprint trails. Between 1973 and 1977, numerous specimens, including the partial skeleton known as Lucy, were recovered from the Hadar Formation sediments in Ethiopia. It was proposed that the Laetoli and Hadar fossils belonged to a primitive species of *Australopithecus* named *A. afarensis. Australopithecus afarensis* fossils are known also from Maka in the Middle Awash of Ethiopia, and from Koobi Fora, Kenya. Fossils that possibly belong to *A. afarensis* (although their specific attribution is open to question) are known from the Ethiopian sites of Belohdelie and Fejej. *Australopithecus afarensis* is dated radiometrically to between 3.7 and 2.9 Ma; it may extend back to 4.0 Ma if the specimens from Belohdelie and Fejej belong to it.

In 1993, the anterior part of a mandible and an upper premolar fragment were discovered in the Bahr el Ghazal region of Chad (Fig. 2). The associated fauna includes several taxa found also in the Hadar Formation of Ethiopia, indicating an age of some 3.0

to 3.4 Ma. It has been suggested that the Chadian fossils represent a separate species, *A. bahrelghazali.* However, the Chad specimens do not afford a definitive diagnostic differentiation from *A. afarensis*; thus, as a conservative approach, the Bahr el Ghazal fossils are considered here as being tentatively attributable to *A. afarensis*.

In 1985, the discovery of a nearly complete skull in the Nachukui Formation on the western side of Lake Turkana, Kenya, led to the recognition of a third "robust" australopithecine species, *P. aethiopicus*. A handful of fossils attributable to this species are known also from southern Ethiopia (Shungura Formation). This species is known from about 2.7 to about 2.3 Ma.

Fossils found in the Aramis region of Ethiopia in 1992 and 1993 were recognized as belonging to a separate taxon, *Ardipithecus* ("ground ape") *ramidus*. It is the most primitive hominid species known. In some morphological features, it is more similar to apes than to other, later hominids. Most of the Aramis bones date to 4.4 Ma. Three fragmentary fossils from Kenya may be attributable to *Ar. ramidus*. One is from the site of Lothagam; it is likely between 5.0 and 5.5 million years old. The other two are from Tabarin and Baringo. Volcanic rocks beneath the Tabarin jaw are 5.1 million years old, and the fauna provides an upper age of 4.2 million years for the hominid. The fauna from Baringo also predates 4.2 Ma.

In 1995, the species *A. anamensis* was named to accommodate fossils that were found at the Kenyan sites of Kanapoi and Allia Bay. These specimens date to between about 3.9 and 4.2 Ma.

Species. Thus, at least seven australopithecine species, belonging to at least three genera, can be recognized in the Pliocene and early Pleistocene of Africa:

Ardipithecus ramidus	(?5.0–4.4 Ma)
Australopithecus anamensis	(3.9–4.2 Ma)
Australopithecus afarensis	(?4.0–2.9 Ma)
Australopithecus africanus	(3.0–2.3 Ma)
Paranthropus aethiopicus	(2.8–2.3 Ma)
Paranthropus boisei	(2.3–1.2 Ma)
Paranthropus robustus	(1.8–1.5 Ma)

All are defined on the basis of craniodental morphology.

Ardipithecus ramidus. This is the most primitive hominid species known. It has large canine teeth coupled with comparatively small molars. The anterior lower deciduous premolar is chimpanzeelike, and the anterior permanent premolar is asymmetrical, dominated by a single large cusp. It is unique among all hominids in having thin tooth enamel. The temporal bone has a very shallow glenoid fossa (jaw joint) without a distinct articular eminence; the external auditory meatus is small, the tympanic bone is tubular, and its lateral edge is positioned far laterally. The temporal bone is extensively pneumatized (filled with air cells). The upper limb bones display some features in common with modern humans, as well as characteristics found in living apes. The latter include a large styloid process at the end of the radius, a strong elbow-stabilizing mechanism of the radius and humerus, and strong muscle ridges. These features almost certainly indicate that *Ar. ramidus* was an accomplished climber.

The Aramis fauna is dominated by primates, and arboreal colobine monkeys are the most common of these. The mammalian remains and the presence of fossil seeds suggest that *Ar. ramidus* occupied a wooded habitat.

Australopithecus anamensis. This species has several primitive features in common with *Ardipithecus*, but it also exhibits some derived traits in common with taxa such as *A. afarensis*. It has a few unique characteristics. Thus, the anterior lower permanent premolar is virtually single-cusped and asymmetrical, and the canine crowns and especially the roots are very large (at least in presumptive males). The temporal bone is also extensively pneumatized, the glenoid fossa is shallow without a distinct articular eminence, the tympanic bone is tubular, and the external auditory meatus is small. The palate is shallow anteriorly. The postcanine tooth rows are nearly parallel to one another and close together. The mandibular symphysis, which has a marked postero-inferior slope, extends as far back as the first molar tooth. Its molar crowns have sloping sides, but its tooth enamel appears to be similar in thickness to that of other *Australopithecus* species.

The tibia has features that clearly indicate that this species was a biped. It is a large bone; its owner is estimated to have weighed 47–55 kg (103–120 lb). A wrist bone (the hamate) has a long hook, suggestive of large, powerful hand flexors running through the carpal tunnel. The radius and distal humerus have several apelike characters that are associated with vertical climbing. These bones suggest that *A. anamensis* was a terrestrial biped that was also heavily engaged in tree climbing. Current reconstructions of the paleoecology at Kanapoi and Allia Bay are consistent with this hypothesis. The fauna suggests a woodland-bushland habitat along a large river (at Kanapoi), and a gallery forest associated with a large river that formed a local delta into a large lake (Lake Lonyumun).

Australopithecus afarensis. This species is characterized primarily by a suite of primitive craniodental features, including a strongly protruding jaw (prognathism), a flat (unflexed) cranial base, and a relatively flat glenoid fossa without a distinct articular eminence. It has a tubular tympanic bone, although the external auditory meatus is large. The parietal mastoid angle is strongly flared, and there is close approximation of the cranial points lambda and inion at the back of the skull. The palate is anteriorly shallow, the mediolaterally and supero-inferiorly convex nasoalveolar clivus is demarcated from the floor of the nose by a horizontal sill, and the upper lateral incisor roots are situated to the side of the lateral walls of the nasal aperture. The canines are relatively large, and the anterior lower premolar tends to be asymmetrical and dominated by a single large cusp. The teeth are relatively large in comparison to body size. Endocranial capacity ranges about 310–500 cm^3 (19–30 in.^{3}). In addition, *A. afarensis* has a high incidence of an occipital-marginal sinus by which blood is drained from the brain.

There is a high degree of (presumed sexual) dimorphism in body size. Estimates range 30.5–80.5 kg (67–177 lb), which is similar to body size ranges for gorillas and orangutans.

In 1978–1979, volcanic-ash layers bearing hominid footprint trails were discovered at Laetoli. These trails and the bones of the postcranial skeleton indicate that *A. afarensis* was a terrestrial biped, but its gait was not exactly the same as that of modern humans. It was not able to employ a full striding gait. Furthermore, a host of postcranial features, such as strongly curved finger bones, a superiorly oriented shoulder joint, a relatively long upper limb, and a relatively long forearm, indicate that *A. afarensis* was well adapted to arboreal climbing. This species almost certainly spent a considerable amount of time in an arboreal milieu, perhaps feeding and sleeping. The knee joints from Hadar suggest that the smaller (presumptive female) and larger (presumptive male) individuals varied in their reliance on arboreality, such as shown by orangutans.

Paleoenvironmental reconstructions indicate diverse habitats, from well-watered and wooded, to forested environments with edaphic grasslands (at Hadar), to more woodland-bushland conditions (at Laetoli). *Australopithecus afarensis* probably had a fairly broad range of locomotor abilities and dietary proclivities.

Australopithecus africanus. This is the type species of the genus *Australopithecus*. Its cranium is more globular and less pneumatized than that of *A. afarensis*. In comparison to *A. afarensis*, this species has a more steeply inclined forehead, a greater separation of the cranial points lambda and inion at the back of the skull, and a high glabella that is separated from the nasion. The nasoalveolar clivus is prognathic but flattened, the palate is anteriorly deep (shelved), and the maxillary lateral incisor roots are situated medial to the lateral margins of the nasal aperture. The glenoid fossa is deep, and there is a distinct articular eminence. The tympanic plate is vertically disposed, and the external auditory meatus is large. The lower anterior premolar is bicuspid. Endocranial capacity estimates for eight specimens range 428–515 cm^3 (26–31 in.^{3}), although the latter value is likely underestimated. In contrast to *A. afarensis*, *A. africanus* has a high incidence of the transverse sinus route for blood drainage from the brain.

Dental and skeletal dimensions indicate that sexual dimorphism in *A. africanus* was also of considerable magnitude. Body weight estimates range 33.5–67.5 kg (67–149 lb) for presumptive females and males. Postcranial elements are generally similar to those of *A. afarensis*, suggesting that *A. africanus* retained adaptations for tree climbing along with the adaptations for bipedal terrestrial locomotion. Some features, such as foot bone structure indicating a greater range of motion of the big toe, limb proportions indicating a relatively long upper limb, and the structure of the proximal end of the tibia, suggest that the skeleton of *A. africanus* may be somewhat more primitive (chimpanzeelike) than that of *A. afarensis*.

Paleoenvironmental reconstructions suggest that *A. africanus* inhabited well-watered environments with notable bush and tree cover. The microscopic details of tooth wear indicate that the diet of *A. africanus* consisted principally of fruits and leaves.

Paranthropus robustus. The three species in the genus *Paranthropus* are characterized by cranial and dental features indicative of powerful chewing. Thus, the *Paranthropus* face is flattened and somewhat "dished," with the cheeks located anterior to the margins of the nasal aperture. The facial skeleton is hafted high onto the neurocranium, which results in a low forehead with a concave frontal trigone. The prominent glabella is situated below the level of the supraorbital margin and is nearly coincident with the nasion. The nasoalveolar clivus is flattened and grades imperceptibly into the floor of the nose. The palate is extremely thick. The cheek bones are very robust, and males possess a sagittal crest from which the large temporalis muscles originate. In addition, the petrous portion of the temporal bone is coronally inclined, and the bulbous mastoid region of the temporal bone is laterally inflated. The premolars are fully bicuspid, and they and the molars are greatly enlarged. Tooth enamel is exceptionally thick.

Paranthropus robustus is the type species of the genus *Paranthropus*. It is further characterized by the presence of a triangular depression on the front of the cheek bone (maxillary trigone), a palate that is anteriorly shallow, and a deep glenoid cavity with a distinct articular eminence. The tympanic bone is a vertically deep plate, and the external auditory meatus is very large. An endocranial capacity estimate of some 530 cm^3 (32 in.^{3}) has been recorded for the single reliable specimen. The upper canines are in the same coronal plane as the incisors, and the anterior teeth not only are absolutely smaller than those of *Ardipithecus* and *Australopithecus* but are especially diminutive in relation to the sizes of the premolars and molars.

Analysis of the postcranial skeleton of *P. robustus* is complicated by the presence of specimens of the genus *Homo* in the same deposits at Swartkrans. Nevertheless, it appears that *Paranthropus robustus* bones share some features with those of *A. afarensis* and *Australopithecus africanus*. These features include a relatively small femoral head with a long and anteroposteriorly flattened femoral neck. The morphology of the proximal radius is like that seen in earlier australopithecine species and living great apes, suggestive of arboreal capabilities. Foot bones indicate bipedal terrestrial locomotion, with evidence of a more humanlike foot than is present in either *A. afarensis* or *A. africanus*. It has also been suggested that the hand of *P. robustus* may have been more similar to that of modern humans, indicating that *P. robustus* had the manual capability to manufacture stone tools. Body size estimates for *P. robustus* range 42.5–65.5 kg (93.5–144 lb); these values suggest that *P. robustus* may not have been larger, on average, than either *A. afarensis* or *A. africanus*. Size (presumed sexual) dimorphism in cranial, dental, and postcranial elements of *P. robustus* appears to be somewhat less prominent than in the other two species of *Paranthropus* (*P. boisei* and *P. aethiopicus*) or in species of *Australopithecus* (such as *A. afarensis* and *A. africanus*). However, it is unclear whether this apparent pattern reflects a specific reduction of body size dimorphism in *P. robustus*, or whether it is a taphonomic artefact that reflects the preferred prey size of the predators (such as leopards) that were primarily responsible for the accumulation of *P. robustus* remains.

Paleoenvironmental reconstructions suggest an open habitat with a nearby river that probably supported a woodland or gallery forest, and also more open, wooded grassland or savanna conditions during the accumulation of *P. robustus* remains. Tooth wear suggests that the diet of *P. robustus* was composed of harder items than were chewed by *A. africanus*. Studies of the carbon isotope signatures of *P. robustus* teeth indicate that this species had an overall reliance on C3-based foods (that is, trees, shrubs, forbs, and tubers). Overall, the cranial and dental anatomy of *P. robustus* indicates a primary adaptation to the mastication of foods that required the application of powerful chewing forces.

Paranthropus boisei. Most of the morphological features that characterize this species are also possessed by *P. robustus*, although the dental and bony characteristics that are associated with mastication appear to be even more exaggerated in *P. boisei*. Thus, *P. boisei* differs from *P. robustus* in having a deeper maxilla, an anteriorly deep (shelved) palate, the absence of a maxillary trigone, zygomatics that tend to be laterally bowed with a visorlike configuration of the cheek, and a heart-shaped foramen magnum with a straight or posteriorly convex anterior margin. The cheek teeth of *P. boisei* tend to be larger than those of *P. robustus*; indeed, *P. boisei* premolars and molars are the largest of any hominid taxon. The canines and incisors, however, are diminutive. Endocranial capacity estimates range about 500–530 cm^3 (30–32 in.^{3}) with an average of approximately 515 cm^3 (31 in.^{3}). Thus, as in *P. robustus*, the brains of *P. boisei* specimens tend to be somewhat larger than those of most *Australopithecus* specimens. *Paranthropus boisei* has a high incidence of an occipital-marginal sinus by which blood is drained from the brain.

Although cranial and dental remains of *P. boisei* are quite abundant, there are comparatively few postcranial bones that can be attributed to this species with certainty. The bones that are reasonably attributed to *P. boisei* reveal a picture of a biped that retained adaptations to arboreal locomotion (for example, it had relatively long arms, such as displayed by *Australopithecus anamensis*, *A. afarensis*, and *A. africanus*, and an apelike proximal radius morphology that suggests a long lever arm of the biceps muscle as is found in *Australopithecus*). The ankle bones (talus and calcaneus) are generally humanlike, and the femur has a relatively small head coupled with a long and anteroposteriorly flattened neck.

Cranial and postcranial elements indicate considerable sexual dimorphism. Body weight estimates from more complete bones range 52–60 kg (114–132 lb), although other, less complete postcranial elements indicate a greater range, from less than 30 kg (66 lb) to greater than 100 kg (220 lb).

Paleoenvironmental reconstructions reveal a variety of potential habitats, including scrub woodlands with extensive wetlands, open woodlands with edaphic grasslands (in which the soil has been periodically inundated with water, for example, a bank overflow), riparian woodlands, and grassland or shrublands, with a possible preference for well-watered sites such as riverine gallery forests. Overall, it appears that *P. boisei* preferred fairly open habitats (woodland to scrub woodland) associated with water and edaphic grasslands. Wear on *P. boisei* cheek teeth suggests the mastication of some hard items, although they may have played less of a role than in the dietary regimen of *P. robustus*. The diet of *P. boisei* may have required the prolonged chewing of tough or fibrous vegetable foods, perhaps with relatively little nutritional value.

Paranthropus aethiopicus. This species is known from comparatively few fossils (one fairly complete cranium, two incomplete mandibles, and a handful of isolated teeth). Like *P. robustus* but unlike *P. boisei*, this species possesses a maxillary trigone and an anteriorly shallow palate. Like *P. boisei* specimens, *P. aethiopicus* has a heart-shaped foramen magnum with a straight anterior margin. The premolars and molars of *P. aethiopicus* are similar in size to those of *P. boisei*; thus, they tend to be larger than most *P. robustus* cheek teeth.

The cranium of *P. aethiopicus* differs from, and is more primitive than, those of *P. boisei* and *P. robustus*, in that it possesses a long, unflexed base, marked alveolar prognathism, a shallow glenoid fossa without a distinct articular eminence, and a strongly flared parietal mastoid angle. In addition, the anterior teeth of *P. aethiopicus* may have been larger than those of the other two *Paranthropus* species. The endocranial capacity of the single known (male) cranium is 410 cm^3 (25 in.^{3}), which is noticeably smaller than estimates for *P. boisei* and *P. robustus* crania. In addition, *P. aethiopicus* appears to have employed the transverse sinus rather than the occipital-marginal sinus to drain blood from the brain.

Cranial remains reveal considerable size (possibly sexual) dimorphism in *P. aethiopicus*. Postcranial bones that may belong to *P. aethiopicus* include a large ulna that is notable for its considerable length and substantial dorsoventral curvature, features also found in apes and associated with arboreal climbing. It seems safe to assume that the postcranium of *P. aethiopicus* will exhibit those characters that are possessed in common by other early hominid species, including considerable differentiation in size, features related to bipedal terrestrial locomotion, and other characteristics associated with arboreal climbing.

Evolutionary relationships. Paleoanthropologists disagree over the assignment of early hominid fossils to different genera and species. Such arguments over what is known as Alpha Taxonomy are to be expected, as different workers view the fossil record from different philosophical perspectives. Such differences also account for disagreements over the phylogenetic relationships of these species, including the issue of which (if any) is most closely related to the human genus, *Homo*. Every phylogenetic hypothesis that has been put forward since the 1950s has been either falsified outright or at least substantially altered by ongoing research and new discoveries.

At present, no scientifically rigorous phylogenetic analysis has been undertaken that incorporates all seven of the species discussed above. The most comprehensive study of "australopithecine" evolutionary relationships to date is that by David Strait and colleagues (1997). That study does not include *Ar. ramidus* or *A. anamensis*. Nevertheless, it is evident from the descriptive account of *Ar. ramidus* that it very likely resembles the stem hominid taxon in its morphology (**Fig. 3**). *Australopithecus anamensis* shares some evolved (derived) traits with later species such as *A. afarensis*, but retains some primitive features that are displayed also by *Ar. ramidus*. Thus, *A. anamensis* most likely evolved from a species that was at least morphologically similar to *Ar. ramidus* in some respects. For the moment, *Ar. ramidus* represents the best candidate for the ancestor of *A. anamensis*. *Australopithecus anamensis*, in turn, possesses some unique features that make it unlikely to be the immediate ancestor of *A. afarensis*, but it is probable that *A. afarensis* evolved from a species that had a strong morphological resemblance to *A. anamensis*.

Australopithecus afarensis likely gave rise to a lineage that provided the ancestry of both *A. africanus* and another lineage that included the common ancestor of the genera *Paranthropus* and *Homo*.

Although *A. africanus* shares a number of derived morphological characters with species that are part of the *Paranthropus* and *Homo* lineage, it is not considered to be directly ancestral to that line because it exhibits derived morphology in several characters (for example, glenoid fossa depth and anterior palatal depth) that are more primitive in both *A. afarensis* and *P. aethiopicus*. It is perhaps more likely that the derived traits which *A. africanus* shares with some species of *Paranthropus* and *Homo* were evolved in parallel. Other workers, however, have argued cogently that *A. africanus* constitutes a reasonable morphotype for the last common ancestor of the lineage that leads to *Homo* and *Paranthropus*. In this case, the primitive features displayed by *P. aethiopicus* would represent evolutionary reversals. At present, then, it is safest to conclude that the phyletic position of *A. africanus* remains ambiguous.

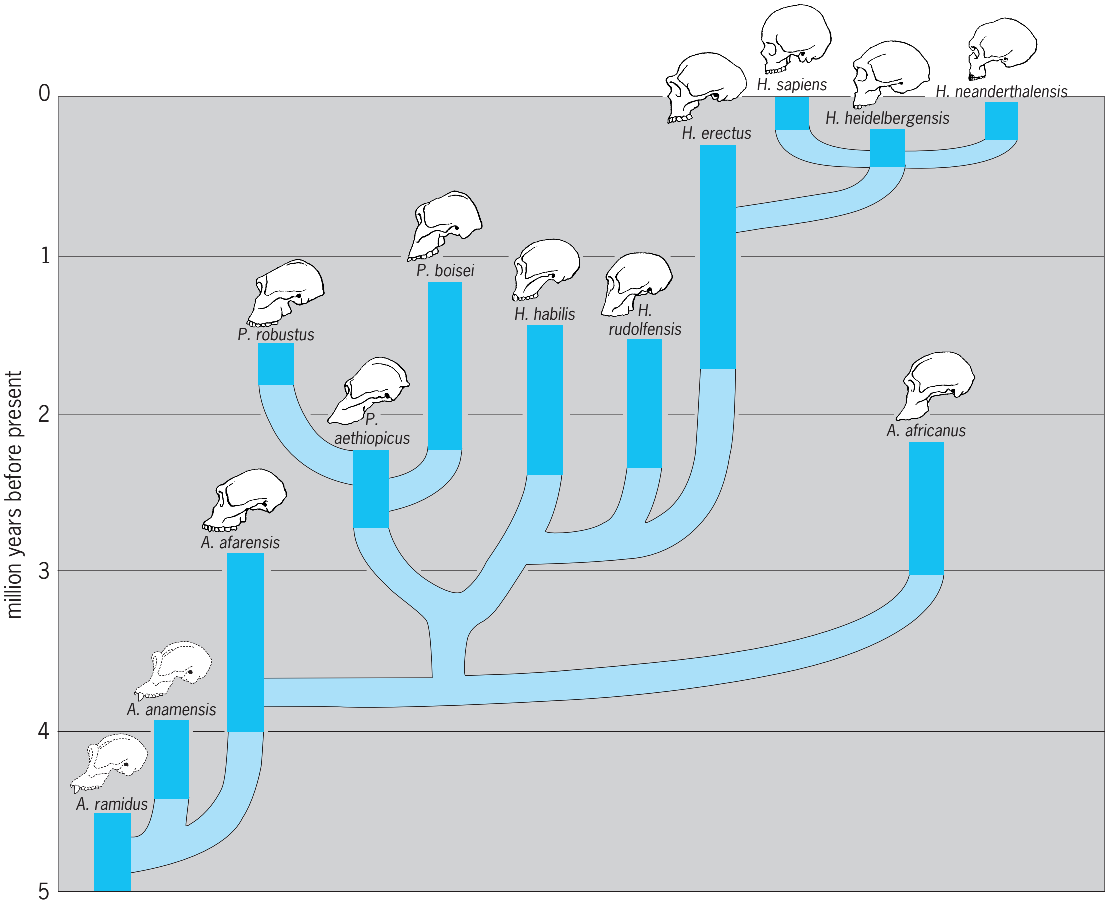


Fig. 3. Phylogenetic tree showing hypothesized evolutionary relationships among species of *Australopithecus*, *Parathropus*, and *Homo*.

Paranthropus aethiopicus is considered to be a likely candidate for the ancestry of both *P. boisei* and *P. robustus*. The reason is that *P. aethiopicus* shares a number of primitive features with *A. afarensis*, but at the same time it shares a host of derived features with the later species of *Paranthropus*, namely *P. robustus* and *P. boisei*.

The lineage leading to *Paranthropus* shares a number of morphological features with that leading to *Homo*. For example, all species of *Homo* and *Paranthropus* share a coronally oriented petrous temporal bone, a foramen magnum that is roughly horizontal in disposition, and a vertically oriented mandibular symphysis. Thus, it is most parsimonious to assume that these two lineages shared a common ancestor at some time prior to 2.8 Ma.

The apparent increase in cranial capacity that is shown by some species of *Paranthropus* (*P. robustus* and *P. boisei*) would appear to parallel the increase in brain size that characterized the evolutionary history of the lineage that led to *Homo sapiens*. Tools made of stone or bone are not known to be associated with *Ardipithecus* or the three species of *Australopithecus*. Stone tools are known from sites that contain *P. boisei* fossils, and both bone and stone tools are known from sites that preserve *P. robustus* remains. However, early members of the genus *Homo* are known also from these same localities. Thus, it is difficult to determine whether *Paranthropus* species may have been responsible for some of the late Pliocene and early Pleistocene archeological record. Indeed, it has been argued that the later species of *Paranthropus* may have been driven to extinction through competition with early *Homo*, because the latter possessed a distinct ecological advantage through the utilization of lithic technology in the procurement of food. While the evidence for this is not compelling, it is possible that ecological interactions between *Paranthropus* and early members of *Homo* may have influenced the evolutionary course of the human genus. *See* APES; FOSSIL HUMANS.

Frederick E. Grine

Bibliography. L. Aiello and C. Dean, *An Introduction to Human Evolutionary Anatomy*, Academic, New York, 1990; D. S. Strait, F. E. Grine, and M. A. Moniz, A reappraisal of early hominid phylogeny, *J. Human Evol.*, 32:17–82, 1997; M. H. Wolpoff, *Paleoanthropology*, 2d ed., McGraw-Hill, New York, 1999.

Authigenic minerals

Minerals that are formed in sediment or a sedimentary rock. Their in-place origin distinguishes them from minerals that are formed elsewhere and transported to the site of deposition (detrital minerals). Authigenic minerals form at the Earth's surface as well as during subsequent burial. The postdepositional processes are referred to as diagenesis, and the resulting minerals are important clues to postdepositional physical and chemical changes in the rock. *See* DIAGENESIS; SEDIMENTARY ROCKS.

Mode of formation. Authigenic minerals precipitate from the overlying water column, pore fluids in the sediment, recrystallization or alteration of preexisting minerals, and structural transformation of one mineral to another. The minerals change in an attempt to equilibrate to the physical and chemical conditions present at any given time. Critical factors in their formation are initial mineral assemblage, temperature, pressure, ionic concentration, pH, electron availability, and the fluid flux through the rock.

In sedimentary rocks it is common to find a record of multiple diagenetic events based on the authigenic minerals. For example, in sediments near the surface, meteoric water may displace original marine pore water, resulting in distinct types of cements. Iron oxide can result from oxidizing fluid. Depletion of oxygen by bacteria may result in the formation of iron sulfides, such as pyrite. During burial, the sediments respond to increasing temperature (up to 200°C; 390°F), pressure (up to 2.5 kilobars; 250 megapascals), and fluid movement from compaction-driven waters or influx of water from the basin flanks. As a result, the sedimentary rock may contain authigenic minerals that record a sequence of events ranging from processes occurring near the sediment-water interface to those forming during deep burial. Unlike metamorphic rocks, the preexisting (detrital) mineral assemblage is at least partially retained, in part due to the sluggish reaction rates at diagenetic conditions. Early cementation processes often seal up the rock, preventing subsequent diagenetic reactions and preserving the original detrital mineral assemblage.

Examples. Authigenic minerals occur in all sedimentary rock and can vary from trace amounts to virtually the total rock (see **table**). The carbonate minerals calcite, dolomite, and siderite are some of the most common types. They form in a wide range of depositional environments and at varying burial depths. Calcite and dolomite form the principal minerals in limestones and dolostones, respectively, as well as cements in sandstones or shales. Carbonate cements result from recrystallization of detrital carbonates and from dissolution of other calcium, iron, and magnesium minerals with carbon dioxide from organic reactions. Much of the calcite in limestones initially consisted of aragonite or magnesium-rich calcite, whereas most dolomite has been formed by the chemical alteration of calcite. Recrystallization may change aragonite to calcite. Aragonite (orthorhombic) is a naturally unstable form of calcium carbonate. With the passage of geologic time, aragonite normally inverts to the more stable calcite (hexagonal). The substitution of magnesium for calcium is responsible for the conversion of calcite or aragonite to dolomite, and it has been shown that dedolomitization (replacement of magnesium by calcium) is also possible. *See* ARAGONITE; CALCITE; CARBONATE MINERALS; CEMENT; DOLOMITE.

Sandstones and shales are often cemented near the sediment-water interface by carbonate, which forms subrounded shapes called concretions (**Fig. 1**). Early-formed iron oxides of hematite or

Common authigenic minerals

Mineral	Formula
Albite	$NaAlSi_3O_8$
Anatase	TiO_2
Anhydrite	$CaSO_4$
Apatite*	$Ca_5(PO_4)_3(F,Cl,OH)$
Aragonite (orthorhombic)	$CaCO_3$
Barite	$BaSO_4$
Boehmite	$AlO(OH)$
Calcite (hexagonal)	$CaCO_3$
Celestite	$SrSO_4$
Clay minerals	
Chlorites*	$(Mg,Fe^{2+},Fe^{3+})_6$-$(Al,Si_3)O_{10}(OH)_8$
Illites*	$K(Al)_2(AlSi_3)O_{10}(OH)_2$
Kaolinite	$Al_2Si_2O_5(OH)_4$
Smectites*	$(Na,0.5Ca)_{0.5}(Al,Mg,Fe)_2$-$(Al,Si_3)O_{10}(OH)_2 \cdot nH_2O$
Dolomite	$CaMg(CO_3)_2$
Gibbsite	$Al(OH)_3$
Glauconite*	$K(Al,Mg,Fe^{2+},Fe^{3+})_2$-$(Al,Si_3)O_{10}(OH)_2$
Goethite	$Fe_2O_3 \cdot n(H_2O)$
Gypsum	$CaSO_4 \cdot 2(H_2O)$
Halite	$NaCl$
Hematite	Fe_2O_3
Leucoxene	TiO_2
Limonite	$FeO(OH) \cdot n(H_2O)$
Opal (amorphous)	$SiO_2 \cdot n(H_2O)$
Orthoclase	$KAlSi_3O_8$
Pyrite (isometric)	FeS_2
Pyrolusite	MnO_2
Quartz	SiO_2
Siderite	$FeCO_3$
Zeolites*	$X_y^{1+,2+}Al_xSi_{1-x}O_z \cdot nH_2O$
Clinoptilolite*	$(Na,0.5Ca,K)_{3.5}\ Al_{3.5}\ Si_{14.5}O_{36} \cdot nH_2O$
Analcime	$NaAlSi_2O_6 \cdot H_2O$
Laumontite	$CaAl_2Si_4O_{12} \cdot 4H_2O$

*Group of minerals characterized by considerable chemical variation.

goethite may coat grains when sediments are exposed to oxidizing waters or iron sulfides (such as pyrite) if the water is low in dissolved oxygen. Quartz cements in sandstones form overgrowths on detrital quartz grains at advanced burial conditions. Potential reactions producing silica include quartz dissolution at quartz-clay grain contacts and breakdown of aluminosilicates such as clay minerals and feldspars. Overgrowth of authigenic feldspar on detrital feldspar is very common. At higher temperatures, calcic plagioclase is replaced by albite.

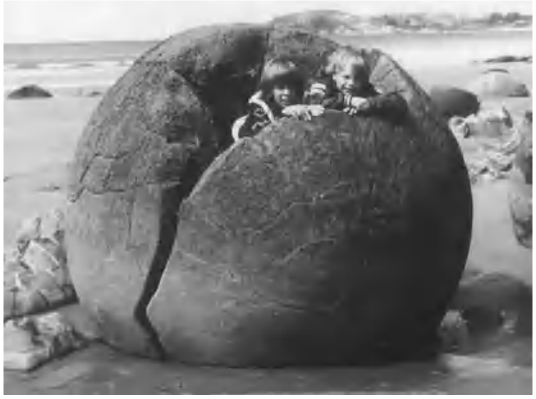


Fig. 1. Calcite concretion formed in Paleocene marine siltstone, Moeraki beach, New Zealand. Uncemented siltstone has been eroded away by wave action. Concretion is about 2 m (7 ft) in diameter, which is unusually large.

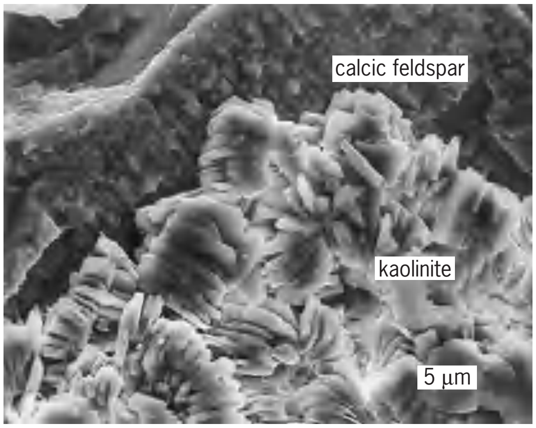


Fig. 2. Scanning electron micrograph of authigenic kaolinite. The clay mineral formed from dissolution of calcic feldspar in a Miocene sandstone, San Joaquin basin, California.

Authigenic clay minerals—chlorite, illite, and smectite—occur as grain coatings and cements. Clay minerals in sandstones commonly result from alteration of feldspars (**Fig. 2**). Commonly the dissolution within the feldspar will result in porosity, and the clay forms a pore-filling cement between the sand grains. In shales, mixed-layer smectite/illite clays are particularly common, and during progressive burial the increases in temperature and time cause the smectite to be converted to illite. Volcanic glass in sediments is often altered to zeolite and smectitic clay minerals (**Fig. 3**). With increasing geologic time the zeolite mineral assemblage becomes simpler due to the breakdown of relatively unstable early-formed zeolites. Pure fine-grain quartz rocks called chert (flint, jasper) crystallize from biogenic opaline silica during burial. *See* CHERT; CLAY; FELDSPAR; ZEOLITE.

Commonly, authigenic mineral assemblages are a result of a series of dehydration and desilicification reactions during progressive burial. Examples are reaction of gypsum to anhydrite, goethite to hematite, opaline silica to quartz, smectite to illite, and transformation of zeolites to feldspars. To balance these reactions, other processes and reactions consume silica and water. Recrystallization in authigenic mineral formation leads to an enlargement of crystal size through the process of dissolution followed by precipitation. Examples are large halite, gypsum, and anhydrite crystals in evaporite deposits and calcite crystals in some limestones. *See* GYPSUM; HALITE.

Identification. Since many of the authigenic minerals shown in the table may be deposited in sediments as detrital minerals, it is important that their authigenic origin be clearly established, especially when these minerals are utilized to interpret diagenetic chemical, temperature, and pressure conditions. When grown into pore space, authigenic minerals often exhibit crystalline shapes. In contrast,

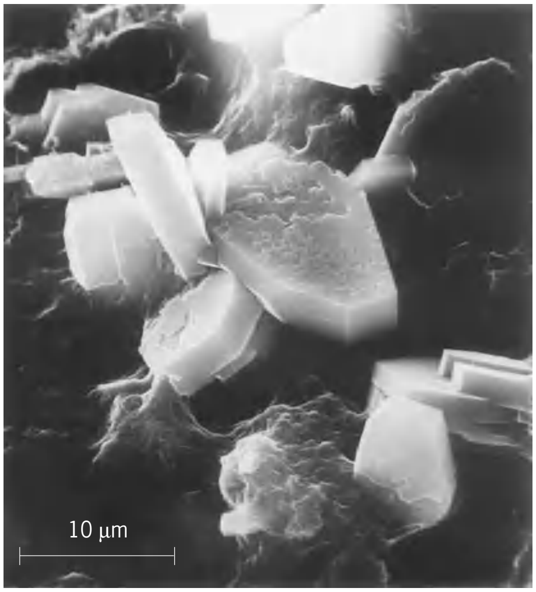


Fig. 3. Scanning electron micrograph of clinoptilolite zeolite (white crystals) on smectite-coated (dark background) sand grains. Sample is from Paleocene volcanic sandstone, Baja California, Mexico.

detrital mineral surfaces are irregular, rough, or rounded because they were leached in a weathering environment, were abraded during transportation, or have been leached further by pore fluids after deposition. Authigenic minerals also form obvious cements that coat grains and completely fill the original pore space in the sediment. In many cases, the authigenic minerals use preexisting crystal structures as templates for their growth. Examples are feldspars and zeolites forming on detrital feldspar, quartz overgrowths on quartz, and carbonates growing on shell tests. Although the authigenic origin of minerals can often be established from the study of thin sections, the scanning electron microscope and electron microprobe are particularly useful for their identification and chemical study. The identification of many fine-grain minerals such as clays and zeolites is possible only with x-ray diffraction and scanning electron microscope techniques (Fig. 2). The high magnification and resolution attainable with the scanning electron microscope make it possible to differentiate these minerals from like species of detrital origin by their crystal form. *See* MICROSCOPE; SCANNING ELECTRON MICROSCOPE; X-RAY DIFFRACTION. James R. Boles

Autism

A neurodevelopmental condition that impairs the way that a person relates to and communicates with other people. Persons with autism also can have unusual behaviors, such as insistence on sameness, obsessions, or stereotypic (purposeless and repetitive) behaviors such as hand flapping, spinning, and toe walking. The condition varies greatly in the presenting symptoms, the timing of presentation, the range and severity of symptoms, and its association with other conditions. While recognition of autism is increasingly common, the cause and the reason for its increase remain unknown. With intensive early intervention, significant improvements in a large percentage of children with autism can be achieved.

Characteristics and diagnosis. Leo Kanner first described autism as a condition in 1943, but descriptions matching autism can be found throughout history. Based on his observation of 11 children, Kanner identified characteristic features of "early infantile autism" that are still considered essential in making an autism diagnosis. Three key features are: (1) social impairments that affect how a person relates to other people, apparent early in life; (2) peculiar use of language, or a failure to use language for interactive conversation; and (3) an obsessive desire for the maintenance of sameness associated with a "limitation in the variety of spontaneous activity." IQ ranged from 50 to 70 in classic Kanner autism, but currently autism is recognized over a broader range of IQ's.

Autism is diagnosed only through behavioral criteria. Typically, children with autism present with language delays, often first prompting a hearing assessment. Stereotypical behaviors such as arm flapping, rocking, and spinning usually develop later, after communication and social impairments are apparent. About a third of parents of autistic children report regression of previously achieved developmental milestones, and sometimes they attribute the loss of milestones to an identifiable stressor, such as a vaccination or the birth of a younger sibling. Two-thirds of the children whose parents reported regression had signs of social impairments or delays noted on review of videotape taken around 12 months of age.

The advantages of early diagnosis are many and generally outweigh concerns about potential harm from "labeling" a child. P. A. Filipek and colleagues authored a consensus statement on the assessment of children suspected of having autism. Developmental surveillance should occur at well-child visits. Concerns about speech, language, or social development warrant further evaluation. First-level evaluations include audiological assessments, testing for lead exposure if indicated by pica behaviors (eating of nonfood substances), and tests specific for autism [such as the Checklist for Autism in Toddlers (CHAT) screen or Autism Screening Questionnaire]. Children who screen positive for autism should be referred for a multidisciplinary assessment, which should include a detailed health and developmental history and a physical exam that looks for identifiable conditions that have autistic features. The history and physical findings should determine the necessity for additional laboratory and imaging studies. The assessment should include additional interviews and observations that specifically assess for autism. A complete evaluation also includes additional developmental screening and speech and language assessment, occupational therapy assessment, behavioral assessment, and a social work assessment.

Etiology. As recently as the late 1960s, autism was attributed to disturbed mother-child relationships.

The term "refrigerator mother" (implying destructive withholding of maternal nurturing) represents a dark period in the history of autism, when misplaced blame added to the anguish experienced by families with autistic children and delayed investigation into its true causes. The recognition that 25% of children with autism developed seizures began the shift in thinking to the more current view that autism is a neurodevelopmental disorder with a biological basis.

Genetics. The causes of autism remains elusive, but data support a strong genetic component. Families with one child with autism are more likely to have another child with autism than previously unaffected families, the risk being 3 to 5% in affected families compared to 0.1 to 0.2% in the general population. Autism occurs in males 3 to 4 times more frequently than females. Studies of families with multiply affected members have identified many chromosomes that are highly associated with autism; however, none are universally found in children with autism. Twin studies have found a concordance rate of 36–91% in identical twins (that is, when one twin has autism, the other twin has autism 36% to 91% of the time) compared with a less than 1% concordance rate in fraternal twins. Autism is more common in some conditions, such as Fragile X, neurofibromatosis, tuberous sclerosis, and Down syndrome, but no such associated condition can be identified in most children with autism. *See* CONGENITAL ANOMALIES; DOWN SYNDROME.

Prenatal and perinatal factors. Autism often manifests after the first year of life, but data from at least two sources suggest that autism often begins at or before the prenatal and perinatal periods. Analyses of neonatal blood spots taken from children later diagnosed with autism showed that 95% of a small sample of children with autism have elevated levels of four neuropeptides (short-chain proteins that can act as neurotransmitters) and neurotrophins (nerve growth factors). In this study, blood spot data for normal children and those with autism, mental retardation, or cerebral palsy were compared. Elevated neurotropin and neuropeptide levels were also found in children with mental retardation but not in normal children or those with cerebral palsy. Specifically, the damage that occurs during pregnancy resulting in cerebral palsy does not produce the changes in blood spot data that were found in autism or mental retardation. Another study found that 42% of children with autism have posteriorly rotated ears (compared with 10% of children without autism), representing changes that occur before birth in the first month of gestation. Other studies have shown neuroanatomic differences. Head sizes of autistic children are on average normal at birth, but become larger than average as the children age. Specific brain structures are differentially affected. Some are larger than normal in autistic children, such as the hippocampus and amygdala, and some are smaller, such as the cerebellar vermis. *See* MENTAL RETARDATION.

Autism spectrum disorders. Autism is a spectrum disorder that also includes Asperger disorder, Rett's disorder, childhood disintegrative disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS). Other autism spectrum disorders (ASD) retain many of the core features of autism.

Asperger syndrome. Deficits in communication and social interactions occur in Asperger syndrome, but to a lesser degree than autism. The significant delay in the onset or early course of language that is seen in autism is not seen in Asperger syndrome. Overall, affected children have a better facility with language and more interest in social activities, but usually have autistic features, such as obsessive preoccupations.

Rett's disorder. Rett's disorder (also known as Rett syndrome), primarily affecting females, has normal early development, then regression some time after 5 months. Girls with Rett's disorder typically carry a single gene marker, the *MECP2* gene, and develop an unsteady gait, loss of or lack of language, constant hand wringing and loss of functional use of their hands, social deficits, and mental retardation. Head sizes of affected children are on average normal at birth, but slowing of head growth during early development results in small head size.

Childhood disintegrative disorder. Children with childhood disintegrative disorder have normal development for 2 to 10 years and then experience severe regression, resulting in complete loss of speech, social interaction, and self-help skills. Recovery has been quite limited in these children.

PDD-NOS. Children with PDD-NOS have some autistic features, such as communication and social impairments and repetitive behaviors, but do not meet the criteria for full syndrome autism and are not better described by one of the other ASD.

Epidemiology. Autism prevalence estimates have steadily increased over the last 20 years. Studies conducted prior to 1985 found that autism affected 4 to 5 per 10,000 children. Studies conducted between 1985 and 1994 show autism prevalence rates closer to 12 per 10,000 children. More recent data suggest autism is even more common, affecting 16 to 40 per 10,000 children. Straffordshire (United Kingdom) preschool children in the late 1990s had an autism prevalence rate of 16.2 per 10,000. In 1998 Brick Township, New Jersey, had autism rates estimated at 40 per 10,000 in 3- to 10-year-olds. Prevalence rates of autism and Asperger syndrome in Cambridgeshire (UK) were 1 in 175 (or ~57/10,000) for 5- to 11-year-olds. Children born in the Atlanta area in 1996 had an estimated autism prevalence rate of 34 per 10,000.

Controversy arises in the interpretation of the observed increase. A 1999 California Department of Developmental Services (DDS) report showed that from 1987 to 1998 autism cases in California's Regional Center System increased by 273%. This report was cited by some as evidence of an autism epidemic. Others challenged that the increase was due to changes in diagnostic criteria, increased awareness, increase in population, migration fluxes, earlier age at diagnosis, recording artifacts, and problems with cross-sectional data. However, a follow-up study that compared Regional Center clients who were born between 1983 and 1985 with those born between 1993 and 1995 concluded that the increase in autism cases was not due to changes in diagnostic criteria, changes in misclassification of autism, or an influx of out-of-state autism cases. In addition, California DDS issued a 2003 report that documents a 100% increase in autism caseload since 1999.

Theories abound to explain the apparent increase in autism rates, given the lack of an identified cause for autism. These theories include: a bowel-brain connection that is altered by some potential insult, such as food allergies, antibiotic use, and immunizations; immunizations with MMR (measles, mumps, and rubella) or thimerosal-containing vaccines; heavy metal or other toxic exposure (such as to PCBs); defective hepatic detoxification capacity; autoimmune disorder or an imbalance between T-cell response; and yeast overgrowth in the bowel. The search for a cure or at least some improvement in their child's condition has led some families to unproven and potentially harmful interventions, which will likely continue until a cause and cure are found.

Epidemiology of other ASD is less well established. Some studies estimated Asperger syndrome as 3 to 4 times more common than autism. Based on a large population study conducted prior to the identification of the *MECP2* gene, Rett's disorder was estimated to occur in about 1 in 23,000. Childhood disintegrative disorder is felt to be rare: <5 per 10,000. PDD-NOS is believed to be more common that autism, but population estimates are hampered by the fuzziness in diagnostic criteria. One small-area estimate from the Brick Township investigation is that autism occurs in 1 in 250 and ASD occur in 1 in 150.

Treatment. Early identification and early intervention with intensive behavioral programs is the cornerstone of effective treatment for autism. In the past, dismal expectations led to minimal intervention resulting in discouraging outcomes. Expectations improved when Ivar Lovaas showed that intensive early intervention could make significant differences for about half of the children enrolled. Subsequently, a number of programs have been developed, including Auditory Integration Training, Daily Life Therapy (Higashi), EarlyBird Programme, LOVAAS/Applied Behavior Analysis (ABA), Treatment and Education of Autistic and Related Communication Handicapped Children (TEACCH), Options Son-Rise Programme, Picture Symbols (PECS) training, and Structure Positive (approaches and expectations) Empathy Low arousal Links (SPELL). Most of these programs are time-intensive and reward mastery of individual small steps that gradually lead to mastery of larger tasks. Adjuncts to intervention programs include diet modification, music therapy, facilitated communication, and computer applications. No one intervention is guaranteed to help every child with autism. Choice of programs is often dictated by program availability and funding. Even though early intensive interventions have been shown to be effective, many families have significant difficulties in obtaining services. There are no known medications that cure autism, but some are used to ameliorate some behaviors and improve functioning.

Long-term outcomes. A diagnosis of autism generally has lifelong implications. Outcomes in adults with autism reveal an overall poor prognosis for independent living and successful employment for the majority. Autistic adults with an IQ below 70 fare poorly, as might be expected by the cognitive limitations alone. In a study of autistic adults with IQ scores of at least 50, among those with more normal intelligence (IQ>70) only 17% achieved a high level of independence, having some friends and a job. Another 15% were working, but required some degree of support in daily living and had some friends or acquaintances. Compared with children currently diagnosed with autism, these adults were diagnosed much later (mean age at diagnosis was 7 years) and had fewer and less extensive therapeutic interventions. Studies of overall competence and language skills have demonstrated better outcomes for recently diagnosed individuals with autism. Continued advances in early diagnosis and development of effective treatments should further improve outcome for the increasing number of children diagnosed with autism. Robert Byrd

Bibliography. S. E. Bryson and I. M. Smith, Epidemiology of autism: Prevalence, associated characteristics, and implications for research and service delivery, *MRDD Res. Rev.*, 4:97–103, 1998; P. A. Filipek et al., The screening and diagnosis of autistic spectrum disorders, *J. Autism Develop. Disorders*, 29(6):439–484, 1999; S. Ozonoff, S. J. Rogers, and R. L. Hendren (eds.), *Autism Spectrum Disorders: A Research Review for Practitioners*, American Psychiatric Publishing, Arlington, VA, 2003; A. M. Wetherby and B. M. Prizant (eds.), *Autism Spectrum Disorders: A Transactional Developmental Perspective*, Paul H. Brookes Publishing, York, PA, 2001.

Autoimmunity

The occurrence in an organism of an immune response to one of its own tissues, that is, a response to a self constituent. Foreign substances to which an organism makes a protective immune response are called nonself. Efficient discrimination between self and nonself, the basis of normal immune function, depends upon a function known as immune tolerance, which means an inertness to substances that could be capable of provoking an immune response. Failure of immune tolerance to self constituents results in an autoimmune response which is often, although not invariably, associated with autoimmune disease. Autoimmune disease occurs when the autoimmune response to self constituents has damaging effects of a structural or functional character. *See* ACQUIRED IMMUNOLOGICAL TOLERANCE.

History. In 1901, P. Ehrlich proposed that, as a condition of life, organisms should not be reactive against their own tissues. This view influenced scientists until the 1950s, when it was found that autoimmunity could be experimentally created in animals by vigorous immunizing procedures. Soon thereafter, laboratory tests in humans showed that immune responses to self tissues could occur spontaneously and could be associated with disease of the blood or thyroid gland. Procedures for laboratory diagnosis of autoimmunity developed rapidly, based on tests for detection of antibodies in the blood plasma. Antibodies are globulin molecules that react with a stimulating substance, the antigen. Autoantibodies are equivalent globulin molecules in blood that react with self antigens; their presence points to an autoimmune reaction in the body. By the 1960s, it was recognized that autoimmunity was a direct or contributing cause of a wide range of human diseases. *See* ANTIBODY; ANTIGEN; IMMUNOGLOBULIN.

Normal immunity mechanisms. When any foreign antigenic substance, such as a microorganism, protein, carbohydrate, nucleic acid, or a complex of these, is introduced into the body, it is recognized by a genetically determined process as being nonself. The foreign antigen is first degraded by phagocytic cells, which then present processed antigen to cells of the lymphoid system. The specialized cells with this function are part of the lymphoid reticuloendothelial system and include dendritic cells and macrophages. Lymphoid cells, after a recognition process based on surface receptor structures, undergo proliferation and generate one population of lymphocytes which reacts specifically with the antigen, and a second group which produces specifically reactive antibody molecules. Long-lived descendants of these populations of lymphoid cells provide for "memory," which allows for accelerated responses in the event of subsequent exposure to the same antigen. Immune responses are exemplified by the neutralization of toxic protein molecules, the killing of infecting microorganisms, the rejection of foreign tissue grafts, and memory by the protective effects of vaccination.

Lymphocytes that participate in immune responses belong to two major groups. One group, which matures in the thymus gland, comprises thymic or T lymphocytes, of which there are several subsets. These subsets have different functions and carry unique surface molecules which can be identified with monoclonal antibodies: (1) The subset of helper T lymphocytes which is marked by the CD4 molecule responds to antigens by releasing stimulatory cytokines (intercellular hormones) that can amplify the number and activity of lymphocytes participating in the immune response; (2) the subset of cytolytic T lymphocytes which is marked by the CD8 molecule can directly recognize and kill cellular targets, usually virus-infected cells; and (3) the subset of suppressor T lymphocytes, which also carries the CD8 molecule, releases molecules that reduce the intensity of immune responses, or switch these off altogether. The other major group of lymphocytes, which mature in the bone marrow, are B lymphocytes. After stimulation with antigen molecules, and under the influence of factors released by helper T cells, B lymphocytes proliferate and they later secrete the antibody molecules which, when circulating in the blood, provide for humoral immune responses. *See* CELLULAR IMMUNOLOGY; IMMUNITY.

Major histocompatibility complex (MHC). This gene complex on human chromosome 6 codes for cell-surface molecules which confer biological uniqueness on cells of an individual. The complex is called HLA in humans: it is important for the matching of transplanted organs, and also has fundamental importance for immune responsiveness. There are two classes of MHC (HLA) molecules. Class I molecules are expressed on all cells and are recognized by $CD8^+$ cytolytic T lymphocytes; this agrees with the fact that any cell may be infected with a virus and thus should be susceptible to killing. Class II molecules are expressed normally only on those cells, B cells or antigen-presenting cells, with the special function of interacting with and stimulating $CD4^+$ helper T lymphocytes. The expression of MHC (HLA) molecules on lymphocytes is amplified by cytokines; for example, the cytokine interferon-gamma can induce MHC class II molecules on cells which normally are incapable of presenting antigen to helper T lymphocytes. Such aberrant expression of MHC molecules is relevant to autoimmunity. *See* HISTOCOMPATIBILITY.

Recognition by the immune system. Concepts on the origin of specific recognition structures on the surface of cells of the immune system developed after the mid-1970s. During embryonic development there is generated information in the genome of cells of the immune system which provides for reactivity with all antigens ever likely to be encountered in later life. Thus there is genetic programming for the structure of the cell-surface receptor molecules that enable antigens to be recognized, and for the specificity of the antibody molecules that combine with antigens in the blood. Antigen selects lymphocytes with the appropriate surface receptor and stimulates these to proliferate as an identical clone of cells: this is the clonal selection concept of immune responsiveness. In the essentially random process of generation of antibody diversity, reactivity for self antigens will be created, and this must be dealt with, as described below.

Regulation of immune responses. The immune system remains in a state of balance conditioned by positive (on) signals and negative (off) signals. Positive signals are provided by antigen in low dose and the amplifying factors released by activated helper T lymphocytes, while negative signals are provided by antigen present in excess which causes an overload paralysis, and by suppressor T lymphocytes which are generated preferentially when self antigens are presented. There is still a lack of full understanding of the mechanism by which unwanted immune responses to self antigens are suppressed so as to provide for natural tolerance to self. The major processes are (1) permanent deletion, or functional inactivation in early life of cells capable of responding to self antigens; and (2) regulatory controls, which inhibit the activity of self-reactive lymphocytes that escape the deletion process. The relative contribution of these two mechanisms for specific self antigens appears to differ, and both probably operate to control autoimmunity. Deletion of self-reactive cells would be the more economical means of achieving self-tolerance, but it is the less flexible. It is likely that these processes apply more particularly to T lymphocytes; since B lymphocytes are dependent on helper T lymphocytes for their activation and maturation, a lack of T-cell help is one way in which B-cell function can be reduced. There are low background levels of immunologic reactivity to many self antigens in healthy subjects, indicating that suppressor activity over immune responses to autoantigens must be continuously operative.

Causes. Failure of immune regulation is responsible for autoimmune disease. Inheritance may account for 25–50% of the risk for autoimmune diseases, since it is known that autoimmune disease, or at least the tendency to produce autoantibodies, runs in families. There are many genetic determinants, and they are poorly understood. One set is in some way associated with MHC (HLA). Since products of HLA genes normally function to direct T lymphocytes to cells with which they should interact, it is not surprising that autoimmune diseases are associated with the presence of particular HLA types; examples include B8 (thyrotoxicosis), DR4 (rheumatoid arthritis, type 1 diabetes mellitus), and DR2 (multiple sclerosis). The reason may be that the autoantigen readily associates with the MHC (HLA) molecule on cells which present antigen to helper T lymphocytes. The MHC influences the occurrence of autoimmunity in other ways. Thus release of cytokines by T lymphocytes may induce aberrant expression of class II molecules on tissue cells which then can present their own antigens, and these become inducers of and targets for an autoimmune response. In addition to the MHC, there are other inherited determinants of autoimmunity, including genes specifying immunoglobulin structure, and genes specifying weakness in the down-regulation of immune reactions, thus allowing autoimmune responses to become amplified. *See* IMMUNOGENETICS.

There may also be somatic genetic causes of autoimmunity, meaning random mutations in later life among genes that code for immunoglobulins that function as recognition structures on the surface of B lymphocytes; such a mutation may generate a cell with a receptor structure with exquisite specificity for a self antigen which is resistant to regulation. Environmental causes could include infection with microorganisms that carry antigenic structures closely resembling those of self; these, when presented to the immune system, could provoke an uncontrolled response to the related self structures of the body.

Any autoimmune response must become self-sustaining, which implies coexisting failure of normal regulatory processes, either by reason of genetic predisposition or by an acquired disruption of immune function. Clearly the induction of autoimmunity is multifactorial, and occurs when there is a synchronous breakdown of several physiologic fail-safe functions that normally operate to maintain homeostasis in the immune system.

Effects of autoimmune reactions. Once self-sustaining, the autoimmune reaction can cause damage or dysfunction in one of several ways. First, autoantibody molecules circulate in the blood and, by attaching to self antigens on cell surfaces, either damage cells or interfere with important cell-surface receptor molecules. Second, antibodies can unite with their autoantigen, which results in the binding of a serum factor, complement, to form immune complexes that are capable of provoking inflammatory responses; for example, immune complexes deposited in the filtering tissue of the kidney cause progressive nephritis. Third, there may be generated T lymphocytes with the capacity for cellular destruction, and these may cause the progressive inflammatory damage that characterizes autoimmune reactions in solid organs.

Autoimmune diseases. Many human diseases can be attributed to autoimmune reactions. Circulating autoantibodies are responsible for diseases in which there is intravascular destruction of elements of the blood, for example, the red blood cells in hemolytic anemia. T lymphocytes may be responsible for some types of thyroid goiter, such as Hashimoto's disease; a stomach mucosal degeneration that results in nonabsorption of vitamin B_{12} and thus the blood disease pernicious anemia; the insulin-dependent or juvenile type of diabetes mellitus; and one type of chronic hepatitis. Immune complexes cause glomerulonephritis and most of the features of systemic lupus erythematosus, in which autoantibodies are formed to various constituents of cell nuclei. In Sjögren's disease, in which salivary and lacrimal glands are destroyed, damage by T lymphocytes within the glands may be accompanied by damage by immune complexes throughout the body. Some autoimmune diseases are caused by antibodies to cell receptors, which either block neuromuscular transmission, as in myasthenia gravis, or stimulate thyroid cells to overactivity, as in Graves' disease. Some important human diseases may be autoimmune disorders, although demonstration of an autoimmune basis is not yet adequate: these include rheumatoid arthritis, multiple sclerosis, and ulcerative colitis. In many autoimmune diseases the autoantigenic trigger has not been identified; perhaps autoimmunity supervenes after an unidentified primary cause has set the process going. *See* ANEMIA; ARTHRITIS; DIABETES; HEPATITIS; MULTIPLE SCLEROSIS; MYASTHENIA GRAVIS.

Animals. Autoimmune diseases are not confined to humans. The numerous examples in the animal world include hemolytic anemia and lupus, which occur as genetically dependent diseases in certain inbred strains of mice, and diabetes in inbred strains of rats; outbred domestic animals, notably dogs, develop autoimmune diseases, such as lupus, equivalent to those seen in humans. Study of these models in animals has been less informative than might have been expected in terms of clarifying causes of autoimmunity and autoimmune disease.

Treatment of autoimmune disease. Autoimmune diseases are alleviated by treatment, though these diseases are seldom actually curable. At the simplest level, replacement of the specific secretions of tissues or organs damaged by autoimmune reactions may help: this applies to treatment of diseases resulting from atrophy of the thyroid gland (myxedema), stomach mucosal degeneration (pernicious anemia), and destruction of pancreatic islet tissue (diabetes mellitus). For multisystem autoimmune disease, such as lupus, there are drugs, particularly cortisone derivatives, that modify the harmful effects of humoral or cellular autoimmune attack on tissues and so allow the body to reestablish immunologic homeostasis. Also used are cytotoxic immunosuppressive drugs, which are given specifically to inhibit the activity of immunologically active cells responsible for autoantibody formation or for cytolytic damage to tissues. In the future, biological modifiers may allow facilitation of normal regulatory control by modifying various stages between the triggering of an autoimmune response and the destructive effect on the target organ. Finally, as techniques for molecular characterization of autoantigens are developed, it may become possible to inject these in such a way as to reestablish specific self-tolerance to the autoantigen responsible for an autoimmune disease. *See* IMMUNOLOGY.

Ian R. Mackay

Bibliography. C. A. Bona et al. (eds.), *The Molecular Pathology Autoimmune Disease*, 1993; I. R. Mackay and F. M. Burnet (eds.), *Autoimmune Diseases: Pathogenesis, Chemistry and Therapy*, 1963; N. R. Rose and I. R. Mackay (eds.), *The Autoimmune Diseases* 3d ed., 1998; N. Talal (ed.), *Molecular Autoimmunity*, 1991.

Automata theory

A mathematical model of computing. The versatility of modern computers and their applications to the study of complex and dynamic systems, such as financial markets, cellular growth, and communication networks, raise questions about the ultimate power of computers. For example, what are the limitations, if any, on the tasks that we can ask computers to perform? Can we use computers to carry out any computational task we want? To answer such questions, we generally make use of abstract mathematical models. One of the most important mathematical abstract models that has been widely used to simulate objects and processes such as computer and digital circuits is automata theory. *See* ABSTRACT DATA TYPE; COMPUTER.

Alan M. Turing, a British mathematician, first examined the notion of an automaton, or abstract machine, in 1936 while studying the limits of human ability to formalize methods of solving problems. At that time, Turing was trying to solve a decision problem to determine whether there is a general method that can be applied to any assertion to determine whether the assertion is true. In his approach to this problem, Turing invented the most general model of a computing machine, the Turing machine. Turing machines are universal in the sense that every known algorithm can be executed by a Turing machine. For purposes of explanation on how an automata works, we will use less general models such as the finite-state machine and pushdown automata. We will consider these two types of automata, along with the Turing machine.

Finite-state machines. A finite-state machine is the simplest of all automata theory models and serves as the control device for all other automata. The term "finite-state machine," depending upon the application, is defined in many different ways in the automata literature. The commonality of these definitions is that finite-state machines model computing devices that have a fixed and finite amount of memory and read sequences of symbols made up from a finite set of input symbols. The main differences in the definitions concern what the machines do in the way of output. Finite-state machines find applications in the computer field in areas such as spell checkers, sequential circuits, and compiler design.

Finite-state machines can be described by (1) a finite and nonempty set of states, with one of these states is designated as the starting state, (2) a finite set of input symbols (or input alphabet), and (3) a transition or next-state function that assigns to a current state, on a given input symbol, one or more states called the set of next states. If the transition function assigns only single next states, the finite-state machine is called deterministic; otherwise, it is called nondeterministic. Other types of finite-state machines may have a finite set of output symbols (or output alphabet) and an output function. The output function may associate an output value (from the output alphabet) with every transition (Mealy machines) or produce an output value based on the current state (Moore machines). Finite-state machines also have a subset of their states designated as final or accepting states.

As indicated in **Fig. 1**, a finite-state machine can be conceptualized as a control unit that has an input tape head that reads symbols from a linear tape in a sequential manner, one character at a time, from left to right. The control unit is made up of states. Initially the machine is in starting state and reads the leftmost symbol of an input string—a sequence of symbols or characters of a given alphabet. Each individual symbol of the input string occupies a single cell on the tape.

In mathematical terms, a finite-state machine, M, with no output can be defined as a five-tuple $M = (Q, \Sigma, \delta, q_0, F)$, where Q is a finite set of control states, Σ the set of input symbols, δ the transition function ($\delta: Q \times \Sigma \rightarrow Q$), q_0 the initial state, and F ($F \subseteq Q$) the set of final states. The term "five-tuple" is used to indicate that there are five components that

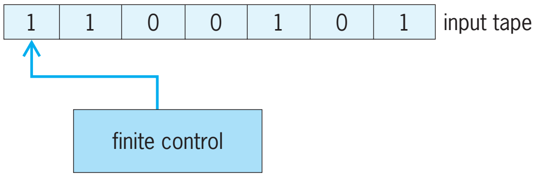


Fig. 1. Abstract representation of a finite-state machine.

State	Input symbol	
	0	1
S_0	S_1	S_0
S_1	S_1	S_2
$\underline{S_2}$	S_1	S_0

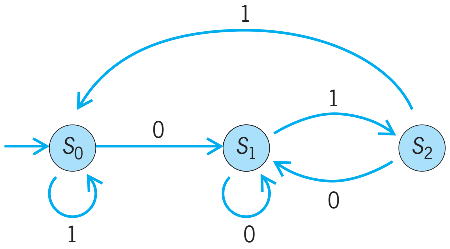


Fig. 2. Transition table and state diagram of a finite-state machine.

define this type of mathematical structure. As long as these elements are clearly identified, the order in which they appear in the five-tuple is immaterial. The expression $\delta\colon Q \times \Sigma \rightarrow Q$ represents a function that maps pairs of elements of Q and Σ (in that order) into elements of Q. The interpretation of $\delta(q, a) = p$, for q and p in Q and a in Σ, respectively, is that M, in state q and scanning the input symbol a, moves its input head one cell to the right and goes to state p.

Finite-state machines are generally represented by state diagrams or transition tables as shown in **Fig. 2**. State diagrams are directed graphs with labeled edges, where circles represent states and arrows represent transitions. Input symbols associated with transitions are shown along the arrows. An incoming arrow identifies the starting state, and the accepting states are shown as double circles. The individual components of the finite-state machine of Fig. 2 are $Q = \{S_0, S_1, S_2\}$, $\Sigma = \{0, 1\}$, $q_0 = \{S_0\}$, $F = \{S_2\}$. The transition function, δ , determines the next state based on the current state and the current input symbol. For example, if the current state of the machine is S_0 and the current input symbol is 0, then S_1 is the next state. Likewise, if the current state is S_1 and the current input symbol is 1, then S_2 is the next state.

Transition tables convey the same information as state diagrams. In transition tables, there is a column for each input symbol and a row for every state. Table entries represent transitions. We will adopt the convention that the first row of the table corresponds to the starting state and that the accepting states are underlined.

The finite-state machine of Fig. 2 is an example of an acceptor. A finite-state machine accepts an input string when, after reading the last character of the string, the machine finds itself in one of its accepting states; otherwise, the machine rejects the input string. The set of all input strings accepted by a finite-state machine is called its language. The finite-state machine of Fig. 2 accepts an input string such as 1100101, and its language consists of all strings that end in 01.

Pushdown automata. Finite-state machines can be used only in computational tasks that require a fixed and finite amount of memory. Processes such as the recognition of valid algebraic expressions by a compiler cannot be carried out under these memory restrictions. Therefore, to perform tasks that require no memory limitations, it is necessary to augment the finite-state machine model and its memory capabilities. One model that accomplishes this is the pushdown automaton. This type of automata extends the finite-state machine with a pushdown list or stack (**Fig. 3**). A stack is a list in which insertions and deletions are possible. Both operations take place at one end of the list called its top. Symbols on a pushdown list can be inserted or deleted in the same way we can add to or take plates from a stack of plates in a cafeteria. We can only put a plate at the top and can only remove the topmost plate.

Formally a pushdown automaton, P, is a seven-tuple $P = (Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$, where Q is a finite set of control states, Σ is a finite input alphabet, Γ is a finite alphabet of pushdown list symbols, and δ defines the control operations. The initial state of the machine is q_0. An element of Γ, Z_0, is the start symbol of the stack. And F ($F \subseteq Q$) is the set of final states.

A pushdown automaton can be in one of a finite set of states. Symbols from Γ are inserted on top of the stack by a "push" operation. The topmost symbol of the stack is deleted by a "pop" operation. As with any finite-state machine, the processing of an input string is carried out in a series of steps. Each step is determined by a control and its current configuration. The latter term refers to a 3-tuple formed by the current state, the current input symbol, and the symbol at the top of the stack. Each step may change the configuration of the machine by changing its state and pushing or popping the stack. However, a pushdown automaton, unlike a finite-state machine, may take several steps to process a single input symbol. At each step, the machine determines whether it is finished processing the current input symbol, in which case it reads the next input symbol, or whether the next step should keep the current input symbol. To process an input sequence, the pushdown automaton should be in its starting state, q_0, with Z_0 on top of the stack; the current symbol should be the first character of the string. The pushdown automaton then carries out operations called for by its control. If a transition is called for, a new current input, top stack symbol, and state are obtained and the control is called to carry out another operation. If an exit is called for, processing stops. The control operations cannot ask for a next symbol past the last character of the input string. Likewise, the stack cannot

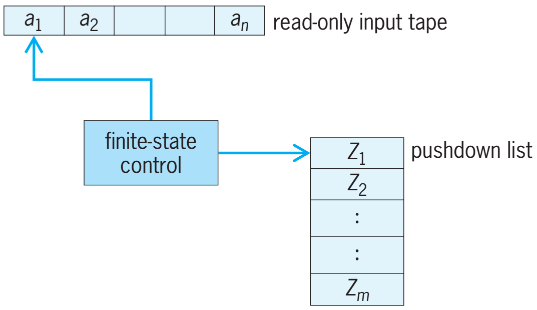


Fig. 3. Conceptual representation of a pushdown automaton.

be popped if Z_0 is at the top of the stack (an empty stack).

Turing machines. The Turing machine is the most powerful of all computational models. Since the publication of Turing's work, many other mathematical systems, such as the production systems of E. Post, the logical systems of A. Church, and the recursive functions of S. Kleene, have been yielded equivalent concepts. The results of these independent works support the argument that the computing power of the Turing machine represents a fundamental limit on the capability of realizable computing devices (Church's thesis). Put simply, Church's thesis states that a Turing machine can describe any computation that can be performed on a computer. In other words, a Turing machine can carry out any algorithm that can be executed by a computer. Thus, if there ever existed a procedure that could be mechanically carried out in a finite sequence of instructions but could not be described by a Turing machine, that procedure could not be programmed for any existing computer. *See* ALGORITHM.

Specifications for the Turing machine have been presented in different ways in the literature. We will consider a basic model that resembles the conceptualization of Fig. 1. It has a finite control with a tape head that, at any time, can read from a cell or write to a cell a single symbol from a finite set of tape symbols. The tape has a leftmost cell but it is infinite to the right. Initially, the n ($n \geq 1$) leftmost cells hold an input string whose symbols belong to a subset of the tape symbols called the input symbols. All cells to the right of the last character of the input string hold a noninput symbol called the blank. The Turing machine can make a move depending on the state of the control and the current symbol. In a move, the Turing machine may change its state, move the tape head left or right one cell, or write a nonblank symbol on the current cell. Any symbol written into a cell replaces its previous content. Pairs of current state and current symbol that do not have a specified move halt the machine. Whenever this happens, the remaining symbols on the tape form the output of the machine. A sequence of moves followed by a halt is called a computation. A Turing machine accepts some input string if it halts in it. The set of all accepted input strings is called the language accepted by the Turing machine.

Mathematically, a Turing machine, T, is a six-tuple, $T = (Q, \Sigma, \Gamma, \delta, q_0, F)$, where Q is a finite set of control states. Γ is a finite set of allowable tape symbols; one particular symbol, denoted by B, is the blank. Σ ($\Sigma \subseteq \Gamma$) is the set of input symbols; Σ does not include the blank. δ defines the control operations and is called the next-move function (δ: $Q \times \Sigma \rightarrow Q \times \Gamma \times \{\text{Left, Right}\}$). The function, δ, may be undefined for some arguments. The initial state of the finite control is q_0. And F ($F \subseteq Q$) is the set of final states.

	0	1	B
q_1	$1\ q_3\ S$	$0\ q_2\ R$	
q_2	$1\ q_1\ R$	$B\ q_2\ L$	
q_3			

Fig. 4. Table representation of a Turing machine.

Turing machines can be described by tables similar to that of the finite-state machines. **Figure 4** shows a table with input symbols 0, 1, and B and states q_1, q_2, and q_3. Entries in the tables represent moves of the machine. A blank entry indicates that the machine halts. The tape head, from its current position, can move to the right cell (R), to the left cell (L), or stay on the same cell (S). The entry, $1q_1$R, for state q_1 and input symbol 1 indicates that the machine writes 0 on the current cell, the next state is q_2, and the tape head moves to the right cell. This Turing machine, for example, for the input string 1010 will produce the output 0101.

Ramon A. Mata-Toledo

Bibliography. J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 2d ed., Pearson Addison-Wesley, 2000; P. Linz, *An Introduction to Formal Languages and Automata*, 3d ed., Jones & Bartlett, 2000.

Automated decision making

The use of computers to automate decision tasks. A decision task is any task requiring the generation or selection of options.

Related to automated decision making (ADM) are techniques for automated control. The term "control" is usually applied to continuously operating systems that are constantly monitored and adjusted. The term "decision making" usually refers to comparatively high-level tasks with discrete decision points. For example, continuous monitoring and adjustment of operating factory equipment may be considered a control task, while the scheduling of factory operations may be considered a decision task. *See* CONTROL SYSTEMS; PROCESS CONTROL.

Another related area is decision support systems. Decision support systems are intended to support human decision making, while automated decision making concentrates on completely computerautomated decision making. Nevertheless, there is a close connection between these two areas since many decision support systems are designed as advisory systems, where the decision support system recommends decisions. To generate a recommendation, the decision support system invokes an ADM model. Consequently, both automated decision making and decision support systems rely on the same underlying ADM models; and it is the maturity and reliability of the ADM models which determine whether the system is used to automate or just to support decision making. *See* DECISION SUPPORT SYSTEM.

Technological foundations. Techniques for automated decision making are drawn from several disciplines. Alternative approaches to automated decision making can be characterized in terms of a trade-off between power and generality. An ADM technique is general to the extent that it can be applied to diverse problems. It is powerful to the extent that it quickly generates good answers.

Optimization. Mathematical optimization techniques can be used to build ADM models that are powerful but of limited generality. For example, linear optimization techniques can be used to assign resources optimally, but only if the resource allocation problem can be described by linear functions and inequalities. *See* OPERATIONS RESEARCH; OPTIMIZATION.

Artificial intelligence. Artificial intelligence techniques are more general but less powerful than optimization techniques. Many artificial intelligence techniques are designed to encode the diverse knowledge that a person might use to make decisions. However, the cost of this generality is that most artificial intelligence systems usually search for a satisficing solution, which is any solution that satisfies a set of constraints describing an acceptable solution. Artificial intelligence techniques have been applied to a broad range of decision-making problems, from medical diagnosis to automated game playing. *See* ARTIFICIAL INTELLIGENCE.

Decision theory. This is the science of decision making. Decision theory is partitioned into two fields. Normative decision theory searches for abstract characterizations of ideal decision-making behavior. The emphasis is on finding a theoretically justified approach to how judgments and decision should be made. Descriptive decision theory is a branch of psychology that performs research on human judgment and decision making. The emphasis is on characterizing how people actually make decisions.

Normative decision theory provides diverse hypotheses and option selection techniques. These techniques are particularly useful when the decision problem can be framed as a few discrete options. The use of techniques from game theory helps select options when the decision maker needs to consider the possible actions of other decision makers (for example, a business competitor). *See* GAME THEORY.

Descriptive decision theory is useful for characterizing the types of decision problems where people tend to make errors and may therefore benefit from good decision support or automated decision making. For example, one line of research in descriptive decision theory focuses on the use of linear models to predict and replace human judgment. This tradition is based on three findings. First, in probabilistic domains simple linear models (for example, adding up pros and cons) are often good predictors of human judgment. Second, this result holds true even for domains where people employ a complex expert knowledge to make judgments (for example, clinical judgment). Third, linear models of expert judgment usually perform better than the experts from which they were derived.

Thus, in the decision-theory approach to automated decision making, an ADM system considers the same variables that a human expert considers in making a decision, but does not emulate human expert decision-making processes. Rather, the variables are integrated into an overall decision procedure using techniques from normative decision theory. *See* DECISION THEORY.

Integration of techniques. During the 1990s there was a trend toward integration of ADM techniques. Decision-theory approaches now strongly influence artificial intelligence research in automated decision making, and there has also been an active interest in integrating optimization and artificial intelligence problem-solving approaches.

Application to simulation. Commercial computer games and many military training simulations involve an active interaction of "players" with other computer-controlled agents in a simulated world. In both cases, the value of the simulation (whether for entertainment or effective training) depends on the realism and intelligence of the simulated agents. Simulated agents must make decisions in their simulated worlds and are therefore executing automated decision making. *See* SIMULATION; VIDEO GAMES.

Autonomous systems. There is an increasing demand for autonomous systems that can perform various complex tasks without human control or intervention. In 1999, for instance, the National Aeronautics and Space Administration (NASA) released control of its *Deep Space One* spacecraft to an autonomous planning and decision-making system. As the demand for autonomy increases, so will the reliance on ADM capabilities. *See* SPACE PROBE.

Paul E. Lehner

Bibliography. D. Brown and C. White (eds.), *Operations Research and Artificial Intelligence: The Integration of Problem-Solving Strategies*, Kluwer Academic, 1990; E. Castillo, J. M. Gutiérrez, and A. S. Hadi, *Expert Systems and Probabilistic Network Models*, Springer-Verlag, New York, 1996; R. Dawes, The robust beauty of improper linear models in decision making, *Amer. Psychol.*, 34(7):571–582, 1979; P. Lehner and L. Adelman, Behavioural decision theory and its implications for knowledge engineering, *Knowl. Eng. Rev.*, 5(1):5–14, 1990; D. Olson and J. Courtney, Jr., *Decision Support Models and Expert Systems*, Dame, Houston, 1997.

Automatic frequency control (AFC)

The automatic control of the intermediate frequency in a radio, television, or radar receiver, to correct for variations of the frequency of the transmitted carrier or the local oscillator. In high-fidelity broadcast receivers AFC keeps distortion due to detuning to a low figure. In the reception of long-haul telegraph signals, AFC reduces the error rate due to signal pulse distortion or interference from lower intensity signals in the same frequency band.

Single-sideband receivers receive signals which are transmitted with a carrier level that is reduced to as small a proportion as 5%, or less, of the sideband (intelligence) amplitude. Proper demodulation requires the generation locally of a carrier-frequency signal synchronized to the transmitted carrier by AFC. Since propagation at frequencies of 3–30 MHz is dependent upon reflections from the ionosphere, motion in this medium will speed up and retard the arrival of the wave, causing Doppler-effect frequency changes. Transmission to and from speeding aircraft will also suffer Doppler frequency drift. To reduce these effects, the carrier is transmitted for synchronization so that the frequency difference between the carrier and side frequency is maintained on reception.

AFC techniques are varied. One uses a discriminator to furnish a voltage whose magnitude and polarity are determined by the frequency change. This voltage is used to adjust the frequency of the local oscillator of the receiver, thereby keeping the intermediate frequency constant. A second technique uses two-polarity pulse accumulation which furnishes a dc potential proportional to frequency error. Another technique for spread-spectrum transmissions uses the recovered chipping frequency for AFC.

To select only the carrier for AFC, very narrow bandwidths are utilized. As an example, a bandwidth of 30 Hz at 70% of maximum response is quite common. This is accomplished in telemetry transmissions by using phase modulation with low index so a carrier component remains for receiver AFC. The response of the control circuits is usually designed to be slow and to be inactive below a determinable level of carrier input. These techniques reduce noise and interference as well as frequency-control capture by undesired adjacent carriers. *See* RADAR; RADIO RECEIVER; TELEVISION RECEIVER. Walter Lyons

Automatic gain control (AGC)

The automatic maintenance of a nearly constant output level of an amplifying circuit by adjusting the amplification in inverse proportion to the input field strength, also called automatic volume control (AVC). Almost all radio receivers in use employ AGC. In broadcast receivers AGC makes it possible to receive incoming signals of widely varying strength, yet have the sound remain at nearly the same volume. In communications receivers a type of AGC circuit called a squelch circuit is used to prevent noise during periods of no transmission, such as in the reception of on-off keying, frequency-shift keying, and phone. AGC is also useful in accelerating the switching action between receivers in diversity connection. *See* RADIO RECEIVER.

AGC action depends on the characteristic, possessed by most electronic tubes and transistors, of adjustment of gain by the variation of the applied bias voltage. If the dc voltage applied to the control grid of a vacuum tube is made more negative, the amplification of that stage will be reduced.

In most broadcast receivers the AGC voltage is taken from the detector. This dc voltage, proportional to the average level of the carrier, adjusts the gain of the radio-frequency (rf) and intermediate-frequency (i-f) amplifiers and the converter (**Fig. 1**). AGC tends to keep the input signal to the audio-frequency (af) amplifier constant despite variations in rf signal strength. There are several modifications of this basic circuit.

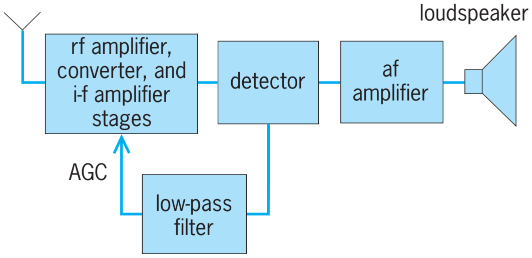


Fig. 1. Block diagram of broadcast receiver using AGC.

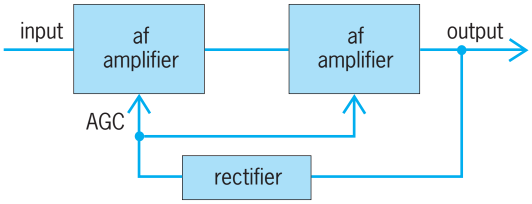


Fig. 2. Block diagram of AGC of an audio amplifier.

Perfect AGC action would provide a constant output for all values of input signal strength. A slightly rising output characteristic with increased signal strength is generally desirable to facilitate proper tuning. The figure of merit applied to AGC action is given as the change in input required for a given output change. An example of a good figure of merit may be seen when an 80-dB change in input carrier signal results in an output change of no more than 3 dB. This applies to the unmodulated carrier strength only, since the modulation of the carrier must always vary as the modulation of the transmitter.

AGC circuits are also used in dictation recording equipment, public address systems, and similar equipment where a constant output level is desirable. **Figure 2** shows a typical block diagram for such equipment. *See* AMPLIFIER. Walter Lyons

Automatic landing system

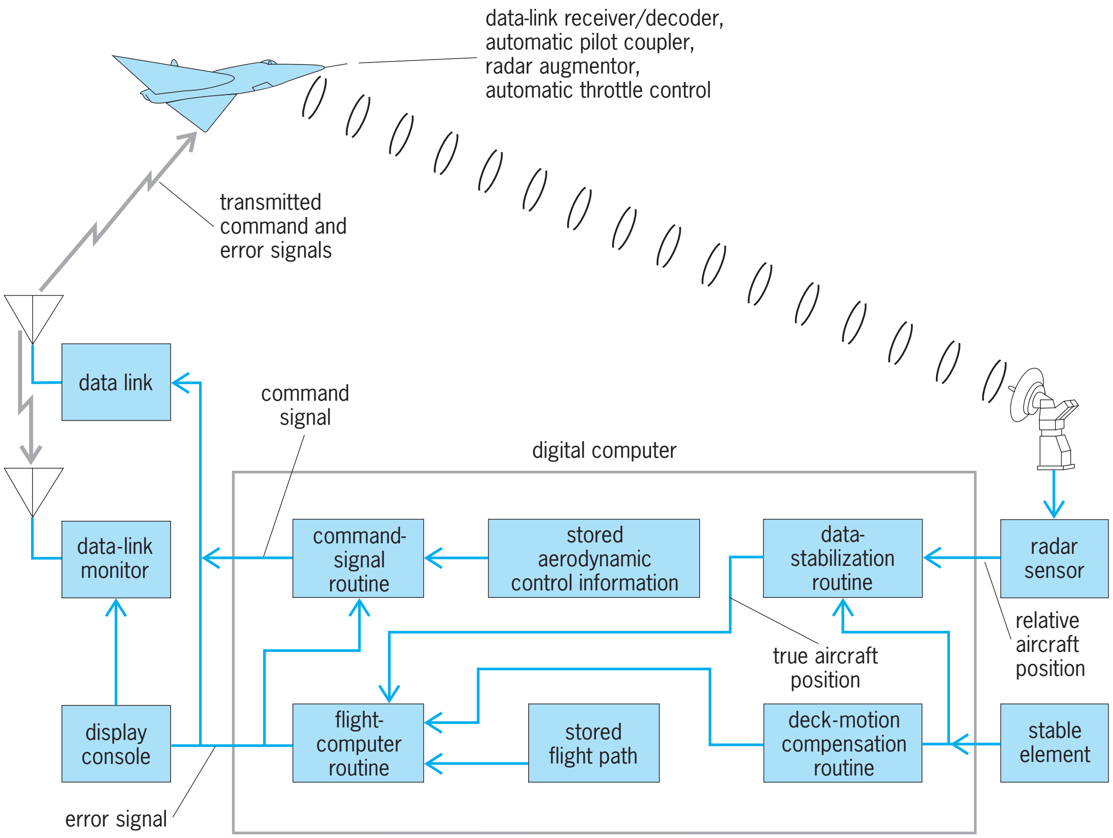
The means for guiding and controlling aircraft from an initial approach altitude to a point where safe contact is made with the landing surface. Such systems differ from low-approach systems in three major respects: (1) They furnish not only guidance but control of the aircraft as well. (2) They furnish information on the aircraft's position with respect to the terrain, and the rate at which the landing surface is being approached. (3) They do not require the pilot to assume manual control near the ground.

Two automatic landing systems have been developed. One is a radar-beam type that was developed to assist pilots landing on an aircraft carrier. The landing guidance is derived on the ship and is communicated to the aircraft. This ground-derived landing system detects the position and rate of change in position of the landing aircraft by means of a radar beam emitted from a ground derived-control (aircraft carrier) complex. The other is a variant of an instrument landing system (ILS) in which the landing solution is derived on the aircraft. The aircraft derives position and rate of change in position by using its onboard instrumentation in addition to radio signals transmitted by ILS-type equipment on the ground. The aircraft instrumentation includes accelerometers (which may be part of an inertial navigation system) and a radio altimeter. Essential to both systems is an autopilot in the aircraft, commanded by a computer on the ground in the radar-beam system and by a computer in the aircraft for the ILS-based system. The radar-beam and ILS-based systems are described below. *See* INSTRUMENT LANDING SYSTEM (ILS).

Ground-derived landing system. The ground-derived system was designed to place the burden of space, weight, and complexity on the aircraft carrier, or at the air station in the case of shore application. The carrier landing system is composed of a K_a-band fire-control-type radar, a stable element (mechanized by inertial instrumentation), a high-speed general-purpose digital computer, a display console, flight-path recorder, data-link coder/transmitter, and data-link monitor on the carrier. In the aircraft a data-link receiver/decoder, automatic-pilot coupler, automatic pilot, and radar augmentor are required. Automatic throttle control is also required in the aircraft, although not directly connected to the automatic landing system. The general system layout is shown in the **illustration**.

An aircraft is flown through a prescribed area, called a gate, where the radar locks on and tracks the aircraft to the touchdown. The angular tracking information of the radar is fed from digital encoders to the computer. The digitally encoded information from a stable element that tracks the ship motion is also fed to the computer. A data stabilization routine in the digital computer removes the effects of ship's roll, pitch, yaw, and heave from the angular tracking data. The corrected data, which precisely locate the aircraft in space, are fed to the flight computer routine in the digital computer, where they are compared with a stored flight path optimized for the type of aircraft under control. Error signals are generated which are transmitted to the aircraft by means of a ground-to-air data link and displayed to the pilot on a cross-pointer indicator or heads-up display. Also stored in the digital computer is sufficient aerodynamic control information for each aircraft type to convert error signals to commands to the aircraft autopilot, taking into account the response characteristics for different aircraft types. Autopilot commands are transmitted to the aircraft by means of a ground-to-air data link in the same message used to transmit error signals. To limit the dispersion in heavy seas, deck motion compensation is used during the last 12 s to fly the aircraft to the vertical position of the carrier deck touchdown point at impact.

Use of two systems. In practice, two complete systems are used on each carrier to double the landing rate and to provide a high systems availability in the event of component failure. Each system uses a conical scan radar antenna, with a choice of


data-link receiver/decoder,
automatic pilot coupler,
radar augmentor,
automatic throttle control
transmitted command and error signals
data link
command signal
digital computer
data-link monitor
command-signal routine
stored aerodynamic control information
data-stabilization routine
radar sensor
relative aircraft position
true aircraft position
display console
flight-computer routine
stored flight path
deck-motion compensation routine
stable element
error signal

Schematic of radar-beam system for aircraft-carrier and shore applications.

vertical or circular polarization. A 4-ft (1.2-m) reflector provides a beam width of 0.5°. The radar operates in the 33.0–33.4-GHz frequency band, with a pulse repetition rate of 2000 pulses per second and a peak power output of 40 kW. A radar augmentor (radar beacon or corner reflector) eliminates target scintillation and aids in penetration of heavy rainfall. *See* RADAR.

The digital computer has an 18-bit-word and 32,000-word memory. The memory recycling time is 2 microseconds. Each computer performs all computations for two approaching aircraft at a rate of 20 times per second. At the same time the computer performs on-line diagnostics throughout both systems. The second computer serves as an off-line monitor. Addressed digital data-link messages containing error and command signals are transmitted to each aircraft at the rate of 10 times per second. In addition, discrete signals are transmitted to indicate to the pilot such things as radar lock-on, 10 s to touchdown, and wave-off.

A display console is provided to monitor automatic approaches. The display provides an AZ-EL presentation which permits the console operator to "talk down" the aircraft, as in ground-controlled-approach (GCA) systems. Talk-down can be provided if the aircraft is not equipped with a data link or if there is a failure in the data link, coupler, or automatic pilot. Flight path, airspeed, rate of descent, and ship's motion and impact velocity are recorded for each landing, and may be used for debriefing pilots when manual landings are made for training purposes. *See* PRECISION APPROACH RADAR (PAR).

The landing gate may be placed 2–8 mi (3–13 km) from touchdown. Assuming a landing speed of 120 knots (60 m/s), the landing rate for aircraft may vary from 1 to 4 min per radar. Provision of two radars permits the landing interval aboard a carrier to be varied from $^1/_2$ to 2 min. For carrier operations, 68% of the landings must be within ±20-ft (6-m) longitudinal and ±10-ft (3-m) lateral dispersion when there is no deck motion. For heavy seas, longitudinal dispersion can be degraded to ±40 ft (12 m).

Air-derived landing system. The air-derived system is designed to take advantage of the large number of ILS installations in worldwide use under the international standards and recommended practices of the International Civil Aviation Organization. Its use, largely in Europe, is confined to the larger multiengine aircraft, because of the size and complexity of the airborne installation. Part of this complexity is necessary to reduce the effect of electrical noise on the localizer and glide-slope beam signals caused by multipath. For example, an aircraft taking off over the localizer antenna may cause wide perturbation of the beam in an approaching aircraft for more than a second. A time-referenced scanning-beam microwave landing system (MLS) is under development and is expected to significantly reduce problems due to multipath.

In the ILS-based system the position of the localizer and glide-slope beams is detected in the aircraft. The autopilot is commanded to seek the ILS fixed-beam centers. This action will describe the runway centerline extension and a fixed glide slope of about 3°. At a flare-initiate height determined by the radio altimeter (about 50 ft or 15 m), the autopilot is commanded to execute a preprogrammed flare to reduce the rate of descent from as much as 10 ft/s (3 m/s) to 1.5–2 ft/s (0.4–0.6 m/s) at touchdown. An automatic throttle control is used (as in the radar-beam system) to maintain the pilot-selected airspeed during changes in the aircraft pitch and bank.

Requirements for the airborne digital computer are similar to those indicated for the ground digital computer of the radar-beam system. The radar sensor is replaced by the ILS receiver, and the stable element by an aircraft inertial navigation system or accelerometers. Measurement of vertical and lateral accelerations is required. The best estimate of aircraft position is determined in the computer by filtering and smoothing the ILS receiver horizontal and vertical inputs. In performing filtering and smoothing, true-motion inputs as determined by the accelerometers are used by the computer to reduce multipath and noise peaks from the ILS receiver inputs. A filter and smoothing routine replaces the data stabilization routine. Rate of change in position is based upon changes in true position after filtering and smoothing. The deck-motion routine of the radar-beam system is not required ashore and is replaced by a flare-path computer routine. During flare, glide-slope information from the ILS receiver is not used. Starting at the flare-initiate height, vertical accelerations are altered to provide a cumulative vertical deceleration along an exponential curve to touchdown. Lateral guidance is still referenced to the ILS localizer input signal. An additional computer routine is required to compensate for crosswinds experienced on the landing path, whereas no crosswind compensation is provided for landing aboard an aircraft carrier, because the carrier is headed into the wind during landing operations.

A pilot normally "crabs" (steers the aircraft toward the wind) to compensate for crosswind. This operation has limitations when used for automatic landings on land. First, the runway is not directly ahead in the pilot's field of vision and is more difficult to locate when trying to establish visual contact with the ground. Also, the aircraft response to rudder control, required to perform de-crab, is marginal to inadequate. This response is particularly poor for aircraft that land with a high-angle-of-attack (nose-up) attitude. ILS-based automatic landing systems use a forward slip maneuver to compensate for the crosswind. In this maneuver the computer controls the ailerons to bank the aircraft while maintaining the aircraft heading along the landing path. This form of compensation keeps the pilot's head pointed to the runway and takes advantage of the more rapid aircraft response to aileron control for removing the wing-down before touchdown. To eliminate a possibility of having a wing contact the ground before the wheels, the amount of wing-down is restricted; hence the forward slip can compensate only for a limited crosswind. If the restricted forward slip is not sufficient to compensate for the crosswind (perhaps above 30 knots or 15 m/s with a wide-body aircraft), crab is added to the forward slip. This is a very unusual condition and crab is held to a minimum. The removal of forward slip and any de-crab must occur very close to the ground to ensure that the aircraft is not blown too far off centerline and that aircraft weight is evenly distributed on the landing gear at touchdown. Crosswind compensation is a complex computer routine and must be carefully tailored to the flight characteristics of each aircraft type.

Since each aircraft computer serves only one aircraft, its computation and memory requirements are about half of those required for the ground computer. The command signal routine directly interfaces the autopilot coupler. Aircraft displays are energized by ILS receiver outputs. *See* ACCELEROMETER; INERTIAL GUIDANCE SYSTEM.

To provide the degree of safety required for airline operations, the ILS-based automatic system uses triple-redundant control systems in the aircraft. This requirement is due in part to the lack of an independent landing monitor for the pilot or on the ground. The use of triple redundancy increases the complexity, size, weight, and cost of the airborne ILS-based system. In addition, civil systems are subject to a very rigorous certification process that further improves their reliability. The ground-based system does permit ground monitoring as a part of the system but not independently. Both classes of systems can provide adaptive computer programs to compensate for reduced aircraft or equipment performance. John L. Loeb; Douglas Hodgkins

Bibliography. T. A. Bombard et. al., Future DOD precision approach and landing architecture, *Proc. ION Annual Meeting*, *June 1998*, pp. 627–636, 1998; M. Kayton and W. Fried, *Avionics Navigation Systems*, 2d ed., 1995; R. J. Kelly and E. F. C. LaBerge, MLS: A total system approach, *IEEE AES*, 5(5):27–39, May 1990; B. R. Peterson, G. Johnson, and J. Stevens, Feasible architectures for Joint Precision Approach and Landing System (JPALS) for land and sea, *Proc. ION GNSS Meeting*, *September 2004*, pp. 544–554, 2000.

Automatic sprinkler system

An integrated arrangement of fixed facilities for protection from combustion by use of water extinguishment. The system comprises an adequate water supply, hydraulically designed internal piping, and sprinklers connected in a systematic pattern over the protected area; the system is activated by a fire to discharge a fine spray of water over the heat.

Essential features of a system are its self-detection of fire, prior installation, and built-in activation. In these respects, the automatic sprinkler system is among the earliest-used architectural features that contribute actively to maintenance of the internal environment (in contrast to the passive fire-resistant contribution of the static structure). Auxiliary to an automatic sprinkler system may be a fire alarm. *See* FIRE DETECTOR.

Effectiveness. Amenable to all occupancy classes of the National Building Code, automatic sprinklers are the most widely used fixed apparatus for fire protection. In extinguishment of class A fires and, in adapted system forms, of class B and class C fires, the water spray acts four ways: (1) it cools burning material by conversion of water to steam; (2) the steam displaces the oxygen supply, thereby tending to smother the fire; (3) by dampening materials in the area the spray limits the supply of new fuel; and (4) while the spray is falling, it lowers the temperature in the vicinity by evaporative cooling.

Through prompt response, an automatic sprinkler system generally requires less water to control a fire than does a hose. Also, the vertical spray produces less mechanical damage than does a horizontal hose stream. A sprinkler system is considered effective if it extinguishes or checks a fire until fire-fighting forces arrive. Failures of such systems are principally where they have been made inoperative during building alteration or disuse, or the occupancy hazard has been increased beyond initial system capability.

Not all hazards are accessible to sprinkler protection. For example, closely packed stock piled over two stories high sheds water so that the bottoms of these piles may not be wetted from overhead sprinklers. Thus, management of materials handling and storage and of housekeeping is necessary to stay within the capabilities of the installed sprinklers.

Elements of a system. The water supply for a sprinkler system is separate from that used by a fire department. Normally no water flows in the supply lines to sprinklers, thus freezing is a greater risk than in mains with continuous flow. Standards require sprinkler mains to be buried well below the frost line. The underground main enters the building in a heated area to supply a riser.

Connected at the riser are valves, meters, and often an alarm to sound when flow exceeds discharge from one sprinkler. At the top of this vertical riser, a horizontal network of pipes extends throughout the fire compartment in the building. Other risers feed distribution networks to systems in adjacent fire compartments. Compartmentalization divides a large building horizontally, on a single floor, and vertically, floor to floor. Thus, several sprinkler systems may serve one building.

In the distribution network, branch lines carry the sprinklers. A sprinkler may extend up from a branch line (upright configuration), placing the sprinkler close to the ceiling, or a sprinkler can be below the branch line (pendant configuration). For use with concealed piping, a flush-mounted pendant sprinkler extends only slightly below the ceiling.

For a high-rise building—one in which a fire must be fought internally because of height—hydraulic design must position a water source to supply sprinklers for not less than 30 min. The flow-actuated alarm produces a location indication at a ground-level register accessible to the fire department.

Sprinklers. The principal component of the system is a thermally sensitive sprinkler with a linkage assembly that holds closed the discharge opening. In various designs the assembly is disrupted through a low-melting-point chemical, a frangible bulb filled with liquid, a bimetallic disk, or—usually—a low-melting-point alloy link. The linkage separates above the operating range, which may be any one of the six standard steps from 100°F (38°C) to 475°F (246°C). Then the sprinkler abruptly opens to discharge water against a deflector so that water falls in a hemispherical spray across the area below. Where water damage ranks in importance with fire damage, stop-and-go sprinklers with bimetallic heat sensors open in response to heat and close automatically when heat subsides; they recycle if necessary to track the fire cycle.

Clearance for the spray to spread is maintained between the top of storage, such as library stacks, and sprinkler deflectors. If clearance cannot be maintained, branch lines follow aisles (alternate aisles in the absence of vertical back-to-back shelf dividers). If ventilation extends through decks in a library, sprinkled aisles are staggered vertically.

Modes of operation. Water reaches the sprinklers variously in five basic setups. In the usual wet-pipe system, for heated buildings, all pipes contain water under pressure for immediate release through any sprinkler that opens.

In a dry-pipe system—for unheated buildings in freezing climates or for cold-storage rooms—branch lines and other distribution pipes contain dry air or nitrogen under pressure. This pressure holds closed a dry-pipe valve at the riser. When heat from a fire opens a sprinkler, the air escapes and the dry-pipe valve trips; water enters branch lines and fire suppression begins.

For occupancies where flash fires are possible, a deluge system is appropriate. All sprinklers are continuously open while the pipes are empty. Separate fire detectors throughout the sprinkled compartment are of a type appropriate to the occupancy hazard. In response to combustion, the detectors actuate a valve at the riser. All sprinklers then deluge the entire area, both to suppress the fire and to wet down unburned combustibles.

In a preaction system, both heat-sensitive sprinklers and separate detectors guard the area; the separate detectors open a preaction valve and sound an alarm in the event of fire. The alarm provides opportunity for local hand control of the fire; otherwise heat fuses the links of local sprinklers, designed to operate above the temperature of the auxiliary detectors. These self-selected sprinklers then discharge.

Another adaptation is the recycling preaction system, applicable to unattended buildings. For example: heat from a fire actuates an auxiliary detector at 140°F (60°C). The detector in turn operates a master valve at the riser to admit water to branch lines. If the temperature continues to rise, at 160°F (71°C), to which these sprinklers are fused, one or more sprinklers respond. After these sprinklers bring the fire under control, the temperature drops and the auxiliary detector closes the riser valve. The water flow recycles if necessary to follow the fire cycle.

For overall building protection, a water sprinkler system may serve as backup to a first-response extinguishment facility that protects fire loads not directly extinguishable by water spray. *See* FIRE EXTINGUISHER; FIRE TECHNOLOGY. Frank H. Rockett

Bibliography. W. K. Bare, *Fundamentals of Fire Protection*, 1977; E. G. Butcher and A. C. Parnell, *Smoke Control in Fire Safety Design*, 1978; M. D. Egan, *Concepts in Building Fire Safety*, 1978; G. P. McKinnon (ed.), *Fire Protection Handbook*, 18th ed., 1997; J. Morris, *Managing the Library Fire Risk*, 1979.

Automation

The process of having a machine or machines accomplish tasks hitherto performed wholly or partly by humans. As used here, a machine refers to any inanimate electromechanical device such as a robot or computer. As a technology, automation can be applied to almost any human endeavor, from manufacturing to clerical and administrative tasks. An example of automation is the heating and air-conditioning system in the modern household. After initial programming by the occupant, these systems keep the house at a constant desired temperature regardless of the conditions outside.

Reasons for automating. The complete or partial automation of any task is usually based upon the following considerations:

1. *Productivity*. Automating a task usually results in an increment in productivity, where productivity is defined as tasks performed per unit time.

2. *Efficiency*. Automation has the potential to increase the efficiency in completing a task, where efficiency is defined as inversely proportional to the amount of raw material (for example, electrical power) consumed.

3. *Quality*. Automating a task can significantly increase the quality of a final product or process. Quality can be defined in many ways with many attributes. An obvious definition is simply the degree of excellence which a thing possesses. Quality could refer to the consistency with which a product is manufactured to meet that degree.

4. *Safety*. By removing direct human intervention or activity, automation has the potential to significantly increase the safety of any task. This is particularly true in the area of manufacturing.

5. *Flexibility*. Automation can significantly enhance the flexibility of a process, where flexibility refers to the ability to alter the number or nature of the tasks completed per unit time.

6. *Cost*. While each of the previous items can be related directly or indirectly to the cost of completing tasks, cost is often the primary motivation behind automating tasks previously undertaken partially or completely by humans. Chief among these cost savings is that associated with labor, including salaries, benefits, training, and so forth.

Enabling technologies. Historically, advances in automation have depended upon advances in the requisite enabling technologies, such as digital computation. The fundamental constituents of any automated process are (1) a power source, (2) a feedback control mechanism, and (3) a programmable command structure. Programmability does not necessarily imply an electronic computer. For example, the Jacquard loom, developed at the beginning of the nineteenth century, used metal plates with holes to control the weaving process. Nonetheless, the advent of World War II and the advances made in electronic computation and feedback have certainly contributed to the growth of automation. While feedback is usually associated with more advanced forms of automation, so-called open-loop automated tasks are possible. Here, the automated process proceeds without any direct and continuous assessment of the effect of the automated activity. For example, an automated car wash typically completes its task with no continuous or final assessment of the cleanliness of the automobile. *See* CONTROL SYSTEMS; DIGITAL COMPUTER.

Categorization of automated tasks. Because of the growing ubiquity of automation, any categorization of automated tasks and processes is incomplete. Nonetheless, such a categorization can be attempted by recognizing two distinct groups, automated manufacturing and automated information processing and control. Automated manufacturing includes automated machine tools, assembly lines, robotic assembly machines, automated storage-retrieval systems, integrated computer-aided design and computer-aided manufacturing (CAD/CAM), automatic inspection and testing, and automated agricultural equipment (used, for example, in crop harvesting). Automated information processing and control includes automatic order processing, word processing and text editing, automatic data processing, automatic flight control, automatic automobile cruise control, automatic airline reservation systems, automatic mail sorting machines, automated planet exploration (for example, the rover vehicle, *Sojourner*, on the *Mars Pathfinder* mission), automated electric utility distribution systems, and automated bank teller machines. *See* ASSEMBLY MACHINES; COMPUTER-AIDED DESIGN AND MANUFACTURING; COMPUTER-INTEGRATED MANUFACTURING; FLEXIBLE MANUFACTURING SYSTEM; INSPECTION AND TESTING; SPACE PROBE; WORD PROCESSING.

Elements of automation process. The relationships of the key elements in an automated task or process, particularly the three fundamental constituents outlined above, can be shown in a schematic diagram (**Fig. 1**). Nearly every automated task can be interpreted in terms of this diagram. If no feedback is associated with the task, that is, if it is open-loop in nature, then the feedback element (2) is simply omitted. The increasing pace of automation is attributable to the increased capabilities of digital computers, which not only provide the programmable command but also form the feedback element when implemented as a digital control system. The precise form and function of these elements depend on the particular automation process, as in the following examples. *See* DIGITAL CONTROL.

Robotic assembly machine. A robotic assembly machine includes the robot itself, a conveyor, an assembly pad, a manipulator, a camera, and associated signal processing equipment (**Fig. 2**). Guided by information from the camera, the robot takes parts from the conveyor, assembles them, and places the assembled parts back on the conveyor. The power source is assumed here to be electrical. The command programming serves as one input to the robot controller. The feedback activity is exemplified by the information received from the camera and fed back to the controller. If a human assembler were completing this task, activities paralleling those just described would occur. That is, the worker's vision system would replace the camera, his or her limbs and hands would replace the robot arm and manipulator, and a feedback loop would be effectively utilized between visual inputs and motion. *See* ROBOTICS.

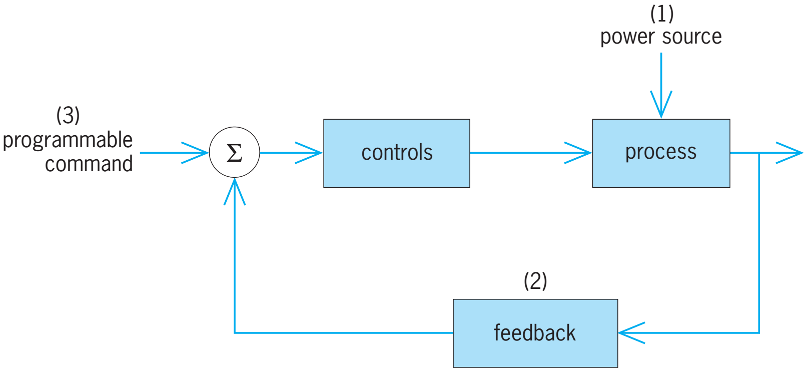


Fig. 1. Elements of an automated system.

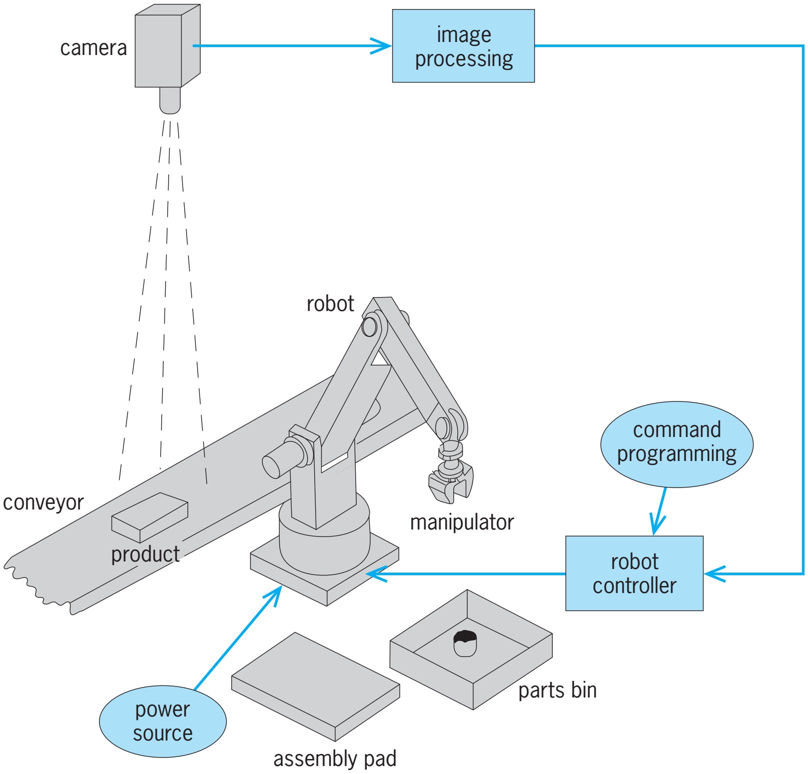


Fig. 2. Simplified representation of a robotic assembly machine.

Automatic flight control system. The constituent elements of an automatic flight control system include the pilot, aircraft, flight control computer, control surface actuator, and control surface (**Fig. 3**). The pilot can be replaced by a partially automated

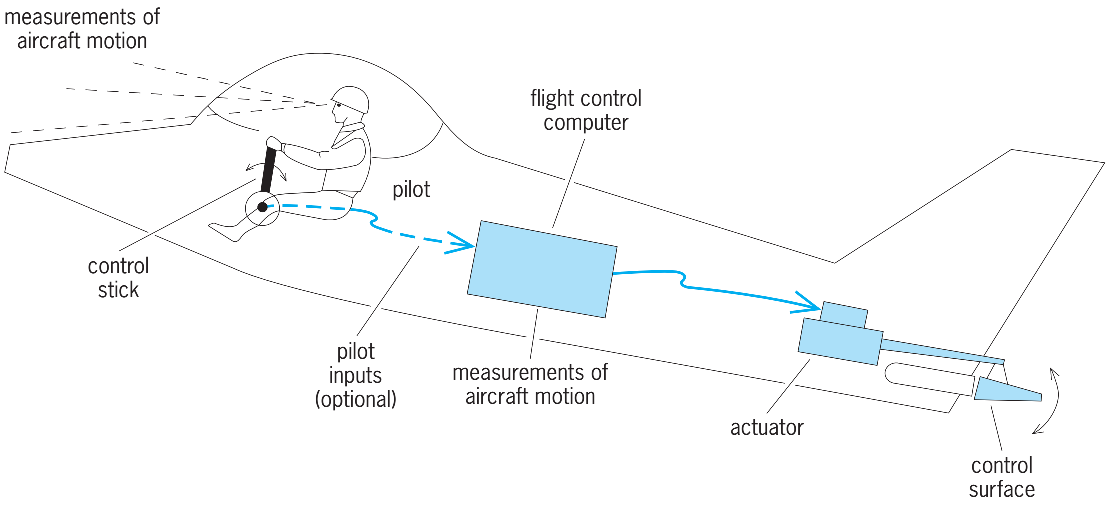


Fig. 3. Simplified representation of an automatic flight control system.

system in which pilot inputs are still possible. The flight control computer receives and processes information from onboard systems (possibly including the pilot), which sense aircraft angular and linear motion (for example, rate of change of pitch-attitude). The processed information is then used to drive the actuator (a hydraulic or electrohydraulic device), which moves the control surface to change the attitude of the aircraft. The power source (not explicitly shown in Fig. 3) would likely be electrical. The command programming is contained in the flight control law, which is coded into the flight control computer. The controller activity is also carried out in the flight control computer, and the feedback activity occurs when the aircraft motion is measured (by inanimate sensors or the pilot) and fed back to the computer. *See* FLIGHT CONTROLS.

Effects on human activity. One obvious societal implication of increasing automation is the elimination and displacement of human labor. While there are positive aspects of such displacement, for example, the elimination of dangerous tasks which required human participation, the negative aspects cannot be overlooked. For example, as automated manufacturing becomes more prevalent, the size of the human work force that supports such enterprises naturally shrinks and the skill requirements of the new work force increase considerably. For example, the type of human activity in an automated factory is significantly different from that in a nonautomated one and can be summarized as follows:

1. *Equipment maintenance*. In an automated manufacturing facility, equipment maintenance becomes a critical issue. Malfunctions or breakdowns in one part of the facility can halt the entire production line.

2. *Computer programming*. Since the digital, programmable computer forms the backbone of the automated manufacturing system, programming the computers, entering pertinent product and process information, and updating manufacturing databases constitute an integral part of the automated factory. *See* COMPUTER PROGRAMMING.

3. *Updating production technologies*. Economic competitiveness requires nearly constant improvement in the production systems in automated manufacturing facilities. This entails equipment redesign, replacement, and upgrades.

4. *Production supervision*. Managing automated manufacturing facilities requires personnel with skills considerably different than those employed in nonautomated plants. Greater emphasis is placed upon the technical skills of the automated plant supervisor as compared to those of the supervisor concerned primarily with personnel management.

The problems and opportunities just outlined are not limited to automated manufacturing. For example, the introduction of automated systems into airline cockpits has reduced the required crew from three to two (with the elimination of the flight engineer). In addition, the required skill set of the airline crew now includes cockpit resource management, which describes the ability of the crew to effectively utilize the capabilities of the modern cockpit, for example, the computerized flight management system. *See* AIRCRAFT INSTRUMENTATION.

Human-machine interaction. While the ability to automate manufacturing and information/control tasks is enhanced by advances in computer technology, the effective application of automation remains a challenge, particularly as it involves tasks and processes requiring the automated system to act in consort with humans. Before the advent of automation, the human often assumed the roles associated with controls and feedback in an automated system (Fig. 1). In the aircraft piloting example, and prior to the introduction of automation, it was the human pilot's visual and vestibular (motion) sensing which provided the feedback elements, and it was the pilot's operation upon the difference between these sensed variables and a command or desired state that resulted in an input to the aircraft through the control stick or throttle. As cockpit automation has progressed, however, the responsibilities for feedback and control have increasingly been assumed by the aircraft automatic flight control system. Human

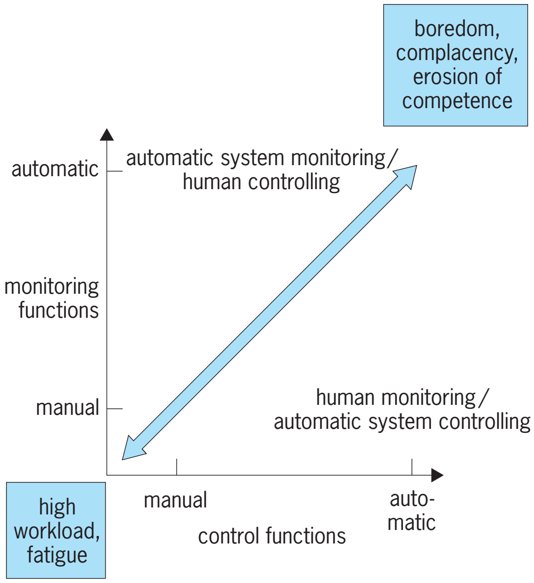

boredom, complacency, erosion of competence
automatic
automatic system monitoring/ human controlling
monitoring functions
manual
human monitoring/ automatic system controlling
high workload, fatigue
manual
control functions
auto-matic

Fig. 4. Trade-off between manual (human) and automatic (machine) monitoring and control.

presence in the cockpit is still required from the standpoint of safety and passenger confidence. The crew must continuously monitor the performance of the automatic flight control system in order to reassume an active feedback and control role should the automatic system fail or situations arise which are beyond the capabilities of this system. In this context, the term "appropriate automation" describes the optimal distribution of responsibility between human and machine in tasks such as airline flight. The effects of this distribution of responsibility between human (pilot) and machine (automatic flight control system) can be demonstrated by means of a simple graph (**Fig. 4**): In the lower left (no automation) human workload is high and fatigue is a concern, while in the upper right (complete automation) human boredom and possible erosion of competence and piloting skills can occur.

In addition to workload and skill atrophy, the successful utilization of a completely or partially automated system depends upon the ease with which the human can communicate with the system. This communication can take the form of human supervision of the automated system, human "training" of the automated system, or humans' reasserting direct control of the task being performed by the automated system. The importance of this concern has led to the concept of human-centered automation, that is, an automation philosophy which considers from the outset that humans will, of necessity, be required to interact with the system. Such a discussion leads to the topic of function allocation between human and machine. *See* HUMAN-MACHINE SYSTEMS.

Function allocation. A major issue in the design of systems involving both human and automated machines concerns allocating functions between the two, that is, answering the question, "What should the (automated) machine do and what should the human do?" This allocation can be static or dynamic. Static allocation is fixed; that is, the separation of responsibilities between human and machine do not change with time. Dynamic allocation implies that the functions allocated to human and machine are subject to change. Historically, static allocation began with reference to lists of activities which summarized the relative advantages of humans and machines with respect to a variety of activities. For example, at present humans appear to surpass machines in the ability to reason inductively, that is, to proceed from the particular to the general. Machines, however, surpass humans in the ability to handle complex operations and to do many different things at once, that is, to engage in parallel processing. Dynamic function allocation can be envisioned as operating through a formulation which continuously determines which agent (human or machine) is free to attend to a particular task or function. In addition, constraints such as the workload implied by the human attending to the task as opposed to the machine can be considered. *See* HUMAN-FACTORS ENGINEERING.

Artificial intelligence. It has long been the goal in the area of automation to create systems which could react to unforeseen events with reasoning and problem-solving abilities akin to those of an experienced human, that is, to exhibit artificial intelligence. Indeed, the study of artificial intelligence is devoted to developing computer programs that can mimic the product of intelligent human problem solving, perception, and thought. For example, such a system could be envisioned to perform much like a human copilot in airline operations, communicating with the pilot via voice input and spoken output, assuming cockpit duties when and where assigned, and relieving the pilot of many duties. Indeed, such an automated system has been studied and named a pilot's associate. Machines exhibiting artificial intelligence obviously render the sharp demarcation between functions better performed by humans than by machines somewhat moot. While the early promise of artificial intelligence has not been fully realized in practice, certain applications in more restrictive domains have been highly successful. These include the use of expert systems, which mimic the activity of human experts in limited domains, such as diagnosis of infectious diseases or providing guidance for oil exploration and drilling. Expert systems generally operate by (1) replacing human activity entirely, (2) providing advice or decision support, or (3) training a novice human in a particular field. *See* ARTIFICIAL INTELLIGENCE; EXPERT SYSTEMS.

Ronald A. Hess

Bibliography. M. P. Groover, Automation, in R. C. Dorf and A. Kusiak (eds.), *Handbook of Design, Manufacturing and Automation*, Wiley, 1994; B. B. Morgan, Jr., et al., Implications of automation technology for aircrew coordination and performance, in W. B. Rouse (ed.), *Human/Technology Interaction in Complex Systems*, JAI Press, 1993; T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, 1992; C. D. Wickens, *Engineering Psychology and Human Performance*, 2d ed., HarperCollins, 1992.

Automobile

A self-propelled land vehicle, usually having four wheels and an internal combustion engine, used primarily for personal transportation (**Fig. 1**). Other types of motor vehicles include buses, which carry large numbers of commercial passengers, and medium- and heavy-duty trucks, which carry heavy or bulky loads of freight or other goods and materials. Instead of being carried on a truck, these loads may be placed on a semitrailer, and sometimes also a trailer, forming a tractor-trailer combination which is pulled by a truck tractor. *See* BUS; TRUCK.

Design. The basic design of the automobile was standardized by 1908, when the Ford Motor Company began production of the Model T. A metal frame served as the main structural member that supported the power train and body. The frame was supported through springs by four wheels and tires. The engine was mounted at the front of the vehicle and transmitted power to the two rear wheels. The body was open, providing little protection from the weather.

By the early 1930s, most automobiles had a closed body made of stamped steel parts that were welded together before attachment to the frame. Improved versions of the front-engine rear-wheel-drive power train dominated automotive production into the 1970s. Until then, this basic configuration was never seriously challenged, although some vehicles had either front-engine front-wheel drive or rear-engine rear-wheel drive.

In the United States, the change from rear drive to front drive was largely necessitated by three acts passed by Congress which established new laws covering automotive air pollution, automotive safety, and automotive fuel economy. The solutions to problems of automotive air pollution and safety were often add-on devices that allowed the conventional automobile to be manufactured in compliance with the new regulations. However, the additional weight and action of some of these devices reduced fuel economy.

When compliance with fuel-economy standards also became a design objective, automotive stylists and engineers began redesigning the car by reducing vehicle weight, downsizing new models, using smaller engines, and changing to lighter, stronger materials. The result was an evolutionary redesign that allowed continuing production while the required new parts and vehicles were developed and phased into each manufacturer's passenger-car fleet.

Automotive air pollution. Historically, the design and manufacture of automotive vehicles in the United States was an unregulated industry. In 1955, the Department of Health, Education, and Welfare was instructed to begin a study of air pollution. By 1962, the automobile engine had been identified as the source of more than 40% of airborne pollutants, making the automobile the largest single contributor to air pollution. *See* AIR POLLUTION.

In 1963, Congress passed the Clean Air Act which, together with its subsequent amendments, directed the Environmental Protection Agency (EPA) to promulgate automotive emission standards and to regulate automotive engine fuels and fuel additives. Because of unusual air-pollution problems in Los Angeles, the Clean Air Act permitted the state of California to set more stringent standards than the rest of the nation. *See* DIESEL FUEL; GASOLINE.

To comply with automotive emission standards, emission-control devices were developed to reduce or eliminate air pollutants from the engine crankcase,

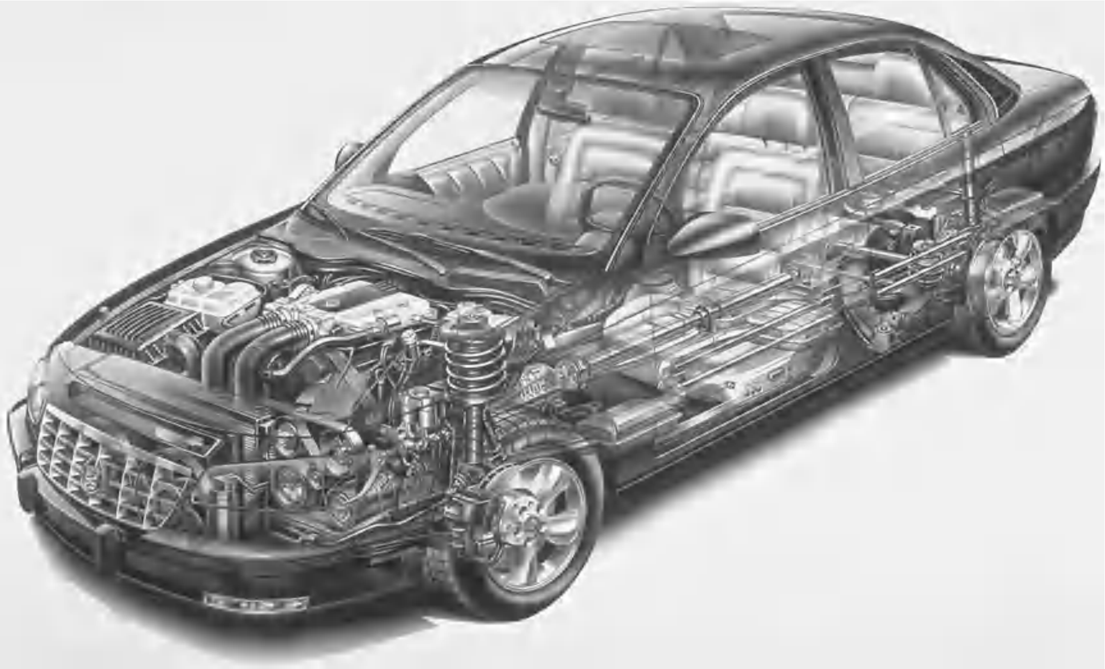


Fig. 1. Phantom view of a front-engine automobile with unitized-body construction. (*Cadillac Division, General Motors Corp.*)

engine exhaust gas, and fuel tank and carburetor. However, some devices reduced fuel economy as well as emissions, resulting in 1968–1974 automobiles having worse fuel economy than earlier models.

In 1973, however, the EPA moved toward requiring automobile engines to run on unleaded gasoline (tetraethyllead was formerly used as an additive for motor fuels to reduce engine knock). Beginning in 1975, a stepped phase-down limited the amount of tetraethyllead allowed in gasoline. Tetraethyllead impairs the operation of some engine emission-control devices and, after the additive becomes airborne in the engine exhaust gas, creates a human and environmental health hazard. *See* LEAD; TOXICOLOGY.

Automotive safety. During the early 1960s, traffic deaths in the United States approached 50,000 per year. Congress began passing legislation in an attempt to reverse the upward trend. The National Traffic and Motor Vehicle Safety Act (1966) required that all new motor vehicles sold in the United States be manufactured in compliance with Federal Motor Vehicle Safety Standards. The first 20 such standards were issued in 1967 and became mandatory on all 1968 models.

Automotive fuel economy. In 1973, Arab nations imposed a total ban on oil exports to the United States. This resulted in a shortage of gasoline and diesel fuel, and caused a consumer demand for smaller cars which were lighter and more fuel-efficient. As the limited amount of fuel that was available rapidly increased in price, the larger, more powerful rear-drive cars built by manufacturers in the United States were gradually replaced by smaller, more fuel-efficient cars imported from Europe and Japan.

To help reduce the dependence of the United States on imported oil, Congress passed the Energy Policy and Conservation Act (1975), which established corporate average fuel economy standards to be met by new cars produced by each automotive manufacturer. To essentially double the fuel economy of the nation's passenger-car fleet in 10 years, Congress established a 1985 corporate average fuel economy standard of 27.5 mi/gal (2 km/liter). This could be achieved by the manufacturers' producing smaller, lighter cars powered by smaller engines.

Weight was reduced by redesigning the basic vehicle structure, eliminating excess material, and continuing the change to lighter and stronger materials. Fuel economy was further improved by introducing more aerodynamic body styles, reducing rolling resistance, and reducing friction and power losses in engines and other components. To maximize these gains, most new car designs had front drive with a transversely mounted integrated power train in which the engine, transmission, final drive, and differential form a single unit.

By the 1980s, the downsized front-engine front-drive car had become the most fuel-efficient and widely produced automobile design. However, some data indicated that the new lighter and smaller car had higher personal injury claims and damage repair costs after a collision than the full-size car.

Manufacturing. In automobile manufacturing, production is the making of the thousands of parts, subsystems, and modular assemblies that comprise the automobile, while assembly is the fitting together of these components. Typical assembly of a passenger car begins when the steel pieces that form the side openings for mounting the doors are assembled and spot-welded together by robot welders. The side assemblies are then welded to the underbody, which consists of pieces of galvanized steel that have been welded together to form the floor pan, engine compartment, and luggage compartment. Welding helps prevent noise—squeaks, rattles, and vibration—that could be caused later by parts moving while the vehicle is on the road. Roof, quarter panels, doors, deck lid, and hood are also welded during fabrication so that a minimum of mechanical fasteners is required.

The body assembly is dipped in a phosphate bath to clean any debris from the surface and then submerged in an anticorrosion solution, such as zinc phosphate. Fillers and sealers are applied to body joints, seams, and cavities that might leak under stress and collect or admit moisture. Sound-deadening material and additional sealer and primer are applied where necessary. A chip-resistant urethane coating is sprayed onto the lower body to provide extra protection against stones flying up from the road. *See* METAL COATINGS.

Robot painters apply the main color, while workers paint spot areas on the body. The paint is sealed with a clear coating to produce a high gloss finish. The hard trim, which includes the instrument panel, steering column, weather stripping, and body glass, is installed. Then, after testing for water leakage, the soft trim is installed. This includes carpeting over a layer of sound-deadening material, seats, door pads, roof insulation, and upholstery. Adhesive bonding may be used, especially when joining various types of nonmetal parts.

The body moves along an assembly line that is above a conveyor line on which the engine and other chassis components are moving (**Fig. 2**). The chassis is the complete operable vehicle, but without the body. A manufacturer's designation for a specific chassis is often called a platform. Changing outside body panels and interiors allows different models to be built on the same platform. All models built on that platform have the same tread width, but wheelbase may vary.

As the engine, complete with transmission or transaxle and exhaust system, is prepared for installation, the fuel tank and bumpers may be fitted to the body. Final assembly occurs when the engine and chassis are raised into position, mounting bolts are installed, and the wheel-and-tire assemblies are attached. Wheel alignment is tested under simulated road conditions as the vehicle undergoes final inspection and drive-away for delivery to a dealer or distribution center.

Body. The automobile body is the assembly of sheet-metal, fiberglass, plastic, or composite-material panels together with windows, doors, seats, trim and upholstery, glass, and other parts that form

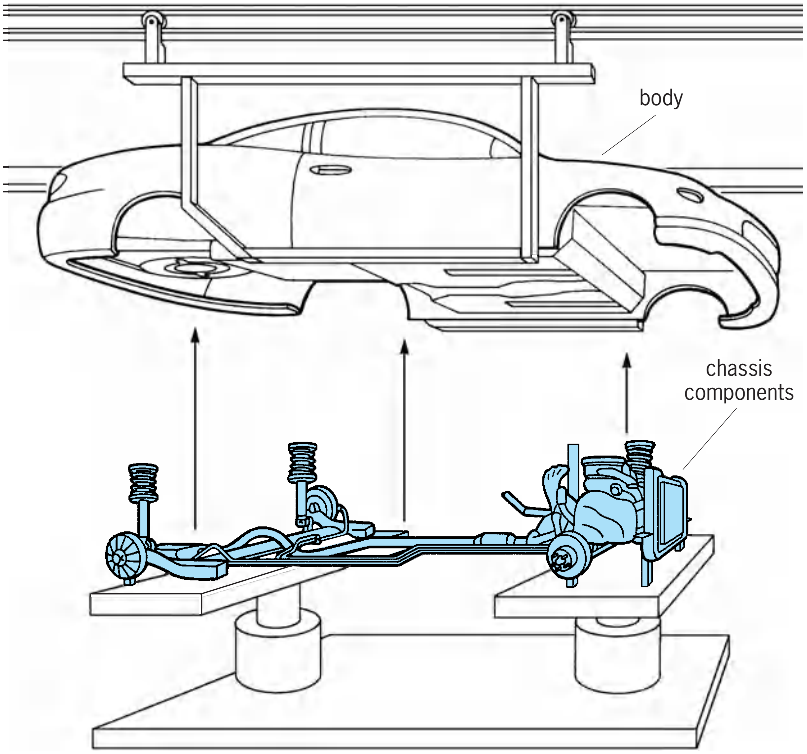


Fig. 2. **Final assembly of the automobile, as the engine and chassis are raised into position and fastened to the body. (*Chevrolet Division, General Motors Corp.*)**

enclosures for the passenger, engine, and luggage compartments. The assembled body structure may attach through rubber mounts to a separate or full frame (body-on-frame construction), or the body and frame may be integrated (unitized-body construction). In the latter method, the frame, body parts, and floor pan are welded together to form a single unit that has energy-absorbing front and rear structures, and anchors for the engine, suspension, steering, and power-train components. A third type of body construction is the space frame which is made of welded steel stampings. Similar to the tube chassis and roll cage combination used in race-car construction, non-load-carrying plastic outer panels fasten to the space frame to form the body. *See* COMPOSITE MATERIAL; SHEET-METAL FORMING; WELDED JOINT.

Size and style. Automobiles are built in several sizes and body styles and may be further classified according to the number of doors and the intended usage. Basic body styles include the hatchback, two-door coupe or sedan, roadster, convertible, four-door sedan, and station wagon. Passenger-car sizes range from the largest full-size car to the compact and small sizes. Light-duty trucks, which include pickups, suburbans, vans, minivans, and sport-utility vehicles, also are available in various sizes.

Many full-size cars are two- or four-door sedans that seat from four to six and have a separate luggage compartment. Some smaller cars are hatchbacks, which have two or four doors, no separate luggage compartment, and a rear lid that lifts up for access to the rear of the passenger compartment. For carrying cargo in a hatchback the available space may be enlarged by folding down the rear seat. The station wagon has two or three seats, and a cargo area that may be enlarged by folding down the rear seat in a two-seat wagon, and the center and rear seats in some three-seat wagons.

Roof options. Several roof options are available, including the vinyl roof, popup sunroof, electric slide sunroof, and T roof. The vinyl roof is a covering of colored vinyl that is laid over the finished roof of the car, primarily for decoration. The sunroof is a panel in the roof above the front seat that can be opened for ventilation; it may pop up, or it may slide open when moved by a hand crank or by an electric motor. A T roof has hatch panels above the driver and passenger that can be removed. Each panel fits between the top of the door glass and a center T bar, which runs from the front section of the roof to a point in back of the front seat. Convertibles are made with a soft fabric top and can be operated as open or closed cars. When the top is down, the convertible is basically a car without a roof.

Safety features. Automobile bodies include many safety features to help protect the occupants, such as the windshield, side windows, and rear window, which are made of laminated or tempered safety glass. The side doors have steel cross bars or beams to resist intrusion into the passenger compartment during side impacts. The body design may include an integral steel safety cage to surround and help protect the occupants, as well as crumple zones front and rear to absorb kinetic energy by deforming during an impact. Energy-absorbing front and rear bumpers, with standardized installed height, prevent or minimize vehicle damage resulting from a low-speed impact or collision. *See* SAFETY GLASS.

Automotive safety features are classified as helping to provide either crash avoidance or occupant protection. Crash-avoidance (active safety) features are often considered more beneficial because in a significant percentage of incidents they help the driver avoid an impending accident or crash. Examples are antilock-braking, traction-control, and vehicle stability control systems; headlights that turn on and off automatically as needed; and electrochromic rear-view mirrors that automatically dim to reduce glare. *See* AUTOMOTIVE BRAKE.

Occupant-protection (passive safety) features include the collapsible energy-absorbing steering column, head restraints, seat belts, air bags, and breakaway inside rear-view mirror. Some front-seat shoulder belts are motorized and move into position automatically. In some vehicles, the seat belt may automatically tighten in a crash to provide additional protection. Air bags are balloon-type supplemental restraints that inflate automatically to help protect the driver and front-seat passenger in a crash (**Fig. 3**). Rapid inflation of the air bag, usually by a pyrotechnic device, prevents the occupant from being thrown forward and injured by striking the steering wheel or windshield. The seat belt and air bag together provide maximum protection against injury in a collision. Additional air bags in the sides of the seatbacks or doors are also used to provide side-impact

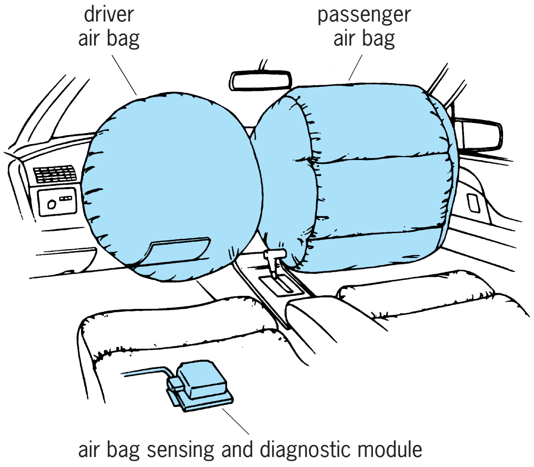


Fig. 3. Front air bags, which provide protection for the driver and front-seat passenger in a collision. (*Cadillac Division, General Motors Corp.*)

protection, primarily for the head and torso, of front- and rear-seat passengers. *See* PYROTECHNICS.

Comfort and convenience features. Comfort features include passenger-compartment ventilation air filter, heating and air conditioning, solar-control glass, electrically heated seats, and tilting and telescoping steering wheel. Convenience features include power windows, power sunroof, remote keyless entry, central locking systems, sound systems with tape players and disc players, voice-activated cellular telephone, lighted sun-visor vanity mirrors, and instrument-panel or console cup holders. *See* AUTOMOTIVE CLIMATE CONTROL; MOBILE RADIO.

Frame. This is the main structural member to which all other mechanical chassis parts and the body are assembled to make a complete vehicle. In older vehicle designs, the frame is a separate rigid structure made of heavy steel stampings welded together to resist the road loads imposed upon the vehicle. Newer passenger-car designs have the frame and body structure combined into an integral unit, or unitized body. The weight of the side rails and other structural members added in construction of the unitized body is less than the weight of the separate frame.

Short or stubby subframes and their assembled components attach to the side rails at the front and rear of the unitized body. The front subframe carries the engine, transmission or transaxle, lower front suspension, and other mechanical parts. The rear subframe, if used, carries the rear suspension and rear axle.

Suspension. The suspension supports the weight of the vehicle, absorbs road shocks, transmits brake-reaction forces, helps maintain traction between the tires and the road, and holds the wheels in alignment while allowing the driver to steer the vehicle over a wide range of speed and load conditions. The action of the suspension increases riding comfort, improves driving safety, and reduces strain on vehicle components, occupants, and cargo. The springs may be coil, leaf, torsion bar, or air. Most automotive vehicles have coil springs at the front and either coil or leaf springs at the rear. *See* AUTOMOTIVE SUSPENSION; SPRING (MACHINES).

The rear suspension includes the springs, control arms, shock absorbers or dampers, struts (if used), and other parts that act together to support the weight of the rear end and absorb road shocks. Similar parts and steering knuckles in the front suspension perform similar functions while also providing steering control.

To produce a safe, stable, and comfortable ride, a shock absorber is located at each wheel to dampen spring oscillations and wheel bounce. Some vehicles have electronic ride control, which allows the driver to electronically adjust the shock absorbers for a softer or firmer ride. An automatic setting enables the system controller to automatically change the firmness of each shock absorber to suit road conditions.

A stabilizer bar, which usually connects to the lower control arms, may be used with the front and sometimes the rear suspension. The stabilizer bar is a transverse-mounted spring-steel bar that helps reduce body roll when making turns and when driving on rough or uneven surfaces. With the bar ends attached to each side of the suspension and the center pivoting in bushings attached to the frame or underbody, the stabilizer bar resists being twisted when spring travel is unequal.

Steering. The steering system enables the driver to turn the front wheels left or right to control the direction of vehicle travel. The rotary motion of the steering wheel is changed to linear motion in the steering gear, which is located at the lower end of the steering shaft. The linear motion is transferred through the steering linkage to the steering knuckles, to which the front wheels are mounted. Steering systems are classed as either manual steering or power steering, with power assist provided hydraulically or by an electric motor.

The force that the driver exerts to turn the steering wheel is multiplied by the mechanical advantage of the steering gear, which usually is a rack-and-pinion or recirculating-ball type. Typical automotive steering ratios range from about 24:1 with manual steering to 14:1 with power steering. Most small cars have a rack-and-pinion steering gear; some full-size and older cars and many trucks have a recirculating-ball steering gear. Some vehicles have four-wheel steering that is either mechanically or electronically controlled. *See* ANTIFRICTION BEARING; AUTOMOTIVE STEERING.

Brake. A brake is a device that uses a controlled force to reduce the speed of or stop a moving vehicle, or to hold the vehicle stationary. The automobile has a friction brake at each wheel. When the brake is applied, a stationary surface moves into contact with a moving surface. The resistance to relative motion or rubbing action between the two surfaces slows the moving surface, which slows and stops the vehicle.

The four-wheel hydraulically operated brake system that the driver's foot controls by pressing and releasing the brake pedal during normal driving conditions is called the service brake. Automobiles and trucks also have a mechanical parking brake, operated by a separate foot pedal or hand lever, which usually acts only on the rear wheels.

Automotive service brakes are either drum- or disc-type, with many vehicles having front disc brakes and rear drum brakes. Some vehicles have disc brakes at all four wheels. In addition, most vehicles have power brakes, which use the power assist of a hydraulic or vacuum booster to increase the service-brake pedal force. Many vehicles have an antilock-braking system that prevents the brake at any wheel from locking, which would cause the tire to skid. *See* AUTOMOTIVE BRAKE; BRAKE.

Engine. The engine supplies the power to move the vehicle. The power is available from the engine crankshaft after a fuel, usually gasoline, is burned in the engine cylinders. Most automotive engines are located at the front of the vehicle and drive either the rear wheels or the front wheels through a drive train or power train made up of gears, shafts, and other mechanical and hydraulic components. In some vehicles, the engine is located at the rear of the vehicle, or in back of the front seat in what is known as a mid-engine location. In vehicles with front engine and rear-wheel drive, the engine is positioned longitudinally; in vehicles with front engine and front-wheel drive, the engine is usually positioned crosswise or transversely, but may be installed longitudinally.

Most automotive vehicles are powered by a spark-ignition four-stroke-cycle internal combustion engine. The inline four-cylinder engine and V-type six-cylinder engine are the most widely used, with V-8 engines also common. Other automotive engines have three, five, ten, and twelve cylinders. Some passenger cars and trucks have diesel engines. Some automotive spark-ignition and diesel engines are equipped with a supercharger or turbocharger. *See* AUTOMOTIVE ENGINE; DIESEL ENGINE; ENGINE; IGNITION SYSTEM; INTERNAL COMBUSTION ENGINE; SUPERCHARGER; TURBOCHARGER.

To improve fuel economy, and to reduce air pollution from the engine exhaust gas and from escaping fuel vapors, most automotive engines have electronic fuel injection instead of a carburetor. A computer-controlled electronic engine control system automatically manages various emissions devices and numerous functions of engine operation, including the fuel injection and spark timing (**Fig. 4**). This allows optimizing power and fuel economy while minimizing exhaust emissions. *See* CARBURETOR; CATALYTIC CONVERTER; CONTROL SYSTEMS; FUEL INJECTION.

Power train. The power available from the engine crankshaft to do work is transmitted to the drive

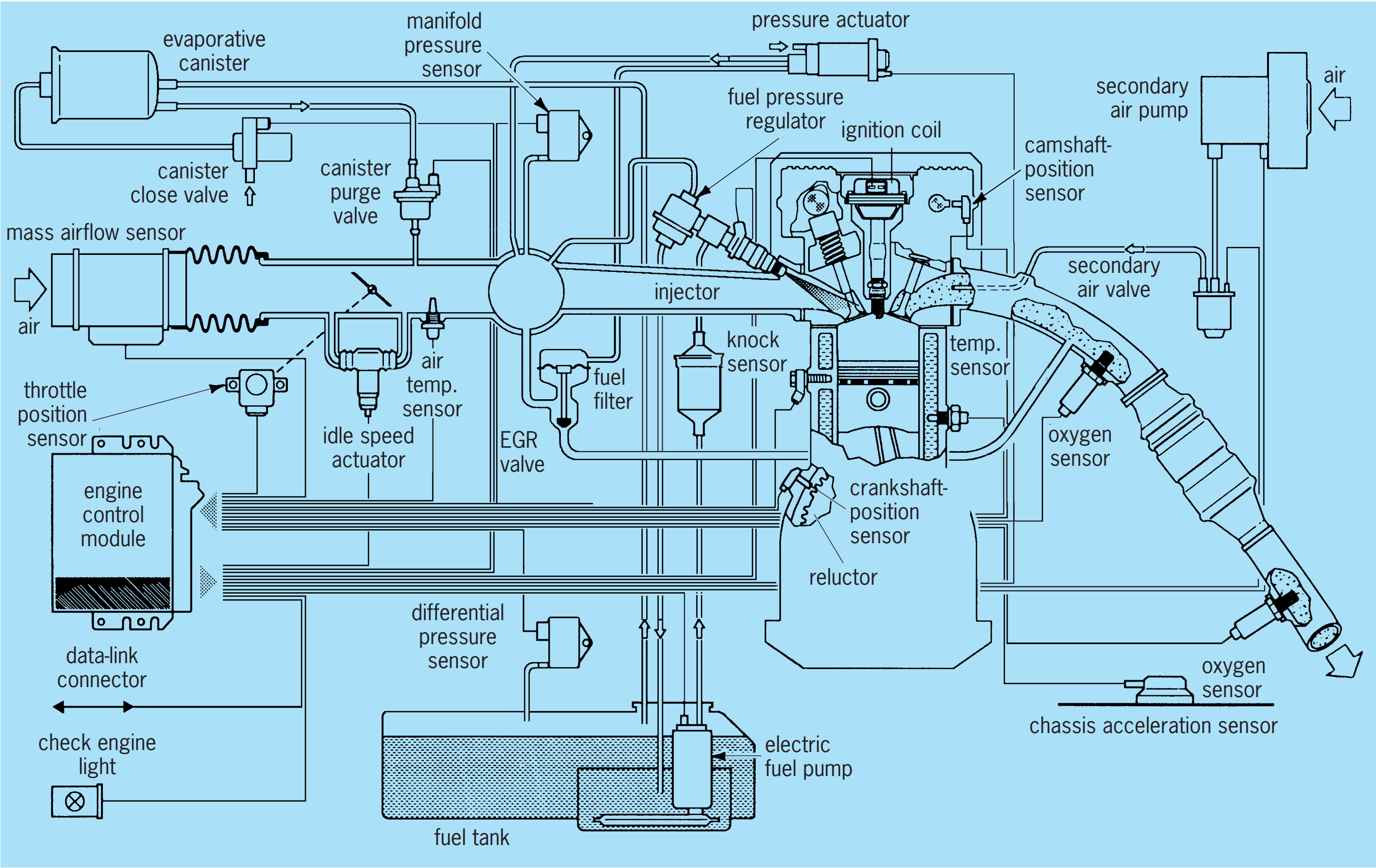

evaporative canister
manifold pressure sensor
pressure actuator
air
secondary air pump
fuel pressure regulator
ignition coil
camshaft-position sensor
canister close valve
canister purge valve
mass airflow sensor
air
injector
secondary air valve
throttle position sensor
air temp. sensor
knock sensor
temp. sensor
fuel filter
idle speed actuator
EGR valve
oxygen sensor
engine control module
crankshaft-position sensor
reluctor
data-link connector
differential pressure sensor
oxygen sensor
chassis acceleration sensor
check engine light
electric fuel pump
fuel tank

Fig. 4. Electronic engine control system that includes onboard diagnostics version II (OBD II), which can identify failure of certain emission-control components and detect deterioration in their performance throughout the life of the vehicle. (*Robert Bosch Corp.*)

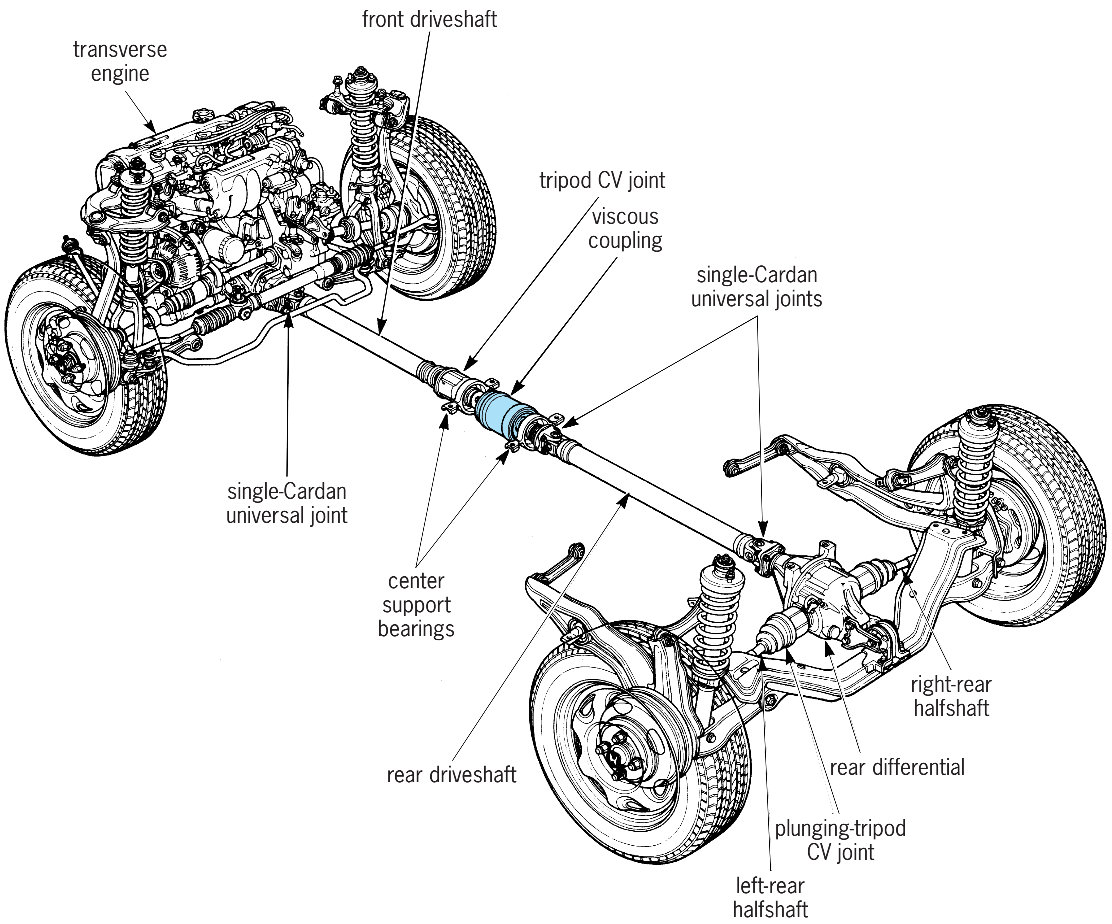


Fig. 5. Power train for an all-wheel-drive vehicle in which the front wheels are normally driven. When the front wheels spin, the viscous coupling locks and sends power to the rear axle. (*American Honda Motor Company, Inc.*)

wheels by the power train, or drive train (**Fig. 5**). In the front-engine rear-drive vehicle, the power train consists of a clutch and manual transmission, or a torque converter and an automatic transmission; driveshafts and Hooke (Cardan) universal joints; and rear drive axle that includes the final drive, differential, and wheel axle shafts. In the typical front-engine front-drive vehicle, the power train consists of a clutch and manual transaxle, or a torque converter and an automatic transaxle. The final drive and differential are designed into the transaxle, and drive the wheels through half-shafts with constant-velocity (CV) universal joints. *See* CLUTCH; GEAR; GEAR TRAIN; UNIVERSAL JOINT.

Most vehicles with four-wheel drive have a two-speed transfer case. This is an auxiliary transmission mounted behind the main transmission, and used to divide engine power and transfer it to both front and rear drive axles, either full or part time. The transfer case in most light-duty trucks has a high and a low range, doubling the number of forward (and reverse) speeds. Some vehicles have part-time four-wheel drive, which operates only when engaged by the driver. Other vehicles have full-time four-wheel drive, which engages and disengages automatically.

Passenger cars and vehicles such as station wagons and minivans may have all-wheel drive (Fig. 5), which is a type of four-wheel drive used to improve traction in on-road vehicles. If a transfer case is utilized, it has only high range. Many all-wheel-drive vehicles are front-drive with power to the rear wheels only when needed. Power transfer is controlled by a viscous coupling, which is a type of self-actuated fluid coupling that can be employed in the transfer case or drive line. If the primary drive wheels have normal traction, no power flows to the other axle. However, if the primary drive wheels lose traction and begin to spin, the viscous coupling locks, transferring power to the secondary drive wheels until the spinning wheels regain traction.

Transmission. The transmission is the device in the power train that provides different forward gear ratios between the engine and drive wheels, as well as neutral and reverse.

The two general classifications of transmission are manual transmission, which the driver shifts by hand, and automatic transmission, which shifts automatically. The transmission may be a separate unit, or it may be combined with the drive axle to form a transaxle which is used in front-drive vehicles. Most manual transmissions provide four, five, or six forward speeds. Automatic transmissions typically have four or five forward speeds.

To shift a manual transmission, the clutch must first be disengaged. However, some vehicles have automatic clutch disengagement for manual transmissions, while other vehicles have a limited manual-shift capability for automatic transmissions. A continuously variable transmission (CVT) is used in some vehicles. It provides an infinite number of gear ratios by transmitting torque through a belt that runs between two variable-diameter pulleys. *See* AUTOMOTIVE TRANSMISSION.

Final drive. In the power train, the final drive is the speed-reduction gear set that drives the differential. The final drive is made up of a large ring gear driven by a smaller pinion, or pinion gear. This provides a gear reduction of about 3:1; the exact value can be tailored to the engine, transmission, weight of the vehicle, and performance or fuel economy desired.

Differential. In drive axles, the differential is the gear assembly between axle shafts that permits one wheel to rotate at a speed different from that of the other (if necessary), while transmitting torque from the final-drive ring gear to the axle shafts. When the vehicle is cornering or making a turn, the differential allows the outside wheel to travel a greater distance than the inside wheel; otherwise, one wheel would skid, causing tire wear and partial loss of control.

The standard or open differential delivers the same torque to each wheel. But if one tire begins to slip and spin, the open differential divides the rotary speed unequally. The tire with good traction slows and stops. This action may also stop the vehicle or prevent it from moving. To limit the amount of slip and provide traction to both drive wheels, some drive axles have a limited-slip differential. It usually has a cone clutch or multiple-disc clutch on the inner end of each axle shaft. When the clutch engages, it locks the inner end of the axle shaft to the differential case so both axles turn at the same speed. A viscous coupling in the drive axle between the axle shafts also can serve as a limited-slip differential by locking both axle shafts together if one wheel slips. Some vehicles have differential locks that can be engaged and disengaged by the driver.

A differential is used in the drive axle of a rear-drive vehicle and in the transaxle of a front-drive vehicle. A vehicle with four-wheel drive has a differential in both front- and rear-drive axles. In addition, a third center or interaxle differential which may be in the transfer case is used to compensate for any difference in front and rear wheel travel in vehicles with four-wheel drive and all-wheel drive. *See* DIFFERENTIAL.

Wheels and tires. A wheel is a disc or a series of spokes with a hub at the center and a rim around the outside for mounting of the tire. The wheels of a vehicle must have sufficient strength and resiliency to carry the weight of the vehicle, transfer driving and braking torque to the tires, and withstand side thrusts over a wide range of speed and road conditions. Wheel size is primarily determined by the load-bearing strength of the tire.

The high speeds encountered in highway driving demand low centers of gravity and have resulted in the adoption of relatively small wheels and tires of matching size. After the tires are mounted on the wheels, or after the wheels are installed on the vehicle, the wheel-and-tire assemblies are spin-balanced. Even a slightly unbalanced wheel-and-tire assembly can cause steering trouble, vibration, and rapid tire wear at high speeds.

Passenger-car wheels are usually the steel disc type. The wheel disc may be a solid plate or a slotted steel disc. Either type can be welded or riveted to the rim. The disc is dished in and usually offset to bring the center point of ground contact under the larger inner wheel bearing. Many vehicles have aluminum wheels, which are lighter than steel and reduce unsprung weight. Molded composite wheels are installed on some vehicles. Composite wheels are made of fiberglass, sheet-molding compound (SMC), and special resins, and are lighter than aluminum wheels. Steel wheels are usually trimmed with wheel covers made of steel or aluminum and attached to the wheel by clips or other fasteners.

Most passenger-car rims have a drop center which improves tubeless tire sealing and bead retention while permitting removal and installation of the tire. The wheels mount to the vehicle with three to five wheel lug bolts or lug nuts. The nose of the lug bolts or nuts usually has either a conical or spherical shape to clamp the wheel in the radial and axial directions. By removing the lug bolts or nuts, the wheel is demountable from the brake drum, axle-shaft flange, or hub which rotates on the axle or steering-knuckle spindle. This serves as the mounting for the wheel to the vehicle.

The tire is the casing-and-tread assembly (with or without a tube) that is mounted on a wheel, filled with pressurized air, and installed on the vehicle to provide pneumatically cushioned contact and traction with the road surface. New automotive vehicles are equipped with tubeless radial tires. In radial tire construction, the plies run parallel to each other and vertical to the tire bead. Tire pressure may be monitored electronically, regardless of vehicle speed, with low pressure in any tire causing a warning light on the instrument panel to illuminate. *See* TIRE.

Some tires are self-sealing to minimize air loss if the tread area is punctured by a nail or other object that has a diameter of 3/16 in. (5 mm) or less. Other tires have stiffened sidewalls that provide limited run-flat capability for up to 200 mi (320 km) after loss of air. Run-flat tires reduce vehicle weight by allowing the jack and spare tire to be eliminated from the vehicle, improving fuel economy.

Automotive electronics. The use of solid-state electronic devices in the automobile began during the 1960s, when the electromechanical voltage regulator of the alternator, with vibrating contact points, was replaced by a transistorized voltage regulator. This was followed in the 1970s by electronic ignition, fuel injection, and cruise control. Since then, electronic devices and systems on the automobile have proliferated. These include engine and power train control, air bags, antilock braking, traction control, suspension and ride control, remote keyless entry, memory seats, driver information and navigation

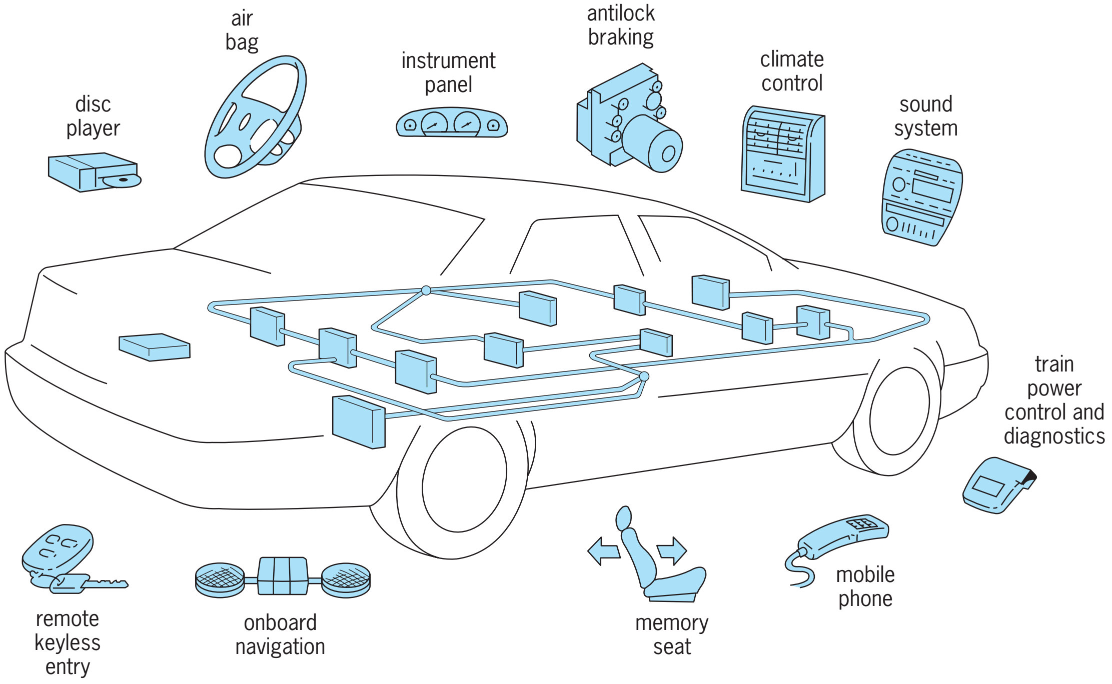


Fig. 6. **Automotive multiplex system that allows data transfer and information sharing by all control modules attached to the wiring loom. (*Cadillac Division, General Motors Corp.*)**

systems, cellular telephone and mobile communications systems, and onboard diagnostics. *See* ELECTRONIC DISPLAY; FEEDBACK CIRCUIT; SATELLITE NAVIGATION SYSTEMS.

The self-diagnostic capability of the vehicle computer, power-train or engine control module (Fig. 4), or system controller may be aided by a memory that stores information about malfunctions that have occurred and perhaps temporarily disappeared. When recalled from the memory, this information can help the service technician diagnose and repair the vehicle more quickly, accurately, and reliably.

As factory-installed electronics in the automobile has increased, so has the number of controllers or microprocessors, each of which may control only a separate stand-alone device or system. This creates duplication of wiring, sensors, and other system elements. To eliminate the redundancy, multiplex systems are being used which allow data transfer over a signal bus and sharing of this information with all other control modules that require it (**Fig. 6**). The multiplexed vehicle has fewer sensors, and the total length and weight of wiring, average wire gauge, and number of terminals and splices are greatly reduced while reliability is increased. *See* MULTIPLEXING AND MULTIPLE ACCESS.

Future progress in automotive electronics requires growth of multiplexing and continuing emphasis on systems integration, which combines separately developed devices so they work together as a complete system. Systems integration allows the signals from a wheel-speed sensor to be used in antilock braking, traction control, vehicle stability control, tire-pressure monitoring, and other systems that need wheel speed information. Drive-by-wire (electronic throttle control), brake-by-wire, steer-by-wire, obstacle detection, collision avoidance, and other automotive innovations will become less expensive and more functional as advances are made in the integration of mechanical, electronic, and data-processing systems.

Donald L. Anglin

Bibliography. K. Newton, W. Steeds, and T. K. Garret, *The Motor Vehicle*, 12th ed., Society of Automotive Engineers, 1996; Robert Bosch GmbH, *Automotive Handbook*, 4th ed., 1996; *SAE Handbook*, Society of Automotive Engineers, 3 vols., annually.

Automotive brake

An energy conversion device used to slow a vehicle, stop it, or hold it in position. The two systems are the service brake and the parking brake (**Fig. 1**), both of friction type. The service brake includes a hydraulically operated brake mechanism at each wheel. These wheel brakes are controlled by movement of the brake pedal, providing braking proportional to the applied pedal force. The parking brake is a mechanical brake operated through a separate hand lever or pedal; it applies parking-brake mechanisms usually at the two rear wheels. Most automotive vehicles have power-assisted braking, where a hydraulic or vacuum booster increases the force applied by the driver to the service-brake pedal. *See* BRAKE.

The two types of wheel-brake mechanisms are drum brakes and disc brakes (**Fig. 2**). Drum brakes are used at all four wheels on older vehicles, and at the rear wheels of many vehicles with front disc brakes. Some vehicles have disc brakes at all four wheels.

Hydraulic system. The four wheel brakes are hydraulically interconnected so they operate together

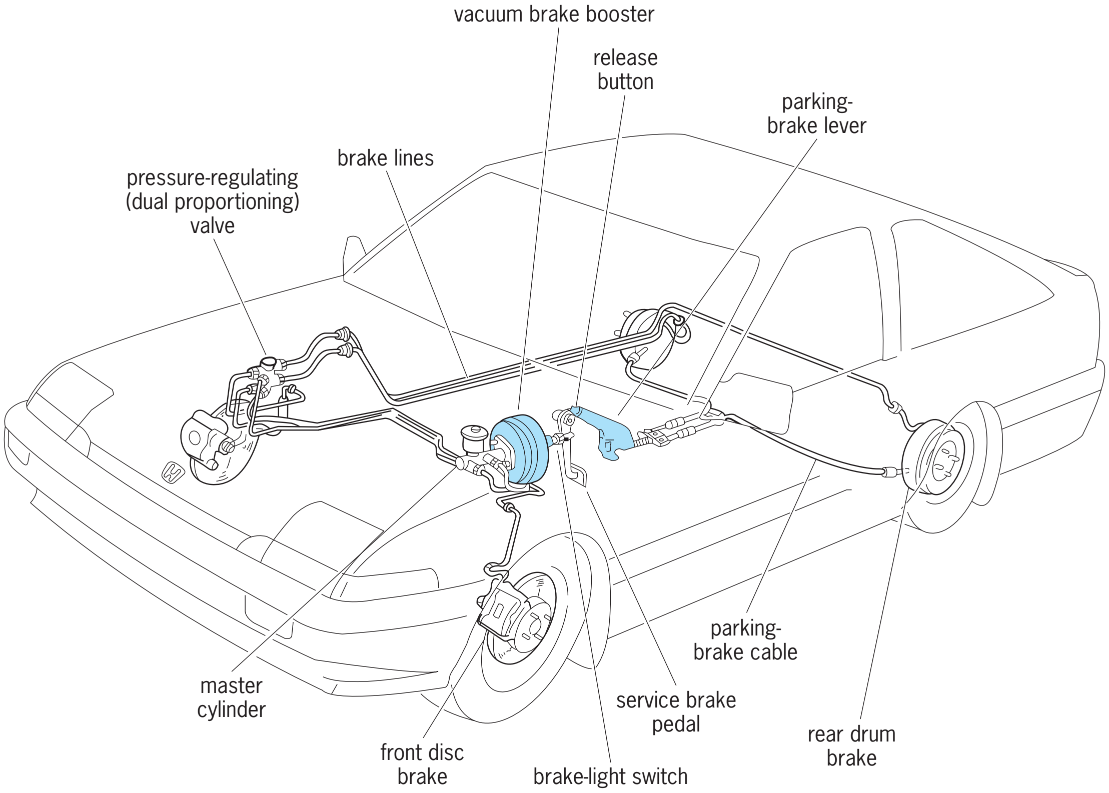


Fig. 1. Automotive brake system, showing a diagonally split hydraulic system with front disc brakes, rear drum brakes, vacuum booster, and hand-operated parking brake. (*American Honda Motor Co., Inc.*)

from one control. When the driver depresses the brake pedal, pistons are forced into fluid chambers in the master cylinder. The resulting hydraulic pressure is transmitted through steel pipe and rubber hose to hydraulic cylinders in the wheel brakes. The pressure forces pistons in the cylinders to move outward, pushing brake friction material, or lining, into contact with the rotating drum or disc to apply the brakes. *See* HYDRAULICS.

For safety, the master cylinder and hydraulic system are divided into two separate sections. Most rear-drive vehicles have a front-rear split, with one section of the hydraulic system operating the front brakes and the other operating the rear brakes. Most front-drive vehicles have a diagonal or X-type split, with each section operating one front brake and the diagonally opposite rear brake. Either arrangement allows some braking by one section if the other section fails because of damage or fluid loss.

In vehicles with front-disc and rear-drum brakes, a pressure-regulating valve is located in the hydraulic line to the rear brakes. During hard braking, this proportioning valve reduces pressure to the rear brakes, where less braking is needed, to avoid rear-wheel lockup and tire skid. A diagonally split hydraulic system may have two proportioning valves, one in each rear-brake line.

Drum brake. In a drum brake, two nonrotating curved steel shoes, faced with heat- and wear-resistant lining, are forced against the inner surface of a rotating brake drum as the driver depresses the brake pedal. When the pedal is released, return springs pull the shoes away from the drum.

Disc brake. In a disc brake, a nonrotating caliper containing one or more pistons and carrying two brake pads, or lined flat shoes, straddles the rotating disc. As the driver depresses the brake pedal, the piston and hydraulic reaction push the brake pads against each side of the disc. When the brake pedal is released, the piston seal, which was deflected as the piston moved out, provides piston retraction. Two types of caliper are the fixed or nonmoving, and the floating or sliding. The floating or sliding type depends on slight inward movement of the caliper, resulting from hydraulic reaction, to force the outer brake pad against the disc.

Power brake. Power-assisted braking is provided by a hydraulic or vacuum booster. As the brake pedal is depressed, the booster furnishes most of the force to push a pushrod into the master cylinder. The power piston in the hydraulic booster is operated by oil pressure from the power-steering pump or from a separate pump driven by an electric motor. In the vacuum booster, a diaphragm usually is suspended in a vacuum supplied from the engine intake manifold or from a vacuum pump driven by the engine or an electric motor. Depressing the brake pedal allows atmospheric pressure to act against one side of the diaphragm. The resulting pressure differential moves the diaphragm and power piston, which forces the pushrod into the master cylinder.

Antilock braking system. Braking is most efficient if lockup is prevented and the brakes are allowed to continue slowing the wheels. In a vehicle with a

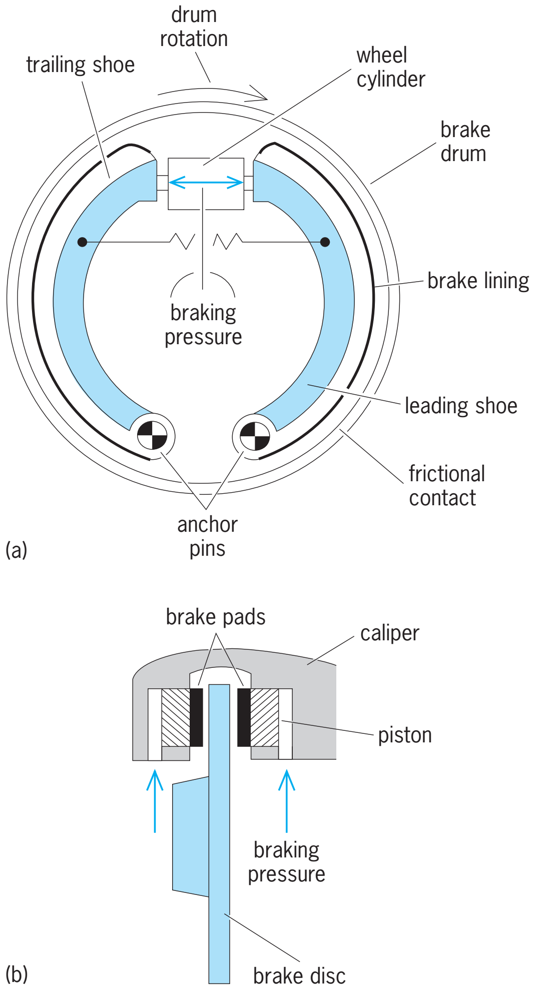


Fig. 2. Friction brakes of (*a*) drum type and (*b*) disc type used in automotive vehicles. (*Robert Bosch Corp.*)

typical antilock braking system (ABS), wheel-speed sensors signal the controller when a wheel is about to lock up. The controller then signals solenoid valves which open and close rapidly to prevent any increase, or begin to decrease, line pressure for that brake. This avoids lockup and tire slip. During normal braking, ABS has no effect on the service brakes.

Several types of ABS are in use. Some light-duty trucks, mostly those with primary rear-wheel drive and a front-rear hydraulic split, have a type of ABS that can prevent lockup of only the two rear wheels. The signal from the vehicle-speed sensor is used by the ABS controller to read the rate of rear-wheel deceleration. If pressure modulation is then necessary, both rear brakes will be affected equally because of the single hydraulic line to the rear axle.

Most vehicles have a wheel-speed sensor for each wheel. These ABS systems may be either three-channel or four-channel. In a three-channel system, the front brakes are controlled individually while both rear brakes are controlled together. In a four-channel system, each wheel brake can be individually controlled.

Traction-control system. When a tire receives more torque than it can transfer to the road, the tire loses traction and spins. To prevent this, a traction-control system (TCS) can be incorporated with ABS. When a wheel is about to spin, the controller applies the service brake at that wheel. This slows the wheel until the chance of wheel spin has passed. On many vehicles, TCS operation can reduce engine speed and torque if braking alone will not prevent wheel spin. Then, the engine controller momentarily retards ignition timing and reduces the amount of fuel delivered to the engine cylinders.

Stability-control systems. ABS and TCS provide braking and acceleration control when the vehicle is moving forward, or longitudinally. Some vehicles have an electronic stability-control system that can actively help control the vehicle during sideways or lateral movements such as cornering and skidding. These systems can automatically reduce engine torque while applying individual wheel brakes to stabilize the vehicle and maintain the driver's desired line of travel. *See* AUTOMOTIVE SUSPENSION.

Donald L. Anglin

Bibliography. Robert Bosch GmbH, *Automotive Handbook*, 4th ed., 1996.

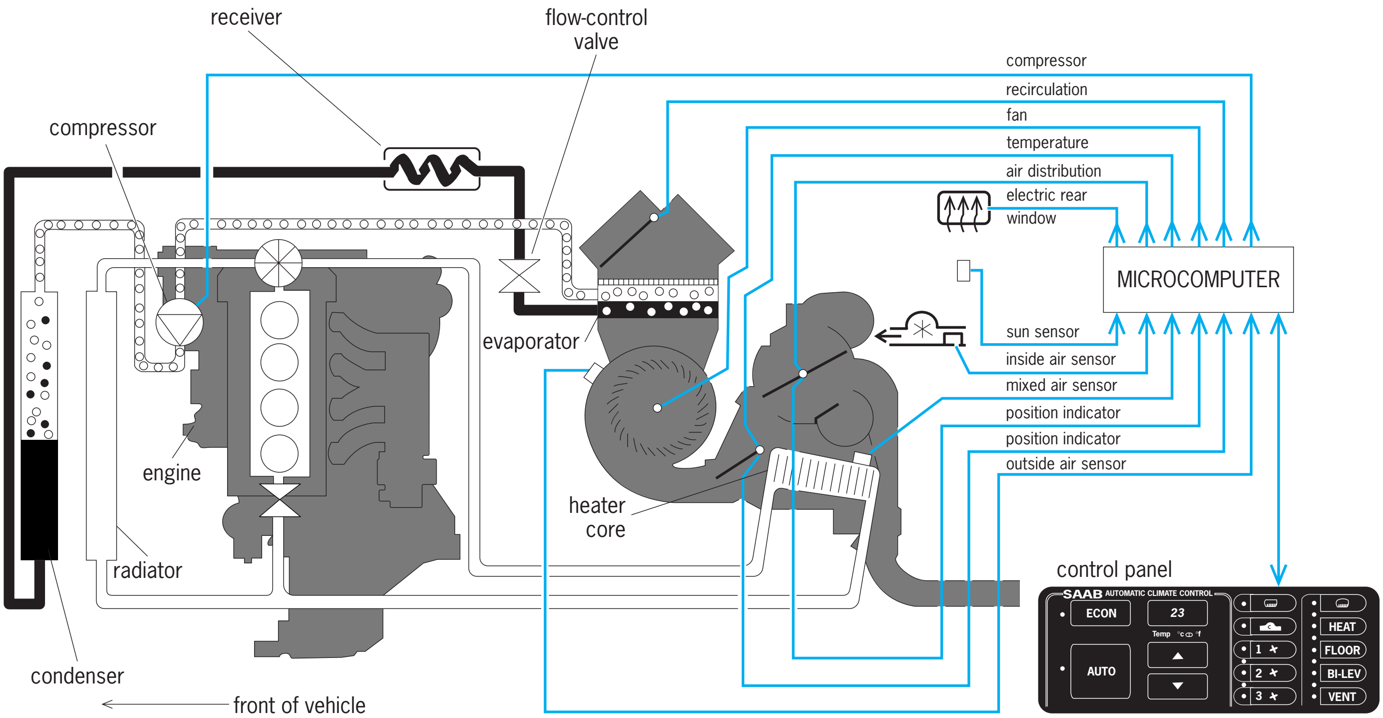
Automotive climate control

A system for providing a comfortable environment within the passenger compartment of a vehicle. Controlled ventilation is utilized, along with a heater, an air conditioner, or an integrated heater and air-conditioner system. Linked to the setup is a windshield defrosting and defogging system capable of clearing the windshield as specified in a timed test, as required for all passenger cars manufactured for sale in the United States. Some vehicles have a ventilation-air filter which cleans the outside air that enters the passenger compartment through the fresh-air inlet. The increasing glass area of many passenger vehicles places an additional load on the air conditioner. Many vehicles incorporate solar-control glass to reduce solar transmission to the interior.

Heating. There are two types of passenger-compartment heaters: engine-dependent and engine-independent. The engine-dependent heater utilizes waste heat from the engine. The engine-independent heater includes a small combustion chamber in which fuel is burned.

Most vehicles have a liquid-cooled engine and an engine-dependent heater through which hot engine coolant flows. The coolant passes through the tubes of a tube-and-fin heater core (see **illus.**) while air flows between the fins. Heat output into the passenger compartment is regulated by controlling either the coolant flow or the airflow. An electric blower motor may run at various speeds to help move the air. When the heater is turned off, a coolant flow-control valve may close to stop the flow of coolant through the heater core. *See* DEWAR FLASK; ENGINE COOLING.

Air conditioning. When the outside temperature is above 68°F (20°C), the passenger compartment may be uncomfortable for the occupants unless


receiver
flow-control valve
compressor
compressor
recirculation
fan
temperature
air distribution
electric rear window
MICROCOMPUTER
evaporator
sun sensor
inside air sensor
mixed air sensor
position indicator
position indicator
outside air sensor
engine
heater core
radiator
control panel
SAAB AUTOMATIC CLIMATE CONTROL
ECON
23
Temp °C ◐ °F
AUTO
1
2
3
HEAT
FLOOR
BI-LEV
VENT
condenser
front of vehicle

Schematic of an automotive air conditioner. (*Saab Cars USA, Inc.*)

the inside air is cooled. The cooling is provided by a mobile, vehicle-mounted refrigeration system known as an automotive air conditioner (see illus.).

The automotive air conditioner combines the refrigeration system with an air-distribution system and a temperature-control system to cool, clean, dry, and circulate passenger-compartment air. Cooling is provided by a mechanical vapor-compression refrigeration system with five major components: compressor, condenser, refrigerant flow-control valve, evaporator, and a receiver or accumulator that includes a desiccant. These parts are connected by hose and tubing, through which the refrigerant flows. Most systems use a fixed-orifice tube as the refrigerant flow-control valve. Other systems use a variable-orifice expansion valve to control refrigerant flow. The condenser and the evaporator are both tube-and-fin heat exchangers. The condenser usually is mounted ahead of the engine radiator at the front of the vehicle. The evaporator is located in the engine compartment. *See* COMPRESSOR; DESICCANT; EVAPORATOR; FLUOROCARBON; HEAT EXCHANGER; REFRIGERATION.

The compressor usually is driven by a belt from the engine crankshaft. When the engine is running and the compressor operating, heat is removed from the passenger-compartment air and transferred to the outside air. This process begins when a small amount of low-pressure liquid refrigerant flows into the evaporator. At the same time, warm passenger-compartment air flows between the evaporator fins. The liquid refrigerant absorbs heat from the warm air and vaporizes, cooling the evaporator and the air passing through it. The heat-carrying low-pressure vapor then flows from the evaporator into the accumulator (in an orifice-tube system), which traps any liquid refrigerant to avoid its damaging the compressor.

From the accumulator, the low-pressure vapor is drawn into the compressor, which compresses the vapor, raising its pressure and temperature. The hot, high-pressure vapor then flows into the condenser, where the hotter vapor loses heat to the cooler outside air flowing across the condenser. As the vapor loses heat, it condenses into a high-pressure liquid that leaves the condenser and flows to the refrigerant flow-control valve. The small opening in the valve allows passage of only a small amount of liquid refrigerant, which now has a lower pressure as it enters the evaporator. Passing through the evaporator, the low-pressure liquid refrigerant again absorbs heat and vaporizes. This cooling cycle repeats as long as the compressor runs.

Operation, air temperature, and air distribution through the passenger compartment may be controlled either automatically or manually by the driver. In some vehicles, conditioned air distribution can be controlled for each seat or seating position. *See* AIR CONDITIONING.

Refrigerant. For years, the refrigerant in automotive air conditioners was dichlorodifluoromethane (CCl_2F_2), a chlorofluorocarbon (CFC). Because of damage to the Earth's ozone layer caused by release of this compound into the atmosphere, R-12 has been replaced by less harmful compounds. The factory-installed refrigerant in most automotive air conditioners is the hydrofluorocarbon 1,1,1,2-tetrafluorethane (CF_3CH_2F), known as R-134a. *See* ATMOSPHERIC CHEMISTRY.

Clean Air Act. In the United States, the 1990 amendments to the Clean Air Act require that any work on an automotive air conditioner involving the refrigerant must be performed by a trained and certified technician using approved refrigerant recycling equipment. The refrigerant must not be discharged or vented into the atmosphere. Recycling the recovered refrigerant cleans it for reuse.

After recycling, contaminants in refrigerant R-12 should not exceed, by weight, 15 parts per million of moisture, 4000 ppm of refrigerant oil, and 330 ppm of noncondensable gases (air). Limits, by weight, for refrigerant R-134a are 15 ppm of moisture, 500 ppm of refrigerant oil, and 150 ppm of noncondensable gases. Donald L. Anglin

Bibliography. P. E. Anglin (ed.), *Automotive Climate Control Systems*, Society of Automotive Engineers, 1991; *SAE Handbook*, Society of Automotive Engineers, 3 vols., annually.

Automotive drive axle

A theoretical or actual crossbar or assembly that supports a motor vehicle and on which one or more wheels turn. The axle is either a live axle or a dead axle. A live axle, or drive axle, drives the wheels connected to it while supporting part of the weight of the vehicle. A dead axle, or nondrive axle, carries part of the weight of the vehicle but does not drive the wheels. *See* AUTOMOBILE.

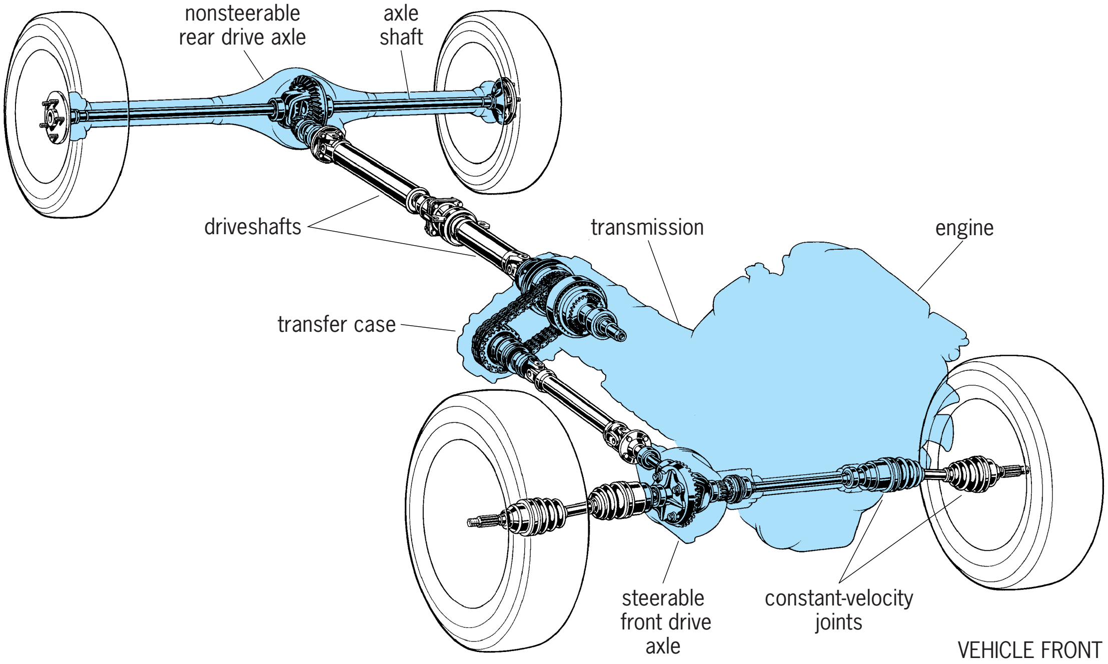
A drive axle on which the wheels can pivot for steering, such as on the front axle of a four-wheel-drive vehicle, is a steerable drive axle (see **illus.**). The rear axle in most automotive vehicles is a nonsteering drive axle.

Automotive configurations include front-wheel drive, rear-wheel drive, and four-wheel or all-wheel drive. All of these include a transmission or transaxle, differential, and driveline or axle shafts. In vehicles with a manual transmission or transaxle, the power flows from the engine crankshaft through a clutch. *See* AUTOMOTIVE TRANSMISSION; CLUTCH.

The wheels that rotate to move the vehicle are on the outer ends of a drive axle. A vehicle with front-wheel drive has a front drive axle, while the rear-wheel-drive vehicle has a rear drive axle. In a vehicle with four-wheel drive or all-wheel drive, power can be sent to all four wheels. The difference is that four-wheel drive usually includes a two-speed transfer case that makes the vehicle suitable for off-road use. All-wheel drive does not have a two-speed transfer case, making the vehicle primarily for on-road use. *See* TRUCK.

Rear drive axle. The rear drive axle is suspended from the vehicle body or frame by springs attached to the axle housing. The housing encloses the final-drive gears, differential, and wheel axle shafts. *See* AUTOMOTIVE SUSPENSION.

The functions performed by the rear drive axle include changing the rotation of the driveshaft by 90° to rotate the wheel axle shafts; providing a final speed reduction between the driveshaft and the axle shafts through the final-drive gears; permitting one wheel to turn at a speed different from the other when necessary (differential action); driving the rear wheels through axle shafts, or halfshafts; and acting as thrust- and torque-reaction members during acceleration and braking. The differential may be either the open type or, for greater traction, the limited-slip type. Differential action between wheels, and between front and rear drive axles (interaxle



Drive train of a four-wheel-drive vehicle, showing the steerable front drive axle and nonsteerable rear drive axle. (*Mazda Motors of America Inc.*)

differential), may also be provided by a viscous coupling. *See* DIFFERENTIAL.

Instead of an axle housing, some vehicles have independent rear suspension. The differential carrier is attached to a body or frame crossmember, and the axle halfshafts are exposed. A constant-velocity universal joint (CV joint) is usually located at the inner and outer ends of each halfshaft. The nonsteerable rear wheels are held in position by rear-suspension control arms. *See* UNIVERSAL JOINT.

Front drive axle. Most four-wheel-drive vehicles have a steerable front drive axle. It is usually similar in construction and function to the rear drive axle, having a housing and enclosed axle shafts. The principal difference is in the provisions made for steering. Because the front wheels must turn in or out on the steering-knuckle pivots, universal joints are located at the outer ends of the axle shafts. This allows the wheels to be driven while moving in and out or up and down. *See* AUTOMOTIVE STEERING.

Automotive vehicles with front-engine and front-wheel drive have independent front suspension. A housing-type front drive axle is not used. Instead, the functions of the transmission and drive axle are combined into the transaxle. The axle halfshafts are exposed, extending from the sides of the transaxle to drive the front wheels. Constant-velocity universal joints are located at the inner and outer ends of each halfshaft.

Steerable rear drive axle. Some automotive vehicles have steerable drive axles on the front and rear which provide both four-wheel drive and four-wheel steering. In these vehicles, the rear wheels are steered in the opposite direction to that of the front wheels at low speed and in the same direction at higher speed. This increases vehicle maneuverability. Donald L. Anglin

Bibliography. W. H. Crouse and D. L. Anglin, *Automotive Mechanics,* 1993.

Automotive electrical system

The system in an automotive vehicle that furnishes the electrical energy to crank the engine for starting, recharge the battery after cranking, create the high-voltage sparks to fire the compressed air-fuel charges, and power the various lights, motors, controllers and control systems, and other signaling and accessory systems.

The vehicle electrical system includes the battery, wiring, starting motor and controls, generator and voltage regulator, electronic ignition, and electronic fuel injection. Also included may be a computerized power-train control system, electronically displayed driver-information system, air bags, power seats and windows, and various types of radios and sound systems.

Most vehicles have a one-wire or ground-return system, with the metal of the vehicle serving as the common ground. To reduce wiring and duplication, some vehicles have interconnected controllers and components that form a multiplex network. By sharing wiring and information over a data bus, the vehicle has fewer sensors, and the amount of wiring and the number of connections are greatly reduced. *See* BATTERY; ELECTRIC SWITCH; ELECTRONIC DISPLAY; ELECTRONICS; FUSE (ELECTRICITY); GENERATOR; REGULATOR; SPEEDOMETER; VOLTAGE REGULATOR. Donald L. Anglin

Bibliography. W. H. Crouse and D. L. Anglin, *Automotive Mechanics*, 1993.

Automotive engine

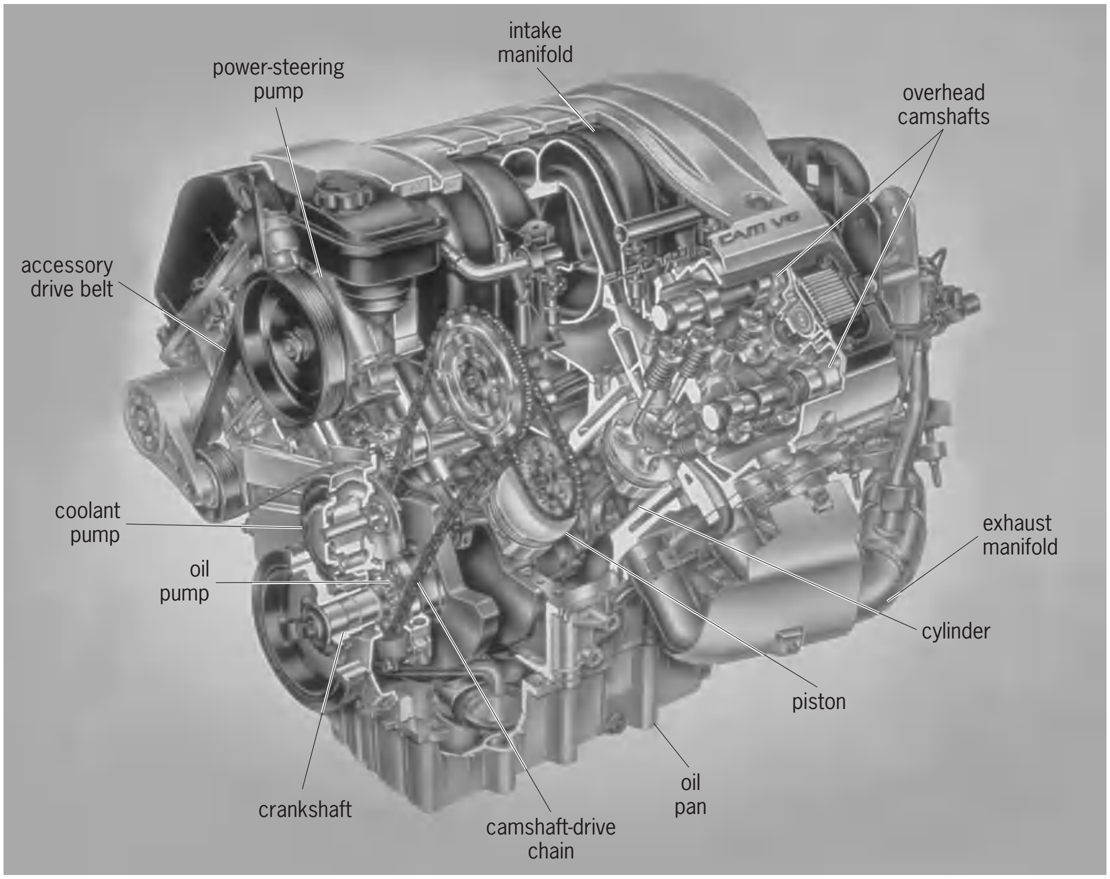
The component of the motor vehicle that converts the chemical energy in fuel into mechanical energy for power. The automotive engine also drives the generator and various accessories, such as the air-conditioning compressor and power-steering pump. *See* AUTOMOTIVE CLIMATE CONTROL; AUTOMOTIVE ELECTRICAL SYSTEM; AUTOMOTIVE STEERING.

Early motor vehicles were powered by a variety of engines, including steam and gasoline, as well as by electric motors. The flexibility of the gasoline engine operating on the four-stroke Otto cycle soon made this engine predominant, and it remains the dominant automotive power plant. The basic modern automotive engine (see **illus.**) is a gasoline-burning, liquid-cooled, spark-ignition, four-stroke-cycle, multicylinder engine. It has the intake and exhaust valves in the cylinder head, and electronically controlled ignition and fuel injection. *See* ENGINE.

Otto-cycle engine. An Otto-cycle engine is an internal combustion piston engine that may be designed to operate on either two strokes or four strokes of a piston that moves up and down in a cylinder. Generally, the automotive engine uses four strokes to convert chemical energy to mechanical energy through combustion of gasoline or similar hydrocarbon fuel. The heat produced is converted into mechanical work by pushing the piston down in the cylinder. A connecting rod attached to the piston transfers this energy to a rotating crankshaft. *See* GASOLINE; INTERNAL COMBUSTION ENGINE; OTTO CYCLE.

Cylinder arrangement. Engines having from 1 to 16 cylinders in in-line, flat, horizontally opposed, or V-type cylinder arrangements have appeared in production vehicles, progressing from simple single-cylinder engines at the beginning of the twentieth century to complex V-12 and V-16 engines by the early 1930s. Increased vehicle size and weight played a major role in this transition, requiring engines with additional displacement and cylinders to provide acceptable performance.

High-volume usage of the V-8 engine began in the mid-1930s and accelerated dramatically after World War II, until it was the predominant engine used in American-built vehicles by the late 1950s. Manufacturers in other countries continued large-volume production of smaller engines with four and six cylinders, primarily because of significantly higher fuel costs. As vehicle size and weight increased, average engine displacement also increased until the early 1970s, when V-8 engines approaching 500 in.^{3}



Automotive engine, which has six cylinders, double-overhead camshafts, 24-valve electronic coil-on-plug spark ignition, and multiport fuel injection. (*Oldsmobile Division, General Motors Corp.*)

(8 liters) displacement were in production. However, oil shortages in 1973–1974 and 1979–1980 reversed this trend, and V-8 engine usage dropped in favor of engines with four and six cylinders.

Turbocharger and supercharger. To provide acceptable vehicle performance with a smaller engine, forced induction may be used. A turbocharger or supercharger forces more air into the intake manifold, allowing the engine to burn more fuel and produce more power. The turbocharger is a centrifugal air compressor driven by an exhaust-gas-powered turbine mounted on a common shaft. The energy in the exhaust gas spins the turbine, which spins the compressor, forcing more air or air-fuel mixture into the combustion chambers. In a typical passenger car, this may increase engine power output by up to 40%.

A supercharger, which is belt-driven from the engine crankshaft, may be used instead of a turbocharger. The supercharger does not have the brief acceleration lag, or so-called turbo lag, that is found objectionable by many drivers of vehicles with turbocharged engines. *See* AUTOMOBILE; COMBUSTION CHAMBER; COMPRESSOR; MUFFLER; SUPERCHARGER; TURBINE; TURBOCHARGER.

Emissions. In the United States, passenger-car emission standards became effective in California in 1966 and in the other 49 states in 1968. These regulations began placing limits on crankcase, exhaust, and evaporative emissions into the atmosphere. The limits became increasingly stringent over the years, requiring the use of catalytic converters and unleaded gasoline beginning with 1975-model cars. Because more accurate fuel metering and ignition timing were required on engines to meet the tightening standards, electronic controls became necessary. As a result, fuel injection replaced the carburetor on automotive engines.

Electronic controls. Ignition, fuel, and emissions systems are integrated under an electronic engine control system. The system utilizes an onboard computer to provide management of various engine-operating parameters and emissions devices. The computer, known as the powertrain control module, may also control shifting of the automatic transmission or transaxle.

Engine design trends. In many automotive engines, the camshaft, which operates the intake and exhaust valves, has been moved from the cylinder block to the cylinder head (see illus.). This overhead-camshaft arrangement allows the use of more than two valves per cylinder, with various multivalve engines having three to five. Some overhead-camshaft engines have only one camshaft, while others have two camshafts, one for the intake valves and one for the exhaust valves. A V-type engine may have four camshafts, two for each bank of cylinders. Some multivalve overhead-camshaft engines have the power and performance of a turbocharged engine of similar size.

Most engines have fixed valve timing, regardless of number of camshafts or their location. Variable valve timing can improve fuel economy and minimize exhaust emissions, especially on multivalve engines. At higher speeds, volumetric efficiency can be increased by opening the intake valves earlier. One method drives the camshaft through an electrohydraulic mechanism that, on signal from the engine computer, rotates the intake camshaft ahead about 10°. Another system varies both valve timing and valve lift by having two cam lobes, each with a different profile, that the computer can selectively engage to operate each valve. Computer-controlled solenoids for opening and closing the valves will allow elimination of the complete valve train, including the camshaft, from the automotive piston engine while providing variable valve timing and lift.

Materials trends. Historically, major engine components have been made from ferrous metals, either by casting or by forging. However, emphasis on weight reduction for improved fuel economy has greatly increased the usage of aluminum for cylinder blocks, cylinder heads, and other engine components. Some engine covers and intake manifolds are made of magnesium. Internal engine parts, such as connecting rods, sprockets, oil-pump rotors, and valve guides, are cast or forged to nearly net shape using powder metallurgy. High-speed engines may use titanium connecting rods to reduce reciprocating mass. *See* POWDER METALLURGY.

Parts such as engine covers, intake manifolds, and oil pans also can be fabricated of plastic or composite materials. These materials provide weight savings while reducing engine noise and vibration. Ceramic engine parts and coatings will allow engine operation at higher temperatures, raising engine efficiency. Ceramic-lined exhaust ports in the cylinder head can lower its temperature while increasing the effectiveness of the catalytic converter.

Fuel-metering trends. With the introduction of electronic controls, a device was added to the carburetor to automatically adjust the air-fuel ratio in response to feedback from an exhaust-gas oxygen sensor. Demand for more accurate fuel metering resulted in the feedback carburetor being replaced by a similarly located throttle-body fuel-injection unit. It meters fuel through the computer-controlled pulsing of one or two solenoid-operated fuel injectors. Further improvements in engine power, fuel economy, and exhaust emissions are provided by multiport fuel injection, which places a fuel injector in each intake port. Solenoid-operated fuel injectors can be pulsed or energized in simultaneous, group, or sequential fashion—the last energizes each injector individually in firing-order sequence.

Ignition trends. On many automotive engines, the ignition distributor has been replaced with computer-controlled distributorless ignition; this in turn is being replaced with coil-on-plug or direct ignition, in which an ignition coil sits directly above, and is connected to, each spark plug. Some engines have two spark plugs per cylinder to provide higher power output with cleaner combustion and less tendency for spark knock, or detonation. Spark knock can be monitored by a knock sensor, which signals the computer for less spark advance to prevent engine damage. The knock sensor also is used, especially with a supercharger or turbocharger, to allow engine operation on a more economical, lower-octane-rated fuel than otherwise would be required.

Onboard diagnostic developments. An onboard computer with self-diagnostic capability has become standard equipment for automotive engine control. The first generation of onboard diagnostics (OBD I) identified the failure of certain emission-control components. The second generation (OBD II), required for 1996 and later model vehicles, has additional capability, including detection of deterioration in performance of emission-control components throughout the life of the vehicle.

Alternative engines. Alternative engine designs have been investigated as replacements for the four-stroke Otto-cycle piston engine, including the two-stroke, diesel, Stirling, Wankel rotary, gas turbine, and steam engines, as well as electric motors and hybrid power plants. However, only two engines are in mass production as automotive power plants: the four-stroke gasoline engine described above, and the diesel engine. Continuing improvements to the Otto-cycle piston engine, such as electronic controls and value actuation and other changes in design and materials, appear to assure its predominance in the short term. *See* BATTERY; DIESEL ENGINE; ELECTRIC VEHICLE; FUEL CELL; GAS TURBINE; MOTOR; POWER PLANT; ROTARY ENGINE; SOLAR CELL; STEAM ENGINE; STIRLING ENGINE.

Donald L. Anglin

Bibliography. H. Heisler, *Advanced Engine Technology*, Society of Automotive Engineers and Edward Arnold, 1995.

Automotive steering

The means by which a motor vehicle is controlled about the vertical axis. It allows the driver to control the course of vehicle travel by turning the steering wheel, which turns the input shaft in the steering gear. The steering system (**Fig. 1**) has three major components: (1) the steering wheel and attached shaft in the steering column which transmit the driver's movement to the steering gear; (2) the steering gear that increases the mechanical advantage while changing the rotary motion of the steering wheel to linear motion; and (3) the steering linkage (including the tie rod and tie-rod ends) that carries the linear motion to the steering-knuckle arms. *See* MECHANICAL ADVANTAGE.

When the only energy source for the steering system is the force that the driver applies to the steering wheel, the vehicle has manual steering. When the driver's effort is assisted by hydraulic pressure from an electric or engine-driven pump, or by an electric motor, the vehicle has power-assisted steering, commonly known as power steering.

Steering column. The steering column (Fig. 1) supports the steering wheel and encloses the steering

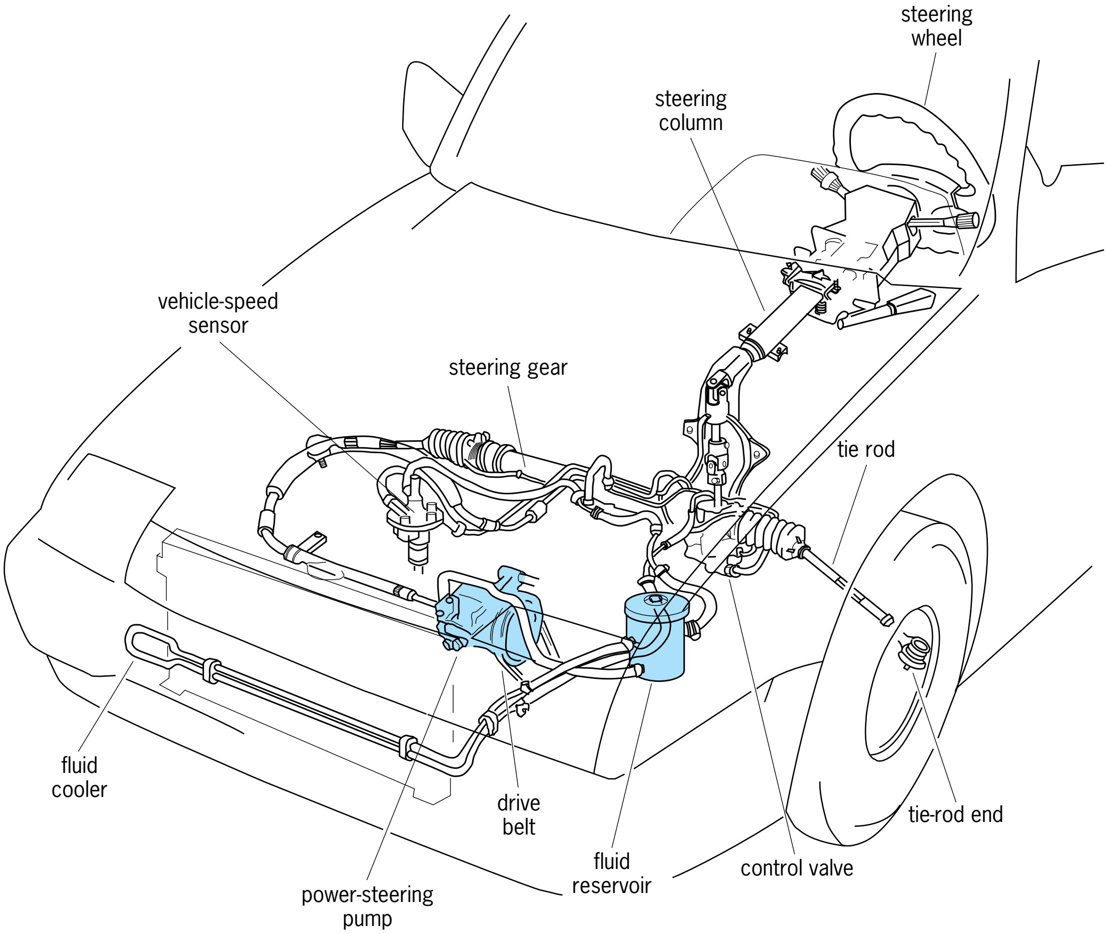


Fig. 1. Speed-sensitive rack-and-pinion power-steering system that provides variable assist. (*American Honda Motor Co., Inc.*)

shaft. Many vehicles have tilting steering wheels, some of which can also be moved in or out of the steering column. In some vehicles, the desired position of the steering wheel can be stored in memory and then automatically reset after the driver enters. The driver's side air bag and inflator are stored in the center of the steering wheel. *See* AUTOMOBILE.

Types of steering gears. Two types of automotive steering gears are rack-and-pinion and recirculating-ball. In a rack-and-pinion steering gear (**Fig. 2**), a tubular housing contains the toothed rack and a pinion gear. The housing is mounted rigidly to the vehicle body or frame to take the reaction to the steering effort. The pinion gear is attached to the lower end of the steering shaft, and meshes with rack teeth. Tie rods connect the ends of the rack to the steering-knuckle arms at the wheels. As the steering wheel turns, the pinion gear moves the rack right or left. This moves the tie rods and steering-knuckle arms, which turn the wheels in or out for steering.

In a recirculating-ball steering gear, a worm gear is attached to the lower end of the steering shaft. The worm gear turns inside a ball nut which rides on a set of recirculating ball bearings. These ball bearings roll in the grooves in the worm and inside the ball nut. Gear teeth on one outside flat of the ball nut mesh with a sector of teeth on the output or sector shaft to which the pitman arm is attached. As the steering wheel is turned, the rotary motion of the worm gear causes the ball nut to move up or down, forcing the sector shaft and pitman arm to rotate. This action moves the steering linkage to the right or left, turning the front wheels in or out for steering. *See* ANTIFRICTION BEARING; GEAR.

Many steering gears provide variable-ratio steering in which the mechanical advantage is less in the straight-ahead, on-center position and greater in any off-center position. This arrangement provides a fast steering ratio with improved response for highway driving, and a slow steering ratio with reduced steering effort for cornering and parking. Some vehicles have four-wheel steering that is controlled either mechanically or electronically.

Power steering. When the force applied to turn the steering wheel is produced by the driver and another energy source, the vehicle has power-assisted steering. This arrangement, commonly known as power steering, allows manual steering to always be available, even if the engine is not running or the power-assist system fails.

There are three types of hydraulic power steering: integral rack-and-pinion (Fig. 2), integral recirculating-ball, and nonintegral or linkage-type

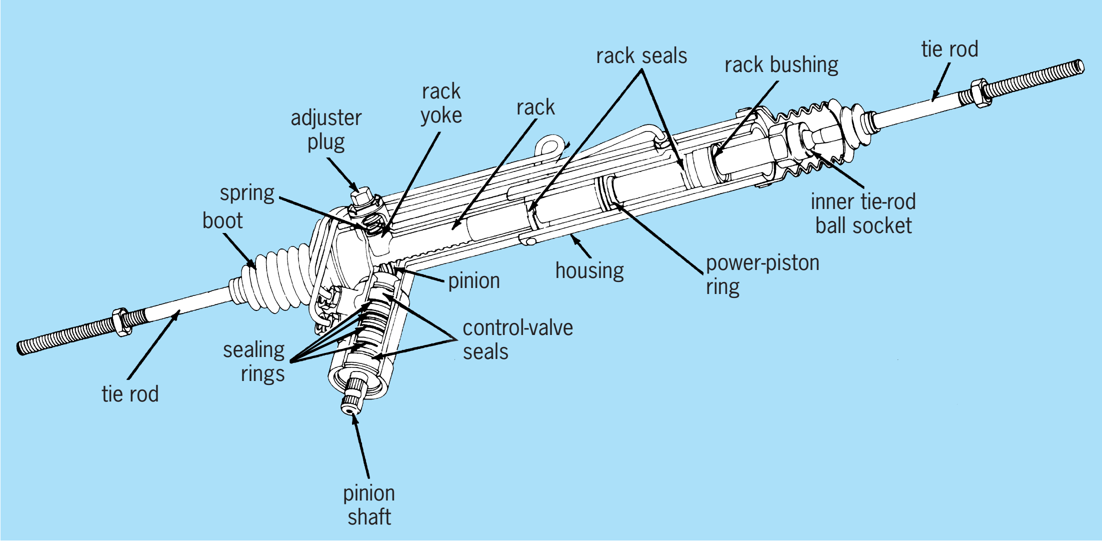

rack seals
rack bushing
tie rod
rack
yoke
rack
adjuster
plug
spring
boot
inner tie-rod
ball socket
pinion
housing
power-piston
ring
control-valve
seals
sealing
rings
tie rod
pinion
shaft

Fig. 2. Construction of a rack-and-pinion power-steering gear. (*Moog Automotive, Inc.*)

recirculating ball. Common components include the pump, fluid reservoir, control valve, power cylinder, and hoses. The fluid may be cooled by passing through a small heat exchanger or fluid cooler. Integral means the power cylinder is built into the steering gear. Nonintegral means a separate power cylinder is attached between the vehicle body or frame and the steering linkage.

Operation of hydraulic power steering is by two basic systems, one that generates and transmits the power assist and the other that controls it. Pressure is developed by the pump, which forces fluid from the reservoir through a pressure hose to the control valve. When the steering wheel is turned, the control valve senses the steering direction desired and directs oil accordingly to one side of the piston in the power cylinder. The fluid then returns under low pressure through a hose to the reservoir. Some systems are electronically controlled by a microprocessor that varies the assist based on vehicle speed and steering-wheel angle. *See* HYDRAULICS; SERVOMECHANISM.

An electronic rack-and-pinion power-steering gear is used on same vehicles. It replaces hydraulic power steering with electronic controls and an electric motor built into the rack housing. Steering-wheel movement is detected by a sensor mounted on the steering-gear input shaft. The electric motor is powered by the vehicle battery, which provides power assist even if the engine is not running.

Ackerman steering. Wheeled motor vehicles use the Ackerman system of steering, in which all wheels roll on circles with a common center. The front wheels are mounted on pivoted steering knuckles, and the steering linkage connects the steering-knuckle arms so the wheels swing together about their pivots. During a turn, the inwardly inclined arms move the inner wheel in more than the outer wheel. This toe-out on turns permits the front tires to round a turn without sideslip. *See* AUTOMOTIVE SUSPENSION.

Wheel alignment. Wheel alignment is the relationship among the wheels, steering parts, suspension angles, and the road that affect the operation and steering of a vehicle. Six basic wheel alignment factors are suspension height, caster, camber, toe, steering-axis inclination, and turning radius or toe-out on turns (**Fig. 3**). Alignment of the front wheels with the rear wheels should provide a common vehicle centerline, geometric centerline, and thrust line.

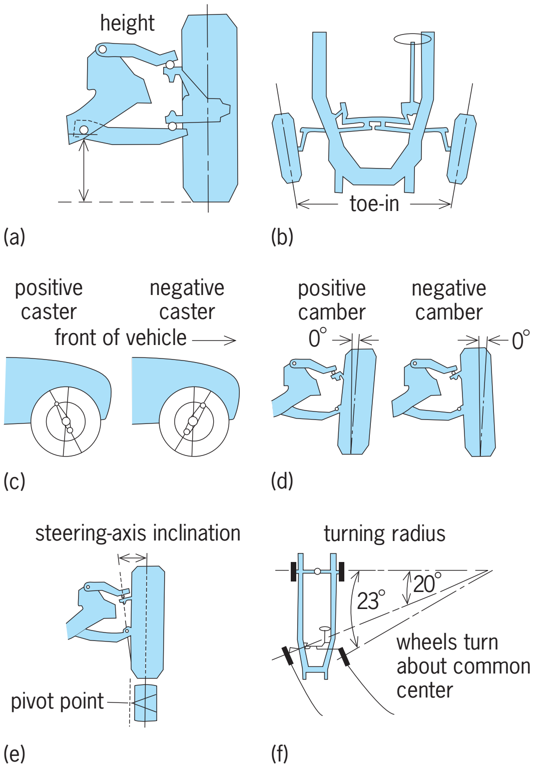

height
(a)
toe-in
(b)
positive
caster
negative
caster
front of vehicle
(c)
positive
camber
negative
camber
0°
0°
(d)
steering-axis inclination
pivot point
(e)
turning radius
23°
20°
wheels turn
about common
center
(f)

Fig. 3. Basic factors in wheel alignment. (*a*) Suspension height. (*b*) Toe. (*c*) Caster. (*d*) Camber. (*e*) Steering-axis inclination. (*f*) Turning radius. (*Chrysler Corp.*)

Suspension height is the distance measured from a specific point on the body, frame, or suspension to the ground. Caster is the angle formed by the forward or rearward tilt of the steering axis from the vertical, when viewed from the side of the wheel. The angle is positive when the steering axis tilts backward. Camber is the amount the top of a wheel tilts in (negative) or out (positive), when viewed from the front of the vehicle. Toe is the amount that the front of the wheels point inward (toe-in) or outward (toe-out). Steering-axis inclination is the angle between the vertical and a line drawn through the centers of the suspension ball joints, when viewed from the front of the vehicle. Turning radius, or toe-out on turns, is the difference between the angle that each steered wheel makes with the vehicle body or frame during turns, usually measured with the outside wheel turned 20°.

Other factors that affect steering are scrub radius or steering offset, wheel setback, and thrust angle. Scrub radius is the distance between the steering axis and the tire contact-area centerline at their intersections with the road surface. Wheel setback is the difference in vehicle wheelbase from one side to the other. Thrust angle is the angle formed between the vehicle and geometric centerlines, and the thrust line. The thrust line runs forward from the midpoint between the two rear wheels and determines the direction that the vehicle will travel if unaffected by the front wheels. If the thrust line does not coincide with the vehicle centerline, a thrust angle is formed and the thrust line then represents the course that the rear wheels will try to travel. Donald L. Anglin

Bibliography. Robert Bosch GmbH, *Automotive Handbook*, 4th ed., 1996.

Automotive suspension

The springs and related parts intermediate between the wheels and the frame, subframe, or side rails of a unitized body. The suspension supports the weight of the upper part of a vehicle on its axles and wheels, allows the vehicle to travel over irregular surfaces with a minimum of up-and-down body movement, and allows the vehicle to corner with minimum roll or loss of traction between the tires and the road. Four types of spring used in automotive suspension are coil, leaf, torsion bar, and air. *See* AUTOMOBILE; SPRING (MACHINES).

In a typical suspension system for a vehicle with front-engine and front-wheel drive (see **illus.**), the weight of the vehicle applies an initial compression to the coil springs. When the tires and wheels encounter irregularities in the road, the springs further compress or expand to absorb most of the shock. The suspension at the rear wheels is usually simpler than for the front wheels, which require multiple-point attachments so the wheels can move up and down while swinging from side to side for steering.

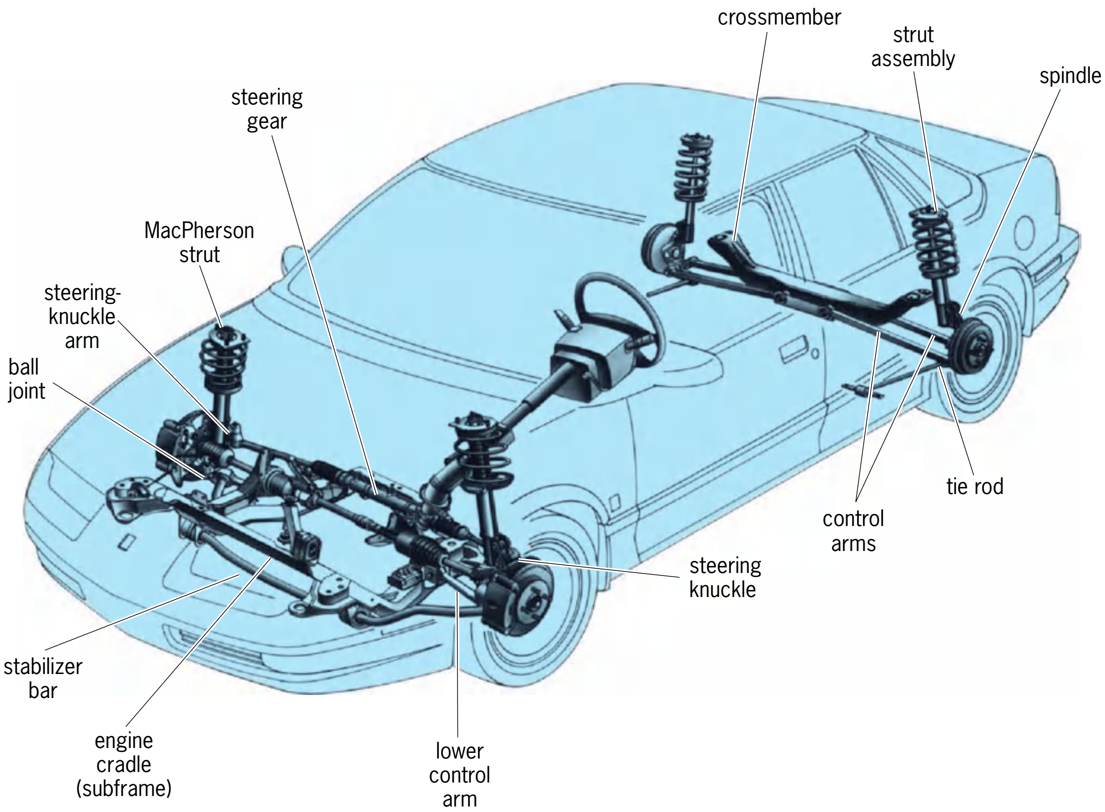
Shock absorber. A telescoping hydraulic damper, known as a shock absorber, is mounted separately or in the strut at each wheel to restrain spring movement and prevent prolonged spring oscillations. The shock absorber contains a piston that moves in a cylinder as the wheel moves up and down with respect to the vehicle body or frame. As the piston moves, it forces a fluid through an orifice, imposing a restraint on the spring. Spring-loaded valves open to permit quicker flow of the fluid if fluid pressure rises high enough, as it may when rapid wheel movements take place. Most automotive vehicles use gas-filled shock absorbers in which the air space above the fluid is filled with a pressurized gas such as nitrogen. The gas pressure on the fluid reduces the creation of air bubbles and foaming. *See* SHOCK ABSORBER.

Electronic ride control. With electronic ride control or variable-damping suspension, the shock absorbers at all four wheels can be electronically adjusted by the driver for softer or firmer operation. When the driver presses the SOFT or FIRM switch, a control module signals a solenoid or small electric motor in the top of each shock absorber. This opens or closes orifices of various sizes in the shock-absorber piston so that the resistance to fluid flow changes within the shock absorber. The switch may have a position that allows the shock absorbers to automatically select the proper damping for the road and driving conditions. Some vehicles have an electronic stability-control system that can selectively apply individual wheel brakes to help control the vehicle during cornering and skidding. *See* AUTOMOTIVE BRAKE.

Front suspension. Most automotive vehicles have independent front suspension, usually using coil springs as part of either a short-arm long-arm or a MacPherson-strut suspension system. In the short-arm long-arm system, two unequal-length control (suspension) arms attach through ball joints to the top and bottom of the steering knuckle. Typically, the coil spring is mounted between the lower control arm and the underbody or frame.

A MacPherson-strut suspension (see illus.) combines a coil spring and shock absorber into a strut assembly that requires only a beam-type lower control arm. The angled ends of a stabilizer bar attach to the outer ends of the lower control arms, holding the arms in position while helping to control the longitudinal loads. The top of the strut attaches through an upper strut bearing to the vehicle underbody, allowing the strut to turn with the steering knuckle. Some vehicles have a modified MacPherson suspension, in which the spring is mounted between the lower control arm and the underbody, instead of on the strut. Other strut-type front suspensions have the spring on the strut, with upper and lower control arms used to position the steering knuckle.

Some vehicles with short-arm long-arm front suspension use either longitudinal or transverse torsion bars for the front springs. One end of the torsion bar is attached to the lower control arm, and the other end is anchored to the vehicle body or frame. As the tire and wheel move up and down, the torsion bar provides springing action by twisting about its long axis. Turning an adjustment bolt at one end of the



Front-wheel-drive car with MacPherson-strut front suspension and strut-type independent rear suspension. (*Saturn Corp*.)

torsion bar raises or lowers the vehicle ride height. *See* TORSION BAR.

Rear suspension. Most automobiles and many light trucks have coil springs at the rear. These may mount on the rear drive axle, on struts, or on various types of control or suspension arms in an independent suspension system. Some rear-drive vehicles have leaf springs at the rear. Others use transverse torsion bars.

Air suspension. The air-suspension system uses air springs at the front or rear of the vehicle, or at all four wheels. The air spring is a rubber cylinder or bag filled with compressed air which provides the springing action. As the load in the vehicle changes and the lower control arm moves up and down, a height sensor in the strut signals the change in position to the system controller. The controller then opens a valve in the air bag to add or release air. An electric air compressor on the vehicle connects to the valve and keeps the air springs properly inflated. The system controls spring rates and provides automatic height and level control.

Active suspension. An active suspension system uses computer-controlled hydraulic actuators (instead of springs and shock absorbers) to keep each tire pushing against the road with a constant force. As a tire meets an irregular road surface, the computer varies the hydraulic pressure to the actuator, slightly raising or lowering the wheel so the tire maintains the same force against the road. This action keeps the vehicle level while providing the best possible ride and handling. Since most of the tire-and-wheel movement is absorbed in the suspension, little shock and vibration reach the vehicle body and its occupants. *See* HYDRAULIC ACTUATOR. Donald L. Anglin

Bibliography. J. C. Dixon, *Tires, Suspension and Handling*, 2d ed., Society of Automotive Engineers, Warrendale, 1996.

Automotive transmission

The device in the power train of a motor vehicle that provides different gear ratios between the engine and drive wheels, as well as neutral and reverse. An internal combustion engine develops relatively low torque at low speed and maximum torque at only one speed, with the crankshaft always rotating in the same direction. To meet the tractive-power demand of the vehicle, the transmission converts the engine speed and torque into an output speed and torque in the selected direction for the final drive. This arrangement permits a smaller engine to provide acceptable performance and fuel economy while moving the vehicle from standstill to maximum speed. The transmission may be a separate unit as in front-engine rear-drive vehicles or may be combined with the drive axle to form a transaxle as in most front-drive vehicles. *See* AUTOMOBILE; AUTOMOTIVE DRIVE AXLE; DIFFERENTIAL.

The two general classifications are manual transmissions that the driver shifts by hand after disengaging the foot-operated clutch, and automatic transmissions that shift with no action by the driver.

However, manual transmissions can have a clutch that is automatically disengaged by an actuator when the driver moves the shift lever, and automatic transmissions can have manual-shift capability which allows the driver to select the shift to the next lower or higher gear ratio by movement of the shift lever. *See* CLUTCH.

Automotive manual transmissions and transaxles typically have four, five, or six forward speeds. Manual transmissions in trucks can have up to 16 forward speeds. However, more than 85% of new cars manufactured in the United States have an automatic transmission or transaxle that has either four or five forward speeds. Both manual and automatic transmissions and transaxles usually have overdrive, a top forward gear ratio that causes the output shaft to overdrive or turn faster than the input shaft. In overdrive, the engine can run slower for improved fuel economy, reduced engine wear, and quieter engine operation while maintaining the same vehicle speed, but with reduced hill-climbing ability and acceleration. Some vehicles have a continuously variable transmission (CVT), which provides an infinite number of gear ratios. In one design, engine torque is transmitted through a belt that runs between two variable-diameter pulleys. By varying the ratio between the pulleys, the engine can operate at its optimum speed more of the time. *See* TRUCK.

Manual transmission. The manual transmission is an assembly of gears, shafts, and related parts contained in a metal case or gearbox partially filled with lubricant. The transmission input shaft connects through the clutch and flywheel to the engine crankshaft (**Fig. 1**). The transmission output shaft connects through a driveshaft to the final-drive gearing in the drive axle. To get the vehicle into motion, reduction or underdrive gearing in the transmission allows the engine crankshaft to turn fast while the drive wheels turn much more slowly but with greatly increased torque. As the vehicle accelerates, and less torque and more speed are needed, the driver shifts the transmission into successively lower numerical gear ratios, known as higher gears. In a typical five-speed manual transmission, gear ratios are approximately 3.35:1 for first gear, 2:1 for second gear, 1.35:1 for third gear, 1:1 (direct drive) for fourth gear, and 0.75:1 (overdrive) for fifth gear. Most transmissions with four or more forward speeds are operated by a floor-mounted shift lever. *See* GEAR; GEAR TRAIN.

Automotive manual transmissions are synchromesh (synchronized constant-mesh) transmissions. The teeth on the forward gears (and sometimes reverse) run in constant mesh, with no gear actually sliding in or out of mesh when a shift is made. Each gear is allowed to turn freely, or is locked to its shaft (usually the output shaft), by movement of the sleeve on a synchronizer clutch. The sleeve fits around the outside of the synchronizer hub, which is splined to the shaft.

When the driver moves the shift lever, the synchronizer clutch equalizes the speeds of the gear and the shaft. Then the synchronizer sleeve can slip over a smaller ring of external teeth which are formed around the face of the gear, locking that gear, through the synchronizer hub, to the shaft. This synchronizing action prevents gear clash during shifting. The sleeve is moved from its center position on the synchronizer hub, or neutral, toward the front or rear of the transmission by a shift fork that is mechanically connected to the shift lever. An interlock device in the transmission prevents two gears from engaging at the same time.

Manual transaxle. Most manual transaxles have either five or six forward speeds. The transmission section of the transaxle is basically the same in function and operation as the manual transmission. However, the final-drive pinion gear is on the transaxle output shaft and meshes with the helical ring gear bolted to

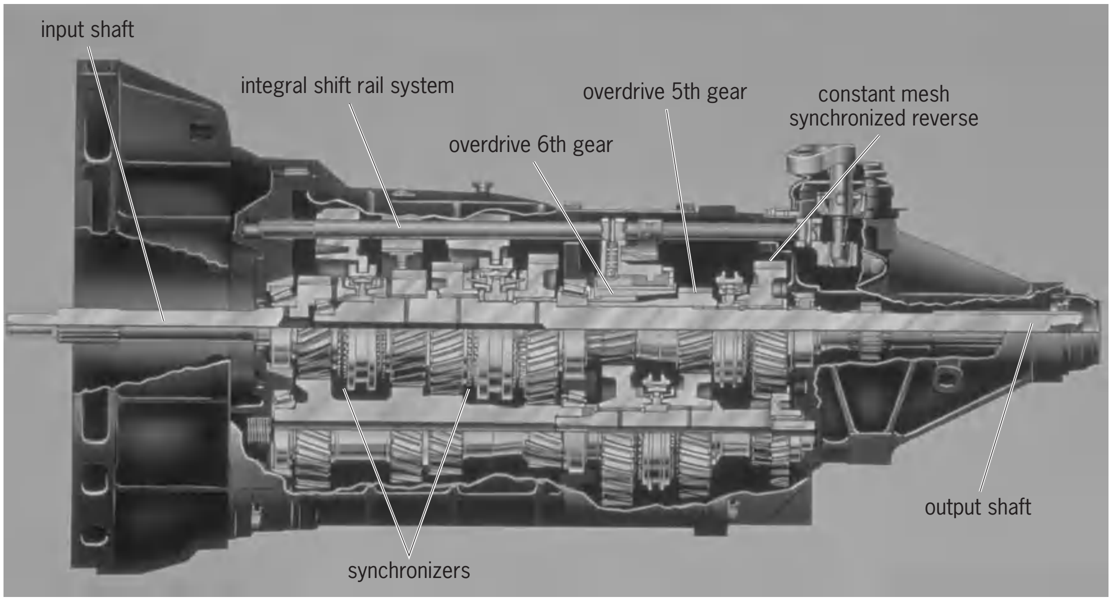


Fig. 1. Six-speed manual transmission for a rear-drive car. (*Pontiac-GMC Division, General Motors Corp.*)

the differential case. As the ring gear rotates, the differential pinion gears drive the differential side gears, which rotate the axle half-shafts to turn the wheels.

Automatic transmission. The automatic transmission provides automatic control of drive-away, gear-ratio selection, and gear shifting through four or five forward speeds. A typical automotive automatic transmission includes a hydrodynamic three-element torque converter with locking clutch, a planetary-gear system that provides overdrive in fourth or higher gear, and a hydraulic or electrohydraulic control system. Shifts are made without loss of tractive power. *See* HYDRAULICS; HYDRODYNAMICS; PLANETARY GEAR TRAIN; TORQUE CONVERTER.

The hydraulic torque converter uses an impeller, a turbine, and a stator to automatically and continuously vary the torque between the input and output shafts. The torque converter also provides for drive-away, and for damping of the shock and vibration caused by gear shifts. The amount of torque multiplication varies, depending on the difference in speed between the impeller (input shaft) and the turbine (output shaft). Torque multiplication decreases from a maximum of about 2.6:1 with the vehicle not moving (turbine stalled) until the ratio approaches 1:1, or the coupling point, as vehicle speed increases. Power loss through the torque converter can be eliminated by use of a centrifugal or electronically-controlled torque-converter clutch to lock the turbine to the converter cover. The clutch is normally unlocked at low speeds, during acceleration, and by a downshift. *See* TORQUE.

The automatic transmission has two or more planetary gear sets, which may be arranged in various ways to provide three or four forward speeds. These gear ratios provide further torque multiplication for acceleration and heavy road loads. In the power flow, the planetary gear sets are positioned after the torque converter. The individual elements of each planetary gear set are controlled by multiple-disc friction clutches and brake bands, which are actuated by oil pressure from an engine-driven oil pump in the front of the transmission. Some elements are controlled mechanically by a one-way clutch.

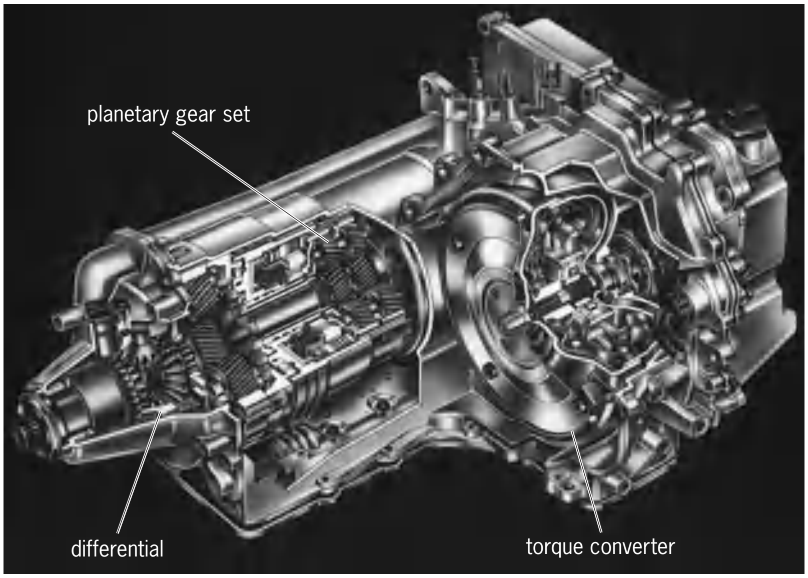


Fig. 2. Electronically controlled four-speed automatic transaxle for a front-drive car. (*Cadillac Division, General Motors Corp.*)

The oil pump supplies oil to fill the torque converter, hydraulic control system and valve body, and shift components. In addition, the oil lubricates and cools the transmission. The oil used in an automatic transmission is one of several types of automatic transmission fluid (ATF), usually dyed red for identification if leakage occurs. *See* LUBRICANT.

The typical automatic-transmission selector lever may be moved to Park, Reverse, Neutral, Drive, or Low (PRNDL). In a hydraulically controlled transmission, upshifts and downshifts are controlled by selector-lever position, engine vacuum, throttle position, and governor action. The governor is a speed-sensitive device that varies hydraulic pressure proportionally to output-shaft speed. *See* GOVERNOR.

Drive is the normal operating range, in which shifts occur primarily as a function of engine load and road speed. A vacuum modulator connected to engine intake-manifold vacuum, or the throttle position, allows the transmission to sense engine load. The governor senses road speed. With the selector lever in Drive and the vehicle at a standstill, the transmission starts out in first gear. As output-shaft speed increases, the governor allows increasing hydraulic pressure to the control-valve assembly. This causes spool valves to open and close oil passages to the hydraulic actuators and clutches, shifting the transmission to second gear. As vehicle speed increases, this basic sequence repeats until the transmission is in normal drive gear—direct drive (third) or overdrive (fourth or fifth). When the driver fully depresses the accelerator pedal, a wide-open-throttle forced downshift usually occurs for increased power or acceleration. *See* HYDRAULIC ACTUATOR.

In an electronically controlled automatic transmission, sensors monitor operating conditions such as vehicle speed, engine load, coolant temperature, and transmission oil temperature. This information is sent to the transmission controller or power-train control module (PCM). The controller decides when (shift timing) and how (shift quality) to make the shift. Signals are sent to electric shift solenoids on the control-valve assembly. The solenoids then open or close fluid passages, which send or release fluid pressure to the clutches and band servos controlling the planetary gear sets. Applying the brake pedal signals the controller to unlock the torque-converter clutch. *See* SOLENOID (ELECTRICITY).

The electronic controls are adaptive, automatically compensating for changes in engine or friction-element torque. Shift quality is not affected by a poorly performing engine. Under normal conditions, the controller "learns" to optimize shift quality during the first few shifts.

Some automatic transmissions and transaxles are similar in construction to manual transmissions and transaxles. Instead of planetary gears and a single centerline, two parallel shafts with spur and helical gears are used. Synchronizers are replaced with hydraulically-operated multiple-disc clutches that are electronically controlled.

Automatic transaxle. Front drive vehicles usually have a four-speed or five-speed overdrive automatic transaxle with locking torque-converter clutch (**Fig. 2**). The final drive and differential arrangement are basically the same as in a manual transaxle. The transmission section is similar in construction and operation to the automatic transmission. Automatic transaxles may be hydraulically or electronically controlled.

To reduce overall length, many automatic transaxles have the torque converter and oil pump offset from the input shaft. Power is transferred from the torque-converter output shaft through sprockets and chain to the transaxle input shaft.

Donald L. Anglin

Bibliography. *Design Practices: Passenger Car Automatic Transmissions*, 3d ed., Society of Automotive Engineers, 1994; Robert Bosch GmbH, *Automotive Handbook*, 4th ed., 1996.

Autonomic nervous system

The part of the nervous system that controls visceral functions of the body. This system innervates smooth and cardiac muscle and the glands, and regulates visceral processes including those associated with cardiovascular activity, digestion, metabolism, and thermoregulation. The autonomic nervous system functions primarily at a subconscious level. It is traditionally partitioned into the sympathetic system and the parasympathetic system, based on the region of the brain or spinal cord in which the autonomic nerves have their origin. The sympathetic system is defined by the autonomic fibers that exit thoracic and lumbar segments of the spinal cord. The parasympathetic system is defined by the autonomic fibers that either exit the brainstem via the cranial nerves or exit the sacral segments of the spinal cord. *See* PARASYMPATHETIC NERVOUS SYSTEM; SYMPATHETIC NERVOUS SYSTEM.

The defining features of the autonomic nervous system were initially limited to motor fibers innervating glands and smooth and cardiac muscle. This arbitrary definition limited the autonomic nervous system to visceral efferent fibers and excluded the sensory fibers that accompany most visceral motor fibers. Although the definition is often expanded to include both peripheral and central structures (such as the hypothalamus), contemporary literature continues to define the autonomic nervous system solely as a motor system. This bias ignores the importance of the afferent pathways. Moreover, it confuses the study of the dynamic regulatory function of the autonomic nervous system, since the regulation of visceral state and the maintenance of homeostasis implicitly assume a feedback system with the necessary constituents of motor, sensory, and regulatory components. Thus, from a functional perspective, the autonomic nervous system includes afferent pathways conveying information regarding the visceral organs and the brain areas (such as the medulla and the hypothalamus) that interpret the afferent feedback and exert control over the motor output back to the visceral organs.

Phylogeny provides insights into the functional and anatomical organization of the autonomic nervous system. The autonomic nervous system may be phylogenetically organized by proceeding from primitive structures that conserve metabolic resources and regulate visceral homeostasis, to structures that mobilize for fight-flight behaviors, and finally to structures found only in mammals that promote social and emotional behavior. Paralleling these functional shifts are increases in the interaction between visceral and somatic neurons and increases in influences from higher brain structures. *See* HOMEOSTASIS; NERVOUS SYSTEM (VERTEBRATE).

Stephen W. Porges

Bibliography. J. N. Langley, *The Autonomic Nervous System*, Heffer and Sons, Cambridge, 1921; S. W. Porges, Love: An emergent property of the mammalian autonomic nervous system, *Psychoneuroendocrinology*, 23:837–861, 1998.

Autopilot

An automatic means for steering an aircraft or other vehicle. The original use of an autopilot, or automatic pilot, was to provide pilot relief during cruise modes. Autopilots now perform functions more rapidly and with greater precision than the human pilot. The functions, designs, and uses of autopilots vary widely depending on the type of vehicle. In addition to controlling various types of aircraft and spacecraft, autopilots are used to control ships or sea-based vehicles and in some cases land-based vehicles. This article discusses autopilots used in aircraft and space vehicles.

Development. Early autopilots were developed with very little theoretical foundation and were based primarily on the ingenuity of such aviators as L. Sperry, who flew the first automatically controlled aircraft in 1914. His designs evolved through the years to the point where a three-axis autopilot was employed in several commercial aircraft in the late 1930s. In 1933 W. Post used a three-axis autopilot in his around-the-world flight. The system included gyros with pneumatic pickoffs and three-axis control with proportional hydraulic servos. During World War II the United States developed electric autopilot designs that could accept maneuver commands either from the pilot or from guidance sensors. Also at that time, two important features evolved that are still essential for the continued growth of automatic flight controls: feedback control theory and electronic computers.

Basic concepts. An autopilot is unique equipment in that it is expected to make the aircraft fly in the same manner as a highly trained, proficient pilot. It must provide smooth control and avoid sudden and erratic behavior. The intelligence for control must

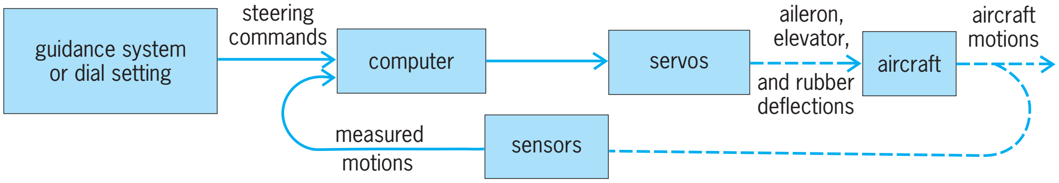

guidance system or dial setting
steering commands
computer
servos
aileron, elevator, and rubber deflections
aircraft
aircraft motions
measured motions
sensors

Fig. 1. Basic elements of an autopilot system.

come from sensors such as gyroscopes, accelerometers, altimeters, airspeed indicators, automatic navigators, and various types of radio-controlled data links. The autopilot supplies the necessary scale factors, dynamics (timing), and power to convert the sensor signals into control surface commands. These commands operate the normal aerodynamic controls of the aircraft. *See* ACCELEROMETER; AIRCRAFT INSTRUMENTATION; ALTIMETER; GYROSCOPE; INERTIAL GUIDANCE SYSTEM.

Autopilots come in varying degrees of sophistication. A simple attitude hold (wing leveler) just barely justifies the term autopilot, while a top-of-the-line system that automatically takes the aircraft from one location to another exceeds the normal capabilities of an autopilot. Sophisticated autopilots are no longer limited to military aircraft but are now common in commercial aircraft and are available for general aviation. In modern fly-by-wire aircraft the autopilot and the flight control system often reside together in the same digital computer, and it is difficult to separate their functions. These advanced systems provide the pilot relief functions plus help to stabilize the aircraft, protect the aircraft from undesirable maneuvers, and provide automatic landings (in some cases on a moving ship). Research aircraft are being tested with backup automatic control concepts that continue to control the aircraft even if the primary controls are damaged and no longer function. *See* FLIGHT CONTROLS.

Aircraft motion is usually sensed by a gyro, which transmits a signal to a computer (**Fig. 1**). The computer commands a control servo to produce aerodynamic forces to remove the sensed motion. The computer may be a complex digital computer, an analog computer (electrical or mechanical), or a simple summing amplifier, depending on the complexity of the autopilot. The control servo can be a hydraulically powered actuator or an electromechanical type of surface actuation. Signals can be added to the computer that supply altitude commands or steering commands. For a simple autopilot, the pitch loop controls the elevators and the roll loop controls the aileron. A directional loop controlling the rudder may be added to provide coordinated turns. *See* AILERON; AMPLIFIER; ANALOG COMPUTER; CONTROL SYSTEMS; DIGITAL COMPUTER; ELEVON; GUIDANCE SYSTEMS; SERVOMECHANISM.

Design. As autopilots become more complex, additional control loops are added. The pitch and roll axis of control must first be stabilized and controlled; then additional control loops can be added. Altitude is supplied by an altimeter; flight path commands come from a navigation system or another automatic system. Autopilots have provisions for pilot control or stick steering. A cruise control designed for fuel economy may command the aircraft to fly at the most economical altitude. An instrument landing system (ILS) or other automatic landing system can also command the autopilot. *See* AUTOMATIC LANDING SYSTEM; INSTRUMENT LANDING SYSTEM (ILS).

Any one of these commands could be selected to supply signals to the computer and are appropriately summed with one or more of the feedback loops. For longitudinal control (**Fig. 2**), pitch rate is the inner loop and is used for damping and stability. An accelerometer may also be used for damping and to limit the acceleration resulting from commands to the aircraft. Pitch angle is fed back to sum with flight-path commands so that the computer can calculate when the proper altitude is reached. The

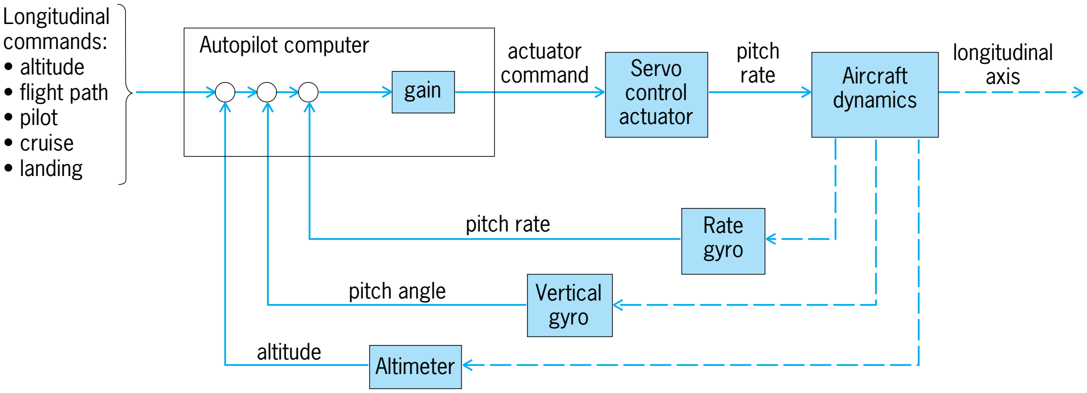

Longitudinal commands:
• altitude
• flight path
• pilot
• cruise
• landing
Autopilot computer
gain
actuator command
Servo control actuator
pitch rate
Aircraft dynamics
longitudinal axis
pitch rate
Rate gyro
pitch angle
Vertical gyro
altitude
Altimeter

Fig. 2. Typical longitudinal autopilot control loop.

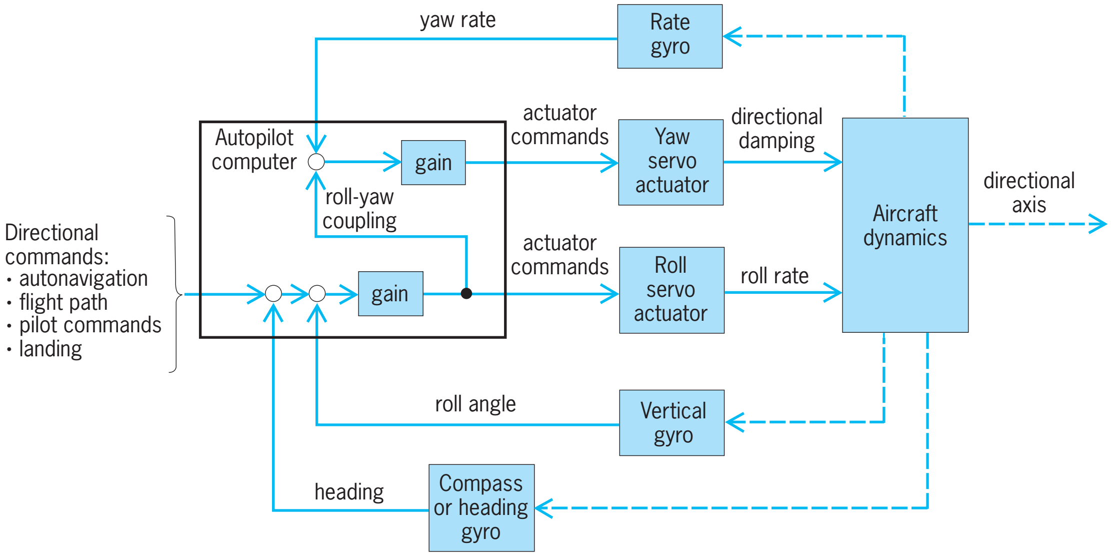


Fig. 3. Typical directional autopilot control loop.

altitude feedback is used to sum with altitude and cruise commands to permit the computer to calculate the proper flight altitude. After summing the appropriate signals, the computer produces an error signal that drives a servo actuator controlling a flight control surface. The control surface produces an aerodynamic force that causes a pitch rate to be generated on the aircraft. The appropriate feedback loop limits the pitch rate, pitch angle, and altitude to suitable values.

Directional control loops (**Fig. 3**) are slightly more complicated, since roll and yaw are coupled in the aircraft dynamics. Basically, the aircraft must be banked (roll) to produce a rate of change in the heading; therefore, roll and yaw must be controlled together. Normally a yaw rate gyro is used to control a vertical surface, such as a rudder, to provide directional damping. A lateral accelerometer may be used to help produce coordinated turns. In some aircraft an interconnecting link is provided between roll and yaw to aid in producing coordinated turns. Typical directional commands (Fig. 3) are similar to the longitudinal commands. The computer sums these signals to produce an error signal to command a roll control surface. This produces a roll rate until the proper roll angle is reached. The magnitude of bank angle is proportional to the required heading correction. As the heading error is removed, the aircraft returns to level.

Modern aircraft are becoming more complex and are using more control surfaces. Both the leading and trailing edges of the wing are now controllable, producing additional methods of pitch and roll control. The aircraft's horizontal surfaces move differentially, allowing the horizontal tail to be used for roll control. These additional control surfaces became practical with the development of fly-by-wire controls where hydraulic surface actuators are controlled with electrical signals. These signals are supplied from a computer, either analog or digital, which controls the aircraft.

Modern control theory. Automatic flight controls now use digital computers, which accept data from multiple sensors, to control an unstable aircraft with multiple feedback loops operating several aerodynamic control surfaces. With this type of system a control failure cannot be tolerated. The system is guarded against failures by having three or four channels operating in parallel with computers monitoring all the channels. When a failure is detected, the faulty unit is corrected or disconnected. This system monitoring adds further complexity to the system design because failures must be corrected without allowing undesirable inputs to the aircraft. Modern control theory enables systems of this type to be analyzed and designed by using computer synthesis. In modern control technique a desired performance criterion is selected and a unique design obtained. The performance of the system is then said to be optimal in terms of the defined criterion. On vehicles using automatic controls of this type the separation of flight controls and autopilot becomes obscure. For purposes of definition, the pilot relief functions are referred to as the autopilot, although the autopilot and flight controls are the same physical units of equipment. *See* OPTIMAL CONTROL THEORY.

Helicopter and VTOL autopilots. Helicopters and vertical takeoff and landing (VTOL) aircraft also use automatic control systems. Automatic control is especially helpful for hovering and low-speed flight since these aircraft are unstable at low speed. The control loops are similar to those discussed above except that the control servos control rotor blade angles on helicopters and vectoring thrust nozzles (or gas jets) on VTOL aircraft. A vertical axis of control may be used to control height. *See* HELICOPTER; VERTICAL TAKEOFF AND LANDING (VTOL).

Missile autopilots. The autopilots that control vehicles without flight crews use the same basic concepts as those described for aircraft autopilots. A missile autopilot is usually designed for use only once and therefore must be inexpensive. Since no crew is involved, fail-safety is not a requirement; however, reliability is still important. Also, with no crew aboard it is possible to perform maneuvers involving higher accelerations. Turns do not have to be coordinated; therefore, the vehicle does not need to bank to turn. Maneuvering is limited only by the structure of the vehicle. The autopilot controls the missile through control surfaces that are powered with hydraulic, pneumatic, or electrical actuators. *See* GUIDED MISSILE; MISSILE.

A special class of uncrewed aircraft is the remotely piloted vehicle (RPV); the pilot stays on the ground or in another aircraft, and actually flies the vehicle with a radio link that controls the autopilot. Remotely piloted vehicles have been used for flight research. In the HI-MAT program of the U.S. Air Force, most of the autopilot components, including the computers, were on the ground. A data link transmitted signals between the computers and the aircraft's control servos and sensors. *See* DRONE.

Launch-vehicle autopilots. The design of launch-vehicle autopilots is very similar to that of aircraft autopilots except for the difference in the devices used to obtain control forces. The rocket-propelled launch vehicles or boosters, used to place spacecraft into orbit and ballistic missiles into trajectories, are controlled by vectoring (controlling the direction) of the thrust. This can be achieved by swiveling the rocket engines or deflecting their thrust with vanes powered by large hydraulic actuators. At low speeds, aerodynamic forces are negligible and all control must come from the thrust. As the vehicle gains speed, the aerodynamic forces increase and provide some damping. Launch-vehicle attitude is sensed by stable platforms, and rate gyros and accelerometers are used for damping. Trajectory commands are provided by the vehicle guidance system. *See* ROCKET PROPULSION.

Space-vehicle attitude control. Attitude-control systems are required in satellites and other space vehicles to maintain correct orientation. In the absence of an atmosphere there are no aerodynamic forces to use for control or to produce disturbing torques. The required control torques are several orders of magnitude lower than those for an aircraft. Attitude disturbances come from irregularities in the gravity and magnetic fields of nearby planets, the Sun's radiation pressure, and impacts of small meteorites.

The autopilot control-loop concept is very similar to that for aircraft. Attitude sensors such as stable platforms, gyros, and accelerometers are used. The principal position sensors are optical, and include star trackers, sun sensors, and horizon sensors. Active attitude control is obtained by optically establishing the pointing direction and measuring the error signal between the pointing desired and the actual vehicle position. The computer applies corrective torques on the vehicle by employing pairs of gas jets. Control by gas jets produces large forces, but gas supplies last only a few days or weeks. Thus gas jets are used typically on short-duration or crewed missions. Where small forces are required, they may be produced by inertial reaction against gyros or wheels, by solar radiation pressure, or by reaction with local gravity or magnetic fields. These devices are useful on missions lasting many months or years. *See* STAR TRACKER.

Passive attitude control is sometimes possible for uncrewed long-duration parts of complex missions and is then preferred because of its higher long-term reliability. The whole vehicle may be spun like a gyro; its spin axis then tends to remain fixed in space. The space vehicle may be designed with a suitable inertia distribution, for example, by deploying long booms that tend to be attracted to the gravity field of the Earth. This system is undamped and requires some form of inertia damping. Another type of passive attitude control is the use of vanes mounted on short booms that react to the photon pressure of the Sun and keep the vehicle pointing at the Sun. This system also requires damping. A system of this type was used on a *Mariner* Mars probe in 1964. *See* SPACE NAVIGATION AND GUIDANCE; SPACECRAFT STRUCTURE.

Lloyd L. Kohnhorst

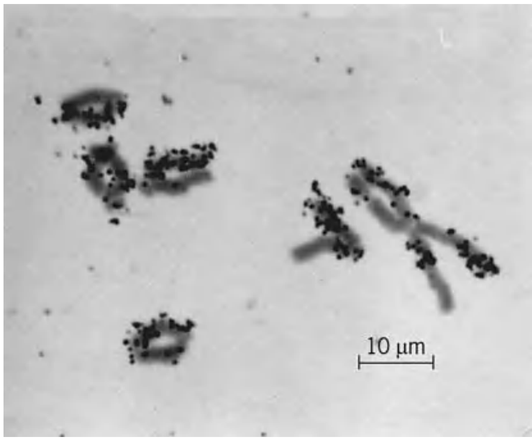
Bibliography. J. H. Blakelock, *Automatic Control of Aircraft and Missiles*, 2d ed., 1991; J. J. D'Azzo and C. H. Houpis, *Linear Control System Analysis and Design*, 3d ed., 1988; P. Garrison, *Autopilots, Flight Directors, and Flight Control Systems*, 1985; N. Radovcich, Active control in tomorrow's marketplace, *Aerosp. Amer.*, 26(8):14–22, August 1988.

Autoradiography

A photographic technique used to localize a radioactive substance within a solid specimen; also known as radioautography.

Procedure. A photographic emulsion is placed in contact with the object to be tested and is left for several hours, days, or weeks, depending on the suspected concentration of the radioactive material to be measured. The emulsion, which is a gel containing silver halide, is then developed, fixed, and washed as in the usual photographic process. At sites where the emulsion was close enough to the radioactive substance, it appears dark because of the presence of silver grains (see **illus.**). When the number of grains is insufficient to darken the film to the unaided eye, the film may be examined with the aid of a microscope. The individual silver grains may then be seen. The pattern formed by the grains depends on the type of radiation and the nature of the photographic emulsion. Alpha particles produce short, straight rows or tracks of grains. Beta particles as well as x-rays and gamma rays, which affect film by producing beta particles, produce tortuous tracks whose lengths and grain densities depend on the energy of the beta particles. Low-energy particles produce shorter tracks with higher grain densities. Very low energy particles like those from tritium (hydrogen-3) may produce only a single grain very close to the site of decay.

Chromatographs are frequently used to prepare autoradiographs. X-ray film is applied in close contact with the surface. The surface is protected by a thin



Photomicrograph of the chromosomes of a plant root cell and its autoradiogram. Black grains indicate the presence of tritium in certain parts of the chromosome.

plastic sheet in the case of wet gels. Such gels have been used extensively for analysis of nucleic acids and are the basis for the widely used techniques for the determination of the nucleotide sequences of nucleic acids. Of course, all molecules must have been made radioactive, usually by attaching a radioisotope or synthesizing the smaller molecules with a radioisotope such as phosphorus-32 or carbon-14 before their separation and use for autoradiography. *See* CHROMATOGRAPHY; NUCLEIC ACID; RADIOACTIVE TRACER; RADIOISOTOPE (BIOLOGY).

The principal advantage of autoradiography is that emulsions can be stored for weeks or months in the cold to reduce fading of the latent image and very small amounts of radioactivity can be detected. Even greater enhancement is possible by the use of scintillators which absorb the energy of radioactivity and emit light. At low temperatures (-70°F or -56°C) with tritium, the enhancement can be 60-fold or greater. The enhancement is useful in autoradiography of chromatograms containing tritium but is not desirable where high resolution is required. *See* BETA PARTICLES; GAMMA RAYS; X-RAYS.

Resolution. Autoradiograms (the picture obtained by autoradiography) can be made with many objects which can be brought in contact with photographic film. The resolution, that is, the precision with which the radioactive substances can be located, is good only when the specimen is flat and in close contact with the film or photographic plate. Whole plants, animals, rocks, or any other objects may be used, but the flat, cut surface of the object provides better contact.

The precision with which the source of radioactivity can be localized (resolution) is dependent on the nearness of the emulsion to the specimen. The highest resolution is obtained by embedding the biological specimen in the emulsion. Flattened cells or thin sections of cells on glass slides can be dipped in melted emulsion, or coated with thin strips of wet emulsion which is then dried before exposure. For very high resolution with the electron microscope, the grids containing the specimen are dipped in dilute melted emulsion, or a loop of emulsion is lowered over the grids. All procedures must be carried out in very low intensity red light or complete darkness. For autoradiography of chromatograms or gels produced by electrophoresis, the specimens are dried and placed in contact with a sheet of x-ray film, with or without a thin protective sheet between. The protective sheet may contain scintillators to enhance the image for specimens that emit low-energy beta particles. Under some conditions it may be possible to get the scintillators into the specimen to improve efficiency, but if the materials are embedded in gels, the scintillators penetrate poorly. Autoradiograms are useful in detecting deoxyribonucleic acid (DNA) for sequencing studies.

Electron microscopy. The technique has been extended to the location of radioactive molecules in specimens photographed with the electron microscope. The emulsion with very small crystals of silver salts is spread on a thin section of the type used for electron microscopy, somewhat as in the dipping technique described. To coat the section, a loop of wire containing a thin film of liquid emulsion is lowered onto the specimen mounted on the electron microscope grid. After drying, the processing is similar to that described for the light microscope except that development is regulated so that the developed silver grains are small enough not to obscure the tiny specimens which are being photographed in the electron microscope. *See* ELECTRON MICROSCOPE.

Applications. Autoradiography can be used to detect, and measure semiquantitatively, the radioactive materials in almost any object that can be placed in contact with film or photographic emulsion in some form. However, in biological research the object may be (1) a whole plant or animal that can be flattened against a film; (2) the cut surface of a plant or animal, or one of its organs; (3) thin sections of tissues or cells; (4) squashed or otherwise flattened cells; (5) surface films produced by spreading on water the protein monolayers containing DNA or ribonucleic acid (RNA) that are picked up on grids for electron microscopy; (6) sheets of paper or other materials on which radioactive substances have been separated by chromatography or electrophoresis; or (7) acrylamide gels in which DNA, RNA, or proteins have been separated by electrophoresis.

High-resolution autoradiograms remain a powerful tool for detecting and measuring small amounts of radioactivity that cannot be detected in any other way. Most radioactive isotopes produce radiations that penetrate the emulsions deeply enough to produce a somewhat out-of-focus picture, but tritium (3-hydrogen) is an exception. Its very low energy beta particles are stopped very close to the site of decay in emulsions, and when the beta particles are in a part of a molecule such as DNA, the site of the molecule can be visualized with high resolution. Tritium-thymidine of very high specific activity is used to label viral and bacterial DNAs so that individual molecules (chromosomes) can be visualized with a light microscope in emulsions by grain patterns. Similar autoradiograms of fragments of DNA from chromosomes of higher cells are used to show patterns of replication. *See* RADIOGRAPHY.

J. Herbert Taylor

Bibliography. J. R. Baker, *Autoradiography: A Comprehensive Overview*, 1989; T. Maniatis et al., *Molecular Cloning: A Laboratory Manual* 2d ed., 1989; M. A. Williams, *Autoradiography and Immunochemistry*, 1978; M. A. Williams (ed.), *Practical Methods in Electron Microscopy: Autoradiography and Immunochemistry*, vol. 6, 1978.

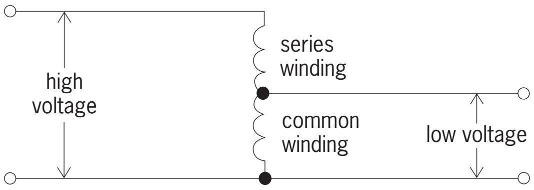
Autotransformer

A special form of transformer having one winding, a portion of which is common to both the primary and the secondary circuits. The current in the high-voltage circuit flows through the series and common windings (see **illus.**). *See* WINDINGS IN ELECTRIC MACHINERY.

The current in the low-voltage circuit flows through the common winding and adds vectorially to the current in the high-voltage circuit to give the common winding current. Thus, an electrical connection exists between high-voltage and low-voltage windings. Because of this sharing of parts of the winding, an autotransformer having the same kilovolt-ampere output rating is generally smaller in weight and dimensions than a two-winding transformer. The equivalent size of a two-winding autotransformer without taps is given by the co-ratio times the kilovolt-ampere output, where the co-ratio equals $(HV - LV)/HV$. When the co-ratio is small, that is, when the high-voltage and low-voltage magnitudes are close together, the cost advantage in favor of an autotransformer is large. As the co-ratio becomes large, the equivalent size and the cost of an autotransformer approach that of a normal transformer.

One possible disadvantage of autotransformers is that the windings are not insulated from each other and that the autotransformer provides no isolation of the primary and secondary circuits.

Types. Autotransformers of large sizes are used for interconnecting high-voltage power systems. They are used in small sizes for intermittent-duty starting of motors. For this use the motor is connected for a short time to the common winding voltage, and then connected to the full line voltage. Small, variable-ratio autotransformers are used in testing and as components of other apparatus. Autotransformers are also used as the induction coil in the ignition system of spark-ignition engines. *See* IGNITION SYSTEM.



Typical autotransformer circuit.

Characteristics. Because of their smaller equivalent size, autotransformers generally have lower no-load loss, exciting current, and load loss than the corresponding transformers. In addition, the impedance and regulation normally are lower because of the connection between the two circuits.

Taps. Taps can be provided in the autotransformer to adjust the turns ratio. This provides control of the output voltage over the operating range of the transformer. These taps may be placed in the series winding or in the common winding. The problems of switching from tap to tap under load are similar to those encountered in tap changing on a two-winding transformer. The switching from tap to tap can be done without interruption of service.

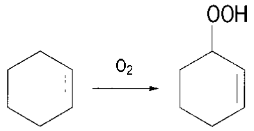
Small variable-ratio autotransformers used for testing have a brush contact which serves as a common line. This contact may be moved across the turns to give a common line voltage from approximately 100 to 0% of the high voltage. *See* TRANSFORMER.

J. R. Sutherland

Autoxidation

The slow, flameless combustion of materials by reaction with oxygen; sometimes spelled autooxidation. Autoxidation is important because it is a useful reaction for converting compounds to oxygenated derivatives, and also because it occurs in situations where it is not desired (as in the destructive cracking of the rubber in automobile tires).

Although virtually all types of organic materials can undergo air oxidation, certain types are particularly prone to autoxidation, including unsaturated compounds that have allylic hydrogens or benzylic hydrogens; these materials are converted to hydroperoxides by autoxidation. This reaction gives particularly good yields (98%) where there is a single allylic hydrogen, a hydrogen atom located on a carbon atom that is adjacent to a double bond, as in cyclohexene, reaction (1).



(1)

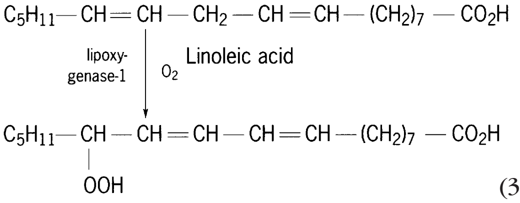
Certain saturated hydrocarbons with unique hydrogens can also be air-oxidized to good yields of hydroperoxides; one example is isobutane, which is oxidized to *tert*-butyl hydroperoxide in commercial practice, reaction (2).

$$(CH_3)_3CH \xrightarrow{O_2} (CH_3)_3—OOH \qquad (2)$$

Isobutane

tert-Butyl hydroperoxide

Both plants and animals contain lipoxygenase enzymes that convert unsaturated fatty acids containing a *cis-cis*-1,4-pentadienoic system to the allylic hydroperoxide. This reaction is analogous to nonenzymatic reactions (1) and (2). For example, lipoxygenase-1, an enzyme isolated from soybeans, converts linoleic acid primarily to the 13-hydroperoxide in which the original double bonds have shifted into conjugation, reaction (3). The enzymes that effect this conversion in animals are associated with the prostaglandin enzyme systems and normally have arachidonic acid, the 20-carbon, tetra-unsaturated fatty acid, as their substrate. The products of such reactions are hydroperoxy-eicosatetraenoic acids and are extremely important intermediates in the biosynthesis of potent physiological regulators called leukotrienes. *See* EICOSANOIDS; ENZYME.



$$\mathrm{C_5H_{11}{-}CH{=}CH{-}CH_2{-}CH{=}CH{-}(CH_2)_7{-}CO_2H}$$

Linoleic acid

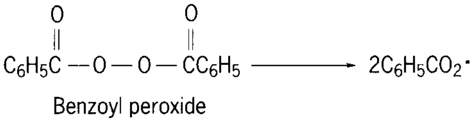
$$\text{lipoxy-genase-1}\ \Big\downarrow\ \mathrm{O_2}$$

$$\mathrm{C_5H_{11}{-}\underset{\displaystyle OOH}{\underset{|}{CH}}{-}CH{=}CH{-}CH{=}CH{-}(CH_2)_7{-}CO_2H} \qquad (3)$$

Reaction mechanism. Autoxidation is a free-radical chain process. Such reactions can be divided into three stages: initiation, propagation, and termination. In the initiation process, some event causes free radicals to be formed. For example, free radicals can be produced purposefully by the decomposition of a free-radical initiator, such as benzoyl peroxide, as shown in reaction (4). (The free radical is indicated by the dot representing its unpaired electron.) In some cases, initiation occurs by a process that is not well understood but is thought to be the spontaneous reaction of oxygen with a material with a readily abstractable hydrogen, reaction (5). This random process is responsible for

$$\mathrm{RH + O_2 \rightarrow R\cdot + HOO\cdot} \qquad (5)$$

the slow discoloration and decay of paper, plastic, and rubber materials left in the air for long periods. Destructive autoxidation processes also are initiated by pollutants such as those in smog.



$$\mathrm{C_6H_5\overset{\displaystyle O}{\overset{\|}{C}}{-}O{-}O{-}\overset{\displaystyle O}{\overset{\|}{C}}C_6H_5 \longrightarrow 2C_6H_5CO_2\cdot} \qquad (4)$$

Benzoyl peroxide

Once free radicals are formed, they react in a chain to convert the material to a hydroperoxide, ROOH. For example, reactions (6) and (7) show the autoxidation chain for the compound RH.

$$\mathrm{R\cdot + O_2 \longrightarrow ROO\cdot} \qquad (6)$$

$$\mathrm{ROO\cdot + RH \longrightarrow ROOH + R\cdot} \qquad (7)$$

The chain is ended by termination reactions in which free radicals collide and combine their odd electrons to form a new bond; for example, reaction (8) shows the termination reaction of two

$$\mathrm{R\cdot + R\cdot \longrightarrow R{-}R} \qquad (8)$$

R· radicals. *See* CHAIN REACTION (CHEMISTRY); FREE RADICAL; ORGANIC REACTION MECHANISM.

Significance. Autoxidation is a process of enormous economic impact, since all foods, plastics, gasolines, oils, rubber, and other materials that must be exposed to air undergo continuous destructive reactions of this type. The slow rancidity that develops in milk and cheese is an example, as is the off-flavor and toxic compounds that develop in fatty meats with long standing in air. The food and plastics industries are actively involved in research to develop acceptable and efficient antioxidants to slow down these natural autoxidation reactions. All plastics and rubber and most processed foods contain antioxidants to protect them against the attack of oxygen. *See* ANTIOXIDANT; FAT AND OIL (FOOD); FOOD MANUFACTURING; PLASTICS PROCESSING; RUBBER.

Autoxidation also causes a slow decay of plant and animal tissue, especially tissue that is subjected to unusually high oxidation environments. For example, emphysema, cataract formation, some types of liver diseases, some types of arthritis, certain chemically induced cancers, and even aging itself are believed to be due in part to autoxidation reactions of susceptible tissue such as the polyunsaturated fatty acids in fats and lipids. *See* COMBUSTION; OXIDATION PROCESS.

William A. Pryor

Bibliography. J. K. Kochi (ed.), *Free Radicals*, vol. 2, pp. 3–62, 1973; W. A. Pryor, *Free Radicals*, 1966; W. A. Pryor (ed.), *Free Radicals in Biology*, vol. 1, 1976, vol. 5, 1982; C. Walling, *Free Radicals in Solution*, 1957.

Auxin

Any of a group of organic compounds which, when applied in low concentration, are able to promote elongation growth of plant shoots excised from a growing region of a young seedling. These substances were the first plant hormones to be studied, and were given the generic name auxins, indicating an increase, by F. Kögl and A. J. Haagen-Smit in 1931. The ability to increase the rate of shoot elongation is a key to the designation of a synthetic or naturally occurring compound as an auxin. However, auxins, and the other plant hormones, influence a variety of plant processes during various stages of plant development.

Discovery of auxin. Studies in the nineteenth century showed that plants were sensitive to light and gravity, and that their growth pattern could be modified by changing the plant's orientation toward such stimuli. These responses were given the name tropisms. This work led directly to the discovery of auxins. In 1872 T. Ciesielski showed that in roots the control of growth and of curvature in response to gravity depends on a stimulus originating from the root tip. His experiments were followed up in 1880 by C. Darwin with experiments on shoots and studies on the transmitted stimulus in phototropism.

In the 1920s and 1930s several investigators began to study the chemical nature of the growth-promoting stimulus coming from plant shoot tips. In 1926 Fritz Went developed the first quantitative

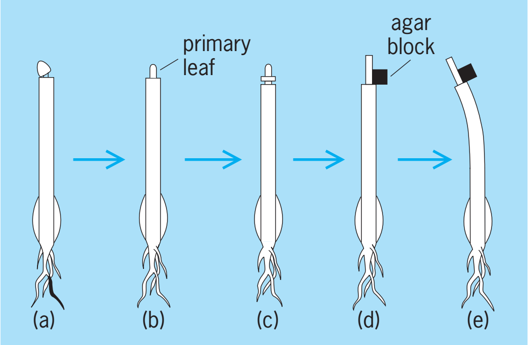


Fig. 1. Oat curvature test for auxin. (*a*) The tip is removed. (*b*) The plant is allowed to remain decapitated for 3 h. (*c*) A second short piece of stump is removed. (*d*) An agar block containing auxin is placed against the plant tissue. (*e*) After 90 min there is curvature due to asymmetric growth on two sides.

biological assay for auxins—the oat coleoptile curvature test (**Fig. 1**). Soon after, the structure and identity of the major naturally occurring auxin, indole-3-acetic acid (IAA), was described (**Fig. 2**). It was not until the late 1940s that the structure was confirmed by adequate chemical methods, and not until the 1970s that the structure of plant-derived indoleacetic acid was found by use of mass spectrometry. *See* MASS SPECTROMETRY; PLANT MOVEMENTS.

Occurrence and biosynthesis in plants. IAA has been found in almost all plant tissues that have been studied. It occurs in minute quantity, usually in less than micromolar amounts. This means that less than 1 oz (28 g) would be present in 150 tons (135 metric tons) of plant material. The understanding of auxins has been closely tied to the successful separation of such small amounts of active compounds from the great bulk of plant constituents. In particular, the development of chromatography has allowed auxins such as indoleacetic acid to be separated from plant material without chemical alteration, which is essential for subsequent biological and chemical tests. The coupling of chromatographic techniques with chemical and spectral analysis (in particular, mass spectrometry) has resulted in the characterization of a number of growth-promoting compounds and in the identification of several important metabolites of the known auxins. *See* CHROMATOGRAPHY.

Perhaps most significant has been the finding that auxins occur in plant tissue in several chemical forms. They may occur as the free active hormone (such as indoleacetic acid), and they may also be present in plant tissue as a number of so-called bound auxins (Fig. 2). The bound forms are auxins linked by a covalent chemical bond to some other chemical compound. These bound forms are thought to be important reserve forms which function to regulate the levels of free hormone in tissue, especially during certain developmental stages such as seedling growth. In addition, it is possible that the compound to which the auxin is linked (usually a sugar or an amino acid) is important in the transport of the hormone within the plant. Thus, auxin conjugates may serve as a type of biological "ZIP code" which determine the distribution of the hormone once it enters the transport system. Conjugation also renders the auxin immune to many of the enzymes which would normally degrade the free auxin. Thus, conjugation prevents further metabolism until the hormone has reached its target site and is released as the free active auxin. When synthetic auxins (for example, naphthalene acetic acid; 2,4-dichlorophenoxyacetic acid, or 2,4-D) are supplied to plant material, they will usually be conjugated to sugars or amino acids by the plant in a manner similar to that which occurs with the natural auxin.

In bean seeds, IAA has been shown to be conjugated to specific plant proteins. Early evidence also indicates that attachment of IAA to cell proteins may be a more general feature of auxin metabolism.

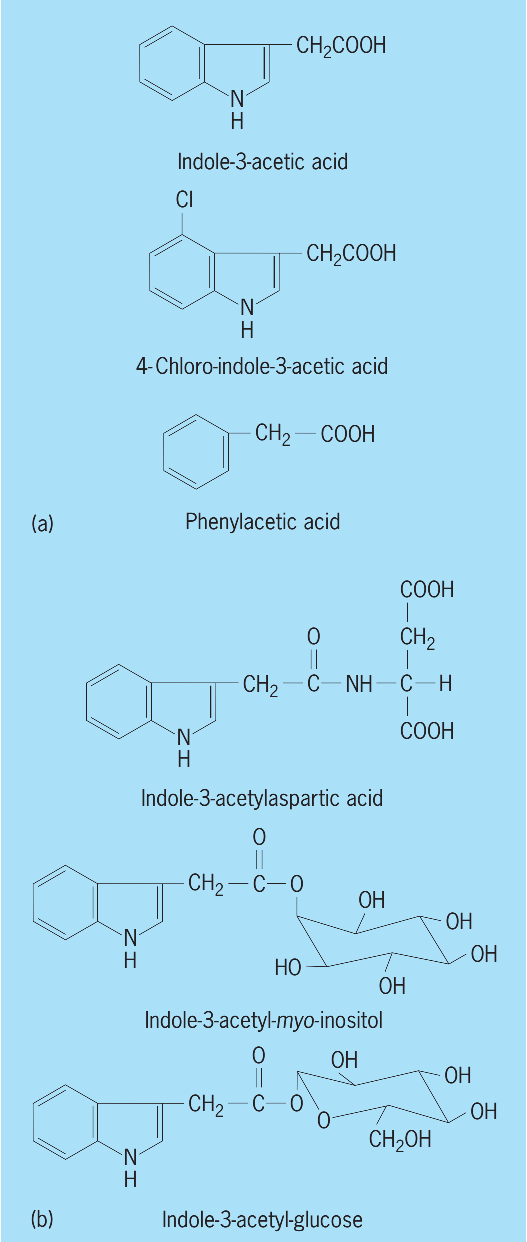


Fig. 2. Chemical structures of auxins. (*a*) Auxins isolated from plants. (*b*) Auxin conjugates found in plants.

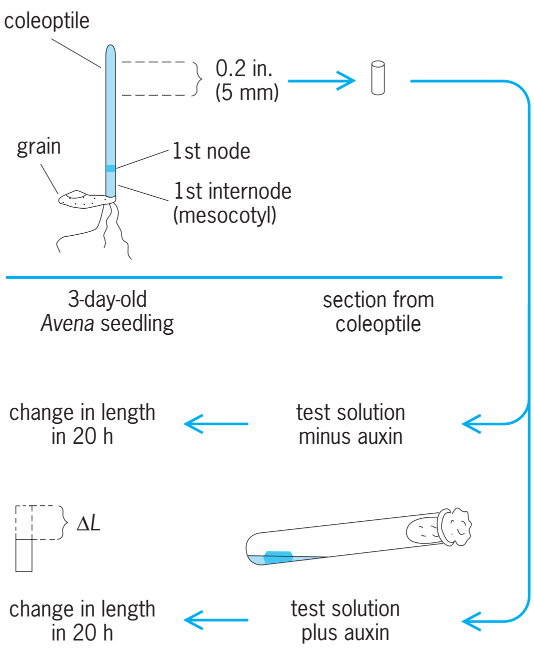


Fig. 3. Oat straight-growth bioassay.

While modifications of proteins after assembly by methylation and phosphorylation have been carefully studied and shown to be an important aspect of metabolic regulation in many organisms, the substitution of proteins by plant hormones is a novel discovery. The role played by these IAA-containing plant proteins is not yet known. Two general hypotheses have been proposed: First, IAA proteins represent an additional conjugate class with a role in controlling the availability of IAA in the plant or, second, IAA proteins are more directly involved in the mechanism by which auxins exert their regulatory function.

Biological activity. A number of experimental systems have been developed to study the auxin response. Most of these rely on the curvature of young seedlings treated on one side with the growth-active compound, or on the increase in growth of short tissue segments treated with the auxin. The use of the coleoptile of 3- or 4-day-old oat seedlings has most often been used, although most seedling stems show a pronounced response when treated with auxin and are thus suitable for such work. Several of the most common biological assays for auxin are shown in Fig. 1 and **Figs. 3** and **4**.

Auxin responses in plant tissue can be artificially divided into two groups, based on the time of their appearance. First, there are a number of very rapid responses which can be measured within minutes after the addition of auxin. Included in these responses are auxin-induced increases in protoplasmic streaming, cell elongation, and an increase in the acidity of the cell-wall free space. The second group of responses includes the long-term effects in which the observable response can be measured only after hours or even days of hormone treatment. Examples of these responses are auxin-induced increases in ribonucleic acid and protein synthesis, initiation of xylem differentiation, and an inhibition of lateral bud growth.

A major challenge to plant physiologists is understanding the mechanisms by which auxin elicits both these long- and short-term responses in plants and how the mechanisms are related. Short-term cell elongation responses to auxins are explained in part by the acid growth theory, which proposes that the first response to auxin is acidification of the cell wall region, resulting in the activation of wall-loosening enzymes involved in elongation growth. The general observations that support this theory are as follows: (1) Acid solutions will cause transient increases in the rate of cell elongation, both in intact tissue segments and in frozen-thawed segments placed under tension. (2) Auxin-sensitive tissues will acidify the cell wall region if given auxin treatment. (3) The rapid kinetics of auxin-induced cell elongation is similar to the kinetics for wall acidification.

Long-term responses to auxin have been attributed to direct interaction with plant deoxyribonucleic acid (DNA) resulting in increased transcription of specific messenger ribonucleic acids (mRNAs). Polysaccharide synthetases are proposed as one product of auxin-induced transcription resulting in increased production of cell wall material. The mechanism of auxin involvement with the plant genome has been theorized to be through binding of auxin, which passes through the plasma membrane to cytosol, with protein mediators. Such soluble auxin-binding proteins have been purified and shown to promote transcription.

Physiologically combining both short- and long-term effects of auxin has been the subject of speculation based on studies of intracellular signaling in cells involving calcium and metabolites of plasma membrane phospholipids known as phosphoinositides. These signals are released from storage into the cytosol by hormone–receptor interactions and effect growth-promoting cell processes. Theories for auxin-induced plant growth involving these signals are taken largely from animal cell systems, which have received more study; however, certain evidence obtained from plant cells has made such theories plausible and encourages research into these processes.

The generalized model for auxin action begins when a membrane receptor responds to auxin by releasing phosphoinositides from the pool of

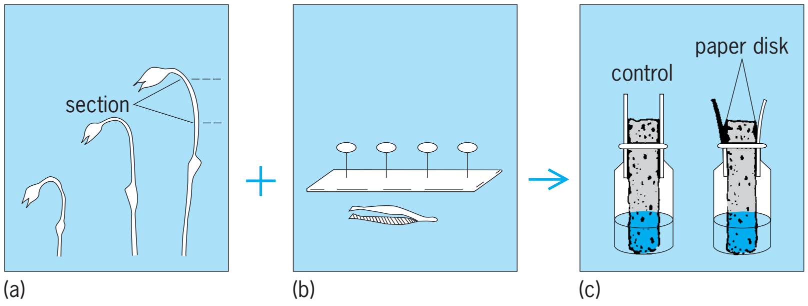
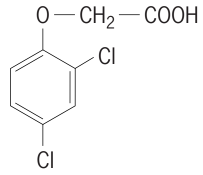


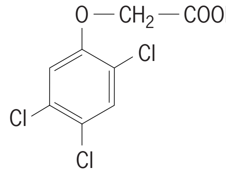
Fig. 4. Bean first-internode bioassay. (*a*) A section (1.5 in. or 4 cm) is cut from the stem between two nodes. (*b*) A test solution is applied to paper disks and placed against the side of the section. (*c*) After 3 h, bending is measured and compared with a control section.

membrane-bound phospholipids. These signals are carried through the cytosol and initiate release of calcium stored in vacuole and endoplasmic reticulum compartments possibly through their own receptor systems. This sudden increase in calcium initiates a response from at least two cell systems. First, the active removal of calcium from cytosol back to the vacuole is begun through the pumping of calcium in exchange for protons in the vacuole. These protons are in turn pumped out of the cytosol and into the cell-wall where acidification occurs. Second, the increase of calcium in cytosol acts to stimulate enzymes known as protein kinases which modify other proteins by phosphorylation. One protein that is modified by phosphorylation in this way is the proposed auxin-binding protein mediator that becomes receptive to auxin and can thus act on plant DNA to promote transcription of mRNAs that are critical for enzymes involved in cell growth. Evidence directly from plant systems for this model is as follows: (1) The release of calcium from subcellular compartments was enhanced by auxin or phosphoinositides in soybean and zucchini cell cultures. (2) The accumulation of calcium into vacuoles, that is, tonoplast vesicles, occurs quickly in carrot. (3) Protein phosphorylation occurs rapidly in the presence of calcium in zucchini and corn preparations. (4) Auxin enhances hydrolysis of the membrane-bound phospholipids into phosphoinositides that are vital in releasing calcium from intracellular storage in *Catharanthus*. This model proposes a direct relationship between the effects of auxin on cell-wall acidification and cell-wall matrix synthesis in growth-limiting epidermal cells. These effects can begin 30 min after auxin treatment. Continued research on the role of phosphoinositides and calcium in plants will undoubtedly test the connection between such signaling systems and the mechanism of action of auxin in plants.

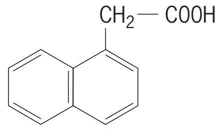
Uses. The most widespread agricultural use of auxins is for weed and vegetation control. Synthetic auxins, such as 2,4-D and 2,4,5-trichlorophenoxyacetic acid (2,4,5-T) and their homologs, are commonly used as herbicides (**Fig. 5**). It has been estimated that the yearly savings to farmers achieved by use of these compounds in one midwestern state alone would more than pay for all of the auxin research in the United States to date. Synthetic auxins have found practical use for other agricultural needs as well. For example, auxin applications are effective for floral thinning of overproductive orchard trees, such as olive, plum, orange, and persimmon. Synthetic auxins have also been used to prevent premature fruit drop and to improve fruit quality in tree crops, such as bananas, apricots, and plums. *See* HERBICIDE.



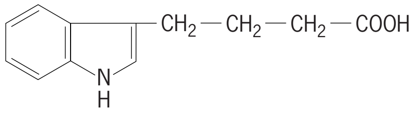
2,4-Dichlorophenoxyacetic acid (2,4-D)



2,4,5-Trichlorophenoxyacetic acid (2,4,5-T)



α-Naphthaleneacetic acid (NAA)



Indole-3-butyric acid (IBA)

Fig. 5. Structure of synthetic auxins.

The use of auxin for the propagation of plants from cuttings dates back to antiquity. Early farmers and gardeners used seed grain as an auxin source for such purposes. Today, commercial preparations, which usually contain the synthetic auxins indole-3-butyric acid or napthaleneacetic acid, are used. Auxin treatments have been used to enhance rooting of over 1000 different plant species, and have been applied on a practical scale to over 30 different species throughout the world. Auxin preparations designed to enhance root formation in cuttings are also available for home use.

The production of large numbers of genetically identical (clonal) plants is now possible by use of plant cell culture techniques. Additions of auxins as well as another type of plant hormone, cytokinin, are usually necessary for the growth of such cultures. Both the type and amount of these two phytohormones are important for regeneration of new intact plants from the cell cultures and seem to differ for each plant type. Over 300 different plant species, ranging from ferns to forest trees, are propagated on a commercial scale by cell culture techniques. In addition, ongoing research on the production of new varieties of agriculturally important plants, using the emerging techniques of molecular biology, rely on cell culture methods and a detailed knowledge of the role of auxins and other plant hormones in the developmental life of plants. *See* CYTOKININS; MOLECULAR BIOLOGY; PLANT GROWTH; PLANT HORMONES; TISSUE CULTURE.

Jerry D. Cohen; Bruce G. Baldi

Bibliography. M. Bopp, *Plant Growth Substances 1985*, 1986; D. A. Brummell and J. L. Hall, Rapid cellular responses to auxin and the regulation of growth, *Plant Cell Environ.*, 10:523–543, 1987; J. D. Cohen and R. S. Bandurski, Chemistry and physiology of the bound auxins, *Annu. Rev. Plant Physiol.*, 33:403–430, 1982; P. J. Davies, *Plant Hormones and Their Role in Plant Growth and Development*, 1987; C. Ettlinger and L. Lehle, Auxin induces

rapid changes in phosphatidylinositol metabolities, *Nature*, 331:176–178, 1988; U. Kutschera and W. R. Briggs, Rapid auxin-induced stimulation of cell wall synthesis in pea internodes, *Proc. Nat. Acad. Sci.*, 84:2747–2751, 1987.

Avalanche

In general, a large mass of snow, ice, rock, earth, or mud in rapid motion down a slope or over a precipice. In the English language, the term avalanche is reserved almost exclusively for snow avalanche. Minimal requirements for the occurrence of an avalanche are snow and an inclined surface, usually a mountainside. Most avalanches occur on slopes between 30 and 45°.

Types. Two basic types of avalanches are recognized according to snow cover conditions at the point of origin. A loose-snow avalanche originates at a point and propagates downhill by successively dislodging increasing numbers of poorly cohering snow grains, typically gaining width as movement continues downslope. This type of avalanche commonly involves only those snow layers near the surface. The release mechanism in this case is analogous to that which would occur in dry sand. The second type, the slab avalanche, occurs when a distinct cohesive snow layer breaks away as a unit and slides because it is poorly anchored to the snow or ground below. A clearly defined gliding surface as well as a lubricating layer may be identifiable at the base of the slab, but the meteorological conditions which create these layers are complex. The thickness and areal extent of the slab may vary greatly, and those slab avalanches with larger dimensions pose the greatest threat to life and property. Both loose and slab types may occur in dry or wet snow. Dry avalanches often entrain large amounts of air within the moving mass of snow and are thus referred to as powder avalanches. Velocities for dry-snow avalanches may exceed 150 mi/h (67 m/s). Wet-snow avalanches occur when liquid water is present in the snow cover at the point of origin. While the wet avalanches move at lower velocities, they often involve greater masses of snow and therefore significant destructive forces. Theoretical calculations and empirical evidence indicate the general range of maximum impact forces to be between 7 and 70 lb/in.^{2} (5 and 50 metric tons/m^2), with extreme values reaching 142 lb/in.^{2} (100 metric tons/m^2). It is frequently the wet-snow avalanche which damages the soil and vegetation cover.

Release mechanism. In the case of the loose avalanche, release mechanisms are primarily controlled by the angle of repose, while slab releases involve complex strength-stress problems. A release may occur simply as a result of the overloading of a slope during a single snowstorm and involve only snow which accumulated during that specific storm, or it may result from a sequence of meteorological events and involve snow layers comprising numerous precipitation episodes. In the latter case, large avalanches may not necessarily be restricted to storms with large amounts of precipitation, but can result from lesser amounts of precipitation falling on older snow layers underlain by an extremely weak structure.

Defense methods. Where snow avalanches constitute a hazard, that is, where they directly threaten human activities, various defense methods have evolved. Attempts are made to prevent the avalanche from occurring by artificial supporting structures or reforestation in the zone of origin. The direct impact of an avalanche can be avoided by construction of diversion structures, dams, sheds, or tunnels. Hazardous zones may be temporarily evacuated while avalanches are released artificially, most commonly by explosives. Finally, attempts are made to predict the occurrence of avalanches by studying relationships between meteorological and snow cover factors. In locations where development has yet to occur, zones of known or expected avalanche activity can be mapped, allowing planners to avoid such areas entirely. Avalanche hazard is small when compared with certain other natural hazards such as floods and tornadoes, but it continues to rise as the popularity of wintertime mountain recreation increases.

Richard L. Armstrong

Bibliography. *Annals of Glaciology*, vol. 26: *Papers from the International Symposium on Snow and Avalanches, Chamonix, France, 26–30 May 1997*, 1998; B. R. Armstrong and K. Williams, *The Avalanche Book*, rev. ed., 1992; S. C. Colbeck, *Dynamics of Snow and Ice Masses*, 1980; D. McClung and P. Schaerer, *The Avalanche Handbook*, 2d ed., 1993.

Aves

The class of animals consisting of the birds. Modern birds are characterized by being feathered, warm-blooded (homeothermic), and bipedal (two-legged) with the forelimb modified into a wing which, together with the tail feathers attached to the short tail, forms the flight mechanism, and by having a very high metabolic rate. Such a characterization, however, as with any group of vertebrates, holds for the living forms and most fossil members of this class, but is blurred by the early fossil record, which contains species with characteristics closer to those of the reptilian ancestors of birds. The feathers of birds are lightweight modifications of the outer skin possessing remarkable aerodynamic qualities. They serve not only as surfaces to generate lift and thrust, and as a streamlined outer surface of the body, but also as insulation to maintain high body temperatures. In addition, birds have lightweight, hollow bones; a well-developed air-sac system and flow-through lungs; a wishbone or furcula (fused clavicles); and a hand reduced to three digits (comparable to digits 2, 3, and 4 of the human hand). Birds have most likely evolved from an ancestor within the large group of ancient diapsid reptiles known as archosaurs (including alligators, snakes and lizards, and dinosaurs, among others). However, debate still centers on whether birds are derived from a basal archosaurian stock or arose later directly from the later and more derived theropod dinosaurs (carnivores such as *Allosaurus* and *Velociraptor*). The latter theory gained support by the discovery of fossils from the end of the Mesozoic Era ("Age of Reptiles") that some researchers reported to be feathered dinosaurs; however, other researchers think they were birds that had already became secondarily flightless and represent "Mesozoic kiwis."

Feathers. Feathers are unique to birds. These lightweight structures made of keratin are the most complex appendages produced by the skin of any vertebrate. The vanes of a feather are supported by a central shaft, or rachis, which is the structural backbone. They are composed of specialized filaments called barbs, which possess secondary, tiny filaments called barbules, or barbs. The barbs are bound together in "Velcro" fashion by small hooklets, or hamuli, located on the barbs. Body-contour feathers of many species have an associated smaller secondary shaft—an aftershaft—which results from the developmental process of the feather and which adds to the insulating property of the plumage. Feathers grow from specialized dermal follicles and are generally renewed once a year during the postbreeding molt. Body feathers of some birds are molted twice a year; the large wing feathers of some raptors are not replaced every year. In addition to the typical vaned form, feathers come in many other forms, such as down, power down, bristles, and filoplumes. However, the tuft of epidermal outgrowths, the beard present on the breast of male turkeys, are not feathers.

Body feathers in most birds are arranged in definite feather tracts, leaving large parts of the skin free of feathers. Muscles attaching to the base of the feather, the calamus, move the feathers, thereby allowing the bird, for example, to fluff the body feathers into a loose ball around its entire body for maximum insulation or to move courtship plumes in precise and complicated ways.

The wing feathers, the remiges, produce lift and forward thrust in flight. They are asymmetric, with a smaller outer vane and a larger inner vane, which are the building blocks of slotted wings. The bird wing comprises two main sets of flight feathers, the outer primary feathers which are attached to the hand, and the inner secondary feathers which are attached to the ulna. The several feathers attached to the moveable second finger constitute the ulna which can form a slot to prevent stalling at low speeds, similar to those in airplane wings.

Contour feathers provide smooth aerodynamic surfaces of the body, resulting in laminar airflow during flight. Typically, the body contour feathers have symmetrical vanes. Most of the vane is stiff and tightly bound, like flight feathers, and is known as the pennaceous portion. However, the part of the vane near the base, known as the plumaceous portion, lacks hooklets and is loosely bound. This basal portion can by fluffed up to trap body heat next to the skin. In warm conditions the body feathers can be flattened to allow heat to escape. Thus, feathers form an insulating plumage to cover the surface of the avian body.

Tail feathers (rectrices) resemble the flight feathers of the wing and provide lift in flight. The tail feathers of modern birds are attached to a vertically flattend bone known as the pygostyle, formed by a number of fused caudal vertebrae. The pygostyle (sometimes called the plowshare bone) also supports the uropygial or oil gland that provides a rich oil used to preen the feathers and to maintain their moistness and flexibility.

Not all feathers have stiff vanes. Typically, newly hatched birds are covered by a coat of plush down feathers, which are replaced by the adult contour feathers. Down feathers lack a central shaft but have long, loosely connected barbules that provide an insulating layer next to the skin.

In addition to flight and insulation, feathers can serve other functions, ranging from providing color patterns and structural forms serving for species recognition and courtship displays (including sound production by wing or tail feathers) to cryptic color patterns for protection (as is the case with the woodcock, which has feathers that blend with the dead leaves of the forest floor). Another unusual adaptation of feathers is seen in the sand grouse, a desert bird of Africa and Asia that makes long flights to a water hole to drink. The male sand grouse also obtains water for the young, storing it in his specialized flattened and coiled barbules on the contour feathers of the belly. This feature allows the male to hold water in flight and transport it to the nest, which may be some 20 mi (30 km) away. The young drink by squeezing the wet feathers in their bills. *See* FEATHER.

Beak. The original toothed jaws of ancestral birds have evolved into the light toothless beak in which the upper and lower jaws are covered by a horny rhamphotheca, and which may vary in texture from the strong beaks of predatory raptors and seed-eating parrots and finches to relatively soft beaks of shorebirds and ducks. Beaks have a great variety of adaptive forms, including the flesh-tearing hooked beaks of hawks and eagles, the filter-feeding straining beaks of flamingos and ducks, the fish-trapping beaks of pelicans, the climbing and nut-cracking beaks of parrots, the hammering beaks of woodpeckers, and the seed-eating beaks of finches. As in most vertebrates (many fish and reptiles), in addition to the usual moveable lower jaw, the upper jaw of all birds is kinetic, that is, moveable with respect to the brain case, which enables many important functional properties of the avian feeding apparatus in contrast to the akinetic jaw apparatus of mammals. The structure of the tongue also varies greatly depending on the diet of the species.

Birds have developed a muscular gizzard (also found in their relatives, crocodiles and dinosaurs) for grinding and processing food into small pieces. The grinding is often assisted by the addition of ingested gizzard stones. The ancient giant moas of New Zealand are preserved with masses of stones in the stomach region, and are known to have subsisted on a diet of leaves.

Wing. Birds are characterized as flying vertebrates with the forelimb modified as a wing with strong flight feathers. They also possess a large, strong pectoral girdle and bony sternum for the attachment of the large flight muscles and for the transfer of the body weight to the wings when the bird is flying. Considerable variation exists in the morphology of the wing (size and shape of the aerodynamic surface formed by the flight feathers) and of the entire muscle-bone system of the pectoral girdle and limb. But it is a mistake to conclude that all activities of birds are associated with flight; indeed for most birds, flight occupies a minor part of their daily activities. The important biological roles of flight in most birds are associated with quick escape from predators, reaching secure nesting places, larger separation of the nesting and foraging sites, and long-distance migration. Yet a few birds (such as swifts and terns) are able to spend many days or even weeks in the air, apparently being able to sleep during flight. And some birds (such as waterfowl and shorebirds) are able to make very long distance migration flights of up to 3000–4000 km (1800–2400 mi).

Hind limb. The leg and foot structures of birds differ greatly depending on their nonflying locomotion: from the long legs of wading birds to the strong, shorter legs of predators and aquatic birds to the almost rudimentary legs of the swifts and hummingbirds. The most primitive avian foot, found in the earliest definitively known bird, *Archaeopteryx*, is the anisodactyl perching arrangement of toes, also found in most modern tree-dwelling birds. Three toes point forward, and a reversed first toe, or hallux, opposes them in perching on a branch.

Other modifications include the zygodactyl feet (fourth toe reversed) of woodpeckers, cuckoos, and parrots, and the heterodactyl foot (second toe reversed) of the trogons with two forward and two rearward-pointing toes. The webbed feet of ducks and many other swimming birds have three forward-pointing toes united by a web that serves as a paddle. Members of the order Pelecaniformes have a foot in which all four toes are united by webbing, a totipalmate foot. Ostriches are unique among living birds in having a foot with only two toes.

Other distinctive features. The avian neck is long and flexible, consisting of many vertebrae with unique saddle-shaped articulating surfaces, allowing the bird to reach most parts of its plumage with its beak for preening. Modern flying birds have a well-developed pectoral girdle including a large, bony sternum with a keel, or carina, for the attachment of the large flight musculature. The pelvic girdle is large with the bones fused together with the synsacrum (a number of completely fused sacral vertebrae) to provide firm support for the bipedal hind leg. The ankle and foot bones are fused and elongated, so that the avian leg consists of three long segments—the femur, the tibiotarsus plus the reduced fibula, and the tarsometatarsus (the fused ankle and foot bones, which is equivalent to the human foot)—and finally the toes. Thus, birds walk on their toes, which are four in most birds but have been reduced to three, with the loss (usually) of the posterior hallux, or two (hallux and fourth) in the ostrich.

In many respects birds have more advanced sensory and physiological systems than mammals. They have a keen sense of vision (color vision based on four different cone cells, allowing color vision into the ultraviolet and fluorescent colors; the plumage of many birds looks very different to a bird than to a human) and hearing. The sense of smell (olfaction) is not particularly well developed, although some birds possess a good olfactory sense (for example, the turkey vulture, *Cathartes aura*). Birds have developed a small, rigid, flow-through lung using tubular air passages for a greater oxygen exchange surface than in mammals, and an extensive air-sac system serving as the pumping mechanism and as an internal system to lose excess heat. Nitrogenous wastes are lost in the form of uric acid, which requires much less water and a simpler kidney than in mammals.

Migration. Birds are found over the entire Earth. One of the most intriguing aspects of avian biology is the ability to migrate exceptional distances. Birds possess highly specialized directional senses for orientation, navigation, homing, and migration, including the ability to detect the Earth's magnetic field. These uncanny abilities permit birds to occupy widely separated wintering and nesting grounds, thus expanding their usable habitats. Some migrations, such as that of the Arctic tern, involve a circum-Atlantic route from Alaska to the South Pole. *See* FLIGHT; MIGRATORY BEHAVIOR.

Taxonomy. There are about 10,000 species of birds living today, of which more than 5500 belong to the order Passeriformes. Many avian species are particularly well known; however, the relationships of the higher categories of birds (orders, etc.) are still debated. Recognition and arrangement of avian orders and of their contained families has traditionally been based on study of morphological, and to a much lesser extent behavioral, features. Over the last three decades, much new information has been added using molecular features and especially with comparison of nucleotide sequences of nuclear and mitochondrial DNA. The number of avian orders, and the contained families and their arrangement are still controversial; even today, the relationships of these orders are almost unknown. Many authors advocate different arrangements. Because the situation is in flux, a fairly conservative system is used below (fossil groups are designated by a dagger). *See* NEOGNATHAE; NEORNITHES; VERTEBRATA; articles on the different avian orders.

Class Aves
  Subclass Sauriurae†
    Infraclass Archaeornithes†
      Order Archaeopterygiformes (Late Jurassic reptile-birds, *Archaeopteryx*)†
        Confuciusornithiformes (Lower Cretaceous, beaked reptile-birds, *Confuciusornis*)†

  - Infraclass Enantiornithes (archaic Mesozoic land birds)†
- Subclass Ornithuriae
  - Infraclass Neornithes (or Carinata)
    - Superorder Incertae Sedis
      - Order Hesperornithiformes (Cretaceous toothed divers, *Hesperornis*)†
      - Ichthyornithiformes (gull-like, Mesozoic toothed birds, *Ichthyornis*)†
      - Order Incertae Sedis (*Gansus*, *Chaoyangia*, etc., archaic modern-type birds)†
    - Superorder Neognathae (modern birds)
      - Order Struthioniformes (58 species)
      - Galliformes (chickens and allies, 282)
      - Anseriformes (waterfowl, 161)
      - Gastornithiformes (giant ground-birds, 2)†
      - Sphenisciformes (penguins, 17 species)
      - Procellariiformes (tube-nose seabirds, 114)
      - Pelecaniformes (pelicans and allies, 66)
      - Ciconiiformes (storks and allies, 119)
      - Falconiformes (hawks, eagles, and vultures, 309)
      - Gruiformes (rails, cranes and allies, 214)
      - Podicipediformes (grebes, 22)
      - Charadriiformes (shorebirds, gulls and allies, 349)
      - Phoenicopteriformes (flamingos, 5)
      - Gaviiformes (loons, 5)
      - Columbiformes (pigeons and doves, 332)
      - Psittaciformes (parrots, 360)
      - Coliiformes (mousebirds, 6)
      - Cuculiformes (cuckoos, 165)
      - Opisthocomiformes (hoatzin, 1)
      - Strigiformes (owls, barn owls, 173)
      - Caprimulgiformes (nightjars and allies, 116)
      - Apodiformes (hummingbirds and swifts, 425)
      - Trogoniformes (trogons, quetzals, 39)
      - Coraciiformes (kingfishers, bee-eaters and allies, 219)
      - Piciformes (woodpeckers and allies, 407)
      - Passeriformes (perching birds, songbirds, 5739)

Fossil record. The classification system presented above agrees with those presented in most major treatises on birds. The subclass Sauriurae contains the archaic birds of the Mesozoic Era, the Age of Reptiles, which includes the toothed fossil *Archaeopteryx*, or Urvogel, the oldest known bird (about 145 million years ago). Sauriurae also contains the subsequently discovered *Confuciusornis*, a beaked bird from Chinese deposits probably of Lower Cretaceous age, but possibly as recent as 120 million years ago. The first recognized specimen of *Archaeopteryx* was discovered in 1861 from the Solnhofen lithographic limestone of Late Jurassic age in Bavaria, and was named *A. lithographica* in reference to the rocks in which it was found. The crow-sized specimen from the bottom of a shallow salt-water lagoon was preserved with its wing and tail feathers in place, illustrating that it was indeed a bird and not a reptile. Another specimen was found in 1876 and became known as the Berlin specimen, as it is housed there. Often called an avian "Rosetta Stone," the Berlin specimen is beautifully preserved, with outstretched wings, and is often cited as the best example of an animal perfectly intermediate between two classes of vertebrates, in this case reptiles and birds. It fulfilled Darwins expectation that such forms should exist, and this specimen has played a large role in the debate on evolution over the years. The jaws of *Archaeopteryx* possess rounded teeth with large root crowns, unlike the flattened, recurved, serrated teeth of theropod dinosaurs. The wing feathers are typical of modern birds, showing that feathers remained essentially unchanged structurally in the 150 million years since *Archaeopteryx* lived. As in modern birds, the wing feathers are divided into primary and secondary flight feathers, which exhibit asymmetric vanes, indicative of aerodynamic function, probably gliding because of the overall weakness of the pectoral girdle suggesting small flight muscles. The wings end in three sharply clawed fingers, the claws being virtually identical to the pedal claws of tree-trunk-climbing birds such as woodpeckers. Because the curvature of the foot claws of *Archaeopteryx* is very similar to that of perching birds, it has been assumed that *Archaeopteryx* was a tree-dweller that used its clawed hands and feet to ascend tree trunks; its flying ability was mainly gliding, with some very limited flapping flight. There are now seven skeletal specimens of *Archaeopteryx*. *See* ARCHAEOPTERYX; ARCHAEORNITHES; REPTILIA.

A controversial fossil, *Protoavis texensis*, was described from the Late Jurassic of Texas from very fragmentary, crushed material. It does have clearly cervical vertebrae with the typical heterocelous (saddle-shaped) articulating surfaces that have been considered to be unique to birds. If this fossil is confirmed as an ancestral bird, it would push the known record of birds back nearly 100 million years and the origin of birds to the earliest members of the archosaurian radiation. But more and better material is needed before the relationships of this fossil can be determined with any assurance.

Confuciusornis sanctus, discovered in 1994 in China, is not as old as *Archaeopteryx* and could be as recent as 130–120 million years, but it shares many features with the Urvogel, including a wing with three sharply clawed fingers. It is clearly a perching bird like *Archaeopteryx* but has better-developed flight architecture and exhibits a typical avian beak without teeth. Many hundreds of specimens of

C. sanctus have been found, some with an elongated pair of narrow central tail feathers and presumed to be males.

During the past 3 decades a large diversity of Cretaceous fossil birds have been discovered in China, including some controversial fossils from the same deposits that produced *Confuciusornis*; these include two feathered but flightless creatures named *Protarchaeopteryx* and *Caudipteryx*. They are proclaimed to be feathered dinosaurs by Chinese and some western scientists, and said to prove the dinosaurian ancestry of birds; however, others believe that these are merely flightless Mesozoic birds and have little to do with avian origins. Despite their superficial resemblance to small dinosaurs, they show many of the features associated with flightlessness in modern birds. Other birds had become flightless early in avian history such as *Hesperornis* (Cretaceous) and *Diatryma* (Paleocene).

The Sauriurae contains a group known as the enantiornithines, or opposite birds. These archaic land birds of the Mesozoic are now known throughout the world from the Cretaceous Period, and are called "opposite birds" because their foot bones fuse in the opposite direction to that of modern birds. They also have a different type of formation of the triosseal canal in the pectoral girdle which is produced by the bones associated with the flight apparatus and through which passes the tendon of the elevator muscle of the wing. These archaic birds are particularly well known from the Lower Cretaceous of Spain and China. All opposite birds became extinct at the end of the Cretaceous Period.

Other Mesozoic birds included the ancient ornithurine birds more closely allied with the modern radiation of birds, among them such forms as the hesperornithiforms (*Hesperornis*, *Baptornis*, etc.), the Cretaceous toothed divers, which superficially resembled loons. They became extinct at the end of the Cretaceous along with their gull-like contemporaries, *Ichthyornis* and *Apatornis*. The Lower Cretaceous *Ambiortus* from Mongolia was a fully volant (capable of flight) ornithurine bird about the size of a pigeon, possessing a well-developed sternal keel and other features of the pectoral region typical of modern birds, which showed that true flying birds existed some 12 million years after the appearance of *Archaeopteryx*.

The Struthioniformes contain the living ratites—flightless birds such as the ostrich, rhea, emu, cassowary, and kiwi—as well as their South American chickenlike relatives, the tinamous, which are fully capable of flight. Recently, a group of fossil birds, the Lithornithidae, has been described from the Paleocene and Eocene of North America and Europe. These chickenlike forms were fully volant, and are closely related to the living tinamous and the ratites. They now are thought to be the ancestral stock that gave rise to the Struthioniformes. The more reasonable explanation for the disjunct distribution of living ratites is that their ancestors, the lithornithids, flew to remote parts of the world and gave rise to various lineages of flightless birds. The huge elephant birds lived on Madagascar contemporaneously with the native peoples, and probably became extinct in historic times. The same was true of New Zealand, where 11 or 12 species of large moas lived. They too survived until the arrival of Polynesians in the late thirteenth century and subsequently became extinct years ago. *See* RATITES; STRUTHIONIFORMES.

By the Eocene, approximately 50 million years ago, all the major orders of modern birds were present. By the Oligocene, most of the families were present, and by the Miocene, some genera of modern birds were well established.

Economic significance. Birds have a huge economic importance in terms of domesticated species, such as chickens and turkeys, and hunting. Today, however, the economics of birds for entertainment, such as birdwatching, ecotourism, and simply backyard feeding and watching, is far more important than for hunting. The interest of humans in observing birds is perhaps the major driving force in conservation efforts for the past century.

Walter J. Bock

Bibliography. W. J. Bock, Aves, in S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, vol. 2, 1982; J. Cracraft et al., Phylogenetic relationships among modern birds (Neornithes), pp. 468–489, in J. Cracraft and M. J. Donoghue (eds.), *Assembling the Tree of Life*, Oxford University Press, 2004; E. C. Dickinson (ed.), *The Howard and Moore Complete Checklist of the Birds of the World*, 3d ed., 2003; J. del Hoyo et al. (eds.), *Handbook of Birds of the World*, vols. 1–9, 1992–2004; A. Feduccia, *The Origin and Evolution of Birds*, 2d ed., 1999; F. B. Gill, *Ornithology*, 2d ed., 1995; D. P. Mindell (ed.), *Avian Molecular Evolution and Systematics*, Academic Press, 1997; N. S. Proctor and P. J. Lynch, *Manual of Ornithology: Avian Structure and Function*, 1993; C. G. Sibley and J. E. Ahlquist, *Phylogeny and Classification of Birds: A Study in Molecular Evolution*, 1990; C. G. Sibley and B. L. Monroe, *Distribution and Taxonomy of Birds of the World*, 1990.

Avian leukosis

A complex of several related and unrelated viruses (both C-type retroviruses and herpesviruses) that are collectively responsible for a variety of benign and malignant neoplasms in chickens and, to a lesser extent, in other avian species. Although most neoplasms observed in avian species are induced by viruses, there are some that are of unknown etiology.

Viruses. The neoplastic diseases induced by the leukosis-sarcoma group of retroviruses include lymphoid leukosis, myeloid or erythroid leukemias or solid tumors, tumors of connective tissue origin (for example, sarcomas, fibromas, and chondromas), epithelial carcinomas, and endothelial tumors. The many viral strains involved have similar physical and chemical characteristics and share a group-specific antigen; some can cause more than one type of neoplasm. The viruses are about 100 nanometers in diameter; have a core composed of ribonucleic acid; contain a reverse transcriptase; mature by budding from the cell membrane; and are divided into subgroups based on envelope glycoproteins. Some strains carry their own specific oncogenes that induce neoplasms within days or weeks. Others lack an oncogene and cause neoplasms less frequently and only after several months, probably by activating a specific cellular oncogene. *See* CANCER (MEDICINE); ONCOGENES; ONCOLOGY.

Diseases. Lymphoid leukosis is the most important of the leukosis sarcoma diseases. The lymphoid leukosis virus is transmitted vertically from hen to chick through the egg. Infection can result in leukotic neoplasms in various visceral organs following metastasis from primary tumors in the bursa of Fabricius. Horizontal transmission from bird to bird is inefficient, rarely causing tumors unless infection occurs within the first week. Large-scale transmission of the lymphoid leukosis virus can be eradicated by eliminating individual infected breeders.

Reticuloendotheliosis virus strains constitute another retrovirus group, unrelated to the lymphoid-sarcoma group, and also may carry a specific oncogene. They can cause a chronic neoplastic form of reticuloendotheliosis or other neoplasms in turkeys, chickens, ducks, geese, quail, and pheasants, and a runting disease has been seen in chickens after accidental contamination of vaccines with reticuloendotheliosis virus.

Marek's disease in chickens is caused by an oncogenic, cell-associated, lymphotrophic, highly contagious herpesvirus. Although the virus is biologically like lymphotrophic gamma herpesviruses, the structure of its genome is similar to alpha herpesviruses. The virus replicates fully only in the epithelial cells surrounding the feather shaft, and is released to the environment associated with dead cells and dander. Inhalation of the virus causes an active infection in lymphoid organs; after about 1 week, a latent infection develops in lymphocytes. T-cell lymphomas may develop within a few weeks or months, depending on age, genetic makeup, virus virulence, and other factors. Degenerative, inflammatory and lymphoproliferative lesions occur principally in the peripheral nerves (causing paralysis), lymphoid tissues, visceral organs, muscle, and skin. Eye involvement (gray eye) can cause blindness. The disease is of great economic importance in chickens, and several vaccines, injected at 1 day of age, have been in worldwide use since about 1970. The vaccines are prepared from attenuated strains of Marek's disease virus or of related, but naturally nononcogenic, herpes viruses isolated from turkeys or chickens. *See* ANIMAL VIRUS; TUMOR VIRUSES.

Bruce W. Calnek

Bibliography. B. W. Calnek et al. (eds.), *Diseases of Poultry* 10th ed., 1997.

Aviation

A general term including the science and technology of flight through the air. Aviation also applies to the mode of travel provided by aircraft as carriers of passengers and cargo, and as such is part of the total transportation system. Aviation also describes the employment of aircraft in such fields as military aviation. The world of the airplane, including the people who manufacture, market, and repair aircraft or who work in allied industries, is frequently spoken of as aviation. *See* AIRPLANE; MILITARY AIRCRAFT.

Aviation in modern usage differs from the gliding or soaring of unpowered aircraft, such as a glider or sailplane, and the flight of lighter-than-air craft, such as a balloon. The pilot of heavier-than-air craft is an aviator, whereas the pilot of a lighter-than-air craft is an aeronaut. *See* BALLOON; GLIDER.

Aerology is the branch of meteorology concerned with the study of free air, that portion of the atmosphere which is undisturbed by objects on the earth, such as mountains or buildings. Because aircraft move through free air, aerology is of great importance to aviation. A person professionally competent in the science of aerology is an aerologist. *See* AERONAUTICAL METEOROLOGY.

Aviation is broadly grouped into three classes: general aviation, air transport aviation, and military aviation. General aviation comprises all aviation not included in military or air-transport aviation. Military aviation includes all forms of aviation in military activities, and air-transport aviation is primarily the operation of commercial airlines essentially as a public utility for the movement of persons and commodities. *See* AIR TRANSPORTATION; GENERAL AVIATION.

Leslie A. Bryan

Bibliography. M. J. Taylor (ed.), *Jane's Encyclopedia of Aviation*, 1993.

Aviation security

A multilayered, multidimensional means to safeguard passengers, crew, ground personnel, and the general public against acts of unlawful interference perpetrated in flight or within the confines of the airport. It specifically refers to the techniques and methods used in protecting aviation from crime or terrorism.

Preventive Measures

Preventive measures are techniques and methods that are in place to prevent potential threats (such as terrorists, criminals, or explosive materials) from getting on the aircraft. All carry-on baggage is screened by explosive detection equipment or "other appropriate means of detection," which could include trace-chemical detection equipment, manual searches, and bomb-sniffing dogs. The high cost of explosive detection machines may account for the approval by the Transportation Security Administration (TSA) of the use of alternative detection systems. The screening of both people and carry-on luggage remains a controversial issue with security breaches occurring and travelers subjected to different rules from airport to airport. Changes being considered include replacing federal screeners with private screeners at some airports; implementing programs that use biometric systems to screen out potential terrorists; employing new explosion detection, trace-chemical detection, and bomb detection systems; increasing video surveillance; and installing more stringent cargo security systems that track airport workers in luggage and loading areas. Many improvements have been made in aviation security in regards to screening passengers and cargo for potential hazards. Various technological advances have been made to develop more rigorous explosive detection systems with highly developed surveillance systems.

Biometric systems. Biometric systems are designed to recognize biological features to facilitate identity verification. The individual whose identity is being verified places his or her finger, hand, retina or face onto or near a scanner and provides data which are compared with a database. The systems can be based on fingerprints (optical scanning of a finger), signature recognition (including measurement of the motion and pressure used in writing the signature), hand geometry (the physical attributes of the hand, such as the length of the figures), speaker verification (utilizing the uniqueness of voice patterns), or the blood vessel pattern of the retina.

Explosive detection systems. An explosive detection machine operates and looks like a medical computerized axial tomography (CAT) scanner. The system first produces an x-ray scan similar to the conventional airport x-ray scanner. An automated inspection algorithm determines the locations within the baggage where the absorption indicates a suspicious area; cross- sectional computerized tomography (CT) slices then determine size, shape, mass density, and texture of any suspect object. If no high-density areas are detected, a single slice through the bag is made to look for any explosives that may not have been seen in the projection scan. After entering the CT chamber, a bag is divided virtually into three-dimensional units called "voxels." Based on their similarities, certain voxels are grouped together as volumes of the same objects within a bag, so that the density of the scanned object can be determined. Based on density and volume, software automatically correlates the mass characteristics of luggage contents to those of potential explosives. If the system finds a match, it alerts the operator by highlighting suspect areas within the CT slice. Since the CT scan produces true cross-sectional slices, it is able to identify objects that are surrounded by other materials or hidden by innocuous objects. Three-dimensional rendering may also be applied. The machine is programmed to recognize numerous chemical compositions. *See* COMPUTERIZED TOMOGRAPHY.

Another type of explosive detection system uses a document scanner to analyze samples collected by swiping the surface of a document over a collection disk and alerts the screener if an explosive's residue is detected. During the pilot program, passengers selected for secondary screening at particular checkpoints had their boarding passes scanned. If the document scanner raised an alarm, additional screening procedures were implemented.

Trace-chemical detection. Trace-chemical detection equipment can analyze a swipe or air sample, detecting and identifying minute traces of explosive substances or biohazards. The detection process takes 4 s to collect the trace particles and another 8 s to analyze them. As the person being inspected stands in the center of an archway, gradually stronger puffs of air come from four surrounding columns, positioned to direct the air from the lower to the upper parts of the body, accelerating the plume at a faster rate than it would naturally rise. The plume is collected in an overhead detector hood, and the collected particles are vaporized. The molecules are either positively or negatively charged, and the resulting ions are pulsed down a drift tube. The equipment measures how fast the ions travel from point to point. This acts as a thumbprint of the substance, since each specific type of ion has its own particular travel time. This enables the machine to identify a broad range of organic matter, including explosives and materials associated with chemical and biological weapons.

Passengers identified as needing additional screening will pass through the trace portals. Some equipment can access the human convection plume to collect any threatening particles. If a person has explosives strapped to their body or has even handled explosives, trace particles will contaminate clothing and register. A computerized voice indicates when a passenger may exit the portal. Screeners will take necessary and appropriate steps to resolve alarms.

Bomb-detection systems. A new type of bomb-detection system being considered is a terahertz radiation (t-ray) device. Visible light provides the naked eye with a very narrow slice of the electromagnetic spectrum. This technology is destined to reveal not only the shape of objects hidden but also their composition. Because different chemical structures absorb them differently, t-rays could be utilized to identify hidden materials, including explosives and dangerous objects. A t-ray machine looks much like a copying machine. An object is placed on the imaging window, the beam passes across it, and a detector measures the transmitted rays. A screen displays the image and a separate probe arm can scan objects that do not fit neatly into the image window.

Neutron technology is being researched to pinpoint chemicals. Pulsed fast neutron analysis (PFNA) technology is based on the detection of signature radiation (gamma rays) induced in material scanned by a beam of neutrons. The PFNA system has the potential to meet TSA goals as it has demonstrated some superior characteristics compared to existing x-ray systems in detecting explosives in cargo containers and passenger baggage. The PFNA system currently requires a long scan time (an average of 90 min per container), needs considerable radiation shielding, is significantly larger than current x-ray systems, and has high implementation costs. These factors are likely to limit installation at transportation facilities, even if the detection capability is improved. Nevertheless, PFNA currently has the good potential for detecting explosives in cargo as well as vehicles approaching a facility.

Video surveillance. New wireless video surveillance systems will enable separate security operations centers to simultaneously monitor distant sites, including secured airport areas, public parking lots, and roadway tunnels. These systems provide a single wireless solution for viewing, storing, and managing real-time video.

Cargo security. The cargo hold remains the most vulnerable part of the aircraft as most cargo is loaded unscreened. In addition, many problems revolve around the shipment of known and unknown cargo. Known cargo is generically defined as cargo presented for carriage by a known freight forwarder who has been a customer for over a year.

Air cargo security programs are in use in Europe, which has a complete air cargo security program. Air cargo agents must have an approved air cargo security program, and regulated agents conduct security investigations not only of the delivery component (usually trucks) but also of the manufacturer. The entire supply chain has a security element authorized by law.

Defensive Measures

Defensive measures are systems and devices that are in place to defend against threats (such as terrorists, criminals, or explosive materials) that are on the aircraft or can attack the aircraft in flight. Many advances have been made in the development of sophisticated defense systems and devices. These advances are providing progressively safer service worldwide for passengers traveling by air transportation. *See* AIR TRANSPORTATION.

Cockpit doors. All U.S.-based passenger airplanes are required to install reinforced cockpit doors to prevent intruders from gaining access to the flight deck. Currently, reinforced and ballistic-resistant doors, composed of a variety of metals and composites, are designed to withstand extreme pounding and a hail of bullets, as well as heat and smoke. Cargo airlines are granted an exception if they file a security plan. *See* COMPOSITE MATERIAL.

Active Denial System (ADS). The Active Denial System is a nonlethal energy weapon that is designed to provide stand-off nonlethal capabilities at ranges beyond the effectiveness of small arms. The ADS uses a transmitter producing energy at a frequency of 950 GHz and an antenna to direct a focused, invisible beam towards a designated subject. The energy reaches the subject and penetrates the skin to a depth of less than 1/64 in. This energy beam induces an instantaneous heat sensation on the target's skin within seconds which is quite intolerable. The reaction is so intense that the targeted individual is repelled without injury. The tactical theory behind the weapon is to provide a means to stop an aggressor without causing injury. The sensation immediately stops when the individual steps out of the beam.

Containing blasts. Blast containment design concepts attempt to completely suppress the results of an explosion within a luggage container. Blast management design concepts consider the container as part of a placement system inside the cargo bay of the aircraft. Factors such as pressurization of the vehicle, load, and in the case of aircraft, aerodynamic effects can play crucial roles in the outcome of an explosion. The baggage can absorb a great deal of the energy of an explosion, lessening the shock waves, but the explosion may also generate a shower of projectiles entirely dependent on the specifics of the individual bags.

Tests have determined that the blast containment concept offers the best alternative for suppressing the potentially catastrophic effects of postblast fires. This system is an independent unit. It stands alone and needs no special handling or placement within the cargo bay. On the other hand, the blast management concept allows a container to essentially fail and bases its control on the ability to vent the detonation products into adjacent containers. The transporter must arrange the cargo appropriately for the system to have any usefulness. This could prove to be tedious and quite time-consuming. Consequently, early on it was decided to focus on a blast containment container which would be constructed of state-of-the-art high-strength composite materials with fragment-penetration-resistant and fire-retardant properties. Blast-resistant containers are only available for wide-body aircraft, which includes only 25% of the aircraft in service.

Fiber optic sensor systems. These systems have provided perimeter security for many types of facilities for many years. They are particularly useful in the protection of airports because of their stability, long-range applicability (optical fibers can carry light over distances of more than 160 km (or 100 mi), suitability for buried applications, and the fact that they are immune to lightning and other electrical interference. These systems are movement activated.

Antimissile defense. The concept of equipping the commercial carrier fleet with antimissile defenses presents an entirely different set of challenges. Officials are addressing the possibility that terrorists could use shoulder-fired missiles to bring down an airliner or attack other forms of transportation. Patrols of perimeters have increased, but facilities remain vulnerable.

An aircraft antimissile defense would be based on a Doppler radar system made up of four antennas at the front of the aircraft, two on the sides and four on the back. The antennas would be capable of giving 360° of radar coverage around the aircraft. Within seconds of a missile being detected, an on-board computer would release flares, firing at different angles to act as a diversion. The system would be completely automated, (i.e., there would be no involvement of the pilots). *See* DOPPLER RADAR.

Shoulder-fired missiles are currently being produced under license in several countries. Several thousand are in circulation and some are thought to have been acquired by terrorist groups. The threat is real but the feasibility of equipping aircraft with antimissile systems has a prohibitive cost in conjunction with significant technological challenges relating to the speed required to react and the fact that commercial planes are not as maneuverable as fighter military aircraft. However, an airborne defense system against shoulder-launched missiles aimed at civilian aircraft underwent a successful test at Palmahim, Israel in May 2004. Kathleen M. Sweet

Bibliography. C. Combs, *Terrorism in the Twenty First Century*, 2d ed., Prentice Hall, Little Saddle River, NJ, 2000; N. M. Denkin, Fiber optics, World Book Online Reference Center, World Book, June 9, 2004; C. E. Simonson and J. R. Spindlove, *Terrorism Today, The Past, The Players, The Future*, Prentice Hall, Upper Saddle River, NJ, 2000; K. M. Sweet, *Terrorism and Airport Security*, Edwin Mellen Press, March 2002; K. M. Sweet, *Aviation and Airport Security: Safety and Terrorism Concerns*, Prentice Hall, Upper Saddle River, NJ, November 2003; K. M. Sweet, *Transportation Security: Threats and Solutions*, Prentice Hall, Upper Saddle River, NJ, August 2005.

Avocado

A tropical and subtropical fruit tree, *Persea americana*, in the Lauraceae family (containing also the laurel of antiquity and the trees that yield commercial cinnamon and camphor). It originated in Central America or adjoining regions of North or South America. It has now spread to much of the near-tropical world. *See* MAGNOLIALES.

Races. The species is divided into three horticultural races with differing commercial qualities. The so-called West Indian race is least tolerant of cold, the Mexican most tolerant, and the Guatemalan intermediate. This same gradation is found in salt tolerance (West Indian most) and oil content (West Indian lowest). But in some other respects the West Indian race is intermediate (skin thickness), or one of the races is different from the other two (the West Indian fruit is less tolerant of cold storage; the Mexican has smaller fruit with a unique anise-like odor; the Guatemalan has a smaller seed ratio, and takes twice as long to mature—14 months or more in California).

Cultivars. In warm and humid areas such as southeast Florida and all world lowlands near the Equator, only the West Indian race is well adapted; but West Indian × Guatemalan hybrids are doing well and becoming increasingly important in extending the harvest season. Cultivar (horticultural "variety") selection for these tropical regions has taken place chiefly in Florida. There the early season (pure West Indian) is dominated by Waldin and Pollock, with Simmonds increasing. Later-maturing cultivars, with an admixture of Guatemalan-race genes, include the Booth 8, first in production among all Florida cultivars and with increasing acreage, and the declining Booth 7. For the still later period, Lula has long predominated and still rivals Booth 8, but is being planted very little because of susceptibility to scab. Choquette and Monroe are two late cultivars of increasing importance. A dozen or so additional cultivars are also of some significance in Florida and other tropical areas.

Cultivar development for drier and colder-winter regions has occurred mostly in California. There, and in similar climates such as most of Mexico, Chile, South Africa, Israel, and Australia, Guatemalan and Guatemalan × Mexican hybrids are best adapted. Fuerte, containing about equal genetic contributions from the two races, was long the leading cultivar but is now declining in importance because of erratic production. Hass, chiefly of Guatemalan-race origin, is increasingly dominating California production, and its success in Mexico makes it easily the world's leading cultivar—in Michoacan State, the proportion planted (or top-worked) to Hass is estimated at a remarkable 90%. A heavy-setting cultivar with superb quality, Hass well deserves its preeminence (**Fig. 1**). Among hundreds of other selections named around the world, the following are most important—California: Zutano, largely Mexican; Bacon, intermediate in race traits, very cold-hardy; Reed, apparently pure Guatemalan; Pinkerton, intermediate, precocious, necked fruit, but very promising; Hawaii: Sharwil (from Australia), intermediate, plus several local selections; South Africa: Edranol (from California), mostly Guatemalan; Israel: Ettinger, mostly Mexican. Major breeding programs in Israel and California should soon produce superior cultivars.

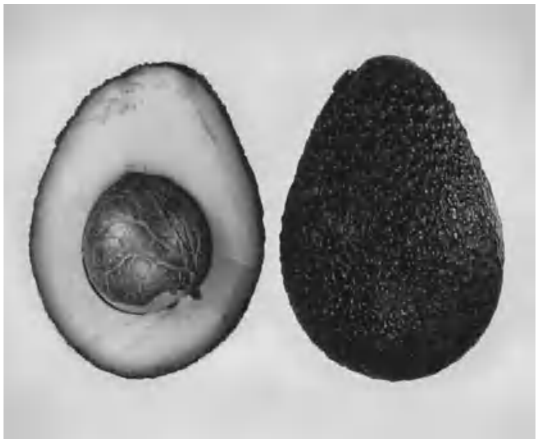


Fig. 1. Hass, the world's leading avocado cultivar.

Culture and management. The avocado is limited especially by its climatic requirements, with their race differences. It is also highly susceptible to drought injury. But excess soil moisture is equally fatal, encouraging the dread Phytophthora root rot in most of the world, and where the fungus is unknown, as in Israel, promoting root asphyxiation. It is likewise unusually sensitive to salts in the root zone. The combined drought and salt susceptibilities mean that, over the 6 months or so without significant precipitation in California, regular irrigation is necessary. It may be needed weekly on light soils, especially when the available water is rather saline. The drip irrigation method is working well, especially with young trees.

Few California orchards need sprays for pest control—predators and parasites normally control such insects and mites as may develop. But in a few areas, during a few seasons, either of two caterpillar genera or thrips build up sufficient numbers to make spray control desirable. In other parts of the world, insects are more of a problem and require regular control measures.

All California orchards need nitrogen fertilizers and probably most need zinc. Laboratory analysis of leaves will reveal their nutritional status for these elements and others that are more rarely present in either deficiency or excess. The experienced grower can often identify nutritional problems from leaf symptoms. For other parts of the avocado world, the only safe guide is local observation and experiment.

Precocity and productivity. Precocity and productivity vary with both cultivar and location. The best California climate is more rigorous than is desirable for the avocado, both in terms of occasional freezes and, more seriously, in terms of average suboptimal temperatures for fruit set. Within California and other growing areas with comparable climate, a location providing good air drainage and southern exposure may well be sufficiently warmer to make the difference between success and failure of the enterprise.

One can expect a light crop set 2 years after planting Pinkerton, 3 years for Hass, 3 or 4 for Bacon, and 4 or 5 for Fuerte. Average production from mature trees is usually correlated with precocity; with spacing at about 7 m (24 ft) square, yields in kilograms per hectare (or pounds per acre) from a well-managed orchard in a good location will be perhaps 7000 for Fuerte in its limited zone of adaptation, and also for Bacon; perhaps 10,000 for Hass; perhaps 15,000 for Pinkerton—although it is not yet adequately tested. The Israeli cultivar Tova may even exceed the Pinkerton level, but Tova quality is lower. Both of these very heavy setters are about on a par with adapted cultivars in more favorable near-tropical regions. Planting arrangements to encourage cultivar cross-pollination will generally, and often strikingly, enhance fruit set because of the avacado's unusual flower behavior. This behavior makes self-pollination difficult by causing nearly all the open flowers of a cultivar, in any given location, to be functionally female one time of day and male a different time; this condition is called synchronous protogynous dichogamy. *See* POLLINATION.

Mexico is the world's leading producer, followed by Brazil and California, then Colombia and Venezuela, countries of eastern South America, Central America, Caribbean Islands, Florida, Philippines, and Zaire (central West Africa). South Africa and Israel have important export industries, primarily to Europe. Many other countries have begun development. The California industry has expanded rapidly, and the avocado has become one of the state's leading fruit crops.

Marketing. Seedling trees have fruits ranging from about 2 oz to 4 lb (50 to 2000 g), with most commercial markets preferring about 10 oz (280 g). However, some local markets, especially in more tropical areas, prefer a somewhat larger fruit. Shape varies from slightly oblate to highly elongate; an intermediate oval or pear shape is preferred. Mature-fruit skin color varies from greenish-yellow to purple or black, through different shades of green and even some red tints. Color preference is usually a reflection of what the consumer has become accustomed to.

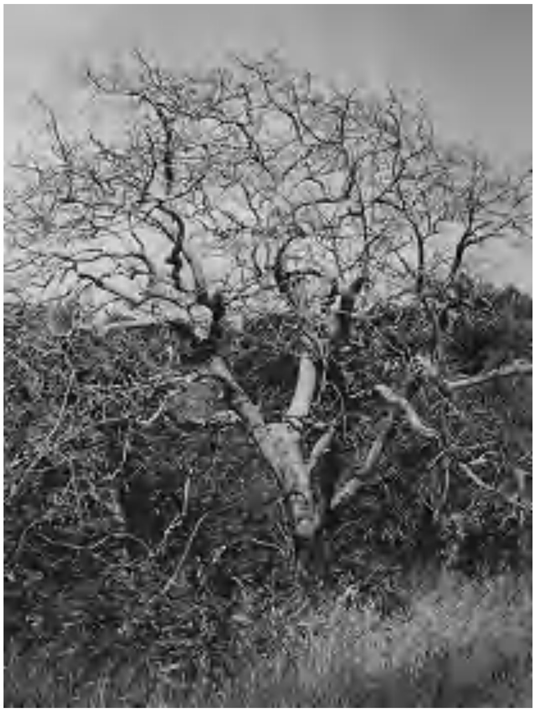


Fig. 2. Avocado tree with Phytophthora root rot.

Marketing is made more difficult by the avocado's unusual flavor. Instead of being in the sweet-to-tart range of most fruits, the avocado flesh has a unique, subtle taste, due to its forming oils rather than carbohydrates. As a result, most consumers are not favorably impressed at their first exposure; only after several experiences does it usually become a prized food delicacy. Therefore, to effectively enter new markets, an extensive program of consumer education and advertising may be necessary. *See* FRUIT; FRUIT, TREE.

B. O. Bergh

Diseases. Phytophthora root rot (causal fungus *Phytophthora cinnamomi*) is the most serious and destructive disease of the avocado worldwide (**Fig. 2**). Resistant rootstocks, soil fungicides, and possibly biological control offer promise against this disease. Other relatively minor problems require action as follows. For Verticillium wilt, preplanting or interplanting avocado with other susceptible crops such as tomato should be avoided. For Armillaria root rot, trees should be removed and the soil fumigated with methyl bromide. Sunblotch, a viroidlike disease, is controlled by using certified seed and budwood sources. Phytophthora trunk cankers necessitate excision and painting affected areas with fungicide, and avoiding moisture accumulation around the trunk. Several leaf and fruit spots in humid tropical areas are caused by the fungi *Colletotrichum, Sphaceloma*, and *Cercospora*, and are controlled by spraying with fungicides. Black streak is a disease of unknown cause affecting Guatemalan varieties. *See* PLANT PATHOLOGY.

George A. Zentmyer

Bibliography. *California Avocado Society Yearbook*, annually; J. Janick and J. N. Moore (eds.), *Advances in Fruit Breeding*, 1975; *Proceedings of the*

1st International Tropical Fruit Short Course: The Avocado, 1977; G. A. Zentmyer et al., *Avocado Diseases*, Univ. Calif. Circ. 534, 1965; G. A. Zentmyer and H. D. Ohr, *Avocado Root Rot*, Univ. Calif. Leaf. 2440, 1978.

Avogadro number

The number of elementary entities in one mole of a substance. A mole is defined as an amount of a substance that contains as many elementary entities as there are atoms in exactly 12 g of ^{12}C; the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles. Experiments give 6.0221367×10^{23} as the value of the Avogadro number. In most calculations the coefficient is rounded off to 6.02. Thus, a mole of ^{12}C atoms has 6.02×10^{23} carbon atoms, a mole of water molecules contains 6.02×10^{23} H_2O molecules, a mole of electrons contains 6.02×10^{23} electrons, and so forth. *See* MOLE (CHEMISTRY).

Significance. The atomic weight (relative atomic mass) of ^{12}C is exactly 12, by definition. Consider 12 g of ^{12}C (which is one mole and contains the Avogadro number of atoms) compared with 4 g of He, whose atomic weight is 4. The 12 g to 4 g ratio of the masses of the two samples is the same as the 12 to 4 ratio of the masses of the atoms of ^{12}C and He. Therefore the two samples must contain the same number of atoms, and 4 g of He contains the Avogadro number of atoms. The same argument holds for any element. Thus, for an element with atomic weight x, a sample with mass x grams contains the Avogadro number of atoms. Similarly, for a substance with molecular weight y, a sample whose mass is y grams must contain the Avogadro number of molecules. For example, 18 g of water contains 6.02×10^{23} H_2O molecules. *See* ATOMIC MASS; RELATIVE ATOMIC MASS.

The Avogadro number is a dimensionless number. The Avogadro constant is defined as the Avogadro number divided by the unit “mole.” The Avogadro constant is usually symbolized by N_A, N_0, or L. Since N_A gives the number of molecules per mole, $N_A = N/n$, where N is the number of molecules present in n moles of a substance.

The Avogadro number relates the mass of a mole of a substance to the mass of a single molecule. For example, for H_2O (whose molecular weight is 18) the mass of one mole is 18 g and the mass of one molecule is $(18\text{ g})/(6.02 \times 10^{23}) \approx 3 \times 10^{-23}$ g. The mass m of one molecule of a substance with molar mass M is $m = M/N_A$.

The Avogadro constant N_A is related to other fundamental physical constants. The Faraday constant F is the absolute value of the charge on one mole of electrons. Therefore $F = N_Ae$, where e is the absolute value of the charge on one electron. Also, $R = N_Ak$, where R is the gas constant and k is the Boltzmann constant. *See* BOLTZMANN CONSTANT; GAS CONSTANT.

Widespread use of the mole concept began only around 1900. The nineteenth-century concept most closely related to the Avogadro number is the number of molecules per unit volume in a gas at 0°C and 1 atm. [The ideal-gas law $PV = nRT = (N/N_A)RT$ gives $N/V = N_AP/RT$, so N/V, the number of gas molecules per unit volume, is proportional to the Avogadro constant N_A at fixed pressure P and temperature T.] Avogadro hypothesized in 1811 that at a fixed temperature and pressure the number of molecules per unit volume is the same for different gases, but he had no way of estimating this number. *See* AVOGADRO'S LAW.

Determination. The first estimate of the number of molecules per unit volume in a gas was given by J. Loschmidt in 1867. Using a kinetic-theory-of-gases equation for the gas viscosity and estimating NV_{molecule} (where V_{molecule} is the volume of a single molecule) from the volume of the liquid formed by condensing the gas, Loschmidt gave a value of N/V which is one-thirtieth the correct value. *See* KINETIC THEORY OF MATTER.

The number of molecules per cubic centimeter in an ideal gas at 0°C and 1 atm is called the Loschmidt number in English-speaking countries. (However, in Germany, the Loschmidt number means the number of molecules per mole.)

In 1900 M. Planck, using the hypothesis of energy quantization, derived the law for blackbody radiation. Planck's law contains the Boltzmann constant k and the Planck constant h. By fitting experimental data to his law, Planck found values for h and k, and from k (which equals R/N_A) he calculated the Avogadro number as 6.18×10^{23}. Using $F = N_Ae$, he calculated $e = 4.69 \times 10^{-10}$ statC. Thus Planck had obtained the first reasonably accurate values of h, k, N_A, and e. *See* BOSE-EINSTEIN STATISTICS.

In 1906 A. Einstein used his molecular theory of diffusion and data for aqueous sugar solutions to estimate $N_A = 4 \times 10^{23}$ mol^{-1}; Einstein subsequently found an error in his work and recalculated this estimate as 6.56×10^{23} in 1911.

J. B. Perrin in 1908 observed the distribution in the Earth's gravitational field of colloidal gamboge particles of uniform size. According to the Boltzmann distribution law, the ratio n_2/n_1 of numbers of particles at heights h_2 and h_1 in the gravitational field is given by Eq. (1). The potential energies of the particles are

$$n_2/n_1 = \exp[-(E_2 - E_1)/kT] \qquad (1)$$

$E_1 = mgh_1$ and $E_2 = mgh_2$, where m is the mass of a particle (corrected to allow for the buoyancy of the suspending fluid) and g the gravitational acceleration. Observation of n_2/n_1 allows calculation of the Boltzmann constant k, and use of $R = N_Ak$ then gives N_A. Perrin found $N_A = 7.0 \times 10^{23}$ mol^{-1}.

From 1908 to 1917, R. A. Millikan and H. Fletcher determined the electron charge by observing the motions of charged oil drops. Using their final result for e and the value of the Faraday constant known from electrolysis experiments, Millikan found $N_A = F/e = 6.06 \times 10^{23}$ mo1^{-1} in 1917. *See* ELECTRON.

A very accurate way to find N_A uses x-rays to measure the lattice spacing in a crystal. The crystal's density ρ equals the mass of a unit cell divided by the volume of a unit cell: $\rho = m_{\text{cell}}/V_{\text{cell}}$. The mass of one formula unit is M/N_A, where M is the molar mass; hence $m_{\text{cell}} = ZM/N_A$, where Z is the number of formula units per unit cell. If the unit cell is cubic, then $V_{\text{cell}} = a^3$, where a is the unit-cell edge length. The crystal density is then given by Eq. (2). Z is readily

$$\rho = MZ/a^3N_A \tag{2}$$

found from x-ray diffraction study of the crystal. *See* X-RAY CRYSTALLOGRAPHY; X-RAY DIFFRACTION.

In the period 1930–1970, several determinations of N_A were done by using Eq. (2), with a a determined from the diffraction pattern produced by x-rays of known wavelength. The wavelength was determined by diffracting the x-rays with a grating having closely ruled lines of known spacing.

The accuracy of the x-ray crystal-density method was improved substantially by R. Deslattes and coworkers in 1974 by using a combination of x-ray and optical interferometry to determine the lattice spacing in a very pure silicon (Si) crystal. These workers also did very accurate measurements of the density and atomic weight of silicon. P. Becker and coworkers in 1981 repeated the interferometry measurement of the Si lattice spacing and obtained a slightly different result from that of Deslattes. When the Becker determination of a is combined with Deslattes group's determination of M and ρ, the value $N_A = 6.02213 \times 10^{23}$ mol^{-1} is found. *See* FUNDAMENTAL CONSTANTS. Ira N. Levine

Bibliography. R. D. Deslattes, The Avogadro constant, *Annu. Rev. Phys. Chem.*, 31, 435–461, 1980; D. Kolb, The mole, *J. Chem. Educ.*, 55, 728–732, 1978.

Avogadro's law

The principle that equal volumes of all gases and vapors, under the same conditions of temperature and pressure, contain an identical number of molecules; also known as Avogadro's hypothesis. From Avogadro's law the converse follows that equal numbers of molecules of any gases under identical conditions occupy equal volumes. Therefore, under identical physical conditions the gram-molecular weights of all gases occupy equal volumes. Avogadro's law is not strictly obeyed by real gases at ordinary temperatures and pressures, although the deviations are only slight. At high pressures the deviations may be large. Avogadro's law can be shown to follow theoretically from the simple kinetic theory of gases. *See* GAS; KINETIC THEORY OF MATTER. Thomas C. Waddington

Axenic culture

The growth and maintenance of a single species in isolation, free from foreign or contaminating species. Isolation is usually achieved by growing the species in an environment that was previously sterilized, and was thereby rid of contaminating organisms. Since, from a practical viewpoint, the contaminating organisms usually encountered are microorganisms, axenic cultures, whether of invertebrates (for example, nematodes) or vertebrates (for example, rodents), are often referred to as germ-free. Indeed, the terms axenic and germ-free are occasionally used interchangeably. Gnotobiotic is also often used interchangeably with axenic; however, in common practice gnotobiotic specifically refers to germ-free conditions.

A principal goal of early studies with axenic cultures was the establishment of nutritional requirements for individual species. Historically, axenic cultures have also been employed to demonstrate that organisms growing in close association can have both direct and indirect effects on each other. Direct effects include competition for nutrients, attacks (for example, by a parasite), or production by one species of toxic compounds that affect a second species. Indirect effects include such phenomena as the production by intestinal microflora of vitamin K (germ-free animals therefore require it as a dietary supplement).

Technology. The two methodological challenges associated with developing axenic culture methods are formulating an appropriate culture medium and designing a functional habitat. Formulating a culture medium for small organisms, including microorganisms and many invertebrates, is the more difficult of the two challenges. For larger organisms, a culture medium is replaced by a carefully formulated diet, which is usually a relatively straightforward task. Initiating the axenic cycle and maintaining a sterile environment can, however, be a formidable endeavor for larger organisms.

For axenic culture of small organisms, such as bacteria, protozoa, and fungi, a typical strategy is to add to a minimal medium [which usually contains only a carbon source (such as glucose), nitrogen, sulfur, and phosphorus, as well as various salts and trace minerals] several supplements that are rich in a large number of growth factors. Such supplements often include water-soluble extracts of meat or yeast containing vitamins and other organic compounds. Once successful growth is achieved, it is usual to attempt to simplify the culture medium by deleting, in a stepwise fashion, each of the more esoteric ingredients (for example, individual vitamins).

Initiating the axenic cycle and designing an appropriate habitat for larger organisms require special effort. First, organisms free of contaminating species need to be obtained; in some cases, cesarean births are employed. Second, sterile containers that have entry and exit ports as well as a comfortable living environment must be constructed.

Methodology. Axenic culture has been achieved for a broad range of organisms. The method employed to remove contaminating organisms in a starter culture or for the initiation of an experiment depends upon the type of organism to be decontaminated and the nature of the contaminating organism.

Decontamination procedures for various types of organisms are summarized below.

Microorganisms. Most bacteria, fungi, and algae can be grown on a solid surface such as that provided by agar supplemented with a variety of nutrients. Initially, either single cells are handpicked and placed on the agar surface, or a collection of cells is first diluted to low density in an appropriate culture medium and then spread on the agar surface. After an appropriate culture period, those single cells will proliferate into a small mass of cells that represents an axenic culture.

Plants. Many types of plant cells can be cultured by using methods similar to those for microorganisms. Axenic culture of whole plants begins with sterilization of the starting material. For example, seeds or small whole plants can be washed with a sterilizing agent, such as alcohol, and then deposited on an agar surface containing several nutrients and a variety of antibiotics. Similarly, some hardy, small whole plants can be sterilized with alcohol or mercuric chloride solution and grown in appropriate liquid or solid media.

Protozoa. Since protozoa are often isolated from water sources laden with particulate soil components, grasses, and other bacteria-infested particles, washing by centrifugation is often employed to provide starting material for culturing. Protozoa are separated from debris and bacteria. Alternatively, a single cell can be isolated with a microscope and transferred to an appropriate culture medium. In both cases it is usually advisable to permit the isolated protozoa to incubate in culture medium for several hours before the final transfer to axenic conditions is made. This permits protozoa, which normally ingest bacteria, to clear any bacteria that they harbor in vacuoles or that cling to their surface structures.

Nematodes. The free-living nematode *Caenorhabditis* has been employed extensively as an experimental system for studies of the influence of gene action on the pattern of embryogenesis. Normally, it ingests bacteria as a food source. It can, however, be grown axenically in a buffered salt solution that contains tissue extracts from mammalian liver or chick embryos. Alternatively, the salt solution can be supplemented with autoclaved (killed) bacteria, sterols, and a heme-containing protein, such as hemoglobin, myoglobin, or cytochrome.

Insects. Various types of insects can be grown axenically, and thus free of both cellular and intestinal microbes. Axenic culture begins with disinfected eggs produced by bathing freshly fertilized eggs in a dilute solution of a sterilizing agent. Disinfected eggs are then raised on a sterilized diet, which can be either undefined (for example, animal parts) or completely synthetic (that is, containing pure organic ingredients). Once developed to the adult stage, axenic insects can be maintained in transparent glass culture containers to which fresh food is regularly added.

Eggs. Various marine invertebrate eggs, as well as the eggs of lower vertebrates such as fish and frogs, can also be disinfected and grown into adults to provide breeding stock for permanent axenic strains.

Mammals. To raise small mammals, such as mice and rats, for long-term (multiple-generation) studies, the whole room that houses the animals is made germ-free. Then cesarean-delivered animals are raised in the room to provide breeding stock. Sterilized bedding and food are delivered to the room; fresh feces are collected regularly and tested for the presence of bacteria. In addition, sterile food and beverage (usually water) must be routinely supplied.

Antibiotic therapy. Since bacteria represent the single most important potential contaminant in axenic cultures, antibiotic therapy is usually employed to maintain the cultures. There is no single antibiotic that inhibits all bacteria, however, so combinations of up to four to six antibiotics are usually added either directly to the culture fluid for axenic stocks of protozoa, plant cells, or nematodes, or to the food or beverage for larger, germ-free animals such as small mammals. Axenic cultures need to be continually monitored for the presence of contaminating microorganisms. *See* BACTERIOLOGY.

Uses. Axenic cultures have been used extensively for both basic research and practical application.

Basic research. Many microorganisms occupy subtle niches in host species. For example, the colon of mammals is coated with a mucous lining that is heavily infested with highly anaerobic bacteria. These bacteria outnumber other bacteria by up to 1000 to 1, but because of their extreme sensitivity to oxygen they earlier evaded detection. Once axenic hosts (for example, mice) became available, it was possible to inoculate them and to identify the interrelationships between microbe and host. An extension of that strategy has been successfully employed in dental research: to study the role of bacterial flora in tooth decay, a germ-free host, usually a mammal such as a small rodent or dog, has a pure culture of bacteria, for example, *Streptococcus*, applied as an oral swab.

Practical applications. Agriculture, medicine, and aerospace research have all benefited tremendously from application of axenic culture methods.

For plants, axenic cultures of peanut, tobacco, corn, and grass, as well as flowering plants, are now routine. Three areas of research are being exploited: (1) determination of plant growth requirements, (2) analysis of organic compounds released from certain plants, and (3) colonization of certain plant roots by bacteria as well as fungi.

Economically important agricultural animals, such as axenic chickens, have been the subject of intense research with the goal of separating the effects of the microflora, which abound in the domestic fowl's digestive tract, from the host's normal digestive enzyme action. Digestion is substantially modified in the axenic chicken. For example, in the crop, the first stages of starch digestion depend on bacteria, which secrete lactic acid. In axenic chickens, that phase of digestion is absent.

Silkworm culture became more practical because of the elucidation of the pathogenic microorganisms that affect the development and productivity of the silkworm.

Parasitology has also benefited. Several parasitic protozoa have been cultivated under axenic conditions and then inoculated into germ-free animals. It was discovered, for example, that when *Entamoeba histolytica* was inoculated into germ-free guinea pigs it failed to induce typical abscesses, as it normally would when injected into a conventional host. Further experiments revealed that in germ-free hosts this parasite was deprived of its natural food source—the bacterial flora of the intestine.

Finally, in aerospace research it is important that flights to other planets or the Moon do not contaminate their surfaces with microorganisms from Earth. Hence, any equipment involved in an actual landing needs to be free of contaminating organisms. Similarly, any extraterrestrial samples that are brought back to Earth should be treated as axenic cultures in order to facilitate their characterization and analyses.

Limitations. One problem is that it is not always possible to be certain that viruses and mycoplasmas are not present as a contaminant in axenic cultures. Another limitation is that since axenic cultures are designed to be pure it is almost impossible to construct a holistic or comprehensive view of an organism with data from axenic cultures alone.

Nevertheless, the benefits that have been obtained from the use of axenic cultures have far outweighed the limitations. Indeed, virtually all aspects of research on pharmaceuticals, nutrition, ecology, agricultural production, and parasitology have benefited enormously from the use of axenic cultures. *See* POPULATION ECOLOGY.

George M. Malacinski

Bibliography. S. S. Bhojwani (ed.), *Plant Tissue Culture: Applications and Limitations*, 1991; R. I. Freshney (ed.), *Animal Cell Culture: A Practical Approach*, 1992; L. Nuzzolo and A. Vellucci, *Tissue Culture Techniques*, 1983.

Aye-aye

A prosimian primate in the family Daubentoniidae inhabiting Madagascar. It is the most endangered of all mammals in Madagascar. A member of the order Primates and the suborder Prosimiae, aye-ayes (*Daubentonia madagascariensis*) are most closely related to the lemurs.

Morphology. Aye-ayes have a long, coarse, dark brown to black coat; long, naked, leathery ears; large eyes surrounded by distinctive dark rings; and an extremely long, slender middle finger with a long claw which is used not only to groom but also to extract insect larvae from beneath the bark of trees and to dig out the pith of bamboo, sugarcane, and coconuts. The nose, cheeks, chin, throat, and spots over the eyes are yellowish white. The hands and feet are black. At times, the aye-aye will suspend itself by its hind feet, thus freeing its hands for feeding or grooming. The tail is long and bushy. Aye-ayes have a head-plus-body length of 360–440 mm (14–17 in.), a tail length of 500–600 mm (19.5–23 in.), and they weigh 2–3 kg (4.5–6.5 lb). There is no sexual dimorphism. The mouth contains 18 teeth (I 1/1, C 0/0, PM 1/0, M 3/3 × 2). The incisors are large, curved, and chisel-like with enamel on only the front surface; they continue to grow throughout the life of the animal in the same manner as the incisors of a rodent. A second, larger species (*D. robusta*), known only from skeletal remains and a few teeth, became extinct several centuries ago.

Ecology. Although aye-ayes are primarily arboreal and inhabit rainforests, mangrove forests, and bamboo thickets, they frequently descend to the ground. They are the largest nocturnal primate and live at very low densities. They are either solitary or live in small family groups. They construct elaborate sleeping nests approximately 500 cm (20 in.) in diameter in the forks of trees. The nests are constructed of twigs and interlaced leaves and are located 10–15 m (30–45 ft) above the ground. Each individual may have 5 to 10 such nests in their territory of approximately 5 hectares (12.5 acres). Male home ranges are overlapping and may be over 200 hectares (500+ acres), although female home ranges have been found to be considerably smaller.

When searching for food, aye-ayes will listen intently for the sounds of insect larvae in decaying wood. They often tap on the wood with their long middle finger, an action thought to help them locate the extensive system of galleries constructed by insect larvae as well as by ants and termites. When a food source is located, the aye-aye will bite through the wood using its strong incisors and extract the insect with its long, clawed middle finger.

Reproduction and development. Females breed every 2 to 3 years and give birth to a single young. The only two recorded gestation periods have been 158 and 172 days. The aye-aye is the only species of the order Primates in which the paired mammary glands and nipples are located in the abdominal, rather than the thoracic, region. Weaning occurs at about 1 year of age, and they reach sexual maturity at 3 years of age. Individuals have lived for more than 23 years in captivity.

Distribution. The aye-aye occurs in a few rainforests on the eastern coast of Madagascar, a rainforest in east-central Madagascar, in several localities in humid northwestern forests, and even in deciduous forests in western Madagascar. These regions contain the older, tall trees which harbor beetle larvae upon which aye-ayes feed. In 1996, 11 aye-ayes were taken to Nosy Mangabe Island off the northeastern coast of Madagascar. This island is a protected wildlife reserve and now contains a relocated population of aye-ayes. It is estimated that the total remaining number of wild aye-ayes numbers between 1000 and 10,000. Their decline has been due mainly to forest destruction and killing by humans. They are designated as endangered by the International Conservation Union (IUCN) and the USDI, and they are on Appendix 1 of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). *See* MAMMALIA; PRIMATES.

Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, vol. 2, McGraw-Hill, 1990; D. Macdonald,

The Encyclopedia of Mammals, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Azeotropic distillation

Any of several processes by which liquid mixtures containing azeotropes may be separated into their pure components with the aid of an additional substance (called the entrainer, the solvent, or the mass separating agent) to facilitate the distillation. Distillation is a separation technique that exploits the fact that when a liquid is partially vaporized the compositions of the two phases are different. By separating the phases, and repeating the procedure, it is often possible to separate the original mixture completely. However, many mixtures exhibit special states, known as azeotropes, at which the composition, temperature, and pressure of the liquid phase become equal to those of the vapor phase. Thus, further separation by conventional distillation is no longer possible. By adding a carefully selected entrainer to the mixture, it is often possible to "break" the azeotrope and thereby achieve the desired separation. *See* AZEOTROPIC MIXTURE; DISTILLATION.

Entrainers fall into at least four distinct categories that may be identified by the way in which they make the separation possible. These categories are: (1) liquid entrainers that do not induce liquid-phase separation, used in homogeneous azeotropic distillations, of which classical extractive distillation is a special case; (2) liquid entrainers that do induce a liquid-phase separation, used in heterogeneous azeotropic distillations; (3) entrainers that react with one of the components; and (4) entrainers that dissociate ionically, that is, salts. *See* SALT-EFFECT DISTILLATION.

Within each of these categories, not all entrainers will make the separation possible, that is, not all entrainers will break the azeotrope. In order to determine whether a given entrainer is feasible, a residue curve map for a mixture undergoing simple distillation is created.

Residue curve maps. The least complicated of all distillation processes is the simple distillation, or open evaporation, of a mixture from an open vessel. The liquid is boiled in such a way that the vapor is removed from contact with the liquid as soon as the vapor is formed. The composition of the liquid will change continuously with time since the vapor is always richer in the more volatile components than the liquid from which it came. The path of liquid compositions starting from some initial point is known as a simple distillation residue curve, or simply, a residue curve. The collection of all such curves for a given mixture is known as a residue curve map. These maps contain the same information as the corresponding phase diagram for the mixture, but they represent it in a way that it is more useful for understanding and designing distillation systems.

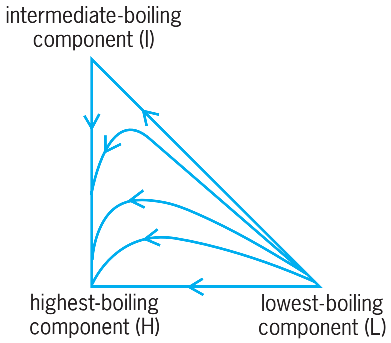

intermediate-boiling component (I)
highest-boiling component (H)
lowest-boiling component (L)

Fig. 1. Schematic representation of the residue curve map for nonazeotropic ternary mixtures.

Mixtures that do not contain azeotropes have residue curve maps that all look the same. For ternary mixtures of this sort, the map looks like the one shown in **Fig. 1**. The residue curves all start at the lowest-boiling pure component (L), they then move toward the intermediate-boiling component (I), and they end at the highest-boiling component (H). These curves show that when such a mixture is boiled the liquid initially gets richer in each of the heavier components but eventually gets richer in only the heaviest component until only the pure heavy component remains. One of the most important properties of residue curves is that they must move in such a way that the boiling temperature of the mixture increases along every curve. The shape of the curve in Fig. 1 reflects this property. The arrows on the edges point in the direction of increasing boiling temperature, and the map in the interior is the only structure that is consistent with the sides of the triangle. Typical systems for nonazeotropic ternary mixtures include hexane (L) + heptane (I) + nonane (H) and acetaldehyde (L) + methanol (I) + water (H).

The presence of even one binary azeotrope destroys the structure shown in Fig 1. If the mixture contains a single minimum-boiling binary azeotrope, three residue curve maps are possible, depending on whether the azeotrope is between the lowest- and highest-boiling components (**Fig. 2***a*), between the intermediate- and highest-boiling components

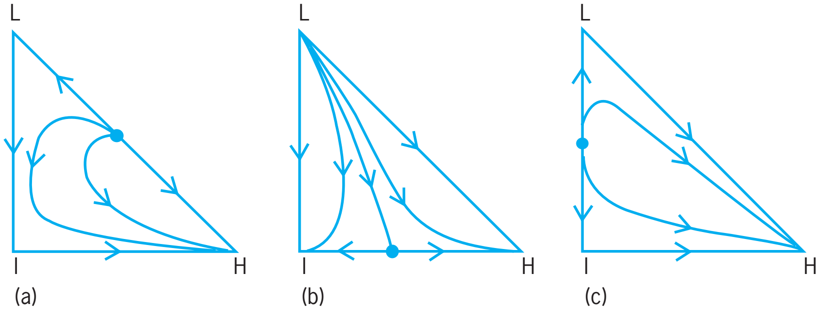


Fig. 2. Schematic representation of the residue curve maps for ternary mixtures with one minimum-boiling binary azeotrope. (*a*) Azeotrope between the lowest- and highest-boiling components. (*b*) Azeotrope between the intermediate- and highest-boiling components. (*c*) Azeotrope between the intermediate- and lowest-boiling components.

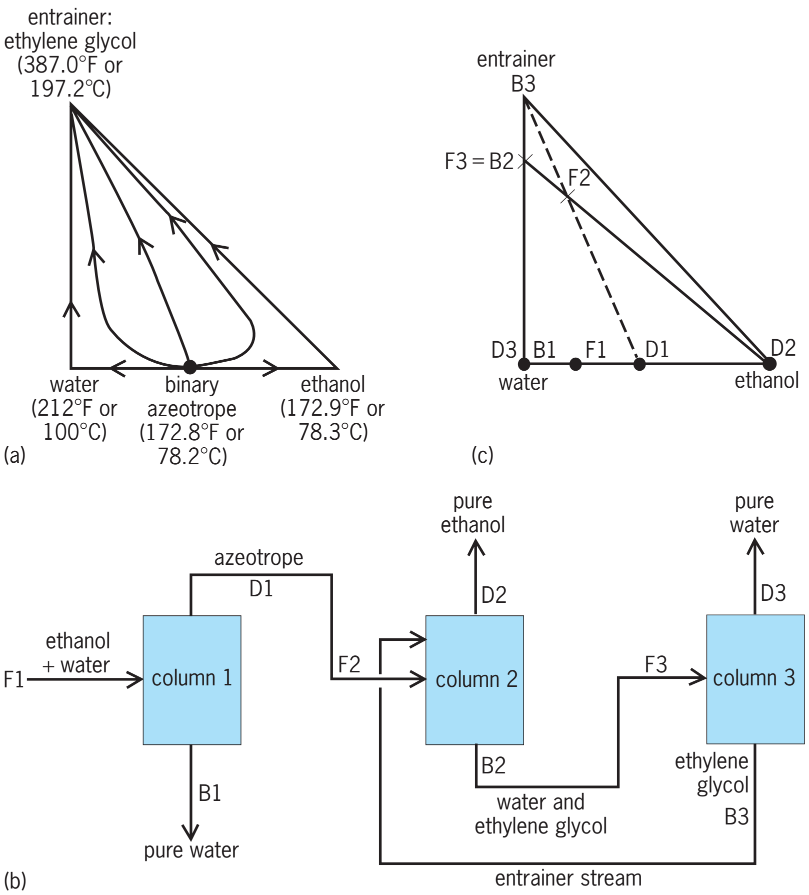

entrainer:
ethylene glycol
(387.0°F or
197.2°C)
water
(212°F or
100°C)
binary
azeotrope
(172.8°F or
78.2°C)
ethanol
(172.9°F or
78.3°C)
(a)
entrainer
B3
F3 = B2
F2
D3
B1
F1
D1
D2
water
ethanol
(c)
pure
ethanol
pure
water
azeotrope
D1
D2
D3
ethanol
+ water
F1
column 1
F2
column 2
F3
column 3
B2
ethylene
glycol
B3
B1
water and
ethylene glycol
pure water
(b)
entrainer stream

Fig. 3. Classical extractive distillation of ethanol and water using ethylene glycol as entrainer. (*a*) Residue curve map. Temperatures shown are the normal boiling points of the pure components and the azeotrope. Binary azeotropic composition is not drawn to scale, for clarity. (*b*) Distillation sequence showing column configuration. (*c*) Overall material balance lines. F = feed. D = distillate. B = bottoms.

(Fig. 2*b*), or between the intermediate- and lowest-boiling components (Fig. 2*c*). It is significant that the residue curves in Fig. 2*a* and *c* all end up at the highest-boiling pure component, but in Fig. 2*b* the curves end at either the I vertex or the H vertex depending on the initial composition. The special curve that divides these two distillation regions is called a distillation boundary. Mixtures with more than one azeotrope exhibit more complex residue curve maps with either single or multiple distillation boundaries. Typical systems include methanol (L) + i-propanol (I) + water (H), as in Fig. 2*b*, and ethanol (L) + water (I) + ethylene glycol (H), as in Fig. 2*c*.

Homogeneous azeotropic distillation. When a mixture is separated continuously in a distillation column, the overall material balance demands that the feed composition, the distillate composition, and the bottoms composition lie on a straight line in the composition triangle. It can be argued that these overall material balance lines cannot cross a simple distillation boundary to any appreciable or useful extent. Thus, nonazeotropic mixtures may be separated into their pure components by using a sequence of distillation columns because there are no distillation boundaries to get in the way. Of course, the situation is quite different when azeotropes are present, as can be seen from Fig. 2. It is possible to separate mixtures that have residue curve maps similar to those shown in Fig. 2*a* and *c* by straightforward sequences of distillation columns. This is because these maps do not have any distillation boundaries. These, and other feasible separations for more complex mixtures, are referred to collectively as homogeneous azeotropic distillations. Without exploiting some other effect (such as changing the pressure from column to column), it is impossible to separate mixtures that have residue curve maps like Fig. 2*b*.

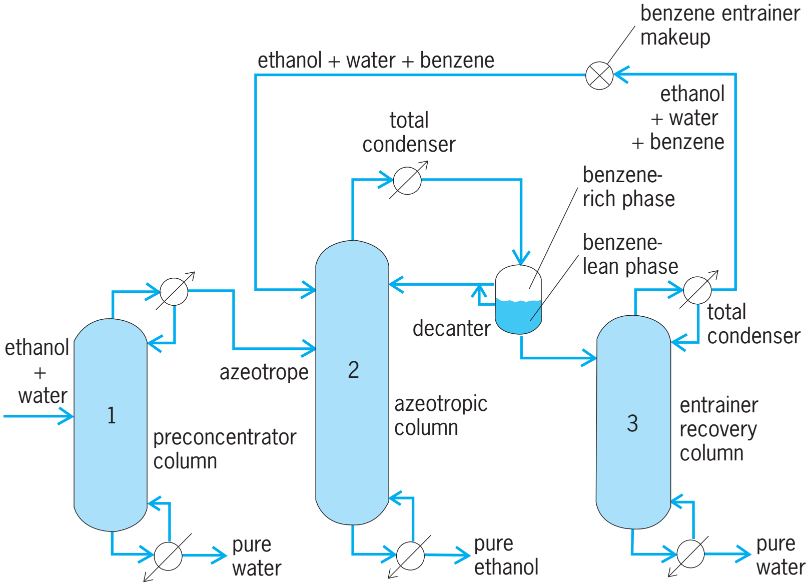

benzene entrainer
makeup
ethanol + water + benzene
ethanol
+ water
+ benzene
total
condenser
benzene-
rich phase
benzene-
lean phase
decanter
total
condenser
ethanol
+
water
1
preconcentrator
column
azeotrope
2
azeotropic
column
3
entrainer
recovery
column
pure
water
pure
ethanol
pure
water

Fig. 4. Typical sequence of columns for heterogeneous azeotropic distillation, for example, ethanol + water + benzene (entrainer).

Extractive distillation. A large number of mixtures have residue curve maps similar to Fig. 2*c*, and therefore the corresponding distillation is given the special name extractive distillation. As can be seen from this residue curve map, the entrainer, H, that must be added to the binary azeotropic mixture of components L and I must have the properties of being the heaviest pure component in the mixture and of introducing no new azeotropes. For example, ethylene glycol acts as just such an entrainer for separating ethanol and water. The residue curve map, the corresponding distillation sequence, and the overall material balance lines for this extractive distillation are shown in **Fig. 3**. Column 1 (Fig. 3*b*) separates the binary mixture of ethanol and water into a pure water stream (B1) and an azeotrope (D1). This azeotrope and the ethylene glycol entrainer are fed into column 2, which is designed to produce pure ethanol as distillate (D2) and a binary mixture of water and ethylene glycol as bottoms (B2). Finally, column 3 performs the easy separation between water (D3) and ethylene glycol (B3), which is recycled to the middle column (2). Typical overall energy requirements to separate a feed mixture consisting of 10 wt % ethanol and 90 wt % water by this method is 22,000 Btu/gallon (6100 kilojoules/liter) of anhydrous ethanol. A well-designed, thermally integrated sequence of distillation columns can carry out this same separation for about 8000 Btu/gallon (2200 kJ/liter) of ethanol. Other azeotropic mixtures that are separated by this technique include (entrainers are in parentheses) acetone + methanol + (water), methanol + methyl acetate + (acetic acid), propanol + allyl alcohol + (water), and methanol + acetone + (tetrahydrofuran).

Heterogeneous azeotropic distillation. Heterogeneous entrainers cause liquid-liquid phase separations to occur in such a way that the composition of each phase lies on either side of a distillation boundary. In this way, the entrainer allows the separation to "jump" over a boundary that would otherwise be impassable. Classic examples include ethanol + water plus; (benzene), or any one of the entrainers pentane, hexane, or heptane instead of benzene; and isopropanol + water + any one of the previous entrainers. The residue curve maps associated with all these mixtures are essentially the same, having two binary minimum-boiling homogeneous azeotropes, one binary minimum-boiling heterogeneous azeotrope, and one ternary minimum-boiling heterogeneous azeotrope. The corresponding sequence of distillation columns is shown in **Fig. 4**.

The first column (preconcentrator) concentrates the binary feed mixture (ethanol + water) into an azeotrope as distillate and water as bottoms. The binary azeotrope is fed to the second column (azeotropic), together with an ethanol + water + benzene stream which is recycled from the third column (entrainer recovery). The bottoms stream is ethanol product, and the overhead vapor is close to the composition of the minimum-boiling ternary azeotrope. The vapor is condensed, whereupon it splits into two liquid phases. The benzene-rich layer is returned to the column as reflux, and the water-rich layer is fed to the final column (entrainer recovery), where it is separated into a bottoms stream of pure water and a distillate stream containing all three components. This stream is recycled to the second column. A typical overall energy requirement to separate a feed mixture consisting of 10 wt % ethanol and 90 wt % water by this method is 23,000 Btu/gallon (6400 kJ/liter) of anhydrous ethanol.

While this configuration is the most common, other heterogeneous residue curve maps have been exploited. The separation of diethoxymethane from a mixture of ethanol and water, has been developed commercially.
M. F. Doherty

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., 1999.

Azeotropic mixture

A solution of two or more liquids, the composition of which does not change upon distillation. The composition of the liquid phase at the boiling point is identical to that of the vapor in equilibrium with it, and such mixtures or azeotropes form constant-boiling solutions. The exact composition of the azeotrope changes if the boiling point is altered by a change in the external pressure. A solution of two components which form an azeotrope may be separated by distillation into one pure component and the azeotrope, but not into two pure components. Standard solutions are often prepared by distillation of aqueous solutions until the azeotropic composition is reached. At 1 atm (760 mmHg or approximately 100 kilopascals) pressure, hydrogen chloride and water form an azeotrope containing 20.24% by weight of HCl. *See* DISTILLATION; SOLUTION.
Francis J. Johnston

Azide

A compound containing the group $—N_3$, which can be represented as a resonance hybrid of two structures, as shown in expression (1).

$$-\overset{\ominus}{N}-\overset{\oplus}{N}\equiv N \longleftrightarrow -N=\overset{\oplus}{N}=\overset{\ominus}{N} \qquad (1)$$

Sodium azide, from which most other azides are prepared, is manufactured by passing nitrous oxide over heated sodium amide, reaction (2). It is a water-soluble, stable compound. Heavy-metal azides are highly explosive and very shock-sensitive; lead azide, $Pb(N_3)_2$, is used as a detonator to set off explosives. Sodium azide, in combination with an oxidizing agent, may be used as a gas generator in motor-vehicle passive-restraint systems. *See* EXPLOSIVE.

$$2NaNH_2 + N_2O \rightarrow NaN_3 + NaOH + NH_3 \qquad (2)$$

A variety of organic azides are known. The most important of these are aryl azides (ArN_3), azidoformates ($ROCON_3$), and sulfonyl azides (RSO_2N_3). These lose N_2 when heated or exposed to ultraviolet light to generate species known as nitrenes, which are so reactive that they will react with almost any organic compound to form amine derivatives, as shown in reaction (3). Aryl azides are widely used to probe the active sites of biological targets by photoaffinity labeling. Difunctional aryl azides are used commercially to prepare photoresists, the nitrenes reacting with a polymer containing double bonds to insolubilize the polymer in the light-struck areas. The insoluble polymer protects the underlying metal from being attacked by an etching solution. Difunctional azido-formates and sulfonyl azides can be used to cross-link polymers, to prepare polymeric foams, and to adhere tire cord to rubber in the manufacture of automobile tires. Compounds containing a sulfonyl azide group and a hydrolyzable silane group, such as $—Si(OCH_3)_3$, in the same molecule are used to bond silaceous fillers (glass fibers, silica, mica, and so forth) to almost any organic polymer. Such coupled systems have much superior properties to simple mechanical mixtures.

$$RN_3 \xrightarrow[-N_2]{\text{heat or light}} R\overset{\cdot\cdot}{\underset{\cdot\cdot}{N}} \xrightarrow{R'CH_3} RNHCH_2R' \qquad (3)$$

Simple acyl azides (R = alkyl or aryl) undergo a Curtius rearrangement on heating to form an isocyanate, reaction (4).

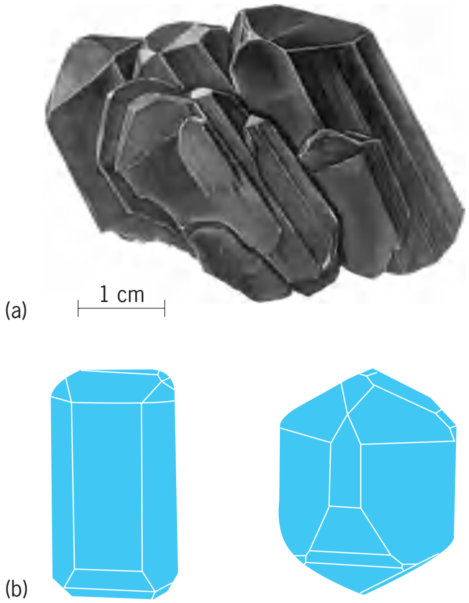
$$\mathrm{RCON_3} \xrightarrow{\text{heat}} \mathrm{RN{=}C{=}O} + \mathrm{N_2} \qquad (4)$$

See NITROGEN; POLYMERIZATION; RESONANCE (MOLECULAR STRUCTURE); RUBBER. David S. Breslow

Bibliography. W. Lwowski (ed.), *Nitrenes*, 1970; S. Patai (ed.), *The Chemistry of the Azido Group*, 1971; E. F. V. Scriven (ed.), *Azides and Nitrenes*, 1984.

Azurite

A basic carbonate of copper with the chemical formula $Cu_3(OH)_2(CO_3)_2$. Azurite is normally associated with copper ores and often occurs with malachite. Azurite is monoclinic. It may be massive or may occur in tabular, prismatic, or equant crystals (see **illus.**). Invariably blue, azurite was originally used extensively as a pigment. Hardness is 3.5–4 (Mohs scale) and specific gravity is 3.8. It can be synthesized by gentle heating of cupric nitrate or sulfate solutions with calcium carbonate in a closed tube. Notable localities for azurite are at Tsumeb, Southwest Africa, and Bisbee, Arizona. *See* COPPER; MALACHITE. Robert I. Harker



Azurite. (*a*) Crystals from Tsumeb, Southwest Africa (*specimen from Department of Geology, Bryn Mawr College*). (*b*) Crystal habits (*after C. Klein, Manual of Mineralogy, 21st ed., Wiley, 1993*).

B

Bacillariophyceae — Binomial theorem

Bacillariophyceae

A class of nonflagellate unicellular algae, commonly called diatoms, with boxlike silicified walls. Diatoms range in maximum dimension from 4 micrometers to more than 1 millimeter. The diatom wall or frustule (**Fig. 1**) comprises several interlocking, usually elaborately sculptured, lightly or heavily silicified pieces overlying a thin polysaccharide layer. The two largest pieces are the upper and lower valves, which fit together like the top and bottom of a petri dish or shoe box. Between the valves (along the side or girdle of the cell), several smaller pieces—hooplike girdle bands—are intercalated. Depending upon which dimension is larger, breadth or depth, a diatom tends to lie on the valve side or on the girdle side.

Valves are honeycombed by perforate chambers arranged in patterns characteristic of individual species. The overall symmetry of the valves and details of their structure and ornamentation provide the basis for classification. More than 200 genera and 12,000 species of living diatoms have been described. Two main groups are recognized: those in which structural features of the valve are arranged with reference to a central pole (centric valve; **Fig. 2**) or to two or more poles (gonioid valve); and those in which the features are arranged with respect to a line, often symmetrically (pennate valve). The first group are usually called centric diatoms (although not all members have centric valves) and are circular, semicircular, elliptical, or polygonal in valve view; the second group, usually called pennate diatoms, are naviculoid (boat-shaped), crescent-shaped, or needle-shaped in valve view.

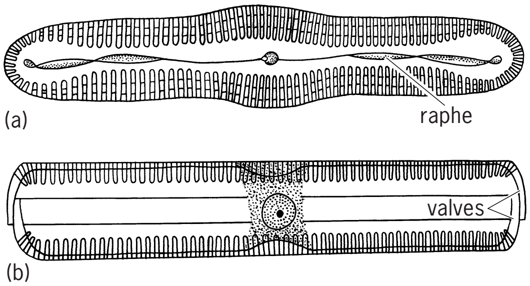


Fig. 1. Pennate diatoms. *Pinnularia*: (*a*) top, or valve view, showing longitudinal slit, or raphe; (*b*) side, or girdle view, showing overlapping halves or valves. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

The diatom cell contains a nucleus and, except in the few colorless species, one to many chloroplasts, which may be minute disks or expanded plates. The chloroplasts are yellow to dark brown as a result of the masking of chlorophylls *a* and *c* by β-carotene and various xanthophylls, especially fucoxanthin. Pyrenoids are present in some species. Thylakoids are grouped in threes, and chloroplast deoxyribonucleic acid is distributed in a ring. A noteworthy feature of the diatom cell is the spindle, which is composed of highly organized microtubular elements and is readily visible with a compound microscope during cell division.

Reproduction. Vegetative reproduction consists of binary fission in the plane of the valves. Following mitosis, two new valves are formed, each lying inside an old valve. One of the daughter cells is thus the same size as the parent cell, while the other is smaller. The average cell size of a population gradually diminishes to a value as small as one-fifth maximum. Eventually, maximum size is restored by sexual reproduction or by an asexual regenerative process.

Vegetative cells of both pennate and centric diatoms are diploid, with meiosis occurring during gametogenesis. Centric diatoms are oogamous, with one or two eggs and up to 128 sperm being produced by separate cells. Each sperm has only one flagellum, which is anteriorly inserted and lacks the central microtubular doublet characteristic of all other flagella. Pennate diatoms are isogamous or anisogamous, with one or two ameboid gametes being produced by each cell. The zygote in both centric and

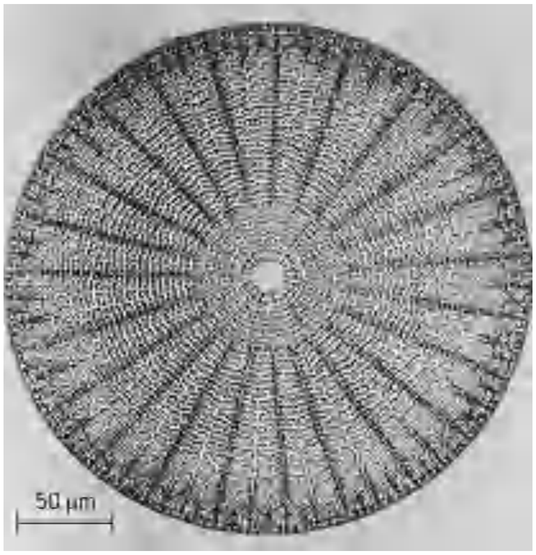


Fig. 2. *Arachnoidiscus ehrenbergii*, a concentric diatom with radial symmetry. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

pennate diatoms enlarges into an auxospore, with an organic wall that may be covered by small siliceous scales or rings. A frustule pattern characteristic of the species is restored in the second generation of daughter cells. The primary inductor of sexuality is a critical minimum size, but high light intensity may also be important.

Motility of vegetative cells is found only in pennate diatoms that possess a raphe, which is a complex longitudinal slit or end-to-end pair of slits passing through the wall of one or both valves. Movement, which can occur only if the raphe-bearing valve is in contact with the substrate, is effected by secretion of polysaccharide along the length of the raphe. A bundle of actin microfilaments, assumed to play a role in movement, parallels the raphe in the adjacent cytoplasm.

Ecology. Most diatoms are photosynthetic. Many, however, are auxotrophic, requiring an external source of certain vitamins. The few colorless species are found where high concentrations of nutrients are available, as in sewers or in the mucilage of brown algae. Silicon, which in the hydrated form $SiO_2 \cdot H_2O$ makes up the frustule, is of critical importance. Diatoms cannot divide in the absence of silicate. Germanium is a competitive inhibitor to the uptake of silicate and is thus a toxin that is selective for diatoms. Food reserves include lipids and the polysaccharide chrysolaminaran.

Diatoms are likely to occur wherever there is moisture. They are free-living or attached, solitary or colonial. Planktonic forms (primarily centric diatoms) bear long spines and bristles that aid flotation. Free-living benthic forms include many motile pennate diatoms. When fixed to a substrate, diatoms may be cemented individually by the lower valve, attached by a gelatinous pad or stalk, or enveloped in a gelatinous film. Colonies take the form of chains of cells linked valve to valve by interlocking spines, mucous or chitinous threads, or mucilaginous pads, or of pseudofilaments or blades composed of numerous individuals (often motile) in a common tubular or flattened mucilaginous matrix.

Terrestrial habitats, which generally support pennate rather than centric diatoms, include wet rocks in the spray zone at the seashore, along streams, and, if there is sufficient light, in caves. Diatoms are also found in the upper layers of soil, in old snow, and among mosses. They inhabit all bodies of still and running fresh water, including hot springs with temperatures as high as 118°F (48°C). They are abundant in both plankton and benthos and may be attached to other algae, higher plants, animals, or sediment particles, or free-living in the sediment. In marine habitats they are often the main constituent of the phytoplankton and the mudflat flora. They are the most abundant eukaryotic organisms on the lower surface of sea ice in the Arctic and Antarctic, where they survive temperatures below 28°F (−2°C). Diatoms grow on the surface of most marine macroalgae, and several species occur on the skin of whales. A few live symbiotically as naked cells within marine invertebrates and foraminifera.

Marine planktonic diatoms have been important primary producers for at least 100 million years, and over the millennia their frustules have accumulated on the ocean floor. From the abundant fossil record, it is known that centric diatoms evolved first (Cretaceous), followed by pennate diatoms (Paleocene). Uplifted deposits (called diatomite, diatomaceous earth, or kieselguhr) are mined at several locations, including Santa Barbara County, California. Diatomite is used as a filter, abrasive, catalyst carrier, absorbent, and insulator.

Because certain species are characteristic of clean water while others are characteristic of organically polluted water, diatoms may be used as an environmental index. *See* ALGAE; CHRYSOPHYCEAE.

Paul C. Silva; Richard L. Moe

Bibliography. P. Bourrelly, *Les Algues d'Eau Douce . . .*, tome 2: *Les Algues Jaunes et Brunes, Chrysophycées, Phéophycées, Xanthophycées et Diatomées*, 1968; N. I. Hendey, *An Introductory Account of the Smaller Algae of British Coastal Waters*, pt. 5: *Bacillariophyceae (Diatoms)*, 1964; D. Werner (ed.), *The Biology of Diatoms*, 1977.

Bacillary dysentery

A highly contagious intestinal disease caused by rod-shaped bacteria of the genus *Shigella*. Bacillary dysentery is a significant infection of children in the developing world, where it is transmitted by the fecal-oral route. The global disease burden is estimated as 165 million episodes and 1.3 million deaths annually. Common-source outbreaks occasionally occur in developed countries, usually as a result of contaminated food. The most common species isolated in developed countries is *S. sonnei*, while *S. flexneri* serotypes predominate in endemic areas.

Epidemics of *S. dysenteriae* 1 occur in equatorial regions, and these outbreaks can involve adults as well as children.

When ingested even in very small numbers, shigellae multiply in the intestine and invade the epithelial lining of the colon. Infection of this tissue elicits an acute inflammatory response (colitis) that is manifested as diarrhea or bloody, mucoid stools (dysentery). The virulence of all *Shigella* species, and *Shigella*-like enteroinvasive *Escherichia coli*, depends on an extrachromosomal genetic element (virulence plasmid) that encodes four invasion plasmid antigen (Ipa) proteins and a secretory system (Type III) for these proteins. Secreted Ipa proteins help shigellae to initiate colonic invasion through specialized endocytic intestinal cells (M cells). After shigellae pass through these M cells, they are phagocytized by tissue macrophages in the underlying lymphoid tissue. Ipa proteins then induce apoptosis (programmed cell death) in infected macrophages, releasing cytokines (primarily IL-1) that initiate an acute, localized inflammatory infiltrate. This infiltrate of polymorphonuclear leukocytes destabilizes tight junctions between absorptive epithelial cells (enterocytes), making the tissue more susceptible to additional *Shigella* invasion. Secreted Ipa proteins induce uptake of shigellae by the colonic enterocytes. The virulence plasmid also encodes an intercellular spread protein (IcsA) that recruits mammalian cytoskeletal elements (primarily actin) to the bacterial surface. This actin is organized into a cytoplasmic motor that facilitates spread of shigellae to adjacent enterocytes. *See* DIARRHEA; ESCHERICHIA.

In otherwise healthy individuals, bacillary dysentery is typically a short-term disease lasting less than a week. The symptoms can be truncated by appropriate antibiotic therapy (such as oral ampicillin or cyprofloxacin) that rapidly eliminates shigellae from the intestinal lumen and tissues. When *S. dysenteriae* 1 is the etiologic agent, however, hemolytic uremic syndrome (HUS) can be manifested as a serious consequence of disease. This species produces a cytotoxin (Shiga toxin or Stx) that is functionally identical to the toxin of enterohemorrhagic *E. coli* (for example, O157:H7). Stx inhibits protein synthesis, damaging endothelial cells of the intestinal capillary bed; the toxin may also damage renal tubules, causing acute renal failure with chronic sequela in up to one-third of HUS patients. *See* MEDICAL BACTERIOLOGY.

Thomas L. Hale

Bibliography. R. Bhimma et al., Post-dysenteric hemolytic uremic syndrome in children during an epidemic of *Shigella* dysentery in Kwazulu/Natal, *Pediat. Nephrol.*, 11:560–564, 1997; E. Egile et al., SopA, the outer membrane protease responsible for polar localization of IcsA in *Shigella flexneri, Mol. Microbiol.*, 23:1063–1073, 1997; R. Menard et al., The secreted Ipa complex of *Shigella flexneri* promotes entry into mammalian cells, *Proc. Nat. Acad. Sci. USA*, 93:1254–1258, 1996; P. J. Sansonetti et al., Role of IL-1 in the pathogenesis of experimental shigellosis, *J. Clin. Invest.*, 96:884–892, 1995.

Background count

The number of counts recorded by a radiation detector from background radiation. The term background radiation refers to the natural ionizing radiation on the Earth. Ionizing radiation refers to all radiations, waves, and particles that are energetic enough to remove electrons from stable atoms; they are stronger than infrared radiation, radio waves, or visible light, which cannot separate electrons from stable atoms. Radiation strong enough to cause ionization of atoms is measured in electrical units which range from 32 electronvolts up to millions of electronvolts. *See* ELECTROMAGNETIC RADIATION; IONIZATION POTENTIAL.

Sources. Background radiation comes from two sources, cosmic and terrestrial. The sum of these radiations is called the background radiation dose. Cosmic radiation comes from outer space and consists of gamma rays and alpha, neutron, and beta particles. Alpha particles from outer space are very energetic, and each consists of two neutrons coupled to two protons. They are relatively massive but do not penetrate deeply into matter such as human tissue. Beta particles are energetic electrons and are more penetrating. However, it is the gamma-ray flux, that is, concentration, that constitutes most of the cosmic-ray dose impacting on the Earth. This dose, the collection of all ionizations, is attenuated as it passes through the Earth's atmosphere; the dose rate roughly doubles for each 2000 m (6500 ft) increase in altitude. At thinner mountain atmospheres the cosmic-ray dose is higher, and on jet flights the dose rate can be 100 times that at sea level. Some of the cosmic rays are so energetic that they can still be detected in the deepest mines. *See* ALPHA PARTICLES; BETA PARTICLES; COSMIC RAYS; GAMMA RAYS.

The Earth is naturally radioactive with the decay products of natural elemental uranium and thorium which are present (at about 1 part per million on average) in every gram of the Earth's mantle. These primordial radionuclides decay through a complex chain, ultimately ending as stable elemental lead. Each step in the decay process results in the emission of some ionizing particles and gamma rays.

Dosage. A radiation detector at sea level, for example, would detect the radiation from both cosmic and terrestrial sources. The measurement of background radiation is in absorbed dose units known as sieverts (Sv). (Previously the unit was the millirem, and 100 millirem = 1 mSv.) The chief concern is with the amount of background radiation absorbed by people's tissues. Most people absorb about 0.3 mSv per year at sea level, roughly 1 microsievert per day. About 10% of this is external radiation from cosmic radiation, and 10% is external radiation from terrestrial sources. Internal radiation from inhaled and ingested radionuclides constitutes the rest. Some 13% of the background radiation comes from the natural terrestrial radionuclide (radioisotope) potassium-40, which is uniformly concentrated in all living cells and is present as 1 in every 2000 potassium atoms. The remaining two-thirds of the background dose comes from inhalation of radon daughter products. Radon is a radioactive noble gas which is derived from the decay of radium, a decay product of uranium. *See* RADON; UNITS OF MEASUREMENT.

In regions where the geology is such that there is a relatively high concentration of uranium, the radon concentration may be elevated, and the background dose rate can be quite high. About 0.1% of the population of the United States may receive radon doses 10 times the average, and some 10–15% may be subjected to levels 5 times higher than the average. One such area is the so-called Reading Prong in Pennsylvania, where the levels of radon in some buildings have been found to be so high as to constitute a possible radiation hazard for the occupants.

Similarly, at high altitudes, such as in Denver, Colorado, at 1.6 km (1 mi), the increased cosmic-ray dose raises the total background dose rate about 10% above that at sea level, but this increase is of minimal significance in terms of radiation risk. There are many areas where the background radiation rate varies because of local geology or altitude. Background radiation is not constant since cosmic flux is influenced by sunspots and radon levels are influenced by humidity and the degassing of soil.

Typically, an average person will receive a lifetime background dose of about 0.25 Sv. The level can be a somewhat lower or up to about 10 times higher, as mentioned above, but it can never be zero.

Instrument correction. All radiation detection instruments are influenced by background radiation from the sources mentioned above. Thus any measurement to determine the level of radiation in a specific environment must correct for the natural level of ionization reaching the detector. Analytical radiation-measuring devices frequently employ a means of automatically subtracting the background level so as to provide information on the net amount of radiation in the sample or environment in question. For very sensitive measurements, the detector is frequently sheathed in a heavy shield of lead, and a lead-collimated tube may also extend from the sample to the detector to shield the sensitive detector from external background radiation and minimize the background count. *See* ENVIRONMENTAL RADIOACTIVITY; RADIATION SHIELDING; RADIOACTIVITY.

Marvin Goldman

Bacteria

Extremely small—usually 0.3 to 2.0 micrometers in diameter—and relatively simple microorganisms possessing the prokaryotic type of cell construction. Although traditionally classified within the fungi as Schizomycetes, they show no phylogenetic affinities with the fungi, which are eukaryotic organisms. The only group that is clearly related to the bacteria are the blue-green algae.

Bacteria are found almost everywhere, being abundant in soil, water, and the alimentary tracts of animals. Each kind of bacterium is fitted physiologically to survive in one of the innumerable habitats created by various combinations of space, food, moisture, light, air, temperature, inhibitory substances, and accompanying organisms. Dried but often still living bacteria can be carried into the air.

One of the few locations in which bacteria are not usually found is within the cells of other healthy organisms, though even this is subject to exceptions, as there are many bacteria that do live intracellularly in a number of eukaryotic organisms.

Bacteria have a practical significance for humans. Some cause disease in humans and domestic animals, thereby affecting health and the economy. Some bacteria are useful in industry, while others, particularly in the food, petroleum, and textile industries, are harmful. Some bacteria improve soil fertility. *See* FOOD ENGINEERING; FOOD MICROBIOLOGY; INDUSTRIAL MICROBIOLOGY; MEDICAL BACTERIOLOGY; PETROLEUM MICROBIOLOGY; SOIL MICROBIOLOGY; TEXTILE MICROBIOLOGY; ZOONOSES.

As in higher forms of life, each bacterial cell arises either by division of a preexisting cell with similar characteristics or through combination of elements from two such cells in a sexual process. The earlier idea, that full-fledged bacteria arise from nonliving material by spontaneous generation, has been disproved by careful elimination of living bacteria from the nonliving material. This does not eliminate the possibility that, sometime during evolution of the universe, life was derived from the nonliving. Separation of matter into living and nonliving is arbitrary, though useful and unambiguous when transitional states are not under consideration. *See* PREBIOTIC ORGANIC SYNTHESIS; REPRODUCTION (ANIMAL).

Cultures

Descriptions of bacteria are preferably based on the studies of pure cultures, since in mixed cultures it is uncertain which bacterium is responsible for observed effects. Pure cultures are sometimes called axenic, a term denoting that all cells had a common origin in being descendants of the same cell, without implying exact similarity in all characteristics. Pure cultures can be obtained by selecting single cells, but indirect methods achieving the same result are more common.

If conditions are suitable, each bacterium grows and divides, using food diffused through the gel, and produces a mass of cells called a colony. Colonies always develop until visible to the naked eye unless toxic products or deficient nutrients limit them to microscopic dimensions. *See* AXENIC CULTURE; BACTERIOLOGY; CULTURE.

Classification

The morphology, that is, the shape, size, arrangement, and internal structures, of bacteria can be distinguished microscopically and provides the basis for classifying the bacteria into major groups. Three principal shapes of bacteria exist, spherical (coccus), rod (bacillus), and twisted rod (spirillum). The coccus may be arranged in chains of cocci as in

Streptococcus, or in tetrads of cocci as in *Sarcina*. The rods may be single or in filaments. Stains are used to visualize bacterial structures otherwise not seen, and the stain reaction with Gram's stain provides a characteristic used in classifying bacteria. *See* STAIN (MICROBIOLOGY).

Many bacteria are not motile. Of the motile bacteria, however, some move by means of tiny whirling hairlike flagella extending from within the cell. Others are motile without flagella and have a creeping or gliding motion. Spiral forms are usually polarly flagellated, that is, with flagella at the end of the cell. Cocci (spheres) are rarely flagellated. Rod-shaped bacteria may lack flagella or have polar or peritrichous (around the entire surface of the cell) flagella.

Many bacteria are enveloped in a capsule, a transparent gelatinous or mucoid layer outside the cell wall. Some form within the cell a heat- and drought-resistant spore, called an endospore. Cytoplasmic structures such as reserve fat, protein, and volutin are occasionally visible within the bacterial cell.

The nucleus of bacteria is prokaryotic, that is, not separated from the rest of the cell by a membrane. It contains the pattern material for forming new cells. This material, deoxyribonucleic acid (DNA), carrying the information for synthesis of cell parts, composes a filament with the ends joined to form a circle. The filament consists of two DNA strands joined throughout their length. The joining imparts a helical form to the double strand. The double-stranded DNA consists of linearly arranged hereditary units, analogous and probably homologous with the "genes" of higher forms of life. During cell division and sexual reproduction, these units are duplicated and a complete set is distributed to each new cell by an orderly but as yet unelucidated mechanism. *See* BACTERIAL GENETICS.

The submicroscopic differences that distinguish many bacterial genera and species are due to structures such as enzymes and genes that cannot be seen. The nature of these structures is determined by studying the metabolic activities of the bacteria. Data are accumulated on the temperatures and oxygen conditions under which the bacteria grow, their response in fermentation tests, their pathogenicity, and their serological reactions. There are also modern methods for determining directly the similarity in deoxyribonucleic acids between different bacteria.

Temperature relationships. Bacteria are said to be psychrophilic if their optimum temperature is below 60°F (20°C), mesophilic if it is 60–113°F (20–45°C), and thermophilic if it is above 113°F (45°C). Some bacteria can grow at temperatures as high as 167°F (75°C). Others, which are not killed but which cannot grow at high temperatures, are called thermoduric.

Oxygen relationships. Bacteria are said to be aerobic if they require oxygen and grow best at a high oxygen tension, usually 20% or more. Microaerophilic bacteria need oxygen, but grow best at, or may even require, reduced oxygen tensions, that is, less than 10%. Anaerobic bacteria do not require oxygen for growth. Obligatorily anaerobic bacteria can grow only in the complete absence of oxygen.

Fermentation and respiration. Fermentation is a term used to indicate processes in which foodstuffs are decomposed in the absence of oxygen. Respiration is the comparable aerobic process, in which oxygen is one of the foods. Some oxygen-utilizing microorganisms cannot completely oxidize the food to water and carbon dioxide, and often form acids as a product of this type of aerobic food utilization. These incomplete oxidative processes are sometimes called fermentations, though they are actually examples of limited respiration. As in all chemical rearrangements, some of the available energy in both respiration and fermentation is dissipated as heat. The remainder is stored in the form of the materials that make up living cells. *See* FERMENTATION.

In respiration as much as 50% of the food material and energy appears as bacterial cell material and the remainder as carbon dioxide, water, and heat. In fermentation the lack of O_2 decreases the energy supply available through rearrangement of the food, and less food (up to 15%) is converted to cells. Less heat is dissipated, and fermentation products are formed, such as CO_2, hydrogen, methane, ethanol, acetone, glycerol, and formic, acetic, propionic, butyric, lactic, and succinic acids. These products, when combined with oxygen by aerobes to form water and carbon dioxide, yield energy equal to the difference between the energy available in respiration and in fermentation.

Fermentation tests. Fermentation tests, which use liquid media, each medium containing a different nutrient, aid in classifying bacteria. Gas or acids or both are always formed when carbohydrates are fermented. Acid is detected by including an indicator of acidity in the culture medium; a small inverted tube is used to trap any gas emitted. Other useful tests measure the acidity (pH) developing during fermentation and the range of acidity permitting growth.

Fermentation of proteins yields products similar to those from carbohydrates plus large quantities of nitrogenous products, such as ammonia and amines. Since ammonia and amines are weak bases, a protein fermentation causes alkalinity instead of acidity. The ability of an organism to ferment carbohydrate or protein is tested by inoculating the organisms into milk containing litmus indicator, called litmus milk. The culture is incubated and then examined for color changes denoting acidity or alkalinity.

Digestion tests. Tests for digestion of protein, starch, fats, cellulose, pectin, and many other insoluble materials disclose other physiological characteristics useful in classification. Ability to digest protein (peptonization) is often tested by examining litmus milk cultures for an increase in transparency, caused by digestion of casein, the protein responsible for the white opacity of milk. The digestion of gelatin, another protein, may be detected by liquefaction.

Other metabolic reactions of bacteria include the oxidation of ammonia to nitrate by nitrifying

-s

bacteria, oxidation of sulfur to sulfates by sulfur bacteria, and oxidation of ferrous to ferric iron by iron bacteria. Some of the products formed within the cell in these oxidations react with carbon from carbon dioxide, with hydrogen from water, and with other elements to form new cells.

The purple sulfur bacteria and the green sulfur bacteria also form new cells from inorganic compounds, but the hydrogen is obtained by splitting water with light (photolysis) instead of with chemical energy. These bacteria form an oxidized substance as the second product of the photolysis of water, as do green plants. But the former cannot convert this to oxygen, which the latter do; hence the photosynthetic bacteria can photosynthesize only if an oxidizable compound, such as hydrogen (H_2), hydrogen sulfide (H_2S), or a suitable organic substance, is present with which the oxidized moiety is continually reduced.

Some bacteria obtain energy from the oxidation of reduced substances with compounds other than oxygen (O_2). The sulfate reducers use sulfate, the denitrifiers nitrate or nitrite, and the methanogenic bacteria carbon dioxide as the oxidizing agents, producing H_2S, nitrogen (N_2), and methane (CH_4), respectively, as reduction products. *See* BACTERIAL PHYSIOLOGY AND METABOLISM.

Pathogenicity. Pathogenicity, the ability to cause disease, is another property used in establishing the relationship between various groups of bacteria. Some bacteria produce disease only in certain species; for example, *Neisseria gonorrhoeae* will cause gonorrhea only in humans. Some bacteria cause only one disease, while others may cause several diseases. An example of the former is *Corynebacterium diphtheriae*, which causes diphtheria; *Staphylococcus aureus* belongs to the latter category and may cause boils, osteomyelitis, and pneumonia. *See* PATHOGEN.

Serological reactions. Serological reactions are very useful in distinguishing closely related bacteria. If two bacteria, A and B, differ, some of their proteins and other complex molecules also differ. When cells of A are injected into an experimental animal, such as a rabbit, some of their constituent molecules (especially proteins) cause production in the rabbit's blood of special proteins called antibodies. Each of these can combine specifically with the molecular species that caused its production. If, after a suitable incubation period, blood is drawn from the animal (in the case of a rabbit usually from an ear vein) and allowed to clot, a clear yellow liquid (blood serum) is extruded as the clot shrinks. It contains antibodies against each protein in the injected A cells. If B cells in excess are added to this antiserum, each B protein occurring also in A reacts with its corresponding antibody, thereby removing from the serum all antibodies against proteins common to both A and B. Addition of A cells gives a further reaction if A contains proteins not found in B. With reciprocal absorption of B antiserum with A cells, and testing with B cells for antibodies restricted to B, any differences in the A and B cells can be detected. *See* BACTERIAL TAXONOMY; SEROLOGY.

Bacterial pathogens have been observed that form a protective capsule which may belong to one of several different serological types. Noncapsulated mutant cells occasionally arise. In the animal they are destroyed by phagocytes, but on artificial media they survive and produce colonies with a rough surface. Such strains are called rough or *R* forms to distinguish them from the capsulated, smooth or *S* types. Heavy inoculation of an *R* strain into a susceptible animal causes a change to an *S* strain since any reverse mutant (that is, a mutation from *R*) *S* cell can multiply, whereas the *R* cells are consumed by phagocytes. Pretreatment of the injected *R* cells with an extract of *S* cells or growing *R* cells in a medium with an extract of *S* cells induces reversion to an *S* strain, serologically identical with the one used to prepare the extract. *See* LYSOGENY; PNEUMOCOCCUS.

Natural defenses against infections depend in part on serological mechanisms. When bacteria enter animal hosts containing antibodies against them, the bacteria become coated with antibodies and are then susceptible to engulfment and digestion (phagocytosis) by host cells. Chicks, mice and rats, aseptically removed from the shell or uterus, can be reared bacteria-free. These axenic animals, when mature, are highly susceptible to infection by bacterial types harmless to normal animals. Antibodies carried over from the mother protect very young animals and, by the time maternal antibodies are depleted, normally reared offspring have developed their own. The isolation of axenic animals deprives them of the bacterial antigens necessary for development of protective antibodies. Every ancestor of living organisms survived bacterial attacks to reach maturity. The resistance, selected in this manner, is the factor most commonly concerned in defense against infections. *See* DISEASE.

Bacterial Variation

Variation in the characteristics of a single cell occurs during cell division, but since cells in a culture divide at different times, unless artificially synchronized, the average for the entire cell population is constant, as long as the environment is constant. A change in any limiting factor in the environment causes a change in the population. In nature environmental changes are often cyclic, and bacteria undergo accompanying changes in morphology interpreted by some as a life cycle. The applicability of this term to bacteria has been disputed.

An extremely small proportion of living organisms undergo sudden genetic changes, usually involving only one characteristic of the cell, which are transmitted through many generations. Because of the tremendous number of bacteria (1 in.^{3} or 16 ml of a culture may contain 50 billion bacteria), their mutations are common in cultures as well as in nature. If, in a given environment, a mutation enables its possessor to grow and divide more rapidly than the type from which it arose, the mutants ultimately predominate. The characteristics of the population are changed by the environment through selection of cells most fitted to survive, rather than by a direct action on all cells as discussed in the preceding paragraph. *See* MUTATION.

Interrelationships

Interrelationships may be close and may involve particular species. Examples are the parasitic association of many bacteria with plant and animal hosts, and the mutualistic association of nitrogen-fixing bacteria with leguminous plants, of cellulolytic bacteria with grazing animals, and of luminous bacteria with certain deep-sea fishes. *See* NITROGEN FIXATION; POPULATION ECOLOGY.

Bacteria are also active in other less intimate, but no less important, natural interrelationships. Bacterial decomposition of the dead bodies of animals, and especially plants, releases for reuse by living plants the carbon dioxide needed in photosynthesis. Many other chemical activities relate bacteria to other organisms through the world pool of materials essential to life, to which all organisms contribute and from which they draw their food. Robert E. Hungate

Gas Vesicles and Vacuoles

Gas vesicles are submicroscopic structures of cylindrical shape with conical ends. They were discovered in electron micrographs of gas vacuole–containing cyanobacteria. These studies revealed that the gas vesicles are the structural units which make up the gas vacuoles recognizable within the cells by ordinary light microscopy. Gas vesicles occur exclusively in prokaryotic microorganisms. *See* VACUOLE.

Characteristics. The gas vacuoles were first described for purple sulfur bacteria in 1888 and called hollow cavities; they were named gas vacuoles in 1895, when it was observed that they conferred buoyancy to the cells containing them. In the light microscope, gas vacuoles appear in the cytoplasm as refractile hollow cavities of irregular shape and pinkish shine.

The gas vesicles are homologous structures in all prokaryotic organisms. The size of the vesicles varies considerably in different systematic groups. In the cyanobacteria and green sulfur bacteria, the vesicles are about 70 nanometers in diameter and, on the average, 400 nm long (maximum length up to 1 micrometer). The gas vesicles of the purple sulfur bacteria and of several species and genera of chemotrophic bacteria are 100–200 nm wide but only up to 300 nm long. The widest gas vesicles, with a diameter of up to 300 nm, occur in the halobacteria.

The gas vesicle membrane is 2 nm thick and exhibits a ribbed fine structure from both the outside and the inside. The ribs are 4.5 nm wide and lie perpendicular to the long axis of the vesicle cylinder. The ribs apparently represent turns of a shallow spiral rather than stacks of concentric rings. The vesicle membrance, permeable to gases but impermeable to water, encloses a hollow space into which gases diffuse freely.

The gas vesicles are fairly rigid structures. However, when suspensions of cells with gas vacuoles are subjected to a sudden increase in pressure, the gas vesicles collapse and lose their gas-filled hollow spaces. Consequently, the gas vacuoles are no longer detectable by light microscopy and the cells lose their buoyancy.

Composition and molecular structure. Intact gas vesicles can readily be isolated from gently lysed cells by centrifugation, which causes the vesicles to float to the supernatant surface, where they are skimmed off. Chemical analysis of isolated gas vesicles from cyanobacteria and halobacteria shows that the vesicles consist exclusively of protein. Only one type of protein was found in the vesicles of cyanobacteria, while the vesicles of *Halobacterium* consisted of two very similar protein types. The molecular weights of the vesicle proteins are between 13,000 and 15,000. The amino acid composition of the vesicle proteins is fairly similar in all species. The proteins have in common a high proportion (more than 50%) of hydrophobic aliphatic amino acids and a low proportion of aromatic amino acids; both cysteine and methionine are absent. *See* CENTRIFUGATION.

It was established by x-ray and neutron diffraction studies that the gas vesicle membrane consists of a monolayer of the vesicle protein. Since the outside of the vesicles is hydrophilic and wettable while the inside is hydrophobic, the protein molecules must be positioned so that their hydrophobic aliphatic amino acids are located toward the inner surface of the vesicles. It is assumed that the hollow space of the vesicle arises simultaneously with formation and enlargement of the vesicle. The gases in the surrounding cytoplasm diffuse into the hollow space, while the hydrophobic inner surface of the vesicle prevents the penetration of water as well as the formation of water droplets inside the structure. *See* CELL MEMBRANES.

Function. Good evidence exists for only one of several suggested biological functions for gas vesicles. The hollow space of the vesicles effectively decreases the specific weight of the cells and, therefore, provides buoyancy. Buoyancy can be of selective advantage only in nonturbulent aquatic habitats. In agreement with this, gas-vacuolated microorganisms are mainly encountered as planktonic cells in lakes and stagnant water bodies, while they are rarely found in soils.

Since the vesicle membrane is permeable to gases, the pressure of the gas in the vesicles is always equal to the atmospheric pressure at the surface of the aquatic habitat. The rigid structure of the vesicle membrane maintains the hollow space of the vesicle against both the turgor pressure of the cells and the hydrostatic pressure of the water at the depth in which the cells float. Depending on the critical pressure that is tolerated by the vesicle membrane of a given species, a kind of buoyancy regulation is feasible by either formation or collapse of gas vesicles. For certain planktonic cyanobacteria (for example, *Oscillatoria agardhii*) which develop their maximum population density at a certain depth in stratified lakes, buoyancy regulation has been experimentally established. When at growth-limiting light intensities the turgor pressure of the cells is low, newly forming gas vesicles cause the cells to rise in the water column. At higher light intensities in the vertical light gradient, a rising turgor pressure may cause part of the gas vesicles to collapse and, consequently, cause the cells to stop rising or even to sink. At the population maximum, gas vesicles render the cells neutrally buoyant.

Under certain conditions in nutrient-rich eutrophic lakes, gas-vacuolated cyanobacteria may become overbuoyant by an excess of gas vesicles. In this case, masses of cells appear at the surface and form a water bloom. *See* EUTROPHICATION.

Norbert Pfennig

Endospores

Endospores are resistant and metabolically dormant bodies produced by the gram-positive rods of *Bacillus* (aerobic or facultatively aerobic), *Clostridia* (strictly anaerobic), by the coccus *Sporosarcina*, and by certain other bacteria. Sporeforming bacteria are found mainly in the soil and water and also in the intestines of humans and animals. Some sporeformers are found as pathogens in insects; others are pathogenic to animals and humans. Endospores seem to be able to survive indefinitely. Spores kept for more than 50 years have shown little loss of their capacity to germinate and propagate by cell division.

The endospore appears as a light-refractile body inside another cell (sporangium), as shown in **Fig. 1**. Each sporangium produces one endospore with a characteristic size, shape, and position within the cell.

Structure and constituents. The mature spore has a complex structure which contains a number of layers. The outermost envelope, surrounding the spore, is called the exosporium and is a thin, membranous covering. Beneath the exosporium lies the spore coat, which is composed of several layers, largely of a protein nature; each is about 2–2.5 nm thick. Beneath these is a thin membrane which separates the spore coat from an area of low electron density called the cortex. The cortex is primarily a modified peptidoglycan structure. It occupies approximately half the volume of the spore. A wall and a thin membrane separate the cortex from the cytoplasm of the dormant spore. The internal structures of the spore appear similar to those seen in the cytoplasm of vegetative cells, as shown in **Fig. 2**.

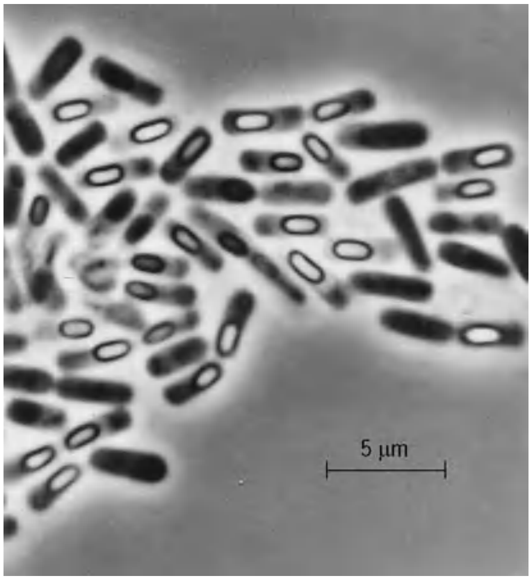


Fig. 1. Sporulating cells of *Bacillus cereus*. (*Photomicrograph by F. C. Fitz-James*)

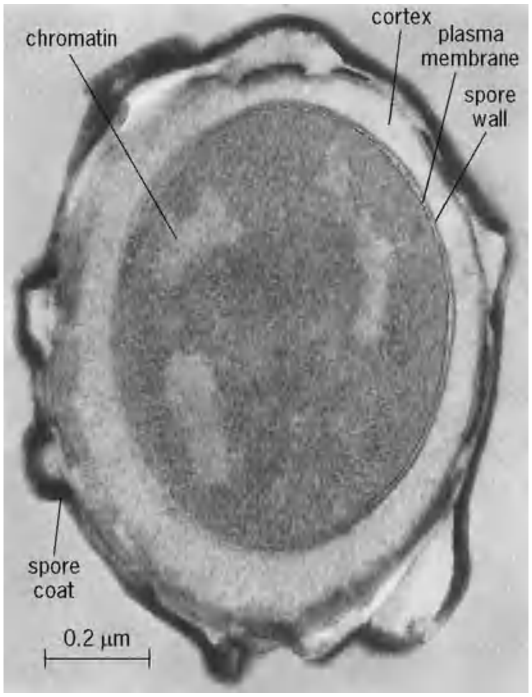


Fig. 2. Section of resting spore of *Bacillus megaterium*. (*Electron micrograph by C. F. Robinow*)

The unique properties of bacterial spores are their extreme resistance to heat, radiation from ultraviolet light and x-rays, organic solvents, chemicals, and desiccation. The most conspicuous chemical component is a chelating agent called dipicolinic acid (2,6-pyridine dicarboxylic acid), which constitutes 5–15% of the dry weight. This compound is absent from vegetative cells. Spores also differ from vegetative cells in containing higher levels of divalent metals and disulfide bonds and little, if any, free water.

Formation. The capacity of a bacterial cell to form a spore is under genetic control, although the total number of genes specific for sporulation is not known. The actual phenotypic expression of the spore genome depends upon a number of external factors. For each species of sporeforming bacteria, there exist optimum conditions for sporogenesis which differ from the optimal conditions for vegetative growth. These conditions include pH, degree of aeration, temperature, metals, and nutrients. Limitations in a variety of substances in the medium can initiate the process of sporulation. Sporulation is an ordered sequence of morphological and biochemical events leading to the formation of a mature spore. Commonly observed stages of sporulation are, in order of appearance, antibiotic release, cortex synthesis, increased light refractility, dipicolinic acid synthesis, formation of coats, resistance to octanol, resistance to heat, and release of the spore from the mother cell.

Sporulation is strongly influenced by the carbon and nitrogen sources available. Metabolizable nitrogen compounds generally repress sporulation. Glucose, in the presence of an available nitrogen source, effectively represses the initiation of spore formation; in general, carbon and nitrogen sources which are rapidly metabolized favor vegetative growth, whereas carbon and nitrogen sources which are more slowly metabolized stimulate spore formation.

Breaking of dormancy. The three processes involved in the conversion of the spore into a vegetative cell are (1) activation (usually by heat or aging), which conditions the spore to germinate in a suitable environment; (2) germination, an irreversible process which results in the loss of the typical characteristics of a dormant spore; and (3) outgrowth, in which new classes of proteins and structures are synthesized so that the spore is converted into a new vegetative cell.

Germination. Germination is an irreversible process in which a number of simultaneous events take place, shortly after the exposure of activated spores to specific stimulants (amino acids, sugars, and nucleotides). Germination is accompanied by a swelling of the spore, either rupture or absorption of the spore coat, and loss of a number of typical properties of the spore. Among these last-mentioned events are a loss of resistance to environmental stress, a loss of refractility, an increase in permeability, a release of spore components (dipicolinic acid, calcium, spore peptides), and an increase in metabolic activity. As a whole, the process is degradative and probably involves a number of enzymes.

Outgrowth. Germination is followed by a period of biosynthetic activity called outgrowth. The endospore coat breaks, and the cortex disappears (degraded by hydrolytic enzymes). A new cell emerges from the spore coat and eventually matures into a vegetative form, as shown in **Fig. 3**. During this period, new proteins and structures characteristic of vegetative cells are synthesized. Outgrowth terminates at the time of cell division and return to vegetative growth. The conditions for outgrowth are usually different from those supporting germination. Germination and outgrowth have different optimum temperatures, and most spores need nutrients for outgrowth that are not required for germination.

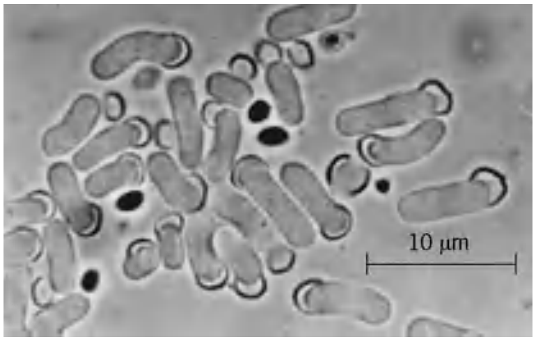


Fig. 3. Germination of spores of *Bacillus megaterium*. (*Photomicrography by F. C. Fitz-James*)

Microcycle sporulation. Spores may be germinated under nutrient-limited conditions. When faced with insufficient nutrients for proliferation, the new vegetative cell does not divide but instead enters the sporulation process, a shortcut leading to the formation of a mature spore.

Harlyn O. Halvorson; Keith Hutchison; Cristian Orrego

Appendages

On the basis of structure and function, bacteriologists recognize several types of bacterial appendages: flagella, pili, acellular stalks, and prosthecae.

Flagella. A bacterial flagellum has three parts, a long helical filament (about 0.01 micrometer in diameter), a short proximal hook (about 0.05 μm long), and a basal body, composed of a rod and a set of rings (about 0.02 μm in diameter) embedded in the cell wall and cytoplasmic membrane. The filament is a polymer or crystal composed of several thousand subunits per turn. Its shape can be changed by altering the structure of the subunit (the protein flagellin), the pH, ionic strength, temperature, or dynamic load. It dissolves in acids or in bases or when heated. Cultured in a solution containing pure monomeric flagellin, the filament grows in one direction at a constant rate. When attached to the living cell, it grows at the free end at a rate that decreases exponentially with length. Evidently, monomeric flagellin passes through the center of the filament and crystallizes on it at the distal end. The proximal hook is made up of a second protein, and the basal body is built up of another dozen or so different proteins. In a gram-positive bacterium, the basal body has two rings, an M ring embedded in the cytoplasmic membrane and an S ring found just outside this membrane. Common gram-negative bacteria have two additional rings, thought to serve as a bushing that carries the rod through the outer membrane. The M ring appears to be the rotor, the S ring the stator, the rod the drive shaft, the proximal hook a flexible coupling, and the filament a propeller.

Howard C. Berg

Pili. The pilus is also a proteinaceous appendage but differs from the flagellum in that it has a hollow core, is generally finer (ranges in width from 3 to 30 nm), and does not cause motility. Nonetheless, several important functions have been attributed to this structure. The most thoroughly documented function performed by pili is their role in bacterial sexual conjugation. The “male” cell of strains capable of conjugation produces a sex pilus that enables it to attach to an appropriate “female” cell containing the specific receptor site for the pilus. Only when cells of these two mating types are physically attached by the pilus can genetic material be transferred from the male to the female cell. *See* BACTERIAL GENETICS.

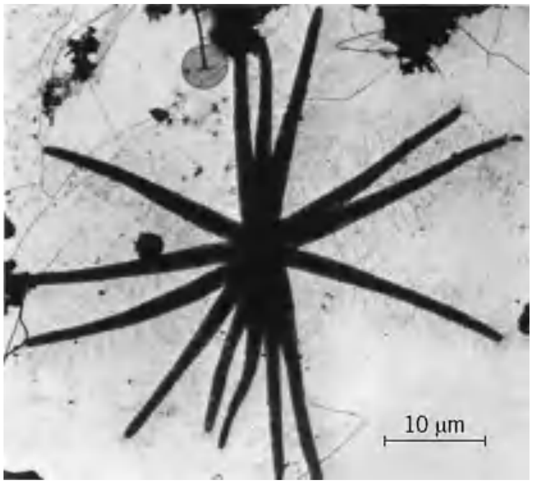


Fig. 4. Electron micrograph of a rosette-forming bacterium found in ponds and lakes.

Pili have also been implicated in the attachment of bacteria to unrelated organisms. For example, piliated strains of *Neisseria gonorrhoeae*, the bacterium which causes the venereal disease gonorrhea, are more frequently pathogenic than nonpiliated strains. It is thought that the piliated strains can use the pili to attach more strongly to host tissues. Certain viruses that infect bacteria attach to specific pili during the initial stages of infection.

Immotile aquatic bacteria commonly possess pili. **Figure 4** shows an unnamed aquatic bacterium that forms rosettes (the rosette shown had 14 cells which are joined together at a common center). Each cell has more than 100 very fine pili emanating from it. The role of these structures in such aquatic bacteria is not known.

Acellular stalks. Flagella and pili are too fine to be observed using ordinary light microscope techniques of observation. Most acellular stalks, on the other hand, are sufficiently wide to be seen by these classical procedures. For this reason it should not seem surprising that stalks were observed long before pili and flagella were seen. Indeed, the stalked bacterium *Gallionella ferruginea* was one of the first bacteria described. Interestingly, however, early investigators thought that the stalk was the bacterium because it was enormous compared to the small, bean-shaped cell which was readily dislodged from its position at the tip of the stalk. *Gallionella* is commonly found in iron springs, where the stalk becomes heavily encrusted with iron oxides that impart a rust-colored appearance. Although these bacteria have been grown in pure culture, the difficulties encountered in cultivating them have precluded extensive studies of their biology. Pure-culture studies indicate that the stalk is made up of several small fibrils, possibly pili, that are extruded during growth from the concave side of the cell. The iron appears to be precipitated in a sheath that surrounds the fibers.

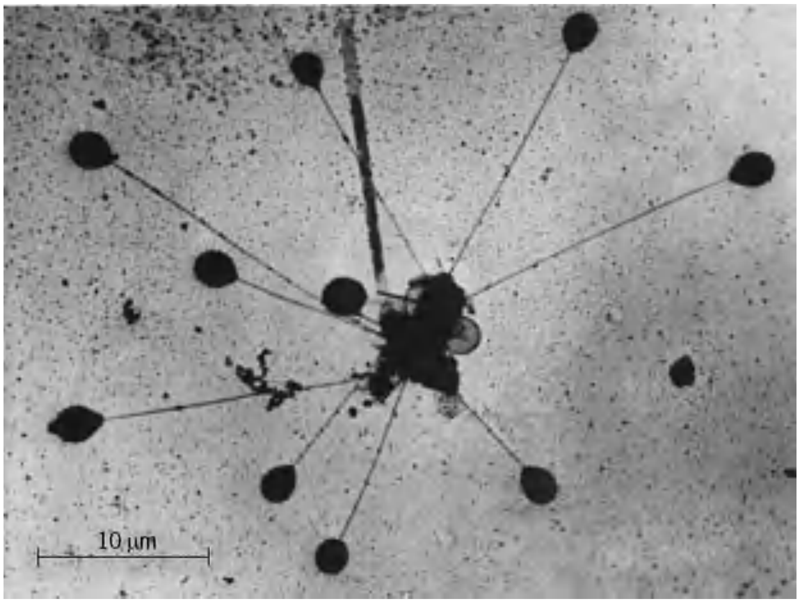


Fig. 5. Electron micrograph of a rosette of a budding bacterium of the *Blastocaulis-Planctomyces* group; note small bud on cell at lower left. Pili can be seen emanating from each cell.

The *Blastocaulis-Planctomyces* group also has members with acellular stalks. One example is shown in **Fig. 5**. This is a rosette containing nine ovoid cells borne at the tips of acellular stalks which are connected together at a common center. The cells reproduce by budding at the opposite, nonstalked pole of the cell. Note also that each cell has numerous pili in addition to the single stalk.

In both of the examples cited above, the appendage is an excretion of the cell and for this reason is termed extracellular or acellular.

Prosthecae. Unlike acellular stalks, the prostheca is an appendage that is actually part of the cell; that is, it is bounded by some or all of the layers of the cell envelope (cell wall and cell membrane). Frequently, species that have these appendages also have nonprosthecate cells that serve as stages in the life cycle of the organism. These nonprosthecate cells are usually motile by flagella and undergo predictable developmental stages during which prosthecae differentiate and daughter cells are formed. Thus, these bacteria are among the simplest unicellular organisms in which cellular developmental processes can be studied at the molecular level.

There are two hypotheses which have been advanced by workers in this field to explain the function of prosthecae: (1) These structures serve to act as "wings" by preventing the cells that have them from settling out of the water column in aquatic habitats; and (2) by increasing the membrane surface area of the organism, prosthecae enable it to take up nutrients more quickly in the dilute environments in which they reside.

To provide some insight into the nature of these bacteria, the life cycles of several of those most thoroughly studied are described below.

Caulobacter. The best-known genus of prosthecate bacteria is *Caulobacter*. These bacteria have a single prostheca, termed a stalk, that extends from one end of the cell (**Fig. 6**). *Caulobacter* cells undergo division at their nonstalked pole by binary fission to

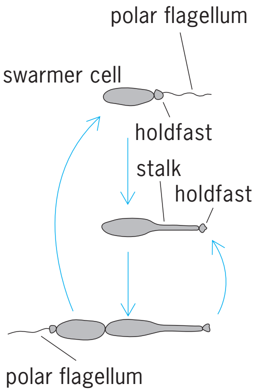


Fig. 6. Life cycle of *Caulobacter* species.

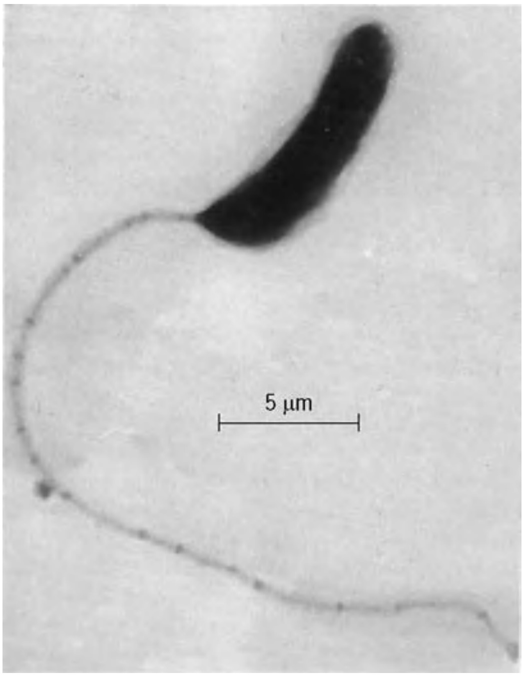


Fig. 7. Electron micrograph of a *Caulobacter* cell showing numerous crossbands on the prostheca; each crossband was formed when a daughter cell was produced.

produce a daughter cell that has a single flagellum. The daughter cell, which has no stalk, becomes motile and separates from the mother cell. After separation, the daughter cell, called a swarmer cell because of its motility, normally attaches to a solid substratum by a sticky holdfast material located at the base of the flagellum. In time the cell loses its motility and synthesizes a stalk at the flagellum-base position so that the holdfast is borne at the tip of the stalk. It is now a mature cell that elongates and produces swarmer cells of its own.

It has been discovered that the age of a *Caulobacter* cell can be estimated by counting the number of crossbands in the stalk of the cell. These crossbands can be seen when cells are observed with the electron microscope (**Fig. 7**). Apparently, each time a cell undergoes binary fission, the mother cell synthesizes a crossband in its stalk. Therefore the age of the cell can be estimated by counting these structures in the same manner that the annual rings of a tree can be used to determine its age. This is the only bacterium whose age can be determined by direct examination of the organism.

Hyphomicrobium. The genus *Hyphomicrobium* is distinctive because of its prosthecae and because of its division by budding. Like *Caulobacter*, it undergoes, for a bacterium, a rather complex life cycle including flagellated, nonprosthecate cells and prosthecate, nonmotile cells (**Fig. 8**). Newly formed oval buds are motile by subpolar flagella. Some species have holdfast material associated with the cell that permits them to attach to particulate material. Eventually the buds lose their motility and develop prosthecae, invariably from one of the cell ends. The tip of the prostheca enlarges to form a bud which separates from the mother cell and undergoes the same cycle. The mother cell, however, has several options available to it: (1) It may produce another bud from the same site at the tip of its prostheca; (2) it may produce another prostheca at the same or opposite pole of the cell and form a bud at its tip; or (3) it may produce a branch from its already existing prostheca and form a bud at its tip. Therefore, the life cycle of this bacterium is more complex than that of *Caulobacter*.

Prosthecomicrobium. Bacteria in the genus *Prosthecomicrobium* have approximately 20 conical prosthecae that extend in all directions from the cell, giving the organism the appearance of a bur of a cocklebur plant when observed in the microscope (**Fig. 9**). The appendages vary in length from one species to another, being as short as 0.2 μm or as long as 1.0 μm, or more. The bacteria reportedly divide by binary fission. Although both motile and nonmotile cells are found in some strains, not all strains have

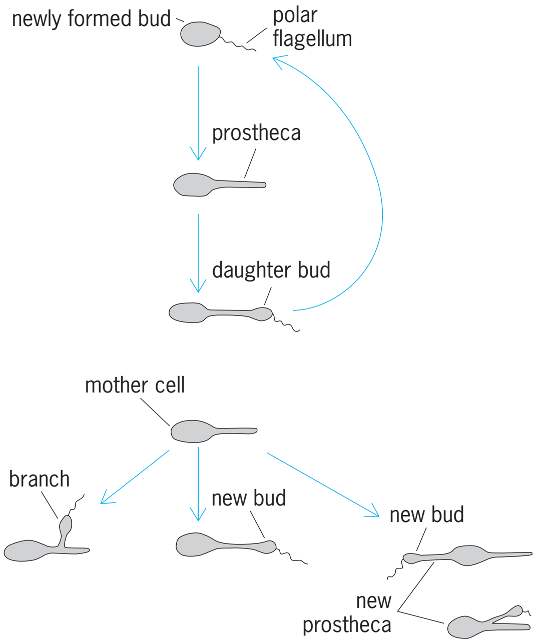


Fig. 8. Life cycle of *Hyphomicrobium* species.

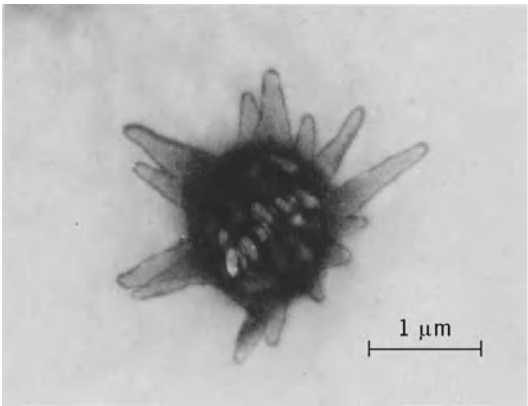


Fig. 9. Electron micrograph of *Prosthecomicrobium pneumaticum*. Note the 14 prosthecae extending from the cell, and the transparent gas vesicles inside the cell.

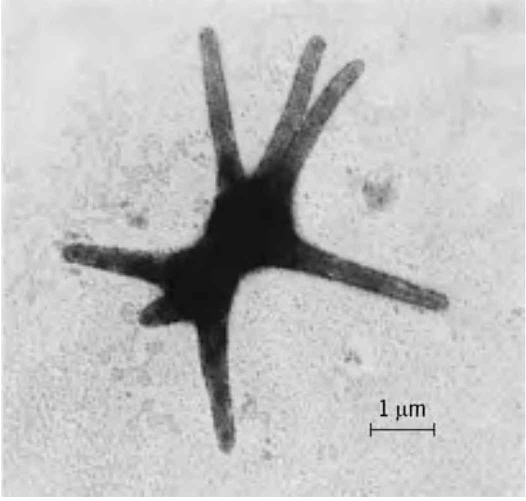


Fig. 10. Electron micrograph of *Ancalomicrobium adetum*. The cell has seven prosthecae.

motile stages, and in those that do, it is not known whether the motile stage is part of the life cycle, as in the case of *Caulobacter* and *Hyphomicrobium*.

Ancalomicrobium. Like *Prosthecomicrobium* cells, cells of the genus *Anacalomicrobium* have several prosthecae per cell, although the number rarely exceeds eight (**Fig. 10**). This nonmotile bacterium produces an outgrowth from one position on the cell surface. This outgrowth, or bud, differentiates to form from two to four prosthecae, each about 3 μm long. These appendages occasionally form branches analogous to those of *Hyphomicrobium*, but buds are not produced at the tips. James T. Staley

Motility

Motile bacteria swim or glide. Common gram-positive and gram-negative organisms swim by rotating flagellar filaments that project from the surface of the cell and extend several micrometers into the surrounding medium. Spirochetes, another group of gram-negative organisms, swim by rotating filaments that run between the protoplasmic cylinder and the outer membrane. Gliding bacteria, a third major group of gram-negative organisms, do not swim or have flagella but creep when in contact with solid surfaces.

Flagellar motion. Each flagellum is driven at its base by a reversible rotary motor; the filament rotates rigidly clockwise or counterclockwise; it does not wave or beat. To demonstrate this, a cell can be fixed to a glass slide by one of its flagellar filaments: the cell body spins alternately clockwise and counterclockwise. Mutants exist that spin in only one direction. Their motion has been followed for millions of revolutions; therefore, a bacterial flagellum does not wind up and unwind—it truly rotates. The flagella are not powered by the high-energy phosphate compounds that drive muscles, but rather by a proton flux. The passage of about a thousand protons from the outside to the inside of the cell carries the motor through one revolution. When connected to the proximal hook, the motor spins about 100 revolutions per second; when also connected to the filament, it spins about half as fast. When loads are light, the motor appears to run at constant speed; when loads are heavy, it runs at constant torque.

The style of locomotion of a bacterium depends on the size and shape of its body and its mode of flagellation. A rod-shaped cell with one or more flagella at either pole (monotrichous or lophotrichous flagellation, either monopolar or bipolar) swims forward or backward in the direction of its long axis. The body rotates (rolls) in a direction opposite to the direction of rotation of the flagella. Thrust generated by the flagellar filaments is balanced by viscous drag due to translation of the cell body, and torque generated by the filaments is balanced by viscous drag due to the roll.

If the cell body is corkscrew-shaped (as in *Spirillum*), the roll also contributes to the thrust. This is the primary mechanism for locomotion in most spirochetes (for example, the genus, *Spirochaeta*), which are long thin helical organisms with loosely attached outer membranes. The rotation of the filaments underneath this membrane is thought to cause the surface of the cell to move circumferentially. Viscous drag, in turn, causes the cell body to roll about its long axis, and the cell screws its way through the medium. This mode of locomotion is particularly effective in gel-like media. *Leptospira*, having a more tightly fitting outer membrane, generates torque by gyrating its ends—the anterior end in a spiral configuration, the posterior end in a hook configuration; additional thrust is gained from propagation of the anterior spiral wave.

Rod-shaped cells with flagella arising at random points on their sides (peritrichous flagellation) swim as in a random walk. When their flagella spin counterclockwise, they coalesce into a synchronous bundle that pushes the cell steadily forward: the cell is said to run. When the flagella spin clockwise, the bundle flies apart, and the flagella turn independently, moving the cell this way and that in a highly erratic manner: the cell is said to tumble. These modes alternate, and the cell executes a three-dimensional random walk. A variety of species swim in this way, including

Bacillus subtilis, *Salmonella typhimurium*, and *Escherichia coli*.

Taxis. Bacteria change the direction in which they swim in response to changes in their environment: they accumulate in regions that they find favorable. Different cells respond to light, heat, oxygen, and a variety of chemicals; and are said to be phototactic, thermotactic, aerotactic, and chemotactic, respectively. Cells of *E. coli*, for example, swim into capillary tubes filled with dilute solutions of simple sugars and amino acids. They do this by biasing their random walk. Runs that happen to carry a cell up the gradient are extended, while those that happen to carry it down the gradient are not. The cells monitor concentrations as a function of time; they compare the occupancy of specific receptors over the past second or so with the occupancy a few seconds before that. The memory required for this comparison involves receptor carboxymethylation. If current receptor occupancy exceeds past occupancy, as indicated by a relatively low level of methylation, the probability that the flagella spin counterclockwise increases and the probability that they spin clockwise decreases.

Gliding. Cells of some gliding bacteria, for example, *Myxococcus*, aggregate to form fruiting bodies. This behavior requires intracellular communication; the way in which one cell glides depends on its contacts with other cells. *Myxococcus* glides very slowly, about 2 μm per minute; other gliding bacteria, for example, *Cytophaga*, glide more rapidly, up to 2 μm per second. The organelles of locomotion for these bacteria have not been identified. Howard C. Berg

Bibliography. J. Adler, The sensing of chemicals by bacteria, *Sci. Amer.*, 234(4):40–47, 1976; R. P. Burchard, Gliding motility of prokaryotes: Ultrastructure, physiology and genetics, *Annu. Rev. Microbiol.*, 35:497–529, 1981; E. Canale-Parola, Motility and chemotaxis of spirochetes, *Annu. Rev. Microbiol.*, 32:69–99, 1978; I. I. Gunsalus and R. Y. Stanier (eds.), *The Bacteria*, 7 vols., 1960–1979; G. L. Hazelbauer and S. Harayama, Sensory transduction in bacterial chemotaxis, *Int. Rev. Cytol.*, 81:33–70, 1983; J. G. Holt (ed.), *The Shorter Bergey's Manual of Determinative Bacteriology*, 8th ed., 1977; J. M. Parry et al., *A Colour Atlas for the Identification of Bacillus Species*, 1983; E. M. Purcell, Life at low Reynolds number, *Amer. J. Phys.*, 45:3–11, 1977; R. Y. Stanier et al., *An Introduction to the Microbial World*, rev. ed., 1979; A. Sussman and H. O. Halvorson, *Spores: Their Dormancy and Germination*, 1966.

Bacterial genetics

The study of gene structure and function in bacteria. Genetics itself is concerned with determining the number, location, and character of the genes of an organism. The classical way to investigate genes is to mate two organisms with different genotypes and compare the observable properties (phenotypes) of the parents with those of the progeny. Bacteria do not mate (in the usual way), so there is no way of getting all the chromosomes of two different bacteria into the same cell. However, there are a number of ways in which a part of the chromosome or genome from one bacterium can be inserted into another bacterium so that the outcome can be studied. *See* GENETICS.

Characteristics of bacteria. The ways of classifying organisms are rapidly changing. Similarities in nucleotide sequences, particularly of the ribosomal ribonucleic acid (rRNA) sequences, have revealed new aspects of the evolutionary tree. All organisms have diverged from a common ancestral prokaryote whose precise location in the evolutionary tree is unclear. This has resulted in three primary kingdoms, the Archaebacteria, the Eubacteria, and the Eukaryotae. All bacteria are prokaryotes, that is, the "nucleus" or nucleoid is a single circular chromosome, without a nuclear membrane. Bacteria also lack other membrane-bounded organelles such as mitochondria or chloroplasts, but they all possess a cytoplasmic membrane. Most bacteria have a cell wall that surrounds the cytoplasmic membrane, and some bacteria also contain an outer membrane which encompasses the cell wall. Duplication occurs by a process of binary fission, in which two identical daughter cells arise from a single parent cell. Every cell in a homogeneous population of bacterial cells retains the potential for duplication. Bacteria do not possess the potential for differentiation (other than spore formation) or for forming multicellular organisms. *See* ARCHAEA; BACTERIA; PROKARYOTAE; RIBONUCLEIC ACID (RNA).

One of the most frequently used organisms in the study of bacterial genetics is the rod-shaped bacillus *Escherichia coli*, whose normal habitat is the colon. Conditions have been found for growing *E. coli* in the laboratory on defined media, and it is by far the best understood of all microorganisms. The single circular chromosome of *E. coli* contains about 4.5×10^6 base pairs, which is enough to make about 4500 average-size genes (1000 base pairs each). The relative positions of almost half of these genes are known. In regions where mapping studies are reasonably complete, the impression is obtained of an efficiently organized genome. Protein coding regions are located adjacent to regulatory regions. There is no evidence for significant stretches of nonfunctional deoxyribonucleic acid (DNA), and there is no evidence for introns in the coding regions. The coding regions from eukaryotic cells frequently contain introns, that is, regions that are removed by splicing the messenger RNA (mRNA) before it is translated into protein. Very little repetitive DNA exists in the *E. coli* chromosome other than the seven sequence-related rRNA genes that are dispersed at different locations on the chromosome. *See* CHROMOSOME; DEOXYRIBONUCLEIC ACID (DNA); GENETIC CODE.

Frequently, genes with a related function are tightly clustered. This clustering results from the fact that genes with a related function are commonly transcribed as a single polygenic (polycistronic) mRNA molecule. A chromosomal region transcribed in this way is called an operon. It contains a single site at one end called the promoter, which is where RNA polymerase binds just prior to the initiation of transcription. *See* OPERON.

Genetic methodology. The first step in performing genetic research on bacteria is to select mutants that differ from wild-type cells in one or more genes. Then crosses are made between mutants and wild types, or between two different mutants, to determine dominance-recessive relationships, chromosomal location, and other properties. Various genetic methods are used to select bacterial mutants, antibiotic-resistant cells, cells with specific growth requirements, and so on.

Selection for bacterial mutants. Spontaneous mutation frequencies in *E. coli* vary from about 10^{-7} to 10^{-9} mutations per genome per generation. Since it is possible to plate 10^7 to 10^9 bacteria on a single standard-size growth plate, a few growth plates should contain a number of mutants of the desired type. The next consideration is to find a suitable selection procedure˜ for detecting and isolating the desired mutants from a large, principally wild-type population. The two most popular selection techniques involve selecting for mutants with altered antibiotic resistance or nutritional requirements. *See* MUTATION.

Selection for antibiotic-resistant cells. Ordinary wild-type cells are sensitive to a number of bacteriostatic and bacteriocidal factors, including antibiotics, toxins produced by certain plasmid-containing cells, and bacterial viruses (bacteriophages). Mutants that are resistant to antibiotics are very useful in the analysis of biochemical pathways because the sites of action, that is, the target sites of such antibiotics, are often highly specific. For example, the target site for streptomycin action is the ribosome, as evidenced by the fact that streptomycin-resistant mutants usually have an altered ribosomal protein. *See* ANTIBIOTIC; BACTERIOPHAGE; DRUG RESISTANCE; PLASMID; TOXIN.

The usual procedure for isolating antibiotic-resistant mutants is simple. Approximately 10^8 cells grown in a liquid medium are spread onto a semisolid agar plate containing growth medium and the antibiotic. Only a few cells will survive and grow on the surface of the agar. These will give rise to clones, that is, a population of genetically identical cells. *See* CULTURE.

Selection for cells with specific growth requirement. Wild-type *E. coli* cells that can grow on minimal medium, such as one containing only glucose ammonium chloride, and other salts, are called prototrophs. By contrast, auxotrophs are cells that require supplementation of the minimal medium with one or more factors. The penicillin enrichment technique can be used to isolate an auxotroph by selecting against the prototroph. For example, a tryptophan auxotroph (*trp*$^-$) will grow on glucose-minimal medium to which the amino acid tryptophan is added. To select for a *trp*$^-$ strain, a large number of cells are grown on glucose-minimal medium in the presence of penicillin. Penicillin is a bacteriocidal agent that kills only growing cells by inhibiting one or more reactions in cell-wall synthesis. The *trp*$^-$ cells, or for that matter any auxotrophs that cannot grow on glucose-minimal medium, will survive penicillin treatment even though they cannot grow in its presence. In fact, they survive precisely because they cannot grow. The wild-type cells that do grow in this medium will be killed by the penicillin treatment. After a suitable period, usually overnight, the cells are removed from the penicillin by centrifugation. They are resuspended in growth medium containing tryptophan and spread on an agar plate. Under these conditions the *trp*$^-$ auxotrophs, which have survived the exposure to penicillin on the minimal medium, will grow.

Back mutations or revertants from an auxotrophic to a prototrophic state may be directly selected for by growth on minimal medium. Starting from a large population of auxotrophs, the only cells capable of forming clones on minimal medium would be revertants. Revertants possess either the exact reverse mutation or another mutation that compensates for the initial mutation and permits growth in the absence of added amino acid or other growth requirements.

Regulatory gene mutations. Certain genes that have the function of modulating the expression of other genes are known as regulatory genes. Mutations that affect the action of regulatory proteins are of two types: those that occur in the genes that encode the regulatory proteins, and those that affect the genetic loci where the regulatory protein interacts to modulate the level of gene expression. *See* PROTEIN.

In the case of tryptophan synthase, which is one of the enzymes required for tryptophan synthesis, a detailed examination has shown that the genes for tryptophan synthase are part of a multigene cluster, or operon. Expression of the operon is regulated at the promoter which is located at one end of the *trp* operon. In the *trp* operon the promoter overlaps with another site known as the operator. Whereas the promoter serves as the locus for initial binding of RNA polymerase and initiation of transcription, the operator serves as the binding site for a so-called repressor protein. When the repressor protein is bound to the operator locus, it prevents the RNA polymerase from binding to the promoter locus, thereby inhibiting the initiation of transcription and gene expression. The repressor is encoded by an unlinked gene located at a distinct site called *trpR*.

Mutations that lead to decreased or enhanced expression of the *trp* operon have been observed and analyzed in great detail. Such mutations result from changes in the DNA at the promoter-operator locus or from changes in the structure of the repressor protein, but they never affect the structure of the enzymes encoded by the operon, only their amounts. Some regulatory gene mutations cause overproduction and some cause underproduction of gene products. This is the hallmark of a mutation that influences the functioning of a regulatory protein or regulatory factor–binding site; it affects the quantity but not the quality of other gene products. Furthermore, regulatory gene mutations are frequently pleiotropic, that is, they influence the rate of synthesis of several gene products simultaneously. *See* GENE ACTION.

Induced mutagenesis. Frequently, geneticists want to increase the number or types of mutants that can be obtained as a result of spontaneous mutagenesis. In such instances, they treat a bacterial population with a mutagenic agent to increase the mutation frequency. This is called induced mutagenesis. The simplest techniques of induced mutagenesis involve measured exposure of the bacteria to a mutagenic agent, such as x-rays or chemical mutagenic agents. Such procedures have a general effect on the increase in the mutation rate. *See* MUTAGENS AND CARCINOGENS.

More sophisticated procedures involve isolating the gene of interest and making a change in the desired location. This is called site-directed mutagenesis. The goal is usually to determine the effects of a change at a specific gene locus. The gene in question is isolated, modified, and reinserted into the organism. Discrete alterations can be made in a variety of ways on any DNA in cell-free culture, and the effect of such alterations can be subsequently tested in the organism. Most of the techniques used rely upon a site-specific restriction enzyme at some stage. The most popular procedures for directed mutagenesis are all variations of a particular procedure. This procedure begins with the limited degradation of one DNA chain of a duplex to yield a gapped molecule. The gapped DNA is annealed to a synthetic oligonucleotide containing discrete differences from the original DNA. After completing the ligation process, the resulting heteroduplex can be reinserted into the organism, and by suitable selection procedures mutants can be isolated that contain specific changes dictated by the sequence in the synthetically produced oligonucleotide. *See* GENETIC ENGINEERING.

Gene mapping. Bacteria do not mate to form true zygotes, but they are able to exchange genetic information by a variety of processes in which partial zygotes (merozygotes) are formed. The first type of genetic exchange between bacteria to be observed was transformation. Naturally occurring transformation involves the uptake of DNA. This phenomenon is observed only for a limited number of bacterial species and is a relatively difficult technique to use for gene manipulation. For these reasons and because more powerful approaches to gene mapping were discovered soon afterward, transformation has received relatively narrow use as a means for gene mapping. In 1946 direct chromosomal exchange by conjugation between *E. coli* cells was discovered by J. Lederberg and E. Tatum, and in 1951 transduction, the virus-mediated transfer of bacterial genes, was discovered. Both conjugation and transduction provide facile, generally applicable methods for moving part of the bacterial chromosome from one cell to another. The discovery of bacterial transposons in the 1970s has been useful in marking and mobilizing genes of interest. The purely genetic approaches to mapping have been supplemented by the biochemical approaches of hybrid plasmid construction and DNA sequence analysis. A detailed genetic and physical map of the entire bacterial chromosome is within grasp. *See* GENETIC MAPPING; TRANSFORMATION (BACTERIA).

Conjugation. The naturally occurring exchange of genetic information in *E. coli* bacteria takes place between cells of different mating types. Donor cells are defined as bacteria capable of transferring genes, and recipient cells are defined as bacteria capable of accepting and integrating such genes. The difference between a donor and a recipient cell is due to a plasmid known as F factor. The F factor is a circular DNA molecule with a molecular weight of about one-fortieth that of the bacterial host chromosome. Donor cells (F^+ cells) arise from bacteria that contain the F factor. Within an F^+ population, cells that transfer bacterial genes represent distinct variants. These variants, called Hfr's, have integrated the F factor into the host chromosome, and the site of integration determines the order of transfer of bacterial genes when an Hfr is mated with an F^- cell. The order of transfer of markers may be determined by interrupting the mating process at different times after mixing the two cell types. The so-called interrupted mating procedure has been carried out on many different Hfr's that have integrated the F factor at different locations and in either location. The results from such conjugations have permitted a rough mapping of the bacterial chromosome. Indeed, it was studies of this sort that led to the prediction that the bacterial chromosome was circular, a prediction that was verified by physical measurements many years later.

On rare occasions, the integrated F factor in the Hfr can reversibly excise to give back the original F^+ cell. On even rarer occasions, the F factor excises imperfectly, taking with it a part of the bacterial chromosome. Such F' factors have been used as a valuable source of discrete segments of the bacterial chromosome; they can readily be transferred to F^- cells. They have been most valuable in determining dominance-recessive relationships for alleles of the same genes.

Transduction. Transfer of bacterial segments can be done by sexduction, that is, the transfer of F' particles, but a far more convenient procedure for most applications involves transport by bacteriophages. Transfer of bacterial DNA by bacteriophages is called transduction. Transduction has provided the most valuable technique for bacterial gene mapping and for the isolation of discrete segments of the bacterial chromosome. A number of bacteriophages are used in transduction. They have different properties, and each one has particular advantages for gene manipulation. Some bacteriophages, such as the λ bacteriophage, transfer only specialized regions of the bacterial genome. For this reason, λ is referred to as a specialized transducing virus. Other bacteriophages, such as P1, which can transfer virtually any part of the bacterial genome are referred to as generalized transducing viruses. To obtain a P1 preparation that carries parts of a particular bacterial chromosome, the bacterial strain is infected with P1. The infection results in the production of a large number of normal P1 virus particles and a small mixed population of P1 transducing particles. The P1 transducing particles carry segments of the bacterial genome instead of the viral genome. When they infect a bacterial cell, the segment of bacterial genome enters the cell.

For a high-resolution mapping of bacterial genes, generalized transduction is a virtually indispensable tool. Bacterial genes that are transduced into *E. coli* by P1 transducing particles cannot replicate or survive unless they become integrated into the host genome. In so doing, they replace the homologous segment in the bacterial genome. Integration requires a minimum of two crossovers. In three-factor crosses, the middle marker may be identified by the least frequent recombinant class. Exclusive exchange of the middle marker requires two crossovers also. It occurs with a lower than average probability because the crossover points are restricted so that only the middle marker exchanges. *See* CROSSING-OVER (GENETICS); TRANSDUCTION (BACTERIA).

Transposons. Chromosomes might be thought of as being composed of two types of genetic elements: those that maintain a fixed location in a larger stable structure, and those that are designed to move from one location to another. The latter type is known by the general term mobile genetic element. Transposons are a class of mobile genetic elements that are commonly found in bacterial populations. Most transposons have three characteristic properties: they carry readily selectable marker genes; they insert at many different locations in the bacterial chromosome; and they are mutagenic in regions adjacent to the site of insertion. *See* TRANSPOSONS.

Expression of bacterial genes. At any given time, only a small percentage of the *E. coli* genome is being actively transcribed. The remainder of the genome is either silent or being transcribed at a very low rate. When growth conditions change, some active genes are turned off and other, inactive genes are turned on. The cell always retains its totipotency, so that within a short time (seconds to minutes), and given appropriate circumstances, any gene can be fully turned on. The maximal activity for transcription varies from gene to gene. For example, a β-galactosidase gene makes about one copy per minute, and a fully turned-on biotin synthase gene makes about one copy per 10 min. In the maximally repressed state, both of these genes express less than one transcript per 10 min.

The level of transcription for any particular gene usually results from a complex series of control elements organized into a hierarchy that coordinates all the metabolic activities of the cell. For example, when the rRNA genes are highly active, so are the genes for ribosomal proteins, and the latter are regulated in such a way that stoichiometric amounts of most of the ribosomal proteins are produced. When glucose is abundant, most genes involved in processing more complex carbon sources are turned off in a process called catabolite repression. If the glucose supply is depleted and lactose is present, the genes involved in lactose breakdown (catabolism) are expressed. In *E. coli* the production of most RNAs and proteins is regulated exclusively at the transcriptional level, although there are notable exceptions. Rapid response to changing conditions is ensured partly by a short mRNA lifetime, on the order of 1 to 3 min for most mRNAs. Some mRNAs have appreciably longer lifetimes (10 min or more), and the consequent potential for much higher levels of protein synthesis per mRNA molecule. These atypical mRNAs, at least in some instances, also may be subject to translational control. Finally, the fine-level control for any particular enzyme system is subject to regulation by activators or inhibitors directly modulating the enzyme activities.

Rel gene. The ribosome is the site of protein synthesis. Bacterial ribosomes are composed of 3 rRNAs and about 50 ribosomal proteins. *Escherichia coli* cells, under conditions of rapid growth, contain about 10,000 ribosomes. There are seven operons located at distinct and separate points on the bacterial chromosome which make similar rRNA transcripts. The transcription of the rRNA operons is controlled by the compound guanosine tetraphosphate, ppGpp, known as magic spot. This compound has pyrophosphates at the 3′ and the 5′ carbons of the ribose sugar. Under conditions of amino acid deprivation, the level of magic spot rises. It combines with the RNA polymerase, changing the structure of the RNA polymerase so that the latter has a low affinity for the promoters of the rRNA genes. The result is that less rRNA is made. The ppGpp compound has a similar effect on the expression of genes for the rRNA proteins. The mRNAs for the ribosomal protein genes are also controlled at the translation level by a feedback inhibition mechanism. These proteins normally combine with the rRNA to make ribosomes. However, if there is an excess of ribosomal proteins over the available rRNA, the ribosomal protein will combine with its own message so as to prevent further translation of the message. In this way the amount of ribosomal protein never reaches large excess but always is synthesized in slight excess over the available level of rRNA. *See* RIBOSOMES.

Gene regulation for DNA repair. The cell possesses an intricate system (the *din* genes) for regulating the synthesis of genes involved in DNA repair. The *din* genes are under the negative control of the lex repressor. Unlike most repressors, including the lac repressor, the lex repressor action is not reversibly controlled by complexing with a small molecule. Rather, it is broken down by a highly specific protease. This protease is carried by the recA protein and is not active under normal growth conditions. However, when the growing cells are exposed to ultraviolet light or other agents that damage DNA, the recA protease becomes active. It is believed that this is due to a complex formed between the recA protease and a DNA fragment. Once the protease becomes active, it breaks down all of the lexA protein in the cell. The genes repressed by lexA then become active and stay active until the DNA damage is repaired. After the DNA damage has been repaired, the recA protease returns to its normal inactive state, and the concentration of lexA repressor reaches a level where it silences the *din* genes and everything returns to normal. *See* BACTERIAL PHYSIOLOGY AND METABOLISM.

Geoffrey Zubay

Bibliography. C. R. Woese, Bacterial evolution, *Microbiol. Rev*., 51:221–271, 1987; G. Zubay, *Genetics*, 1987.

Bacterial growth

The processes of both the increase in number and the increase in mass of bacteria. Growth has three distinct aspects: biomass production, cell production, and cell survival. Biomass production depends on the physical aspects of the environment (water content, pH, temperature), the availability of resources (carbon and energy, nitrogen, sulfur, phosphorus, minor elements), and the enzymatic machinery for catabolism (energy trapping), anabolism (biosynthesis of amino acid, purines, pyrimidines, and so forth), and macromolecular synthesis [proteins, ribonucleic acid (RNA), and deoxyribonucleic acid (DNA)]. Cell production is contingent on biomass production and involves, in addition, the triggering of chromosome replication and subsequent cell division. The cells may or may not separate from each other, and the division may partition the cell evenly or unevenly. Alternatively, growth may occur by budding (unequal division). Most cells so produced are themselves capable of growing and dividing; consequently, viability is usually very high when growth conditions are favorable. Moreover, in many cases the incidence of death is surprisingly low in the absence of needed nutrients. Many bacteria differentiate into resistant resting forms (such as spores); others may simply reduce their rate of metabolism and persist in the vegetative state for long times.

Exponential balanced growth. If bacteria have been growing under well-mixed, low-cell-density, constant conditions for many generations, then cell-cell interactions are negligible and all aspects of the growth process come to have rates balanced so that all cell components double in a doubling time. However, because bacteria can grow so fast, it is usually necessary to keep diluting the culture to achieve such balanced growth in order to prevent the population from entering the stationary phase. The mathematics of balanced growth are readily expressed. Let B stand for the amount of bacteria per milliliter of culture. Clearly, if the bacteria behave independently of each other, growth would be twice as fast if B were twice as large. Therefore, as shown in Eq. (1),

$$\frac{dB}{dt} = rB \tag{1}$$

the rate of production is proportional to B. Thus bacteria grow in an autocatalytic way because the product of the reaction is the catalyst for more growth. The rate constant r is usually called the specific growth rate. (The symbol μ is used in certain subfields of microbiology.)

Integration of Eq. (1) between the limits $B = B_0$ and $B = B_t$, where B_0 is the cell concentration at time zero and B_t is the cell density at time t, yields Eq. (2), where e is the base of natural logarithms.

$$B_t = B_0 e^{rt} \tag{2}$$

Taking natural logarithms of both sides results in Eq. (3). The specific growth rate can be estimated

$$\ln B_t = \ln B_0 + rt \tag{3}$$

from plots of data according to Eqs. (2) or (3), and sophisticated statistical procedures can be employed. Often, the doubling time t_d is of interest. Setting B_t equal to $2B_0$ in Eq. (2) yields Eq. (4) or (5). The latter

$$2 = e^{rt_d} \tag{4}$$

$$\ln 2 = rt_d \tag{5}$$

provides the way to convert values of r into values of t_d, and vice versa.

Evidently, many bacteria have been selected in nature for the ability to grow rapidly in their natural habitat. Doubling time may be many hours; however, wild-type strains of *Escherichia coli* under optimal conditions can double once every 16 min ($t_d = 16$ min, $r = 2.6$/h). At a rate this fast, one cell would grow in one day into 10^{27} organisms. It is no wonder that exponential growth lasts only a short time.

Culture cycle. After an extended culturing period, cells have exhausted their external and internal resources and are usually in the stationary phase. If a culture in this phase is diluted in fresh nutrient-rich medium at a temperature at which growth can occur (**Fig. 1**), there is often a lag phase during which the cell number does not increase and cells are preparing

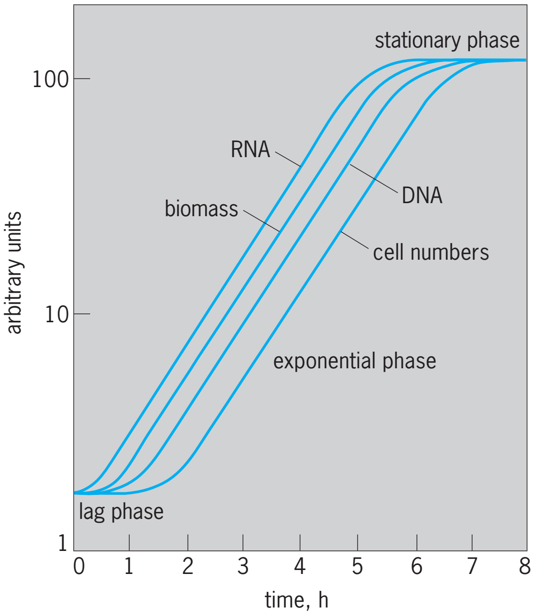


Fig. 1. Growth cycle, showing the consequence of diluting a stationary-phase culture into fresh medium.

for active growth. During the lag phase, cells are accumulating some metals and detoxifying others and synthesizing cofactors (containing vitamins). They are also synthesizing ribosomes and the factors necessary for protein synthesis and the enlargement of the cell. After some time, chromosomal DNA synthesis is initiated, followed by cell division. Eventually, the exponential balanced phase is reached. Some nutrient becomes limiting or some toxic product accumulates, and growth again slows and stops. If this transition is relatively gradual or if the cells have internal reserves, the bacteria may be able to respond morphologically to the impending “hard times” by forming spores or cysts, which are characteristic of the stationary phase. *Escherichia coli* continues DNA replication and cell division after the rate of protein synthesis slows; this results in more but smaller cells. In some cases, ribosomes and certain proteins are torn down to make components of the stationary cells. In early stationary-phase cultures, the vast majority of the countable cells are viable in the sense that they can give rise to a colony. Subsequently, a phase occurs in which viability is lost and the biomass is reduced by metabolism and autolysis. In nature, bacteria would be subject to digestion by protozoans and metazoans, or would be killed by physical conditions or host immunity mechanisms. *See* CELL CYCLE.

Dependence of growth rate on substrate. Many bacteria have very complex nutrition, but others grow on simple media. For the latter types to grow in varying environments, they must be capable of many molecular transformations to make all the cell’s building blocks from available resources. However, if the concentration of a single irreplaceable nutrilite (nourishing compound) is lowered, eventually the growth rate must fall. An approximate kinetic description (called the Monod model) used extensively in microbiology, ecology, and biotechnology is drawn from an analogy to enzyme kinetics [Eq. (6)]. K_m is the concentration supporting half-maximal growth and S is the concentration of of substrate, or, in this case, necessary nutrilites.

$$r = r_{\max}\frac{S}{K_m + S} \tag{6}$$

It follows from this relationship that r is proportional to S at low S and is independent at high S, when $r = r_{\max}$. In the intermediate region, studies have shown that a better approximation is given by Eqs. (7). S_c is the critical concentration where growth becomes independent of nutrilite concentration, and $\alpha = r_{\max} S_c$.

$$\begin{aligned} r &= \alpha S \quad & S \le S_c \\ r &= r_{\max} \quad & S \ge S_c \end{aligned} \tag{7}$$

A more elaborate form is needed in other cases, and additional modifications are needed for oligotrophs, that is, organisms living and growing in very dilute environments. Oligotrophs have very effective uptake systems so that they can scavenge scarce nutrients; sometimes they have extra appendages (prosthecae) to give them extra surface. But even *E. coli*, an inhabitant of the mammalian gut, has the capability of taking up nearly every molecule of glucose that impinges on its cytoplasmic membrane. For organisms capable of coping with both abundance and scarcity, the uptake systems must be adequate when the concentration is low, but would be excessive when the substrate is not limiting. This is one of the reasons that Eqs. (7) are, in many cases, a better representation of cellular growth than Eq. (6).

Alternative carbon resources. When presented with two suitable carbon (and energy) sources, and under certain circumstances, a bacterium may elect to consume one compound and commence synthesizing the enzymes and other metabolic machinery needed to use the second substance, only when the first one has been depleted. For example, *E. coli* can be given both glucose and lactose (milk sugar). After the glucose is consumed, there is a lag in growth while β-galactosidase and galactoside permease are formed, allowing renewed growth. β-Galactosidase is an enzyme that splits lactose into glucose and galactose; galactoside permease is a membrane protein that functions in the entry of lactose into the cell. This growth pattern is called diauxic (double-fed) growth.

Catabolite repression, a concept in molecular biology, is now understood in molecular terms through the study of diauxic growth. **Figure 2** shows the DNA coding for the regulatory and structural part of *E. coli* lactose metabolism. In the presence of high levels of cyclic adenosine monophosphate (cAMP) resulting from glucose shortage, the catabolite activator protein (CAP protein) binds the “upstream” region of the regulatory site of the *lac* operon about 60 base pairs before the site of transcription initiation. This allows the promoter region to increase the binding and functioning of RNA polymerase. RNA polymerase synthesizes a polycistronic message that specifies β-galactosidase and galactoside permease. More directly favoring lactose metabolism is the effect of a lactose derivative binding to the repressor and unblocking the operator region.

This regulatory system is sufficiently complex that it functions appropriately and sometimes quite differently under a series of other circumstances. When two substrates are each present in low concentration and when one serves as a nitrogen source as well as a carbon and energy source, metabolism of both types can occur simultaneously. *See* ALLOSTERIC ENZYME; ENZYME INHIBITION; GENE; GENE ACTION; OPERON.

Cell division. The small size of bacteria provides a high surface-to-volume ratio, which assures effective uptake of substances from the environment. Since small size increases the number of propagules per unit of biomass, cell division is an important aspect of growth. Seemingly, it is not necessary for cell division to be very precise, because the important thing is to keep the living “packages” small. However, in some cases cell division is quite accurate; the bacteria divide nearly in the middle and, during balanced growth, nearly at the same size. Except in laboratory-created mutants, cells almost never form without receiving a complete chromosome. But even in the

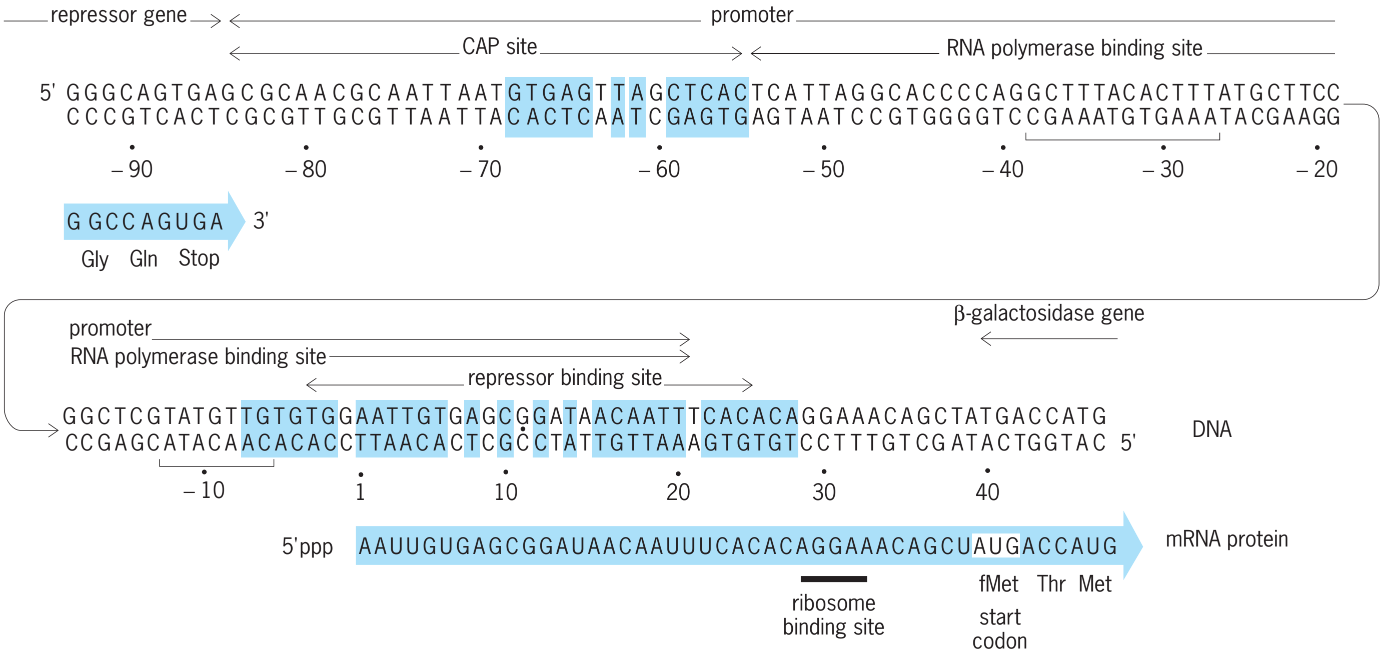

repressor gene
promoter
CAP site
RNA polymerase binding site
5' GGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCC
CCCGTCACTCGCGTTGCGTTAATTACACTCAATCGAGTGAGTAATCCGTGGGGTCCGAAATGTGAAATACGAAGG
– 90
– 80
– 70
– 60
– 50
– 40
– 30
– 20
GGCCAGUGA 3'
Gly Gln Stop
promoter
RNA polymerase binding site
repressor binding site
β-galactosidase gene
GGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATG
CCGAGCATACAACACACCTTAACACTCGCCTATTGTTAAAGTGTGTCCTTTGTCGATACTGGTAC 5'
DNA
– 10
1
10
20
30
40
5'ppp AAUUGUGAGCGGAUAACAAUUUCACACAGGAAACAGCUAUGACCAUG
mRNA protein
fMet Thr Met
ribosome binding site
start codon

Fig. 2. The *lac* operon. The genetic region of *Escherichia coli* K12 is shown indicating the region interacting with effector molecules and enzymes. The promoter region contains the part of the DNA binding to the CAP protein, the RNA polymerase, and a portion of the repressor-binding site. The part binding to repressor is called the operator. Various elements of symmetry are noted with the boxes. The underlined regions in the RNA polymerase–binding site are functionally important in initiating transcription. On the upper right, the tail end of the repressor gene is shown. On the lower right, the messenger RNA and the first three translated codons of β-galactosidase are shown.

most precise cases of cell division there is some variation in size and evenness at division, which translates into a larger variability of the age that cells attain before they divide. The doubling time is a suitable average of the individual division times of cell cycles.

Continuous culture. Bacteria in nature are usually engaged in activities similar to the culture cycle described above. A spore or stationary cell finds its way to a new habitat, grows, and divides until the clone has utilized the available resources as effectively as its metabolic capabilities permit. Later, as spores or stationary-phase cells become dispersed to new habitats, the cycle starts over. Another major growth mode occurs where there is flow of nutrient through the ecosystem and growth is more or less continuous. This second mode is idealized by systems of continuous culture. One device that is used for continuous culture is the chemostat. It operates by the introduction of sterile medium to a constant-volume culture vessel at a constant rate. When growth is adequate to deplete the medium (usually of one limiting nutrilite), the system becomes self-regulating. For a specified culture volume, V, and flow rate of medium, ω, the bacteria deplete the limiting nutrilite until its concentration, S, becomes such a value that r from Eqs. (1)–(6) equals ω/V. This is a stable situation, because if the biomass were to temporarily increase, S would decrease and then r would decrease until the biomass returned to its original steady-state value. *See* CHEMOSTAT.

In another form of continuous culture, the turbidostat, the biomass concentration is monitored photoelectrically. New medium is added and an equal volume of culture is removed, so that the turbidity stays constant and the steady-state rate of growth on that particular medium is maintained. *See* CULTURE.

Growth on surfaces. Most organisms, even if they grow well in dispersed cultures, are capable of binding to living cells and dead solids. This is advantageous in that they are not easily removed from a favorable environment. In addition, the surface may serve either to provide nutrients or to sequester nutrients. In continuous cultures, adherence provides a way to remain near the inflow of nutrient-rich medium. In fact, most systems of continuous culture are plagued by adherence of the strain under study, or of its mutants or contaminants.

Secondary growth phase. Another aspect of the culture cycle is that many organisms produce special chemicals at certain times and conditions. This has led microbiologists, particularly those working in the fermentation industries, to divide the growth process into the trophophase (feeding phase) and the idiophase (special or personal phase). It is during the latter phase that bacteria and fungi produce large amounts of antibiotics and other chemicals of economic importance. It appears that when the stationary phase approaches, different organisms redirect their metabolism toward the production of different exotic chemicals, which probably have some benefit to the organism. For industrial purposes, the idiophase is extended by strain selection and by careful feeding of appropriate resources (for example, glucose) at a rate that allows very slow growth to continue and the culture to persist for a long time at the border of the exponential and stationary phase.

In some cases, the production at idiophase depends on the concentration of specific metals and, in all cases, the production depends on the absence of exogenous inorganic phosphate. In some, it precedes the processes of differentiation into spores or conidia.

Organismal response to stress. The known organizational functions of the cell together with the stimulatory and regulatory processes use about 10% of the coding capacity of the genome of an organism, such as *E. coli*. Sequencing the entire genome shows that there are many regions that code for unknown proteins. Although the genes for a wide range of functions have been studied, these open reading frames have been encountered; some can be assigned probable functions, while some cannot.

There are two interpretations for the roles of open reading frames. First, they are genes that were needed in the past, then silenced, and in the future may eventually be removed by deletion. Second, they are genes that are functional but are part of systems that function only when stimulated. Such functions are hard to study because the challenges that elicited them are not known. There are, however, three stimulons which are being actively studied that may serve as prototypes for many of the unassigned genes. As distinct from an operon, such as the lactose operon shown in part in Fig. 2, a stimulon is a system of genes not physically linked together. These systems can be studied because they have long-term and specialized functions for the cells. Therefore, stimulons that protect the cell against high temperature, genetic damage, starvation, and oxidation can all be studied by looking for those genes being uniquely transcribed when the cells are challenged. *See* OPERON.

Heat stimulon. Approximately 46 genes are involved in protecting *E. coli* against temperature shocks. Some have been known for decades with regard to how they function and relate to the corresponding genes in flies, humans, and other organisms. Many are synthesized when new sigma subunits (σ^{32} or σ^{e}) for the RNA polymerase are formed in response to the temperature challenge. The sigma factor, σ^{e}, is formed at higher temperatures than σ^{32}. Besides the functions of some of these genes, it is known how some respond to other types of challenges. Thus, treatment with alcohol elicits the formation of a subset of the heat shock genes.

SOS stimulon. When the DNA of the cell is damaged, a set of processes that behave in different ways is initiated. Some are involved in repair; for example, they correct lesions. Photoreactivation, for example, is catalyzed by a gene that uses ultraviolet-A-light (UVA) and visible-light energy to undo the changes in adjoining pyrimidine stretches in the DNA that have been chemically cross-linked by ultraviolet B light (UVB) and ultraviolet C light (UVC). In addition, there are excision repair processes, in which DNA is removed from one strand and then reconstructed complementary to the other partner. Also, there is postreplication repair that can occur if there are two copies of the genome in the cell by recombining fragments between them until a complete, functional genome is generated.

Besides the repair function of this stimulon, the SOS signal inhibits cell division. The inhibition may give time for the repair to take place before the signaling is reversed and the cell attempts to replicate the DNA and to divide. *See* ULTRAVIOLET RADIATION (BIOLOGY).

Starvation stimulon. Some bacteria sporulate when all of an available resource has been consumed, but others such as *E. coli* elicit the formation of up to 50 different proteins. Cells prevented from forming these proteins die very rapidly when starved. These 50 proteins are probably only the tip of the iceberg. For the long-term survival of the species, the death of most of the cells in a colony is irrelevant. However, the creative change of one in a million or a trillion cells so that one cell can cope with a severe environment and grow is all that is needed. Quite likely, such processes involve cell mutations making extreme changes in their DNA that mostly lead to even more immediate death, but may (although very rarely) provide an answer to the cause of the starvation. *See* ANTIBIOTIC; BACTERIAL PHYSIOLOGY AND METABOLISM; INDUSTRIAL MICROBIOLOGY. Arthur L. Koch

Bibliography. M. DePedro, J. V. Holtje, and W. Loffelhardt, *Bacterial Growth and Lysis*, 1993; P. Gerhardt (ed.), *Methods for General and Molecular Bacteriology* 2d ed., 1994; A. L. Koch, *Bacterial Evolution, Growth and Form*, 1995; F. C. Neidhardt, J. L. Ingraham, and M. Schaechter, *Physiology of the Bacterial Cell*, 1990; D. White, *Microbial Physiology*, 1995.

Bacterial luminescence

The production of visible light by bacteria; with very few exceptions this light is blue-green. The phenomenon is seen in many species of several genera, including *Vibrio, Photobacterium, Alteromonas*, and *Xenorhabdus*.

Habitats. Luminous bacteria are primarily marine, but there are some genera with terrestrial (*Xenorhabdus*) and fresh-water (*Vibrio*) species. In the marine environment the bacteria are found in various habitats, including planktonic (free-floating), saprophytic (on a variety of marine proteinaceous materials), parasitic (on a number of marine invertebrates), and symbiotic. Marine shellfish left at room temperature will often begin to glow after a few hours due to the colonies of saprophytic luminous bacteria growing on them.

The symbiotic habitat can take one of several forms. The symbiotic bacteria may be loosely associated as gut symbionts in many different marine organisms; they may be specifically and more tightly associated in the light organs of marine fishes and squids; or they may be very tightly associated as intracellular symbionts in luminescent pyrosomes (light-omitting organelles). When associated as light-organ symbionts, the bacteria are used by the host fish or squid as a biological light bulb. Under these conditions the bacteria are maintained in specialized organs where they are cultured by the host organism, kept free from contaminants, and continuously emit light. The actual light emission is then controlled physically by the host's use of shutters, chromatophores, or other mechanisms. These symbiotic relationships are probably the most common habitats in which bacterial luminescence is observed in the ocean.

Biochemistry and physiology. The chemistry of bacterial luminescence is unique among luminous organisms. The enzyme that catalyzes light emission is luciferase; it combines with a riboflavinlike substance called flavin mononucleotide ($FMNH_2$). This complex then reacts with a long-chain aldehyde, and with molecular oxygen to form an excited state capable of emitting light. The molecule that actually emits the light is an altered form of the flavin. This unique biochemistry has been used as an indicator of the presence of luminous bacteria in cases where the symbiotic bacteria could not be obtained in pure culture or could not be grown free from their host. *See* RIBOFLAVIN.

Physiological studies have shown that the luminous activities are regulated by many different factors, including the carbon and energy sources used for growth, oxygen, salt (osmotic) concentration, and iron. These controls can differ markedly between bacterial species. Autoinduction, another control system, makes the bacteria luminous whenever they are in high concentration, and shuts down the luminescence when the cells are dilute. This mechanism is apparently found in all species of luminous bacteria.

Many mutants have been isolated after treatment with chemical mutagens, and these have been used to elucidate the physiology and biochemistry of light emission. Mutants have been found in the synthesis of the luminescent system, in the enzyme luciferase, and in several other key components. For three different species of luminous bacteria, the genes for bioluminescence have been cloned and shown to express functionally in *Escherichia coli*, a common intestinal bacteria of humans. Cloning the genes into *E. coli* has allowed many genetic and biochemical manipulations that were previously impossible, and has greatly enhanced the knowledge of bacterial luminescence. Cloning as few as seven genes into *E. coli* can make these organisms luminous and in many ways analogous to the marine luminous bacteria. This raises questions about why bacterial luminescence is not more widespread, and what the use of luminescence is to the bacteria. *See* BACTERIAL PHYSIOLOGY AND METABOLISM; BIOLUMINESCENCE; GENETIC ENGINEERING. Kenneth H. Nealson

Bibliography. G. Leisman, D. Cohn, and K. H. Nealson, Bacterial origin of luminescence in marine animals, *Science*, 208:1271–1273, 1980; K. H. Nealson and J. W. Hastings, Bacterial bioluminescence: Its control and ecological significance, *Microbiol. Rev.*, 43:469–518, 1979; M. P. Starr (ed.), *The Prokaryotes*, 1981.

Bacterial physiology and metabolism

The biochemical reactions that together enable bacteria to live, grow, and reproduce. Strictly speaking, metabolism describes the total chemical reactions that take place in a cell, while physiology describes the role of metabolic reactions in the life processes of a bacterium.

The study of bacteria has significance beyond the understanding of bacteria themselves. Since bacteria are abundant, easily grown, and relatively simple in cellular organization, they are ideal for the study of basic life processes and thus have been used extensively in biological research. Functional analyses of bacterial systems have provided a foundation for much of the current detailed knowledge about molecular biology and genetics.

Bacteria are prokaryotes, lacking the complicated cellular organization found in higher organisms; they have no nuclear envelope and no specialized organelles. Yet they engage in all the basic life processes—transport of materials into and out of the cell, catabolism and anabolism of complex organic molecules, and the maintenance of structural integrity. To accomplish this, bacteria must obtain nutrients and convert them into a form of energy that is useful to the cell.

This article discusses cell chemistry, enzymes, coenzymes, catabolism including oxidative metabolism and fermentative metabolism, and anabolism.

Cell Chemistry

Bacterial cells contain a number of distinctive chemical constitutents, many of which have been found in the cells' capsules, walls, or membrane structures. Despite the presence of the variety of unusual compounds in the major classes of chemical structures in bacteria, the quantitative chemical analyses are broadly similar to those for other cell types. Thus, on a weight basis, bacteria contain 70–90% water, and a mineral content of 1–10%. Much of the mineral matter is accounted for by potassium, sodium, calcium, magnesium, and phosphorus. Proteins and nucleic acids, the major organic constituents of bacteria, are present in more constant amounts than other organic compounds, such as the lipids and carbohydrates. Protein, determined by the percentage of nitrogen, accounts for 40–60% of the dry weight of the bacterial cell. The nucleic acid content is 10–20%. The amounts of carbohydrate and lipid vary widely and are dependent on the growth conditions. Carbohydrate content is 10–30% and that of lipids 1–50%.

Chemical composition. The overall chemical composition of the bacterial cell is very similar to that of all other types of cells of animal, plant, and microbial origin, which are capable of growth and replication. Bacteria thus possess the proteins ribonucleic acid (RNA) and deoxyribonucleic acid (DNA), which are the major classes of chemical constituents. These are required for the biochemical activities and the continuity and expression of the genetic characteristics of the cell. All cells possess a plasma membrane structure responsible for permeability properties, and one of the universal features of cellular membranes is their protein-lipid nature.

Bacterial cytoplasmic membranes generally have a high ratio of protein to lipid (3:1), a feature similar to that of eukaryotic mitochondrial membranes. Thus, bacteria contain lipids of the phospholipid class as essential chemical constituents of their cell membranes. Lecithin is one of the commonest phospholipids in higher organisms, and it is found in several bacterial groups, but much less frequently. Although sterols, such as cholesterol, are widely distributed in nature, occurring in membranous structures from animal, plant, and higher microbial cells, they have not been found in bacterial cells (except the mycoplasma group of pathogenic bacteria where sterols are sequestered into the membranes from an exogenous supply in the complex growth medium) or in blue-green cyanobacteria. Hopanoids of the triterpenoid family have been found in a wide variety of bacteria, and are structurally related to sterols and may be the structural equivalents to sterols of eukaryotic organisms. On the other hand, carotenoids, which are also lipid-soluble compounds, are found in many bacteria and in the closely related cyanobacteria where they are localized in the cell membranes and photosynthetic membranes (thylakoid equivalents in prokaryotes). Isoprenoid quinones, a class of terpenoid lipids, occur in the membranes of bacteria, cyanobacteria, and Archaebacteria where they function as respiratory quinones; and as chemical entities they have chemotaxonomic relevance. The other major class of chemical components of cells, the carbohydrates or polysaccharides, also occurs in a wide variety of bacterial groups. The polysaccharides may occur in the bacterial cell as starch- or glycogenlike storage granules, or as surface capsular or slime layers or polymers covalently linked to peptidoglycan in gram-positive bacteria, or to lipid structures of lipopolysaccharides in gram-negative bacteria.

Many of the unusual compounds of biological origin have been found in bacterial cell structures and products. Although bacterial proteins appear to be quite normal with respect to the variety and isomeric form of the amino acid building blocks, the cell walls, certain capsules, and antibiotics formed by bacteria possess peptides containing D isomers of amino acids, such as alanine, glutamic and aspartic acids, and phenylalanine. The amino acid diaminopimelic acid, first detected and isolated from the bacterium *Corynebacterium diphtheriae* in 1950, was subsequently found to be one of the characteristic cell-wall amino acids of many bacterial species. It occurs in the wall peptidoglycan but is not an amino acid component of the bacterial proteins. *See* AMINO ACIDS; ANTIBIOTIC; CAROTENOID; CELL MEMBRANES; STEROL.

Cell walls of bacteria contain one of the most distinctive bacterial products, the amino sugar muramic acid (3-*O*-carboxyethyl-glucosamine), in combination with glucosamine in the peptidoglycan polymer forming the rigid wall structure. The distribution of muramic acid in nature is confined to the bacteria, cyanobacteria, and rickettsias. Muramic acid has not been detected in higher microorganisms or in plant and animal cells. It has also not been found in Archaebacteria, but certain members of this group possess an analogous wall structure which contains talosaminuronic acid in its glycan; this amino sugar serves as the linkage molecule to the peptides. *See* ARCHAEA.

The amino sugars generally found in higher organisms are glucosamine, galactosamine, and the sialic acids. In addition to these amino sugars and muramic acid, bacteria also contain a variety of other amino sugars in their polysaccharide and lipopolysaccharide structures. Polysaccharides of the pneumococcus capsular substances and lipopolysaccharides of gram-negative bacteria have been shown to contain the following variety of amino sugars: mannosamine, fucosamine, pneumonsamine (an aminodideoxytalose), amino uronic acids, amino dideoxy sugars, and diamino sugars.

The bacterial lipopolysaccharides possess a unique lipid tail consisting of lipid A, a glucosamine moiety to which β-hydroxy fatty acids are linked, with the amino sugar linked to the core and polysaccharide chains. They may also contain a unique class of sugars, the dideoxy sugars—colitose, abequose, paratose, and tyvelose. The other very characteristic product of the lipopolysaccharides is the sugar acid ketodeoxyoctonic acid. Bacterial cell walls also possess a group of polymers called the teichoic acids, whose occurrence so far is confined to bacteria. The teichoic acids are polymers of ribitol or glycerol phosphate with additional compounds such as glucose, *N*-acetylglucosamine, and D-alanine linked to the polyol backbone. The D-alanine in the teichoic acid is linked through an ester bond and thus represents another example of the unusual features of the bacterial wall compounds. In addition to wall teichoic acids, gram-positive bacteria have either a glycerol lipoteichoic acid or a lipomannan anchored in their membranes. This chemically unique class of compounds is amphiphilic, possessing a glyceride lipophilic tail. *See* AMINO SUGAR.

Proteins and peptides. Most of the cellular proteins formed by bacteria are similar in amino acid composition to those of other organisms. The extracellular proteins secreted by gram-positive bacteria show one conspicuous difference from many other proteins in that they are devoid of, or exceptionally low in, the amino acid cysteine. Thus it appears that these proteins lack the S-S type of bridges linking the peptide chains together. In addition to the polypeptide antibiotics, the bacterial cell produces unusual peptide structures. The substance forming a thick, viscid layer, or capsule, surrounding the cells of *Bacillus anthracis* is a polymer of D-glutamic acid.

Although glycoproteins similar to those of animal cells have not been found in bacteria, some of the Archaebacteria possess cell envelopes assembled from glycoprotein subunits, and certain eubacteria have surface layers composed of glycoproteins. *See* PROTEIN.

Nucleic acids. As in other cells, the bacterial cell contains both RNA and DNA. The RNA is largely in the form of ribosomes (RNA-protein particles) of somewhat smaller dimensions than those found in higher organisms. The DNA of the bacterial cell gives the Feulgen reaction, which is the cytochemical basis for the demonstration of the bacterial nucleus. The DNA in bacteria is not enclosed within a nuclear membrane as it is in other types of cells. The base composition (G + C, or guanine + cytosine) values vary from one bacterial species to another and range from 20 to 74%. *See* NUCLEIC ACID.

Carbohydrates. A great variety of carbohydrates are synthesized by bacteria. They range from simple polymers of glucose, such as the cellulose produced by *Acetobacter xylinum*, to complex substances composed of a number of sugar units, such as glucose, galactose, rhamnose, and amino sugars. Unusual features of certain bacterial carbohydrates include the presence of the D isomer of arabinose in some, the occurrence of heptose sugars, and the detection of new amino sugars. Neuraminic acid, an important constituent of carbohydrate-protein complexes of animal cells and tissues, has been reported in polymeric form in *Escherichia coli* and is also found in the surface polysaccharide of group B streptococci and in meningococcal capsular polysaccharides. *See* CARBOHYDRATE.

Lipids. Bacteria produce a variety of lipids and lipid complexes, waxes, fats, glycolipids, and peptidoglycolipids. Although phospholipids constitute the major class of lipid, smaller amounts of lipid-soluble compounds including carotenoids, isoprenoid quinones, sterollike hopanoids, and hydrocarbons are found in bacteria. Despite the lower amounts of these compounds, they are functionally important for the bacteria.

Phospholipids found in bacteria include lecithin (phosphatidyl choline); phosphatidyl derivatives of ethanolamine, serine, and inositol; phosphatidic acid; phosphatidyl glycerol; and cardiolipin. An unusual lipid first detected in *Bacillus megaterium* is a polymer of β-hydroxybutyric acid. It is present in the form of granules that may be stained with the fat-soluble dye Sudan black, and has been found in many different kinds of bacteria.

The lipids of Archaebacteria are chemically quite unique and differ from the glycerol–fatty acid lipids characteristic of eubacteria and eukaryotic cells. The archaebacterial lipids are isopranyl glycerol ether lipids and are distinctive chemical markers for the identification of this group of bacteria.

The lipids of bacteria contain the common fatty acids found in lipids from other types of cells. However, bacteria lack the ability to form polyunsaturated fatty acids and thus differ from higher microorganisms, plants, and animals. Certain bacteria contain high proportions of branch-chain fatty acids in their lipids. Other unusual features of the bacterial lipids include the occurrence of lipoamino acids in a variety of species and the derivatives of glycerol diether in the extremely halophilic bacterium *Halobacterium cutirubrum*. *See* LIPID.

Chemical anatomy. Anatomy, the science of bodily and cellular structure, has been applied to the bacterial cell. The major structural elements of bacteria have been isolated and chemically characterized. This method of studying the chemical anatomy of the bacterial cell avoids the difficulties of applying cytochemical tests, such as the Feulgen reaction, to such small cells.

Flagella and pili. The filamentous, locomotive appendages of bacteria are composed of flagellin, a protein related chemically and physically to the hairlike proteins of other organisms. Common enteric bacteria and other groups of gram-negative organisms possess long, hairlike appendages (pili, or fimbriae) composed of proteins called pilins that function in the adhesion of bacterial cells. *See* CILIA AND FLAGELLA.

Capsules. The thick, mucous envelope, or capsules, may be composed of polypeptide or carbohydrate material. The viscous capsular carbohydrates may be composed of a variety of sugars, including hexoses, pentoses, methylpentoses, uronic acids, and amino sugars.

Cell walls. Bacteria possess a rigid wall structure that defines the shape of each cell. The principal structural component of the wall is a covalently bonded peptidoglycan, the glycan portion of which consists of alternating molecules of *N*-acetylmuramic acid and *N*-acetylglucosamine linked by $\beta 1 \rightarrow 4$ glycosidic bonds. The peptide component is linked by an amide bond to the muramic acid. Alanine, D-glutamic acid, lysine, or diaminopimelic acid form the basic part of the peptide, and in some bacterial walls these peptide strands may be cross-linked by either glycine, alanine, or threonine peptide bridges or sometimes all three. In the gram-positive bacteria, the teichoic acids and polysaccharides may be covalently linked to the rigid peptidoglycan.

The wall, or more strictly speaking, the cell envelope, of the gram-negative bacteria is a more complex structure with a rigid layer of peptidoglycan upon which the lipopolysaccharide-protein-lipid complexes are anchored together with a cysteinyl glycerol lipoprotein covalently linked to peptidoglycan. The peptidoglycan has the basic building blocks of the structure found in the walls of gram-positive organisms. Certain Archaebacteria have a rigid layer or sacculus which is chemically different from the peptidoglycan layer of the common bacteria (eubacteria) in that the former layer has talosaminuronic acid instead of muramic acid and only L-isomers of amino acids in the peptide moiety.

Protoplasmic membrane and mesosome. The plasma membrane and its intrusions within the cell (called the mesosome) are composed of 20–30% lipid and 50–70% protein. In some instances, small amounts of carbohydrate have been found in isolated membranes. Virtually all of the cell's lipid, phospholipid, carotenoid, and electron-transport components are localized in the membrane-mesosome structures, in addition to which the membranes of gram-positive organisms contain the membrane amphiphiles, lipoteichoic acids, or lipomannans.

The membranes of bacteria perform mitochondrial, transport, and biosynthetic functions and thus contain a variety of proteins separable by polyacrylamide gel electrophoresis. M. R. J. Salton

Enzymes

A list of bacterial enzymes (organic catalysts) would include many of the enzymes found in mammalian tissues, as well as many enzymes not found in higher forms of life. By combining with such enzymes, many antibiotics are able to exert a selective killing or inhibition of bacterial growth without causing toxic reactions in the mammalian host. The great capability of the bacterial cell to metabolize a wide variety of substances, as well as to control to some extent the environment in which the cell lives, is reflected in its ability to form inducible enzymes. The majority of bacterial enzymes require cofactors for activity. These cofactors may be inorganic cations of organic molecules called coenzymes. *See* COENZYME; ENZYME.

Classification. Bacterial enzymes may be classified in numerous ways, for example, on the basis of (1) whether they are inducible or constitutive; (2) whether they are degradative (catabolic; resulting in the release of energy) or synthetic (anabolic; using energy to catalyze the formation of macromolecules); or (3) whether they are exoenzymes (enzymes secreted from the cell to hydrolyze insoluble polymers—wood, starch, protein, and so on—into smaller, soluble compounds which can be taken into the cytoplasm of the bacterium).

In addition, bacterial enzymes are involved in the transport of substrates across the cell wall, in the oxidation of inorganic molecules to provide energy for the cell, and in the destruction of a large number of antibiotics.

Inducible and constitutive enzymes. Constitutive enzymes are defined as those enzymes formed by the bacterial cell under any or all conditions of growth, whereas inducible enzymes are formed by the bacterial cell only in response to an inducer. In nature the inducer is usually the substrate for the specific inducible enzymes involved. However, in the laboratory it is possible to use inducers which are not substrates, provided the molecular configurations are similar to those of true substrates. Galactokinase, an enzyme necessary for the phosphorylation of galactose, is an example of an inducible enzyme, since it is formed only if the bacteria are grown on a medium containing D-galactose or D-fucose as a carbon source. If a culture of bacteria grown on galactose is subsequently transferred to a galactose-free medium but otherwise sufficient for protein synthesis, the enzyme (galactokinase) will no longer be synthesized by the bacterial cell. Although bacteria vary considerably in their ability to utilize various substrates, most are versatile in their ability to form inducible enzymes. However, such enzymes are not ordinarily synthesized by the bacterium because a repressor substance produced by the cell reacts with the operator gene to prevent the transcription of the specific gene involved and, hence, the ultimate synthesis of the enzyme. When inducer is present (normally the substrate), it reacts directly with the repressor, preventing its reaction with the operator gene. The gene is then able to function, and transcribes its message to messenger RNA for synthesis of the enzyme or enzymes being induced. Since this process involves the removal of a repressor, it is also referred to as derepression.

An additional control was described when it was observed that, when cells were actively metabolizing a substrate such as glucose, it was not possible to induce them to synthesize inducible enzymes even in the presence of the specific inducer. This phenomenon, called catabolite repression, can be overcome by the addition of adenosine 3′,5′-cyclic monophosphate (cAMP); it is now known that before the operon can be transcribed a catabolite repressor protein (CR-protein) must first bind to cAMP and then react with the promotor gene for the enzyme. When a cell is rapidly metabolizing glucose, cAMP is decreased, and the CR-protein cannot react with the promoter to permit transcription. Thus, these controls accomplish two functions: (1) inducible enzymes are not synthesized unless the specific substrate is available; and (2) inducible enzymes are not synthesized when the cell is already rapidly metabolizing a better substrate.

Catabolic and anabolic enzymes. Many bacteria can oxidize various organic compounds in the presence of oxygen to CO_2 and H_2O, and other organisms carry out incomplete oxidations or fermentations resulting in the formation of stable oxidation or fermentation products. Examples of a few such end products of bacterial metabolism would include acetic acid, lactic acid, formic acid, butyric acid, ethanol, isopropanol, acetone, butanediol, butanol, hydrogen, and carbon dioxide. These degradative reactions yield energy which can be trapped as adenosine triphosphate (ATP) to supply the energy for macromolecular syntheses.

Anabolic enzymes are necessary to synthesize proteins, nucleic acids, lipids, cell walls, and so on. Many of these enzymes carry out reactions which are similar to those occurring in mammalian cells, but others are extremely specific for a bacterial cell, such as those enzymes involved in cell wall biosynthesis and capsule synthesis. Furthermore, many bacterial enzymes involved in the transcription of DNA (DNA-dependent RNA polymerase) and synthesis of proteins either are not found in mammalian cells or are sufficiently different from their mammalian counterparts that they can be inhibited by various antibiotics without harm to the mammalian cell.

Some of the Common reaction types catalyzed by bacterial enzymes are listed in **Tables 1** and **2**; the reactions given as examples illustrate either a general or a specific case.

Many enzymes are recognized as composed of a protein moiety, also called carrier or apoenzyme, and a small molecular species, the coenzyme, also known as the prosthetic or active group. Neither component acts as a catalysts by itself; only the combination, or holoenzyme, does. The coenzyme is responsible for

TABLE 1. Reactions involving carbohydrates

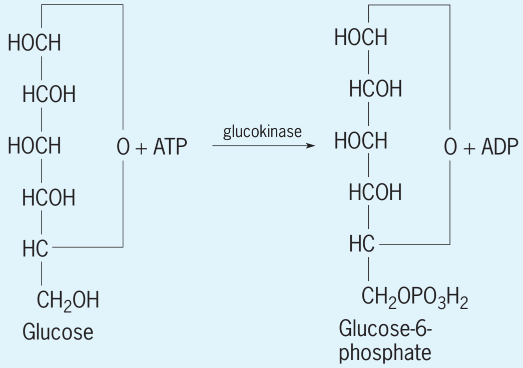
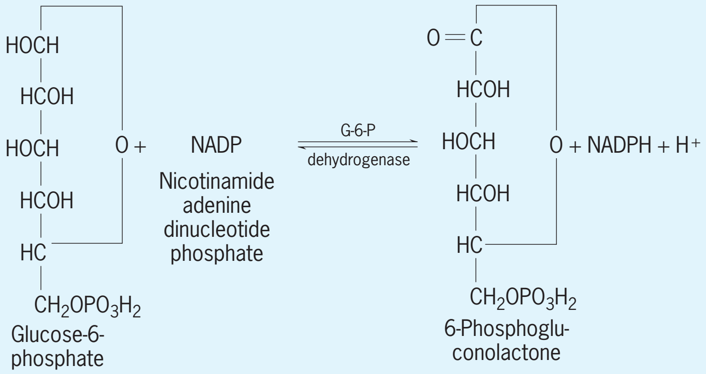
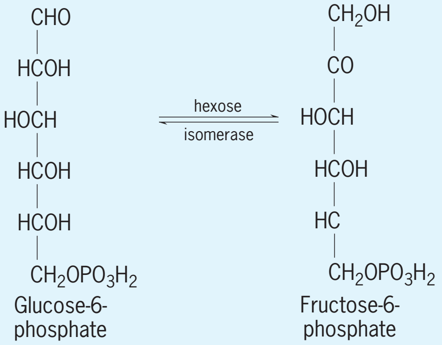
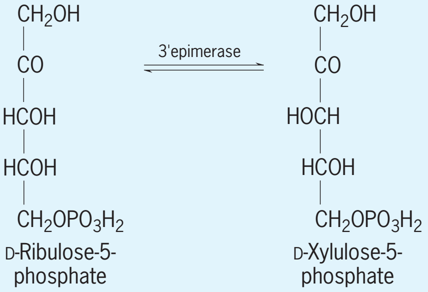
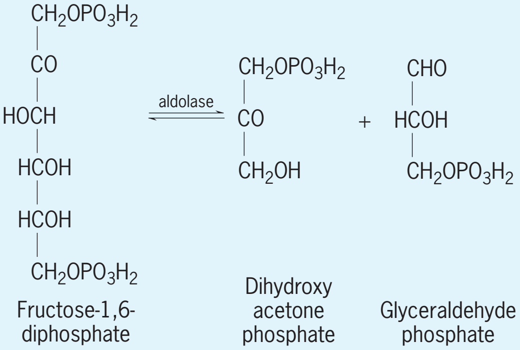
Enzyme type	Type of reaction catalyzed	Example
Kinases	Phosphorylation of substrate	 glucokinase O + ATP O + ADP Glucose Glucose-6-phosphate
Dehydrogenases	Oxidation or reduction of substrate	 O + NADP Nicotinamide adenine dinucleotide phosphate G-6-P dehydrogenase O + NADPH + H+ Glucose-6-phosphate 6-Phosphoglu-conolactone
Isomerases	Molecular rearrangement of substrate to form ketose from aldose or aldose from ketose	 hexose isomerase Glucose-6-phosphate Fructose-6-phosphate
Epimerases	Rearrangement of hydroxyl groups on substrtae	 3'epimerase D-Ribulose-5-phosphate D-Xylulose-5-phosphate
Aldolases	Cleavage of substrate into an alcohol and an aldehyde	 aldolase + Fructose-1,6-diphosphate Dihydroxy acetone phosphate Glyceraldehyde phosphate

TABLE 1. Reactions involving carbohydrates (*cont.*)

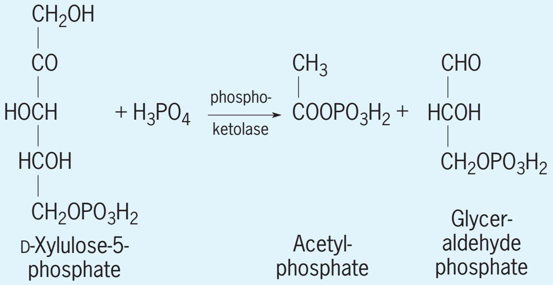
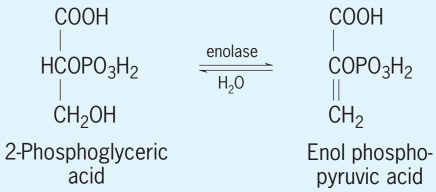
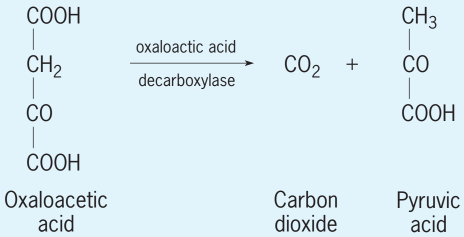
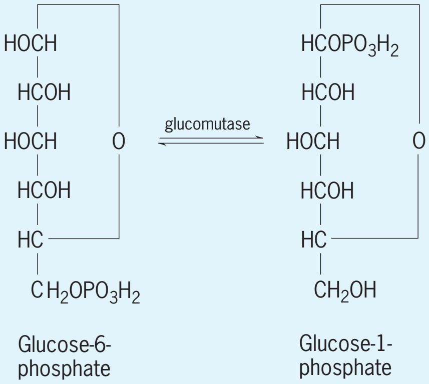
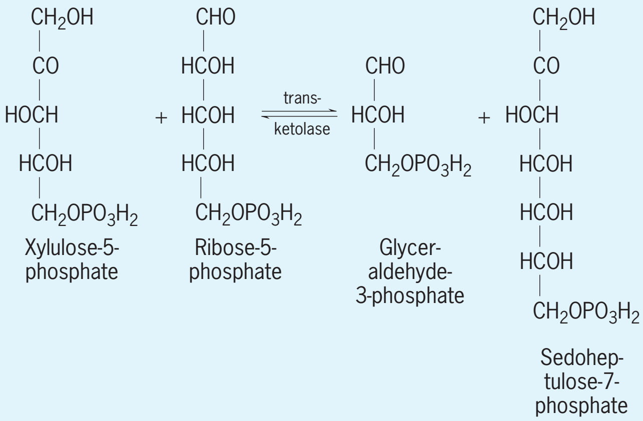
Enzyme type	Type of reaction catalyzed	Example
Ketolases	Cleavage of substrate at position of carbonyl carbon	 phospho- ketolase D-Xylulose-5-phosphate Acetyl-phosphate Glyceraldehyde phosphate
Enolase	Convert keto form of substrate to enol form by addition of H_2O	 enolase H_2O 2-Phosphoglyceric acid Enol phosphopyruvic acid
Decarboxylases	Remove carboxyl carbon from substrate	 oxaloactic acid decarboxylase Oxaloacetic acid Carbon dioxide Pyruvic acid
Mutases	Result in transposing phosphate from one carbon of substrate to a different carbon atom of same substrate molecule	 glucomutase Glucose-6-phosphate Glucose-1-phosphate
Transketolase	Similar to ketolase except 2-carbon fragment must be transported by the enzyme to an acceptor compound to form a new compound	 trans- ketolase Xylulose-5-phosphate Ribose-5-phosphate Glyceraldehyde-3-phosphate Sedoheptulose-7-phosphate

TABLE 1. Reactions involving carbohydrates (*cont.*)

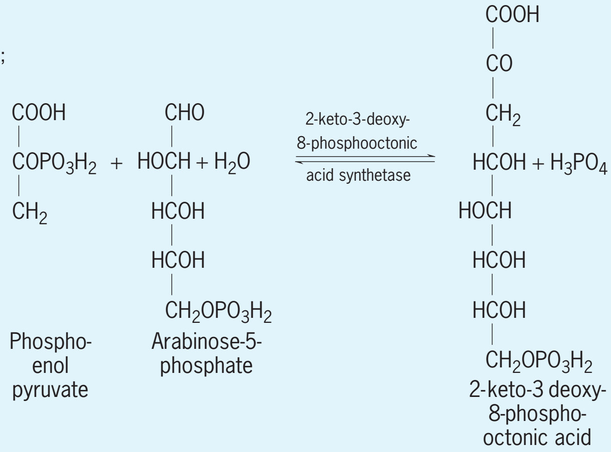
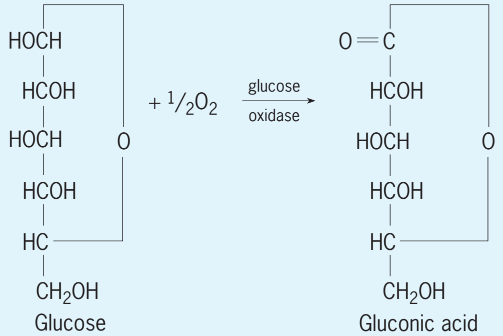
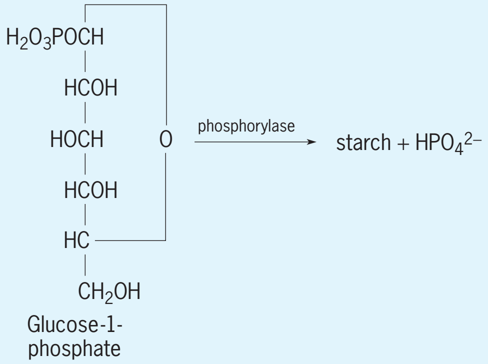
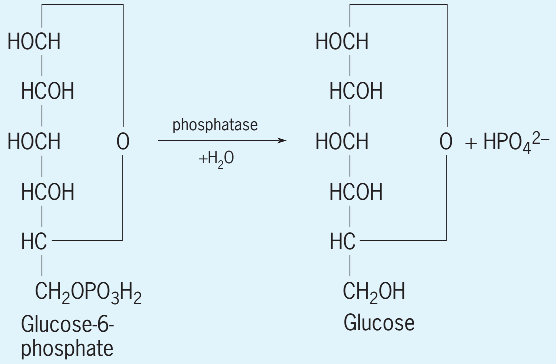
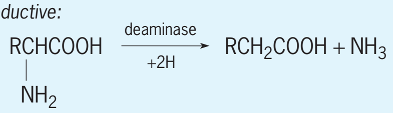
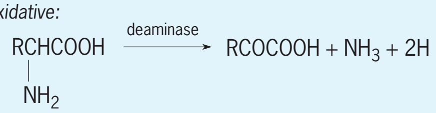
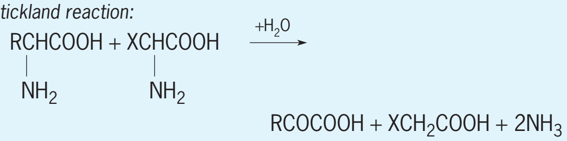
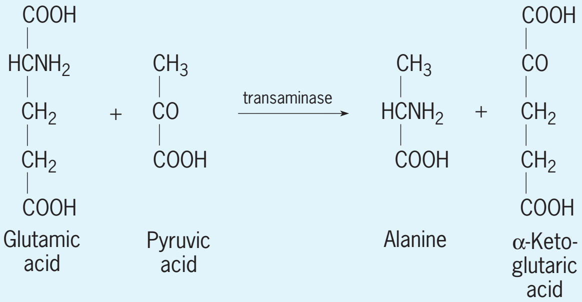
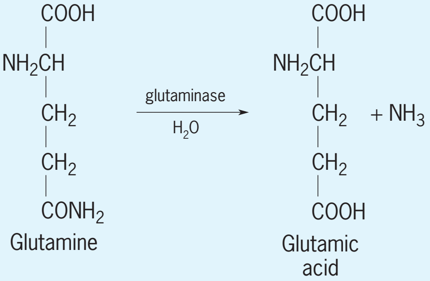
Enzyme type	Type of reaction catalyzed	Example
Synthetase	Condensation of phosphoenol pyruvate with CO_2 or other phosphorylated carbohydrates; each synthetase is specific for one condensation	 2-keto-3-deoxy-8-phosphooctonic acid synthetase Phospho-enol pyruvate Arabinose-5-phosphate 2-keto-3 deoxy-8-phospho-octonic acid
Oxidases	Yield same overall result as dehydrogenases; however, no coenzyme is involved required and molecular oxygen is	 glucose oxidase Glucose Gluconic acid
Phosphorylases	Condenses smaller phosphorylated sugars into polysaccharides	 phosphorylase starch + HPO4 2– Glucose-1-phosphate
Phosphatases	Remove phosphate from aldehyde or acid substrate	 phosphatase +H2O + HPO4 2– Glucose-6-phosphate Glucose

TABLE 2. Enzymes involving amino acid metabolism

Enzyme type	Type of reaction catalyzed	Example
Amino acid deaminases	Remove amino group from amino acid; different deaminases may carry out a reductive, or oxidative, deamination; members of genus *Clostridium* carry out an anaerobic dismutative deamination referred to as the Stickland reaction; it represents a combination of the two kinds	*Reductive:*  $\underset{\substack{\mid\\ NH_2}}{RCHCOOH} \xrightarrow[+2H]{\text{deaminase}} RCH_2COOH + NH_3$ *Oxidative:*  $\underset{\substack{\mid\\ NH_2}}{RCHCOOH} \xrightarrow{\text{deaminase}} RCOCOOH + NH_3 + 2H$ *Stickland reaction:*  $\underset{\substack{\mid\\ NH_2}}{RCHCOOH} + \underset{\substack{\mid\\ NH_2}}{XCHCOOH} \xrightarrow{+H_2O} RCOCOOH + XCH_2COOH + 2NH_3$
Amino acid transaminases	Removes amino group from one amino acid and carries it to an α-keto acid to form a new amino acid and a new α-keto acid	 $\begin{array}{c}COOH\\ \mid\\ HCNH_2\\ \mid\\ CH_2\\ \mid\\ CH_2\\ \mid\\ COOH\end{array} + \begin{array}{c}CH_3\\ \mid\\ CO\\ \mid\\ COOH\end{array} \xrightarrow{\text{transaminase}} \begin{array}{c}CH_3\\ \mid\\ HCNH_2\\ \mid\\ COOH\end{array} + \begin{array}{c}COOH\\ \mid\\ CO\\ \mid\\ CH_2\\ \mid\\ CH_2\\ \mid\\ COOH\end{array}$ Glutamic acid, Pyruvic acid, Alanine, α-Keto-glutaric acid
Amino acid decarboxylases	Remove carboxyl group leaving basic amine	$\underset{\text{Amino acid}}{RCHNH_2COOH} \longrightarrow \underset{\text{Amine}}{RCH_2NH_2 + CO_2}$
Amino acid oxidases	Same as oxidative deaminases (see amino acid deaminases above)	
Amidases	Remove amide group from some amino acids	 $\begin{array}{c}COOH\\ \mid\\ NH_2CH\\ \mid\\ CH_2\\ \mid\\ CH_2\\ \mid\\ CONH_2\end{array} \xrightarrow[H_2O]{\text{glutaminase}} \begin{array}{c}COOH\\ \mid\\ NH_2CH\\ \mid\\ CH_2\\ \mid\\ CH_2\\ \mid\\ COOH\end{array} + NH_3$ Glutamine, Glutamic acid

the particular type of catalytic activity displayed by the holoenzyme, such as the dehydrogenation of an alcohol to an aldehyde or of an aldehyde to a carboxyl group or the transfer of an amino group from one molecule to another. The apoenzyme, on the other hand, confers on the holoenzyme its pronounced specificity. Therefore, a single kind of coenzyme can act in combination with any one of a variety of apoenzymes, so that each combination of coenzyme and apoenzyme is functional in the conversion of one specific substrate.

Exoenzymes. In addition to the enzymes of bacteria which participate directly in the metabolic cycles of the cell, there are numerous enzymes which function as depolymerizing enzymes, making micromolecules available to the cell. The most common examples include the cellulases which degrade cellulose to cellobiose, amylases which degrade starch to dextrins and maltose, and the proteinases which hydrolyze proteins to polypeptides and amino acids. The ability of any given organism to use a particular substance of high molecular weight depends upon its ability to produce hydrolytic enzymes that are capable of attacking the substance in question. The exoenzymes are generally referred to as hydrolases because they catalyze the hydrolytic breakdown of their substrates (**Table 3**). *See* AMYLASE.

Inducible exoenzymes. Many exoenzymes are constitutive. Nevertheless, their production may be

TABLE 3. Extracellular bacterial enzymes

Organism	Enzyme	Substrate	End products
Clostridium	Cellulase	Cellulose	Cellobiose
Beneckea	Chitinase	Chitin	*N*-Acetyl glucosamine derivatives
Erwinia	Pectinase	Pectin	Galacturonic acid derivatives
Bacillus	Amylase	Starch Amylose Glycogen	Limit dextrins, maltose
Bacillus	Proteinase	Protein	Polypeptides
	Lipases (esterases)	Fats	Glycerol and fatty acids

enhanced by cultivation of the appropriate bacteria in the presence of the corresponding di-, oligo-, or poly-saccharides; hemicelluloses; and proteins. In some cases exoenzymes are formed only in response to the presence of the specific substrate.

Medical importance. Many pathogenic microorganisms excrete enzymes which may play an important role in pathogenesis in some cases. The α-toxin (lecithinase) of *Clostridium perfringens* illustrates a highly active enzyme which is responsible for the necrotizing action associated with gas gangrene infections due to this microorganism. *Streptococcus pyogenes* excretes hyaluronidase which degrades ground substance (polymer of hyaluronic acid), and streptokinase which activates plasmin resulting in a system that lyses fibrin. Other examples include coagulase of the *Staphylococcus* which activates clotting of plasma, urease of *Proteus vulgaris* which splits urea to ammonia and carbon dioxide, and collagenase of *Clostridium* which hydrolyzes collagen. A summary of the more common enzymes excreted by pathogenic microorganisms is presented in **Table 4**. *See* DIPHTHERIA; GANGRENE; HYALURONIDASE; STAPHYLOCOCCUS.

Enzymes destroying antibiotics. Many bacteria are able to synthesize enzymes which will hydrolyze or modify an antibiotic so that it is no longer effective. Essentially all of these enzymes are coded by DNA that exists in bacterial plasmids. As a result, the ability to produce enzymes which destroy antibiotics can be rapidly passed from one organism to another either by conjugation in gram-negative organisms or by transduction in both gram-negative and gram-positive bacteria. Bacterial plasmids can code for enzymes that (1) destroy penicillin by hydrolysis of the β-lactam bond; (2) destroy streptomycin by adding adenyl groups to the antibiotic; (3) inactivate neomycin and kanamycin by phosphorylating the antibiotics; (4) destroy chloramphenicol by adding acetyl groups; and (5) render tetracycline ineffective, probably by reducing the permeability of the bacterial cell to the antibiotic.

Autolytic enzymes. Autolytic enzymes seem to be a necessary component of the cell to hydrolyze covalent bonds in the rigid cell wall, allowing insertion of new structural units into the growing wall. Autolysis (dissolution of the cell) occurs when autolytic enzymes and cell wall synthetic enzymes do not work in unison. Thus, if too many bonds are opened in the rigid cell wall before new subunits are ready to occupy these spaces, internal osmotic pressures within the bacteria cause the cell to burst. In some organisms the autolytic system takes over following injury to the cell, and in others (such as *Streptococcus pneumoniae*) this system can be easily activated by treatment with surface-active agents such as bile salts.

Control of enzymes. In the absence of a control mechanism, a bacterial cell would synthesize much more of a specific metabolite than is necessary for sustained growth. Therefore the cell has developed two types of control for the regulation of bacterial enzyme activity: enzyme repression and feedback inhibition.

Enzyme repression. This is a rather coarse control that acts at the genetic level. This system acts in a manner that is somewhat the reverse of enzyme induction. Thus, if an organism that is synthesizing its own amino acid, histidine (which requires 20 different enzymes), is transferred to a medium containing adequate histidine, the organism will cease to synthesize the enzymes needed for histidine biosynthesis.

TABLE 4. Enzymes excreted by microorganisms of medical importance

Organism	Enzyme	Substrate	End products
Clostridium	Lecithinase	Lecithin	Diglyceride, phosphoryl choline
Clostridium	Collagenase	Collagen	?
Streptococcus	Hyaluronidase	Hyaluronic acid polymer	Hyaluronic acid
Streptococcus	Streptodornase	Deoxyribonucleic acid	Nucleotides
Streptococcus	Streptokinase	Activates plasminogen to plasmin	Results in lysis of fibrin clots
Staphylococcus	Coagulase	Coagulase reacting factor	Results in coagulation of plasma
Proteus	Urease	Urea	Ammonia and carbon dioxide
Corynebacterium diphtheriae	Diphtheria toxin	Nicotinamide adenine dinucleotide (NAD)	Splits NAD and adds ADP-ribose to elongation factor 2 to prevent protein synthesis by freezing ribosome movement

In other words, the presence of histidine has stopped the transcription of the entire portion of DNA which is responsible for histidine biosynthesis. A model for this effect supposes that the histidine reacts with an inactive repressor, resulting in the formation of an active repressor which can now react with the operator gene of the chromosome so as to prevent the transcription of the operon. Should the histidine level fall too low, the repressor is released, and the operon will again be transcribed. Many bacterial enzymes are subject to enzyme repression at the genetic level.

Feedback inhibition. This is a much finer control in which the end product of a series of reactions reacts directly with the enzyme of the first reaction and temporarily inactives it so that no more end product can be made. Note that in this case the end product reacts directly with an enzyme, and not with a repressor or operator gene on the chromosome. Thus, feedback inhibition does not prevent synthesis of the enzyme, but only inhibits its activity. As the end product is used up by the bacterium, the enzyme is freed to make more product. Enzymes that are subject to feedback inhibition are called allosteric enzymes since they appear to have two active sites, that is, one for their enzymatic activity, and one which reacts with a final end product, causing the inhibition of the enzymatic site. *See* ALLOSTERIC ENZYME; BACTERIAL GENETICS.

Coenzymes

Organic molecules that participate directly in a bacterial enzymatic reaction and may be chemically altered during the reaction are called coenzymes. Although many enzymes do not need specific cofactors, the majority of bacterial endoenzymes do. These cofactors may be simple inorganic cations such as magnesium, manganese, or calcium, which are commonly referred to as activators, or more complex organic molecules (coenzymes).

Coenzymes are the functional units, also called prosthetic or active groups, of an enzyme. Their catalytic activity depends on their association with a protein moiety, the apoenzyme or carrier, which is responsible for the high degree of specificity of the complex, or holoenzyme. Many of these combinations are readily dissociable: holoenzyme $\rightleftharpoons$ coenzyme + apoenzyme. Good examples of such enzymes are the pyridine nucleotide enzymes whose coenzymes are nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP); and the flavin nucleotide enzymes, whose coenzymes are flavin mononucleotide and flavin adenine dinucleotide (FMN and FAD). These coenzymes act as hydrogen acceptors and can, in combination with a wide variety of specific apoenzymes, catalyze the dehydrogenation of a large number of substrates.

Some holoenzymes, particularly the iron-porphyrin enzymes such as catalases, peroxidases, and cytochromes, can be dissociated only by drastic means; their coenzymes are far more firmly bound to the protein. This is in keeping with the much more restricted range of their activity, which appears to be limited to specific reactions that are common to the majority of bacteria and other living organisms. *See* ENZYME; NICOTINAMIDE ADENINE DINUCLEOTIDE (NAD); NICOTINAMIDE ADENINE DINUCLEOTIDE PHOSPHATE (NADP).

Coenzymes and growth factors. The universal occurrence in the metabolism of all living organisms of hydrogenations, dehydrogenations, transhydrogenations, aminations, deaminations, transaminations, and various other general reaction types implies that the coenzymes for these processes must be present in the cells. This has been important in discovering the often-encountered special nutrient requirements, especially for vitamins of the B group. These vitamins are substances that are structurally closely allied to coenzymes. Thus the ability of a particular bacterium to develop only if supplied with a certain B vitamin can be readily interpreted to mean that the organism cannot synthesize the corresponding coenzyme from other substances. For example, *Hemophilus influenzae* cannot synthesize and must consequently be supplied with the intact coenzyme before it can grow. Other organisms can synthesize the pyridine nucleotides if they are provided with simpler building blocks, such as nicotinamide or nicotinic acid. If a bacterium can grow in a medium devoid of such building blocks, the implication is that it can perform the synthesis of the coenzyme from still more remote, and often quite simple, ingredients. Where this has been experimentally tested, the results have been uniformly positive. *See* CULTURE.

Many organisms have become obligatory parasites because of a spontaneous loss in their ability to synthesize a needed vitamin. As a result, such organisms become completely host-dependent for a source of growth factors. This has been the basis for the development of ideas on physiological specialization and evolution. *See* VITAMIN.

Function. The mode of action of coenzymes generally involves their role as temporary acceptors for particular atoms or atom groups derived from a substrate, and subsequently as donors of the same entities to another kind of acceptor under the influence of specific apoenzymes. This is shown in **Fig. 1**, where H_2A represents an appropriate oxidizable substrate and A its oxidation product.

Similar diagrams can be constructed to illustrate the role of ATP as a phosphorylating agent with the production of adenosine diphosphate (ADP), and the subsequent reformation of ATP from ADP with phosphate groups produced in special positions during metabolism; of thiamine pyrophosphate as the coenzyme of certain decarboxylations; of pyridoxal

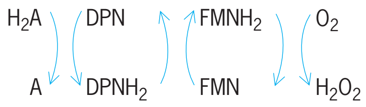


Fig. 1. Diagram of role of pyridine nucleotide and flavin nucleotide coenzymes in hydrogen transfer.

TABLE 5. Some organic coenzymes with their functions and related nutritional factors

Coenzyme	Function	Related nutritional factor or vitamin
Nicotinamide adenine dinucleotide (NAD)	Hydrogen acceptor in dehydrogenases	Nicotinamide
Nicotinamide adenine dinucleotide phosphate (NADP)	Hydrogen acceptor in dehydrogenases	Nicotinamide
Flavin mononucleotide (FMN)	Hydrogen acceptor in dehydrogenases	Riboflavin
Flavin adenine dinucleotide (FAD)	Hydrogen acceptor in dehydrogenases	Riboflavin
Coenzyme A (CoA)	Condensing enzymes	Pantothenic acid
Pyridoxal phosphate	Transaminases, amino acid decarboxylases	Pyridoxine, pyridoxal, pyridoxamine
Thiamine pyrophosphate	Cocarboxylase	Thiamine
Tetrahydrofolic acid	Transformylation	Folic acid
6,8-Dithio-*n*-octanoic acid (lipoic acid)	Oxidative decarboxylation	Required by some microorganisms
Iron porphyrin	In catalase, peroxidase, cytochromes	Protoporphyrin
Glucose-1,6-diphosphate	Phosphoglucotransferase	None
Cobalamin	Methylmalonyl CoA mutase	B_{12}

phosphate in amino group transfer; of tetrahydrofolic acid in the transfer of formyl groups; and of coenzyme A in the transfer of acyl groups. In the last-mentioned case acetyl groups, derived from the oxidation of alcohol or lactic, pyruvic, or fatty acids, may be transferred to oxaloacetate with the production of citrate, which in turn is subject to further oxidation by way of the tricarboxylic acid or Krebs cycle. *See* ADENOSINE DIPHOSPHATE (ADP); ADENOSINE TRIPHOSPHATE (ATP); BIOLOGICAL OXIDATION; CITRIC ACID CYCLE; COENZYME; TRANSAMINATION.

Reaction specificity. As already mentioned, a particular coenzyme may act in concert with any one of a large number of substrate-specific apoenzymes. The nature of the coenzyme then determines the general type of reaction that the holoenzymes can catalyze (**Table 5**). Wesley A. Volk

Bacterial Catabolism

Bacterial catabolism comprises the biochemical activities concerned with the net breakdown of complex substances to simpler substances by living cells. Substances with a high energy level are converted to substances of low energy content, and the organism utilizes a portion of the released energy for cellular processes.

Endogenous and exogenous catabolism. Endogenous catabolism relates to the slow breakdown of nonvital intracellular constituents to secure energy and replacement building blocks for the maintenance of the structural and functional integrity of the cell. This ordinarily occurs in the absence of an external supply of food. Exogenous catabolism refers to the degradation of externally available food. The principal reactions employed are dehydrogenation or oxygenation (either represents biological oxidation), hydrolysis, hydration, decarboxylation, and intermolecular transfer and substitution. The complete catabolism of organic substances results in the formation of carbon dioxide, water, and other inorganic compounds and is known as mineralization. Catabolic processes may degrade a substance only part way. The resulting intermediate compounds may be reutilized in biosynthetic processes, or they may accumulate intra- or extracellularly. The degree of breakdown depends on the genetic constitution of the particular bacteria, and also can be markedly influenced by environmental conditions. Catabolism also implies a conversion of the chemical energy into a relatively few energy-rich compounds or “bonds,” in which form it is biologically useful; also, part of the chemical energy is lost as heat.

Intermediary metabolism. Bacterial intermediary metabolism relates to the chemical steps involved in metabolism between the starting substrates and the final product. Normally these intermediates, or precursors of subsequent products, do not accumulate inside or outside the bacterial cell in significant amounts, being transformed serially as rapidly as they are formed. The identification of such compounds, the establishment of the coenzymes and enzymes catalyzing the individual reaction steps, the identity of active forms of the intermediates, and other details of the reaction mechanisms are the objectives of a study of bacterial intermediary metabolism.

In studying intermediary metabolism, the usual objective is to interrupt the normal reaction sequence artificially so that the formation of one or more intermediates from their precursors continues. Since the intermediates cannot then be further transformed, they accumulate. Intermediates may be detected qualitatively and quantitated by fairly specific procedures. In some cases, they are isolated in pure form and identified. A number of chemical, biological and physical methods are applied to this kind of study and usually are complementary. Frequently it is desirable to simulate the reaction and study it with enzymes extracted from the living cell. The respiratory quotient RQ (moles CO_2 produced/moles O_2 consumed) provides a rough index of the chemical nature (relative state of reduction) of a material being oxidized by a population of live bacteria. By comparison with the theoretical values for complete oxidation, valuable information relative to the extent of oxidation may be obtained. The instrument universally employed for measuring the gas exchange on a micro scale is the Warburg respirometer. Changes in pressure of CO_2 and O_2 are measured at constant gas volume and temperature in individual manometers attached to vessels containing substrate and suspensions of cells.

Uniform physiochemical environment is ensured by constant shaking in a constant-temperature water bath. Chemical or enzymatic reactions in which any gas is formed or consumed can also be studied kinetically and quantitatively.

The Thunburg technique is used to study oxidation of a substrate occurring by dehydrogenation reactions. A reversibly oxidizable indicator, methylene blue, substitutes for molecular oxygen as the ultimate hydrogen acceptor (oxidant), becoming reduced to the colorless leuco form. Rigorous exclusion of gaseous oxygen is essential.

Triphenyl tetrazolium compounds can be used instead of methylene blue. The oxidized form is colorless; the reduced form is a red formazan. Tetrazolium indicators have the advantage over methylene blue that they are reoxidized by air only very slowly; hence anaerobic conditions are not essential.

Metabolite utilization determines the probable involvement of a supposed intermediate in a metabolic pathway. If the hypothetical intermediate is metabolized, it must be examined further. However, negative results may only indicate the failure of an intermediate to penetrate the permeable membrane of the cell. In this event, cell-free enzyme preparations of the bacteria are helpful.

Metabolite replacement by analogs or antimetabolites, chemicals structurally related to the natural intermediates, is a method for interfering fairly specifically with a given reaction or pathway. Usually a degree of inhibition of a reaction can be achieved and the effect on the formation of an end product or of a function studied.

Inhibition of a specific reaction on a pathway by any of a number of means, including metabolite analogs or nonspecific poisons, frequently results in the accumulation of a precursor at the blocked reaction. The accumulation may be intracellular, extracellular, or both. *See* POISON.

Isolation and determination of the identity of the accumulated intermediates is then possible, whereas normally the intermediates exist in unisolable concentrations.

Paper, vapor phase, and column chromatographic techniques are invaluable physical procedures for the separation, purification, isolation, and sometimes identification of compounds from multicomponent mixtures such as extracts of bacterial cells or cultural filtrates. Tracer techniques employ metabolites labeled uniformly or in particular atoms with radioactive or stable isotopes. The metabolic path of the metabolite is traced by following the distribution and fate of the labeled elements, usually in various other metabolites, with suitable instruments. Concentrations of intermediates below the sensitivity of chemical isolation or detection procedures can readily be studied with tracers. *See* CHROMATOGRAPHY; RADIOISOTOPE (BIOLOGY).

Biochemical mutation is a tool for securing an inheritable block in a metabolic pathway. A gene controlling the synthesis or function of a specific enzyme is permanently altered mutationally, usually via irradiation or treatment with nitrogen mustard gas. This alteration implies that the corresponding enzyme does not function in the mutated cell. Since each of the steps in a metabolic chain reaction is mediated by its own specific enzyme, the result is that the chain is interrupted. If, for example, a biochemical process normally proceeds by a consecutive conversion of substance S_1 to product S_n via the steps

$$S_1 \xrightarrow[E_1]{} S_2 \xrightarrow[E_2]{} S_3 \xrightarrow[E_3]{} S_4 \xrightarrow[E_4]{} S_5 \xrightarrow[E_5]{} S_6 \xrightarrow[E_6]{} \cdots \xrightarrow[E_{(n-1)}]{} S_n$$

and the individual step reactions involve the enzymes E_1, E_2, E_3, E_4, E_5, ... , E_{n-1} respectively, then a mutated cell lacking a functioning E_4 can convert S_1 only as far as S_4. Hence this substance, now no longer subject to conversion to S_5, tends to accumulate, often in sufficient quantity to permit its isolation and hence identification as an intermediate product in a reaction chain. *See* MUTATION.

Adaptive or induced enzymes are synthesized by the bacterial cell only in the presence of a specific inducing substance. The product of the action of the first induced enzyme in turn functions as an inducer of a second enzyme, and so on, until the original substrate is converted to products acted on by the central constitutive enzymatic machinery of the cell. If the hypothetical intermediate is utilized nonadaptively by a bacterium preadapted to a given substrate, that is, if the cell is simultaneously adapted to attack it, this is taken as presumptive evidence that the postulated intermediate is on the pathway of utilization of the original substrate.

Coenzyme action can be deduced by determining the biochemical or metabolic alterations caused in bacteria by affecting the function of specific enzymes through interference with the participation of essential inorganic or organic coenzymes. This may be done by creating a deficiency, by limiting the amount of coenzyme furnished as exogenous nutrients, or through vitamin or coenzyme analogs. In cell-free extracts, the coenzyme deficiency may be obtained by means of dialysis.

Catabolism: Oxidative Metabolism

Bacterial oxidative metabolism is manifested by the decomposition of a tremendous number of different substances by one or another species of bacteria. Virtually any naturally occurring substance theoretically capable of yielding energy through oxidation is utilized for that purpose by some bacterial organism or other. Some bacteria are extremely versatile and can utilize any one of a great number of different organic substrates as sources of energy and carbon for growth. Other bacteria attack limited numbers of compounds, and some are extremely specific for a single compound or for very few compounds. Some are obligately linked to the oxidation of the reduced forms of a single inorganic element. In the great majority of cases, the substance, or substrate, undergoing oxidative attack is converted for the most part to bacterial cell, carbon dioxide, and complex mucoid substances. The accumulation in appreciable yields of incomplete oxidation or degradation products by aerobic bacteria is not common, although several instances are known. In terms of substrate utilized the energy requirements of bacteria are considerably in excess of the carbon requirements for cell synthesis, and so it is characteristic of bacterial oxidations that relatively large amounts of substrates are oxidized in relation to the amount of cellular material synthesized. The more completely a substrate is decomposed by bacteria, the less of it needs to be oxidized to yield a given amount of bacterial growth.

Aerobic oxidations. Aerobic oxidations are reactions in which the substrate is oxidized with concomitant reduction of oxygen to water. The oxidizable substrate may be represented by the generalized symbol H_2A. Most bacterial oxidations take place via dehydrogenation of the hydrogen donors. This occurs by transfer of two electrons *e* and two protons, resulting in the formation of the oxidized product A as shown in reaction (1). The dehydrogenation reaction will proceed only when the equilibrium is shifted to the right by removal of one or more products of the reaction. This is done by hydrogen acceptors, or oxidants, which take up the protons and electrons, thereby becoming reduced. The oxidant may be denoted by the symbol B, and its reduction by reaction (2).

$$\underset{\substack{\text{Hydrogen}\\ \text{donor}\\ \text{(substrate)}}}{H_2A} \xrightarrow[\text{(oxidation)}]{\text{dehydrogenation}} A + 2H^+ + 2e \tag{1}$$

$$B + 2H^+ + 2e \xrightarrow[\text{(reduction)}]{\text{hydrogenation}} H_2B \tag{2}$$

Coupling the two reactions, the net transformation for the bacterial oxidation of any substance may be represented as the sum of these reactions, that is, reaction (3). The transfer of protons (active hydrogen) and electrons is mediated by a series of carriers present in catalytic concentrations, and the equation merely represents the overall transformation by which all bacteria secure energy.

$$H_2A + B \longrightarrow A + H_2B + \text{energy} \tag{3}$$

Another less common mechanism of bacterial oxidation is oxygenation, the direct incorporation of molecular oxygen in certain substrates in the initial oxidative step. These reactions are catalyzed by oxidases and may be represented by the generalized reactions (4) and (5).

$$\underset{\text{Substrate}}{A} + O_2 \longrightarrow \underset{\substack{\text{Oxidized}\\ \text{substrate}}}{AO_2} \tag{4}$$

$$\underset{\text{Substrate}}{AH} + 2[H] + O_2 \longrightarrow \underset{\substack{\text{Oxidized}\\ \text{substrate}}}{AOH} + H_2O \tag{5}$$

Depending on the bacterial species, or even strain, and the environmental conditions, H_2A may be any one of a great variety of organic and a lesser number of inorganic substances.

Different organisms are capable of employing different hydrogen acceptors, or oxidants, for the oxidation of a substrate, and this provides a basis for classification of the bacteria as well as for distinguishing types of bacterial oxidations. Most bacteria utilize molecular oxygen as the ultimate hydrogen acceptor; these are, accordingly, aerobic bacteria, or aerobes. In the absence of oxygen, that is, under anaerobic conditions, they can use other hydrogen acceptors in lieu of O_2, such as methylene blue or tetrazolium compounds. Some of them can also use nitrate, nitrite, or nitrous oxide (N_2O) in that capacity. Some bacteria that cannot live in the presence of oxygen (anaerobic bacteria) use sulfate or carbon dioxide as hydrogen acceptors (oxidants). As in the case of oxygen, which yields H_2O, the reduced forms of these oxidants accumulate as the result of the bacterial transformations; each plays the role of B under appropriate conditions, reaction (6).

$$B + 2H^+ + 2e \xrightarrow[\text{(reduction)}]{\text{hydrogenation}} H_2B \tag{6}$$

See BIOLOGICAL OXIDATION; FERMENTATION.

Metabolism of inorganic substances. When the oxidation of the reduced forms of inorganic substances is utilized as the sole source of energy for growth at the expense of carbon dioxide as the sole source of carbon, the process is known as autotrophy. The use of organic substances for energy and carbon assimilation is known as heterotrophy.

1. Molecular hydrogen. Molecular hydrogen is used autotrophically by certain bacteria, all of which are capable of growing either autotrophically or heterotrophically; they are facultative autotrophs. The growth-supporting reaction (sometimes called the Knallgasreaktion, that is, the explosive gas reaction) which these bacteria carry out is shown in reaction (7).

$$2H_2 + O_2 \longrightarrow 2H_2O + \text{energy} \tag{7}$$

The best-known hydrogen bacteria belong to the genera *Hydrogenomonas* and *Mycobacterium*. Many heterotropic bacteria can also carry out the previous reaction, but they are incapable of growing at the expense of the energy derived from that reaction exclusively.

2. Iron compounds. Iron may occur in nature in the form of salts and complexes of organic compounds. Bacteria that decompose the organic moiety will therefore cause the formation of ferric carbonate or hydroxide. Such organisms have often been designated as "iron bacteria" because their activities lead to the precipitation of $Fe(OH)_3$.

Nevertheless, this term was first proposed for the members of a special physiological group of bacteria, characterized by their ability to oxidize ferrous to ferric ion and to live at the expense of the energy liberated in this oxidation; they can therefore grow as strict autotrophs. Thus far, this ability has been definitely established for only one bacterial species, *Ferrobacillus ferrooxidans*, and made probable for another, *Gallionella ferruginea*. Still others, such as the filamentous, sheathed bacteria known as *Leptothrix ochracea* and related species, that regularly occur along with *Gallionella* in localities where ferrous ion–containing seepage water runs into an oxygen-rich environment and $Fe(OH)_3$ precipitates may also belong to this physiological group. However, the *Leptothrix* species are now recognized as special environmental forms of *Sphaerotilus* species, which can also thrive in media not containing ferrous iron where they grow at the expense of the oxidation of organic substances.

The oxidation of ferrous to ferric iron may be represented by reaction (8).

$$4FeCO_3 + O_2 + 6H_2O \longrightarrow 4Fe(OH)_3 + 4CO_2 + \text{energy} \qquad (8)$$

3. Reduced sulfur compounds. Inorganic sulfur compounds at various levels of reduction, such as sulfite, tetrathionate, thiosulfate, elementary sulfur, and hydrogen sulfide, can be used as oxidizable substrates by many kinds of bacteria. Generally the oxidation yields sulfate as the end product, as in reactions (9) and (10); if H_2S is the substrate, it is oxidized initially to elementary sulfur, as in reaction (11). In the case of the relatively large representatives, such as species of *Beggiatoa, Thiothrix, Thiovulum*, and *Chromatium*, this sulfur accumulates in the bacterial cells in the form of characteristic highly refractile droplets which apparently can serve as a reserve source of substrate and disappear if the supply of H_2S gets too low.

$$2S + 3O_2 + 2H_2O \longrightarrow 2H_2SO_4 + \text{energy} \qquad (9)$$

$$\underset{\text{Sodium thiosulfate}}{Na_2S_2O_3} + 2O_2 + H_2O \longrightarrow Na_2SO_4 + H_2SO_4 + \text{energy} \qquad (10)$$

$$2H_2S + O_2 \longrightarrow 2S + 2H_2O + \text{energy} \qquad (11)$$

Many sulfur bacteria are obligately autotrophic. Physiologically the best known are the small, rod-shaped *Thiobacillus* species. *Thiobacillus thioparus* thrives in neutral media and *T. thiooxidans* in acid media; the latter can still grow in solutions containing as much as 5% sulfuric acid. *Thiobacillus novellus* is a facultatively autotrophic thiosulfate oxidizer.

The oxidation of sulfide is postulated to proceed via sulfur (hydrate), sulfoxylic acid, and sulfurous acid; that of thiosulfate via tetrathionate, trithionate, and sulfite.

4. Nitrogen compounds (nitrification). The best-known bacteria which oxidize reduced inorganic nitrogen compounds are the members of the obligately autotrophic rods comprising two genera, *Nitrosomonas* and *Nitrobacter*. *Nitrosomonas europaea* oxidizes ammonia to nitrous acid via hydroxylamine; nitroxyl (NOH) has been postulated as the subsequent intermediate oxidation stage, as shown by reactions (12*a*) and (12*b*). *Nitrobacter winogradskii* oxidizes nitrite to nitrate by way of reaction (13).

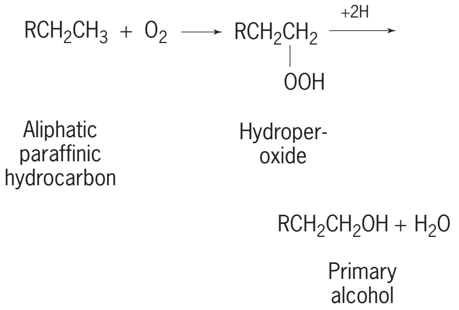
$$\underset{\text{Ammonia (hydrate)}}{NH_4OH} + {}^1/_2O_2 \longrightarrow \underset{\text{Hydroxylamine}}{NH_2OH} + H_2O + \text{calories} \qquad (12a)$$

$$\underset{\text{Hydroxylamine}}{NH_2OH} + O_2 \longrightarrow \underset{\text{Nitrous acid}}{HNO_2} + H_2O + \text{calories} \qquad (12b)$$

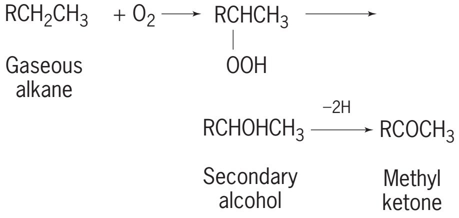
$$\underset{\text{Nitrous acid}}{HNO_2} + {}^1/_2O_2 \longrightarrow \underset{\text{Nitric acid}}{HNO_3} + \text{energy} \qquad (13)$$

Nitrogen-free aliphatic compounds. Although the number of oxidizable compounds and bacteria capable of effecting their oxidation are almost infinitely large, patterns of oxidation common to particular types of substrates appear to be the rule. In general, the oxidation pathways shape the substrate so that it can feed into the central terminal oxidation mechanisms almost universal in bacteria. In the main, the product entering the terminal oxidation cycle is active acetate, that is, acetyl coenzyme A (acetyl-CoA). Thus the special pathways for the decomposition of the diverse organic substrates in reality are mechanisms for their conversion to acetic acid. Most carbon assimilated by the bacterial cell is drained from the oxidation pathway at or following the acetate stage.

1. *Hydrocarbons.* A variety of bacteria, especially in the genera *Mycobacterium* and *Nocardia*, oxidatively attack paraffinic and olefinic hydrocarbons, including gases, liquids, and solids. The available information indicates that paraffinic hydrocarbons are oxidized monoterminally, probably via the addition of molecular oxygen to form the hydroperoxide, which in turn yields the corresponding primary alcohol by way of reaction (14). The gaseous alkanes can also be oxidized subterminally; in that case they yield methyl ketones. The proposed mechanism is shown in reaction (15). Palmitate waxes (esters) are



$$\underset{\text{Aliphatic paraffinic hydrocarbon}}{RCH_2CH_3} + O_2 \longrightarrow \underset{\text{Hydroperoxide}}{RCH_2\underset{|}{CH_2}\atop \quad OOH} \xrightarrow{+2H} \underset{\text{Primary alcohol}}{RCH_2CH_2OH} + H_2O \qquad (14)$$



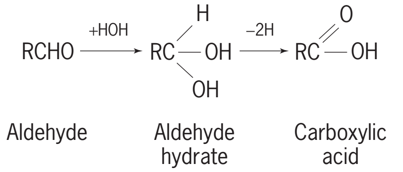
$$\underset{\text{Gaseous alkane}}{RCH_2CH_3} + O_2 \longrightarrow \underset{OOH}{\underset{|}{RCHCH_3}} \xrightarrow{+2H} \underset{\text{Secondary alcohol}}{RCHOHCH_3} \xrightarrow{-2H} \underset{\text{Methyl ketone}}{RCOCH_3} \qquad (15)$$

produced by certain bacteria during growth on long-chain liquid alkanes. The alcohol moiety of the ester corresponds to the substrate hydrocarbon.

2. *Alcohols.* Primary low-molecular-weight alcohols are oxidized by many bacteria, the first stage, reaction (16), being the corresponding aldehyde. The

$$\underset{\text{Primary alcohol}}{RCH_2OH} \xrightarrow{-2H} \underset{\text{Aldehyde}}{RCHO} \qquad (16)$$

enzyme alcohol dehydrogenase is responsible. The aldehyde is almost invariably rapidly oxidized to the corresponding carboxylic acid by hydration and dehydrogenation, as in reaction (17).



$$\underset{\text{Aldehyde}}{RCHO} \xrightarrow{+HOH} \underset{\text{Aldehyde hydrate}}{RC(H)(OH){-}OH} \xrightarrow{-2H} \underset{\text{Carboxylic acid}}{RC({=}O){-}OH} \qquad (17)$$

Although most bacteria that can carry out this type of oxidation rapidly convert the carboxylic acids to further oxidation products, some, notably members of the genera *Acetobacter* and *Acetomonas*, cause the accumulation of such acids, often in high yields. *See* VINEGAR.

Primary alcohols with more than five carbon atoms in their chain are toxic, and can be oxidized only if present in low concentration.

Secondary alcohols are typically oxidized to the corresponding ketones, as shown in reaction (18).

$$\underset{\text{Secondary alcohol}}{R{-}CHOH{-}R'} \xrightarrow{-2H} \underset{\text{Ketone}}{R{-}CO{-}R'} \qquad (18)$$

The ketones often accumulate. Tertiary alcohols are also decomposed, but nothing is known about the consecutive oxidation stages.

3. *Ketones.* Only the oxidation of the short-chain methyl ketones, particularly of acetone, has been studied in some detail. Acetone is photosynthetically carboxylated by ***Rhodopseudomonas gelatinosa*** to acetoacetic acid, which is then oxidized to acetic acid. This mechanism is represented by reaction (19).

$$\underset{\text{Acetone}}{CH_3COCH_3} + CO_2 \xrightarrow{\text{light}}$$
$$\underset{\text{Acetoacetic acid}}{CH_3COCH_2COOH} \longrightarrow \underset{\text{Acetic acid}}{2CH_3COOH} \qquad (19)$$

This mode of attack may also occur in the nonphotosynthetic bacterial oxidation of methyl ketones. However, a different type of mechanism, reaction (20), has been discovered with an acetone-oxidizing diphtheroid bacterium. This causes an initial oxidation to acetol, which in turn is broken down to acetaldehyde and an active one-carbon fragment, presumably formaldehyde or formate.

$$\underset{\text{Acetone}}{CH_3COCH_3} + {}^1\!/_2O_2 \longrightarrow \underset{\text{Acetol}}{CH_3COCH_2OH} \longrightarrow$$
$$\underset{\text{Acetaldehyde}}{CH_3CHO} + \underset{\text{Active carbon fragment}}{C_1} \qquad (20)$$

4. *Fatty acids.* Fatty acids are decomposed by oxidative attack at the beta carbon atom, resulting in the formation of acetic acid (active acetate; acetyl-CoA). The acetate represents the carboxyl and the alpha carbons of the carbon chain. A fatty acid with two carbons less is always formed, and the process is repeated serially. The first step is the activation by formation of the fatty acid thioester with coenzyme A (RCO—SCoA). As in **Fig. 2**, two mechanisms are known for the activation of a substrate fatty acid.

The first involves enzymatic displacement of acyl kinase of the pyrophosphate group (PP) of ATP by the fatty acid, forming acyl adenylate. Magnesium ions are required. An enzymatic reaction with CoA displaces adenylic acid, resulting in the thioester (acyl-CoA). The second mechanism of fatty acid activation takes place by direct enzymatic transfer of CoA by CoA transphorase from a donor acyl-CoA which itself may be the product of fatty acid oxidation. The acyl-CoA is dehydrogenated between the α- and β-carbons by acyl-CoA dehydrogenase. At least two forms of this enzyme are known. Both are metalloflavoproteins. The prosthetic group of both is FAD. The one which attacks short-chain fatty acids is a copper enzyme; the one attacking longer chains may be an iron enzyme. The addition of the elements of water to the double bond is accomplished by a hydrase, specifically crotonase. This enzyme has no known coenzyme, and it also provides a means of reversible conversion of β,γ-unsaturated acids to α,β-unsaturated acids via a common β-hydroxy acid intermediate shown in reaction (21).

$$RCH{=}CHCH_2CO{-}SCoA \xrightleftharpoons{HOH}$$
$$RCH_2CHOHCH_2CO{-}SCoA \xrightleftharpoons{HOH}$$
$$RCH_2CH{=}CHCO{-}SCoA \qquad (21)$$

The β-hydroxy acyl-CoA is oxidized by β-hydroxyl acyl-CoA dehydrogenase, the enzyme-containing NAD. All acids containing 4–12 carbon atoms (C_4 to C_{12}) are oxidized at approximately the same rate. Only the *d*-isomer is oxidized; the *l*-isomer is inert. The β-ketoacyl-CoA is then cleaved by a β-ketothiolase (acetyl-CoA transacetylase) between the α- and β-carbons with the concomitant uptake of 1 mole of CoA. The original CoA remains on the two-carbon cleavage fragment as acetyl-CoA. The thiolase reaction adds the second CoA molecule to the remaining carbon chain, at the carbonyl group, forming an acyl-CoA chain which is shorter by two carbons than the original fatty acid. *See* OPTICAL ACTIVITY.

This shorter acyl-CoA moiety is further oxidized by the same general mechanism. Hence the fatty acids that contain an even number of carbon atoms and have the general composition $C_{2n-1}H_{4n-1}COOH$ eventually yield a corresponding number, *n*, of

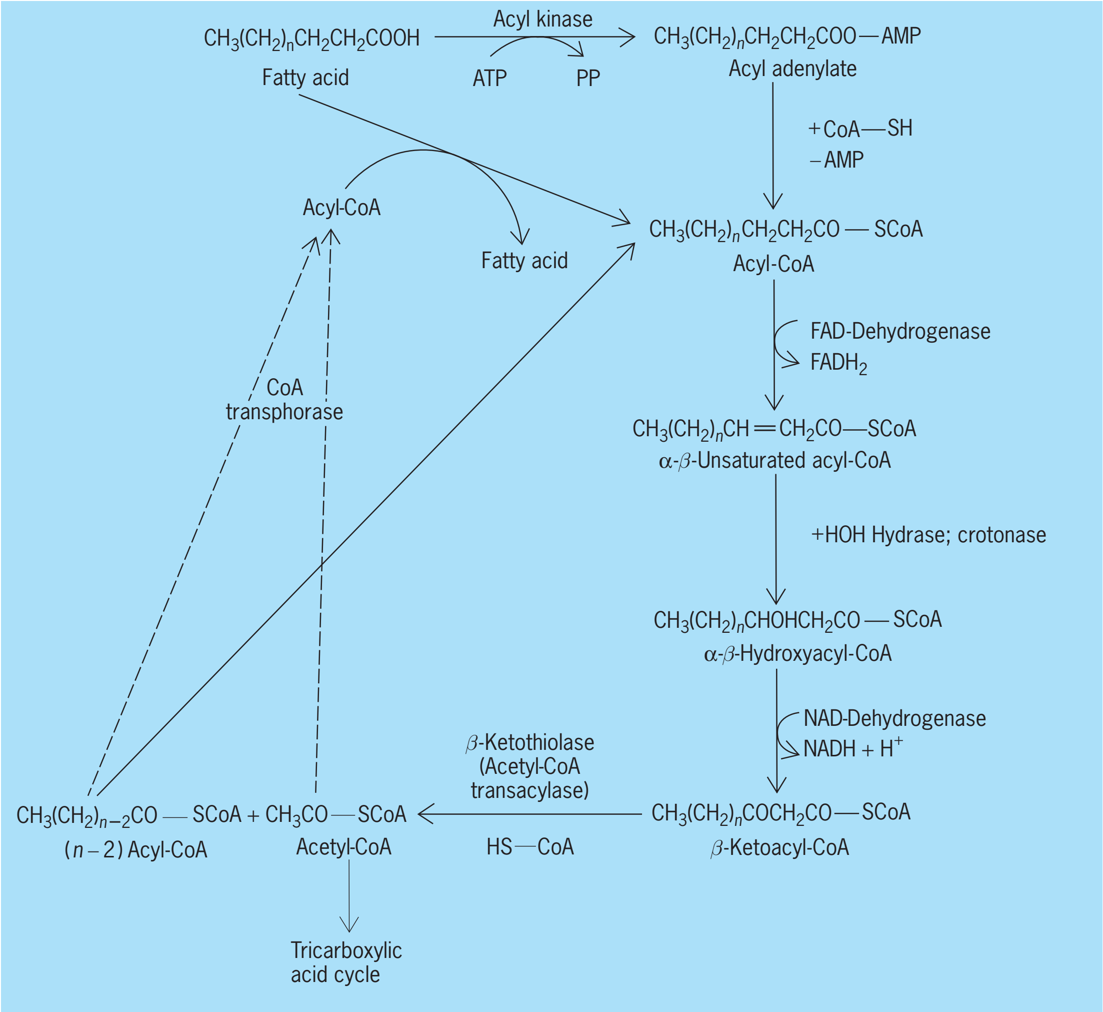
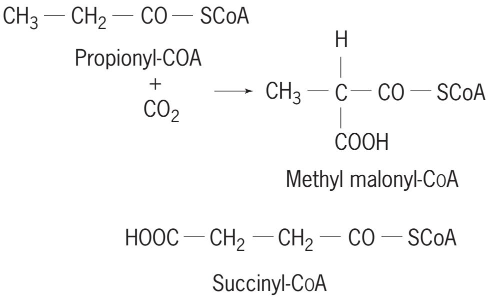

$CH_3(CH_2)_nCH_2CH_2COOH$
Fatty acid
Acyl kinase
ATP
PP
$CH_3(CH_2)_nCH_2CH_2COO{-}AMP$
Acyl adenylate
+CoA—SH
−AMP
Acyl-CoA
Fatty acid
$CH_3(CH_2)_nCH_2CH_2CO{-}SCoA$
Acyl-CoA
FAD-Dehydrogenase
$FADH_2$
CoA transphorase
$CH_3(CH_2)_nCH{=}CH_2CO{-}SCoA$
α-β-Unsaturated acyl-CoA
+HOH Hydrase; crotonase
$CH_3(CH_2)_nCHOHCH_2CO{-}SCoA$
α-β-Hydroxyacyl-CoA
NAD-Dehydrogenase
$NADH + H^+$
β-Ketothiolase (Acetyl-CoA transacylase)
$CH_3(CH_2)_{n-2}CO{-}SCoA + CH_3CO{-}SCoA$
(n − 2) Acyl-CoA
Acetyl-CoA
HS—CoA
$CH_3(CH_2)_nCOCH_2CO{-}SCoA$
β-Ketoacyl-CoA
Tricarboxylic acid cycle

Fig. 2. The two known mechanisms for the oxidative decomposition of fatty acids.

acetyl-CoA molecules. The fatty acids with an odd number of carbon atoms, $C_{2n}H_{4n+1}$ COOH, yield one molecule of propionyl-CoA in addition to $n - 1$ molecules of acetyl-CoA.

Acetyl-CoA is further oxidized via the citric acid or glyoxylic acid cycle, discussed later. It is probable that propionyl-CoA is degraded by an initial carboxylation to methyl malonyl–CoA, which is next converted to succinyl-CoA, as in reaction (22). The succinate can then be oxidized to acetate via fumarate, malate, and oxaloacetate.



$$\underset{\text{Propionyl-COA}}{\mathrm{CH_3{-}CH_2{-}CO{-}SCoA}} + \mathrm{CO_2} \longrightarrow \underset{\text{Methyl malonyl-CoA}}{\mathrm{CH_3{-}\overset{H}{\underset{COOH}{C}}{-}CO{-}SCoA}} \longrightarrow \underset{\text{Succinyl-CoA}}{\mathrm{HOOC{-}CH_2{-}CH_2{-}CO{-}SCoA}} \tag{22}$$

Certain bacteria produce acetyl phosphate, apparently as an energy reservoir. Acetyl phosphate is readily converted to acetyl-CoA by the enzyme phosphotransacetylase in reaction (23).

$$\underset{\text{Acetyl phosphate}}{\mathrm{CH_3CO{-}OPO_3^{2-}}} + \mathrm{HSCoA} \xrightarrow{\text{phosphotrans-acetylase}} \underset{\text{Acetyl-CoA}}{\mathrm{CH_3CO{-}SCoA}} + \mathrm{HOPO_3^{2-}} \tag{23}$$

As in biological oxidations in general, the pattern of bacterial oxidations is to chip large and more complex substrates into small carbon chain fragments. As a rule, these fragments are then broken down in stepwise fashion, by fairly complex cyclic systems, to CO_2, water, and energy in a process known as terminal oxidation.

The Krebs tricarboxylic acid (TCA) system is the most common cyclic mechanism of terminal oxidation in biological systems. It was conceived and clarified by H. A. Krebs, a British biochemist who received a Nobel prize for this contribution. The mechanism appears to be identical in bacterial and

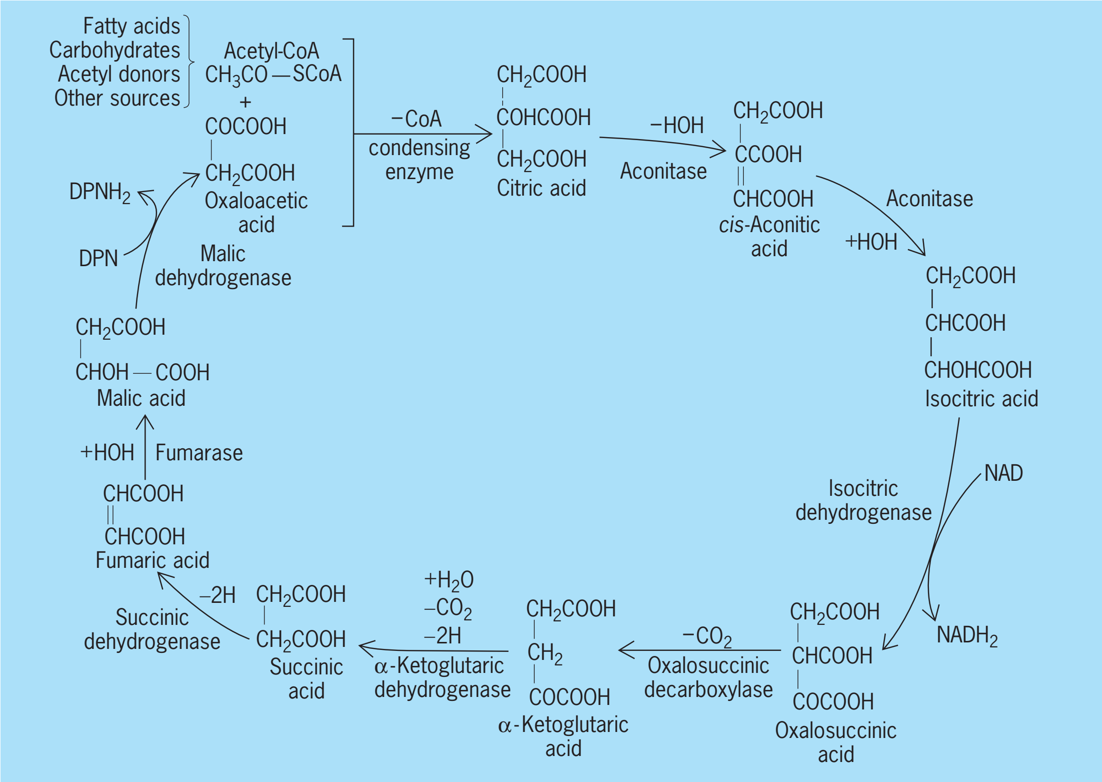

Fatty acids
Carbohydrates
Acetyl donors
Other sources
Acetyl-CoA
$CH_3CO—SCoA$
+
COCOOH
CH_2COOH
Oxaloacetic acid
−CoA
condensing enzyme
CH_2COOH
COHCOOH
CH_2COOH
Citric acid
−HOH
Aconitase
CH_2COOH
CCOOH
CHCOOH
cis-Aconitic acid
Aconitase
+HOH
CH_2COOH
CHCOOH
CHOHCOOH
Isocitric acid
Isocitric dehydrogenase
NAD
$NADH_2$
CH_2COOH
CHCOOH
COCOOH
Oxalosuccinic acid
$-CO_2$
Oxalosuccinic decarboxylase
CH_2COOH
CH_2
COCOOH
α-Ketoglutaric acid
$+H_2O$
$-CO_2$
−2H
α-Ketoglutaric dehydrogenase
CH_2COOH
CH_2COOH
Succinic acid
−2H
Succinic dehydrogenase
CHCOOH
CHCOOH
Fumaric acid
+HOH
Fumarase
CH_2COOH
CHOH—COOH
Malic acid
Malic dehydrogenase
DPN
$DPNH_2$

Fig. 3. Pathways of the Krebs tricarboxylic acid cycle, a common biological oxidative mechanism.

other microorganisms and in plant and animal cells. Minor modifications or adaptations of the Krebs cycle are known. The TCA system is a cyclic pathway for the oxidation of acetyl-CoA derived from fatty acid or carbohydrate breakdown, from acetyl donors, or from other types of metabolism (**Fig. 3**) to CO_2 and water, producing high-energy bonds. In the presence of assimilable nitrogen and mineral salts, various components are drained off from the cycle to serve as precursors of amino acid and other cellular components. Under resting cell conditions the substrate may be oxidized completely via the citric acid cycle, or variable amounts may be drained off and converted to products of oxidative assimilation. The oxidative formation of acetyl-CoA from pyruvic acid will be discussed later under triose pathway in connection with carbohydrate oxidation. Oxaloacetic acid is the key component of the cycle; it functions as the initial acceptor of acetate into the cycle. A transfer of acetate is effected by a specific enzyme, oxaloacetate transacetylase, better known as citrate-condensing enzyme, which links the methyl carbon of the acetyl group with the carbonyl carbon of oxaloacetate, splits out CoA, and forms citric acid.

In the next step, the elements of water are removed from citrate by the enzyme aconitase, leaving the asymmetrical unsaturated *cis*-aconitic acid. The elements of water are then added back to the double-bond carbons, but in the reverse positions, forming isocitric acid. There is some evidence for two different aconitases effecting the dehydration and the hydration. The simultaneous dehydrogenation of isocitrate at the carbinol group and decarboxylation of the tertiary carboxylic group are believed to be catalyzed simultaneously by an NAD-dependent enzyme, isocitric dehydrogenase, and by oxalosuccinic decarboxylase. The product, α-ketoglutaric acid, is oxidatively decarboxylated to succinic acid. The coenzymes diphosphothiamine, lipoic acid, and CoA are sequentially involved in this overall transformation, which comprises three intermediate enzymatic steps. This system is similar to that which oxidatively decarboxylates pyruvic acid and which is discussed below. Succinyl-CoA is the immediate precursor of the succinic acid; the acid is liberated by succinyl-CoA deacylase. Succinic acid is then oxidized to fumaric acid by succinic dehydrogenase, a flavin-dependent enzyme. The fumarate is next hydrated by fumarase, forming L-malic acid, which in turn regenerates oxaloacetate after its oxidation by means of the NAD-dependent malic dehydrogenase. The hydrogens, or electrons, are converted stepwise to water via the terminal electron transport system. High-energy phosphate bonds are thereby generated at various stages; their formation constitutes oxidative phosphorylation.

In the living cell, where the keto acid components are siphoned off for the synthesis of amino acids, the cycle would quickly run down because there would no longer be a continuous supply of oxaloacetic acid.

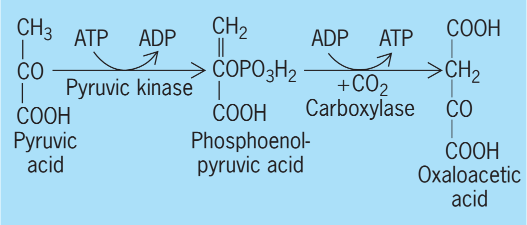


Fig. 4. Wood-Werkman reaction.

In addition, this substance itself can be readily decarboxylated to carbon dioxide and pyruvic acid, which constitutes an additional drain. The level of oxaloacetate can, however, be maintained by a carboxylation of pyruvic acid derived from other sources, mainly carbohydrate. One such reaction is the so-called Wood-Werkman reaction (**Fig. 4**), in which pyruvic acid is converted to phosphoenolpyruvic acid, the actual acceptor of CO_2. The fixation is coupled with the regeneration of an energy-rich pyrophosphate bond; either ADP or inosine diphosphate functions as the phosphate acceptor. The second carboxylation reaction is that accomplished by the malic enzyme; it represents a simultaneous reduction and carboxylation of pyruvic acid to malic acid by a NAD-dependent enzyme, and is shown in reaction (24).

$$\underset{\text{Pyruvic acid}}{\mathrm{CH_3COCOOH + CO_2}} \xrightarrow[\text{"malic enzyme"}]{\text{NADPH}_2\text{ or NADH}_2 \;\curvearrowright\; \text{NADP or NAD}} \underset{\text{L-Malic acid}}{\mathrm{HOOCCH_2CHOHCOOH}} \qquad (24)$$

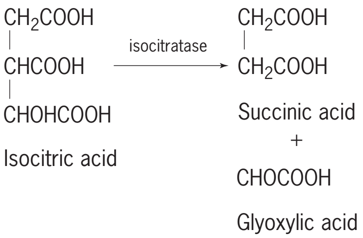
It has been recognized, however, that these carboxylation reactions are inadequate to generate sufficient oxaloacetate, particularly in those cases where acetate is the main or only oxidizable substrate. This has long presented a problem, which has been solved by the discovery of the glyoxylic acid cycle, which permits the formation of oxaloacetate, and hence also of citrate, from two-carbon fragments exclusively. The cycle comprises a condensation of acetyl-CoA with glyoxylic acid to malic acid under the influence of the enzyme malate synthetase, as in reaction (25). The malate is then oxidized to oxaloacetate.

$$\underset{\text{Glyoxylic acid}}{\mathrm{HOOCCHO}} + \underset{\text{Acetyl-CoA}}{\mathrm{CH_3CO — SCoA}} \xrightarrow[\text{malate synthetase}]{-\text{CoA}} \underset{\text{Malic acid}}{\mathrm{HOOCCHOHCH_2COOH}} \qquad (25)$$

A continuous supply of glyoxylic acid is ensured by a conversion of isocitrate other than its oxidation to oxalosuccinate. It is a cleavage, under the influence of the enzyme isocitratase, to glyoxylate and succinate according to reaction (26). The cycle is schematically represented in **Fig. 5**. The net change in one turn of the glyoxylic acid cycle is represented by reaction (27). The succinate is metabolized via

$$2\ \text{Acetate} + {}^1\!/_2 O_2 \longrightarrow \text{succinate}$$

the Krebs cycle or other pathways.



$$\underset{\text{Isocitric acid}}{\begin{array}{l}\mathrm{CH_2COOH}\\ |\\ \mathrm{CHCOOH}\\ |\\ \mathrm{CHOHCOOH}\end{array}} \xrightarrow{\text{isocitratase}} \begin{array}{l}\mathrm{CH_2COOH}\\ |\\ \mathrm{CH_2COOH}\\ \text{Succinic acid}\\ +\\ \mathrm{CHOCOOH}\\ \text{Glyoxylic acid}\end{array} \qquad (26)$$

5. *Hydroxy acids.* The β-hydroxy monocarboxylic acids are discussed under fatty acids. The α-hydroxy acids are oxidized to the corresponding α-keto acid by zinc-containing dehydrogenases, as in the conversion of L-lactic acid to pyruvic acid under the influence of lactic dehydrogenase in reaction (28).

$$\underset{\text{L-Lactic acid}}{\mathrm{CH_3CHOHCOOH}} \xrightarrow[\text{lactic dehydrogenase}]{-2\text{H}} \underset{\text{Pyruvic acid}}{\mathrm{CH_3COCOOH}} \qquad (28)$$

Lactate dehydrogenase acts on several α-hydroxy acids. The enzymes are stereospecific for the L forms. Three lactic dehydrogenases are known which differ in their coenzyme requirements. One utilizes NAD, a second uses flavin nucleotide, and the third cytochrome *c*; the last one is oxygen linked.

6. *Keto acids.* The most important keto acids are the α- and β-ketomonocarboxylic acids. The α-keto acids may be successively decarboxylated to the next smaller aldehyde by a diphosphothiamine

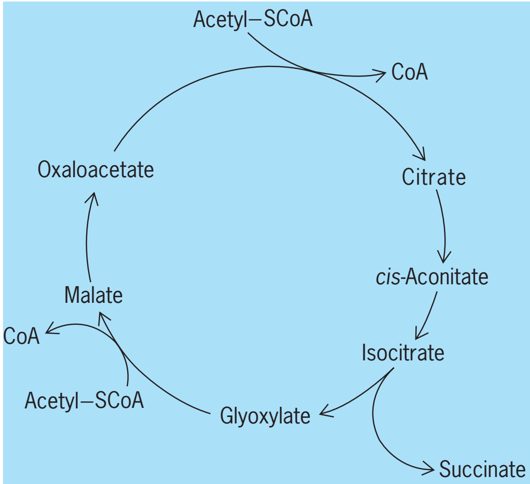
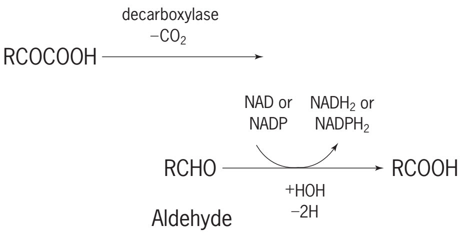


Fig. 5. Glyoxylic acid cycle.

decarboxylase, and then oxidized, reaction (29). The oxidative decarboxylation of pyruvic acid involves a series of coenzymes as shown in **Fig. 6**.



$$RCOCOOH \xrightarrow[-CO_2]{\text{decarboxylase}}$$

$$\underset{\text{Aldehyde}}{RCHO} \xrightarrow[+HOH \quad -2H]{NAD \text{ or } NADP \rightarrow NADH_2 \text{ or } NADPH_2} RCOOH \qquad (29)$$

For *β*-ketomonocarboxylic acids oxidation, see the discussion of fatty acids; for *β*-ketoadipic acid oxidation, see the section on aromatic compounds.

7. *Carbohydrates.* The oxidation of carbohydrates occurs by different routes, all of which may occur in any one bacterial species. However, one or more may be lacking in some bacteria.

The triose pathway is responsible for the production of pyruvic acid by the Embden-Meyerhof mechanism of glycolysis, described later in this article. The pyruvic acid is oxidatively decarboxylated to acetyl-CoA as described above, which in turn is oxidized by way of the citric acid cycle.

In aerobic bacteria, the rupture of the carbon chain of glucose occurs after oxidation to 6-phosphogluconate. This compound may be oxidatively decarboxylated to ribulose-5-phosphate, which is further metabolized via the pentose phosphate pathway described later. Alternatively, the 6-phosphogluconate may be dehydrated at carbon atoms two and three, forming 2-keto, 3-deoxy, 6-phosphogluconic acid. This pathway is also described later (Entner-Doudoroff).

The net effect of six turns of the hexosemonophosphate cycle is the complete oxidation of one molecule of glucose, as represented by reactions (30*a*) and (30*b*).

$$6\ \text{Glucose} + 6O_2 \longrightarrow 5\ \text{glucose} + 6H_2O + 6CO_2 \qquad (30a)$$

$$\text{Net: Glucose} + 6O_2 \longrightarrow 6H_2O + 6CO_2 \qquad (30b)$$

The oxidative steps are those between glucose and ribulose-5-phosphate (RP) [**Fig. 7**]. Glucose is regenerated from RP by interconversion of five-, six-, and seven-carbon sugar phosphate esters. The terminal two carbons of one RP are transferred by transketolase (a diphosphothiamine enzyme) to a second RP, forming a seven-carbon sugar (D-sedulose-7-phosphate) and leaving glyceraldehyde-3-phosphate (G-3-P). By transaldolase, the terminal three carbons of the sedoheptulose are condensed with the G-3-P to produce fructose-6-phosphate. The tetrose remaining from the seven-carbon sugar is condensed with a terminal two-carbon fragment from another RP, forming fructose-6-phosphate and leaving G-3-P. The latter is converted to an equilibrium mixture with dihydroxyacetonephosphate by triose phosphate isomerase. These trioses are then condensed to hexose diphosphate by the action of aldolase.

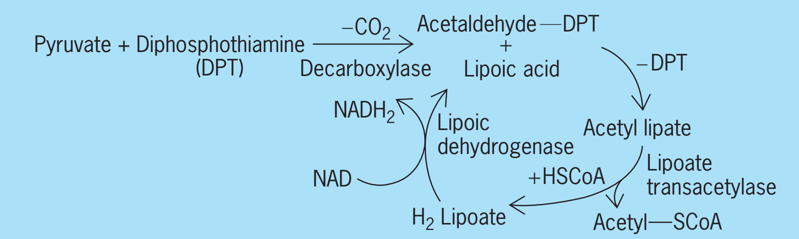

Pyruvate + Diphosphothiamine
(DPT)
−CO₂
Decarboxylase
Acetaldehyde—DPT
+
Lipoic acid
−DPT
NADH₂
Lipoic
dehydrogenase
NAD
Acetyl lipate
+HSCoA
Lipoate
transacetylase
H₂ Lipoate
Acetyl—SCoA

Fig. 6. Oxidative decarboxylation of pyruvic acid showing coenzymes involved.

Oxidation of aromatic compounds. This property is found mainly in aerobic bacteria belonging to the genera *Pseudomonas, Vibrio*, and *Mycobacterium*. As a rule, the aromatic compounds are converted to a benzene ring compound containing two adjacent carboxyls. The catechol and the protocatechuic acid pathways are the best known, both eventuating in *β*-ketoadipic acid (*β*-KAA). Phenol is oxidized directly to catechol. Mandelic acid, tryptophan, and indoleacetic acid are converted to catechol via the

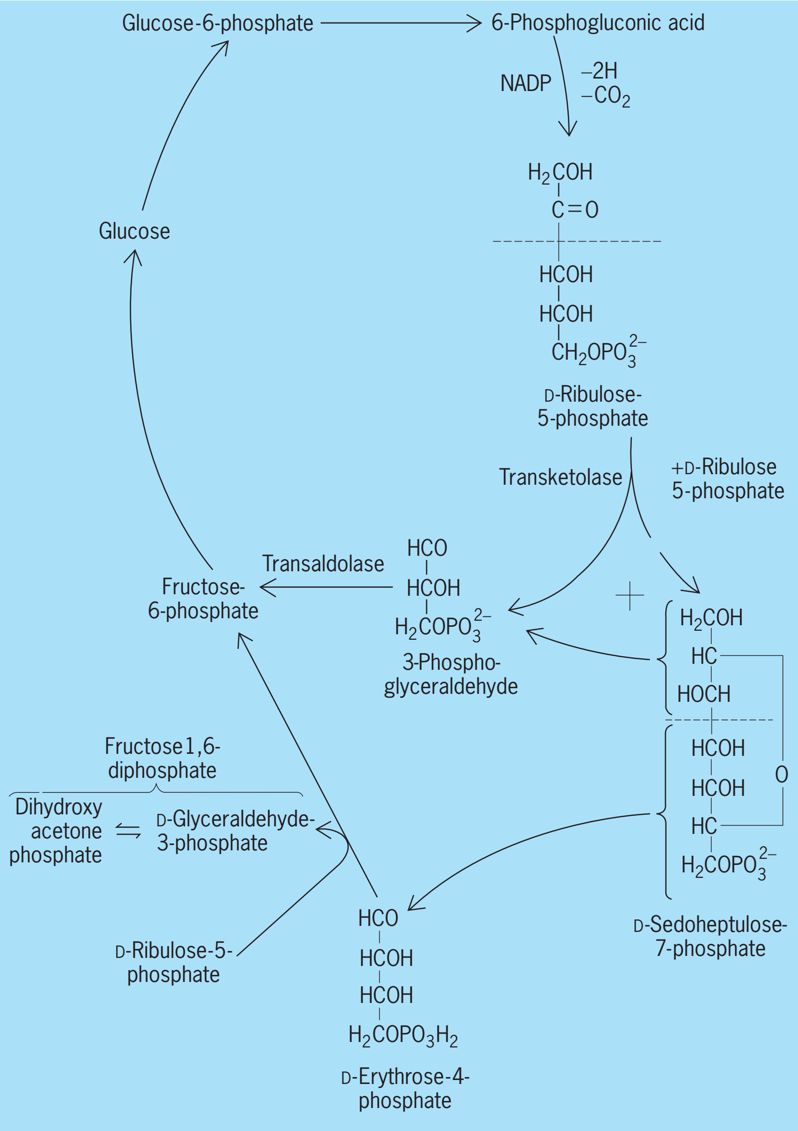

Glucose-6-phosphate
6-Phosphogluconic acid
NADP
−2H
−CO₂
H₂COH
C=O
HCOH
HCOH
CH₂OPO₃²⁻
D-Ribulose-
5-phosphate
Glucose
Transketolase
+D-Ribulose
5-phosphate
Transaldolase
HCO
HCOH
H₂COPO₃²⁻
3-Phospho-
glyceraldehyde
Fructose-
6-phosphate
+
H₂COH
HC
HOCH
HCOH
HCOH
O
HC
H₂COPO₃²⁻
D-Sedoheptulose-
7-phosphate
Fructose 1,6-
diphosphate
Dihydroxy
acetone
phosphate
D-Glyceraldehyde-
3-phosphate
D-Ribulose-5-
phosphate
HCO
HCOH
HCOH
H₂COPO₃H₂
D-Erythrose-4-
phosphate

Fig. 7. Carbohydrate oxidation in aerobic bacteria.

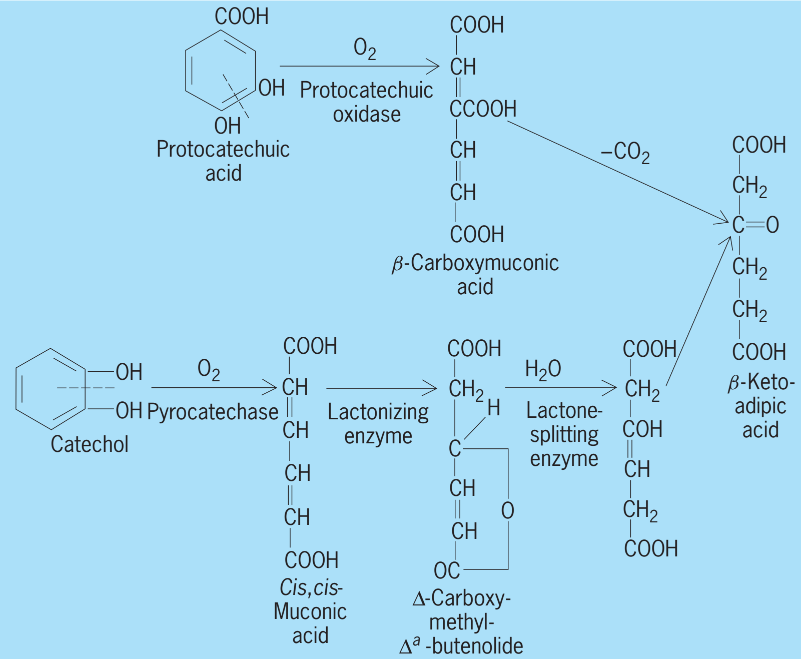

COOH
OH
OH
Protocatechuic acid
O_2
Protocatechuic oxidase
COOH
CH
CCOOH
CH
CH
COOH
β-Carboxymuconic acid
$-CO_2$
COOH
CH_2
C=O
CH_2
CH_2
COOH
β-Keto-adipic acid
OH
OH
Catechol
O_2
Pyrocatechase
COOH
CH
CH
CH
CH
COOH
Cis,cis-Muconic acid
Lactonizing enzyme
COOH
CH_2
H
C
CH
CH
O
OC
Δ-Carboxy-methyl-$Δ^a$-butenolide
H_2O
Lactone-splitting enzyme
COOH
CH_2
COH
CH
CH_2
COOH

Fig. 8. **Cleavage of dihydroxy ring structures in protocatechuic acid and catechol.**

sequences shown in reactions (31), (32), and (33).

$$\text{D}(-)\text{–Mandelic acid} \xrightleftharpoons[\text{racemase}]{} \text{L}(+)\text{–mandelic acid} \xrightarrow[\text{L–mandelic dehydrogenase}]{-2H} \text{benzoylformic acid} \xrightarrow[\text{benzoylformic carboxylase}]{-CO_2} \text{benzaldehyde} \xrightarrow[\text{benzaldhyde dehydrogenase}]{+H_2O_3 -2H} \text{benzoic acid} \longrightarrow \text{catechol} \quad (31)$$

$$\text{Tryptophan} \xrightarrow[\text{tryptophan peroxidase}]{O_2,\ H_2O} \text{formylkynurenine} \xrightarrow[\text{formylase}]{H_2O} \text{kynurenine} \xrightarrow[\text{kynureninase}]{-\text{alanine}} \text{anthranilic acid} \longrightarrow \text{catechol} \quad (32)$$

$$\text{Indoleacetic acid} \longrightarrow \text{skatole} \longrightarrow \text{indoxyl} \longrightarrow \text{salicylic acid} \longrightarrow \text{catechol} \quad (33)$$

The protocatechuic acid pathway is represented by reaction sequence (34).

$$\begin{array}{l} p\text{-Nitrobenzoic acid} \longrightarrow p\text{-aminobenzoic acid} \\ \qquad\qquad\qquad\qquad\qquad\qquad \downarrow \\ p\text{-cresol} \longrightarrow p\text{-hydroxybenzyl alcohol} \longrightarrow \\ p\text{-hydroxybenzoic acid} \longrightarrow \text{protocatechuic acid} \quad (34) \end{array}$$

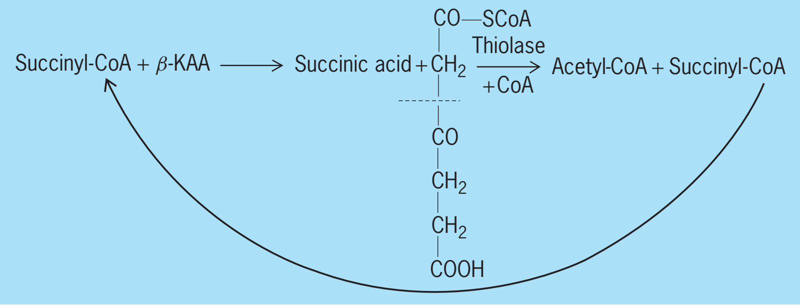

Succinyl-CoA + β-KAA
Succinic acid + CH_2
CO—SCoA
Thiolase
+CoA
Acetyl-CoA + Succinyl-CoA
CO
CH_2
CH_2
COOH

Fig. 9. **Splitting of β-ketoadipic acid to form acetyl-CoA and succinyl-CoA.**

The dihydroxy ring structures are then cleaved between the adjacent hydroxyls to form aliphatic dicarboxylic acids. The two pathways converge at β-ketoadipic acid, as shown in **Fig. 8**. Beta-ketoadipic acid is split to acetyl-CoA and succinyl-CoA. This reaction sequence is shown in **Fig. 9**. The succinyl-CoA activates β-ketoadipic acid, and the process is a cyclic one whereby the net products are acetyl-CoA and succinic acid, which are then oxidized via the citric acid cycle.

The oxidative degradation of the amino acids phenylalanine and tyrosine proceeds by a different pathway, as shown in reaction (35).

$$\text{Phenylalanine} \xrightarrow[\text{phenylalanine hydroxylase}]{} \text{tyrosine} \xrightarrow[\text{L-amino acid oxidase}]{+H_2O,\ -2H,\ -NH_3} p\text{-hydroxyphenylpyruvic acid} \xrightarrow[\text{oxidase}]{+O_1,\ -CO_2} \text{homogentisic acid} \xrightarrow[\text{oxidase}]{O_2} \text{4-maleylacetoacetic acid} \xrightarrow[\text{isomerase}]{} \text{4-fumarylacetoactic acid} \xrightarrow[\beta\text{-diketonase}]{H_2O} \text{fumaric acid} + \text{acetoacetic acid} \quad (35)$$

Polycyclic aromatic hydrocarbons are oxidized by certain bacteria by routes which have not been clarified enzymatically. As a rule, *o*-hydroxycarboxylic acids containing one ring less than the starting compound have been isolated and are considered to be intermediate breakdown products.

Nitrogen-containing organic substances. The complete breakdown of this class ordinarily results in release of all of the nitrogen as ammonia.

1. *Amino acids.* Enzymes are specific for L- and D-amino acids. The key reaction is the splitting out of ammonia by deamination. Four types of deamination are known, the applicable one depending on the amino acid and the bacterial species.

Amino acid oxidases are flavoproteins which attack a number of different amino acids according to reaction (36). The imino acid hydrolyzes spontaneously. The flavoprotein is linked to oxygen and produces hydrogen peroxide. Glycine oxidase is a separate specific amino acid oxidase.

$$\underset{\alpha\text{-Amino acid}}{RCHNH_2COOH} \xrightarrow[\text{amino acid oxidase}]{-2H} \underset{\text{Imino acid}}{RC{=}NHCOOH} \xrightarrow{H_2O} \underset{\alpha\text{-Keto acid}}{RCOCOOH} + \underset{\text{Ammonia}}{NH_3} \quad (36)$$

Glutamic and alanine dehydrogenase are NAD- or NADP-linked enzymes which convert the substrates to α-ketoglutaric acid and pyruvic acid, respectively, with the liberation of ammonia. Oxygen is not essential.

Aspartase catalyzes the deamination of aspartic acid to fumaric acid and ammonia.

Beta-substituted amino acids, that is, serine, threonine, and cysteine, are dehydrated through the process shown in reaction (37).

$$\underset{\substack{\beta\text{-Substituted}\\ \alpha\text{-Amino acid}}}{\mathrm{RCHOHCHNH_2COOH}} \xrightarrow[\text{dehydrase}]{-\mathrm{H_2O}} (\mathrm{RCH{=}CNH_2COOH}) \longrightarrow \underset{\text{Imino acid}}{\mathrm{RCH_2C{=}NHCOOH}} \xrightarrow{\mathrm{H_2O}} \underset{\alpha\text{-Keto acid}}{\mathrm{RCH_2COCOOH}} + \underset{\text{Ammonia}}{\mathrm{NH_3}} \qquad (37)$$

Histidine and hydroxy L-proline are oxidized as shown by reactions (38) and (39).

$$\text{Histidine} \xrightarrow[\text{deamination}]{-\mathrm{NH_3}} \text{urocanic acid} \xrightarrow[\text{uricanase}]{\text{HOH}}$$

$$\begin{tikzcd} & \text{glutamic acid + formylglycine} \arrow[r] & \text{formic acid + ammonia} \\ \text{formimino-glutamic acid} \arrow[ur] \arrow[dr] & & (38) \\ & \text{glutamic acid + formamide} \arrow[r] & \text{formic acid + ammonia} \end{tikzcd}$$

$$\text{Hydroxly-L-proline} \longrightarrow \text{allohydroxy-D-proline} \longrightarrow \Delta'\text{-4-hydroxypyrroline-2-carboxylic acid} \longrightarrow \text{2-keto-4-hydroxy-5-aminovaleric acid} \longrightarrow \text{glutamic acid} \qquad (39)$$

2. *Purines and pyrimidines.* The oxidation of purines is best known in the case of uric acid and is shown in reaction (40). The formic acid and CO_2 produced from glyoxylic acid are derived from the carboxyl and α-carbons, respectively.

$$\text{Uric acid} \xrightarrow{+\mathrm{O,\ H_2O}} \text{hydroxyacetylene} \longrightarrow \text{allantoin} \xrightarrow{+\mathrm{H_2O}} \text{allantoic acid} \longrightarrow \text{2 urea + glyoxylic acid} \longrightarrow \mathrm{CO_2} + \text{formic acid} \qquad (40)$$

The initial step, reaction (41), in pyrimidine degradation is a reduction. From uracil the intermediates are β-carbamylaminopropionic acid and β-aminopropionic acid, respectively; from thymine they are β-carbamylaminoisobutyric acid and β-aminoisobutyric acid. The oxidase pathways shown in reactions (42) and (43) have been described in a species of *Mycobacterium*.

$$\text{Pyrimidine} \xrightarrow{+2\mathrm{H}} \text{dihydropyrimidine} \longrightarrow \beta\text{-carbamylamino acid} \longrightarrow \beta\text{-amino acid} + \mathrm{CO_2} + \mathrm{NH_3} \qquad (41)$$

$$\text{Uracil} \xrightarrow[\text{oxidase}]{+\mathrm{O}} \text{barbituric acid} \xrightarrow[\text{barbiturase}]{+\mathrm{H_2O}} \text{urea + malonic acid} \longrightarrow \text{acetic acid} + \mathrm{CO_2} \qquad (42)$$

$$\text{Thymine} \xrightarrow[\text{oxidase}]{+\mathrm{O}} \text{5-methylbarbituric acid} \qquad (43)$$

3. *Pyridines.* The scanty information on the mechanism of degradation of pyridine compounds is limited to the oxidation of pyridoxine (vitamin B_6), nicotinic acid, and nicotine. These appear to be oxidized by way of the pathways shown in reactions (44), (45), and (46).

$$\text{Pyridoxine} \longrightarrow \text{isopyridoxal} \longrightarrow \text{5-pyridoxic acid} \longrightarrow \textit{a}\text{-hydroxymethyl-}\textit{a}\text{-(}\textit{N}\text{-acetylaminomethylene)-succinic acid} \qquad (44)$$

$$\text{Nicotinic acid} \xrightarrow{\mathrm{H_2O}} \text{6-hydroxynicotinic acid} \xrightarrow{-\mathrm{CO_2},\ +\mathrm{O_2}} \text{2,5-dihydroxynicotine} \longrightarrow \text{fumaric acid + formic acid} \qquad (45)$$

$$\text{Nicotine} \longrightarrow \text{3-(4-methylaminobutyryl)-pyridine} \longrightarrow \text{3-succinyl-6-hydroxypyridine} \longrightarrow \ ? \qquad (46)$$

4. *Proteins.* Proteins are hydrolytically broken down to the constituent amino acids by means of proteolytic enzymes and peptidases. Polypeptides of gradually diminishing size are produced; free amino acids are ultimately liberated. These are oxidized as described above.

Anaerobic oxidations. A number of different bacteria are capable of utilizing inorganic oxidants as alternatives to oxygen or organic substances as the ultimate electron, or hydrogen, acceptors. The nature of the oxidant and the change it undergoes distinguish the various kinds of anaerobic oxidations. The reduction of the oxidant is part of the energy-liberating process as contrasted to reduction for purposes of assimilation as a nutrient. In the former case, the amount of oxidant that is reduced per unit of growth is very large, and the reduced product accumulates extracellulary.

Nitrate reduction. A great many aerobic bacteria, including species belonging to a variety of different genera, can facultatively utilize nitrate in the absence of oxygen. The enzyme nitrate reductase is a molybdenum-dependent flavin-linked inducible enzyme. Usually, the specific substrates that an organism can oxidize with nitrate as the ultimate oxidant are the same as those it can oxidize with oxygen. The nitrate is reduced to nitrite, which may accumulate or may be further reduced to ammonia by way of hydroxylamine. Nitrite reductase and hydroxylamine reductase are also flavoproteins. Cytochrome *c* appears to mediate the flow of the electrons to nitrite.

Denitrification. The reduction of nitrate or nitrite to gaseous products constitutes denitrification. Nitrogen gas is the most common product, although nitrous oxide, N_2O, and nitric oxide, NO, are also formed under some conditions. The latter two may be further reduced to N_2, but they are not obligatory intermediates in the denitrification. Nitrite is reduced to nitric oxide by a flavoprotein that requires copper and iron. All of the enzymes in denitrification are capable of receiving electrons from pyridine nucleotide enzymes.

Sulfate reduction. The bacteria effecting sulfate reduction belong to the genus *Desulfovibrio, D. desulfuricans* being the most common species. *Clostridium nigrificans* also is a sulfate reducer. The organisms are obligate anaerobes which reduce sulfate to hydrogen sulfide during oxidation of various substrates. Blackening of the medium results from the interaction of the H_2S with iron. Sulfate reduction is mediated by a cytochrome-linked enzyme, and the active form of sulfate that undergoes reduction is adenosine-3′-phosphate-5′-phosphosulfate. Sulfate-reducing bacteria cause corrosion of iron because, by creating a strongly reduced environment, they locally create anodic conditions relative to the more oxidized surroundings, thus producing a corrosion cell.

Carbonate reduction. Two different types, distinguishable by the nature of the reduction products, are known. They are the following:

1. *CO_2 reduction to acetic acid.* This is performed by certain obligately anaerobic sporeforming bacteria that use CO_2 as the hydrogen acceptor for the oxidation of various substrates. *Clostridium aceticum* thus produces acetic acid from molecular hydrogen and CO_2 in the manner of reaction (47)

$$2CO_2 + 4H_2 \longrightarrow CH_3COOH + 2H_2O \qquad (47)$$

and, like *C. thermoaceticum*, oxidatively decarboxylates pyruvic acid with a concomitant reduction of CO_2 to acetic acid by way of reactions (48). These

$$\begin{array}{rcl} 4CH_3COCOOH + H_2O & \longrightarrow & CH_3COOH + CO_2 + 2[H] \\ 2CO_2 + 8[H] & \longrightarrow & CH_3COOH + 2H_2O \\ \hline 4CH_3COCOOH + 2H_2O & \longrightarrow & 5CH_3COOH + 2CO_2 \end{array} \qquad (48)$$

organisms decompose sugar in accordance with reaction (49), and tracer studies with *C. thermoaceticum* have shown that this conversion proceeds by an initial decomposition of the sugar via pyruvic acid, as shown in reactions (50). *Clostridium acidi-urici* carries out an anaerobic degradation of purines in which the oxidation of the substrate is likewise coupled with the reduction of CO_2 to acetic acid.

$$C_6H_{12}O_6 \longrightarrow 3CH_3COOH \qquad (49)$$

$$\begin{array}{rcl} C_6H_{12}O_6 & \longrightarrow & 2CH_3COCOOH + 4[H] \\ 2CH_3COCOOH & \longrightarrow & 2CH_3COOH + 4[H] + 2CO_2 \\ 2CO_2 + 8[H] & \longrightarrow & CH_3COOH \\ \hline C_6H_{12}O_6 & \longrightarrow & 3CH_3COOH \end{array} \qquad (50)$$

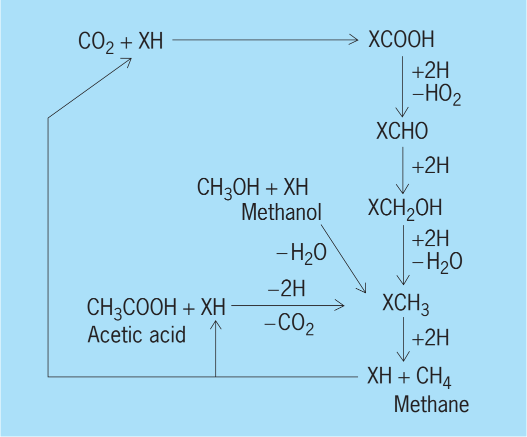


Fig. 10. A possible mechanism for methane fermentation. XH = intracellular reduced carrier or coenzyme.

2. *CO_2 reduction to methane.* The anaerobic decomposition of organic matter by complex microbial populations invariably results in the production of methane from CO_2 by the activities of the so-called methane bacteria. One stage in the disposal of sewage is marked by an intense methane fermentation. The lower primary and secondary alcohols, as well as organic acids, are the most common substrates whose oxidation is linked to the reduction of CO_2 to methane. Hydrogen and carbon monoxide also undergo a methane fermentation. In some cases it has been found that carbon monoxide, formic acid, formaldehyde, and methanol are not intermediates in the bacterial reduction of CO_2 to methane. The bacteria that carry out this process are strict anaerobes and are dependent on CO_2. Four genera are known: *Methanobacillus, Methanobacterium, Methanosarcina,* and *Methanococcus*. In addition to morphological differences, the bacteria display substrate specificities. Some types do produce methane by reduction of methanol, or from the methyl group of acetate, proving that a mechanism independent of CO_2 reduction exists for formation of methane. Experiments with deuterium-labeled acetate have shown that the methane evolved may comprise the intact methyl groups, suggesting a reductive demethylation of acetate. *See* METHANOGENESIS (BACTERIA).

The scheme in **Fig. 10** represents a possible mechanism of the methane fermentation.

Jackson W. Foster; R. E. Kallio

Catabolism: Fermentative Metabolism

Many bacteria are able to decompose organic compounds and to grow in the absence of oxygen gas. Such anaerobic bacteria obtain energy and certain organic compounds needed for growth by a process of fermentation. This consists of an oxidation of a suitable organic compound, using another organic compound as an oxidizing agent in place of molecular oxygen. In most fermentations both the compounds oxidized and the compounds reduced (used as an oxidizing agent) are derived from a single fermentable substrate. In other fermentations, one substrate is oxidized and another is reduced. Different bacteria ferment different substrates. Many bacteria are able to ferment carbohydrates such as glucose and sucrose, polyalchohols such as mannitol, and salts of organic acids such as pyruvate and lactate. Other compounds, such as cellulose, amino acids, and purines, are fermented by some bacteria.

Bacterial fermentations are rather arbitrarily classified according to either the nature of the substrate or the nature of the characteristic product. These two types of classification are not mutually exclusive and lead to different associations of genera and species catalyzing fermentative reactions. In the following discussion, the primary classification is based upon the nature of the substrate, which is either nonnitrogenous (carbohydrates and related compounds) or nitrogenous (amino acids, purines). The further subdivision of the carbohydrate fermentations is based largely on the nature of the characteristic fermentation product.

Bacterial carbohydrate fermentations. Some of the characteristic products of bacterial carbohydrate fermentations are ethyl alcohol, lactic acid, acetic acid, propionic acid, butyric acid, butyl alcohol, acetone, and butanediol. Because certain microorganisms form predominantly one or another of these compounds, the name of the product has come to be associated with the type of fermentation. In this way such terms as alcoholic fermentation, lactic acid fermentation, and propionic acid fermentation have become established, though often of more historical than scientific value.

Alcoholic fermentation. Alcoholic fermentation is the process by which certain yeasts decompose sugars in the absence of oxygen to form ethanol and carbon dioxide, according to reaction (51). This fermentation provides the basis for the commercial production of ethanol and alcoholic beverages such as wine and beer.

$$\underset{\text{Glucose}}{C_6H_{12}O_6} \longrightarrow \underset{\text{Ethanol}}{2CH_3CH_2OH} + \underset{\text{Carbon dioxide}}{2CO_2} \qquad (51)$$

The conversion of glucose to ethanol and carbon dioxide is a complex enzymatic process involving many chemical steps. The main features of this process are the phosphorylation of the hexose (six-carbon) sugar to form a hexose diphosphate, a cleavage of this compound into two molecules of triose (three-carbon) phosphate, the oxidation of the latter and its further conversion to pyruvate, the decarboxylation of pyruvate to carbon dioxide and acetaldehyde, and the reduction of the latter to ethanol. An essential feature of this process is a coupled conversion of inorganic orthophosphate to a high-energy phosphate group in ATP. Two moles of phosphate are converted in this way for each mole of glucose fermented. The resulting ATP is used to provide energy for the biosynthesis of essential cellular components.

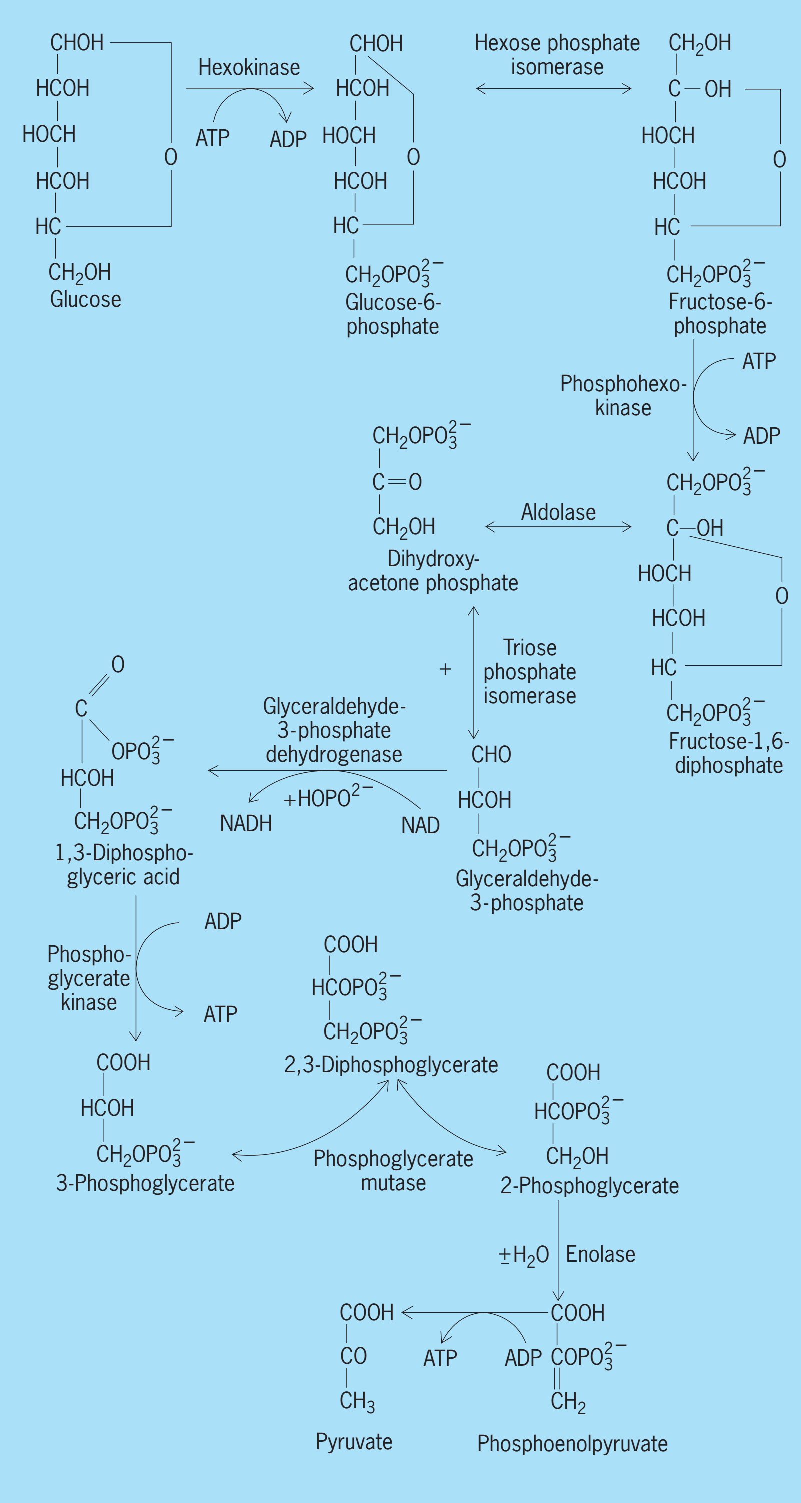

Glucose
Hexokinase
ATP
ADP
Glucose-6-phosphate
Hexose phosphate isomerase
Fructose-6-phosphate
Phosphohexokinase
ATP
ADP
Fructose-1,6-diphosphate
Aldolase
Dihydroxyacetone phosphate
Triose phosphate isomerase
Glyceraldehyde-3-phosphate
Glyceraldehyde-3-phosphate dehydrogenase
+HOPO$_3^{2-}$
NADH
NAD
1,3-Diphosphoglyceric acid
Phosphoglycerate kinase
ADP
ATP
3-Phosphoglycerate
2,3-Diphosphoglycerate
Phosphoglycerate mutase
2-Phosphoglycerate
±H$_2$O
Enolase
Phosphoenolpyruvate
ATP
ADP
Pyruvate

Fig. 11. Embden-Meyerhof pathway of glycolysis and alcoholic fermentation.

Glycolysis is the conversion of glucose to lactic acid. Glycolysis and alcoholic fermentation by yeast involve the same chemical steps from glucose to pyruvate. This sequence of enzymatic reactions, also known as the Embden-Meyerhof pathway, is shown in **Fig. 11**. The first step is a transfer of a phosphate group from ATP to glucose to form glucose-6-phosphate. This is rearranged to form fructose-6-phosphate, which reacts with ATP to form fructose-1,6-diphosphate. This compound is cleaved by the enzyme aldolase into 1 mole each of dihydroxyacetone phosphate and D-glyceraldehyde-3-phosphate. The enzyme triose phosphate isomerase interconverts the two triose phosphates. After these preparatory reactions, glyceraldehyde-3-phosphate is oxidized and NAD is reduced. This oxidation-reduction reaction, catalyzed by the enzyme glyceraldehyde phosphate dehydrogenase, requires the presence of orthophosphate. During the reaction, orthophosphate disappears and 1,3-diphosphoglyceric acid is formed. The phosphate group attached to carbon atom 1 of this compound is transferred to ADP to form 3-phosphoglyceric acid and ATP. The enzyme phosphoglyceromutase transfers the phosphate group from the 3 to the 2 position of glyceric acid in a reaction involving 2,3-diphosphoglyceric acid. Enolase dehydrates 2-phosphoglyceric acid to phosphoenol pyruvic acid, a compound containing a “high-energy” phosphate group. Finally the phosphate group is transferred to ADP to yield pyruvic acid and ATP. The conversion of 1 mole of glucose to 2 moles of pyruvic acid is accompanied by an uptake of 2 moles of orthophosphate, P_i, and a net formation of 2 moles of ATP, according to reaction (52).

$$C_6H_{12}O_6 + 2NAD + 2ADP + 2P_i \longrightarrow 2CH_3COCOOH + 2ATP + 2NADH \tag{52}$$

In glycolysis by muscle tissue, the reduced diphosphopyridine nucleotide (NADH) is reoxidized back to the oxidized form (NAD) by reaction (53), in

$$\underset{\text{Pyruvate}}{CH_3COCOOH} + NADH + H^+ \longrightarrow \underset{\text{Lactate}}{CH_3CHOHCOOH} + NAD^+ \tag{53}$$

which pyruvic acid is simultaneously reduced to lactic acid.

In alcoholic fermentation by yeast enzymes the acetaldehyde which is formed in the decarboxylation of pyruvic acid, reaction (54*a*), reoxidizes the NADH, reaction (54*b*). The overall reaction is shown in reaction (55).

$$\underset{\text{Pyruvic acid}}{CH_3COCOOH} \longrightarrow \underset{\text{Acetaldehyde}}{CH_3CHO} + CO_2 \tag{54a}$$

$$\underset{\text{Acetaldehyde}}{CH_3CHO} + NADH + H^+ \longrightarrow \underset{\text{Ethanol}}{CH_3CH_2OH} + NAD^+ \tag{54b}$$

$$C_6H_{12}O_6 + 2ADP + 2P_i \longrightarrow 2CH_3CH_2OH + 2CO_2 + 2ATP \tag{55}$$

In living yeast the ATP (present in catalytic amounts) is used in various biosynthetic reactions. In cellfree yeast extract, in which conditions are not favorable for the occurrence of many biosynthetic reactions, ATP is used mainly to phosphorylate excess glucose. Both types of reaction continuously regenerate ADP, which is essential for the oxidation of glyceraldehyde phosphate.

The pentose phosphate pathway of glucose decomposition involves hexose monophosphates and pentose monophosphates but not fructose diphosphate. This pathway is used by some anaerobic bacteria. It is shown in **Fig. 12**.

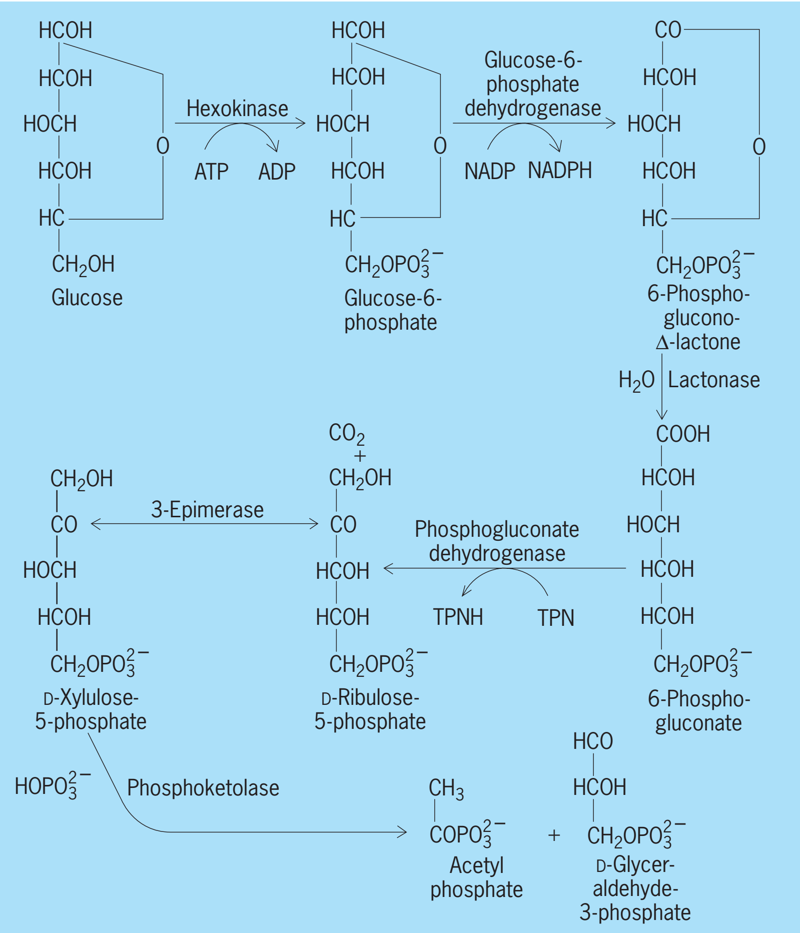
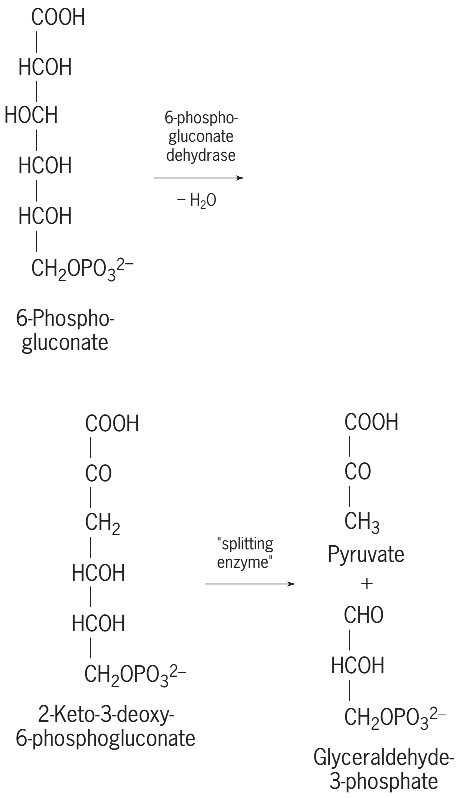


Fig. 12. Pentose phosphate pathway of glucose decomposition in some anaerobic bacteria. Pathway involves hexose and pentose monophosphates.

The first step is the same as in glycolysis: the conversion of glucose to glucose-6-phosphate. The latter compound is oxidized by a dehydrogenase to 6-phosphoglucono-Δ-lactone; the oxidizing agent is NADP. The lactone is hydrolyzed to 6-phosphogluconic acid which is then oxidized by NADP and a specific dehydrogenase to ribulose-5-phosphate and carbon dioxide. A rearrangement at carbon atom 3 under the influence of an epimerase results in the formation of xylulose-5-phosphate. This compound is cleaved between carbon atoms 2 and 3 in an enzymatic reaction that consumes orthophosphate, P_i, and produces glyceraldehyde-3-phosphate and acetyl phosphate. The glyceraldehyde phosphate is converted by a portion of the Embden-Meyerhof pathway to pyruvate. The further fate of pyruvate and acetyl phosphate depends upon the enzymatic makeup of the organism using this metabolic path.

The 2-keto-3-deoxy-6-phosphogluconate (Entner-Doudoroff) pathway represents another method of degrading glucose with the liberation of energy. This pathway has been studied mainly in an aerobic organism, *Pseudomonas saccharophila*, but probably also functions in some anaerobic bacteria. The distinguishing feature of the Entner-Doudoroff pathway of glucose metabolism is shown in reaction sequence (56).



$$\begin{array}{c}\text{COOH}\\|\\\text{HCOH}\\|\\\text{HOCH}\\|\\\text{HCOH}\\|\\\text{HCOH}\\|\\\text{CH}_2\text{OPO}_3{}^{2-}\\\text{6-Phospho-}\\\text{gluconate}\end{array}\xrightarrow[-\,H_2O]{\text{6-phospho-gluconate dehydrase}}$$

$$\begin{array}{c}\text{COOH}\\|\\\text{CO}\\|\\\text{CH}_2\\|\\\text{HCOH}\\|\\\text{HCOH}\\|\\\text{CH}_2\text{OPO}_3{}^{2-}\\\text{2-Keto-3-deoxy-}\\\text{6-phosphogluconate}\end{array}\xrightarrow{\text{"splitting enzyme"}}\begin{array}{c}\text{COOH}\\|\\\text{CO}\\|\\\text{CH}_3\\\text{Pyruvate}\\+\\\text{CHO}\\|\\\text{HCOH}\\|\\\text{CH}_2\text{OPO}_3{}^{2-}\\\text{Glyceraldehyde-}\\\text{3-phosphate}\end{array}\qquad(56)$$

The initial steps, from glucose to 6-phosphogluconate, are the same as in the pentose phosphate pathway. The 6-phosphogluconate is then dehydrated and rearranged under the influence of a dehydrase to form the characteristic compound of this pathway, 2-keto-3-deoxy-6-phosphogluconate. A "splitting enzyme" cleaves this compound between carbon atoms 3 and 4 to yield pyruvate from glucose carbon atoms 1–3 and glyceraldehyde-3-phosphate from carbon atoms 4–6. The conversion of the triose phosphate to a second mole of pyruvate follows the Embden-Meyerhof pathway. The further reactions of pyruvate vary with the organism.

Alcoholic fermentation by bacteria is relatively uncommon. *Pseudomonas lindneri* (*Zymononas mobilis*), a bacterium found in pulque, ferments glucose or fructose to 45% carbon dioxide, 45% ethanol, and 7% lactic acid. The fermentation appears to follow the Entner-Doudoroff pathway. This is indicated by the fact that the carbon dioxide is derived from glucose carbon atoms 1 and 4, and the primary alcohol group of ethanol is derived from glucose carbon atoms 2 and 5. In alcoholic fermentation by yeast, carbon dioxide is derived from glucose carbon atoms 3 and 4, and the alcohol group of ethanol is derived from glucose carbon atoms 2 and 5 (**Fig. 13**).

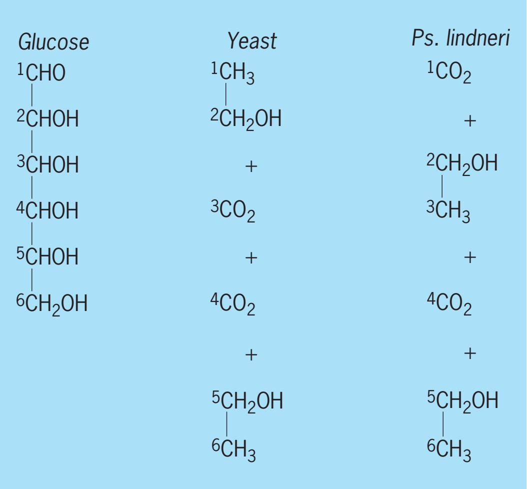


Fig. 13. Comparison of the products of alcoholic fermentation in yeast and *Pseudomonas lindneri*.

Lactic acid fermentation. Lactic acid fermentations are processes by which glucose and some other carbohydrates are converted to lactic acid and sometimes other products. These fermentations are caused by species of *Streptococcus, Leuconostoc, Lactobacillus*, and some other genera. These bacteria are responsible for the souring of milk and cream, for the preservation of silage, and for some types of food spoilage. *See* MILK.

Lactic acid fermentations may be divided into two types, the homolactic and heterolactic fermentations.

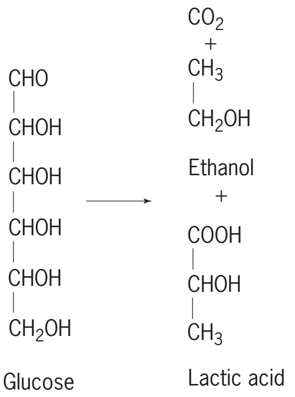
Homolactic fermentations, caused mainly by *Streptococcus* and certain *Lactobacillus* species, convert sugar almost entirely to lactic acid, as in reaction (57).

$$\underset{\text{Hexose}}{C_6H_{12}O_6} \longrightarrow \underset{\text{Lactic acid}}{2CH_3CHOHCOOH} \qquad (57)$$

Only when the fermentations occur in an alkaline medium do other products such as formate, acetate, and ethanol accumulate in considerable amounts. *Streptococcus* species usually form L(+)-lactic acid, the isomer metabolized most readily by humans. *Lactobacillus* species may form either L(+)-, D(−)-, or racemic lactic acid. Homolactic fermentations occur by the glycolytic (Embden-Meyerhof) pathway. The pyruvic acid is reduced to lactic acid by reduced nicotinamide adenine dinucleotide (NADH) under the influence of the enzyme lactic dehydrogenase by reaction (58).

$$\underset{\text{Pyruvic acid}}{CH_3COCOOH} + NADH + H^+ \longrightarrow \underset{\text{Lactic acid}}{CH_3CHOHCOOH} + NAD^+ \qquad (58)$$

Heterolactic fermentations, caused mainly by *Leuconostoc* and certain *Lactobacillus* species, produce lactic acid in smaller yield and convert a considerable part of the sugar to carbon dioxide and other products, such as ethanol, acetic acid, glycerol, and mannitol. *Leuconostoc mesenteroides* ferments glucose according to reaction (59). The glycolytic pathway of sugar decomposition cannot account for the formation of equimolar amounts of the three products. The origin of the carbon atoms of the products, established by experiments with ^{14}C-labeled glucose, is consistent with the occurrence of the pentose phosphate pathway. The acetyl phosphate produced in this sequence is reduced first to acetaldehyde and then to ethanol by use of reducing agents (NADPH and NADH) produced during the oxidation of glucose to acetyl phosphate, pyruvate, and carbon dioxide. Apparently all typical heterolactic fermentations involve the pentose phosphate pathway. Heterolactic fermentations of fructose, but not glucose, characteristically produce mannitol in good yield.



$$\begin{array}{c} \mathrm{CHO} \\ | \\ \mathrm{CHOH} \\ | \\ \mathrm{CHOH} \\ | \\ \mathrm{CHOH} \\ | \\ \mathrm{CHOH} \\ | \\ \mathrm{CH_2OH} \\ \\ \text{Glucose} \end{array} \longrightarrow \begin{array}{c} \mathrm{CO_2} \\ + \\ \mathrm{CH_3} \\ | \\ \mathrm{CH_2OH} \\ \text{Ethanol} \\ + \\ \mathrm{COOH} \\ | \\ \mathrm{CHOH} \\ | \\ \mathrm{CH_3} \\ \\ \text{Lactic acid} \end{array} \qquad (59)$$

Several homolactic and heterolactic species are able to ferment the pentose sugars, xylose, arabinose, and ribose. The products of pentose fermentation by both types of lactic acid, bacteria, reaction (60), are lactic and acetic acids.

$$\underset{\text{Pentose}}{C_5H_{10}O_5} \longrightarrow \underset{\text{Lactic acid}}{CH_3CHOHCOOH} + \underset{\text{Acetic acid}}{CH_3COOH} \qquad (60)$$

Propionic acid fermentation. The propionic acid fermentation converts sugars, polyalcohols, salts of organic acids such as lactate and pyruvate, and some amino acids such as alanine and serine to propionic acid and often other products, including acetate, succinate, propanol, and carbon dioxide. Most of the bacteria that cause the propionic acid fermentation are species of *Propionibacterium*. A few other bacteria, including *Clostridium propionicum* and *Micrococcus (Veillonella) lactilyticus*, cause a similar fermentation.

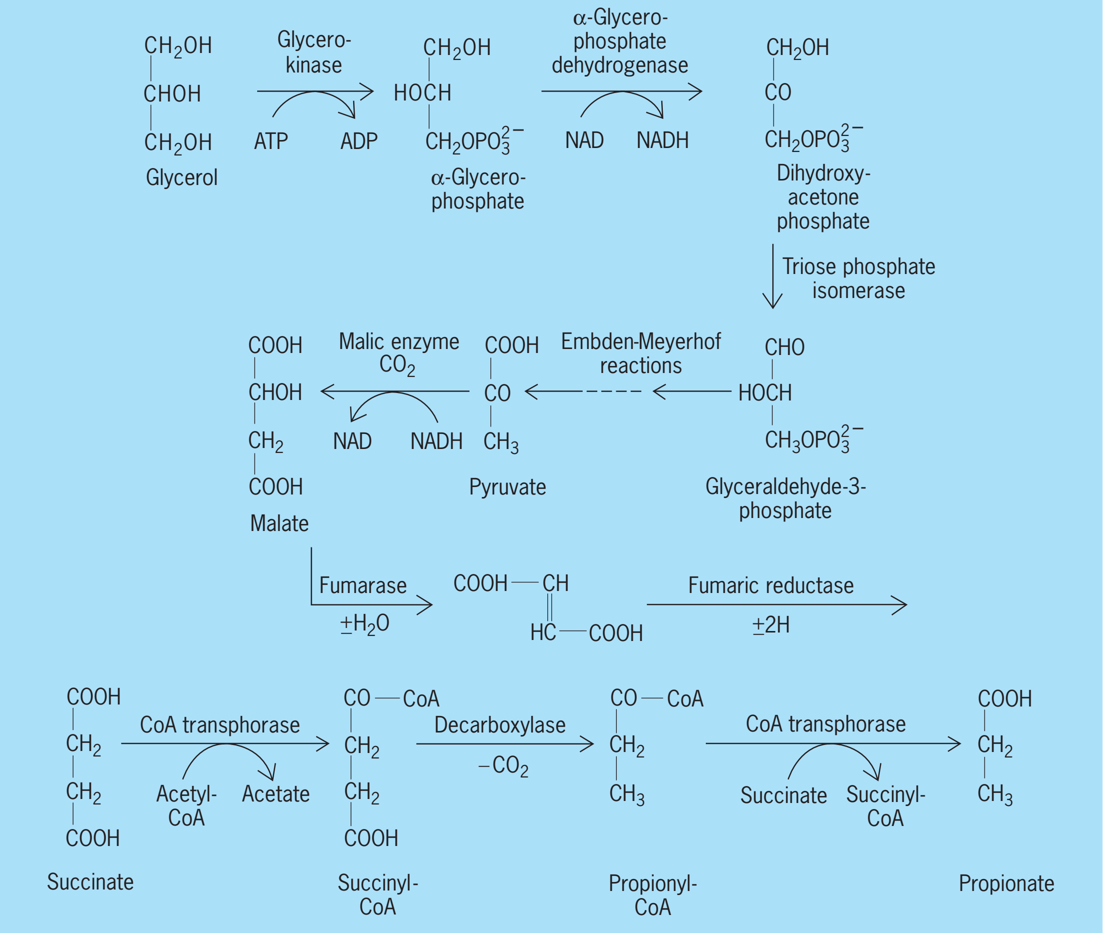

CH$_2$OH
CHOH
CH$_2$OH
Glycerol
Glycero-kinase
ATP
ADP
CH$_2$OH
HOCH
CH$_2$OPO$_3^{2-}$
α-Glycero-phosphate
α-Glycero-phosphate dehydrogenase
NAD
NADH
CH$_2$OH
CO
CH$_2$OPO$_3^{2-}$
Dihydroxy-acetone phosphate
Triose phosphate isomerase
CHO
HOCH
CH$_3$OPO$_3^{2-}$
Glyceraldehyde-3-phosphate
Embden-Meyerhof reactions
COOH
CO
CH$_3$
Pyruvate
Malic enzyme
CO$_2$
NAD
NADH
COOH
CHOH
CH$_2$
COOH
Malate
Fumarase
±H$_2$O
COOH—CH
HC—COOH
Fumaric reductase
±2H
COOH
CH$_2$
CH$_2$
COOH
Succinate
CoA transphorase
Acetyl-CoA
Acetate
CO—CoA
CH$_2$
CH$_2$
COOH
Succinyl-CoA
Decarboxylase
−CO$_2$
CO—CoA
CH$_2$
CH$_3$
Propionyl-CoA
CoA transphorase
Succinate
Succinyl-CoA
COOH
CH$_2$
CH$_3$
Propionate

Fig. 14. Probable sequence of reactions for propionic acid fermentation from glycerol.

The simplest propionic acid fermentation is that of glycerol; 90% of the substrate may be converted to propionic acid, as shown in reaction (61). When

$$\underset{\text{Glycerol}}{CH_2OHCHOHCH_2OH} \longrightarrow \underset{\text{Propionic acid}}{CH_3CH_2COOH} + H_2O \qquad (61)$$

the medium contains a considerable concentration of carbon dioxide, succinic acid is also formed and the yield of propionic acid is reduced. For each mole of succinate accumulating, 1 mole of carbon dioxide is consumed. Experiments with ^{14}C-labeled carbon dioxide show that carbon dioxide carbon is incorporated into the carboxyl groups of both propionic and succinic acids. There is considerable evidence that succinic acid and succinyl coenzyme A are intermediates in the formation of propionic acid. Propionic acid is probably formed from glycerol by the sequence shown in **Fig. 14**.

The acetyl-CoA required for the formation of succinyl-CoA is probably formed by the reaction shown in reaction (62). Lactate, a favorite substrate

$$\underset{\text{Acetic acid}}{CH_3COOH} + ATP + CoA \longrightarrow \underset{\text{Acetyl-CoA}}{CH_3CO - CoA} + AMP + PP \qquad (62)$$

of propionic acid bacteria, is fermented approximately according to reaction (63). A little succinic

$$\underset{\text{Lactic acid}}{3CH_3CHOHCOOH} \longrightarrow \underset{\text{Propionic acid}}{2CH_3CH2COOH} + \underset{\text{Acetic acid}}{CH_3COOH} + CO_2 + H_2O \qquad (63)$$

acid is also formed. Hexose and pentose sugars are converted to the same products.

The propionic acid fermentation is essential for the preparation of Swiss cheese. The characteristic flavor is partially attributable to propionic acid and the holes are caused by carbon dioxide formed by the propionic acid bacteria. *See* CHEESE.

Butyric acid fermentation. The butyric acid fermentation results in the formation of butyric acid and usually other products which may include hydrogen, carbon dioxide, acetic acid, and caproic acid. This type of fermentation is usually caused by anaerobic spore-forming bacteria such as *Clostridium butyricum* and *C. kluyveri*. A few nonsporulating bacteria, such as *Butyribacterium rettgeri* and *Sarcina maxima*, also produce butyric acid. Suitable substrates for the butyric acid fermentation are hexose and pentose sugars, polyalcohols, polysaccharides, salts of organic acids (tartrate, lactate), and occasionally ethanol plus a short-chain fatty acid.

The simplest butyric acid fermentation is caused by *C. kluyveri*, which converts ethanol and acetate to caproic acid. The fermentation involves two stages; in the first stage butyric acid is formed, reaction (64*a*), and in the second stage it is converted to caproic acid, reaction (64*b*). These reactions

$$\underset{\text{Ethanol}}{CH_3CH_2OH} + \underset{\text{Acetic acid}}{CH_3COOH} \longrightarrow \underset{\text{Butyric acid}}{CH_3CH_2CH_2COOH} + H_2O \qquad (64a)$$

$$\underset{\text{Ethanol}}{CH_3CH_2OH} + \underset{\text{Butyric acid}}{CH_3CH_2CH_2COOH} \longrightarrow \underset{\text{Caproic acid}}{CH_3CH_2CH_2CH_2CH_2COOH} + H_2O \qquad (64b)$$

provide the energy needed by the bacteria.

The conversion of ethanol and acetate to butyrate, like all fermentations, is a relatively complex process. The individual enzymatic steps are shown in the reaction sequence in **Fig. 15**.

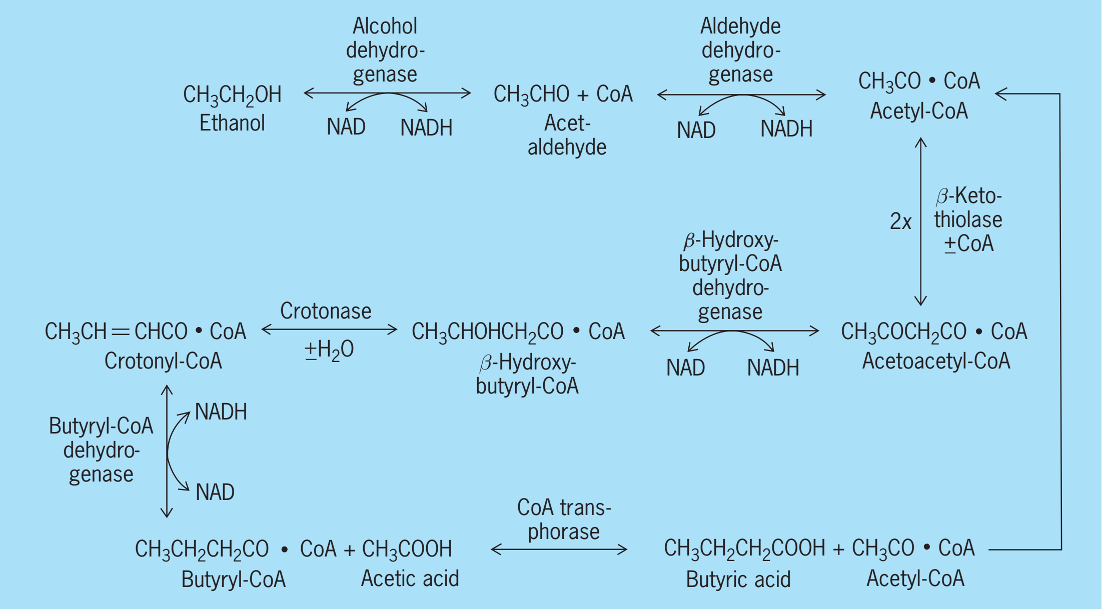

Alcohol dehydrogenase
CH₃CH₂OH
Ethanol
NAD
NADH
CH₃CHO + CoA
Acetaldehyde
Aldehyde dehydrogenase
NAD
NADH
CH₃CO • CoA
Acetyl-CoA
2x
β-Ketothiolase
±CoA
β-Hydroxybutyryl-CoA dehydrogenase
CH₃CH=CHCO • CoA
Crotonyl-CoA
Crotonase
±H₂O
CH₃CHOHCH₂CO • CoA
β-Hydroxybutyryl-CoA
NAD
NADH
CH₃COCH₂CO • CoA
Acetoacetyl-CoA
Butyryl-CoA dehydrogenase
NADH
NAD
CH₃CH₂CH₂CO • CoA + CH₃COOH
Butyryl-CoA
Acetic acid
CoA transphorase
CH₃CH₂CH₂COOH + CH₃CO • CoA
Butyric acid
Acetyl-CoA

Fig. 15. Enzymatic steps in the conversion of ethanol and acetate to butyrate.

Caproic acid is formed by a similar sequence of reactions, in which a molecule of acetyl-CoA is replaced by butyryl-CoA in the β-ketothiolase reaction. Elongation of the carbon chain occurs by addition of two carbon moieties onto the carboxyl carbon of butyric acid.

Similar reactions occur in the butyric acid fermentation of glucose. The sugar appears to be converted to pyruvate by the Embden-Meyerhof pathway. Pyruvate is then oxidized to acetyl-CoA with the simultaneous formation of carbon dioxide and hydrogen according to reaction (65). As a consequence of this reaction, 2 moles of carbon dioxide and 2 moles of hydrogen are formed per mole of glucose fermented. The acetyl-CoA is converted to butyric acid with the aid of NADH produced during the oxidation of triose phosphate.

$$\underset{\text{Pyruvic acid}}{CH_3COCOOH} + CoA \longrightarrow \underset{\text{Acetyl-CoA}}{CH_3CO - CoA} + CO_2 + H_2 \qquad (65)$$

Butanol-acetone fermentation. The butanol-acetone fermentation yields considerable amounts of the solvents butanol and acetone and smaller amounts of acetic and butyric acids along with the gases hydrogen and carbon dioxide. This fermentation, catalyzed by *C. acetobutylicum*, has been extensively used for the commercial production of butanol and acetone from starch and sugars.

The butanol-acetone fermentation results from a small modification of the chemistry of the butyric acid fermentation of carbohydrates. Initially the same products, mainly butyric and acetic acids, carbon dioxide, and hydrogen, are formed in both fermentations. Then, if the medium is permitted to become acid, a leak develops in the sequence of reactions for converting acetyl-CoA to butyrate. This leak develops as a result of the formation of an enzyme system that converts acetoacetyl-CoA to acetoacetate and then decarboxylates the latter to acetone, as shown in reaction (66). The destruction of acetoacetyl-CoA both causes the accumulation of acetone and prevents the further formation of butyrate. The reducing agents, initially used to reduce acetoacetyl-CoA to butyryl-CoA, are then used for the reduction of butyryl-CoA via butyraldehyde to *n*-butanol, asshown in reaction (67).

$$\underset{\text{Acetoacetyl-CoA}}{CH_3COCH_2CO - CoA} \longrightarrow CoA + \underset{\text{Acetoacetate}}{CH_3COCH_2COOH} \xrightarrow[\text{decarboxylase}]{\text{acetoacetate}} \underset{\text{Acetone}}{CH_3COCH_3} + CO_2 \qquad (66)$$

$$\underset{\text{Butyryl-CoA}}{CH_3CH_2CH_2CO \cdot CoA} \xrightarrow[\text{NADH} \quad \text{NAD}]{} CoA + \underset{\text{Butyraldehyde}}{CH_3CH_2CH_2CHO} \xrightarrow[\text{NADH} \quad \text{NAD}]{} \underset{\text{Butanol}}{CH_3CH_2CH_2CH_2OH} \qquad (67)$$

Consequently, acetone and butanol are formed simultaneously. During the solvent-producing stage of the fermentation, a large increase in the formation of carbon dioxide relative to hydrogen results from acetoacetic acid decarboxylation.

Clostridium butylicum, a species closely related to *C. acetobutylicum*, forms an additional product, isopropanol. This branched-chain secondary alcohol is formed by reduction of acetone, reaction (68).

$$\underset{\text{Acetone}}{CH_3COCH_3} + 2H \longrightarrow \underset{\text{Isopropanol}}{CH_3CHOHCH_3} \qquad (68)$$

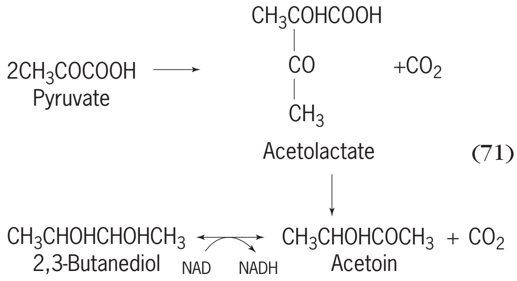
Ethanol-acetone fermentation. The ethanol-acetone fermentation of carbohydrates is catalyzed by *Aerobacillus macerans*. This species forms ethanol, acetone, acetic and formic acids, carbon dioxide, and hydrogen from glucose or starch. The chemistry of the fermentation has not been studied extensively but is probably similar to that of the butanol-acetone fermentation, except for the absence of reactions from acetoacetyl-CoA to butyrate. The acetone is formed by decarboxylation of acetoacetic acid; and the ethanol, a major product, originates by reduction of acetyl-CoA via acetaldehyde, a reaction analogous to the reduction of butyryl-CoA to butanol by *C. acetobutylicum*.

The 2,3-butanediol fermentation of carbohydrates is caused by *Aerobacter* and *Serratia* species, *Aerobacillus polymyxa*, and some strains of *Bacillus subtilis*. This type of fermentation produces large yields of carbon dioxide, ethanol, and 2,3-butanediol; moderate yields of formate and hydrogen; and small yields of acetate, lactate, and succinate. In *Aerobacter* species the coversion of glucose to pyruvate appears to follow the Embden-Meyerhof pathway. Some of the pyruvate undergoes a cleavage reaction involving coenzyme A with formation of acetyl-CoA and formate, reaction (69). The acetyl-CoA is partly converted to acetate and partly reduced via acetaldehyde to ethanol. The formate is partly broken down by the formic hydrogenylase enzyme system to carbon dioxide and hydrogen, as shown in reaction (70). The remainder of the pyruvate is mostly converted to acetolactate and carbon

$$\underset{\text{Pyruvate}}{CH_3COCOOH} + CoA \longrightarrow \underset{\text{Acetyl-CoA}}{CH_3CO - CoA} + \underset{\text{Formate}}{HCOOH} \qquad (69)$$

$$\underset{\text{Formate}}{HCOOH} \longrightarrow CO_2 + H_2 \qquad (70)$$

dioxide through reaction (71). The acetolactate is rapidly decarboxylated by a specific enzyme to acetoin. The latter compound is an effective oxidizing agent under anaerobic conditions; it is readily reduced under the influence of a specific dehydrogenase to 2,3-butanediol. Under aerobic conditions, some bacteria can slowly oxidize acetoin to diacetyl, reaction (72), a slightly volatile compound responsible for the flavor of butter.



(71)

$$\underset{\text{Acetoin}}{CH_3CHOHCOCH_3} + {}^{1}\!/\!{}_{2}O_2 \longrightarrow \underset{\text{Diacetyl}}{CH_3COCOCH_3} + H_2O \qquad (72)$$

Fermentations of nitrogen compounds. Under anaerobic conditions, bacteria are able to decompose many organic nitrogenous compounds, including amino acids and various heterocyclic compounds such as the purines. Fermentations of such compounds are described as putrefactive because they frequently result in the formation of substances having unpleasant odors. The bacteria that ferment nitrogenous compounds are mostly species of *Clostridium* (*C. botulinum, C. perfringens, C. sporogenes, C. tetani*, and *C. tetanomorphum*), but several species belonging to the genera *Fusobacterium, Diplococcus*, and *Micrococcus* also cause such fermentation.

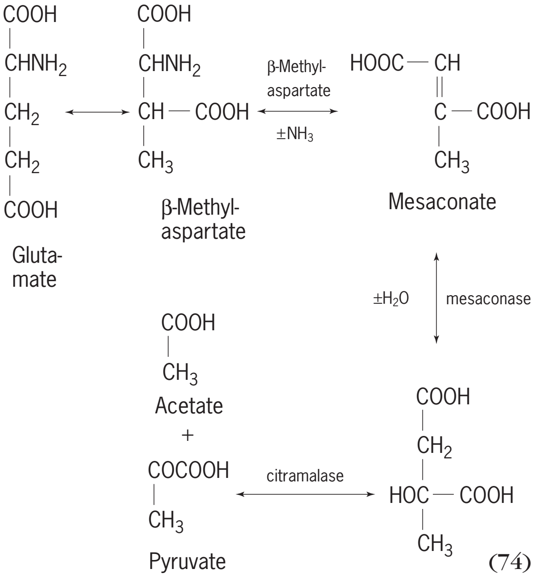
Amino acid fermentation. Amino acid fermentations involve either the anerobic decomposition of a single substrate or the coupled decomposition of two or more substrates. The latter type of decomposition is known as the Stickland reaction.

Most of the common amino acids can be fermented singly by certain anaerobic bacteria. Alanine, cysteine, glutamate, glycine, histidine, serine, and threonine are some of the amino acids known to be fermented. The products usually include one or more of the short-chain fatty acids (acetic, propionic, butyric), carbon dioxide, hydrogen, and ammonia.

The decomposition of alanine by *C. propionicum* is one of the simpler amino acid fermentations. The process, described by reaction (73), is similar to the propionic acid fermentation of lactate by the same organism except for the additional formation of ammonia. The chemical steps in the fermentation of alanine have not been completely elucidated, but they include an oxidation of 1 mole of alanine to acetate, carbon dioxide, and ammonia, via pyruvate, and a reduction of 2 moles of alanine to propionate and ammonia, probably via the coenzyme A derivative of acrylate.

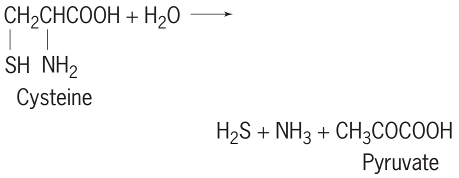
$$\underset{\text{Alanine}}{3CH_3CHNH_2COOH} + 2H_2O \longrightarrow \underset{\text{Propionate}}{2CH_3CH_2COOH} \longrightarrow \underset{\text{Acetate}}{CH_3COOH} + CO_2 + 3NH_3 \qquad (73)$$

The fermentation of glutamate by *C. tetanomorphum* results in the formation of ammonia, acetate, butyrate, carbon dioxide, and hydrogen. This modified butyric acid fermentation involves the transient formation of several branched-chain dicarboxylic acids shown in reaction (74). These reactions result in the conversion of glutamate to acetate and pyruvate. The latter compound undergoes oxidative decarboxylation to form an acetyl compound which is converted partly to acetate and partly to butyrate. Energy for synthetic reactions is produced during the fermentation of pyruvate as in other butyric acid fermentations.



(74)

The aerobic decomposition of the sulfur-containing amino acid cysteine by *Proteus vulgaris* and other bacteria results in the formation of hydrogen sulfide (H_2S), ammonia, and pyruvate, as shown in reaction (75). By an analogous reaction another sulfur-containing amino acid, methionine, gives methylmercaptan (CH_3SH), ammonia, and α-ketobutyrate as shown in reaction (76). Both hydrogen sulfide and methylmercaptan contribute



$$\underset{\text{Cysteine}}{\underset{SH\ \ NH_2}{CH_2CHCOOH}} + H_2O \longrightarrow H_2S + NH_3 + \underset{\text{Pyruvate}}{CH_3COCOOH} \qquad (75)$$

$$\underset{\text{Methionine}}{CH_3S{-}CH_2CH_2CHNH_2COOH} + H_2O \longrightarrow CH_3SH + NH_3 + \underset{\alpha\text{-Ketobutyrate}}{CH_3CH_2COCOOH} \qquad (76)$$

to the putrefactive odor of decomposing proteins. The two keto acids are further fermented to fatty acids and carbon dioxide.

A decarboxylation of certain amino acids, according to reaction (77), is catalyzed by a number of

$$\mathrm{RCHNH_2COOH} \longrightarrow \mathrm{RCH_2NH_2 + CO_2} \qquad (77)$$

bacteria. The amino acids most commonly decarboxylated are histidine, tyrosine, arginine, ornithine, lysine, tryptophan, and glutamate. The decarboxylation of lysine, for example, produces cadaverine, reaction (78), a basic substance having a very un-

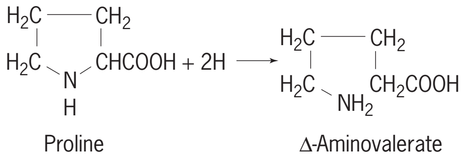
$$\underset{\text{Lysine}}{\mathrm{NH_2(CH_2)_4CHNH_2COOH}} \longrightarrow \underset{\text{Cadaverine}}{\mathrm{NH_2(CH_2)_4CH_2NH_2}} + \mathrm{CO_2} \qquad (78)$$

pleasant odor. The decarboxylation reaction occurs mainly in bacteria exposed to a relatively acidic medium. Because the products of decarboxylation are more basic than the amino acids, the reaction helps to maintain a neutral environment.

The Stickland reaction, the fermentation of pairs of different amino acids, is caused by many *Clostridium* species such as *C. sporogenes* and *C. botulinum*. One amino acid in the pair is oxidized whereas the other is reduced. The simplest example is the coupled decomposition of alanine and glycine according to reaction (79). In this process the oxidation of

$$\underset{\text{Alanine}}{\mathrm{CH_3CHNH_2COOH}} + \underset{\text{Glycine}}{\mathrm{2CH_2NH_2COOH}} + \mathrm{2H_2O} \longrightarrow \mathrm{3NH_3} + \underset{\text{Acetate}}{\mathrm{3CH_3COOH}} + \mathrm{CO_2} \qquad (79)$$

1 mole of alanine to acetate, carbon dioxide, and ammonia is coupled with the reduction of 2 moles of glycine to acetate and ammonia. Other amino acids that are oxidized include leucine, isoleucine, valine, and histidine. The products are always ammonia, carbon dioxide, and a carboxylic acid with one less carbon atom. Other amino acids that are reduced include proline, hydroxyproline, ornithine, and tryptophan. The reduction of proline (or hydroxyproline) converts a ring compound to a straight-chain compound without formation of ammonia, for example, reaction (80). With ornithine or tryptophan,



(80)

as with glycine, the α-amino group is removed by a reductive deamination.

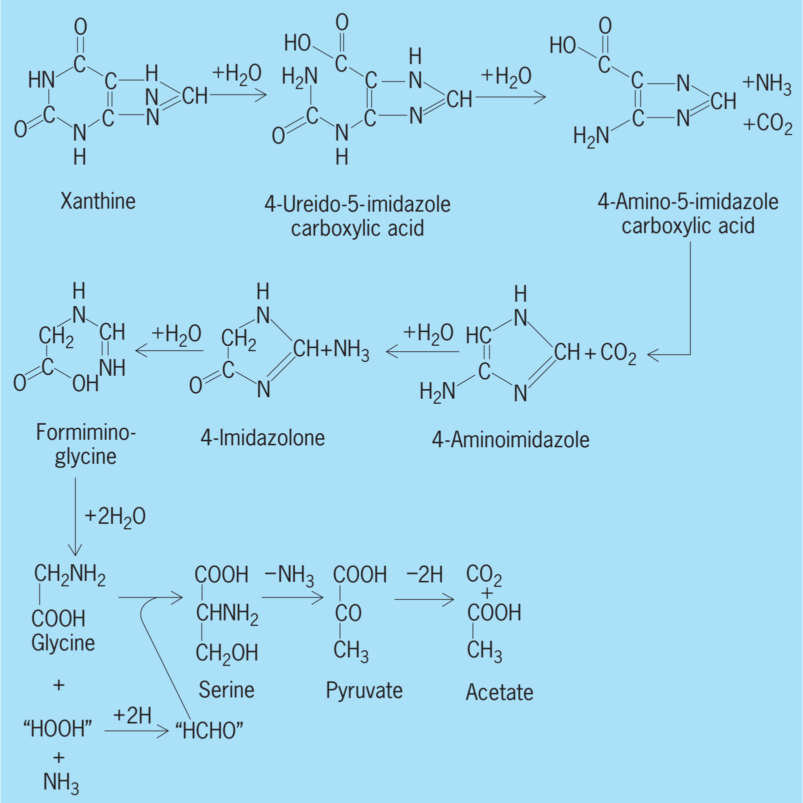
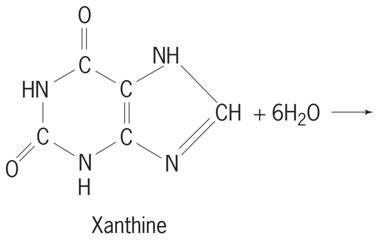


Fig. 16. Steps in the fermentation of the purine xanthine by *Clostridium acidi-urici*.

Purine fermentation. Purine fermentations are caused by a few *Clostridium* and *Micrococcus* species. *Clostridium acidi-urici*, for example, rapidly ferments uric acid, xanthine, and certain other purines to ammonia, carbon dioxide, acetate, and very small amounts of glycine and formate. The fermentation of xanthine is approximately described by reaction (81). Other purines are converted to xanthine be-



$$\mathrm{4NH_3 + 3CO_2 + CH_3COOH} \qquad (81)$$

fore the ring system is ruptured. Chemical steps in xanthine fermentation are shown in **Fig. 16**.

The fermentation involves destruction of the six-membered ring, followed by conversion of the imidazolone ring to the straight-chain compound form-iminoglycine. The latter is converted to glycine, ammonia, and activated formic acid, actually a derivative of tetrahydrofolic acid. The activated formic acid is reduced to the formaldehyde level of oxidation and condensed with glycine to form serine. This amino acid is deaminated and rearranged to form pyruvate, which is oxidized to acetate and carbon dioxide. High-energy phosphate is probably formed during the oxidation of pyruvate and stored as ATP for use in biosynthetic reactions. Horace A. Barker

Bacterial Anabolism

Bacterial anabolism comprises the physiological and biochemical activities concerned with the acquisition, synthesis, and organization of the numerous and varied chemical constituents of a bacterial cell. Clearly, when a cell grows and divides to form two cells, there exists twice the amount of cellular components that existed previously. These components are drawn, directly or indirectly, from the environment around the cell, and (usually) modified extensively in the growth processes when new cell material is formed (biosynthesis). This build-up, or synthesis, begins with a relatively small number of low-molecular-weight building blocks which are either assimilated directly from the environment or produced by catabolism. By sequential and interrelated reactions, they are fashioned into different molecules (mostly of high molecular weight, and hence called macromolecules), for example, lipids, polysaccharides, proteins, and nucleic acids, and many of these molecules are in turn arranged into more complex arrays such as ribosomes, membranes, cell walls, and flagella. Other typical anabolic products, of lower molecular weight, include the pigments, vitamins, antibiotics, and coenzymes. The enzymes responsible for the sequential reactions in any one biosynthetic pathway or assembly sequence are often located on or in cellular structures and thus in physical proximity to the preceding and succeeding enzymes, and their products, and to the site(s) where cellular structures are to be formed. Anabolism also includes the transport of molecules into cells, of building blocks to reaction sites, energetic activations, and the transfer and incorporation of the finished products to their ultimate sites in or outside the cell.

Energy relations in metabolism. During the metabolism of a substrate, bacteria perform numerous biochemical transformations whose overall effect is the liberation of energy in the form of heat. This can be illustrated by the spectacular rise in temperature accompanying the microbial decomposition of large masses of plant material(s) which are sufficiently well insulated to prevent a rapid release of heat into the environment, such as in a compost pile. The temperature in such a mass may rise to 80°C (176°F) or more in a matter of hours. Similarly, in the large culture vats in which microbes are grown, for example, for the commercial production of antibiotics, heat production is so intense that the contents must be cooled continuously to keep the temperature within the physiologically tolerable range.

Prior to 1930, it was customary to attribute the energy liberation exclusively to those metabolic processes collectively called catabolism. The anabolic reactions were grouped together in a separate category, because they could all be considered as processes involving the uptake of energy.

Reactions that release free energy and thus can proceed spontaneously are called exergonic. Here the sum of the chemical (free) energy of the reactants is greater than that of the products. Reactions in which the total free energy of the products exceeds that of the reactants are called endergonic. They can occur only if enough energy is supplied to compensate for the difference. If a proper mechanism exists whereby the energy released in an exergonic reaction can be transferred to an endergonic one, the former can drive the latter, and in this case the two processes are said to be energetically coupled.

The sharp distinction between catabolic and anabolic processes in bacterial metabolism could be justified on this basis. For example, the degradation of a fatty acid to the corresponding number of acetic acid molecules is an energy-yielding process. Its opposite, the formation of fatty acids from acetate, which is quite common in bacterial metabolism, must therefore require energy. This kind of argument could also be applied to practically every aspect of anabolism. Anabolism is an immensely complex phenomenon. This is clear from the fact that many bacteria can live and multiply in a simple inorganic medium supplied with but a single organic substrate, such as acetate. This implies that these organisms can synthesize all of their cell constituents—carbohydrates, fats, proteins, enzymes, nucleic acids, and so on—from the substrate. So bewildering are the number and diversity of the components of a bacterial cell that it seemed hopeless to attempt an explanation of the manner in which the necessary syntheses are accomplished. The best that could be accomplished was to recognize the principle of energetic coupling, which could be expressed in the form of two general reactions (82*a*) and (82*b*).

$$\text{Substrate} \xrightarrow{\text{catabolism}} \text{breakdown products + energy} \qquad (82a)$$

$$\text{Substrate + energy} \xrightarrow{\text{anabolism}} \text{products of synthesis} \qquad (82b)$$

These equations show that catabolic reactions may occur without concomitant anabolic ones, while the reverse is obviously impossible; anabolism is dependent on catabolism. After 1930, however, when a sound beginning had been made with a more detailed analysis of catabolic processes, it gradually became clear that anabolic reactions might be amenable to this same approach, and after 1950 enormous strides were made in understanding the mechanism by which they proceed. It is now universally accepted that anabolic processes can also be interpreted as being composed of series of simple, consecutive-step reactions, and that each of these can be represented by reaction (83), which is fundamentally similar to the one given earlier as a general expression of catabolic step reactions, reaction (3).

$$H_2A + B \longrightarrow A + H_2B \qquad (83)$$

This significant change in outlook, which led to a progressively refined understanding of the mechanism of biosynthetic reactions, began with the recognition that certain step reactions in such a process need not be regarded as endergonic. The seeming contradiction of this statement with the argument presented earlier can be readily explained. It was learned that the actual reactants in a biosynthetic reaction, the building blocks, are not the simple substances like the substrate molecules previously considered, but rather are derivatives of the latter, in which they occur combined with a reactive moiety. As a result of this combination, they become more reactive; this can also be expressed by saying that the substrate molecules have thus been lifted to a higher energy level. Consequently they can then participate in spontaneous reactions for which the unchanged substrate molecules are unfit.

Such reactive derivatives can be formed in either of two ways.

1. They may originate directly as intermediate products during a breakdown process. This happens if a substrate remains tied to the enzyme or coenzyme that mediates its decomposition. In that event the catabolic product first emerges as a coenzyme-bound substance.

2. They may also be formed by a reaction in which a general energy-storage product, such as ATP, is used to transfer an energy-rich (reactive) fragment to a substrate molecule and so to raise the reaction potentials of the latter.

The example of the degradation and synthesis of fatty acids can again serve to illustrate these two aspects. In the section on fermentative metabolism, it has been mentioned that the breakdown proceeds under the influence of enzymes in association with coenzyme A. The fatty acid first becomes attached to the latter as an acyl-CoA compound from which, after dehydrogenation, two-carbon fragments are split off as acetyl-CoA. This substance is far more energy-rich, and hence reactive, than acetate itself, and two molecules can readily be transformed without additional energy supply, to acetoacetyl-CoA. On the other hand, acetoacetic acid cannot be so formed from acetate; this is an endergonic process. Thus reaction (84) illustrates the direct catabolic formation

$$\text{Fatty acid} - \text{CoA complex} \xrightarrow[\text{genation}]{\text{dehydro-}} \text{acetyl-CoA} \qquad (84)$$

of a reactive entity that can function as a reactant in a synthesis that does not require an input of energy.

The second type of mechanism is illustrated by the spontaneous, enzyme-mediated reaction between acetate and ATP which yields acetyl phosphate and ADP. Acetyl phosphate, as a mixed anhydride comparable to acetyl chloride, is an unstable substance on the same energy level as acetyl-CoA, and its acetyl group can be enzymatically transferred, without extra energy supply, to CoA, yielding the genuine building block for the synthesis of the long-chain fatty acids by way of reactions (85*a*) and (85*b*). The

$$\text{Acetate} + \text{ATP} \longrightarrow \text{acetyl phosphate} + \text{ADP} \qquad (85a)$$

$$\text{Acetyl phosphate} + \text{CoA} \longrightarrow \text{acetyl-CoA} + \text{phosphate} \qquad (85b)$$

formation of acetyl phosphate from acetate represents an activation of the latter; it is one instance of many similar processes by which a relatively stable, that is, unreactive, substance can be converted into an unstable, reactive one. Because of this activation, the substance can then participate in a spontaneously occurring synthetic process; the unactivated one cannot.

This is the fundamental principle of the mechanism by which the energy-liberating catabolic processes can be coupled to the energy-requiring anabolic reactions.

Role of phosphorylated compounds. The above paragraphs have shown that ATP, by transfer of one of its terminal phosphate groups to another substance, can be used to raise the energy level of the latter. Such reactions are quite common in the metabolism of bacteria, as well as in that of other microbes, plants, and animals. In fact, most phosphorylations are accomplished in this manner, and in earlier parts of this article many examples have been given. This kind of reaction is possible because ATP itself is an unstable substance which contains two energy-rich (commonly called high-energy) phosphate bonds. Upon hydrolytic cleavage, each of these can release an amount of energy of the order of magnitude of 8000–10,000 cal/mole. If such a phosphate group is transferred to a metabolically unreactive molecule, the latter acquires some or all of the bond energy, which is thus not lost as heat. This fundamental notion of the preservation of bond energy was first formulated by F. Lipmann in 1939.

However, not all phosphorylations at the expense of ATP are equally effective in preserving bond energy. In addition to the energy-rich compounds, of which acetyl phosphate is an example, there are other types of phosphorylated substances in which the energy of the phosphate bond is only of the order of magnitude of 3000 cal/mole, as shown by the amount of heat liberated during their hydrolysis. These are the phosphate esters, such as glycerophosphate, glucose-6-phosphate, and fructose-1,6-diphosphate, all of which can be formed by transfer of a phosphate group from ATP according to reaction (86). The transfer of phosphate groups from

$$\text{R} - \text{CH}_2\text{OH} + \text{ATP} \xrightarrow[\text{kinase}]{\text{specific}} \text{R} - \text{CH}_2\text{O} - \text{PO}_3\text{H}_2 + \text{ADP} \qquad (86)$$

ATP to substrate molecules is mediated by enzymes that are called kinases. They are specific for a particular substrate, and are correspondingly named glucokinase (or hexokinase), 6-phosphofructokinase, and so on.

Because the phosphate-bond energy in the esters is far less than that in ATP, the esterification reactions involve the loss of a large proportion of the chemical energy as heat. Nevertheless, the esterifications are metabolically significant, for they also yield compounds with potentialities greater than those possessed by nonphosphorylated substances. These phosphate esters are used in polysaccharide biosynthesis.

In the activation of substrate molecules by transphosphorylation from ATP the latter is consumed, and it will be evident that in a metabolizing cell the stock of this material must be continuously replenished. This is accomplished by the catabolic formation of high-energy phosphate bonds, including ATP itself. It may occur in a number of different ways as discussed earlier in the section on bacterial catabolism.

Adenosine phosphates. Three adenosine phosphates, which are derivatives of the glucoside adenine riboside or adenosine, are known. They are adenosine monophosphate (AMP), also referred to as adenylic acid, ADP, and ATP. The structures and relations of these substances are shown in **Fig. 17**. It will be seen that the phosphate group of AMP is attached to adenosine by an ester (low-energy) linkage, while the additional phosphate groups of ADP and ATP represent anhydride (high-energy) bonds. Consequently ADP and ATP, but not AMP, can act as high-energy phosphate donors, and AMP and ADP are both high-energy phosphate acceptors.

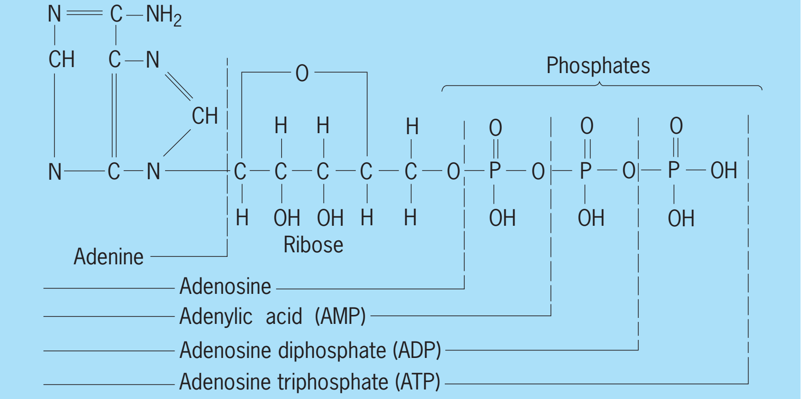


Fig. 17. Relationship of adenosine and its three adenosine phosphate derivatives.

Although ADP and, particularly, ATP are undoubtedly the most common high-energy phosphate donors in bacterial metabolism, there are others, such as the previously mentioned acetyl phosphate, phosphoenolpyruvic acid, and 1,3-diphosphoglyceric acid. The last two are known only as intermediate metabolic products, while ADP and ATP generally are, and acetyl phosphate occasionally may be, high-energy phosphate storage products as well. This also applies to guanosine and other di- and triphosphates (GDP, GTP) which play a role in the Krebs or tricarboxylic acid cycle.

Apart from being a high-energy phosphate donor, ATP can also participate in anabolic processes in another way. It can transfer the adenyl (AMP) moiety to a substrate molecule, with the formation of an adenylate and pyrophosphate (PP). The andenylates, too, represent activated substrates and are, as will be seen later, the genuine reactants in certain biosynthetic processes. The significance of such adenylations has been sought in the concomitant formation of pyrophosphate by the enzyme pyrophosphatase. In a situation where the equilibrium of the adenylase reaction is highly in favor of the starting materials (indicated by $\overset{\rightarrow}{\longleftarrow}$) and thus where adenylate would be formed in only extremely low concentration, the removal of the pyrophosphate in a subsequent reaction would cause a shift in the equilibrium, with a larger concentration of adenylate being formed, as shown in reactions (87*a*) and (87*b*).

$$\text{Substrate} + \text{ATP} \ \underset{\longleftarrow}{\overset{\text{adenyltrans-ferase}}{\rightarrow}} \ \text{substrate-adenylate} + \text{PP} \qquad (87a)$$

$$\text{PP} + H_2O \xrightarrow{\text{pyrophosphatase}} \text{orthophosphate} \qquad (87b)$$

Energy-rich coenzyme-substrate. The adenylation reactions, mediated by enzymes known as adenylyltransferases, are a special instance of a more general type of substrate activation which yields an energy-rich substrate-enzyme complex. Such complexes may be formed in one of two ways.

1. They may arise directly from the metabolic conversion of a substrate-coenzyme compound that does not initially contain an energy-rich bond. This matter of origin is illustrated by the formation of the coenzyme complex of 3-phosphoglyceric acid, with its high-energy bond, with the addition product of 3-phosphoglyceraldehyde and the sulfhydryl-containing coenzyme by dehydrogenation.

2. Alternatively, they may result from the transfer of a moiety from an already activated to a nonactivated substance. Examples of this mode of formation are the above-mentioned transadenylations, and the transacetylation by which the acetyl group of acetyl phosphate is transferred to coenzyme A with the formation of acetyl-CoA and orthophosphate. The former represent cases where the energy-rich bond initially resides in the coenzyme, ATP, rather than in the substrate; the second exemplifies the reverse.

In all such reactions the principle of the preservation of bond energy is manifest. The same principle is further illustrated by those processes in which a metabolic waste product originates in a coenzyme-bound form with an energy-rich bond. If the waste product were split off from the complex by hydrolysis, the bond energy would be lost as heat. Generally, it is not so released, but rather through transfer of the coenzyme moiety to a substrate molecule which thereby becomes activated.

Catabolism and anabolism. The metabolic significance of the energy-rich coenzyme-substrate complexes is that they represent the genuine building blocks that take part in the spontaneous biosynthetic reactions. As shown in the preceding sections, these complexes are formed, directly or indirectly, as a result of processes by which substrates are degraded. Hence it now becomes possible to present a general picture of the connection between the catabolic and anabolic aspects of metabolism that is much more revealing than the earlier invoked concept of energetic coupling. However useful this concept may have been at a time when nothing was known about the mechanism of energy transfer, it was unsatisfactory precisely because it avoided this major problem.

Now that the general principles of that mechanism have been elucidated, the notion of energetic coupling has lost its element of mystery. At the same time, it has become evident that the two aspects of metabolism that had formerly been sharply distinguished need no longer be viewed in this manner. On the contrary, the connection between them can now be interpreted far more adequately by considering the breakdown of a substrate as the means by which the genuine building blocks for the synthetic processes are manufactured. This can be expressed by notation (88), which emphasizes the common origin of waste products and of products of synthesis from the same intermediate products rather than the energy relations.

$$\begin{tikzcd}
\text{Substrate} \arrow[dr] & & \text{waste products} \\
& \text{intermediate products} \arrow[ur] \arrow[dr] & \\
& & \text{products of synthesis}
\end{tikzcd} \qquad (88)$$

Essential for these advances in the understanding of metabolic processes has been the recognition that many of the enzyme-controlled biochemical reactions are reversible, and that the same intermediate products, required as building blocks for the synthesis of different classes of compounds, are produced during the metabolism of a wide variety of substrates. This has been amply documented in the sections on oxidative and fermentative catabolism of this article.

The reversibility of enzymatic reactions calls for an additional comment. The equilibrium of such processes is determined by the free-energy changes involved. Thus, if for a reaction of the type shown in reaction (89) the total free energy of A and B is greater

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C} + \mathrm{D} \qquad (89)$$

than that of C and D, the equilibrium will be in favor of the latter. The greater the difference, the farther will the reaction proceed to the right. But if this system were combined with one representing an even greater free-energy change, such as reaction (90), the

$$\mathrm{A} + \mathrm{E} \rightleftharpoons \mathrm{F} + \mathrm{G} \qquad (90)$$

removal of A by this reaction may bring about an increased formation of B from C and D. This situation has been touched upon in discussing the adenylation of substrates by reaction (91). There it was sug-

$$\text{Substrate} + \text{ATP} \rightleftharpoons \text{substrate-adenylphosphatelate} + \text{PP} \qquad (91)$$

gested that the subsequent hydrolysis of pyrophosphate to orthophosphate by pyrophosphatase could drive the adenylase reaction to the right. The reason is that pyrophosphate contains an anhydride (that is, high-energy) bond whose energy is released as heat during hydrolysis. This makes the pyrophosphatase reaction for all practical purposes irreversible, so that its inclusion in a complex enzymatic processes can be an effective method of energetic coupling by a simple mechanism.

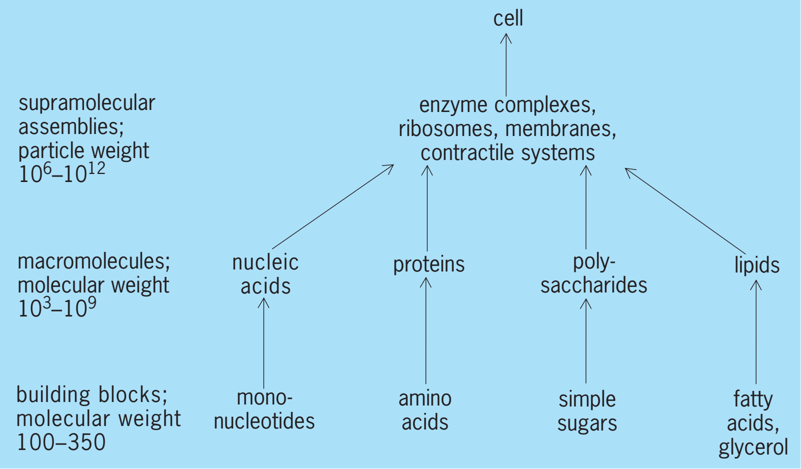

cell
supramolecular assemblies; particle weight 10^6–10^{12}
enzyme complexes, ribosomes, membranes, contractile systems
macromolecules; molecular weight 10^3–10^9
nucleic acids
proteins
poly-saccharides
lipids
building blocks; molecular weight 100–350
mono-nucleotides
amino acids
simple sugars
fatty acids, glycerol

Fig. 18. Levels of organization in bacteria. (*After W. B. Wood, The Molecular Basis of Metabolism* [Unit 3, *Biocore*], *McGraw-Hill, 1974*)

Oxidative and fermentative assimilation. In a growing bacterial culture the number and diversity of synthetic processes comprising the formation of all cell constituents are so vast that a detailed analysis of the mechanism by which any one of them is formed is virtually impossible. This requires a simpler situation, which can be realized by using suspensions of bacteria in a medium in which they cannot multiply because it lacks some ingredient, such as a nitrogen source, necessary for the synthesis of a quantitively important cell material. If a small amount of an oxidizable or fermentable substrate is added to such a suspension, it will be quickly and completely metabolized. At this point the quantity of carbon that disappeared as substrate is, however, much greater than that recoverable as CO_2 or fermentation products, and the missing fraction is found as reserve material in the cells; it may amount to 20–80% of the substrate carbon. Thus the metabolism entails a partial conversion of the substrate into assimilation products; this is referred to an oxidative or fermentative assimilation, depending upon whether it occurred during an oxidative or fermentative decomposition of the substrate.

In the presence of substances such as 2,4-dinitrophenol or sodium azide, which prevent the metabolic generation of ATP, the substrate is entirely converted into metabolic waste products. The absence of assimilation under these conditions indicates that the assimilatory processes involve transphosphorylations. This conclusion is entirely consistent with current knowledge of biosynthetic mechanisms.

The nature of the assimilation product depends on the organism as well as on the substrate. For example, *Pseudomonas saccharophila* produces mainly poly-β-hydroxybutyric acid, a fatlike reserve material, from acetate, butyrate, and glucose; *Rhodospirillum rubrum* also forms this polymer from acetate and butyrate, but forms a glycogenlike polysaccharide when metabolizing propionate, pyruvate, succinate, or malate.

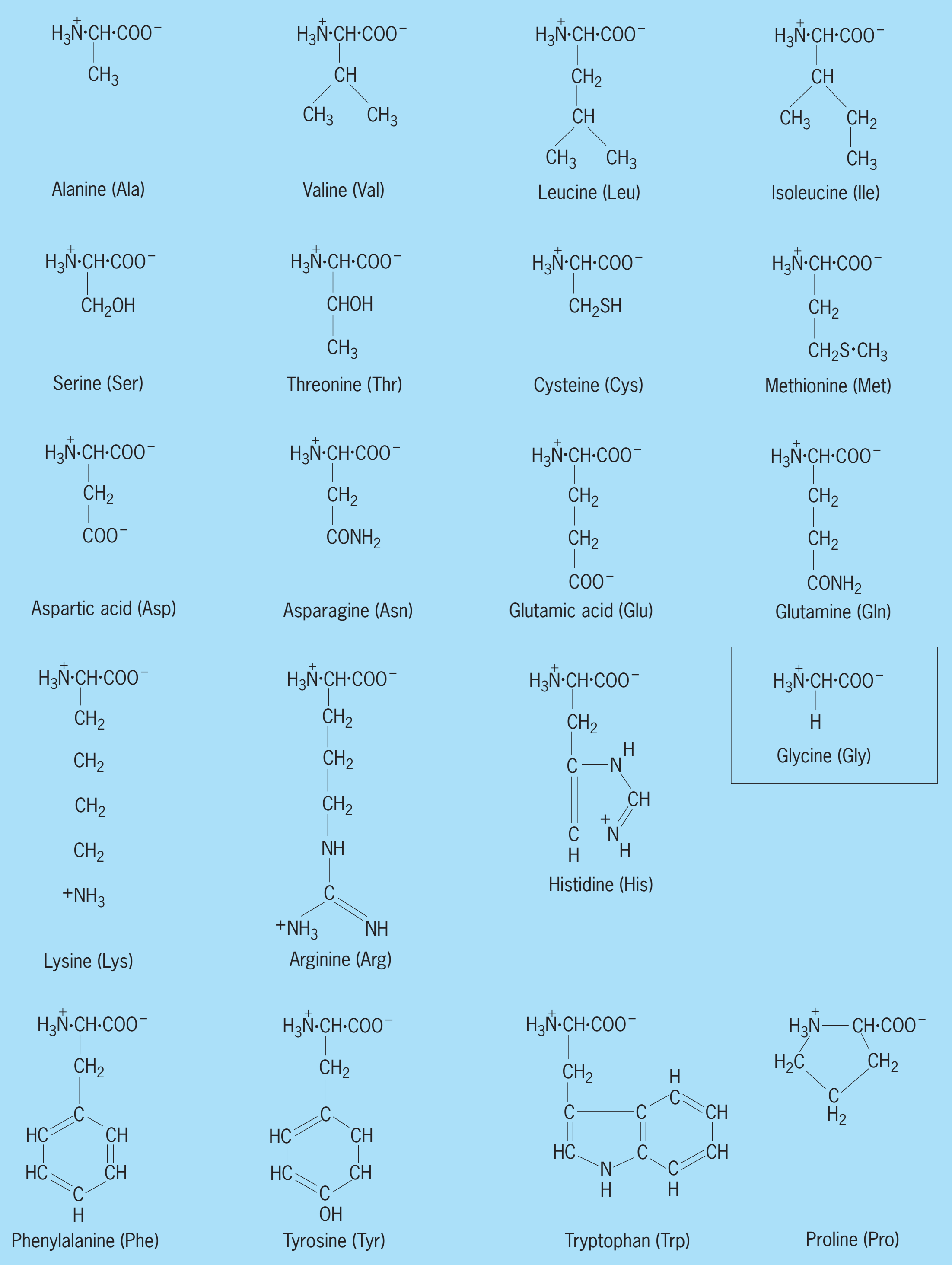

Alanine (Ala)
Valine (Val)
Leucine (Leu)
Isoleucine (Ile)
Serine (Ser)
Threonine (Thr)
Cysteine (Cys)
Methionine (Met)
Aspartic acid (Asp)
Asparagine (Asn)
Glutamic acid (Glu)
Glutamine (Gln)
Lysine (Lys)
Arginine (Arg)
Histidine (His)
Glycine (Gly)
Phenylalanine (Phe)
Tyrosine (Tyr)
Tryptophan (Trp)
Proline (Pro)

Fig. 19. The 20 amino acids commonly occurring in proteins. All the amino acids have the l configuration when they occur in proteins, except glycine which is not optically active. (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed., Halsted Press* [*John Wiley*], *1973*)

Group transfer. In the preceding sections several examples have been given of reactions in which smaller or larger groups of atoms are transferred from one molecule to another. In this manner special building blocks, arising as intermediate products during the metabolism of a particular substrate, can be used in biosynthetic processes for the modification of other compounds, and thus aid in the eventual synthesis of cellular constituents, many of which bear little or no resemblance to the substrates from which they can be manufactured.

The groups in question range from a pair of hydrogen atoms, used for the reduction of some intermediate product and generally donated by reduced pyridine nucleotides, to sulfhydryl, sulfate, amino, methyl-carbamyl, formyl, acetyl, succinyl, glucosyl,

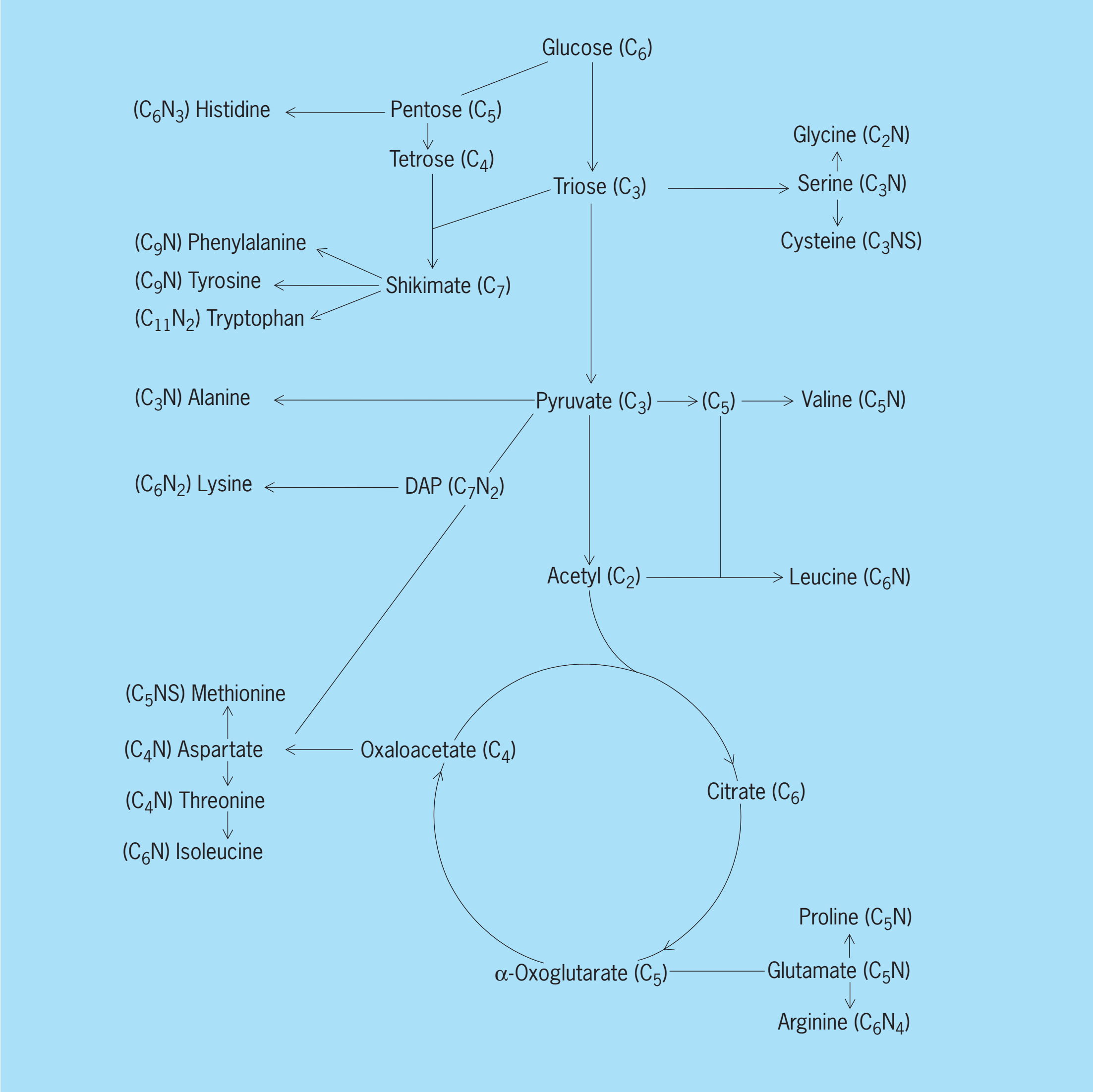

Glucose (C_6)
(C_6N_3) Histidine
Pentose (C_5)
Tetrose (C_4)
Triose (C_3)
Glycine (C_2N)
Serine (C_3N)
Cysteine (C_3NS)
(C_9N) Phenylalanine
(C_9N) Tyrosine
($C_{11}N_2$) Tryptophan
Shikimate (C_7)
(C_3N) Alanine
Pyruvate (C_3)
(C_5)
Valine (C_5N)
(C_6N_2) Lysine
DAP (C_7N_2)
Acetyl (C_2)
Leucine (C_6N)
(C_5NS) Methionine
(C_4N) Aspartate
Oxaloacetate (C_4)
(C_4N) Threonine
(C_6N) Isoleucine
Citrate (C_6)
Proline (C_5N)
α-Oxoglutarate (C_5)
Glutamate (C_5N)
Arginine (C_6N_4)

Fig. 20. Synthesis of amino acids. The carbon is supplied by glucose, the nitrogen by NH_4^+, and the sulfur by SO_4^{2-} (diaminopimelate is represented by DAP). (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed., Halsted Press* [*John Wiley*], *1973*)

fructosyl, and even entire amino acid groups.

Cellular biosyntheses. The following account of some specific biosynthetic processes contains several examples of group transfer.

Of what is a bacterial cell composed? In addition to water, which accounts for about 70% of the cell's total weight, there are present over 3000 different protein molecules (which account for some 15% of

TABLE 6. Some building blocks and macromolecules of bacterial cells

Building blocks	Macromolecules
Amino acids	Proteins
Nucleotides	Nucleic acids
Monosaccharides	Polysaccharides
Fatty acids, glycerol	Lipids
Amino sugars, amino acids	Peptidoglycan
Lipids, monosaccharides, amino sugars, glycerol or monosaccharides	Lipopolysaccharide

TABLE 7. The amino acid families

Precursor	Family name	Member acids
Oxaloacetic acid[*] or fumaric acid[*]	Aspartate	Asparagine, lysine, threonine, methionine isolecuine, aspartic
α-Ketoglutaric acid[*]	Glutamate	Glutamine, proline, arginine, glutamic
3-Phosphoglyceric acid[†]	Serine	Glycine, serine, cysteine
Pyruvic acid[†]	Pyruvate	Alanine, valine, leucine
Erythrose phosphate[‡] Phosphoenolpyruvate[†]	Aromatic	Phenylalanine, tyrosine, tryptophan

[*] An intermediate of the Krebs (tricarboxylic acid) cycle.

[†] An intermediate in the Embden-Meyerhof-Parnas pathway for carbohydrate degradation.

[‡] An intermediate in the hexose monophosphate pathway for carbohydrate degradation.

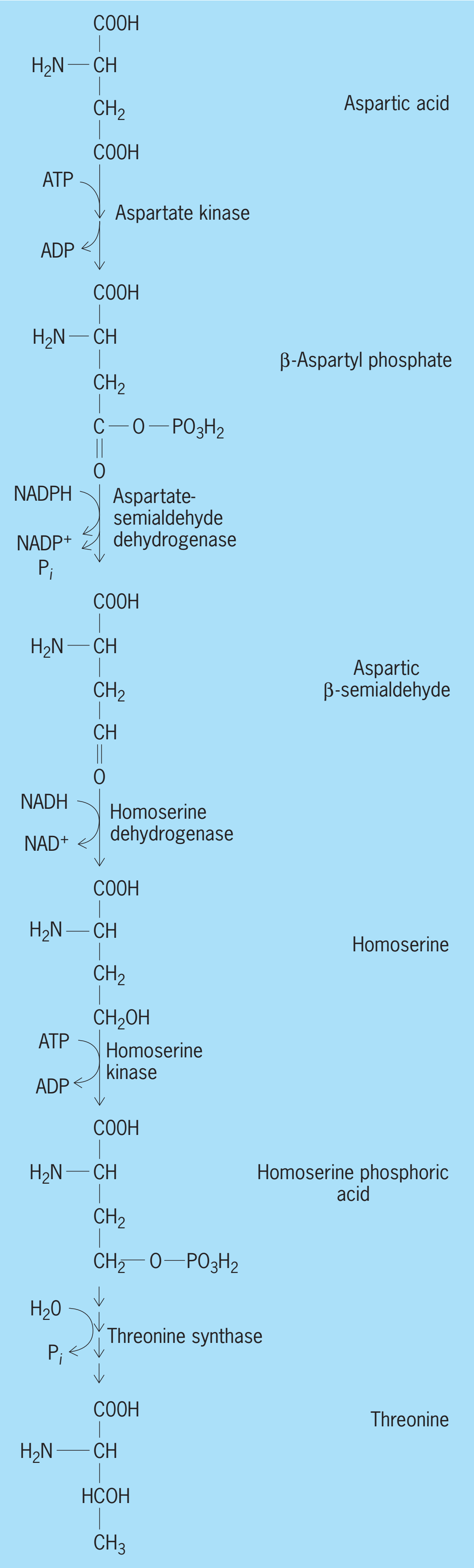
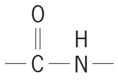

COOH
H2N—CH
CH2
COOH
Aspartic acid
ATP
Aspartate kinase
ADP
COOH
H2N—CH
CH2
C—O—PO3H2
O
β-Aspartyl phosphate
NADPH
Aspartate-semialdehyde dehydrogenase
NADP+
Pi
COOH
H2N—CH
CH2
CH
O
Aspartic β-semialdehyde
NADH
Homoserine dehydrogenase
NAD+
COOH
H2N—CH
CH2
CH2OH
Homoserine
ATP
Homoserine kinase
ADP
COOH
H2N—CH
CH2
CH2—O—PO3H2
Homoserine phosphoric acid
H2O
Threonine synthase
Pi
COOH
H2N—CH
HCOH
CH3
Threonine

Fig. 21. Biosynthesis of threonine from aspartate. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

the total weight), deoxyribonucleic acid (one or two kinds, about 1% of the total weight), ribonucleic acid (some 1000 different molecular species, totaling about 6% of the weight), carbohydrates (perhaps 50 different kinds, 3% of cell weight), lipids (about 50 different kinds, 2% of total cell weight), the building blocks or molecules of intermediary metabolism (500 kinds, 2% of total cell weight), and inorganic ions (about 12 kinds, 1% of cell weight).

Thus the task facing the colon bacterium *Escherichia coli* or the soil inhabitant *Pseudomonas putida* is to utilize the carbon and energy inherent in, for example, acetate



$$-\overset{\overset{\displaystyle O}{\|}}{C}-\overset{H}{N}-$$

so as to form the diverse array of biological

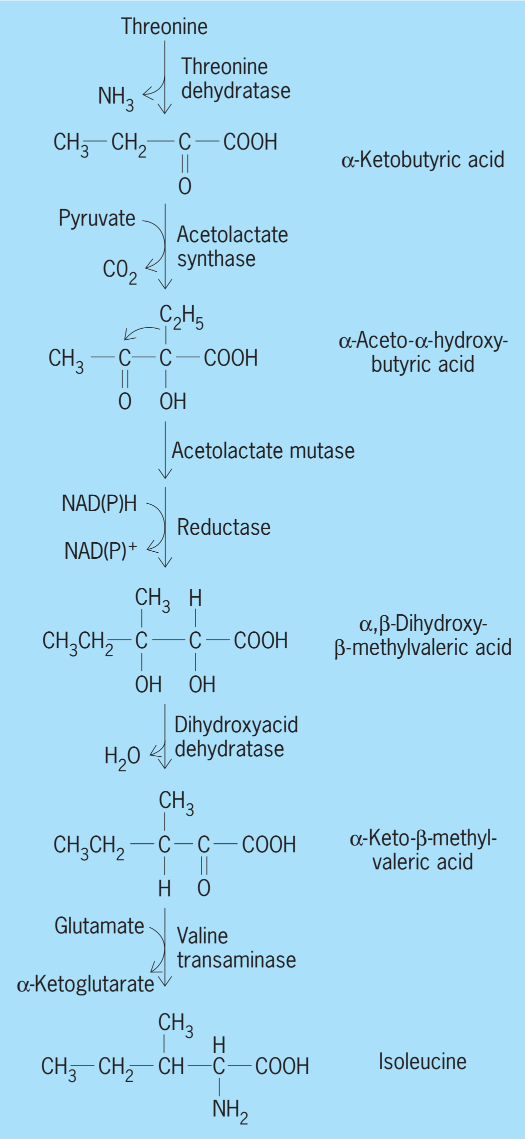
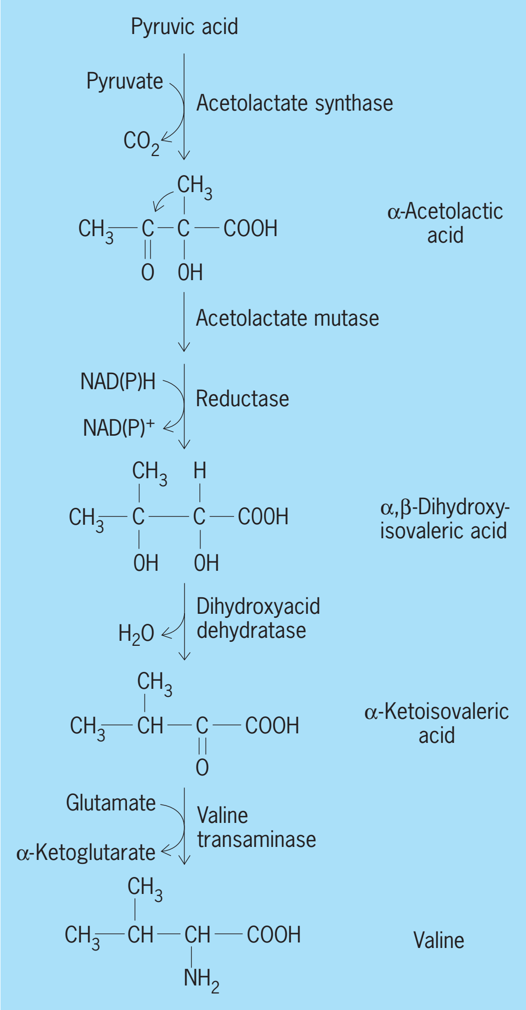

Threonine
Threonine dehydratase
NH3
CH3—CH2—C—COOH
O
α-Ketobutyric acid
Pyruvate
Acetolactate synthase
CO2
C2H5
CH3—C—C—COOH
O OH
α-Aceto-α-hydroxy-butyric acid
Acetolactate mutase
NAD(P)H
Reductase
NAD(P)+
CH3 H
CH3CH2—C—C—COOH
OH OH
α,β-Dihydroxy-β-methylvaleric acid
Dihydroxyacid dehydratase
H2O
CH3
CH3CH2—C—C—COOH
H O
α-Keto-β-methyl-valeric acid
Glutamate
Valine transaminase
α-Ketoglutarate
CH3
H
CH3—CH2—CH—C—COOH
NH2
Isoleucine

Fig. 22. Pathway for biosynthesis of isoleucine from threonine. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)



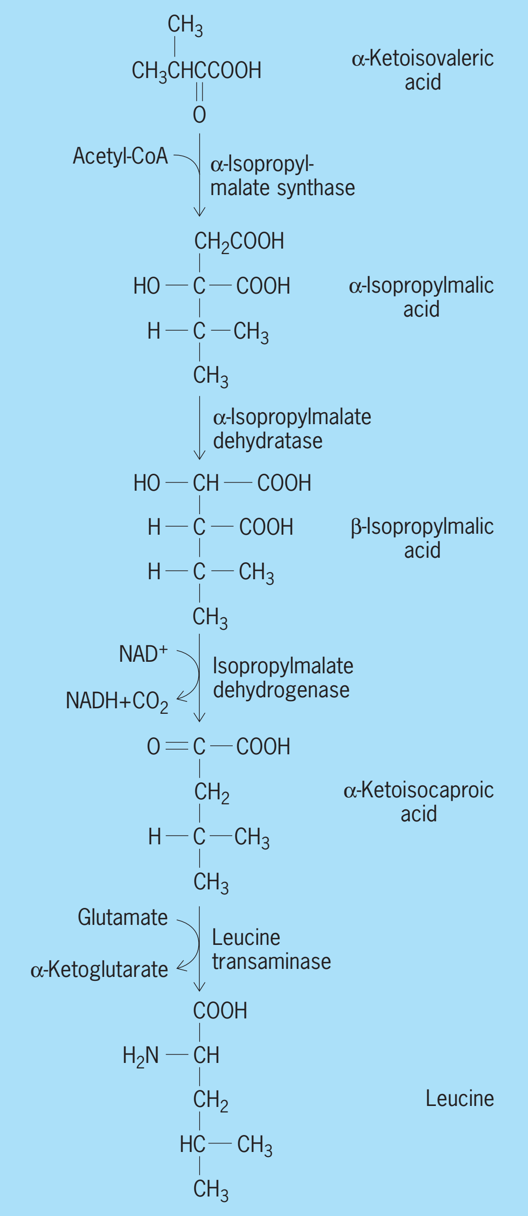
Pyruvic acid
Pyruvate
CO_2
Acetolactate synthase
α-Acetolactic acid
Acetolactate mutase
NAD(P)H
$NAD(P)^+$
Reductase
α,β-Dihydroxy-isovaleric acid
H_2O
Dihydroxyacid dehydratase
α-Ketoisovaleric acid
Glutamate
α-Ketoglutarate
Valine transaminase
Valine

Fig. 23. Pathways to valine and isoleucine. (*After A. Lehringer, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

macromolecules which are present in a cell, and which must be duplicated in order for the cell to divide. In an environment with a sufficient supply of water and acetate, ammonium, phosphate, sulfate, and other ions, the growing cell assimilates these materials and fashions them into the building blocks (**Table 6**) from which the macromolecules, the supramolecular assemblies, and, ultimately, the cell is formed (**Fig. 18**). The energy required for these metabolic conversions is made available by catabolic reactions, and used in the biosynthetic, or anabolic, reactions. Of the latter, it is useful to speak of two main types: one concerned with the conversion of substrate molecules, or molecules derived from the substrate, into the building blocks and other small molecules (vitamins, pigments, coenzymes); and a second type, involved with conversion of the small (low-molecular-weight) molecules into macromolecule and molecular aggregates. All classes or reactions—those of catabolism, as well as those of anabolism—are, of course, specifically catalyzed by a highly efficient class of proteins called enzymes.

Amino acids. The proteins of an organism contain over 20 different amino acids, some of which are acidic, others basic, others essentially neutral, while others differ by virtue of the aromatic rings which are part of their structure (**Fig. 19**). The individual amino acids are formed by different biosynthetic routes, or pathways, as a result of the sequential action(s) of many different enzymes. However, because many of the carbon atoms in the different amino acids are derived from common precursor molecules, it is possible to group the amino acids into families so as to reflect their common biosynthetic origins (**Table 7**); histidine is not in this grouping, primarily because its biosynthesis is more akin to that of purines and pyrimidines than to that of any other amino acid.

A detailed examination of the individual, sequential reactions involved in the synthesis (**Fig. 20**) of each amino acid is beyond the scope of this section,



α-Ketoisovaleric acid
Acetyl-CoA
α-Isopropyl-malate synthase
α-Isopropylmalic acid
α-Isopropylmalate dehydratase
β-Isopropylmalic acid
NAD^+
$NADH+CO_2$
Isopropylmalate dehydrogenase
α-Ketoisocaproic acid
Glutamate
α-Ketoglutarate
Leucine transaminase
Leucine

Fig. 24. Biosynthetic pathway to leucine. (*After A. Lehringer, Biochemistry, 2d ed., Worth Publishing Co. 1975*)

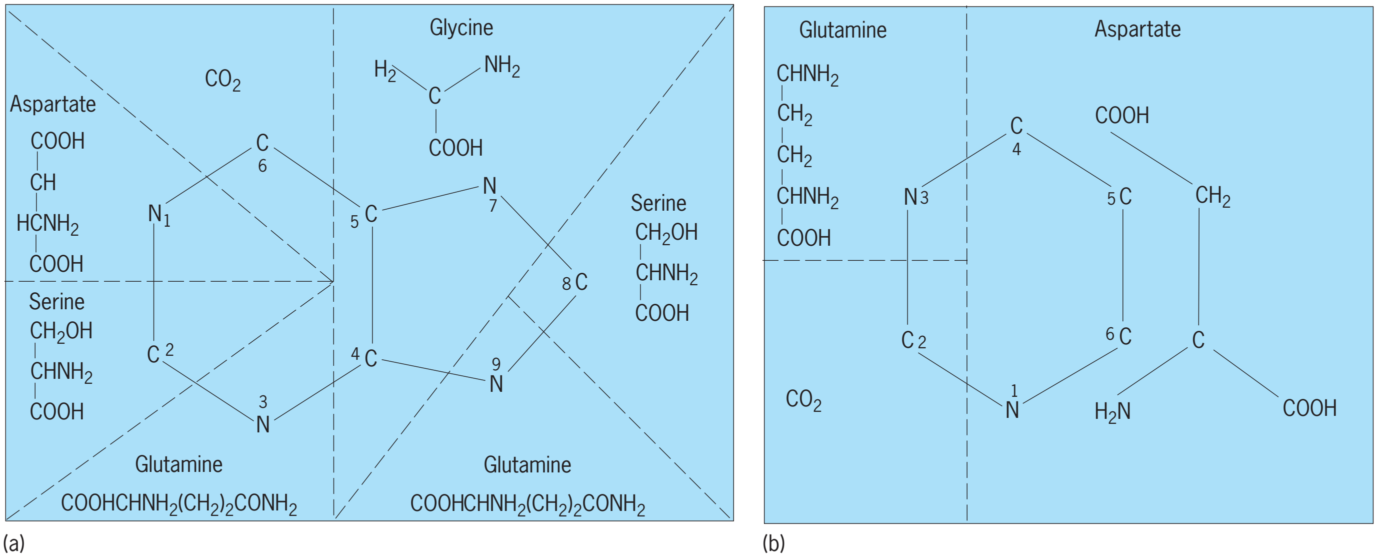

Glycine
H_2
NH_2
C
COOH
CO_2
Aspartate
COOH
CH
$HCNH_2$
COOH
Serine
CH_2OH
$CHNH_2$
COOH
N_1
C 6
5 C
N 7
8 C
C 2
4 C
9 N
3 N
Serine
CH_2OH
$CHNH_2$
COOH
Glutamine
$COOHCHNH_2(CH_2)_2CONH_2$
Glutamine
$COOHCHNH_2(CH_2)_2CONH_2$
(a)
Glutamine
$CHNH_2$
CH_2
CH_2
$CHNH_2$
COOH
Aspartate
COOH
C 4
N3
5 C
CH_2
C 2
6 C
C
CO_2
1 N
H_2N
COOH
(b)

Fig. 25. Origins of the atoms of (*a*) purine rings and (*b*) pyrimidine rings. (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed., Halsted Press* [*John Wiley*], *1973*)

but several principles are especially relevant: (1) the energy currency of the cell, ATP molecules, or the so-called high-energy phosphate compounds such as phosphoenolpyruvate are used often in the stepwise synthesis of many amino acids, as are also reductants (electron donors) such as NADH or NADPH which, like the high-energy phosphate compounds, are generated as a result of catabolic activities; (2) several of the amino acids are synthesized in the form of alpha-keto acids, and only in the final enzymatic step in a pathway is the amino (—NH_2) group introduced into the molecule by a set of reactions termed transamination. *See* AMINO ACIDS.

In transamination an amino group of one amino acid (the amino donor) is transferred to an alpha-keto acid (the amino receptor). This receptor thus becomes an amino acid, and the donor is transformed into an alpha-keto acid. The amino donor is, quite

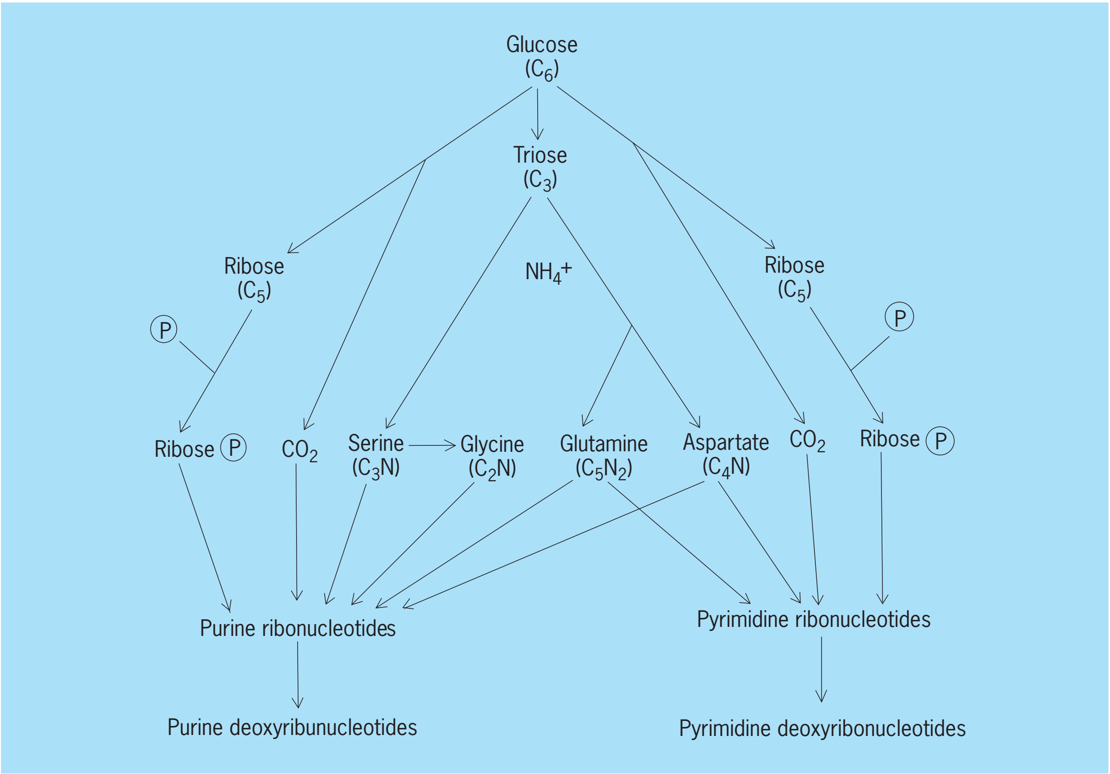

Glucose
(C_6)
Triose
(C_3)
Ribose
(C_5)
NH_4^+
Ribose
(C_5)
(P)
(P)
Ribose (P)
CO_2
Serine
(C_3N)
Glycine
(C_2N)
Glutamine
(C_5N_2)
Aspartate
(C_4N)
CO_2
Ribose (P)
Purine ribonucleotides
Pyrimidine ribonucleotides
Purine deoxyribunucleotides
Pyrimidine deoxyribonucleotides

Fig. 26. Diagram of the pathways in the synthesis of nucleotides. (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed., Halsted Press* [*John Wiley*], *1973*)

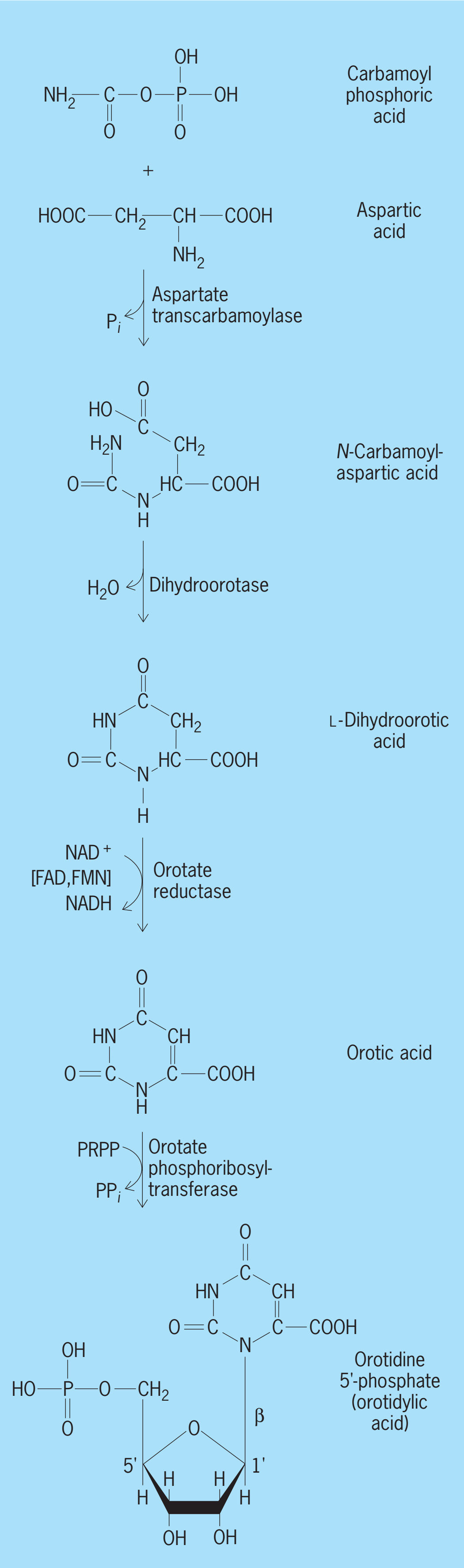

Carbamoyl phosphoric acid
+
Aspartic acid
P_i
Aspartate transcarbamoylase
N-Carbamoyl-aspartic acid
H_2O
Dihydroorotase
L-Dihydroorotic acid
NAD+
[FAD,FMN]
NADH
Orotate reductase
Orotic acid
PRPP
PP_i
Orotate phosphoribosyl-transferase
Orotidine 5'-phosphate (orotidylic acid)
β
5'
1'

Fig. 27. **Biosynthesis of uridylic acid. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)**

often, glutamic acid. This amino acid is synthesized not by a transamination reaction, but by a reductive amination catalyzed by the enzyme glutamic dehydrogenase which has among its substrates the Krebs cycle intermediate, alpha-keto glutaric acid, ammonium ion, and NADH, as in reaction (92).

$$NH_4^+ + \alpha\text{-ketoglutarate} + NADH + H^+ \rightleftharpoons \text{L-glutamate} + NAD^+ + H_2O \qquad (92)$$

Since it is clear that the amino groups of alanine and aspartic acid are also introduced by transamination reactions (where pyruvic acid and oxaloacetic acids are the respective "acceptors") which use glutamic acid as the amino donor, obviously one of the

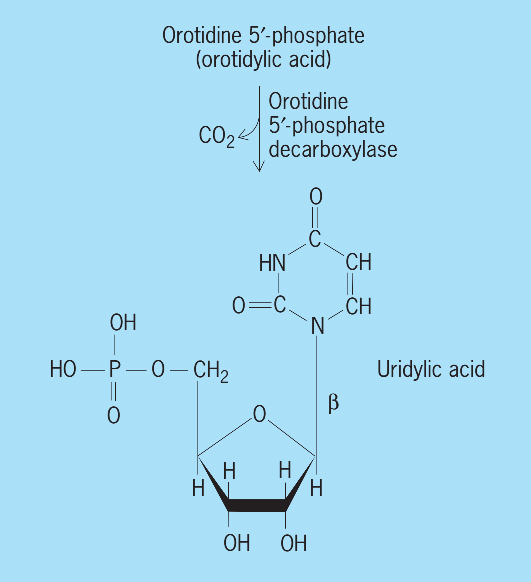

Orotidine 5′-phosphate (orotidylic acid)
CO_2
Orotidine 5′-phosphate decarboxylase
Uridylic acid
β

Fig. 28. Formation of uridylic acid (UMP). (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

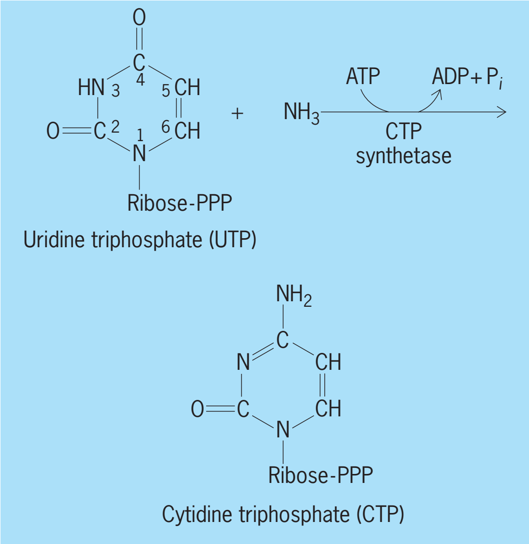

+
NH_3
ATP
ADP+P_i
CTP synthetase
Ribose-PPP
Uridine triphosphate (UTP)
Ribose-PPP
Cytidine triphosphate (CTP)

Fig. 29. Amination of UTP to CTP by CTP synthesis. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

main pathways by which NH_4^+ is converted to organic nitrogen is via the reductive amination of alpha-keto glutaric acid.

Special enzymatic steps are involved in the introduction of the S atom in amino acids such as cystine; cysteine, and methionine, for the introduction of the methyl (CH_3—) group which is part of methionine, and for the introduction of the additional N and C atoms found in arginine.

Selected examples of amino acid biosynthesis will illustrate some key features. **Figure 21** depicts the pathway for biosynthesis of the amino acid threonine from the amino acid aspartate, and **Fig. 22** depicts the pathway by which some threonine molecules so formed undergo further sequential conversion(s) to form isoleucine. The biosynthetic pathway for valine is depicted in **Fig. 23**, while **Fig. 24** indicates that this pathway branches (after synthesis of alpha-keto valeric acid) and thus leads also to the synthesis of leucine. The synthesis of these and other amino acids, as well as other cellular components, is energetically expensive, but the activities of many enzymes are subject to rapid regulation, and thus unnecessary, wasteful syntheses may be avoided.

The amino acids either formed by these biosynthetic sequences or assimilated from the extracellular milieu are used largely for the biosynthesis of the

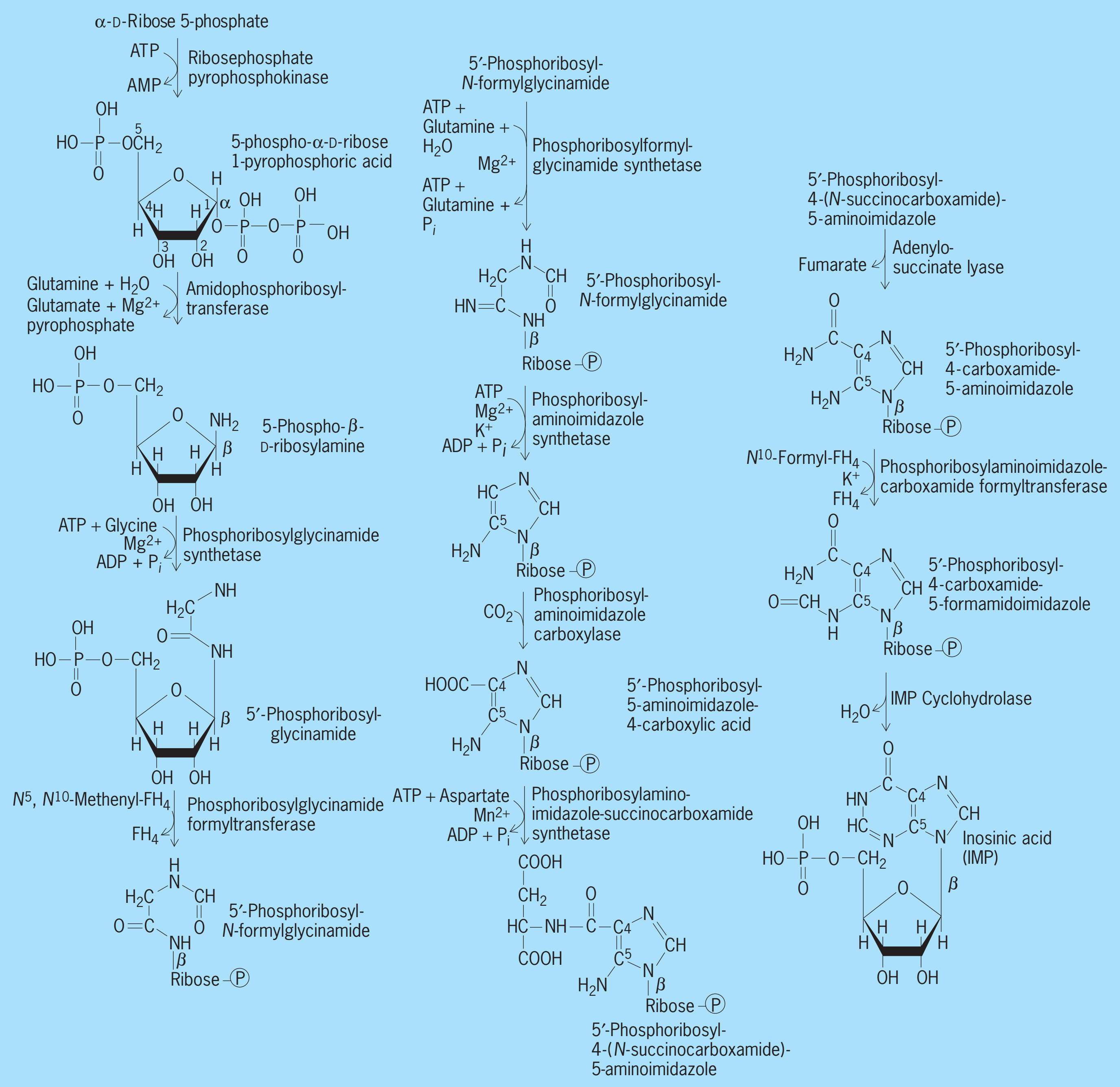

α-D-Ribose 5-phosphate
ATP
AMP
Ribosephosphate pyrophosphokinase
5-phospho-α-D-ribose 1-pyrophosphoric acid
Glutamine + H2O
Glutamate + Mg2+ pyrophosphate
Amidophosphoribosyl-transferase
5-Phospho-β-D-ribosylamine
ATP + Glycine
Mg2+
ADP + Pi
Phosphoribosylglycinamide synthetase
5′-Phosphoribosyl-glycinamide
N5, N10-Methenyl-FH4
FH4
Phosphoribosylglycinamide formyltransferase
5′-Phosphoribosyl-N-formylglycinamide
5′-Phosphoribosyl-N-formylglycinamide
ATP +
Glutamine +
H2O
Mg2+
ATP +
Glutamine +
Pi
Phosphoribosylformyl-glycinamide synthetase
5′-Phosphoribosyl-N-formylglycinamide
ATP
Mg2+
K+
ADP + Pi
Phosphoribosyl-aminoimidazole synthetase
CO2
Phosphoribosyl-aminoimidazole carboxylase
5′-Phosphoribosyl-5-aminoimidazole-4-carboxylic acid
ATP + Aspartate
Mn2+
ADP + Pi
Phosphoribosylamino-imidazole-succinocarboxamide synthetase
5′-Phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole
5′-Phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole
Fumarate
Adenylo-succinate lyase
5′-Phosphoribosyl-4-carboxamide-5-aminoimidazole
N10-Formyl-FH4
K+
FH4
Phosphoribosylaminoimidazole-carboxamide formyltransferase
5′-Phosphoribosyl-4-carboxamide-5-formamidoimidazole
H2O
IMP Cyclohydrolase
Inosinic acid (IMP)
Ribose-Ⓟ

Fig. 30. Biosynthesis of inosinic acid (IMP). (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

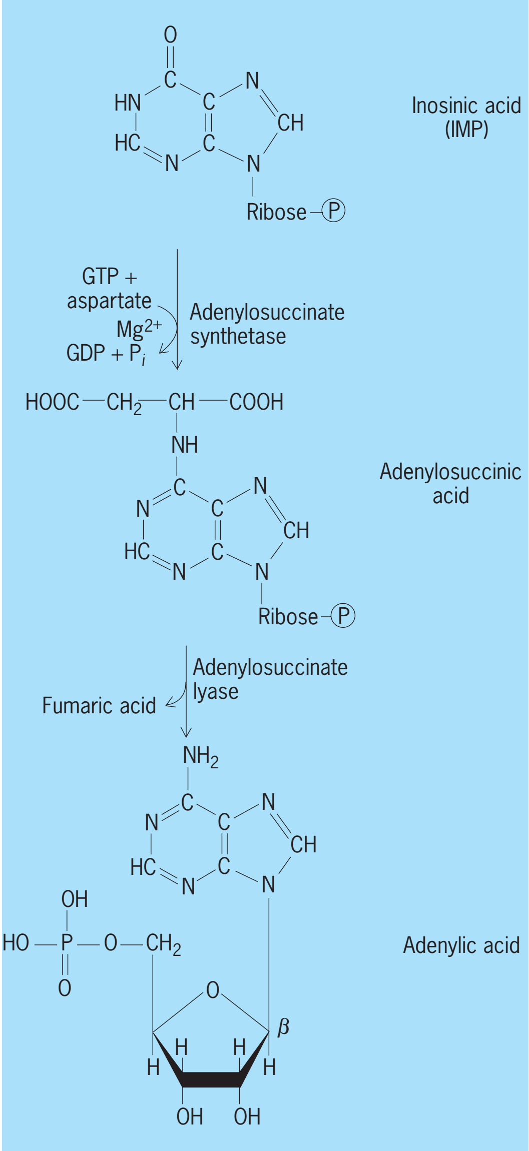

Inosinic acid (IMP)
Ribose—Ⓟ
GTP + aspartate
Mg2+
GDP + Pi
Adenylosuccinate synthetase
HOOC—CH2—CH—COOH
NH
Adenylosuccinic acid
Ribose—Ⓟ
Fumaric acid
Adenylosuccinate lyase
NH2
Adenylic acid
HO—P—O—CH2
OH
β

Fig. 31. The pathway from inosinic acid to adenylic acid. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

many protein molecules of the cell by a series of endergonic reactions (discussed later).

Just as intermediate metabolites resulting (usually) from carbohydrate metabolism are the starting points for amino acid synthesis, so are other catabolic intermediates the raw materials for biosynthesis of the purine and pyrimidine bases found in the nucleotides, which themselves are precursors of the nucleic acids. Different multienzyme pathways are involved in the formation of the two bases. The diverse compounds which contribute specific atoms of the purine and pyrimidine rings are shown in **Figs. 25** and **26**.

Pyrimidine nucleotide biosynthesis. Largely as a result of the work of Arthur Kornberg and colleagues, the pathway for the conversion of aspartate to uridylic acid, and thence to uridine triphosphate (UTP) and cytidine triphosphate (CTP), was unraveled. As shown in **Fig. 27**, this pathway involves the condensation of aspartate and carbamyl phosphate, and the splitting out of water, for the synthesis of the first ring compound (L-dihydroorotic acid) in the reaction sequence; following an oxidation involving the coenzymes FAD, FMN, and NAD^+, the addition of an energy-rich form of the pentose sugar ribose (PRPP) results in the formation of the nucleotide, orotidine-5′-phosphate. This nucleotide is then decarboxylated to yield uridylic acid (UMP; **Fig. 28**), which is then phosphorylated with ATP to form UTP, as in reactions (93*a*) and (93*b*), and some of this

$$\text{UMP} + \text{ATP} \rightleftharpoons \text{UDP} + \text{ADP} \tag{93a}$$

$$\text{UDP} + \text{ATP} \rightleftharpoons \text{UTP} + \text{ADP} \tag{93b}$$

UTP aminated, with the expenditure of another ATP, to form cytidine triphosphate (CTP; **Fig. 29**).

Purine nucleotide biosynthesis. There is a much more complex biosynthetic pathway for synthesis of the purine nucleotides adenosine triphosphate (ATP) and guanosine triphosphate (GTP). Ribose-5-phosphate, formed from glucose by enzymes catalyzing the initial steps of the hexose monophosphate pathway, is activated, that is, converted to an energy-rich form (**Fig. 30**), and to the resulting product (PRPP) an amino group (donated by

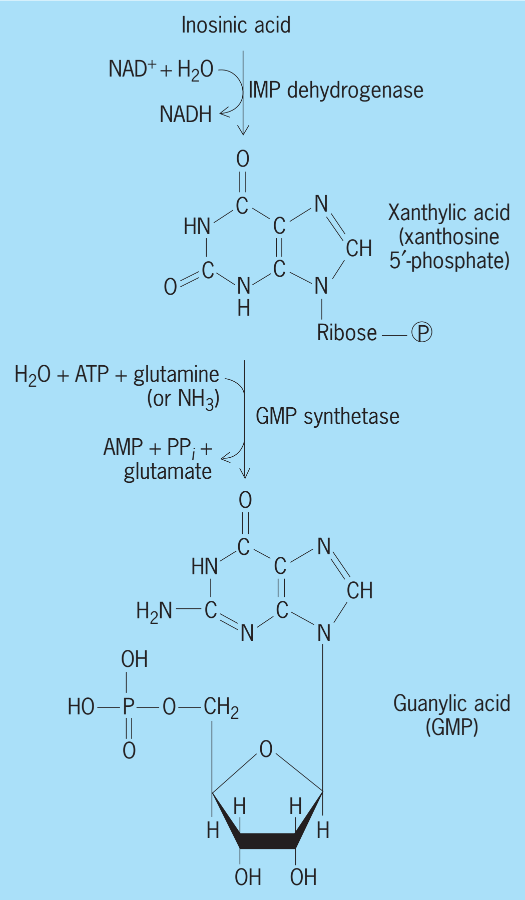

Inosinic acid
NAD+ + H2O
NADH
IMP dehydrogenase
Xanthylic acid (xanthosine 5′-phosphate)
Ribose—Ⓟ
H2O + ATP + glutamine (or NH3)
AMP + PPi + glutamate
GMP synthetase
H2N
HO—P—O—CH2
Guanylic acid (GMP)

Fig. 32. The pathway from inosinic acid to guanylic acid. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

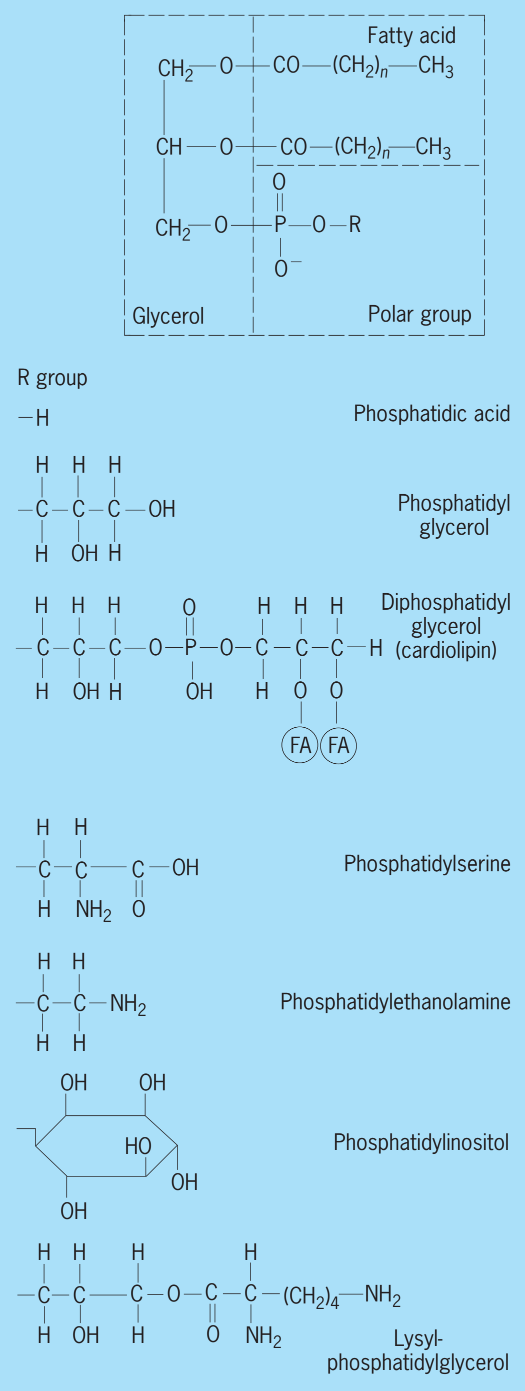
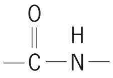

Fatty acid
Glycerol
Polar group
R group
Phosphatidic acid
Phosphatidyl glycerol
Diphosphatidyl glycerol (cardiolipin)
FA
FA
Phosphatidylserine
Phosphatidylethanolamine
Phosphatidylinositol
Lysyl-phosphatidylglycerol

Fig. 33. Phospholipids found in bacteria. (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed, Halsted Press* [*John Wiley*]*, 1973*)

glutamine) is added, the amino acid glycine is linked via a peptide bond to that amino group, and then a reduction (involving a reduced form of the cofactor folic acid) is carried out to form 5-phosphoribosyl-*N*-formyl glycinamide. Later steps in this enzymatic sequence involve addition of more N (from glutamine) at the expense of ATP, synthesis of the imidazole ring, addition of a carbon atom in the form of CO_2, incorporation of aspartic acid, further folate-involved reductions, and finally formation of the pyrimidine portion of the purine ring and the nucleotide inosinic acid (IMP; Fig. 30). Additional enzymes convert this IMP to adenylic acid (AMP; **Fig. 31**) or guanylic acid (GMP; **Fig. 32**) or both. These nucleotides are then converted to their triphosphate forms (ATP, GTP) again at the expense of ATP generated in catabolism; the triphosphates are used for synthesis of nucleic acids and proteins.



$$-\overset{\overset{\displaystyle O}{\|}}{C}-\overset{H}{N}-$$

Lipid biosynthesis. In bacterial cells, lipids (**Fig. 33**) usually occur in two main forms: as the lipoproteins and phospholipids which are part of the cell's cytoplasmic membrane and as triglycerides which are storage forms. Triglycerides are formed from fatty acids and glycerol, the latter derived from glucose via reduction of the dihydroxyacetone phosphate (produced in glycolysis) to alpha-glycerol phosphate, as in reaction (94).

$$\text{Dihydroxyacetone phosphate} + \text{NADH} + \text{H}^+ \rightleftharpoons \underset{(\alpha\text{-glycerol phosphate})}{\text{L-glycerol-3-phosphate}} \qquad (94)$$

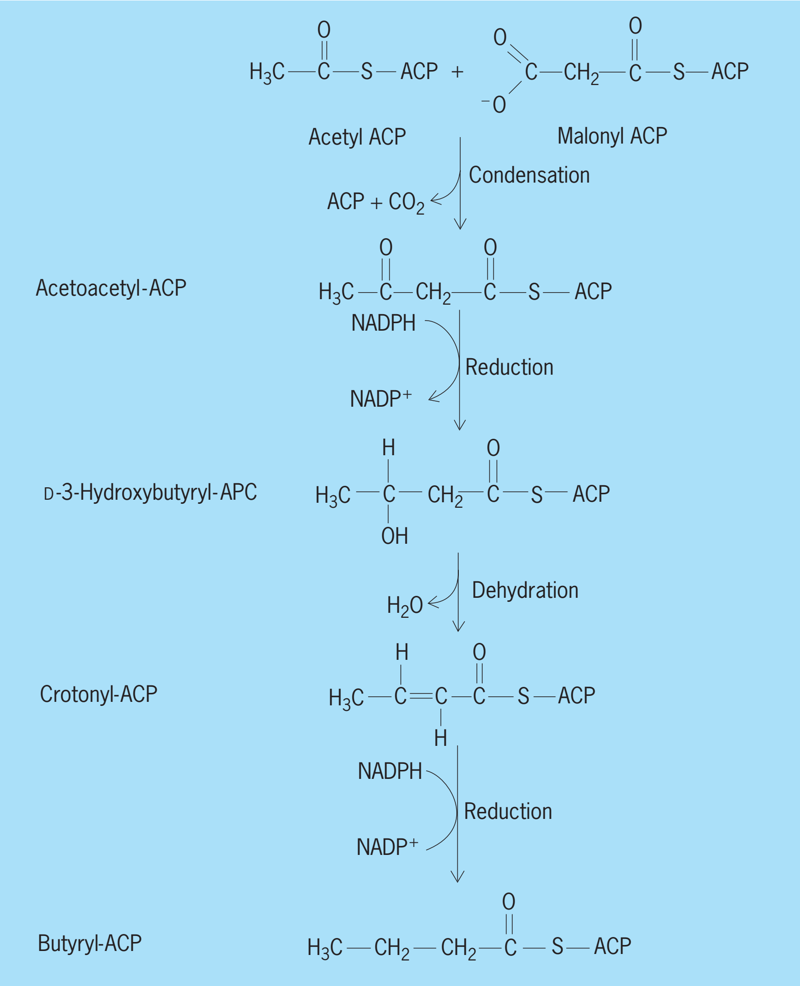

Acetyl ACP
Malonyl ACP
Condensation
ACP + CO2
Acetoacetyl-ACP
NADPH
Reduction
NADP+
D-3-Hydroxybutyryl-APC
Dehydration
H2O
Crotonyl-ACP
NADPH
Reduction
NADP+
Butyryl-ACP

Fig. 34. Reaction sequence in the synthesis of fatty acids: condensation, reduction, dehydration, and reduction. The intermediates shown here occur in the first round of synthesis. (*After L. Stryer, Biochemistry, W. H. Freeman, 1975*)

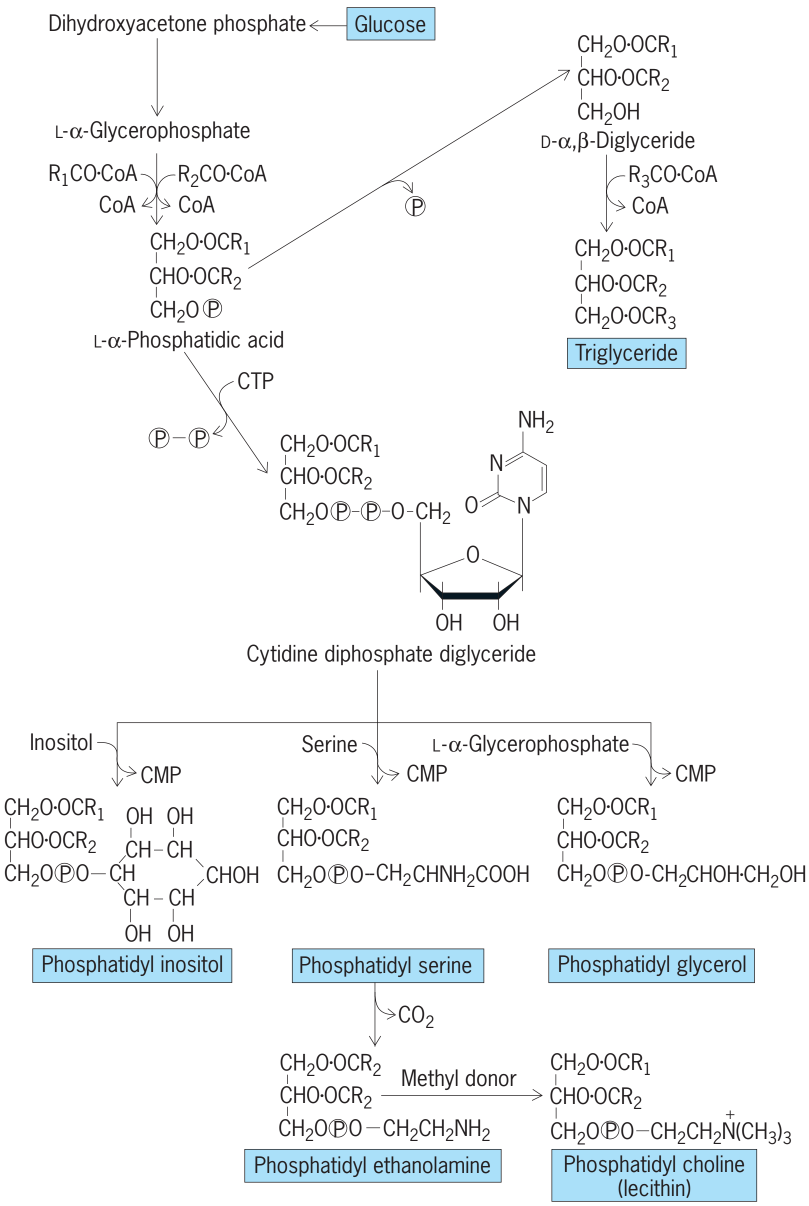
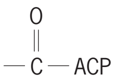

Dihydroxyacetone phosphate
Glucose
L-α-Glycerophosphate
R₁CO·CoA
R₂CO·CoA
CoA
CoA
CH₂O·OCR₁
CHO·OCR₂
CH₂O Ⓟ
L-α-Phosphatidic acid
Ⓟ
CH₂O·OCR₁
CHO·OCR₂
CH₂OH
D-α,β-Diglyceride
R₃CO·CoA
CoA
CH₂O·OCR₁
CHO·OCR₂
CH₂O·OCR₃
Triglyceride
CTP
Ⓟ–Ⓟ
CH₂O·OCR₁
CHO·OCR₂
CH₂OⓅ-Ⓟ-O-CH₂
NH₂
N
O
N
O
OH OH
Cytidine diphosphate diglyceride
Inositol
CMP
Serine
CMP
L-α-Glycerophosphate
CMP
CH₂O·OCR₁
CHO·OCR₂
CH₂OⓅO–CH
OH OH
CH–CH
CHOH
CH–CH
OH OH
Phosphatidyl inositol
CH₂O·OCR₁
CHO·OCR₂
CH₂OⓅO–CH₂CHNH₂COOH
Phosphatidyl serine
CH₂O·OCR₁
CHO·OCR₂
CH₂OⓅO-CH₂CHOH·CH₂OH
Phosphatidyl glycerol
CO₂
CH₂O·OCR₂
CHO·OCR₂
CH₂OⓅO–CH₂CH₂NH₂
Phosphatidyl ethanolamine
Methyl donor
CH₂O·OCR₁
CHO·OCR₂
CH₂OⓅO–CH₂CH₂N⁺(CH₃)₃
Phosphatidyl choline (lecithin)

Fig. 35. Biosynthesis of triglycerides and phospholids. (Note that fatty acyl S · ACP rather than fatty acyl S · CoA functions as the fatty acyl donor in some microorganisms.) (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed., Halsted Press [John Wiley], 1973*)

Fatty acid synthesis utilizes acetyl coenzyme A (acetyl-CoA), formed from pyruvate, which is then carboxylated, with the involvement of CO_2, ATP, and an enzyme which contains the cofactor, biotin, to form malonyl-CoA, as in reaction (95). The

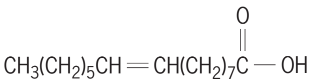
$$\text{Acetyl} - \text{CoA} + CO_2 + H^+ + \text{ATP} \rightleftharpoons \text{malonyl} - \text{CoA} + \text{ADP} + P_{\text{inorganic}} \qquad (95)$$

malonyl-CoA and additional acetyl-CoA molecules are then converted to the forms in which they are active in fatty acid biosynthesis by reacting with a sulfhydryl protein (acyl-carrier-protein, ACP-SH); these ACP derivatives are enzymatically condensed to form acetoacetyl-ACP, CO_2, and ACP-SH (**Fig. 34**). Following the multistep reduction of the acetoacetyl-ACP to butyryl-ACP, to the latter another malonyl-ACP is added to form a six-carbon ACP derivative (and, again, CO_2 and an ACP-SH), which again is reduced to form the saturated fatty acid–ACP compound. Repetition of the malonyl-ACP addition and reduction events with 8, 10, 12, etc., carbon fatty acid–ACP molecules results in the formation of the long-chain saturated fatty acids. For the synthesis of palmitic acid (16 carbon atoms), 8 acetyl-CoA molecules, 7 of ATP, and 14 of NADPH are utilized. Formation of some of the long-chain monounsaturated fatty acids (those containing a double bond, —CH═CH—) diverges from the biosynthetic pattern at the 10-carbon–ACP-derivative stage: a specific dehydratase forms a double bond between the beta and gamma carbons, and then three additional two-carbon units are successively incorporated and reduced, but this addition is to the



$$-\overset{\overset{\large O}{\|}}{C}-\text{ACP}$$

portion of the molecule, rather than, as before, to the terminal methyl (CH_3—) carbon, thus eventually resulting in the formation of an acid such as palmitoleic (C_{16}, monounsaturated):



$$CH_3(CH_2)_5CH=CH(CH_2)_7\overset{\overset{\large O}{\|}}{C}-OH$$

Other pathways for synthesis of monounsaturated fatty acids may well exist.

The amounts of triglycerides—glycerol which is esterified with three different long-chain fatty acids—are rather small in bacteria. Enzymatic synthesis proceeds using a molecule of L-alpha-glycerophosphate and two different (depicted as R_1 and R_2 in **Fig. 35**) molecules of fatty acids in the CoA esterified form resulting in the formation of a phosphatidic acid which, by loss of phosphate, is converted to a diglyceride. Triglyceride formation involves addition of a third fatty acid, via its CoA derivative (Fig. 35).

The phosphatidic acid is a branching point in this biosynthetic pathway, for synthesis of phospholipids proceeds by the enzymatic linkage of part of a cytidine triphosphate molecule to the unesterified hydroxyl group of the phosphatide (Fig. 35). The resulting coenzyme (cytidine diphosphate glyceride) is the active intermediate which reacts with either inositol, serine, or L-alpha-glycerophosphate to form, finally, phosphatidyl-inositol, -serine, -glycerol, -ethanolamine, or -choline, which constitute the major cellular phospholipids, and are found mainly in the cell membrane or membrane wall (envelope) complex of gram-positive or gram-negative organisms, respectively.

Glycolipids, compounds which contain sugar(s) linked to the unesterified hydroxyl group of a diglyceride, are present in bacterial cells, again most likely in the membranes or in the membrane wall complexes. Mannose, glucose, galactose, glucosamine, and rhamnose are among the sugars present in different glycolipids. The sugars may be present singly or linked to form di- or trisaccharide chains which are joined by glycosidic linkage to the free hydroxyl group of the diglyceride. Details of glycolipid biosynthesis remain unclear, although it seems likely that sugars in nucleotide-coenzyme form (linked to guanosine diphosphate) are intermediates, and the sugar (glycosyl) moiety of this nucleotide is transferred by the responsible enzyme to the diglyceride, with the concomitant release to guanosine diphosphate.

Glycophospholipids, which contain both carbohydrate and phosphate residues, are being found with increasing regularity in diverse bacterial groups, and may be as abundant and as significant as glycolipids. Details of the biosynthesis of these compounds is even less clear than for glycolipids, but the same general synthetic features and types of intermediates may be anticipated.

Other lipids of cellular significance are the terpenes, among which are the carotenoid pigments and the long-chain primary alcohols such as bactoprenol (a compound of importance in wall synthesis). Both types of compounds are synthesized via condensation of acetyl-CoA molecules to form mevalonic acid (**Fig. 36**) which then undergoes extensive modification(s) to form isoprenyl-phosphate compounds which themselves are further modified and joined to form the terpenes. *See* CAROTENOID; CHOLESTEROL.

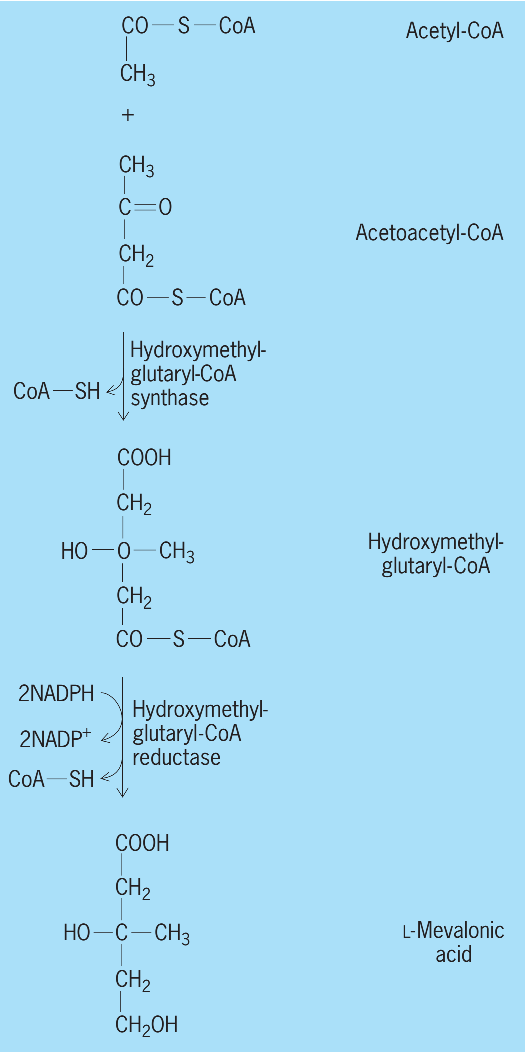

CO—S—CoA
CH3
Acetyl-CoA
+
CH3
C=O
CH2
CO—S—CoA
Acetoacetyl-CoA
CoA—SH
Hydroxymethyl-glutaryl-CoA synthase
COOH
CH2
HO—O—CH3
CH2
CO—S—CoA
Hydroxymethyl-glutaryl-CoA
2NADPH
2NADP+
CoA—SH
Hydroxymethyl-glutaryl-CoA reductase
COOH
CH2
HO—C—CH3
CH2
CH2OH
L-Mevalonic acid

Fig. 36. Synthesis of mevalonic acid. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

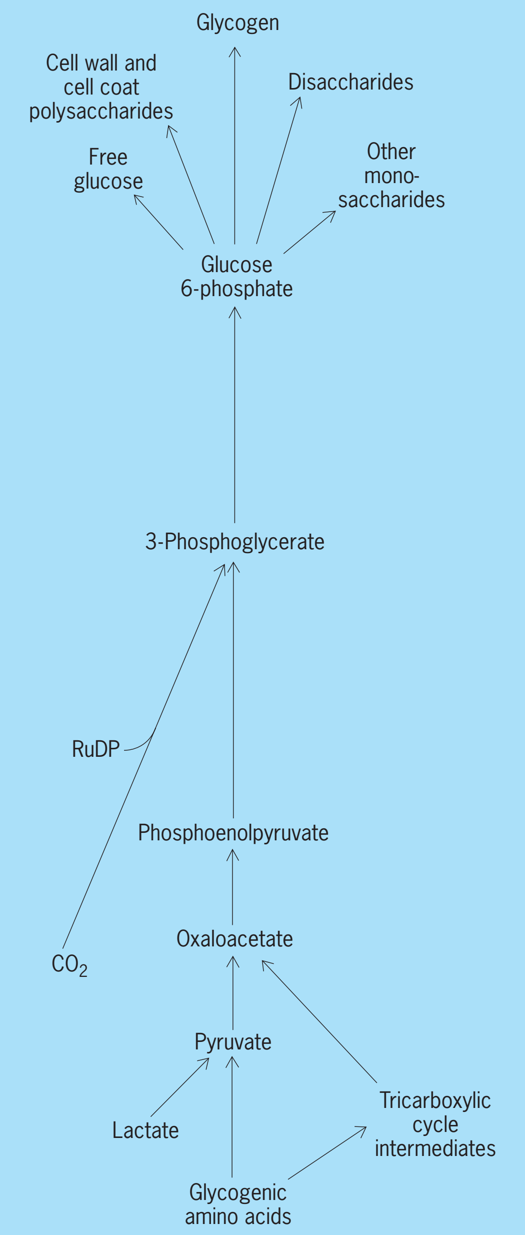

Glycogen
Cell wall and cell coat polysaccharides
Disaccharides
Free glucose
Other mono-saccharides
Glucose 6-phosphate
3-Phosphoglycerate
RuDP
Phosphoenolpyruvate
Oxaloacetate
CO2
Pyruvate
Lactate
Tricarboxylic cycle intermediates
Glycogenic amino acids

Fig. 37. Central pathway of hexose biosynthesis. Auxiliary pathways feed into the central pathway from CO_2, lactate, and amino acids; diverging pathways lead from glucose 6-phosphate to other carbohydrates. RuDP is ribulose 1,5-diphosphate, an important intermediate in autotrophic formation of hexose. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

Acetyl-CoA molecules also undergo condensation to form acetoacetyl-CoA, which after reduction becomes beta-hydroxybutyric acid, several molecules of which may be joined by ester linkages (between the hydroxyl group of one molecule and the carboxyl group of another) to form the polymer, poly-beta-hydroxybutyrate; this is a storage form of lipid

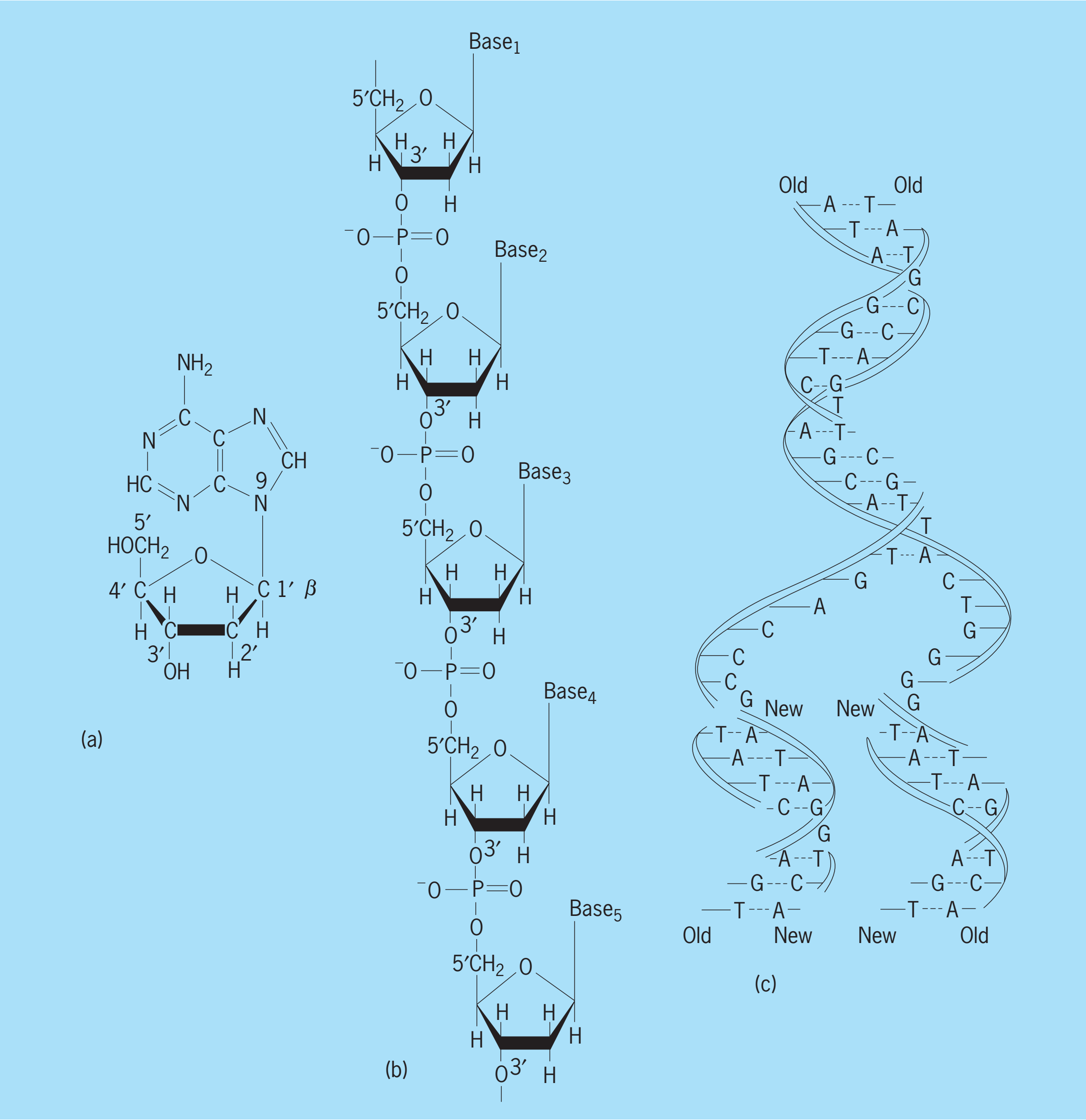


Fig. 38. Deoxynucleotide and polydeoxynucleotide structures. (*a*) 2′-Deoxyadenosine. (*b*) A polydeoxynucleotide chain. (*c*) Configuration of the DNA double helix, showing mechanism of self-duplication. (*After W. B. Wood, The Molecular Basis of Metabolism* [*Unit 3, Biocore*], *McGraw-Hill, 1974*)

in many bacteria, particularly in situations where substrate carbon is in ample supply, and nitrogen is limited.

A rather different type of storage polymer, polyphosphate, is often found within bacterial cells. Whether these polymers function primarily as an energy or as a phosphorus reserve is not yet certain. Clearly, they differ from other known storage polymers by their lack of carbon and nitrogen. They are, thermodynamically, rather high-energy phosphate compounds and synthesized by one of two ways, involving either a polyphosphate kinase which catalyzes transfer of the terminal phosphoryl group to polyphosphate, reaction (96) by a polyphosphate

$$(P_i)_n + \text{ATP} \rightleftharpoons (P_i)_{n+1} + \text{ADP} \qquad (96)$$

transferase which uses 1,3-diphosphoglycerate as phosphate donor, reaction (97).

$$\text{1,3-diphosphoglycerate} + (P_i)_n \rightleftharpoons \text{3-phosphoglycerate} + (P_i)_{n+1} \qquad (97)$$

Carbohydrate synthesis. Synthesis of this class of molecules has several aspects—the synthesis of simple sugars (monosaccharides) by organisms utilizing compounds such as CO_2, fatty, organic, or amino acids, or other noncarbohydrate materials as their growth substrate(s), and the synthesis of polysaccharides such as glycogen, starch, cell walls, and extracellular polymers.

As indicated in **Fig. 37**, glucose may be synthesized from noncarbohydrate substrates other than CO_2 by comparable, but not identical, intermediates and pathways which are involved in carbohydrate catabolism. Different bacterial types may vary in the precise details followed.

Of special significance is the synthesis of triose and hexose phosphates from CO_2 by photosynthetic and chemosynthetic bacteria. In the former the required energy in the form of ATP and NADH or NADPH results from the trapping of radiant energy and its conversion to ATP and so forth, while in the latter group these compounds are formed as a result of the oxidation of compounds such as NH_4^+, NO_2^-, H_2, H_2S, SO, $S_2O_3^{2-}$, Fe^{2+}, and so forth. The biosynthetic pathways for CO_2 assimilation and conversion into sugars is analogous to the Calvin cycle involved in green plant photosynthesis. *See* PHOTOSYNTHESIS.

Macromolecular biosyntheses. The polymerization of the different monomeric building blocks, whose synthesis has just been sketched, into macromolecules such as polysaccharides, proteins, nucleic acids, cell walls, and extracellular polymers is invariably an energy-requiring process. As will be seen, the requisite energy for driving these endergonic reactions comes, again, from the large amount of ATP formed in catabolism.

Deoxyribonucleic acid. Deoxyribonucleic acid, the genetic material of the cell, exists as double-helical structures of complementary strands of polynucleotides. Four different purine and pyrimidine deoxyribonucleotides—dAMP, dGMP, dCMP, and dTMP—are present in each strand (**Fig. 38**). Replication of a preexisting DNA strand proceeds with a series of enzymes (DNA polymerases) which link the deoxyribonucleoside triphosphates, dATP, dGTP, dCTP, and dTTP, to form a polymer of complementary base composition to the "primer" strand serving as a template; pyrophosphate is eliminated from each triphosphate as this polynucleotide formation occurs. Each strand of the existing double-helical DNA is, presumably, being separated and copied simultaneously, thus resulting, when separation is complete, in the formation of two identical molecules of DNA, each of which contains a single strand of the original, parental DNA. The genetic information contained in the original DNA has thus been duplicated. *See* DEOXYRIBONUCLEIC ACID (DNA).

Ribonucleic acid. The information inherent in the sequence of nucleotides in DNA is transcribed into the polyribonucleotides of an RNA molecule (**Fig. 39**) when one is synthesized from the purine and pyrimidine ribonucleoside triphosphates (ATP, GTP, CTP, UTP), by a polymerizing enzyme (DNA-directed RNA polymerase) which "copies" only one strand of DNA.

The RNA formed by RNA polymerase using DNA as a directive template is of three types: messenger, transfer, and ribosomal (mRNA, tRNA, rRNA). The mRNA serves as a template for protein biosynthesis, and hence each mRNA molecule must contain the information transcribed from a gene, or group of genes; thus this class of molecules is of heterogeneous composition. Transfer RNAs are rather specific for each of the some 20 amino acids with which they interact, and in addition have a portion (the anticodon) of the molecule which is complementary to a specific region(s) (the codons) of mRNA with which they will eventually combine. Ribosomal RNA is incorporated along with proteins into the subcellular particles, ribosomes, which serve as the sites of protein synthesis. *See* RIBONUCLEIC ACID (RNA).

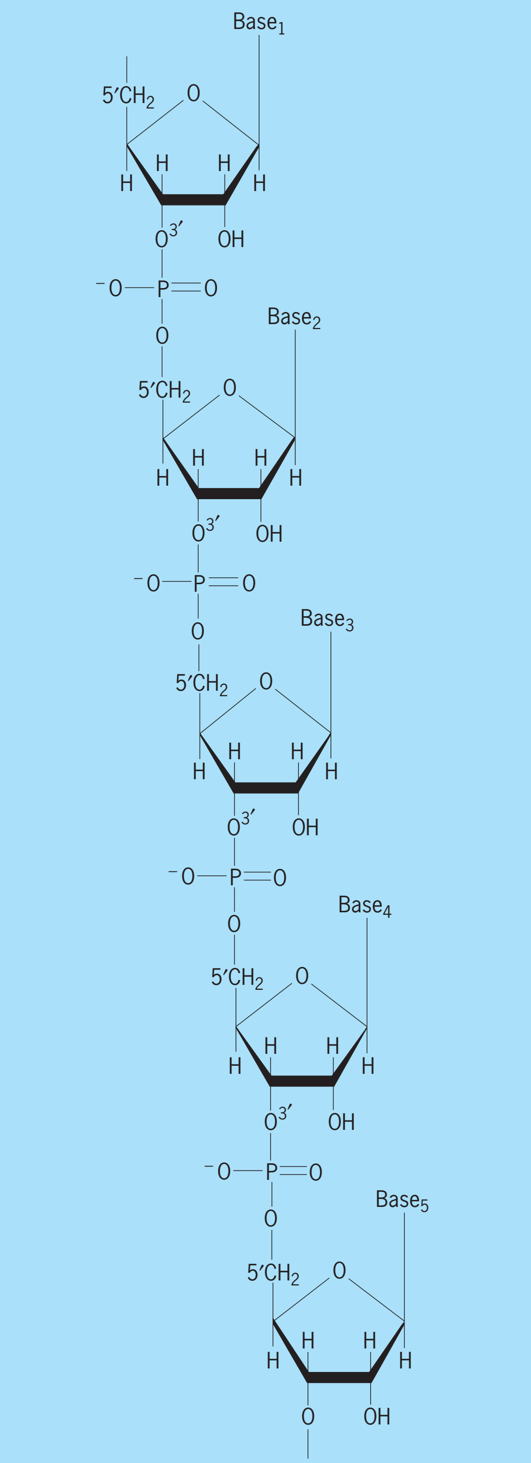


Fig. 39. Structure of a polyribonucleotide. (*After W. B. Wood, The Molecular Basis of Metabolism* [*Unit 3, Biocore*], *McGraw-Hill, 1974*)

Proteins. Protein biosynthesis proceeds after each of the some 20 types of amino acids is activated by means of an enzyme specific for both the amino acid and its corresponding species of tRNA. In the reaction, ATP is cleaved into AMP and pyrophosphate, and the energy is thus provided to drive the reaction.

The resulting amino-acyl-tRNA molecule is endowed with a specificity (in the tRNA moiety) which has its complement(s) in the mRNA strand; thus each amino acid (in the form of its tRNA derivative) is ensured of being located next to the correct neighboring amino acids, that is, the genetic information of DNA and RNA is translated into the correct positioning of amino acids in the polypeptide chain (protein) to be formed.

The site of protein synthesis is the ribosomes which are complex aggregates of RNA ($^{2}/_{3}$ by weight) and protein ($^{1}/_{3}$ by weight). Once the mRNA, the tRNAs, and the ribosome have been appropriately complexed, the peptide bond formation between adjacent amino acids (that is, protein elongation) occurs as the mRNA is translocated with respect to the ribosome in a complex manner of which all the details are not yet clear. However, the complete process ensures the collinearity of the base sequence (information) in a gene and the amino acid sequence of the product of the gene—a polypeptide (protein). *See* RIBOSOMES.

Although the bulk of cellular protein is synthesized on ribosomes, some rather small proteins, such as polypeptide antibiotics, are not; instead, steps analogous to certain of these fatty acid syntheses occur in the cytoplasm.

Polysaccharides. The bacteria synthesize a variety of polysaccharides, including intracellular reserve, or storage material such as glycogen or glycogenlike materials, as well as extracellular polymers such as the homopolymers levan and dextran, a variety of heteropolysaccharides, and structural polymers such as the cell wall components peptidoglycan and lipopolysaccharide.

The synthesis of intracellular homopolymers such as glycogen usually occurs when growth is limited by the supply of a suitable nitrogen source, and where there is an ample supply of extracellular carbon substrate (note the parallel with the conditions which favor formation of the lipid reserve polymer, poly-beta-hydroxybutyrate). Glycogen biosynthesis proceeds by enzymatic conversion of glucose to glucose-1-phosphate, and the latter compound then is a substrate for the enzyme ADP-glucose pyrophosphorylase which utilizes ATP and the glucose-1-phosphate to form the nucleoside diphosphate derivative of the sugar, adenosine diphosphate glucose (ADP-glucose) and pyrophosphate. The ADP-glucose is the activated, or carrier, form of the glucose and substrate for the enzyme glycogen synthase which repetitively adds the glucose moiety to an existing (primer) bit of glycogen and thus extends the chain by one glucose unit, and releases the ADP.

Nucleoside diphosphate sugar derivatives (**Fig. 40**) are widely used for biosynthesis of other polysaccharides, but more often than not the carrier, or activated, form is a derivative of uridine diphosphate (UDP) rather than ADP as was the case for glycogen synthesis. This is true, for example, for synthesis of a last part of the lipopolysaccharides, and the cell wall polymer, peptidoglycan. But, by contrast, the synthesis of another type of polymer found in bacterial cell walls, the teichoic acids, involves the cytidine diphosphate (CDP) derivatives of the sugar alcohols, glycerol or ribitol, to form glycerol-teichoic acid or ribitol-teichoic acid respectively. In other instances guanosine diphosphate (DP) appears to serve as the carrier, or activated, form. These differences in the particular nucleoside triphosphate used in polysaccharide biosynthesis should not, however, obscure the significant role that these compounds play in biosynthesis of macromolecules other than the nucleic acids.

Because such a major portion (up to 15%) of a cell's total dry weight may be in the form of the bag-shaped macromolecule peptidoglycan, and because this molecule is unique to the bacterial world, it is important to consider the general features of its synthesis (**Fig. 41**). Peptidoglycan, as the name implies, consists of peptides and sugars; the linear chains of polysaccharides are crosslinked to one another by means of short peptides. The repeating units of the polysaccharide chains are two glucose derivatives, *N*-acetyl glucosamine and *N*-acetyl muramic

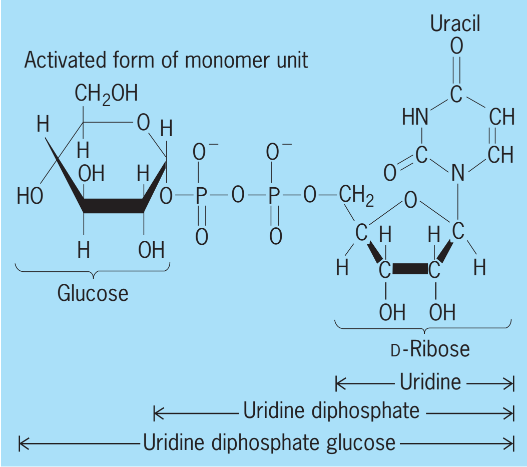


Fig. 40. Structure of the nucleoside diphosphate sugar, urdine diphosphate glucose. (*After W. B. Wood, The Molecular Basis of Metabolism* [*Unit 3, Biocore*], *McGraw-Hill, 1974*)

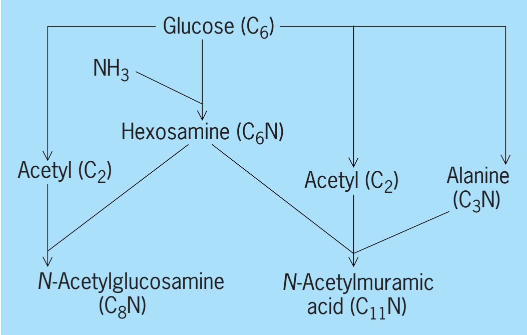


Fig. 41. Synthesis of hexosamines. The acetylated hexosamines are constituents of the peptidoglycans which are essential components of bacterial cell walls. (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed., Halsted Press* [*John Wiley*], *1973*)

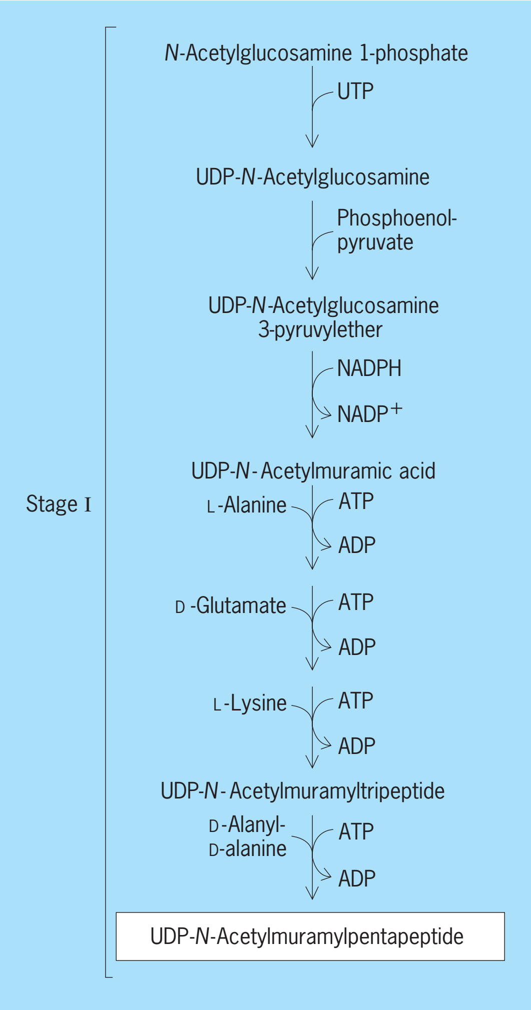

Stage I
N-Acetylglucosamine 1-phosphate
UTP
UDP-N-Acetylglucosamine
Phosphoenol-pyruvate
UDP-N-Acetylglucosamine 3-pyruvylether
NADPH
NADP+
UDP-N-Acetylmuramic acid
L-Alanine
ATP
ADP
D-Glutamate
ATP
ADP
L-Lysine
ATP
ADP
UDP-N-Acetylmuramyltripeptide
D-Alanyl-D-alanine
ATP
ADP
UDP-N-Acetylmuramylpentapeptide

Fig. 42. Steps in the biosynthesis of the peptidoglycan of a bacterial cell wall. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

acid which, as their UDP derivatives, are linked to form a disaccharide. To the UDP-*N*-acetyl muramic acid is added (**Fig. 42**), successively, the amino acids L-alanine, D-glutamate, L-lysine, and finally a dipeptide, D-alanyl-D-alanine, thus forming a pentapeptide chain linked to the UDP-*N*-acetyl muramic acid (the formation of each peptide link requires the expenditure of an ATP; ribosomes, however, are not involved).

The next stage in synthesis involves enzymatic transfer of the UDP–muramic acid–pentapeptide to a terpene, undecaprenyl phosphate, which is bound in the cell membrane. This lipid acts as a carrier molecule which is able to effect the transport of the polar (charged) UDP across the lipid-rich cell membrane. To the *N*-acetyl muramic acid portion of the complex is first added *N*-acetyl glucosamine (as its UPP derivative), and then to the resulting dissacharide-pentapeptide complex is added short peptide chains which will act as cross links between these peptide chains attached to the muramic acid moieties (**Fig. 43**).

The final step in peptidoglycan formation is cleavage of this entire disaccharide unit from the lipid carrier and concomitant transfer to a preexisting (primer) polysaccharide chain. Cross-linking between parallel peptidoglycan chains occurs via the short peptide chains (**Fig. 44**). Repetitive addition of the disaccharide repeating unit (including side chains and cross-linking chains) and repeated cross-linking (it is this latter reaction that is sensitive to the antibiotic penicillin) brings about the encircling of the cell with the single molecule of enormous size and strength. To this rigid backbone are attached the various teichoic acids, and in gram-negative organisms an additional wall layer, the complicated lipopolysaccharide.

Extracellular polymers. Although the synthesis of many extracellular polysaccharides involves nucleoside diphosphate carriers, others, usually homopolymers (such as dextran and levan), are synthesized without involvement of these coenzymes. Particularly when dextran or levan is synthesized by cells growing on sucrose (a disaccharide of glucose and fructose), a rather different principle is used: the energy of the glycosidic bond linking the two

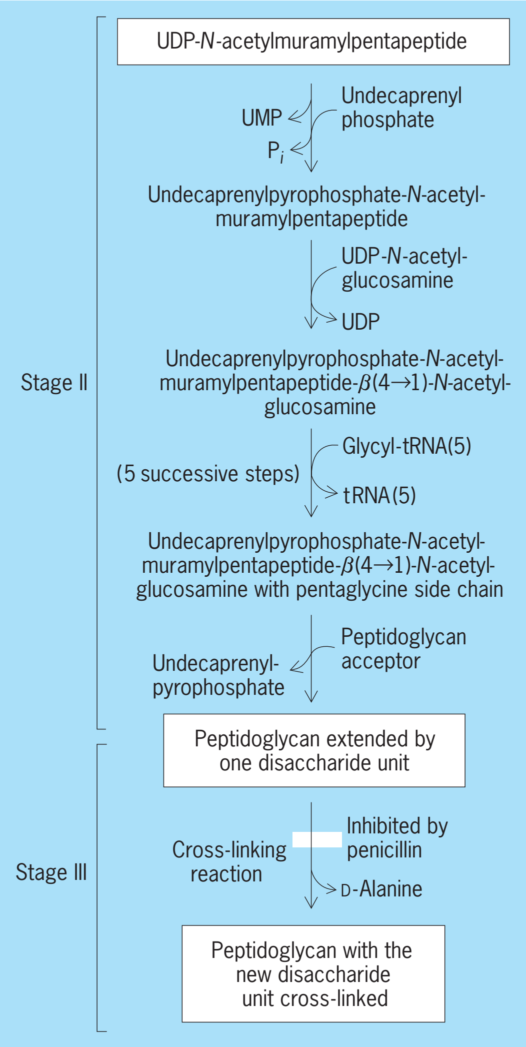

UDP-N-acetylmuramylpentapeptide
UMP
Undecaprenyl phosphate
Pi
Undecaprenylpyrophosphate-N-acetyl-muramylpentapeptide
UDP-N-acetyl-glucosamine
UDP
Stage II
Undecaprenylpyrophosphate-N-acetyl-muramylpentapeptide-β(4→1)-N-acetyl-glucosamine
(5 successive steps)
Glycyl-tRNA(5)
tRNA(5)
Undecaprenylpyrophosphate-N-acetyl-muramylpentapeptide-β(4→1)-N-acetyl-glucosamine with pentaglycine side chain
Undecaprenyl-pyrophosphate
Peptidoglycan acceptor
Peptidoglycan extended by one disaccharide unit
Stage III
Cross-linking reaction
Inhibited by penicillin
D-Alanine
Peptidoglycan with the new disaccharide unit cross-linked

Fig. 43. Stages II and III in peptidoglycan biosynthesis. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

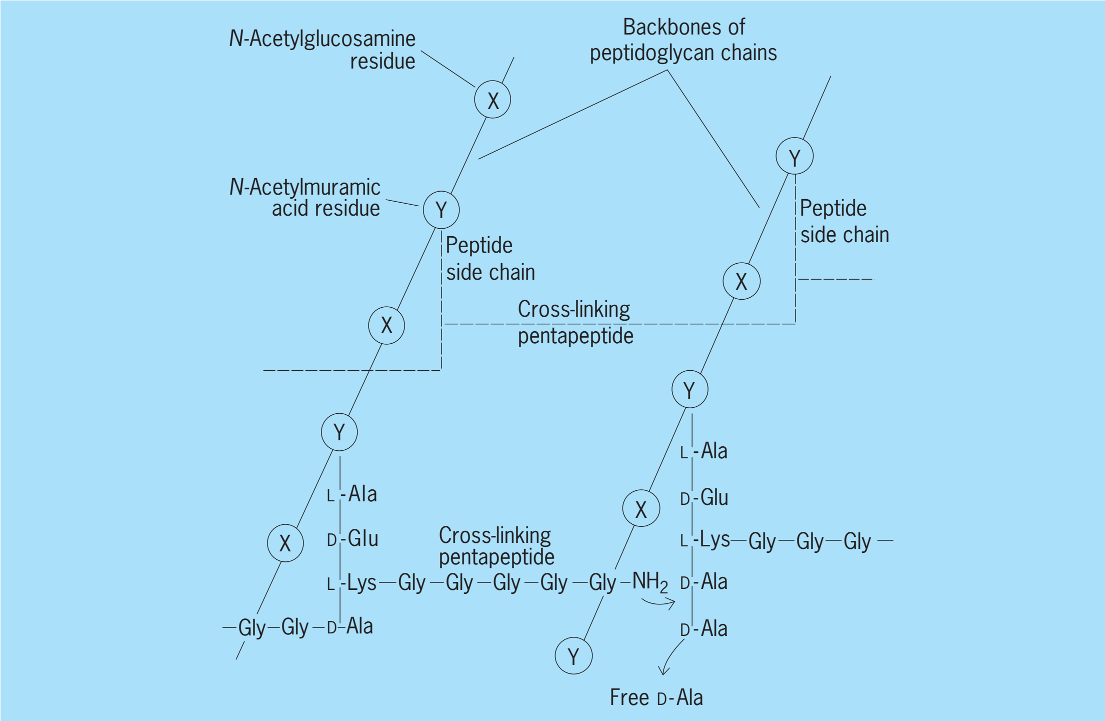

N-Acetylglucosamine residue
Backbones of peptidoglycan chains
N-Acetylmuramic acid residue
Peptide side chain
Peptide side chain
Cross-linking pentapeptide
X
Y
X
Y
X
Y
X
Y
X
Y
L-Ala
D-Glu
L-Lys—Gly—Gly—Gly—Gly—Gly—NH2
—Gly—Gly—D-Ala
Cross-linking pentapeptide
L-Ala
D-Glu
L-Lys—Gly—Gly—Gly—
D-Ala
D-Ala
Free D-Ala

Fig. 44. Completion of a cross link between two adjacent peptidoglycan chains in the bacterial cell wall. This reaction is blocked by penicillin. (*After A. Lehninger, Biochemistry, 2d ed., Worth Publishing Co., 1975*)

monosaccharides is preserved in an enzyme-glucose complex formed when the enzyme dextransucrase, for example, cleaves sucrose with the release of fructose. The energy inherent in the enzyme-bound glucose moiety is then utilized when this glucosyl unit is added to a dextran molecule, thus increasing the latter's size. Similarly when the enzyme levansucrase cleaves sucrose, with the release of glucose, the enzyme-bound fructosyl moiety is in an activated form and may be added to a levan molecule without utilization of an additional energy source.

Synthesis of extracellular heteropolysaccharides, where two or more different carbohydrates or carbohydrate-derivatives are polymer constituents, invariably involves the successive, stepwise addition of the monomers as nucleoside diphosphate and complexes.

Energy and growth. It should thus be clear that synthesis of cell materials (anabolism), an endergonic process, utilizes massive amounts of ATP. Although it would be very difficult to calculate the amount of ATP needed by an organism synthesizing all its cell material from a single substrate such as acetate, it is a somewhat easier task to calculate the ATP requirements for synthesis of macromolecules from their monomeric units, where only polymerization reactions are involved. Such calculations indicate that for every 0.1 g (dry weight) cell material formed, over 3 millimoles of ATP are needed, and that, accordingly, 1 mole of ATP should be sufficient energy to form over 30 g cell material from preformed monomers.

Experimental assessment of the molar growth yields of bacteria and yeasts reveals that only a scant 10 g dry weight of cell material is actually formed per mole of ATP generated in catabolism. The basis for the discrepancy between this 10 g, as measured experimentally, and the some 30 g, as calculated, is not yet clearly understood, but the "loss" may represent energy expenditure for cellular repairs, active transport, motility, and, of course, as heat.

Transport across membranes. The very composition of the cell or plasma membrane, rich in phospholipids, proteins, and combinations thereof, renders it an effectively impermeable barrier particularly to highly polar molecules and ions, which are of course characteristic intermediate metabolites in the cell; the membrane thus serves a useful purpose in preventing diffusion of these metabolic intermediates, formed intracellularly, from the cell. The same membrane feature simultaneously excludes a variety of polar organic molecules (such as glucose, acetate and amino acids) from gaining entrance to the cell interior by simple diffusion across the cell membrane, and in addition, prevents the loss or release, by diffusion, of metabolic end products which may be toxic to the cell if appreciable amounts of them accumulate. The rapid growth rate of bacteria demands a very high metabolic rate, and hence of substrate availability and end-product removal. These problems are met by bacteria by the use of transport mechanisms which effect a movement of glucose, amino acids, and so on, not only across the cell membrane but against a concentration gradient. Because such activity results in an increase in free energy, the process must be driven by energy-yielding reactions which in bacteria are known to be of several types.

Located in the periplasmic space—the area between the cell membrane and cell wall—are a number of different proteins that bind, selectively, different inorganic ions, sugars, or amino acids. These binding proteins are thought to play an important role in transport processes.

One type of sugar transport mechanism known to be present in some anaerobic bacteria involves the phosphorylation of incoming sugar by an enzyme located in the cell membrane using as the energy source a specific phosphorylated protein formed in the interior cytoplasm. Transported across the membrane, thus, is not the free sugar, but the phosphorylated form. In some aerobic bacteria a different tactic is employed: amino acids and sugars are not chemically modified (for example, phosphorylated), but their transport is intimately coupled to transfer of electrons from substrate to oxygen. The details of this coupling are not fully known, but one view, that of the chemoosmotic hypothesis, presumes a proton-gradient generated by electron transport as the driving force for such active transport.

Cell organization and enzyme location. It has been mentioned for enzymes involved in transport process and for final stages of peptidoglycan synthesis that some enzymes occupy particular locales in the bacterial cell. It is roughly accurate to say that most enzymes of substrate catabolism and monomer, or building-block, biosynthesis are located in the cytosol, and hence termed "soluble" enzymes, while many of those concerned with polymerization or synthetic reactions of macromolecules involved in

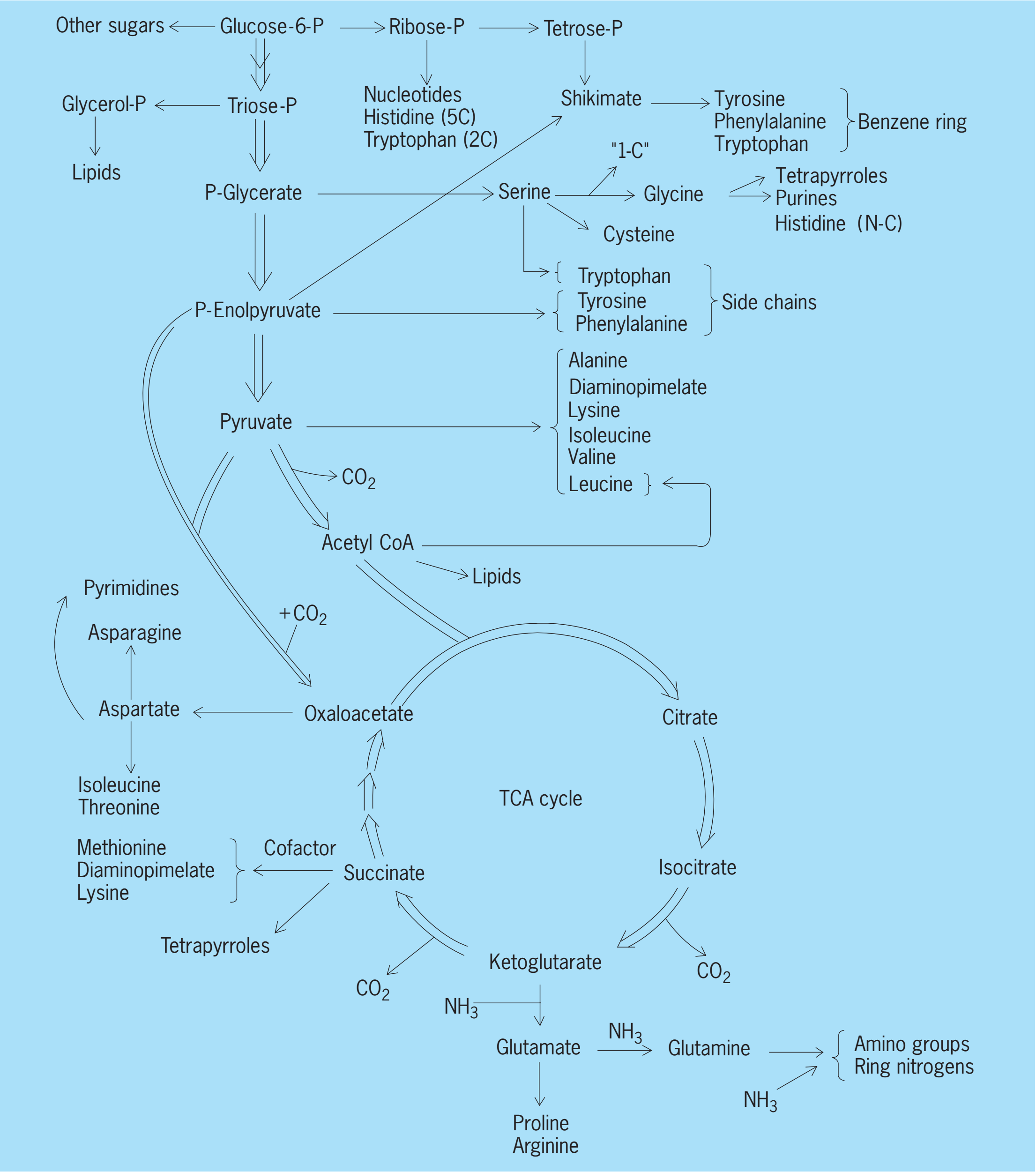

Other sugars
Glucose-6-P
Ribose-P
Tetrose-P
Glycerol-P
Triose-P
Nucleotides
Histidine (5C)
Tryptophan (2C)
Shikimate
Tyrosine
Phenylalanine
Tryptophan
Benzene ring
Lipids
"1-C"
P-Glycerate
Serine
Glycine
Tetrapyrroles
Purines
Histidine (N-C)
Cysteine
Tryptophan
Tyrosine
Phenylalanine
Side chains
P-Enolpyruvate
Alanine
Diaminopimelate
Lysine
Isoleucine
Valine
Leucine
Pyruvate
CO2
Acetyl CoA
Lipids
Pyrimidines
+CO2
Asparagine
Aspartate
Oxaloacetate
Citrate
Isoleucine
Threonine
TCA cycle
Methionine
Diaminopimelate
Lysine
Cofactor
Succinate
Isocitrate
Tetrapyrroles
Ketoglutarate
CO2
CO2
NH3
NH3
Glutamate
Glutamine
Amino groups
Ring nitrogens
NH3
Proline
Arginine

Fig. 45. Relation of the amphibolic pathways (indicated by the double arrows) to the main anabolic pathways. (*After B. Davis et al., Microbiology, 2d ed., Harper and Row, 1973*)

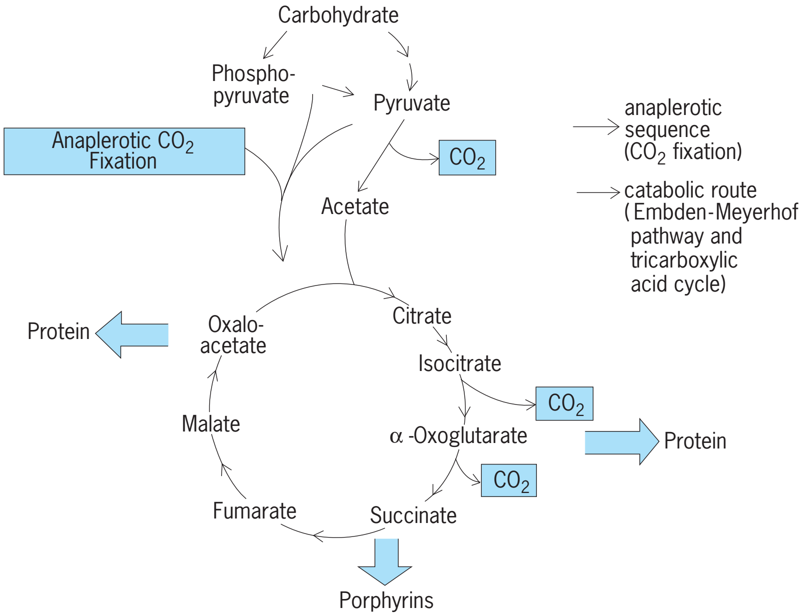


Fig. 46. Routes for the provision of energy and of cell components during the utilization of carbohydrates. (*After P. N. Campbell and G. D. Greville, eds., Essays in Biochemistry, vol. 2, Academic Press, 1966*)

structural features of the cell are bound to, or located in, those structural components. In addition to enzymes involved in synthesis and degradation of phospholipids, biosynthesis of glycolipids, peptidoglycans, lipopolysaccharides, exopolysaccharides, teichoic acids, and other macromolecules significant in cell structure, there are membrane-associated enzymes involved in other processes: ATPases, electron transport components (such as oxidases, dehydrogenases and cytochromes) and, perhaps, enzymes for initiation of DNA synthesis. In contrast, enzymes involved in carbohydrate degradation, biosynthesis of amino acids, purines and pyrimidines, and so on, are thought not to be bound to cell structural components. Some enzyme activities involved with synthesis or degradation of storage polymers such as poly-beta-hydroxybutyrate or glycogen may be associated with particles composed of these polymers which are found in the cytosol.

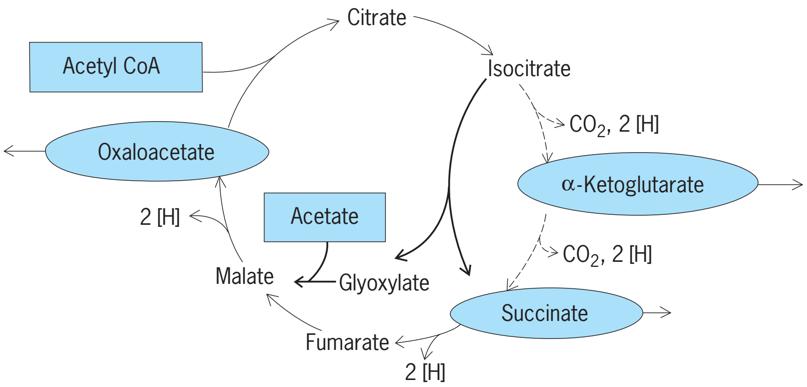


Fig. 47. Krebs cycle, with the glyoxylate cycle reactions (heavy arrows) inserted. The broken arrows indicate reactions which may be bypassed when the glyoxylate cycle is being used by organisms growing with acetate as sole source of carbon and energy. (*After B. Wood et al., Microbiology, 2d ed., Harper and Row, 1973*)

Anaplerotic sequences. It should be clear from the discussions of catabolism and anabolism that a great deal of substrate carbon is used in the formation of one of a cell's major products—another cell—as well as in the compounds more commonly regarded as end products, CO_2 for aerobes, acids or alcohols for anaerobes. The source(s) of the cellular macromolecules are, of course, the building blocks derived from the pathways of carbohydrate breakdown or Krebs cycle activities. But how, for example, is the latter cyclical series of reactions to continue operation if intermediates in this cycle (such as ketoglutarate, oxaloacetate, and succinate) are being depleted as they are used, as they almost constantly are, for amino acid and purine and pyrimidine biosynthesis?

Clearly there must be ancillary reactions, or pathways, which effect a net synthesis and replenishment of certain metabolic intermediates and thus make possible the maintenance of cyclical activities. Such replenishment reactions have been termed "anaplerotic" (from the Greek, for "filling up") by the British biochemist Hans Kornberg.

One such key reaction(s) in organisms metabolizing carbohydrates or other compounds which are degraded to pyruvate is the enzymatic condensation of CO_2 and either pyruvate or phosphoenolpyruvate to form oxaloacetate, which thus generates an acceptor molecule for acetyl-CoA and keeps the cycle functioning, as in reaction (98). The interrelationship of this synthesis and other related activities is shown in **Fig. 45**.

$$\text{Pyruvate} + CO_2 + \text{ATP} \longrightarrow \text{oxaloacetate} + \text{ADP} + P_i \qquad (98)$$

For organisms growing on substrates which are not degraded to pyruvate (for example, long-chain fatty acids) or on compounds containing only one, or no, carbon-to-carbon bonds (for example, acetate and methanol), the same problem, that of providing a sufficient supply of intermediates for the Krebs cycle, as well as the additional one of providing precursors for carbohydrate synthesis, must be solved. In the case of organisms growing at the expense of acetate, or compounds degraded to acetate, this is done by means of the glyoxylate cycle, which brings about the net synthesis of a four-carbon dicarboxylic acid, malate, from two acetate molecules (**Fig. 46**). For organisms growing at the expense of one-carbon compounds such as methane, methanol, or methylamine, oxidized forms of these substrate molecules undergo condensation reactions with, on the one hand, glycine, to form serine, and eventually, phosphoglycerate (which is of course readily convertible to pyruvate), or on the other hand, with a pentose phosphate, to form, eventually, a hexose phosphate which can be cleaved to give rise to phosphoglyceraldehyde and thence pyruvate. In both instances some of the triose phosphates formed must be diverted in a cyclical fashion to regenerate the initial acceptor molecules (glycine, or a pentose phosphate) for the oxidized one-carbon compounds.

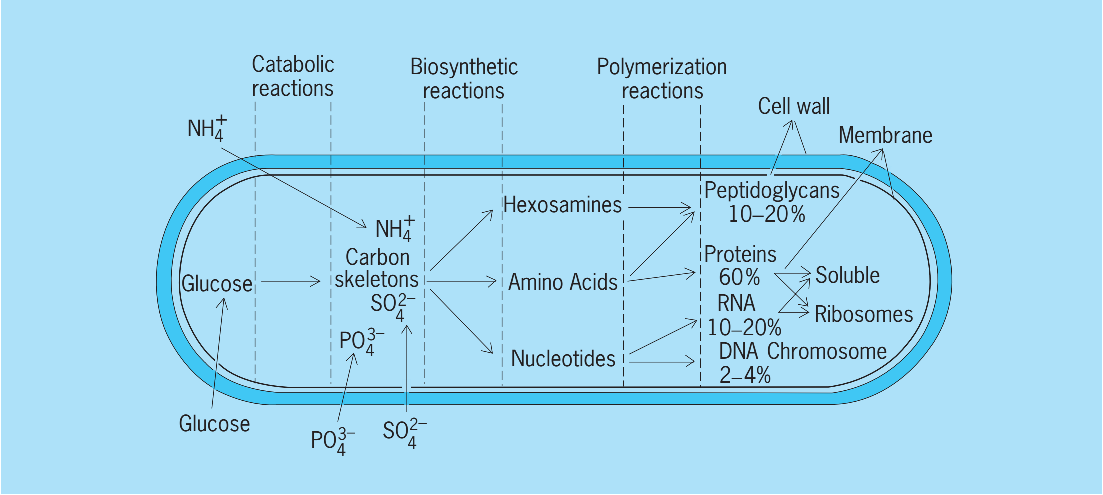


Fig. 48. Generalized flow diagram for the synthesis of the bacterial cell components. (*After J. Mandelstam and K. McQuillen, Biochemistry of Bacterial Growth, 2d ed., Halsted Press* [*John Wiley*], *1973*)

In a larger sense, the reductive pentose pathway, or Calvin cycle, operative in photo- or chemosynthesis may be considered an anaplerotic sequence, for the net result is to provide multicarbon compounds for conversion to building blocks for macromolecular biosynthesis.

Amphibolic pathways. Although initially convenient for purposes of categorization, the terms "catabolism" and "anabolism" may now be seen to have considerable overlap. For example, the pathways of carbohydrate degradation to pyruvate and the Krebs cycle act both to provide energy as well as key biosynthetic intermediates. Those metabolic sequences (**Fig. 47**) which function in this dual capacity are often termed amphibolic, to indicate that they may serve both roles, or under some conditions, either role. This convergence of biosynthetic and catabolic pathways, with common intermediates, is one expression of the close interrelationships which exist between energy generation and consumption in the functioning cell.

Other aspects of anabolism. It is not only multiplying cells that carry out endergonic syntheses. Cells about to enter dormant or resting states (endospores, cysts), into particular stages in a life cycle (stalk forms), or undergoing conversion to carry out specific functions (heterocysts) also undergo particular metabolic or intra- or extracellular structural changes, as do also dormant cells when germinating or others when becoming motile. The energy expenditures connected with these biosyntheses, as well as those connected with N_2 fixation, nitrate reduction, either swimming or gliding motility, photo- and chemotactic behavior, or colonial morphogenesis are no less significant to the economy of the cell than are those which have been enumerated in more detail in the foregoing paragraphs, and summarized in **Fig. 48**.

E. R. Leadbetter

Bibliography. H. A. Barker, *Bacterial Fermentations*, 1957; V. H. Cheldenlin, *Metabolic Pathways in Microorganisms*, 1961; G. N. Cohen, *Biosynthesis of Small Molecules*, 1967; G. N. Cohen, Regulation of enzyme activity in microorganisms, *Annu. Rev. Microbiol.*, 19:105, 1965; B. D. Davis et al., *Microbiology*, 3d ed., 1980; H. W. Doelle, *Bacterial Metabolism*, 2d ed., 1975; I. C. Gunsalus and R. Y. Stanier (eds.), *The Bacteria*, 7 vols., 1960–1979; P. E. Hartman and S. R. Suskind, *Gene Action*, 1965; M. Inouye, *Bacterial Outer Membranes as Model Systems*, 1986; H. A. Krebs and H. L. Kornberg, Energy transformation in living matter, *Ergeb. Physiol.*, 49:212–298, 1957; M.T. Madigan et al., *Brock's Biology of Microorganisms*, 8th ed., 1996; J. Mandelstam et al., *Biochemistry of Bacterial Growth*, 3d ed., 1982; W. G. Murrell and I. R. Kennedy, *Microbiology in Action*, 1988; D. Nelson and M. Cox, *Lehninger Principles of Biochemistry*, 3d ed., 2000; E. L. Oginsky and W. W. Umbreit, *An Introduction to Bacterial Physiology*, 2d ed., 1959; M. J. Pelczar, Jr., et al., *Microbiology*, 5th ed., 1986; J. M. Reiner, *The Organism as an Adaptive Control System*, 1968; H. J. Rogers et al., *Microbial Cell Walls and Membranes*, 1980; M. R. J. Salton, *The Bacterial Cell Wall*, 1964; M. R. J. Salton, *Biochemistry of Bacterial Membranes*, 1975; R. Y. Stanier, E. A. Adelberg, and J. Ingraham, *The Microbial World*, 4th ed., 1976; L. Stryer, *Biochemistry*, 4th ed., 1995; W. A. Volk et al., *Essentials of Medical Microbiology*, 5th ed., 1995; W. A. Volk and M. Wheeler, *Basic Microbiology*, 8th ed., 1997; J. D. Watson, *Molecular Biology of the Gene*, 4th ed., 1987.

Bacterial taxonomy

The classification, nomenclature, and identification of bacteria; sometimes used as a term to indicate the theory of classification. The bacteria are members of the kingdom Prokaryotae, which is defined in terms of the unique structural and biochemical properties of their cells; more specifically, the organization of the deoxyribonucleic acid (DNA) in the nucleus, the lack of a nuclear membrane, the lack of independent membrane-bounded cytoplasmic organelles, the lack of endocytosis and exocytosis, and the chemical nature of some components of plasma membrane and cell walls. *See* BACTERIA.

Bacteria inhabit a great diversity of ecologic niches. The individual kinds of bacteria within these bacterial populations exhibit a variety of physiologic capabilities and accomplish remarkable chemical transformations. The systematic study of the component bacteria in these populations demands recognition of these properties, and a classification of those bacteria that can be cultivated so that they may be assigned to already named species or be described as a new organism to be classified and named.

Classification involves the recognition of similarities and relationships as a basis for the arrangement of the bacteria into taxonomic groups or taxa. The basic taxon is the species. Identification involves the recognition of a bacterium as a member of one of the established taxa, appropriately named, by the comparison of a number of characters with those in the description. *See* TAXONOMIC CATEGORIES.

Species. A bacterial species is a conceptual entity that is hard to define, despite its role as the basic taxonomic grouping. Bacteriologists accept the imprecision and recognize that a species represents a cluster of clones exhibiting some variations in minor properties. They have developed a formal approach to the description of the taxon while trying to solve the problems encountered in the process of recognizing and naming species.

The description is an assembly of such structural, chemical, physiologic, genetic, and ecologic characteristics as can be determined for the available strains that closely resemble each other. A strain is any pure culture of an organism isolated from nature, and the collected strains may then be conserved as cultures in the laboratory for study and comparison. In addition to the description, one strain must be designated by the author and preserved in a culture collection as a type strain, or permanent example, of the species and available to all who study bacteria. If that type strain is lost or succumbs, a formal proposal of a substitute strain (neotype) must be published. In general, bacterial taxonomy is built around the living type specimen: a species consists of the type strain and, whenever available, all other strains sufficiently similar to the type strain to be considered as included in the species. There is a provision for the description and naming of a distinctive species that is not yet cultivable, with the requirement for a suitably preserved type specimen.

A new species, validly described, must be assigned to a genus in order to accord with the binomial system of nomenclature initiated by C. Linnaeus. Thus, a species assigned to the genus *Bacillus* would be referred to as, for example, *Bacillus subtilis*. Such formal names of taxa are italicized to indicate that they are considered to accord with the formal description. If there is no appropriate genus available, a new genus must be named in accord with the *International Code of Nomenclature of Bacteria* and provided with a description that circumscribes the included species, and a type species must be designated as the exemplary representative of the genus.

The lowest nomenclatural rank that is recognized by the *Code* is subspecies, which is a subdivision of the species recognizing consistent variations in otherwise stable characters in the species description, for example, *Bacillus cereus* ssp. *mycoides*. There are times, however, when even finer but unofficial subdivisions of the species are useful and contribute to science, for example, for the epidemiology of pathogenic species. Then, groups of strains may be recognized by some special character as a variety of the species. These may be based on a biological property (biovar), antigenic variation (serovar), pathogenicity (pathovar), or susceptibility to particular bacterial viruses (phagovar). These characters have no formal standing in nomenclature.

Higher taxa. The rules of nomenclature provide for higher taxa, and many have been created and are in use. However, it must be admitted that the relationships implied by their association in a practical scheme of classification, based on resemblances in characters (phenotype), can be very misleading. It may be better to admit that it is not known where some taxa belong. The result is an avoidance of a complete hierarchical classification because of insufficient data. Instead, recognized taxa are grouped within parts or chapters, using vernacular names as titles to express a sharing of substantial and readily determined characters—for example, “gram-negative facultatively anaerobic rods,” or “the gliding bacteria,” or “gram-positive cocci.” Groupings such as these may be diverse in evolutionary or phylogenetic terms, but they are practical for the purpose of identification. It is necessary to include in virtually all such chapters one or more “genera of uncertain affiliation” that cannot be assigned to any family or higher taxon, but at least accord with the broad definition of the chapter. As more is learned to define the relationships at any level, a stable and useful scientific classification of the higher taxa will develop. It has started with the recognition of the Archaebacteria as being phylogenetically separate from all other bacteria. *See* ARCHAEA; BACTERIA.

A provisional arrangement of higher taxa has been proposed (see below), which utilizes some reliable and paramount characters: the cell wall and its constitution (in part recognized by the gram reaction) or its absence, and photosynthesis or its absence.

Kingdom Prokaryotae
  Division I: Gracilicutes (gram-negative bacteria)
    Class I: Scotobacteria—Bacteria that do not utilize light energy and are not phylogenetically related to class II
    Class II: Proteobacteria—Photosynthetic bacteria that do not produce oxygen together with their nonphotosynthetic, phylogenetic relatives
    Class III: Oxyphotobacteria—Photosynthetic bacteria that produce oxygen, including those prokaryotes known as cyanobacteria (formerly called blue-green algae)

- Division II: Firmicutes—Gram-positive bacteria
  - Class I: Firmibacteria—Gram-positive bacteria of simple shapes
  - Class II: Thallobacteria—Gram-positive, branching, rod-shaped bacteria
- Division III: Tenericutes—Bacteria lacking a cell wall
  - Class I: Mollicutes—Single class of Tenericutes, the mycoplasmas
- Division IV: Mendosicutes—Bacteria with walls of unusual composition
  - Class I: Archaebacteria—Bacteria with walls, membrane lipids, and ribosomes of unusual or novel composition (including methanogenic and halophilic bacteria)

It is apparent that, with one exception, shape or form is not used for these ranks. Modern taxonomic studies have shown a number of examples in which morphology is misleading. However, up to now, the gram reaction with the other attendant properties has proved to be relatively reliable in all walled groups, except the class Archaebacteria. Such an arrangement can be used as a bridge between a classification for determinative purposes and a phylogenetic arrangement that will surely arise from current taxonomic research. It is also apparent from the application of nucleic acid sequence data to the determination of relatedness that there are phylogenetic reasons for associating some nonphotosynthetic Gracilicutes with the purple anoxygenic photosynthesizing bacteria. This association defies, so far, a simple phenotypic characterization.

The inclusion of the blue-green algae (the Cyanophyta of the botanical literature) in this taxonomy as cyanobacteria is now less contentious than heretofore. There is general agreement to classify among the bacteria those that can appropriately be activated, described, and typified by bacteriological methods, and to avoid nomenclatural confusion by using whenever possible the names that are also maintained under the botanical code. The blue-green algae are prokaryotic in nuclear organization, have a gram-negative type of cell wall, possess a murein peptidoglycan, and have ribosomes of prokaryotic size and have ribosomal ribonucleic acid (rRNA) that is definitively related to that of the true bacteria. However, many are not yet cultivated and form complex associations or morphological transformations in nature, and are so recognized with preserved specimens as species in the botanical taxonomy. *See* ALGAE.

Each of the classes listed above contains one or more orders and the ranks appropriate to displaying the species included in them. So, for example, a complete taxonomic description of *Neisseria gonorrhoeae*, the cause of the venereal disease gonorrhoea, would be:

Kingdom	Prokaryotae	Prokaryotic
Division	Gracilicutes	Gram-negative
Class	Scotobacteria	Nonphotosynthetic
Order	Not assigned, but grouped with the gram-negative aerobic, rod-shaped, or spherical bacteria	
Family	Neisseriaceae	Included in the family typified by the genus
Genus	*Neisseria*	*Neisseria* Trevisan 1885, 105
Species	*Neisseria gonorrhoea*	*N. gonorrhoea* (Zopf 1885) Trevisan 1885, 106^{AL}
	Type strain, American Type Culture, Collection #19424	

The ascription translates to mean it is the species described by W. Zopf in 1885 and amended by V. Trevisan in 1885 on page 106 of his paper, and the name was included in the *Approved Lists*. A name placed in any rank has a similar detailed ascription.

The higher taxa and the major ranking of organisms within them represent the practical groupings referred to above. But new understanding and changes in approach are producing a steadily accelerating change in bacterial taxonomy. Of greatest importance are systematic studies of the molecular constitution of the major biological polymers and their variations among the prokaryotes.

Nature of characters. Bacterial cells do not exhibit a sufficient variety of reliable (that is, stable) shapes for a useful classification or to make the fragmentary fossil record effective for integration into the higher taxa assigned to existent organisms. There is microscopic evidence of their presence and geological evidence of some of their activities for as much as 3.8×10^9 years before present. But all that can be observed are simple shapes; the rest is inference. Many fossil forms have been named, but usually in accord with the rules of the *Botanical Code*.

Reproduction is almost exclusively vegetative (either by binary fission or by budding); life cycles are absent or infrequent and simple in nature; gametes are not formed and conjugation is not a regular function in reproduction. The determinative features useful in the classification of plants and animals are not available. Bacteriologists were then forced to classify on the basis of any character that could be observed: shape, staining properties, motility, growth requirements, substrate utilization, fermentation products, reserve substances, enzymes, toxins, and so on. The number of characters so generated is rather great and, unfortunately for taxonomists, the selection determined for particular organisms was often dictated by the interests of the bacteriologist; for example, organisms of interest to the medical bacteriologist were subjected to tests that differed from those applied to similar organisms studied by plant pathologists, and vice versa. However, it is realized now that the great diversity of applicable tests represents an extensive sampling of the genome. In a sense this is one form of genetic analysis, and it records the expression of available genes, the phenotype. Formal genetic analysis has been applied to only a few species of bacteria, but in many ways this function is now served by study and comparison of the sequences of nucleotides in the nucleic acids, by recognition of structural similarities in proteins, or by recognition of the functional similarities in metabolic pathways. The results of such comparisons may be interpretable in terms of relationships and phylogeny (evolutionary sequence). These newer approaches must now be discussed, because in them lie the seeds of change. *See* BACTERIAL GENETICS.

Modern approaches. Several new techniques are presently used in modern approaches to taxonomy.

Numerical taxonomy. Numerical taxonomy (taxometrics) is a first approach for the analysis of phenotype. It requires a larger number of tests (say 150) on each of a number of bacterial strains under investigation, together with a collection of type strains and veritable reference strains (preferably not less than 50), both closely and remotely resembling the grouping under study. An analysis of the similarities and dissimilarities requires the comparison of each of the strains (Operational Taxonomic Units or OTUs) with each of the other strains. The data can then be assembled in a similarity table based on the count, for any pair of OTUs, of the number of characters in which they are identical (positive or negative) expressed as a percentage. Other matching coefficients are possible, but this simple matching (S_{SM}) coefficient is the one most often used in bacteriology. Such analyses only became possible with the advent of computers, which allow the table of similarities to be further manipulated to enhance the appreciation of taxonomic structure. The clusters of closely similar OTUs can be assembled in a treelike diagram or dendrogram, in which the branch points represent the S_{SM} that separates each pairing of clusters. Another useful expression of the results is a sorted similarity matrix, in which the clusters are arranged in a two-dimensional format and the similarity is expressed by shading, allowing the relative proportion, size, and separation of the clusters to be recognized by eye. The major separable similarity clusters thus generated are termed phenons, which may or may not be equated with a taxon. But most phenons formed at the 80% similarity level include all or most of the representatives of a species, and so have taxonomic usefulness and form a method for objective analysis of phenotype. It is a useful exercise for the assessment of taxa, and a basis for further and more elaborate statistical analysis. Interpretation, however, requires well-founded scientific judgment despite the overt intention to apply the adansonian principle of not assigning undue weight to any character. In fact, one of the useful returns of taxometric analysis is the formal assessment of the reliability of selected characters for classificatory keys and identification.

Numerical taxonomy implies the existence of programs for computer-assisted identification, either as recognizable phenons or by relation into a computer-stored classification and key program. Either of these methods has found considerable application in dealing with masses of isolates (for example, in studies of pollution or of sediments), or in dealing with results of automated systems for identification of pathogens in clinical bacteriology. The artificial keys and the reference data included in the memory for computer analysis can be infinitely more complex than is practical for the unaided bench worker. Identification can include alternative diagnostic possibilities and a probability estimate. Thus, intuitive judgment in practical bacteriology can be supplemented with artificial keys, and in academic bacteriology with attempts to develop natural or phylogenetic classifications. *See* NUMERICAL TAXONOMY.

Chemotaxonomy. Chemotaxonomy applies systematic data on the molecular architecture of components of the bacterial cell to the solution of taxonomic problems. This has been a powerful tool since the 1950s, and a number of chemotaxonomic markers have been identified, ranging from molecules unique to the Prokaryotae (for example, the murein peptidoglycans) or specific groups of bacteria (for example, the ether-linked lipids of the Archaebacteria) to mechanisms or products of metabolism that characterize genera or species. The availability, and relative simplicity of techniques for amino acid analysis, for sequential analysis of polymers, for gas and thin-layer chromatography, for fermentation products and lipids, and so on, have made the systematic studies possible. These have led to a more effective definition of taxonomic groups based on biochemical assessment of cell wall composition, lipid composition of membranes, the types of isoprenoid quinones, the amino acid sequences of select proteins, and the characterization of proteins, such as the cytochromes and many other macromolecules.

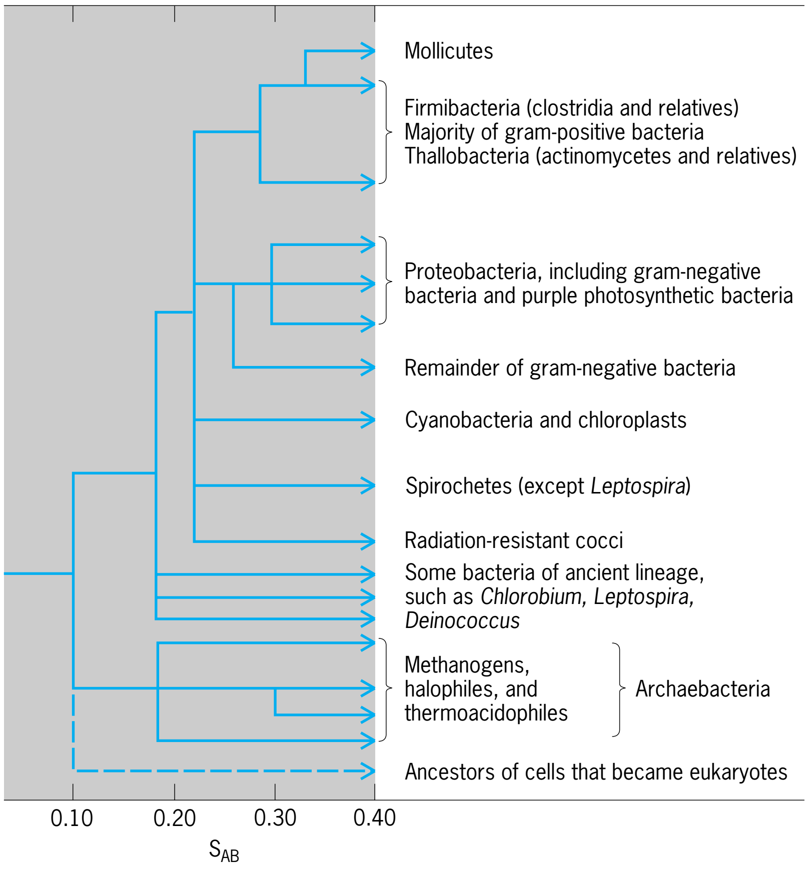
Of all the various chemotaxonomic approaches, the most exemplary is that based on the chemistry of the peptidoglycans of the cell walls. These heteropolymers are unique to the Prokaryotae and are possessed by all taxa except those in the Tenericutes and the Mendosicutes, which have no walls or strangely constructed walls. The great majority of genera are consistent in the amino acids that make up the pentapeptides that are associated with the amino-sugar polymeric backbone, and the simpler peptides that cross-link the strands into a strong network. But some genera and groups, especially among the gram-positive Firmicutes, have peptidoglycan types defined by distinctive diamino acids and amino acids forming their cross-links. These are now found to be consistent with phylogenetic groups recognized on the basis of RNA analysis. *See* CHEMOTAXONOMY.

Nucleic acid studies. These studies have been by far the most potent generators and arbiters of data on relatedness, with distinct capability for applications to the phylogenetic assessment of taxonomic arrangements. *See* DEOXYRIBONUCLEIC ACID (DNA); NUCLEIC ACID; RIBONUCLEIC ACID (RNA).

The DNA base composition, which is most simply expressed as the content of one of the base pairs (as mol % guanine + cytosine), was first applied to species and genera. It is useful information to this extent: an excessive spread of values for a species or a genus should warn the taxonomist that the group is probably more complex than a single taxon; a strain that is divergent in mol % guanine + cytosine from the type and reference strains is probably misidentified. Conversely, similar values are no measure of identity because the overall composition cannot be related to the sequential arrangement of the nucleotides.

Nucleic acid homology experiments are designed to give a measure of the overall similarity in the sequence of the bases in DNA or RNA from one organism compared to that from another organism. This can be done with genomic DNA because the complementary strands can be dissociated into single strands by an appropriately poised high temperature, and reassociated at a lower temperature with any single strands having appropriate sequences, either from the same organism or a related one. The reassociation is proportional to the overall similarity of the sequences. This is usually measured as a percentage of the uptake of radioactively labeled DNA from the test organisms compared to the uptake of labeled homologous DNA by the reference DNA (taken as 100%). There are several well-established methods which give comparable results. Single-stranded RNA molecules will also hybridize with a complementary DNA strand, and the degree of association can be similarly assessed. Consequently, DNA/DNA association allows the use of the whole genome for assessment of the genetic similarity of two organisms. A much smaller fraction of the genome is assessed by DNA/RNA homology, using either rRNA or transfer RNA (tRNA), because relatively few cistrons determine those RNA sequences. However, evolutionary changes in DNA (and messenger RNA, or mRNA) sequences have been far more rapid than in those of rRNA or tRNA, probably because the latter determine the assembly and function of a crucial and evolutionarily stable organelle, the ribosome, which has recognizable and only somewhat variable elements throughout the kingdom. As a result, DNA/DNA homology is useful only for assessing relatedness within a genus and for establishing or separating species (better than 70% homology) and subspecies. DNA/RNA homology is more useful in the assessment of the relationships of genera because the homologous elements involved have been more highly conserved over that evolutionary period. The experience with these techniques reinforced the realization that exceptionally stable determinants, universally available in all kinds of cells, are needed for the assessment of relatedness encompassing the broad scope of biological evolution.

Biochemical phylogenetic analysis is a thrilling development in bacterial taxonomy, and has been markedly stimulated by the introduction of methods allowing assessment of similarities in the 16S rRNA sequences of diverse taxa (see **illus.**). The 16S RNA is chosen rather than the 5S or the 23S because it is neither too large for analysis nor so small as to have limited information content. The approach has been developed and applied to many bacteria by C. R. Woese and his collaborators. In essence, the radioactively labeled and isolated 16S cRNA is cut into small nucleotides with T_1 ribonuclease. Each of these nucleotide sequences ends in a guanine residue and can be recognized by its radioactivity and position after two-dimensional electrophoresis. The very short pieces (five residues or less) have no useful comparative value, but those of six residues or more can be identified as to sequence and recorded in a catalog of oligomers. Each catalog of oligonucleotides is then compared individually to all other catalogs to provide a similarity coefficient (S_{AB}) for each pair of organisms. The table of S_{AB} values is then subjected to cluster analysis, giving the average linkage S_{AB} value between merged groups, and this can be expressed in the form of a dendrogram (see illus.). The most distantly related have about 50 oligonucleotides larger than hexamer in common (S_{AB} about 0.10), and the closely related have as many as 400 in common. It can be assumed that the differences in the average linkage S_{AB} values approximate to the phylogenetic separation, so that groups of organisms separated by deep clefts (that is, low S_{AB}) in the dendrograms have evolved independently for a long time. The precise relationship to real time is uncertain. The result is that an early divergence of the Archaebacteria from the mainstem of prokaryotic evolution can be identified. A third major line of


Mollicutes
Firmibacteria (clostridia and relatives)
Majority of gram-positive bacteria
Thallobacteria (actinomycetes and relatives)
Proteobacteria, including gram-negative bacteria and purple photosynthetic bacteria
Remainder of gram-negative bacteria
Cyanobacteria and chloroplasts
Spirochetes (except Leptospira)
Radiation-resistant cocci
Some bacteria of ancient lineage, such as Chlorobium, Leptospira, Deinococcus
Methanogens, halophiles, and thermoacidophiles
Archaebacteria
Ancestors of cells that became eukaryotes
0.10
0.20
0.30
0.40
SAB

Composite dendrogram showing the major lineages of prokaryotes deduced from 16S rRNA oligonucleotide catalogs.

descent from the ancestral clones would be the as yet unidentified cell types that were later parasitized by chloroplast or mitochondrial prokaryotes and became the ancestors of eukaryotic plants and animals. Woese has called these three lineages "primary kingdoms"; however, others think that this is a premature assignment.

The lineages of the main phylogenetic groupings of the extant true bacteria (that is, all except the Archaebacteria and a few separate and insufficiently studied groups) go back to a major separation expressed as an S_{AB} value of about 0.22. This means that much higher resolution will be required to estimate the exact order of branching of the evolutionary lines. For the moment this is of small consequence; the fascinating and exciting fact is that a phylogenetic order can be established. The technique (albeit somewhat difficult and expensive at this time) is applicable to all the kingdoms, not just the Prokaryotae, and it gives useful data concerning the relationships of genera within groups as well as the genealogy of the groups.

Bacterial taxonomy in transition. Therefore, current taxonomic schemes are departing from the entirely artificial arrangements based on expressed phenotype, and are moving toward the ideal of a natural order expressing evolutionary relationships. This revolution derives from the developments in molecular biology and the availability of computer techniques to generate, systematize, and compare the data on molecular sequences. It depends on recognizing that the evolutionary record is preserved in the sequences of the component molecules of the most stable and most slowly changing complex macromolecules. These sequences are the "semantides" of E. Zuckerkandl and L. Pauling, which they called "the documents of evolutionary history." Few macromolecules can be expected to show the whole story; some of them express subtle systematic shifts of molecular arrangement in some groups of organisms (for example, in the structure of cytochrome *c*). Others show distinctive molecular changes in each of a number of arborizations of a specific major taxon (for example, the peptidoglycan types of the gram-positive bacteria) when other major segments of the classification show no clear distinctions in the same class of macromolecule; other information-laden macromolecules shift too much in sequences and in meaning (for example, chromosomal DNA and mRNA) to be more than informative about that species or genus; and only a few (notably the rRNA molecules) are so stable in evolutionary time that some resemblances can be followed and put into phylogenetic order, whatever the origin. It is fortunate that a large part of the rRNA molecules have to be stable and change only slowly with time in order to preserve the effective and vital organization of ribosomal proteins for effecting protein synthesis from coded nucleotide determinants.

It is now evident that a phylogenetic arrangement of taxa will be accomplished because effective arrangement of many of the broader groups is possible. In fact, the evolutionary resolution will improve, as will the capability (already evident) of the same sequence data to resolve inter- and intrageneric relationships. The integration of the data from 16S rRNA and from 5S rRNA will eventually be joined, as techniques of analysis and data handling improve, by data from the larger and more complex 23S rRNA. Other sources of evolutionary data are coming out of the study of proteins and enzyme pathways, and even if results are not as broad in scope as that from rRNA, they will (and do) provide powerful assistance in the resolution of problems presented by specific taxonomic groups, both small and large. So far, it is encouraging to find a correlation between the different approaches to molecular phylogeny of the bacteria.

Molecular systematics reveals genealogical associations that run counter to the arrangements in the artificial keys based on morphological and physiological characters. For instance, the 16S rRNA catalogs suggest that some species of *Micrococcus* are more closely related to the genus *Arthrobacter* than to other species of *Micrococcus*. Also, there are three well-separated phylogenetic groups of purple photosynthetic bacteria, and each of these shows clear relationships with representatives of a variety of well-known gram-negative genera, some of which are represented in at least two of the groupings. There is the strange possibility (one should remember the prejudice that photosynthesis was an early and unique development) that a great variety of nonphotosynthetic bacteria derived from photosynthetic stock. The resolution of these and other strange associations will require not only confirmatory data using different systems of analysis but the inclusion of many more reference cultures than has been possible up to now. But it is clear that the traditional characters for classification of morphology and physiology present some taxonomic hazards, despite their usefulness in diagnostic bacteriology.

It has to be recognized that a practical classification or a key that enables the identification of an organism may not be (and does not have to be) capable of expressing phylogeny. The essential thing is to be able to arrive at a correct diagnosis. In fact, at this time, the characters (sequences and the like) required for a phylogenetic assignment are too complex and slow in accomplishment for the practicalities of determination. So it is entirely possible that both practical and academic classifications will be needed; and in a way that time has already come. Taxonomy can and must serve all the requirements of bacteriologists and integrate new knowledge to best effect understanding and practical needs. *See* BACTERIAL PHYSIOLOGY AND METABOLISM; BACTERIOLOGY.

R. G. E. Murray

Bibliography. M. J. Carlile, J. F. Collins, and B. E. B. Moseley (eds.), *Molecular and Cellular Aspects of Microbial Evolution*, 1981; P. Gerhardt et al. (eds.), *Methods for General and Molecular Bacteriology*, 1993; J. G. Holt and N. R. Krieg (eds.), *Bergey's Manual of Systematic Bacteriology*, vol. 1, 9th ed., 1993; S. P. Lapage et al. (eds.), *International Code of Nomenclature of Bacteria*, 1975; K. H. Schleifer and E. Stackebrandt, Molecular systematics of prokaryotes, *Annu. Rev. Microbiol.*, 37:143–187, 1983; V. B. D. Skerman et al. (eds.), *Approved Lists of Bacterial Names*, amended ed., 1989; P. Sneath (ed.), *Bergey's Manual of Systematic Bacteriology*, vol. 2, 1986; J. T. Staley (ed.), *Bergey's Manual of Systematic Bacteriology*, vol. 3, 1989; S. T. Williams (ed.), *Bergey's Manual of Systematic Bacteriology*, vol. 4, 1989; C. R. Woese, Archaebacteria, *Sci. Amer.*, 244:98–122, 1981; E. Zuckerkandl and L. Pauling, Molecules as documents of evolutionary history, *J. Theoret. Biol.*, 8:357–366, 1965.

Bacteriology

The study of bacteria. While bacteria are different in some important respects from all other kinds of cells, their basic processes of physiology and genetics are the same as in all forms of life. One unusual property of bacteria as a whole is their physiological diversity: Some live in the total absence of oxygen and convert complex carbohydrates to acids and alcohols (fermentation), sulfate to hydrogen sulfide, nitrate to nitrogen gas, and hydrogen plus carbon dioxide to methane gas; others carry out photosynthesis by mechanisms nearly identical to plants; some bacteria can grow and multiply by using energy obtained from oxidation of sulfur, ammonia, hydrogen, or iron while obtaining carbon for cell synthesis from carbon dioxide; and some can obtain their needed nitrogen from the gas in air. *See* BACTERIA; BACTERIAL PHYSIOLOGY AND METABOLISM; FERMENTATION; METHANOGENESIS (BACTERIA).

Humans and the planet Earth are profoundly affected by bacteria. There is solid evidence that oxygen first appeared in the air some 2 billion years ago as a result of the activity of bacteria. The bacteria are critically important in the recycling of materials essential to plants and animals. The degradation of complex substances such as carbohydrates, proteins, and lipids to carbon dioxide allows plants to grow. Conversion of ammonia to nitrate and of nitrogen gas to amino acids is also essential to plants. Great numbers of bacteria live on human skin surfaces as well as in the mouth and intestinal tract; most of these are benign or even beneficial. Some bacteria, given the opportunity, can cause severe diseases of humans, other animals, and plants.

Spontaneous generation. The discovery of bacteria was made by Antony van Leeuwenhoek in 1676. As a result of examination of teeth scrapings and of an infusion of pepper (the purpose of the latter was to learn if hotness of pepper was due to something biting the tongue), van Leeuwenhoek saw and accurately described microscopic organisms that are now known to be bacteria. However, scientists did not realize the significance of bacteria for 200 years. The idea that these tiny organisms might be involved in disease occurred to some early scientists, but, curiously, the most intense interest and controversy during this period was over the question of where bacteria came from.

A group of scientists, particularly J. T. Needham and G. L. de Buffon, showed that a boiled infusion of meat would be teeming with bacteria when allowed to stand in a closed container for a day or two. They, and others, concluded that bacteria arose spontaneously from this dead organic matter. In 1776, Lazarro Spallanzani demonstrated that more careful boiling of the infusions followed by immediate hermetic sealing of the vessels produced a broth that remained free of organisms for days and weeks. In spite of this seemingly convincing demonstration, the battle of spontaneous generation in infusions continued for another hundred years. The question was finally settled by the careful and tedious experiments of Louis Pasteur and John Tyndall. They proved conclusively that life begets life and that bacteria do not originate spontaneously from dead organic matter.

It is important to note that the current consensus of scientific thinking is that bacteria were the first forms of cellular life on Earth and that they arose spontaneously as the culmination of a long period of chemical evolution. The important point is that life arose spontaneously in a very special primordial atmosphere over a 1- to 2-billion-year period and not overnight from organic materials such as a heated meat infusion.

Bacteria and disease. One of the most important events in the history of civilization was the discovery that bacteria cause disease. In 1876, Robert Koch provided convincing proof that anthrax, an important disease of cattle and sometimes of humans, was caused by bacteria. Koch had already developed methods for growth and isolation of pure cultures of bacteria. To prove that bacteria cause disease, he first observed rod-shaped bacteria in the blood of anthrax-diseased animals and then isolated the organism as a pure culture from the blood. Next, the bacteria were injected into a healthy animal, and when the symptoms of anthrax appeared, the same kind of bacteria were once again isolated. This series of procedures—observation and isolation of an organism from an animal suffering from a disease, inoculation into a healthy animal with development of the original symptoms, and reisolation of the same kind of organism—became known as Koch's postulates. These have since been accepted as the criteria necessary to prove the etiology of diseases not only caused by bacteria but also by viruses, protozoa, and fungi.

Soon after Koch's monumental discovery, Pasteur, while studying cholera in chickens, discovered the basis for acquired immunity to disease. He found that old cells of the disease-causing bacteria became attenuated, or changed in such a way that when they were inoculated into a healthy chicken the disease did not occur. Moreover, injection of young virulent bacteria no longer caused the disease. The chicken had become immune to cholera.

These discoveries by Koch and Pasteur were the impetus for a subsequent golden era of study of the cause and control of bacterial disease of humans and of immunization as a means to combat diseases. A result of the work in this era was a tripling of the life expectancy of humans between 1876 and 1926. *See* IMMUNITY; MEDICAL BACTERIOLOGY.

Models for biochemistry and genetics. At the molecular level, the basic physiological processes associated with metabolism and growth are nearly the same for all forms of life. Because of their rapid growth rate (some bacteria divide every 20–30 min), and the ease of obtaining large amounts of material in a controlled atmosphere in a short time, bacteria were favorite subjects for studies of biochemistry and cell physiology. Much of the current knowledge of the properties and activities of enzymes, the mechanism and control of degradation of foodstuffs, and the biosynthesis of proteins, nucleic acids, carbohydrates, and lipids was gained from studies of bacteria. *See* BACTERIAL GROWTH; BIOSYNTHESIS.

The same properties that made bacteria useful for biochemical studies made them ideal subjects for studies of genetics. The ability to obtain population increases from a few hundred to millions of cells overnight and in a test tube made studies of mutation and selection feasible. It was studies of behavior of populations of bacteria that led Oswald Avery, M. McCarty, and C. M. MacLeod to the proof that deoxyribonucleic acid (DNA), and not protein, is the cell macromolecule of which genes are made. *See* GENETICS.

Relationship to other life forms. Bacteria were considered to be primitive plants until approximately 1955, when improved technology of electron microscopy made it possible to observe their intracellular structure. The organisms were embedded in plastic, then cut into slices as thin as 0.1 micrometer. The intracellular structures of bacteria were found to be much different and simpler than all other cells, both plant and animal. Bacteria have no discrete, membrane-bound nucleus and no special intracellular membrane-bound organelles. Animal and plant cells contain organized nuclei and organelles called mitochondria, and plant cells contain chloroplasts. The ribosomes, particulate structures involved in protein synthesis, are smaller in bacteria than are the ribosomes in the cytoplasm of plant and animal cells. From such observations, it became clear that bacteria are not plants but are as different from plants as plants are from animals. C. B. Van Neil and R. Y. Stanier in 1955 formalized these observations and ideas into a classic paper, “The Concept of a Bacterium.” The bacteria were to be considered a completely different evolutionary line from other life. Bacteria were called prokaryotic cells, meaning primitive or primordial; all other cells were termed eukaryotic, denoting more advanced or true cells. *See* CELL (BIOLOGY).

Biochemical evolution. The idea is now generally accepted among biologists that the first cells to originate on Earth were prokaryotic. These arose spontaneously some 3–3.5 billion years ago in an environment completely devoid of free oxygen. Geochemical evidence verifies that the Earth was anaerobic (there was no oxygen in the atmosphere) until 2–2.5 billion years ago. The prokaryotic cells went through a period of biochemical evolution for a billion years or so, wherein they adapted by mutation and selection to become increasingly adept at obtaining substrates and energy for growth. It was the evolution of oxygenic photosynthetic bacteria some 2.5 billion years ago that brought about the appearance of free oxygen in the air. Bacteriologists theorize that much of the biochemical diversity of present-day bacteria represents processes that have changed but little for billions of years. Thus, some anaerobic bacteria found in soil, sediments, and intestines where there is no free oxygen are thought to be living fossils of early biochemical processes that occurred before free oxygen existed. *See* ATMOSPHERE, EVOLUTION OF; PREBIOTIC ORGANIC SYNTHESIS.

Bacteria and endosymbiosis. The theory now most widely accepted for explaining the origin of eukaryotic cells is most intriguing. Comparisons of the basic physiological properties and cell structure of eukaryotic and prokaryotic cells continued to show how different the cells were. A surprising result of these studies was the demonstration of how similar bacteria are to the eukaryotic cell organelles, that is, mitochondria and chloroplasts. The ribosomes of the bacteria and of the organelles are the same size, and function virtually identically in protein synthesis. The nuclear genetic material of bacteria and of the organelles was found to be organized in similar fashion. In fact, the genetic material, DNA, of organelles is more similar to DNA of bacteria found in nature than it is to the DNA comprising the central nuclear region in the cytoplasm of the eukaryotic cells.

These observations are the basis for the endosymbiosis theory of evolution of eukaryotic cells. The theory suggests that some prokaryotic cells during the period of the primordial anaerobic atmosphere were predators. These cells became larger and engulfed and digested smaller cells. Eventually, some of the engulfed cells were not digested but instead took up permanent residence in the cytoplasm and supplied energy to the host cell while remaining in this protected environment. The cell line with a symbiotic relationship with respiratory bacteria developed into animals. Another cell line, with a symbiotic relationship with both the respiratory bacteria and oxygenic photosynthetic bacteria, became plants. The purpose of mitochondria in animals and plants is to produce energy by coupling oxidation of food to reduction of oxygen to water, a process called respiration. The function of chloroplasts is photosynthesis—the conversion of solar energy to biologically useful energy with formation of oxygen. If the theory of endosymbiosis is correct, bacteria not only are all around but are an essential and integral part, as mitochondria, of every cell of the human body. *See* CELL PLASTIDS; MITOCHONDRIA.

Archaebacteria. Biologists had barely become comfortable with the concept of three major avenues of cellular evolution (prokaryotic cells and eukaryotic plant and animal cells), when Carl Woese discovered the existence of microorganisms that do not fit into this scheme. In an attempt to develop a phylogenetic history of evolution, Woese focused on the constituent nucleotide bases of the ribosomal 16-S ribonucleic acid molecule. He reasoned that this molecule is present in all cells and performs the same function for all. Changes, caused by occurrence of mutations, in the kinds and sequences of the nucleotide sequence would be expected to provide a record of genetic relatedness with respect to evolutionary time. Very different organisms should have very different 16-S RNA nucleotide sequences, while similar groups of organisms should have fewer differences.

The initial data verified the theory and reinforced the concept that prokaryotic cells are more ancient than eukaryotic cells. However, some anaerobic methane-forming bacteria were found to have 16-S RNA with a greatly different nucleotide sequence from that of other prokaryotic cells. Woese deduced that these methane-forming bacteria are as different from the prokaryotic bacteria as are prokaryotic cells from eukaryotic plant and animal cells. These different organisms were named archaebacteria and were considered to be a fourth major evolutionary cell line that originated very early.

Some thermophilic bacteria, growing at near-boiling temperatures, and halophilic bacteria, growing at saturated salt concentrations, were also found to be archaebacteria. Physiological studies of members of this group of bacteria have verified their difference from other bacteria. The true bacteria are surrounded by cell walls containing an unusual sugar—muramic acid—and their membranes are composed of ester-linked straight-chain aliphatic fatty acid molecules. Archaebacteria cell walls contain no muramic acid, and their membranes are unique in nature in being made of ether-linked branched aliphatic fatty acids. The current concept of bacteria is that they actually belong to two widely different and evolutionary distinct groups, the eubacteria, or true bacteria, and archaebacteria. Both are considered to be prokaryotic cells. *See* ARCHAEA.

Biotechnology and genetic engineering. Bacteria are the basis of many important industrial processes. They are involved in production of cheese, fermented food products such as sauerkraut, pickles, and sausage, and formation of methane gas from sewage and other wastes. Most medically important antibiotics are produced by bacteria. Exploitation of bacteria for detoxification of environmental pollutants and for production of useful materials promises to increase in the future. New industries are based on biotechnology, where bacteria play a key role. *See* ANTIBIOTIC; FOOD MICROBIOLOGY; INDUSTRIAL MICROBIOLOGY.

The basis for the new biotechnology is genetic engineering. This involves the ability to transfer specific genes from one kind of organism to another. As a consequence, new biological properties are produced, or engineered. It was discovered that many bacteria possess small pieces of DNA called plasmids that exist in the cell cytoplasm and are not associated with the cell chromosome or nucleus. Plasmids generally contain genetic information for a relatively few cell properties. Often, the ability of bacteria to become resistant to antibiotics is due to plasmid genes. The plasmids are capable of being transmitted rapidly between bacterial populations. Plasmids have been successfully isolated from one kind of bacterium and then inserted into a different kind of bacterium. Often, the cell receiving the plasmids acquires properties from the donor cell due to the plasmid-borne genetic information.

In order to genetically engineer a cell, the plasmid molecule is cut by using a specific enzyme called restriction endonuclease. Next, the genetic material of another cell containing information for formation of the desired product is cut into fragments by the endonuclease. The plasmid and gene-fragment molecules are mixed together and treated with an enzyme that joins fragments of DNA together. The result is reformation of the plasmids containing pieces of the genetic information of the cell. The recombinant plasmid is next introduced into a bacterium, usually the one from which it was originally isolated. The bacteria are allowed to grow and multiply, and analyses are made to determine if the desired property is being expressed.

The gene for production of insulin has been transferred in this way from animal cells into the bacterium *Escherichia coli*, and insulin is being produced on a commercial scale by using the bacteria. The amino acids phenylalanine and aspartic acid that constitute the sugar substitute aspartame are produced by genetically engineered bacteria. The ability to transfer specific genes from one kind of organism to another has opened virtually unlimited possibilities for creation of new kinds of bacteria tailored to perform some useful function. It is now possible to consider large-scale production of anticancer and antiviral agents and of hormones that are present in the human body in infinitesimally small amounts on a commercial scale by using genetically engineered bacteria. *See* BACTERIAL GENETICS; GENETIC ENGINEERING.

J. C. Ensign

Bacteriophage

Any of the viruses that infect bacterial cells. They are discrete particles with dimensions from about 20 to about 200 nanometers. A given bacterial virus can infect only one or a few related species of bacteria; these constitute its host range. Bacteriophages consist of two essential components: nucleic acid, in which genetic information is encoded (this may be either ribonucleic acid, in the RNA phages, or deoxyribonucleic acid, in the DNA phages), and a protein coat (capsid), which serves as a protective shell containing the nucleic acid and is involved in the efficiency of infection and the host range of the virus.

The description of a bacterial virus involves a study of its shape and dimensions by electron microscopy (**Fig. 1**), its host range, the serological properties of its capsid, the kind of nucleic acid which it contains,

-S

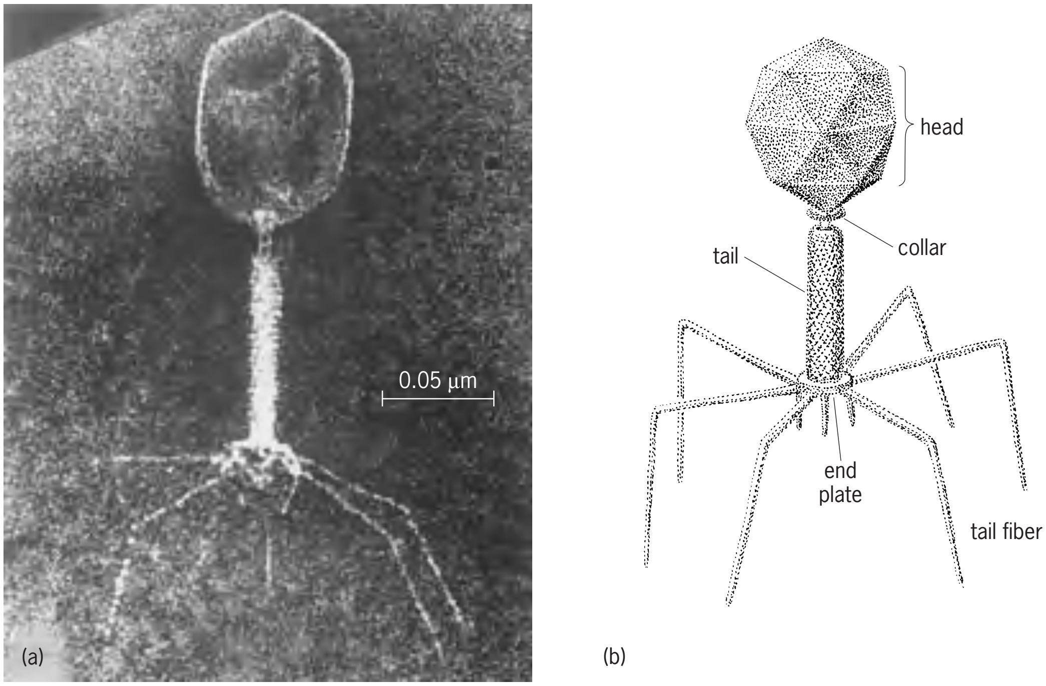


Fig. 1. Bacteriophage structure. (*a*) Electron micrograph of negatively stained T2 bacteriophage showing head and tail components (*courtesy of H. Fernandez-Moran*). (*b*) Diagram of similar T4 bacteriophage.

and the characters of the plaques which it forms on a given host.

Both the nucleic acid and the capsid proteins are specific to the individual virus; in the case of the capsid proteins this specificity is the basis for serological identification of the virus.

The most striking form of phage infection is that in which all of the infected bacteria are destroyed in the process of the formation of new phage particles. This results in the clearing of a turbid liquid culture as the infected cells lyse. When lysis occurs in cells fixed as a lawn of bacteria growing on a solid medium, it produces holes, or areas of clearing, called plaques (**Fig. 2**). These represent colonies of bacteriophage. The size and other properties of the plaque vary with individual viruses and host cells (**Fig. 3**).

The most systematically investigated of the bacteriophages are a series of DNA phages designated T1 through T7, which lyse strains of the gram-negative bacterium *Escherichia coli* and its relatives. These are referred to as coliphages. Certain facts established with these viruses and their hosts have had general application for the study of all viruses. *See* COLIPHAGE.

The effect of phage infection on a population of susceptible host cells is not always the same. RNA phages invariably lyse the cells they infect, with the production of new phage progeny. A few cells in a population may survive; these are found to be mutants resistant to phage infection. DNA phages are of two types. The obligately lytic phages, like the RNA phages, give rise following infection to productive lysis, and surviving bacterial cells prove to be resistant mutants. Other DNA phages, called temperate, are capable of a more complex relationship with their host cells. Infection of a population by these phages results in productive lysis of most of the bacteria, but the few survivors are now of two types: (1) the noninfected resistant mutants and (2) cells in which the phage DNA is integrated in the bacterial

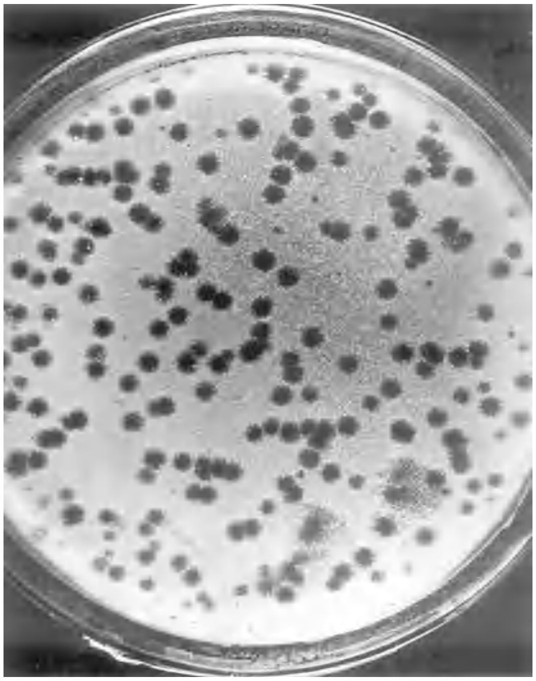


Fig. 2. Plaques of *Rhizobium* bacteriophage; clear, dark areas produced by lysis of infected bacteria. (*From J. Kleczkowska, The production of plaques of Rhizobium bacteriophage on poured plates and its value as a counting method, J. Bacteriol., 50:71–79, 1945*)

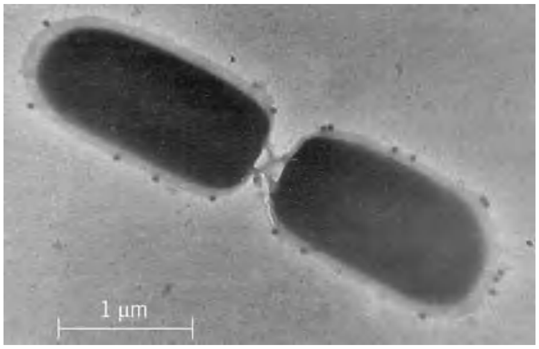


Fig. 3. Dividing *Escherichia coli* cells with 19 particles of coliphage adsorbed to the cell walls. (*From S. E. Luria, M. Delbrück, and T. F. Anderson, Electron microscope studies of bacterial viruses, J. Bacteriol., 46:75,1943*)

chromosome, replicates with it, and endows the bacterial cell with new properties. *See* LYSOGENY.

The lytic cycle involves five steps: (1) adsorption, which involves a receptor site (a specific molecular configuration) on the surface of the host bacterium which reacts with an equally specific taillike structure of the phage particle (Fig. 3); (2) penetration, in which the germinal substance (nucleic acid) of the virus reaches the interior of the bacterium; (3) multiplication, in which the genetic information encoded in the nucleic acid is transcribed and translated to direct the synthetic processes of the cell toward the making of phage components; (4) maturation, involving the assembly of the newly synthesized phage components into mature particles, with nucleic acid enclosed in capsids; (5) lysis of the infected cell with liberation of phage progeny. The specific details of productive lysis differ among the coliphages and other phages as well. The time from infection to lysis with several of the T phages is as short as 13 min. For other phage-host cell systems the latent period is longer; under standard conditions it is characteristic for each system. The average number of new infectious particles produced by *E. coli* infected by a T phage is from 100 to 300 for each infected cell. The average yield for other combinations may be lower or higher: Some corynebacteriophage (*Corynebacterium diphtheriae*) systems produce an average of only 30 virus particles for each infected cell, while *E. coli* infected with some RNA phages may produce more than 10,000 progeny in each host cell. *See* LYTIC INFECTION.

Rare mutants occur among the virus particles produced by phage infection. Mutants of temperate phages, known as virulent, are no longer capable of giving rise to lysogenic bacteria but are obligately lytic. Other mutants are detected by extensions of host range among related bacteria, plaque characteristics, ability to multiply at elevated temperatures, and so on. Related bacteriophages and their mutants are capable of undergoing genetic recombination in mixed infected cells, allowing the construction of genetic maps of a number of bacterial viruses. Studies of the translation of genetic information and its products during the process of phage multiplication have been of inestimable value in establishing the various interactions of DNA, RNA, and enzymes in biosynthetic processes. *See* ACTINOPHAGE; BACTERIAL GENETICS; RECOMBINATION (GENETICS); VIRUS.

Lane Barksdale

Bibliography. E. A. Birge, *Bacterial and Bacteriophage Genetics: An Introduction*, 1983; J. Douglas, *Bacteriophages*, 1975; M. S. DuBow, *Bacteriophage Assembly*, 1981; R. Hendrix et al. (eds.), *Lambda II*, 1983; G. S. Stent, *Molecular Biology of Bacterial Viruses*, 1963.

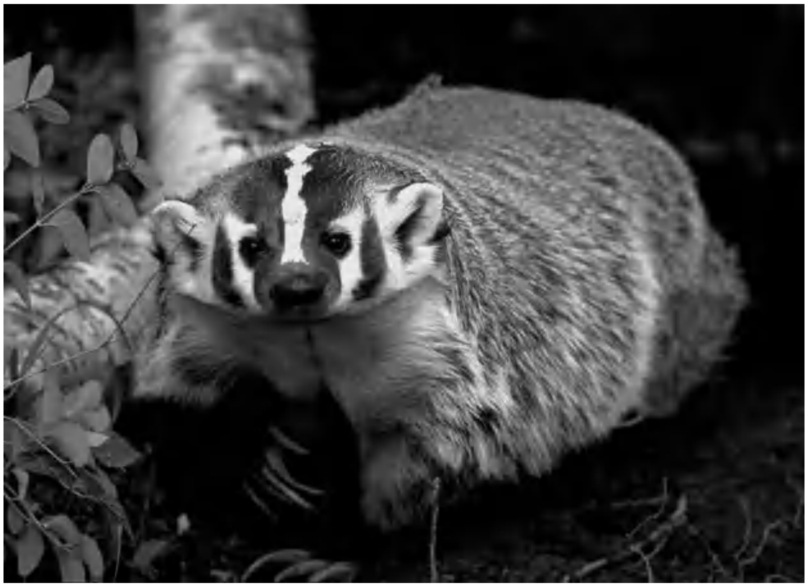
Badger

A carnivorous mammal in the family Mustelidae along with weasels and otters. Badgers originated in Asia and are currently found from Ireland to Japan and the Philippine Islands as well as in North America. They occupy a wide range of habitats including open plains, semidesert, and the boreal forests of Scandinavia and Russia. In parts of Russia, they occur as far north as the Arctic Circle. Worldwide there are ten species of badgers classified in six genera (see **table**).

American badger. The American badger (*Taxidea taxus*) is found from northern Alberta and southern British Columbia to Ohio, central Mexico, and Baja California. Throughout most of its range, this species prefers open country, living in the prairies and plains where its principal foods—ground squirrels, prairie dogs, and other burrowing animals—abound. The American badger is a heavy-bodied, medium-sized mammal with a broad head; a short, thick neck; short legs; and a short, bushy tail. The ears are low and rounded. Five toes are present on each foot. A prominent nonretractile claw is present on each toe, with the claws on the front feet being very long and exceeding 1 in. in length. The thick fur is rather short on the back but moderately long on the sides, thus giving the badger a characteristic

Names and geographic distribution of badgers

Name	Range
Mellivora capensis (Honey badger, or ratel)	Africa, India
Meles meles (Old World badger)	Europe, Asia
Arctonyx collaris (Hog badger)	Southern Asia, Sumatra
Mydaus javanensis (Malay stink badger, teledu)	Malay Archipelago
Mydaus marchei (Palawan stink badger)	Philippines
Taxidea taxus (American badger)	North America
Melogale moschata (Chinese ferret badger)	China
Melogale personata (Indian ferret badger)	Nepal, Indochina
Melogale orientalis (Oriental ferret badger)	Java, Bali
Melogale everetti (Everett's ferret badger)	Borneo



American badger (*Taxidea taxus*). (*Photo by Gerald and Buff Corsi; © 1999 California Academy of Sciences*)

squat appearance (see **illustration**). The general coloration is grayish to reddish with a slight yellowish cast. The brownish face is marked with a white stripe reaching from near the nose to the crown of the head and sometimes onto the neck and back. Paired white areas extend from around the mouth onto the cheeks and insides of the ears, and a prominent, vertical black bar, or "badge," occurs in front of each ear. The backs of the ears and the feet are black; the tail is yellowish-brown; and the underparts are predominately yellowish-brown. Both sexes are colored alike. The dental formula is I 3/3, C 1/1, PM 3/3, and M 1/2 × 2 for a total of 34 teeth. Adults have a head and body length between 420 and 720 mm (16–28 in.), a tail length of 100–155 mm (4–6 in.), and a weight of 4–12 kg (9–26 lb). Males are heavier than females. *See* DENTITION.

Badgers have two pairs of scent glands, one on the belly and the other pair near the anus. The anal scent glands secrete a strong, sometimes pungent, odor and are probably used mostly during the breeding season for sexual attraction and at other times for defense. Badgers hiss, grunt, growl, and snarl when fighting or cornered by an aggressor. They can walk and run, attaining a topmost speed of 10 to 15 miles per hour.

Badgers are excellent diggers, as evidenced by their heavy body, powerful muscles, strong front feet, and long claws. They can dig at a faster rate than a human can dig with a shovel. When digging, badgers loosen the soil with their front feet, pass the dirt under their belly, and kick it out of the hole with their hind feet. Sometimes they use their mouth in digging. Some ferret badgers are good climbers and may sleep in trees.

Reproduction and development. Breeding in American badgers generally occurs in August or September. Several days following fertilization, the embryos become dormant in the uterus for several months before implantation occurs. This process of delayed implantation is common in several other members of the family Mustelidae. In February or March, the embryos become implanted in the uterus and complete their development in about 5 weeks. Thus, while the entire gestation period may last from 6 to 9 months, the period of actual development is completed in about 6 weeks. A single annual litter consisting of 1 to 7 young (usually 2 or 3) is born in late spring. At birth, the altricial young are furred and blind. Their eyes open between 4 and 6 weeks of age. Weaning occurs when the young are about half-grown. The young stay in the vicinity of their home burrow until fall before dispersing.

Ecology. Most badgers are solitary mammals coming together only to mate. European badgers, however, live in groups and are territorial. American badgers usually spend their entire lives within a home range of 1 or 2 square miles and normally do not go far from some hole into which they can make a hasty retreat if danger threatens. They are active mostly at night but sometimes forage in early morning or late evening. They often sunbathe near the entrance of their burrows.

Badgers are active during all seasons, although in more northern portions of their range they may spend considerable time sleeping in their burrows during the colder winter months. During most of the year, the badger's home is a shallow burrow (known as a sett) about 1 ft in diameter which it digs in its daily quest for food. Ground squirrels are a staple food of the American badger during the warmer months of the year, but they also consume mice, rabbits, insects, moles, lizards, snakes, turtle eggs, ground-nesting birds and their eggs, snails, and fish. Both live animals and carrion may be consumed. At times, American badgers and coyotes may forage together for their mutual benefit. During the badger's digging operations, a rodent may escape only to be pounced upon by the agile coyote. Hawks have also been observed following a badger, presumably on the alert for some escaping rodent. While the American badger is a specialist predator, the other nine species are generalist omnivores feeding on a variety of insects and invertebrates as well as small vertebrates, cereals, and tubers.

Indiscriminate poisoning and destruction of their rodent prey has caused a marked decline in American badgers. Dogs and coyotes also prey on badgers. Badgers are a valuable part of the ecosystem because they serve as a control on rodent populations, and their digging aerates and mixes the soil. *See* CARNIVORA; HIBERNATION AND ESTIVATION; MUSTELIDAE.

Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, McGraw-Hill, 1990; D. Macdonald, *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Balance

An instrument for weighing. The word "balance," derived from the Latin *bilanx* ("having two pans"), is often used interchangeably with "scale" or "scales" (Old English, meaning dishes or plates). However, balance is the preferred term for an instrument used for the precise measurement of small weights or masses in amounts ranging from micrograms up to a few kilograms. *See* MASS; WEIGHT.

Classification. Balances are differentiated according to design, weighing principle, and metrological criteria (see **table**). For a given weighing task, a balance is selected primarily for its maximum weighing load (Max) and for the finest graduation or division (d) of its weight-reading device (scale dial, digital display, readout).

The jewelry trade uses balances graduated in carat (1 carat = 200 mg), troy ounce (1 troy ounce = 1.1 avoirdupois ounce), and pennyweight (1 pennyweight = 1.555 g) for weighing diamonds and precious metals. Except for the special weight units, these instruments are identical with macroanalytical and precision balances.

Balances can be roughly differentiated from scales by their resolution or number of scale divisions, $n = \text{Max}/d$. Balances typically have a resolution of more than 10,000 divisions, and scales for the most part have less.

Traditional mechanical balances. In its classic form a balance consists of a symmetric lever called a balance beam, two pans suspended from its ends, and a pivotal axis (fulcrum) at its center (**Fig. 1**). The object to be weighed is placed on one pan, whereupon the balance is brought into equilibrium by placing the required amount of weights on the opposite pan.

Classification of balances

Type	Division (d)	Typical capacity (max.)
Ultramicroanalytical	0.1 μg	3 g
Microanalytical	1 μg	3 g
Semimicroanalytical	0.01 mg	30 g
Macroanalytical	0.1 mg	160 g
Precision	$\geq$1 mg	160 g–60 kg

Thus the weight of an object is defined as the amount represented by the calibrated standard masses that will exactly counterbalance the object on a classic equal-arm balance. Although this is not self-evident with modern balances and scales, the measurement of weight continues to be based on this original understanding.

In the design of an equal-arm balance (**Fig. 2*a***), it is critical that the pan suspension end pivots (E) be equidistant from, and in a straight line with, the center fulcrum (F). A rigid, truss-shaped construction of the beam minimizes the amount of bending when the pans are loaded. The center of gravity (C) of the beam is located a fraction of an inch below the center fulcrum, which gives the balance the properties of a physical pendulum, letting the balance swing about its equilibrium position. With a slight difference in pan loads, the balance will come to rest at an inclined position, the angle of inclination being proportionate to the load differential (Fig. 2*b*). By reading the pointer position on a graduated angular scale, it is possible to determine differential amounts of weight in between the discrete step values of the weight set.

In so-called trip balances, the beam has its center of gravity above the fulcrum. This type of balance

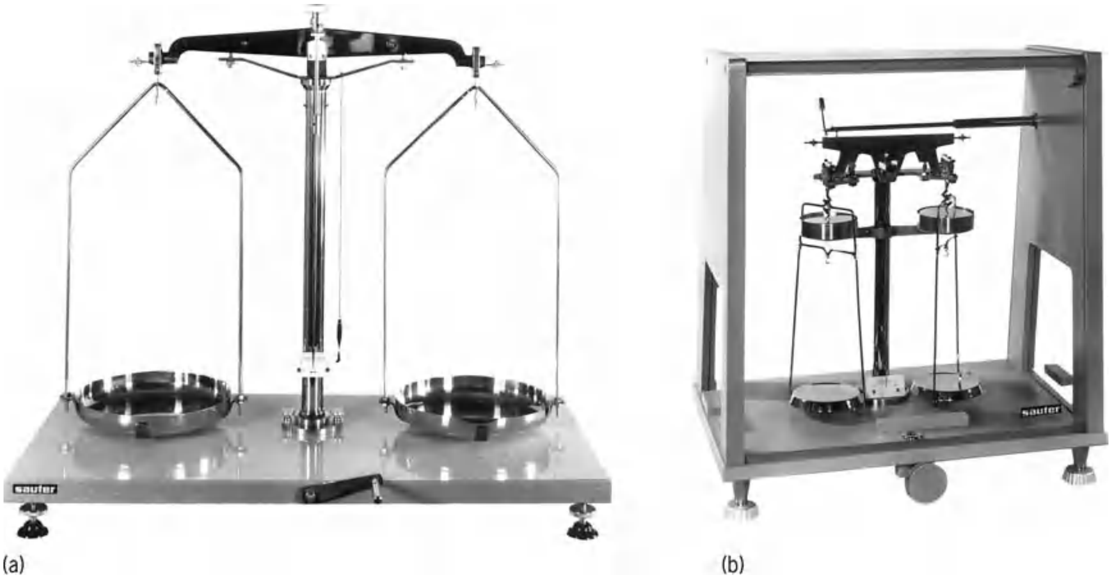


Fig. 1. Mechanical balances. (*a*) Traditional type: crank at center of base plate operates arrestment device for protection of knife-edge pivots, and thumbscrews at ends and center top of beam serve to adjust horizontal equilibrium position and vertical location of center of gravity of balance beam. (*b*) Mechanical analytical balance with several refinements: enclosure to protect balance from air draft disturbance; remote-operated rider poise on top of beam; air dashpots for oscillation damping. (*Mettler Instrument Corp.*)

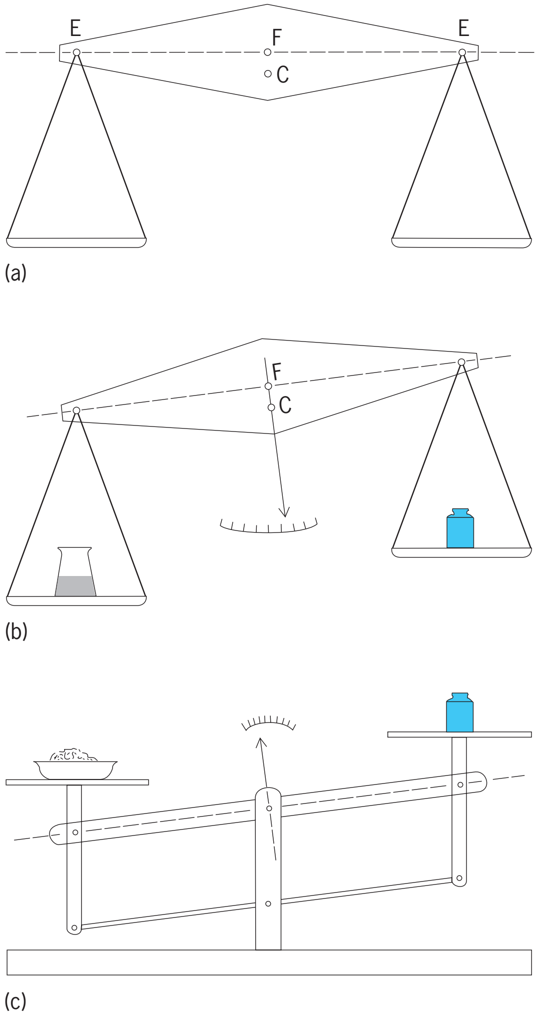


Fig. 2. Balance design considerations. (*a*) Critical design aspects for an equal-arm balance. (*b*) Weighing small weight differentials with an equal-arm balance. (*c*) Top-loading equal-arm balance.

trips to the side of the heavier load, but will not seek or maintain a horizontal equilibrium position with equal loads.

Notable variations and refinements of the mechanical balance include knife-edge bearings of hardened steel, agate, or synthetic sapphire; air damping or eddy-current (magnetic) damping of beam oscillations; sliding-weight poises or riders; built-in weight sets operated by dial knobs; microscope or microprojector reading of the angle of beam inclination; arrestment devices to disengage and protect pivots; and pan brakes to stop the swing of the balance pans.

Top-loading balances (Fig. 2*c*) are based on the same principles and equipped with many of the same features and refinements. With the pans stabilized by lever linkages [so-called Roberval (parallelogram) linkages] they remain horizontal as the pans move up and down. Thus, top-loading balances are more convenient to operate and give unobstructed access for tall or wide weighing loads.

Single-pan substitution balance. First described by the French physicist Jean Charles de Borda (1733–1799), and introduced in industrially manufactured balances by Erhard Mettler in 1946, the substitution principle represented the conclusive step in the evolution of the mechanical balance. Substitution balances (**Fig. 3***a* and *c*) have only one hanger assembly, incorporating both the load pan and a built-in set of weights on a holding rack. The hanger assembly is balanced by a counterpoise which is rigidly connected to the other side of the beam.

The weight of an object is determined by lifting weights off the holding rack until the balance returns to an equilibrium position within its angular, differential weighing range. Small increments of weight in between the discrete dial weight steps are read from the projected screen image of a graduated optical reticle which is rigidly connected to the balance beam.

With only two pivot axes, this design disposes of the problem of keeping pivots in a straight line with, and equidistant from, the fulcrum. Furthermore, the elastic deformation of the beam stays unchanged as the loads on the beam remain essentially constant.

Substitution balances were built as analytical balances with a hanging pan (Fig. 3*b*) and as top-loading precision balances with a stabilized pan (Fig. 3*d*). With minor exceptions, the technical development and production of substitution balances ended because electronic balances have superior accuracy and operating convenience. However, substitution balances will remain in widespread use in laboratories until replacement by new models.

Electronic balances. The evolution of electronic (more accurately, electromechanical) balances (**Fig. 4**) started in the late 1960s and has extended over several generations of electronic technology. Among a number of technical possibilities, one operating principle, electromagnetic force compensation, emerged early as the standard in high-precision weighing. First described by K. Ångström in 1895, the principle of electromagnetic force compensation became feasible for technical application as a result of the advancements in solid-state electronic components. Besides improved accuracy, reliability, and speed of operation, the main benefits from this technology are human-engineered design for optimized interaction between operator and instrument, and numerous operating conveniences such as push-button zero setting, automatic calibration, built-in computing capabilities for frequently used work procedures, and data output to printers and computers.

In every electromechanical weighing system, there are three basic functions (**Fig. 5**): (1) The load-transfer mechanism, composed of the weighing platform or pan, levers, and guides, receives the weighing load on the pan as a randomly distributed pressure force (*P*) and translates it into a measureable single force (*F*). (2) The electromechanical force transducer, often called load cell, converts the mechanical input force into an electrical output, for example, voltage, current, or frequency. (3) The electronic signal-processing part of the balance receives

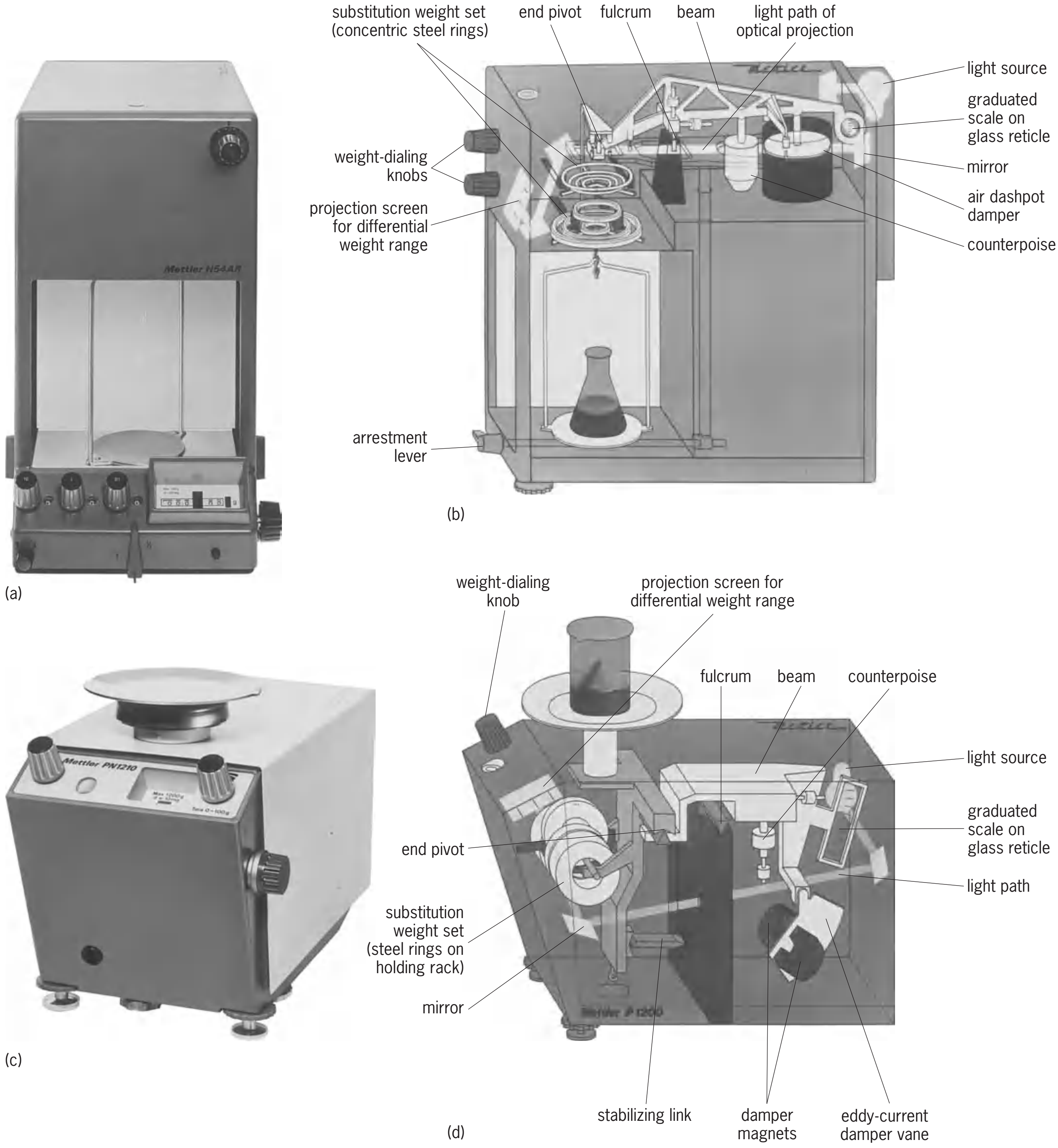

substitution weight set
(concentric steel rings)
end pivot
fulcrum
beam
light path of
optical projection
light source
graduated
scale on
glass reticle
mirror
air dashpot
damper
counterpoise
weight-dialing
knobs
projection screen
for differential
weight range
arrestment
lever
Mettler H54AR
(a)
(b)
weight-dialing
knob
projection screen for
differential weight range
fulcrum
beam
counterpoise
light source
graduated
scale on
glass reticle
end pivot
light path
substitution
weight set
(steel rings on
holding rack)
mirror
stabilizing link
damper
magnets
eddy-current
damper vane
Mettler PN1210
Mettler P1200
(c)
(d)

Fig. 3. Substitution balances. (*a*) The standard in analytical laboratory weighing from about 1950 to 1980. (*b*) Cutaway side view of a substitution analytical balance. (*c*) A top-loading substitution balance, readable to 1 mg or coarser. (*d*) Cutaway side view of a top-loading substitution balance. (*Mettler Instrument Corp.*)

the output signal, converts it to numbers, performs computation, and displays the final weight data on the readout.

Load-transfer mechanism. In electronic balances load-transfer mechanisms are lever linkages similar to those used in mechanical substitution balances. As a refinement, the traditional knife-edge pivots have been replaced by elastic flexure pivots (**Fig. 6**). Levers and other parts that rotate relative to each other are connected by flexibly bending strips or bands of high-grade spring metal. Flexure pivots for balances were first described by Wilhelm Weber in 1841, but the breakthrough in their application came with the development of electronic scales and balances. Pivot points in precision mechanisms are more accurately defined and maintained by flexure pivots than was previously possible with knife edges. As a result, precision balances can be built in an ergonomically correct, low-profile form, and all balances, except for the microgram range, use the top-loading pan arrangement, the pan being nonpendulous and supported and stabilized by flexure-pivoted levers.

Electromechanical force transducer. In a laboratory balance the transducer in many cases resembles the configuration of an electrodynamic loudspeaker. A sound-modulated current in the speaker coil interacts with the magnetic field in the circular gap of the permanent magnet, generating an oscillating force on the coil and voice cone assembly and thus producing sound waves.

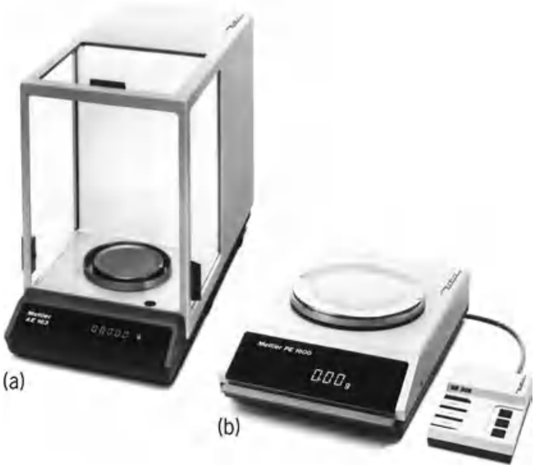


Fig. 4. Electronic balances. (*a*) Analytical balance. The weighing pan is nonpendulous and stabilized by a parallelogram linkage. Operating functions are controlled by single touch bar (integrated in readout panel) and include full-scale tare, automatic calibration, selection of weighing range, averaging period, and stability detector setting. (*b*) Electronic precision balance. Typically, balances reading to 1 mg or coarser require no enclosed weighing compartment, as they are less affected by air drafts than analytical balances. Operating features are similar to those of the analytical balance. The balance is operated through control bar along the front. (*Mettler Instrument Corp.*)

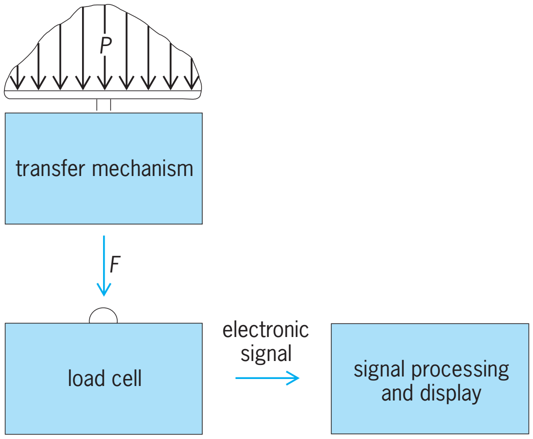


Fig. 5. Three-part structure of an electronic balance.

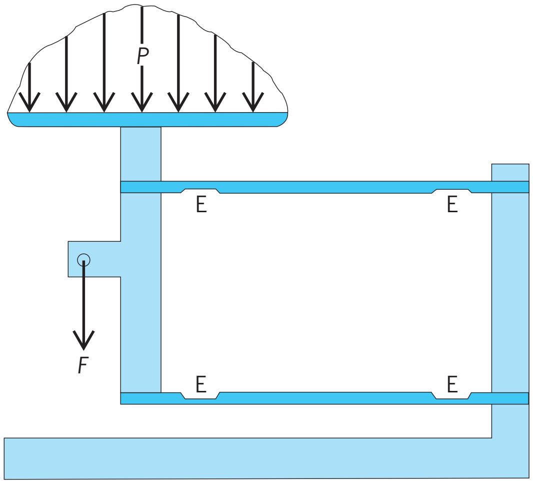


Fig. 6. Diagram of the transfer mechanism of electronic balance. Random force action of weighing load (*P*) is converted into a single force (*F*). Platform is stabilized by flexure-pivoted guides. Pivots are formed by the elastically flexible sections (*E*) in the horizontal guide members (called check links).

In the electromagnetic force compensator (**Fig. 7**), a unidirectional current generates a static force on the coil which, in turn, counterbalances the weight force from the object on the balance pan, usually aided by one or more force-reduction levers. The amount of coil current is controlled by a closed-loop servo circuit which monitors the vertical deflections of the pan support through a photosensor and adjusts the coil current as required to maintain equilibrium between weighing load and compensation force.

The interactive force (F) between a magnetic field and a current-carrying conductor (at a right angle to each other) is determined by the equation below,

$$F = IlB$$

whereby I represents the current, l is the total length of coil wire, and B is the magnetic flux density in the air gap. The force F points in the third direction in space, that is, perpendicular to the directions of current and magnetic field. The force F, and with it the weight of the object on the balance pan, now results from a measure of the current I, for example, by measuring the output voltage V across a precision resistor R wired in series to the force coil, and using the relationship $V = IR$.

Electronic signal processing. This part of a balance can be thought of as a voltmeter whose dial face or digital indicator reads in units of weight. However, the accurate resolution of the measuring range into several million parts, for example, 30 to 0.00001 g, exceeds by far the capabilities of all but the most sophisticated voltmeters. Signal processing in balances involves the most advanced methods of electrical measurement, as well as special computation routines that are applied to each weight value before it is displayed. There are five basic computations: (1) Tare subtraction; subtracts container weight or other preload from each weight value. (2) Automatic

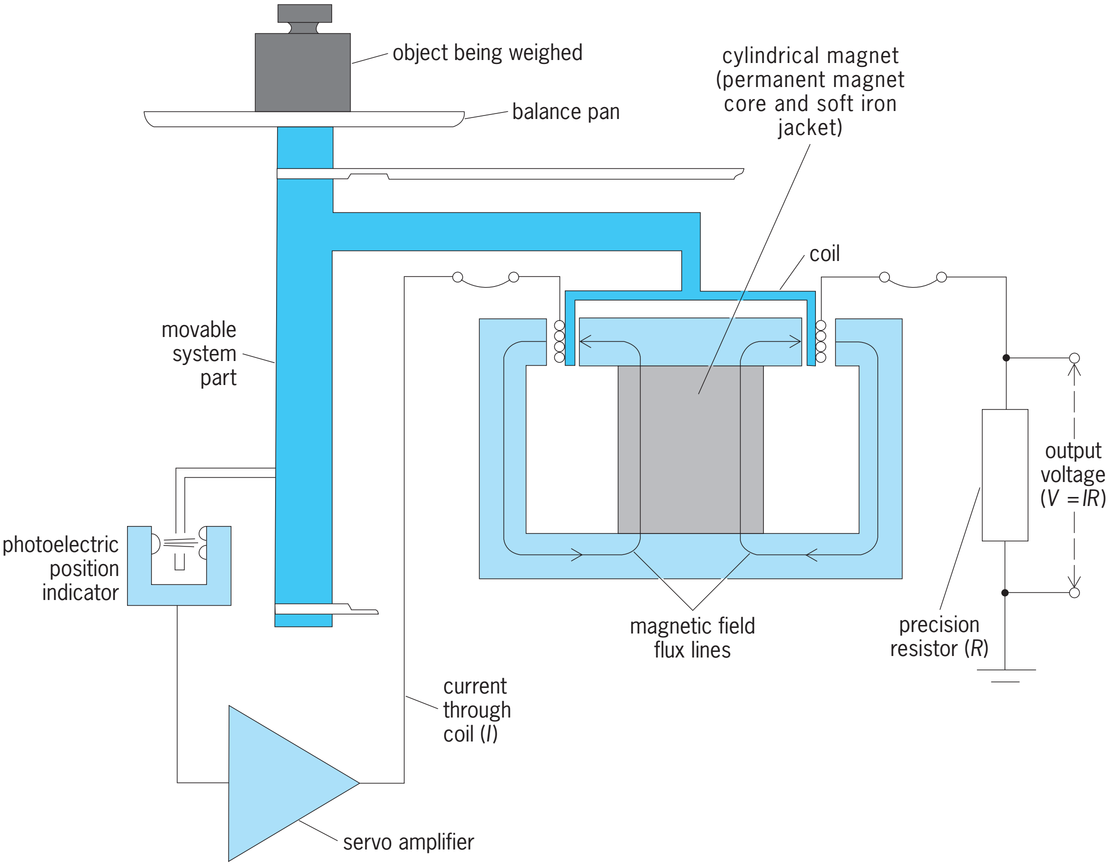

object being weighed
balance pan
cylindrical magnet (permanent magnet core and soft iron jacket)
coil
movable system part
output voltage ($V = IR$)
photoelectric position indicator
magnetic field flux lines
precision resistor (R)
current through coil (I)
servo amplifier

Fig. 7. Diagram of the electromagnetic force compensation principle. The output voltage is proportionate to the weight on the pan.

zero correction; compensates for tendencies of the measuring system to gradually wander off a set zero, even by a fraction of a display division. (3) Vibration filter; applies averaging or other noise-reduction techniques to measurement values in order to produce stable weighing results in the presence of building vibrations or other disturbances, for example in weighing live animals. (4) Stability detector; compares consecutive weighing results with each other, and blocks data output and signals a warning light when weighing results are transient or fluctuating. (5) Automatic calibration; as the operator presses a key and deposits a built-in or external mass standard on the balance, the instrument calibration factor is recalculated and stored in a nonvolatile (power-independent) memory.

In addition to these signal-processing features, state-of-the-art balances offer a number of useful conveniences either built in or available as options on some models, such as display in nonmetric weight units, parts counting, weight statistics, percentage calculation, or recall of net total.

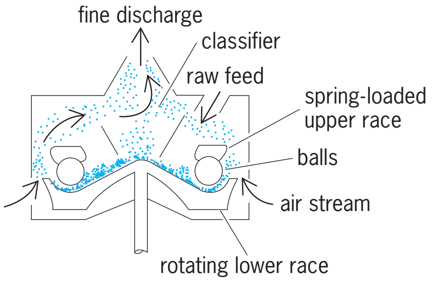
Accuracy. As with all scientific measurements, weighing results should not be taken as unconditionally accurate. Manufacturers' specifications indicate tolerance limits for the most common instrument errors, that is, imprecision, nonlinearity, and temperature drift. However, the single largest error in determining absolute amounts of mass lies outside the balance. It is made when the effects of air buoyancy on the weighing object are neglected. For example, when weighing 1 liter (1.06 qt) of water at sea level, an amount of 1.05 g (0.037 oz) should be added to the net weight as read from the balance in order to obtain the mass of the water. This represents the difference between the buoyancy forces acting on a kilogram (2.2 lb) of water on the one hand, and a kilgram of steel (by which the balance was calibrated) on the other. While the buoyancy correction is necessary in the absolute determination of mass, it is of no concern in any other applications where weight, in the traditional sense, is taken as a direct measure for quantities of any weighable matter. Errors entering into experiments from other sources, for example, volume vessels, are generally much larger, making balance errors negligible in comparison. *See* PHYSICAL MEASUREMENT. Walter E. Kupper

Bibliography. G. D. Christian, *Analytical Chemistry*, 5th ed., 1993; B. Kisch, *Scales and Weights: A Historical Outline*, 1965; R. O. Leonard, Electronic laboratory balances, *Anal. Chem.*, 48(11):879A–894A, September 1976; P. E. Pontius, *Mass and Mass Values*, Nat. Bur. Stand. Monogr. 133, 1974.

Ball-and-race-type pulverizer

A grinding machine in which balls rotate under pressure to crush materials, such as coal, to a fine consistency. The material is usually fed through a chute to

-S



Coarse raw material is ground by crushing and attrition between balls and races and is then withdrawn from the pulverizer by an airstream.

the inside of a ring of closely spaced balls. In most designs the upper spring-loaded race applies pressure to the balls, and the lower race rotates and grinds the coarse material between it and the balls (see **illus.**). The finely ground material discharges along the outer periphery of the ball races. For the pulverization of coal, hot air, introduced between the lower race and the pulverizer housing, lifts or carries the fines to a cyclone classifier at the center of the pulverizer. There the finer particles discharge from the pulverizer while the larger particles return to the grinding zone for further reduction in size. Two or more rings of balls can be cascaded in one machine to obtain greater capacity or output. Counterrotating top and bottom rings also are used to increase pulverizer capacity. Such pulverizers are compact and the power required per ton of material ground is relatively low. *See* CRUSHING AND PULVERIZING. George W. Kessler

Ballast resistor

A resistor that has the property of increasing in resistance as current flowing through it increases, and decreasing in resistance as current decreases. Therefore the ballast resistor tends to maintain a constant current flowing through it, despite variations in applied voltage or changes in the rest of the circuit. The ballast resistor acts as a variable load on the system; therefore it differs from load resistors, which have a constant resistance. Ballast resistors are now mainly of historical interest; electronic devices fulfilling the same function have replaced them. *See* RESISTOR.

The ballast action is obtained by using resistive material that increases in resistance as temperature increases. Any increase in current then causes an increase in temperature, which results in an increase in resistance and reduces the current. Ballast resistors may be wire-wound resistors. Other types, also called ballast tubes, are usually mounted in an evacuated envelope to reduce heat radiation.

Ballast resistors have been used to compensate for variations in line voltage, as in some automotive ignition systems, or to compensate for negative volt-ampere characteristics of other devices, such as fluorescent lamps and other vapor lamps. *See* FLUORESCENT LAMP; VAPOR LAMP; VOLTAGE REGULATOR.

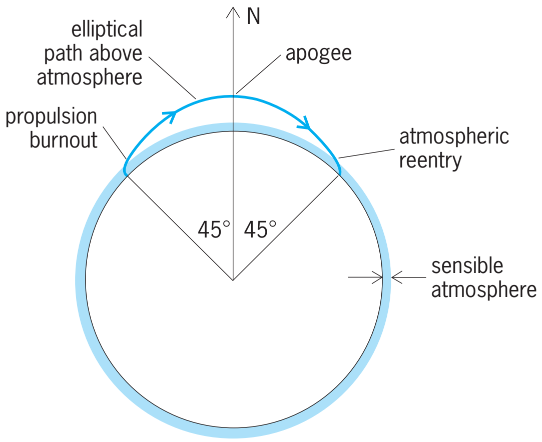
Donald L. Anglin

Ballistic missile

A weapon that consists of integral rocket propulsion, means of pointing or guiding the weapon's velocity vector to a prescribed orientation at the position and time of rocket engine shutoff or burnout, and a warhead. In certain applications, means of deploying multiple warheads or submunitions may be incorporated. Ballistic missiles are conceptually simple weapons whose implementation becomes more complex with increasing accuracy, range, and defense penetration requirements.

Trajectory. The term ballistic means that part or most of the missile's trajectory is not subject to propulsion or control. In its ballistic phase of flight, a missile's motion is affected only by gravitation and uncontrolled aerodynamic interactions with the atmosphere.

The ballistic missile follows an elliptical path due to action of the Earth's gravitational field (see **illus.**). The vertical component of velocity after propulsion burnout decreases as kinetic energy is converted to potential energy through the action of gravity, and minor velocity losses occur due to momentum decrease in the presence of aerodynamic resistance. If both the burnout velocity and burnout altitude are large, then an upwardly slanted flight path will cause the missile's trajectory to rise high above the sensible atmosphere, thereby eliminating the retarding and disturbing influences of the Earth's atmosphere for most of the trajectory. When all of the vertical component of burnout kinetic energy has been converted to potential energy, a condition of minimum speed and maximum altitude associated with the apogee of the trajectory is achieved, and the missile begins to descend while transforming potential energy back into kinetic energy as the downward component of velocity increases because of gravity. If propulsion burnout occurs in the upper reaches of the atmosphere, then the maximum speed of the missile upon descent will be about the same as achieved at propulsion burnout because of conservation of energy in the absence of atmospheric retardation. *See* BALLISTICS; CELESTIAL MECHANICS.



Typical ICBM trajectory superimposed on Earth's circumference.

Propulsion. The purpose of the propulsive function is to impart velocity to the missile so as to achieve an energy state at propulsion burnout that is sufficient for the missile to coast to its target under the influence of gravitation and aerodynamic resistance. Multiple stages of propulsion may be employed to maximize burnout velocity. Each stage of a ballistic missile is an autonomous propulsion unit consisting of propellants, propellant tanks and liquid propellant feed equipment (or casing if solid propellants are used), one or more exhaust nozzles, and means of controlling the direction of the thrust vector. Each stage is operated serially, and when the propellant of a given stage has been consumed the entire empty stage is jettisoned to reduce the system mass that must be accelerated by subsequent stages, and to enable ignition of the subsequent stage. Each stage is successively smaller in propellant mass and thrust magnitude, consistent with the reduced mass each step of the way. *See* ROCKET PROPULSION.

Guidance and control. All ballistic missiles incorporate means of pointing or guiding their velocity vectors so that their trajectories end coincidentally with the intended target. The simplest instance involves launching a missile from a guide rail or tube which the weapon operator points toward the intended target and upward at predetermined elevation angle that will result in a missile impact in the target area. The Army's Multiple Rocket Launching System (MRLS) is a good example of simple pointing as the means of initial-conditions guidance. The most complex guidance schemes are employed with intercontinental ballistic missiles (ICBMs) wherein propulsion durations of several minutes are typical and precise pointing of the velocity vector at propulsion burnout is essential to achieve the desired accuracy in hitting the target area after intercontinental flight. ICBMs and theater nuclear weapons typically employ control over the direction of thrust from their rocket motors to change the orientation of the missile in response to guidance commands. Guidance is based upon the principles of inertial navigation, wherein precise measurements of the missile's acceleration and rotation rate are used to very accurately compute its velocity, orientation, and position. The resulting information is used to apply precise corrections to the missile flight path so as to assure impact very close to its target, regardless of disturbances from the atmosphere or tolerable anomalies in propulsion performance. Control over the burnout velocity may be achieved by varying the thrust or terminating it when the desired velocity is achieved. Additional measurements, such as observing the positions of stars, or other known references, may be made to supplement inertial navigation. *See* GUIDANCE SYSTEMS; INERTIAL GUIDANCE SYSTEM; STAR TRACKER.

Warhead. The warhead of a ballistic missile consists of a single or multiple explosive devices designed to destroy or neutralize the usefulness of the intended target. The explosive devices may consist of conventional chemical explosives with fragmentation material, or nuclear or thermonuclear materials and equipment. When the warheads of ballistic missiles are released above the atmosphere at nearly orbital (3–4 mi/s or 5–6 km/s) speeds with flight path inclinations in the 17–30° range, their trajectories arch high above the sensible atmosphere, so that the vehicles that carry the warheads must be specially designed to protect them from high aerodynamic heating loads upon reentry into the atmosphere at such high speeds.

Modern ICBMs employ multiple reentry vehicles carrying thermonuclear warheads. Each of these vehicles is spin-stabilized and released from a so-called bus vehicle with an added velocity and resultant flight path angle so that each impacts a separate target within a specified region of the Earth's surface. When reentry vehicles have accomplished reentry and descend to altitudes where aerodynamic maneuvering is efficient for changing their flight paths, such maneuvers can be applied as a means of penetrating defenses or of reducing the time to impact. In some designs, on-board means of locating specific targets and guiding the reentry vehicles to them are employed. *See* ATMOSPHERIC ENTRY; ATOMIC BOMB; HYDROGEN BOMB.

Types. Ballistic missiles are either land-based or sea-based.

Land-based missiles. Land-based versions are commonly categorized according to the distance they can fly. *See* ARMY ARMAMENT.

1. Battlefield ballistic missiles can hit targets from 12 to 300 mi (20 to 500 km) from the launch point, and generally employ conventional (nonnuclear) or submunition warheads.

2. Intermediate-range ballistic missiles (IRBM), which are sometimes referred to as intermediate nuclear forces (INF) or theater nuclear weapons, come in a variety of sizes, and can hit targets 300 to 3000 mi (500 to 5000 km) from the launch site. The 3000-mi (5000-km) maximum range is determined by the geometry and extent of Europe, significantly exceeding the distance from London to Moscow. These missiles invariably carry one to three nuclear warheads, and modern versions emphasize vehicle-borne mobility.

3. ICBMs have flyout ranges between 5500 and 7500 mi (9000 and 12,000 km), the larger range being all that is necessary to reach the southern border of China from the northernmost United States, encompassing the former Soviet Union. ICBM survivability in nuclear war is a major issue in determination of basing; the viable options range from so-called hardened underground emplacements to measures that exploit mobility and deceit. Modern ICBMs carry from one to ten nuclear warheads in reentry vehicles that are independently targetable (multiple independently targeted reentry vehicles, or MIRVs).

Sea-based missiles. Sea-based ballistic missiles are invariably based on submarines (submarine-launched ballistic missiles, or SLBMs), providing considerable uncertainty as to their location as an important element of survivability. The flyout range of SLBMs has systematically increased from a few thousand kilometers to more than 5500 mi (9000 km), equaling the capability of ICBMs. Nuclear-powered submarines are employed to maximize duration of deployment without surfacing. SLBM deployments of the early 1960s required submarines to approach coastal regions nearest their targets to reach interior targets. However, SLBM range improvements of the late 1970s and early 1980s allow deployments relatively close to the home ports of the submarines. SLBMs have employed MIRV warheads since the early 1970s, and accuracy enhancements of the 1980s provide SLBMs with effectiveness levels comparable to those of their land-based counterparts. *See* GUIDED MISSILE; MISSILE; NAVAL ARMAMENT; SUBMARINE.

M. Michael Briggs

Bibliography. R. H. Battin, *An Introduction to the Mathematics and Methods of Astrodynamics*, 1987; B. G. Levi, M. Sakitt, and A. Hobson (eds.), *The Future of Land-Based Strategic Missiles*, 1989; P. G. Savage, *Introduction to Strapdown Navigation Systems*, 1985; G. Siouris, *Aerospace Avionic Systems*, 1993.

Ballistics

That branch of applied physics which deals with the motion of projectiles and the conditions governing that motion. Commonly called the science of shooting, it is, for practical purposes, subdivided into exterior and interior ballistics.

Exterior ballistics begins at the instant the projectile leaves the muzzle of the gun barrel or the instant after impact; interior ballistics, logically, deals with the events preceding this instant, that is, the events inside the gun barrel or the details of the impact that produces the motion. In addition to "shooting," the field of ballistics also applies to many games such as football, baseball, golf, volleyball, and pool or billiards.

Exterior Ballistics

Exterior ballistics is the science dealing with the motions of projectiles not under propulsive power, as of a shell after it has left the cannon, of a rocket after burning has ceased, or of a baseball after being hit. If the Earth stood still and had no atmosphere, such trajectories would be simple to describe mathematically. The motion of a body projected at an angle to the Earth's surface is perturbed by several forces, so that the trajectory differs from the parabolic, hyperbolic, or elliptical form it would otherwise have. The perturbing forces are drag, cross wind forces, and the Coriolis effect. Their magnitude depends upon the shape and mass of the body, the density of the air, the wind currents, and the length of the path along the trajectory. These effects seriously complicate the calculation of the trajectory.

The trajectory of a projectile under the influence of the Earth's gravitational field and in vacuum would be an arc of an ellipse with the center of the Earth as one of the foci. This arc can be closely approximated by a parabola if the motion does not carry the projectile so high that the gravitational force differs significantly from the force at the Earth's surface. The gravitational force varies with height, as shown in Eq. (1), where g is the acceleration of gravity at

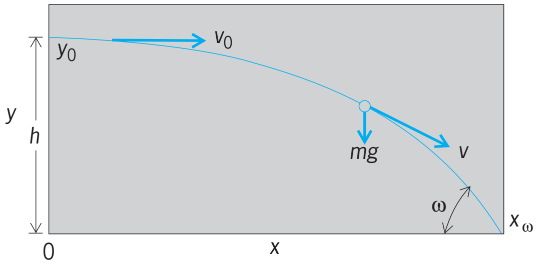


Fig. 1. Parabolic trajectory of a body projected horizontally, without resistance.

$$g = g_0\left(\frac{R}{R+y}\right)^2 \qquad (1)$$

height y, g_0 is its value at the surface ($y = 0$), and R is the Earth's radius. In order that g fall to $0.99g_0$, y must be 20 mi (32 km).

Horizontal projectile. This trajectory is illustrated in **Fig. 1**. After the projectile of mass m leaves the muzzle at $x = 0$, the only force acting upon it will be gravity; this force will have a magnitude mg and will be directed downward. It is assumed here that the range is so short that the Earth may be considered flat, that the height is so low that g is the same as at the Earth's surface, and that the projectile is in a vacuum; complications due to the atmosphere will be dealt with later.

The motion relationships are given in Eqs. (2).

$$\begin{aligned} m\frac{d^2x}{dt^2} &= 0 \\ m\frac{d^2y}{dt^2} &= -mg \end{aligned} \qquad (2)$$

The initial conditions (subscript 0) are given in Eqs. (3).

$$\begin{aligned} &x_0 = 0 && y_0 = h \\ &v_{x_0} = \left(\frac{dx}{dt}\right)_0 = v_0 && v_{y_0} = \left(\frac{dy}{dt}\right)_0 = 0 \\ &\left(\frac{d^2x}{dt^2}\right)_0 = 0 && \left(\frac{d^2y}{dt^2}\right)_0 = -g \end{aligned} \qquad (3)$$

Solutions to these equations are given in Eqs. (4).

$$\begin{aligned} x &= v_0 t \\ y &= h - g\frac{t^2}{2} \end{aligned} \qquad (4)$$

The curve of the trajectory, obtained by eliminating t, is the parabola shown in Eq. (5). The range of

$$y = h - \frac{g}{2v_0^2}x^2 \qquad (5)$$

the point of fall x_ω (the value of x when $y = 0$) is given by Eq. (6).

$$x_\omega = v_0\sqrt{\frac{2h}{g}} \tag{6}$$

The time of flight t_ω (the value of t when $y = 0$) is independent of v_0 and is given by Eq. (7). The angle of fall ω is given by Eq. (8).

$$t_\omega = \sqrt{\frac{2h}{g}} \tag{7}$$

$$\tan\omega = -\left(\frac{dy}{dx}\right)_{y=0} = \frac{\sqrt{2gh}}{v_0} \tag{8}$$

Angular projectile. This trajectory is illustrated in **Fig. 2**. Let a body with mass m be projected in a vacuum with an initial velocity v_0 at an angle θ_0 to the horizontal, and assume a flat Earth with g constant. The origin of coordinates is chosen at the point of projection, the y axis is vertical, and the x direction is so chosen that the trajectory lies in the xy plane. The equations of motion are given by Eqs. (9). The initial conditions are given by Eqs. (10).

$$\begin{aligned} m\frac{d^2x}{dt^2} &= 0 \\ m\frac{d^2y}{dt^2} &= -mg \end{aligned} \tag{9}$$

$$\begin{aligned} \left(\frac{dx}{dt}\right)_0 &= v_0\cos\theta_0 \quad x_0 = 0 \\ \left(\frac{dy}{dt}\right)_0 &= v_0\sin\theta_0 \quad y_0 = 0 \end{aligned} \tag{10}$$

Integrating the equations of motion and inserting the initial conditions, one obtains Eqs. (11).

$$\begin{aligned} x &= v_0 t\cos\theta_0 \\ y &= v_0 t\sin\theta_0 - \tfrac{1}{2}gt^2 \end{aligned} \tag{11}$$

To obtain the shape of the trajectory, eliminate t, as shown in Eq. (12). This is the equation of a

$$y = x\tan\theta_0 - \frac{gx^2}{2v_0^2\cos^2\theta_0} \tag{12}$$

parabola whose vertex, the summit (x_s, y_s) of the trajectory, can be found by equating dy/dx to zero, as in Eqs. (13).

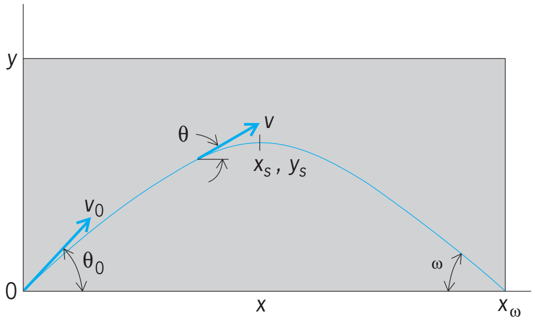


Fig. 2. Parabolic trajectory of a body projected at an angle, without resistance.

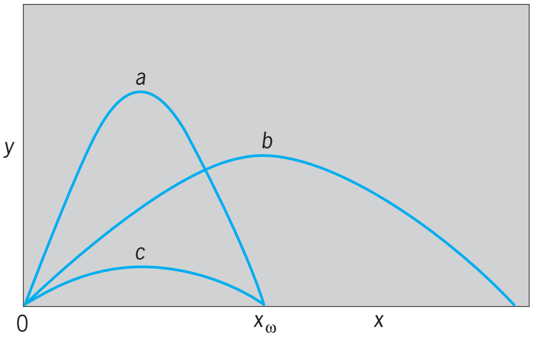


Fig. 3. Possible trajectories of body projected at same initial velocity but at different angle θ_0.

$$\begin{aligned} x_s &= \frac{v_0^2}{g}\sin\theta_0\cos\theta_0 \\ y_s &= \frac{v_0^2}{2g}\sin^2\theta_0 \end{aligned} \tag{13}$$

The time of flight to the summit is then given by Eq. (14). The range of fall x_ω can be determined by setting $y = 0$, as in Eq. (15). The time of flight t_ω, Eq. (16), is found by substituting $y = 0$ or $x = x_\omega$ in the equations involving t.

$$t_s = \frac{v_0}{g}\sin\theta_0 \tag{14}$$

$$x_\omega = \frac{2v_0^2}{g}\sin\theta_0\cos\theta_0 \tag{15}$$

$$t_\omega = \frac{2v_0}{g}\sin\theta_0 \tag{16}$$

Both the range and the time of fall are just twice the values at the summit, as would be expected from symmetry. To obtain the condition for maximum range for a given v_0, let $dx_\omega/d\theta_0 = 0$. Then Eq. (17) holds, and from it Eq. (18) is derived. With a fixed

$$\cos\theta_0 = \sin\theta_0 \tag{17}$$

$$\theta_0 = \frac{\pi}{4} = 45^\circ \tag{18}$$

initial velocity and for a less than maximum range, two trajectories are possible (**Fig. 3**), one with $\theta_0 > \pi/4$ and one with $\theta_0 < \pi/4$.

Motions of bodies with resistance. Resistance to the motion of a body through a fluid such as air derives from the fact that the body must move the fluid out of its way. The fluid particles are accelerated as the forward-moving body collides with them, and this results in a force, called drag, opposing the motion of the body and robbing it of some of its energy. The drag is a function of the density of the fluid ρ, of the area of the body, and of its velocity. The expression for the drag is given by Eq. (19), where

$$D = \rho d^2 v^2 K_D \tag{19}$$

d is the diameter of the body (assuming a circular cross section), v the velocity, and K_D the drag coefficient. The drag coefficient itself is a function of velocity and undergoes a rapid change as the velocity approaches that of sound. Both D and K_D are functions of height, since both the density and temperature of air vary markedly with height. The velocity of sound varies with temperature according to Eq. (20),

$$a = a_s\sqrt{\frac{T}{T_s}} \tag{20}$$

where a is the velocity at absolute temperature T, and a_s is the velocity at standard temperature T_s (15°C or 59°F). The temperature in the atmosphere normally decreases with height for the first several miles.

The effect of the drag is to modify the theoretical vacuum trajectory so that the projectile falls short of the impact point it would have reached in a vacuum and impacts with a larger value of ω. Extensive tables have been determined empirically for the drag coefficient under wide variations of temperature and density. The equations of motion in the presence of drag are given by Eqs. (21).

$$\begin{aligned} m\frac{d^2x}{dt^2} &= -D\cos\theta \\ m\frac{d^2y}{dt^2} &= -D\sin\theta - mg \end{aligned} \tag{21}$$

The initial conditions are shown by Eqs. (22).

$$\begin{aligned} \left(\frac{dx}{dt}\right)_0 &= v_0\cos\theta_0 \\ \left(\frac{dy}{dt}\right)_0 &= v_0\sin\theta_0 \end{aligned} \tag{22}$$

In the equations of motion, the drag force D is a function of the velocity along the trajectory and is thus a complex function of x, y, and t. For this reason, these equations cannot be integrated by normal means, and approximate numerical methods have been developed.

The form of the drag coefficient, when plotted against the projectile velocity expressed in terms of the velocity of sound, is given in **Fig. 4**.

Prior to World War I, when shells were fired at low angle of departure (less than 15 or 20°), the Siacci method of calculating trajectories was commonly used. During that war the use of trajectories with large angles of departure (quadrant angles) became common, and new methods that were based on numerical integration were introduced—in the United States by F. Moulton. These methods were adapted to ballistics from earlier uses in astronomy.

The adequacy of the various approximations which can be used for the calculation of missile trajectories is determined, in the main, by the velocity of the missile and the elevation of the trajectory. In order of complexity, the methods are as follows.

Vacuum approximation. When the velocity is low, the quadrant angle small, and the projectile rather heavy (so that the range is short and atmospheric effects are small), the equations developed earlier for the vacuum trajectory may be used with good accuracy.

Didion-Bernoulli method. This method may be used when the quadrant angle is low (less than about 20°) and the velocity is less than about 800 ft/s (240 m/s), so that the drag may be considered proportional to the square of the velocity. The variables in the equations of motion are separated by employing, in certain critical places in the solution, the approximation that $\sec\theta =$ constant. Once this assumption is admitted, the solution is exact and is found in formulas written around four functions. These functions and the appropriate values of the constant representing $\sec\theta$ are tabulated and, in conjunction with the formulas representing the solution of the equations of motion, make possible the nearly exact calculation of low-velocity, low-angle trajectories.

Siacci method. This method is accurate and useful for trajectories of high-velocity missiles with low quadrant angles of departure. The basic assumptions are that the atmospheric density anywhere on the trajectory is approximately constant and that the angle of departure is less than about 15°. The equations of motion can then be put in a form permitting a solution in terms of formulas for t, x, y, and $\tan\theta$. These formulas, used in conjunction with the tables of four functions involved in the solution, yield the trajectories of projectiles fired at low angles.

Otto-Lardillon method. For trajectories of missiles with low velocities (so that the drag is proportional to the velocity squared) and quadrant angles of departure that may be high, exact solutions of the equations of motion are arrived at by numerical integration and are then tabulated.

Numerical integration. When projectiles with high velocities and high quadrant angles of departure were introduced, all earlier methods used in ballistics failed to give sufficiently accurate solutions. The introduction of high velocities and high angles meant increased complexity both in the variation of the drag coefficient, so that drag was no longer simply a function of the velocity squared, and in the variations

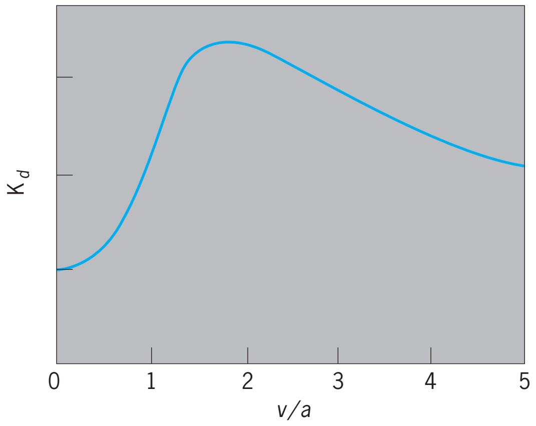


Fig. 4. Plot of drag coefficient K_D. Quantity v is velocity of projectile, and a is velocity of sound.

Methods of trajectory computation

Velocity	Maximum range	Quadrant angle	Method
Low	Short	Low	Vacuum approximation
Low	Medium	Low	Didion-Bernoulli
High	Medium-long	Low	Siacci
Low	Medium	Any	Otto-Lardillon
High	Long	Any	Numerical integration

of atmospheric density along the trajectory. Moulton and his associates introduced the method of approximate numerical integration, long used in astronomy, to obtain solutions for these trajectories. Computers are used extensively to perform the necessary calculations. *See* NUMERICAL ANALYSIS.

Since the advent of rockets which travel high in the atmosphere, the method of numerical integration has been very widely used. Although the problem is complex, even a desktop personal computer can obtain the solution.

Drag effects are very important in many sports. The atmosphere can be considered to have a constant density in sports such as football, golf, and baseball. Both golf balls and baseballs often have initial speeds in excess of their terminal speeds.

The basic approach in solving the rocket trajectory problem is to determine the conditions x, y, dx/dt, dy/dt, and θ at the end of burning and to go back along a hypothetical trajectory to determine the coordinates, velocity, and quadrant angle of departure of an equivalent cannon shell which will pass through the same point with the same velocity vector as the spent rocket. Any of the five methods outlined in the preceding discussion may then by used to calculate the resulting trajectory. An indication of the best method for use with various combinations of the initial conditions of velocity and quadrant angle is given in the **table**.

Effect of Earth's curvature. The above discussion of trajectories is based on the assumption that the Earth may be considered flat over the range of the projectile, and that the force of gravity is directed at all times parallel to its original direction. When the range of the missile or projectile becomes great, this assumption can lead to serious error; thus the analysis must be modified to accommodate the fact that the force of gravity varies with height, is directed toward the center of the Earth, and changes direction as the projectile moves forward. Also, the impact point is depressed because of the Earth's curvature, and the range therefore is extended.

These effects are customarily taken care of by the method of differential corrections but will be discussed here in a somewhat different fashion to clarify the point. Assume a nonrotating Earth for the time being. A projectile hurled along a trajectory away from the surface of the Earth is, while in motion, a satellite of the Earth and moves in its trajectory, which is really a section of an elliptical orbit about the center of the Earth. This is shown diagrammatically in **Fig. 5**. All trajectories or orbits with velocities less than about 25,000 ft/s (7600 m/s) intersect the surface of the Earth. The equations of motion for this two-body problem have been solved in celestial mechanics, except for the complications of drag and varying air density.

The ballistic problem could be solved by treating it as one in celestial mechanics and determining the intersection of the Earth's surface with the orbit to give the impact point. For projectiles of intermediate range, this is not done, but rather the technique of differential corrections on existing methods is used. However, for extremely long-range missiles, the approach of celestial mechanics employing numerical integration to take account of changing air densities and drag coefficients is used. *See* CELESTIAL MECHANICS.

Effect of Earth's rotation. Once the projectile is clear of the muzzle of the gun or the rocket is clear of the launching stand, it is a separate entity free of the Earth's rotation. One component of its initial velocity is the velocity of the surface of the Earth at the point of firing. For this reason, when the projectile arrives at the impact point, it will miss by an amount dependent on the azimuth of the trajectory, the range, and the height of the trajectory. This is called the Coriolis effect. For intermediate trajectories, such as those with pronounced Earth curvature, the Coriolis effect is handled by means of differential corrections. For very long-range missiles, the approach of celestial mechanics with nonrotating Earth-centered coordinates is used. Here the rotation is taken care of by adding the proper vector velocity for the latitude of launch and by taking account of the rotation of the Earth during the time of flight. *See* CORIOLIS ACCELERATION.

Stabilization by spinning. In order to stabilize a projectile in flight, the projectile is frequently spun about its longitudinal axis. The resulting gyroscopic

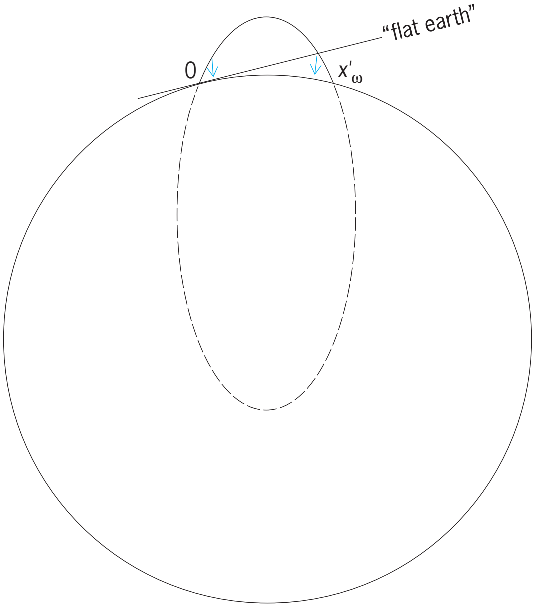


Fig. 5. Elliptical trajectory of a long-range missile. The range of fall is x'_w rather than x_w.

stabilization keeps the projectile from tumbling and hence maintains the proper orientation in flight to keep drag at a minimum. The spin of a shell is imparted by rifling the gun barrel, that of a rocket by canting the exhaust nozzles or the fins. While spin does tend to stabilize the orientation, in practice the spin axis will precess so that the nose of the projectile follows a spiral path. This causes troublesome complexities in the trajectory calculations, in that oscillating lift and yaw terms must be introduced. The equation of motion of a spinning projectile can be integrated by high-speed computing machines.

A further difficulty in spinning projectiles fired with large quadrant angles of departure is that the spin, in stabilizing the orientation along the original direction, causes the projectile to arrive at the point of impact at an inclination. that is different than the trajectory direction. Billiards and golf both depend upon understanding the effects of spin on the motion.

John P. Hagen; John L. Safko

Interior Ballistics

Interior ballistics deals with the motions of projectiles while they are still inside the gun barrel or cannon, and are accelerated by the pressure of expanding gases resulting from the burning of the change. The term is also used to refer to the similar events inside the casing of a solid-fuel rocket, in which case the exhaust blast is regarded as equivalent to the projectile.

Gas pressure changes. If the propelling charge of the cartridge is changed into gas instantaneously, the gas pressure curve, from the instant of ignition to the instant the projectile leaves the muzzle, would be very nearly an adiabatic curve, quite similar to the curve showing the diminishing potential of a gravitational field. In reality, it takes some time for the powder charge to burn; furthermore, various resistance factors must be overcome. Therefore, the gas pressure in the gun barrel behind the projectile does not reach its maximum value until after the projectile has started moving inside the barrel. The pressure should reach a minimum value at the instant the projectile leaves the gun barrel, as shown in **Fig. 6**; hence, there is a definite relationship between powder charge, gas pressure, and barrel strength. If the pressure is still rather high even when the projectile "unseals" the gun barrel, a portion of the propelling charge has been wasted. If the pressure drops too low before the projectile reaches the muzzle, the projectile is slowed down inside the barrel; in this case, shortening the barrel would actually improve the performance.

Minimizing of recoil. Even when the pressure drop has been properly calculated (and achieved), the exit velocity of the powder gases is still 2 to $2^{1}/_{2}$ times the muzzle velocity of the projectile. This fact can be utilized to prevent or to minimize the recoil of the weapon by means of the muzzle brake first developed by Schneider-Creusot in France. The muzzle brake deflects the direction of motion of the gases by 120–180°, thereby creating what might be called a gas recoil opposite in direction to the projectile recoil. Recoilless guns eliminate recoil by discharging a quantity of the propelling gases through the rear.

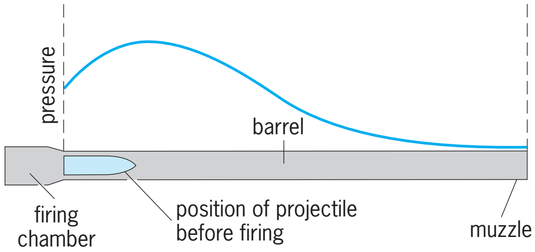


Fig. 6. Gas pressure in gun barrel.

Resistance factors. The expanding gases in the gun barrel have to overcome the following main resistance factors (in addition to their own inertia): (1) the inertia of the projectile, (2) the friction in the barrel caused by the copper driving bands against the lands and grooves of the rifling, and (3) the inertia of the air inside the barrel between the nose of the projectile and the muzzle of the gun. This last factor is rather unimportant for small guns or for short-barreled pieces of any caliber, and is often overlooked. However, it does influence performance in large and long-barreled pieces, especially when the muzzle velocity of the projectile is much larger than Mach 1, since in that case the air is not only accelerated but also compressed. The existence of this "plug" of compressed air often makes the experimental determination of the muzzle velocity difficult.

Further energy losses are caused by gas leakage, by the recoil motion of the gun barrel, and by the heating of the barrel. This last phenomenon not only absorbs energy but may increase the leakage. Obviously, there are many interdependent factors which make an exact calculation impossible. Empirical values must often be used.

Measuring devices. At first, the only tool of the ballistician was the ballistic pendulum (**Fig. 7**), which caught the projectile and indicated its momentum by the magnitude of its swing. This method was first suggested by J. D. Cassini in 1707 and was brilliantly utilized by Benjamin Robins some 30 years later. However, it permitted only indirect conclusions as to gas pressure inside the barrel. The first useful device for

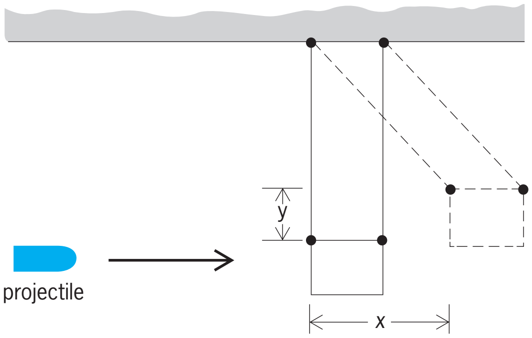


Fig. 7. Ballistic pendulum before (solid lines) and after (broken lines) projectile impact. Measurements of x and $\dot{y}$ and knowledge of mass and initial velocity of projectile give its momentum.

measuring gas pressure was A. Nobel's copper pressure gage, in which a copper plug was deformed by the gas pressure. The copper pellet (a lead pellet is used for weapons where the gas pressure is expected to be low) has remained the most useful tool since its introduction in 1860. Not until about 1940 was it supplemented by the piezoelectric gage (suggested originally by J. J. Thomson), which consists of a tourmaline disk that generates electric currents proportional to the pressure to which it is subjected.

In addition, strain gages have shown that a gun barrel momentarily expands slightly as the projectile moves toward the muzzle.

Basic equations. The main problems of interior ballistics have been defined by H. Athen as follows: "For a gun of which caliber, barrel length, firing chamber volume, weight of projectile, and type and weight of the propelling charge are known, the following factors must be determined: duration of burning of the powder charge, duration of travel of the projectile inside the barrel, gas pressure inside the barrel, and projectile velocity. Of these factors, the peak gas pressure and the muzzle velocity are especially important, since they can most easily be checked experimentally and are also of most practical value to the designer."

Investigations of the problems led to three equations, which are known as the energy equation, the combustion equation, and the equation of motion.

Energy equation. The derivation of the energy equation is based on the van der Waals equation (23),

$$\left(p + \frac{\beta}{s_v^2}\right) \cdot (s_v - \alpha) = RT \qquad (23)$$

where p is the pressure, T the absolute temperature, R the gas constant, s_v the specific volume of the gas, α the volume of the gas for $T = 0$, and β a constant. It is empirically known that $\beta/{s_v}^2$ is nearly zero; it is further known that for each type of powder, RT_c (T_c is the combustion temperature) is a constant value which is designated f and is called the specific pressure. If the volume of the closed space in which the powder is burned is designated as S and the weight of the powder as W, W/S might be called D (density of the charge), and the much simplified Abel equation (24), is obtained. It has been shown that

$$p = f\frac{D}{1 - \alpha D} \qquad (24)$$

this equation holds true for pressures from 17,000 to 64,000 lb/in.^{2} (120 to 440 megapascals or 1200 to 4500 kgf/cm^2). By introducing a few additional factors, Resal's energy equation is obtained.

Equations of combustion and motion. The combustion equation is shown in (25), where y is the distance

$$\dot{y} = A(y)p^k \qquad (25)$$

that the explosion has propagated, $\dot{y} = dy/dt$ is the velocity of the pressure front, (y) is a function determined by the grain shape of the charge and the location of the pressure front, A is a constant characteristic for the propellant under investigation, and p is the gas pressure. The factor k is generally taken to be less than unity.

The equation of motion is derived from Newton's second law. It is shown, taking $k = 1$, in Eq. (26),

$$\dot{v} = v\frac{dv}{dx} = \frac{qp}{\mu} \qquad (26)$$

where v is the velocity of the projectile inside the launcher, μ is the mass of the projectile, and q is the projectile cross section. *See* EXPLOSIVE; GAS.

If the interior ballistic is an impact such as a bat hitting a ball, the angle of impact and the structure of the projectile must be used to calculate the initial velocity and direction.

Willy Ley; David Wollersheim; John L. Safko

Bibliography. R. K. Adair, *Physics of Baseball*, 3d ed., Harper-Collins, New York, 2001; P. J. Brancazio, *Sport Science*, reprint edition, Touchstone, New York, 1985; C. L. Farrar and D. W. Leeming, *Military Ballistics: A Basic Manual (Land Warfare: Brassey's New Battlefield Weapons Systems & Technology Series into the 21st Century)*, Brassey's Publishers, London, 2002; R. L. McCoy, *Modern Exterior Ballistics: The Launch and Flight Dynamics of Symmetric Projectiles*, Schiffer Publications, Atglen, PA, 1999.

Balloon

A nonporous envelope of thin material filled with a lifting gas and capable of lifting any surrounding material and usually a suspended payload into the atmosphere. A balloon which is supported chiefly by buoyancy imparted by the surrounding air is often referred to as an aerostat. The balloon rises because of a differential displacement of air according to Archimedes' principle, which states that the total upward buoyant force is equal to the weight of the air displaced. The upper practical limit for useful ballooning is approximately 34 mi (55 km). Beyond this, the exponential nature of the atmosphere would require balloons of enormous size and delicately thin skin. *See* ARCHIMEDES' PRINCIPLE; ATMOSPHERE.

Balloons have been configured in many geometrical shapes, but the most common are spheres, oblate spheroids, and aerodynamic configurations. The materials used in the manufacture of the balloon envelope have been paper, rubber, fabric, and various plastics. Several types of lifting gases have been used to inflate balloons, but the most common in use are helium, hydrogen, and heated air.

History of ballooning. Crewed ballooning began on November 21, 1783, in Paris, France, when a balloon built by the brothers J. E. and J. Montgolfier carried the first two aeronauts, Pilâtre de Rozier and the Marquis François Laurent d'Arlandes, into the atmosphere. The balloon was constructed of paper and linen and used smoke and heated air for the lifting gas. Early balloons found their primary applications as a war weapon and for sport, and serious scientific balloon experiments were not conducted until the late nineteenth century.

In 1934 and 1935 in the United States the National Geographic Society and the Army Air Corps cosponsored two crewed balloon flights to conduct research in the stratosphere by measuring cosmic radiation, vertical ozone content, and composition of the atmosphere. Following this activity, a surge of crewed ballooning was conducted, with records being established, only to be broken on upcoming flights. Altitude records in excess of 100,000 ft (30 km) were recorded in crewed pressurized capsules. The flurry of crewed scientific ballooning continued until 1961, when it abruptly ended and yielded to the space era.

During the early 1950s considerable research was conducted on the design of balloons and improved materials for balloon manufacture. This research was primarily sponsored by the joint military services of the United States and resulted in the development of the modern zero-pressure balloon constructed of thin polyethylene film. This balloon is by far the most common vehicle for carrying large scientific experiments to the fringes of space. The polyethylene balloon has increased in volume from 1×10^6 to 7×10^7 ft^3 (3×10^4 to 2×10^6 m^3). Payload weight increased from hundreds of pounds to over 11,000 lb (5000 kg).

Balloon types. The many types of balloons in use fall into two main categories, extensible (expandable) and nonextensible. There are three methods of balloon operation: free balloons which are released into the atmosphere, tethered or moored balloons, and powered or controlled free balloons. The various types of balloons in use are the hot-air balloon, meteorological balloon, zero-pressure balloon, superpressure balloon, tethered balloon, and powered balloon.

Hot-air balloon. Historically, this balloon was used for many of the earliest achievements, and contributed significantly to science and aerostatic research. However, since about 1960 its primary function has been as a free balloon in sport, to carry people aloft. The gas bag is constructed of a nonextensible material and is usually spherical. The heated air required to produce lift is generated as a hydrocarbon gas burner attached above the basket. This basket, or gondola, is used to carry the passengers into the atmosphere. Lift can be controlled by adjusting the burning rate of the gas. A valve at the top of the balloon has an attached rope that allows a passenger to control descent. A rip panel and rip cord are attached to the balloon to allow rapid release of the gas on landing, to prevent dragging the load on impact.

Meteorological balloon. The meteorological or sounding balloon is used to carry radiosondes and other lightweight instruments aloft. It is an extensible balloon made from natural or synthetic rubber and filled with either hydrogen or helium. Hundreds of these balloons are launched daily by meterological stations throughout the world in support of the upper-air program of the World Meteorological Organization, as well as local weather forecasting. These flights are vertical sounding probes of the atmosphere for measuring pressure, temperature, humidity, and wind velocity. They can carry up to 40 lb (80 kg) of useful payload. High-altitude types carrying much lighter payloads may exceed 25 mi (40 km) before expanding to the bursting point. *See* METEOROLOGICAL INSTRUMENTATION.

Zero-pressure balloon. This is the workhorse of scientific ballooning, with a worldwide launch rate of several hundred per year. This is a nonextensible balloon constructed of thin-film polyethylene in a tailored natural shape (inverted teardrop). At float, it assumes the theoretical shape of a load-bearing balloon with zero circumferential stress. The polyethylene is manufactured by a blown extruding technique which produces a continuous tube of thin film. The thickness of the film can be adjusted by varying the extruder die, temperature, and pressure to produce films as thin as 9 micrometers (0.00035 in.). The tube is slit, unfolded, and rerolled on spools which are approximately 10 ft (3 m) wide. The material is then laid out on an assembly table, which is the full length of the balloon, and is cut to conform to one vertical segment of balloon, which is referred to as a gore. Additional gores are then added, with a polyester load tape heat-sealed to each seam. When finished, the balloon consists of 60 to 200 polyethylene gores, with load-bearing tapes extending from the apex top-fitting to the bottom end-fitting. The bottom of the balloon is left open through an arrangement of ducts, and hence the term zero-pressure balloon.

During inflation, the lifting gas is contained in a bubble which occupies less than 1% of the total balloon volume (**Fig. 1**). Approximately 10% additional lift over that required for neutral buoyancy is added to the balloon to provide the necessary vertical force to achieve an ascent rate of 800 ft/min (250 m/min).

Typically, these balloons require 3 h to reach a float altitude of 100,000 to 140,000 ft (30 to 42 km). After reaching float, the excess lift is allowed to vent through the ducts. To maintain a constant altitude, this system requires the release of ballast whenever a decrease in heating due to solar radiation is experienced. This effect is most noticeable when entering sunset condition, but can also be significant when passing over extensive cloud cover or thunderstorm activity.

A dynamic launch technique is generally employed when releasing these balloons. The payload is suspended on a launch vehicle, while the balloon and deployed parachute are laid out on the ground. A restraining spool holds the inflated bubble of lifting gas at the far end of the launch site. After sufficient lifting gas has been transferred to the bubble, the polyethylene fill tubes are tied off. On command from the launch director, the spool is released, allowing the bubble of gas to accelerate in a vertical direction. When the entire system is in a vertical position, the payload is released from the launch vehicle, which has been positioned directly under the ascending balloon train.

The float duration of a typical zero-pressure balloon experiment is from 8 to 40 h, depending on the time of year of launch. Recovery of the scientific equipment is essential; therefore coastal regions and international borders determine the extent of usable range. The upper-altitude winds are somewhat

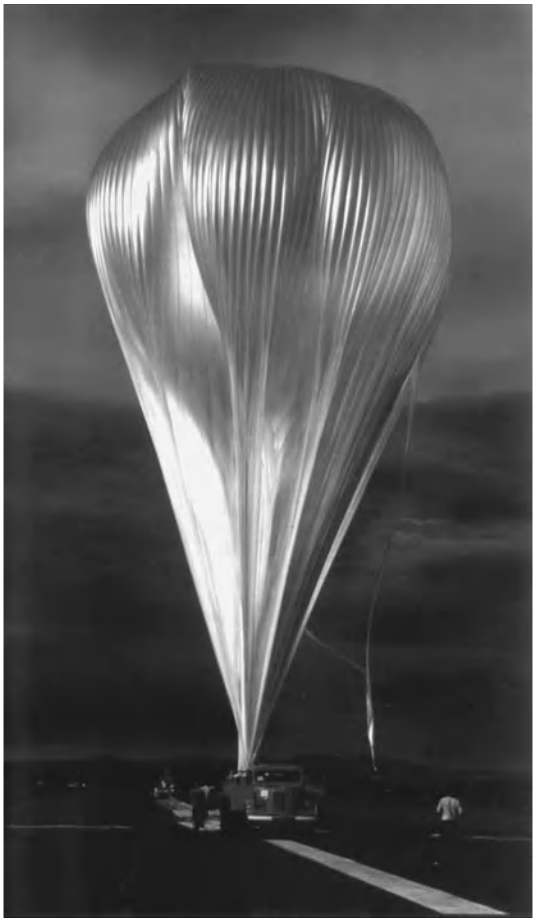


Fig. 1. Inflation of zero-pressure balloon. (*NASA*)

predictable and experience two 180° east-west reversals or "turn-arounds" during the year. The turnaround periods exist in May and September in the United States. During these periods the upper-altitude winds are less than 20 mi/h (10 m/s), and flights of several days duration can be accomplished. The winds become easterly during the summer and reach maximum velocities of 70 to 90 mi/h (30 to 40 m/s) during July. In the winter the winds become westerly and may exceed 60 m/s (136 mi/h).

The zero-pressure balloon is capable of carrying scientific instruments weighing thousands of pounds to an altitude of 140,000 ft (42 km), which is above 99.8% of the Earth's atmosphere. These scientific instruments have become very sophisticated and resemble a satellite in complexity. They utilize the latest technology in electronics; in fact, new technology is often proved in balloon experiments before being flown on satellites. The reasons are the shorter development time for the balloon experiment and the lower cost of the system, which allows a higher risk. The subsystems contained in a typical automated gondola are a high-bit-rate pulse-coded-modulated telemetry system for sending the scientific data to the ground tracking station, a digital command system used to convey information from the ground to the instrument, a battery power system, a temperature-control system, and a pointing system.

Superpressure balloon. A balloon that has been the subject of considerable research is the superpressure or constant-level balloon (**Fig. 2**). This is a nonextensible (constant-volume) free balloon which converts its free lift into superpressure by the time the float level is reached. As long as the balloon is superpressurized, it will continue to float at a constant density level. The main advantage of the superpressure balloon is the possibility of long-duration flights which can circumnavigate the globe. Small superpressure balloons 13 ft (4 m) in diameter have been released in the Southern Hemisphere, and some have remained aloft for over a year. Balloons having diameters of over 100 ft (30 m) have been flown with limited results. The objective of the research has been to produce a reliable superpressure balloon capable of carrying 700 lb (300 kg) to an altitude of 140,000 ft (42 km). The primary failure mode of the superpressure balloon is small pinholes as a result of manufacture and handling, which lead to the premature loss of superpressure. To minimize this effect, bilaminated materials usually of Mylar or some other strong polyester film are used.

Tethered balloon. This type of balloon is attached to a mooring line and allowed to ascend by releasing a winch mechanism (**Fig. 3**). The tethered balloon is generally configured as an aerodynamic body to provide lift, achieve lower drag, and increase lateral stability. Although 20,000 ft (6 km) is generally considered the upper altitude for this concept, the National Center for Space Studies in France reported flying a tethered balloon carrying an 800-lb (350-kg) gondola to an altitude of 55,000 ft (17 km) for 12 h.

The tethered balloon has primarily been used for military purposes to provide acoustic, radar, and optical surveillance, tall antennas, and communication relays. The inherent danger of the mooring cable to aircraft has prevented the application of the high-altitude tethered balloon for commercial and scientific uses. However, a commercial application is in

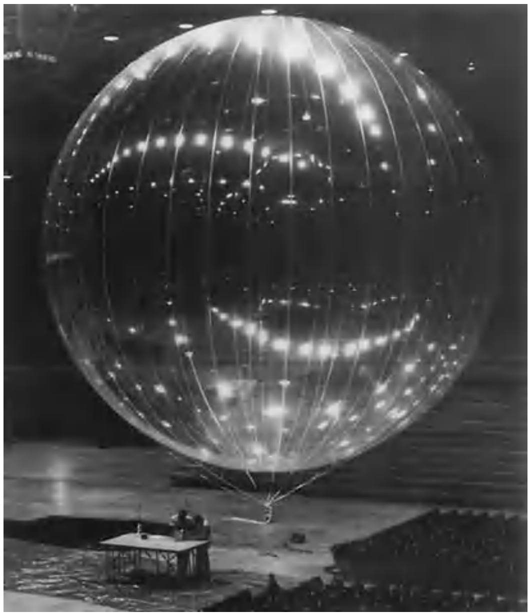


Fig. 2. Superpressure balloon. (*Raven Industries*)

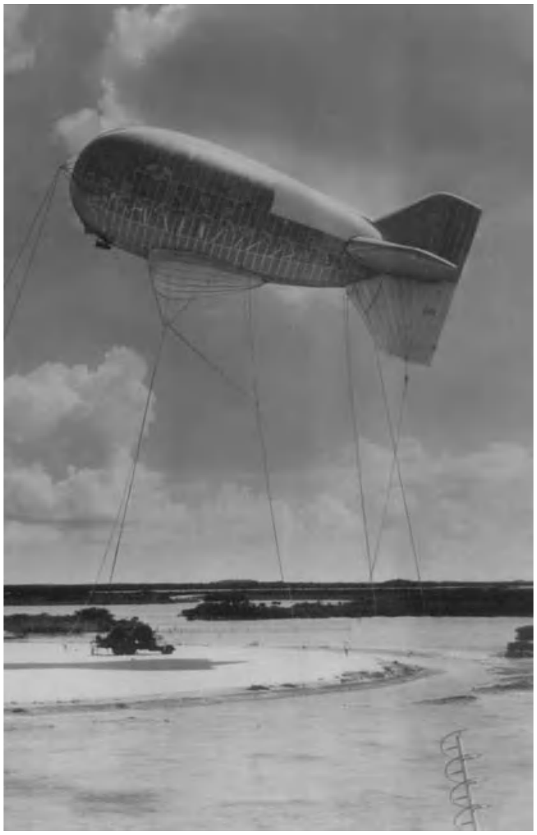


Fig. 3. Tethered balloon. (*U.S. Air Force*)

the logging industry, where heavy-lift, low-altitude balloons provide a means of removing inaccessible timber. This balloon has a lift capability of 22,000 to 44,000 lb (10,000 to 20,000 kg).

Powered balloon. Many large, crewed, lighter-than-air ships or dirigibles of both the rigid and nonrigid type were produced in the United States and Germany prior to World War II. These craft can be propelled and steered. The rigid dirigible was popular until the *Hindenburg* disaster of 1937. The nonrigid airships or blimps were used by the U.S. Navy during World War II for antisubmarine detection. With the advancement of aircraft design providing all-weather operation and superior performance, the crewed airships were driven into obscurity, but these craft are again being seriously considered for a number of civil and military applications. *See* AIRSHIP; BLIMP.

Efforts were undertaken to develop an uncrewed high-altitude (65,000-ft or 20-km) powered balloon which would take advantage of the low wind field and remain stationary over given areas for long periods of time. These platforms would be used for communication purposes, earth resource survey sensors, and military applications.

Space application. The zero-pressure and superpressure balloons have been used for space-related programs. The most extensive use has been development and testing of new detectors, experiments, and sensors which are later flown on satellites or are to be flown on the space shuttle. The balloon provides an ideal vehicle to carry x-ray, gamma-ray, cosmic-ray, infrared, and ultraviolet experiments above the Earth's atmosphere to obtain scientific data. These instruments can be recovered, redesigned, and reflown many times before commitment to a satellite program.

NASA conducted a program using zero-pressure balloons to perform in-place and remote measurements of stratospheric constituents of ozone, nitric acid, water vapor, nitrogen dioxide, and temperature. These balloon measurements were conducted simultaneously with the overpass of the *Nimbus-G* satellite, with its sensors providing remote measurements of the same data. These correlative measurements provided a means of validating the satellite data.

Another NASA program, the Global Atmospheric Research Program, involves the release of hundreds of superpressure balloons to provide measurements of pressure, temperature, and wind direction of the atmosphere. These data are transmitted to the *Tiros-N* satellite and later retransmitted to satellite ground stations to provide worldwide meteorological data. *See* METEOROLOGICAL SATELLITES. Walter R. Nagel

Bibliography. T. D. Crouch, *The Eagle Aloft: Two Centuries of the Balloon in America*, 1983; W. Reidler and K. M. Torkar (eds.), *Balloon Technology and Observations*, 1993.

Balsa

A fast-growing tree, *Ochroma lagopus*, widely distributed in tropical America, especially in Ecuador. The leaves are simple, angled, or lobed, and the flowers are large and yellowish-white or brownish and they are terminal on the branches. *See* MALVALES.

With plenty of room for growth in a rich, well-drained soil at low elevations, trees may attain a height of 50–60 ft (15–18 m) and a diameter of 24–30 in. (60–75 cm) in 5–6 years. Under such conditions, the wood is very light and soft, weighing about 6–8 lb/ft^3 (95–130 kg/m^3). Culture is important, for if the trees are injured only slightly, the wood develops a hard and fibrous texture, thereby losing its commercial value. To secure a uniform product the trees must be grown in plantations.

The wood decays easily in contact with the soil and is subject to sap stain if not promptly dried. Seasoned lumber absorbs water quickly, but this can be largely overcome by waterproofing. Balsa became popular during World War I, and large quantities were used for life preservers and mine buoys.

Balsa owes most of its present commercial applications to its insulating properties. Balsa has sound-deadening qualities, and is also used under heavy machinery to prevent transmission of vibrations. Certain other woods are lighter, but they lack the strength of balsa and are not available in sufficient size or quantity to be commercially important. The heartwood of balsa is pale brown or reddish, whereas the sapwood is nearly white, often with a yellowish or pinkish hue. Luster is usually rather high and the wood is odorless and tasteless. *See* FOREST AND FORESTRY; TREE. Arthur H. Graves/Kenneth P. Davis

Baltic Sea

A semienclosed brackish sea located in a humic zone, with a positive water balance relative to the adjacent ocean (the North Sea and the North Atlantic). The Baltic is connected to the North Sea by the Great Belt (70% of the water exchange), the Øresund (20% of the water exchange), and the Little Belt. The total area of the Baltic is 147,414 mi^2 (381,705 km^2), its total volume 4982 mi^3 (20,764 km^3), and its average depth 181 ft (55.2 m). The greatest depth is 1510 ft (459 m), in the Landsort Deep.

The topography of the Baltic is characterized by a sequence of basins separated by sills and by two large gulfs, the Gulf of Bothnia (40,100 mi^2 or 104,000 km^2) and the Gulf of Finland (11,400 mi^2 or 29,500 km^2). More than 200 rivers discharge an average of 104 mi^3 (433 km^3) annually from a watershed area of 637,056 mi^2 (1,649,550 km^2). The largest river is the Newa, with 18.5% of the total freshwater discharge. From December to May, the northern and eastern parts of the Baltic are frequently covered with ice. On the average, the area of maximum ice coverage is 82,646 km^2 (214,000 km^2). The mean maximum surface-water temperature in summer is between 59 and 63°F (15 and 17°C).

Geologic history. The Baltic Sea in its present form is a relatively young postglacial sea which developed from the melting Fenno-Scandian glaciers about 14,000 years ago. Depending on the balance between the elevation of the Fenno-Scandian land area (due to the decreasing ice weight) and the rising of the oceanic water level, the connection to the open ocean was several times closed and then reopened. Among the stages were the Baltic Ice Lake (12,000–8000 B.C.; arctic, fresh water), the Yoldia Sea (8000–7250 B.C.; arctic, oceanic), the Ancylus Sea (7250–2000 B.C.; warm, fresh water), the Littorina Sea (2000 B.C–A.D. 500; moderate, brackish to saline), and the Mya Sea (A.D. 500 to present; moderate, brackish) [see **illus.**].

Temperature. As the Baltic stretches from the boreal to the arctic continental climatic zone, there are large differences between summer and winter temperature in the surface waters, ranging from about 68 to 30°F (20 to −1°C) in the Western Baltic and 57 to 32°F (14 to −0.2°C) in the Gulf of Bothnia and the Gulf of Finland. Usually one or more sharp thermoclines develop April through June between 16 and 80 ft (5 and 25 m), separating the brackish winter water from the surface layer. Below the permanent thermohalocline, at a depth of 190 to 240 ft (60 to 75 m), the water temperature is almost constant (42 to 44°F or 5.5 to 6.5°C), not influenced by seasonal warming or cooling. Vertical convection due to the cooling of the surface water in autumn and winter penetrates to the bottom only in the inner Gulf of Bothnia and the Gulf of Finland. A significant increase in the temperature in the deep water has been observed since 1900; the increase is about 3.6°F (2°C) in the Bornholm Basin and 2.7°F (1.5°C) in the Central Baltic.

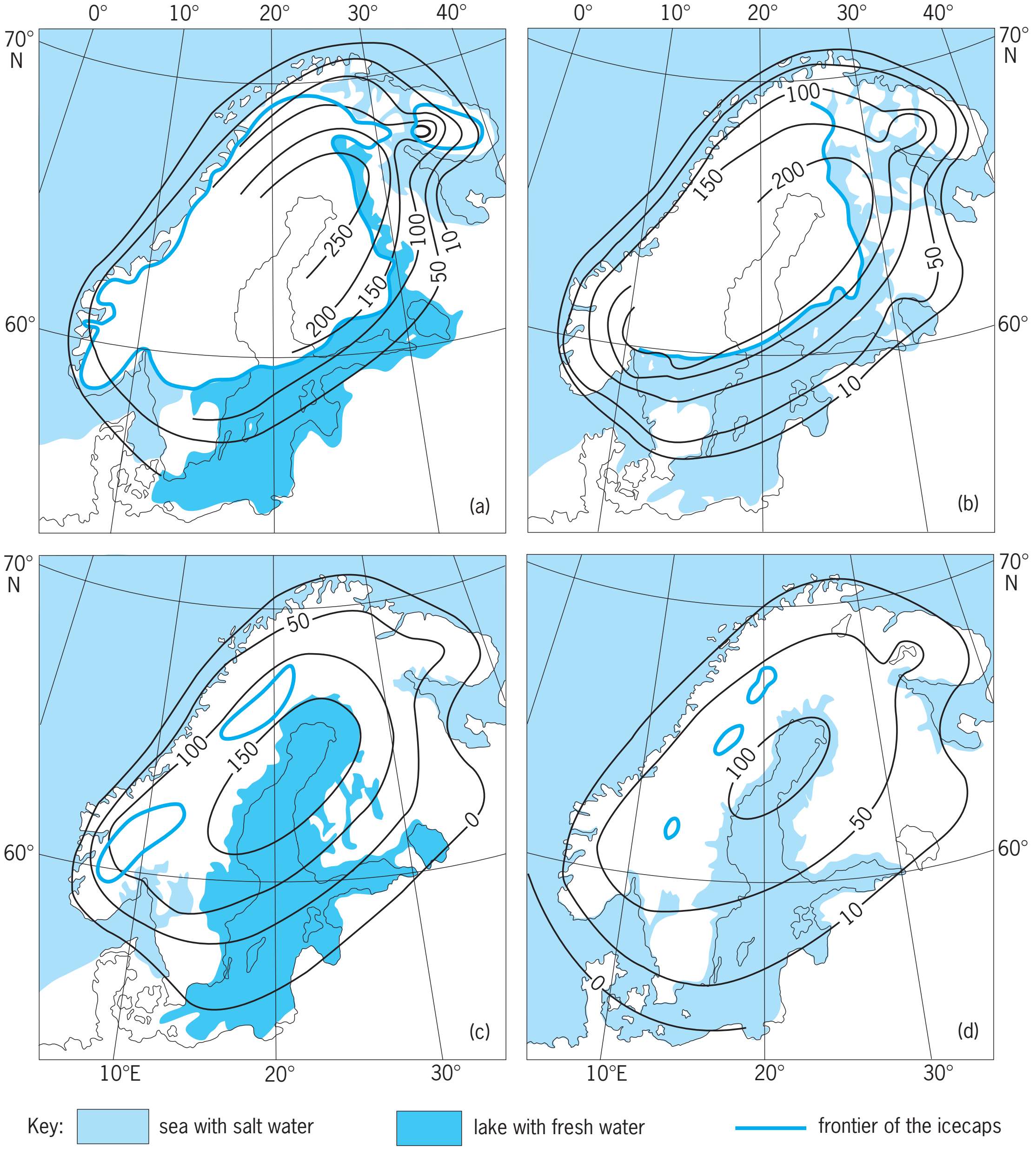
Salinity. The salt content of the Baltic waters is characterized by two major water bodies; the brackish surface water and the more saline deep water. Salinities for the surface water range from 8 to 6‰ in the Western and Central Baltic and 6 to 2000 in the Gulf of Bothnia and the Gulf of Finland; salinities for the deep water range from 18 to 13‰ in the Western and Central Baltic and 10 to 4‰ in the Gulf of Bothnia and the Gulf of Finland. The boundary between these two major water bodies is marked by a relatively steep and permanent halocline which to a large extent prevents vertical exchange of water and dissolved substances. The total average annual outflow of brackish surface water from the Baltic is 220 mi^3 (942 km^3), and the salt balance of the Baltic is maintained by an average annual inflow of 113 mi^3 (471 km^3). Most of the oceanic water entering the Belt Sea does not reach the Baltic proper. Major intrusions of oceanic water happen only every 5 to 10 years; they are caused by the coincidence of several meteorologic factors during the cold season.

Circulation. The surface currents of the Baltic are dominated by a general counterclockwise movement and by local and regional wind-driven circulations. A complex system of small- and medium-scale gyres develops especially in the central parts of the Baltic. The currents in the Belt Sea are dominated by the topography; they are due to sea-level differences between the Baltic proper and the North Sea. Tides are of minor importance, ranging between 0.8 and 4.7 in. (2 and 12 cm). Water-level changes of more than 6 ft (2 m) occur occasionally as a result of onshore or offshore winds and the passage of cyclones over the Baltic Sea area. The frequency of longitudinal sea-level oscillations is about 13.5 h. *See* OCEAN CIRCULATION.

A significant increase of 1–2‰ in both surface and deep water salinities, probably caused by long-term changes in meteorological conditions and by the decrease in precipitation, has been observed since 1900.

Oxygen and nutrients. Due to the limited exchange of deep water with the North Sea and the permanent halocline, the oxygen content of the water below the halocline is considerably reduced. Long-term records show an average annual oxygen deficit of 4.9×10^{10} oz (1.4×10^{12} g) for the Baltic proper. Anaerobic conditions develop frequently in depths below 380 ft (120 m), wiping out all benthic and aquatic life except anaerobic bacteria in large parts of the Central Baltic deep waters. There is a significant increasing trend in the nutrient content of Baltic deep, intermediate, and surface waters, resulting in increased eutrophication. This increase is caused partly by increasing stagnation and partly by human activities. *See* EUTROPHICATION.

Ecology. The flora and fauna of the Baltic are those of a typical brackish-water community, with considerably reduced numbers of species compared to an oceanic community. The productivity is relatively low compared to other shelf seas. The major commercially exploited species are cod, herring, sprat, flounder, eel, and salmon, and some fresh-water species such as whitefish, pike, perch, and trout. The total annual catch amounts to about 880,000 tons (800,000 metric tons). The Baltic is completely



Postglacial development of the Baltic: (*a*) Baltic Ice Sea, 12,000–8000 B.C.; (*b*) Yoldia Sea, 8000–7250 B.C.; (*c*) Ancylus Sea, 7250–2000 B.C.; (*d*) Littorina Sea, 2000–A.D. 500. (*After K. Grasshoff, in S. P. Riley and G. Skirrow, eds., Chemical Oceanography, vol. 2: The Hydrochemistry of Landlocked Basins and Fjords, Academic Press, 1975*)

divided into fishery zones, with exclusive fishing rights belonging to the respective countries.

Pollution. The water pollution of the Baltic is generally decreasing as a result of the Convention for the Protection of the Marine Environment of the Baltic; this covered both land-based and air-borne pollution, and was signed in 1974 by all seven Baltic states. It came into power on May 3, 1980. The major sources of pollution are domestic and industrial wastewater. *See* WATER POLLUTION.

Resources. Other than fish the only major resources that have been exploited are sand and gravel in the Western Baltic Sea. It is believed that the deeper layer under the Gotland Basin contains mineral oil, but so far only exploratory drilling has been carried out in the near-coastal regions. Limited amounts of mineral oil have also been located in the Gulf of Kiel.

K. Grasshoff

Bibliography. G. Alexandersonn, *The Baltic Straits*, 1982; S. P. Riley and G. Skirrow (eds.), *Chemical Oceanography*, vol. 2: *The Hydrochemistry of Landlocked Basins and Fjords*, 1975; A. Voipio (ed.), *The Baltic Sea*, 1981.

Bamboo

The common name of various perennial, ornamental grasses (Gramineae). There are five genera with approximately 280 species. They have a wide distribution, but occur mainly in tropical and subtropical parts of Asia, Africa, and America, extending from sea level to an elevation of 15,000 ft (4572 m). Their greatest development occurs in the monsoon regions of Asia. The plants grow very rapidly. From the jointed rhizome, numerous straight, usually erect, stems arise, which at their full height produce dense masses of horizontal branches. The giant bamboo (*Dendrocalamus giganteus*), the largest known grass, attains a height of 120 ft (36 m) and a diameter of 8–12 in. (20–30 cm). Most are woody; a few are herbaceous or climbing. The economic uses of bamboo are numerous and varied. The seeds and young shoots are used as food and the leaves make excellent fodder for cattle. In varying sizes, the stems are used for pipes, timber, masts, bows, furniture, bridges, cooking vessels, buckets, wickerwork, paper pulp, cordage, and weaving. Entire houses are made of bamboo stems. Certain bamboos have been naturalized in California, Louisiana, and Florida. *See* CYPERALES. Perry D. Strausbaugh/Earl L. Core

Banana

A large tropical plant; also its edible fruit, which occurs in hanging clusters, is usually yellow when ripe, and is about 6–8 in. (15–20 cm) long. The banana belongs to the family Musaceae. The banana of commerce (*Musa sapientum*), believed to have originated in the Asian tropics, was one of the earliest cultivated fruits. For commercial production the plant requires a tropical climate within the temperature range 50–105°F (10–40°C) and a constant supply of moisture by rainfall or irrigation. Bananas are subject to mechanical injury by strong winds which tear the leaves or blow down the plants. *See* IRRIGATION (AGRICULTURE); ZINGIBERALES.

The plant portion above the ground is a false stem (pseudostem) consisting of several concentrically formed leaves, from the center of which develops the inflorescence stalk (**Fig. 1**). The rhizome or true stem is underground. Near the tip of the flower stalk are several groups of sterile male flowers subtended by brilliant purple bracts. The lower female flower clusters on the same stalk give rise to the fruit and contain aborted stamens (male organs). The single fruits are called fingers, a single group of 8–12 fingers is termed a hand, and the several (6–18) hands of the whole inflorescence make up the stem.

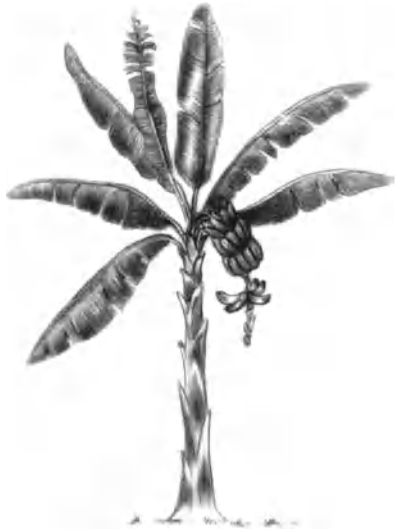


Fig. 1. Pseudostem of commercial banana plant (*Musa sapientum*), showing characteristic foliage and single stem of bananas.

After the single fruiting stalk has produced the fruit bunch, the whole pseudostem is cut off at the ground, allowing one of several new buds or suckers from the underground stem to develop into a new plant. The fruit bunch requires 75–150 days to mature and must be removed from the plant to ripen properly. Chilled banana fruits do not soften normally; hence for best edibility the fruit is kept well ventilated at room temperature.

Kinds and distribution. Banana fruits of commerce set without pollination, by parthenocarpy, and hence are seedless. When mature, most varieties are yellow, although fine red-skinned types are well known. There are several hundred varieties grown throughout the world. The Cavendish banana (*M. nana*, variety Valery) is becoming important in the American tropics. The more starchy bananas, known as plantains, must be cooked before they can be eaten.

The yearly world production averages nearly 100,000,000 bunches, over 60% of which is consumed in the United States. Three-fourths of the world's bananas is grown in the Western Hemisphere. The greatest production occurs in Ecuador, Guatemala, Honduras, and other tropical Central and South American countries. Commercial production of bananas has been attempted in Florida, but climatic conditions there do not allow continuous economical cultivation of the fruit. *See* FRUIT; FRUIT, TREE. Charles A. Schroeder

Diseases. F. L. Wellman indicates that there are between 200 and 220 diseases of bananas in the tropics and some 80 organisms have been associated with banana diseases, but only 3 or 4 have been involved in major epidemics. A new form of fungus leaf spot more damaging than the Sigatoka disease has occurred in several areas and has destroyed many plantings before adequate control measures could be undertaken. This disease is referred to as the black Sigatoka. The less destructive but important Sigatoka disease occurs in most banana-growing districts, and it becomes a major factor in banana losses in the high-rainfall areas. Without adequate controls and somewhat resistant cultivars, bananas could not be grown commercially. The Panama disease, caused by a soil fungus, has forced plant breeders to produce new clones so that growers could cope with the rapid spread and destruction of susceptible bananas. New varieties such as Valery have replaced the once popular Gros Michel in Central and South America. A bacterial disease called Moko has caused widespread damage in many banana-growing areas of Central and South America, and has made it necessary for growers to seek soils that have not been previously planted to bananas and to use disease-free

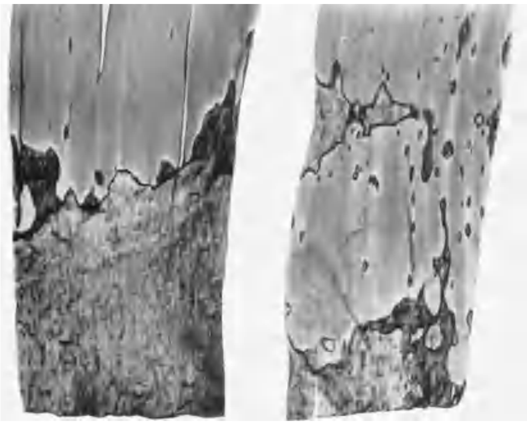


Fig. 2. Sigatoka disease, banana with spotting and necrosis of marginal areas on leaves of Gros Michel variety. (*Division of Tropical Research*, *United Fruit Co.*)

planting stock, and to follow strict sanitary cultural practices in order to maintain plantings. Tremendous efforts by all segments of the banana industry, coupled with millions of dollars spent on research and development, have made it possible to produce bananas economically despite the devastating effects of these diseases.

Etiological agents. Diseases of the banana roots, rhizome, pseudostem, leaves, and fruit are caused by 7 species of bacteria, 57 fungi, 2 viruses, and 14 nematodes. The most important and destructive diseases in commercial plantings include: Sigatoka disease or leaf spot, caused by the fungus *Mycosphaerella musicola* (**Fig. 2**) and its close but more virulent relative *M. fijiensis* var. *difformis*; Fusarium wilt or Panama disease, caused by the soil fungus *Fusarium oxysporum* f. sp. *cubense*; Moko disease or bacterial wilt, caused by *Pseudomonas solanacearum* (**Fig. 3**); bunchy top, caused by the aphid-transmitted bunchy top virus; infectious chlorosis, caused by a strain of cucumber mosaic virus; and rhizome and root rots, caused by the burrowing nematode (*Radopholus similis*) and other nematodes. Fruit spotting and decay are important in postharvest handling and shipping, often to distant markets. Many fungi are involved; some are wound parasites so that care should be taken to avoid injuries.

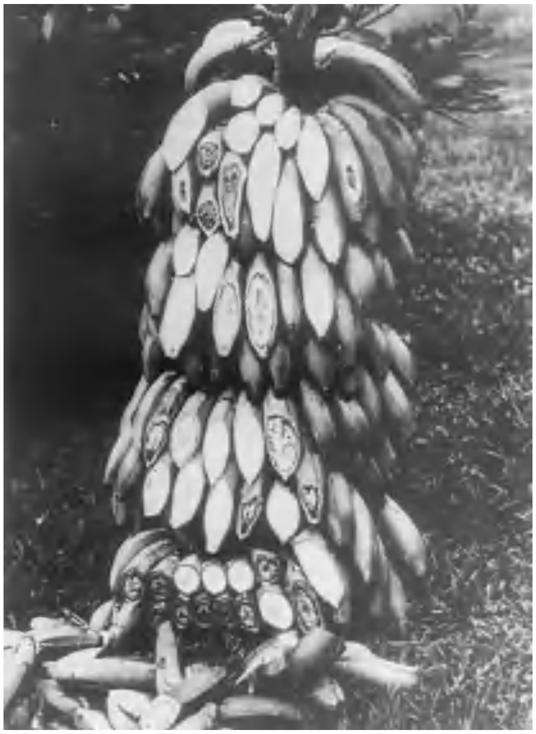


Fig. 3. Moko disease, with internal browning of banana fruit (Gros Michel variety).

Pathogenic forms and host reactions. Many of the organisms causing banana diseases are specific types that attack only bananas and their near relatives, plantains and abaca. Some of these organisms form distinct races that may occur in different geographic areas and may attack different banana cultivars or clones and some of their mutants.

Control methods. Plant breeders have been busy worldwide producing an acceptable fruit that is resistant to the devastating diseases. *Musa* Gros Michel AAA (a commercially excellent type of banana) and its tetraploid dwarf *Musa* Highgate AAAA are both susceptible to Panama disease, whereas *Musa* Giant Cavendish AAA and *Musa* Dwaft Cavendish AAA are somewhat resistant. All varieties of these genetic types are susceptible to the Sigatoka disease. If Gros Michel is crossed with a wild-type (*M. acuminata*), the progeny is resistant to both Panama and Sigatoka diseases but is not commercially acceptable. Some resistance to the burrowing nematode has been found in the cultivar Pisang Jari Buaya, and selections of the bluggoe-type plantain have been used in breeding because of its resistance to both the Panama and Moko diseases.

Despite early banana breeding programs and the awareness of resistance to Panama disease in the triploids of the Cavendish-Lacatan clones, these types were not used extensively because of their poor shipping and storage qualities. Studies in fruit physiology and handling procedures have made it possible to use these and other resistant clones, and Panama disease is no longer important in most banana-growing regions. Black Sigatoka and Sigatoka leaf spots continue to take their tolls unless protective fungicides are used on a regular basis, since all commercial banana types are susceptible to these leaf spots. In some humid areas, a foliar application of fungicide may be necessary on a weekly basis. Commercial companies have pioneered in the development of sophisticated equipment to apply control chemicals. Benzimidazoles, chlorothalonils, dithiocarbamates, copper compounds, imidazoles, and oils have all been used effectively against the leaf diseases, but fungicides do not weather well and the fungicidal activity is soon lost to heavy rainfall in many banana-growing areas. There is some hope that systemic fungicide formulations will give additional control in the future. Presently used banana clones are susceptible to Moko disease and many of the species of nematodes. These diseases are controlled by sanitation and the use of disease-free planting stock. Nematicides have been effective in reducing populations in the field and eliminating them from the offshoots used as planting stock. The inability to “cure” virus diseases has made it necessary to eradicate diseased plants, eliminate weed hosts, and control the insect vectors. Storage and transit losses are reduced by controlling the post-harvest environment by using fungicidal dips and cut-surface applications, polyethylene wraps, wax coatings, and careful handling and boxing. *See* PLANT PATHOLOGY.

J. E. Dimitman

Bibliography. N. W. Simmonds, *Bananas*, 2d ed., 1982; R. H. Stover, *Banana, Plantain, and Abaca Diseases*, 1972.

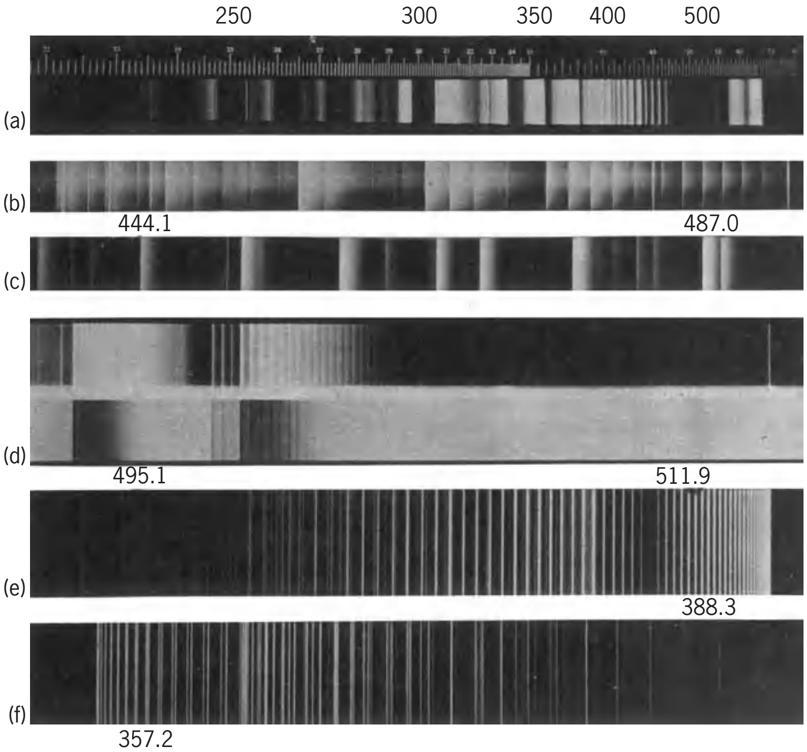
Band spectrum

A spectrum consisting of groups or bands of closely spaced lines. Band spectra are characteristic of molecular gases or chemical compounds. When the light emitted or absorbed by molecules is viewed through a spectroscope with small dispersion, the spectrum appears to consist of very wide asymmetrical lines called bands. These bands usually have a maximum intensity near one edge, called a band head, and a gradually decreasing intensity on the other side. In some band systems the intensity shading is toward shorter waves, in others toward longer waves. Each band system consists of a series of nearly equally spaced bands called progressions; corresponding bands of different progressions form groups called sequences.

Six spectra of diatomic molecular fragments are shown in the **illustration**. The spectrum of a discharge tube containing air at low pressure is shown in illus. *a*. It has four band systems: the γ-bands of nitrogen oxide (NO, 230–270 nanometers), negative nitrogen bands (N_2^+, 290–350 nm), second-positive nitrogen bands (N_2, 290–500 nm), and first-positive nitrogen bands (N_2, 550–700 nm). The spectrum of high-frequency discharge in lead fluoride vapor in *b* has bands in prominent sequences. The spectrum in *c* shows part of one band system of SbF, and was obtained by vaporizing SbF into active nitrogen. Emission from a carbon arc cored with barium fluoride (BaF_2) and absorption of BaF vapor in an evacuated steel furnance are illustrated in *d*. These spectra were obtained in the second order of a diffraction grating, as were the spectra in *e* and *f*. The photograph *e* is that of the CN band at 388.3 nm from an argon discharge tube containing carbon and nitrogen impurities, and *f* is a band in ultraviolet spectrum of NO, obtained from glowing active nitrogen containing a small amount of oxygen.

When spectroscopes with adequate dispersion and resolving power are used, it is seen that most of the bands obtained from gaseous molecules actually consist of a very large number of lines whose spacing and relative intensities, if unresolved, explain the appearance of bands of continua (parts *e* and *f* of the illustration). For the quantum-mechanical explanations of the details of band spectra *See* MOLECULAR STRUCTURE AND SPECTRA.

W. F. Meggers; W. W. Watson



Photographs of band spectra of (*a*) a discharge tube containing air at low pressure; (*b*) high-frequency discharge in lead fluoride vapor; (*c*) SbF (*b* and *c* taken with large quartz spectrograph, after Rochester); (*d*) BaF emission and absorption; (*e*) CN; and (*f*) NO. The measurements are in nanometers. (*From F. A. Jenkins and H. E. White, Fundamentals of Optics, 3d ed., McGraw-Hill, 1957*)

Band theory of solids

A quantum-mechanical theory of the motion of electrons in solids. Its name comes from the fact that it predicts certain restricted ranges, or bands, for the energies of electrons in solids.

Suppose that the atoms of a solid are separated from each other to such a distance that they do not interact. The energy levels of the electrons of this system will then be those characteristic of the individual free atoms, and thus many electrons will have the same energy. Now imagine the atoms being slowly brought closer together. As the (equal) distance between atoms is decreased, the electrons in the outer shells begin to interact with each other. This interaction alters their energy and, in fact, broadens the sharp energy level out into a range of possible energy levels called a band. One would expect the process of band formation to be well advanced for the outer, or valence, electrons at the observed interatomic distances in solids. Once the atomic levels have spread into bands, the valence electrons are not confined to individual atoms, but may jump from atom to atom with an ease that increases with the increasing width of the band.

Although energy bands exist in all solids, the term energy band is usually used in reference only to ordered substances, that is, those having well-defined crystal lattices. In such a case, a basic theorem due to F. Bloch asserts that an electron energy state can be classified according to its crystal momentum $\mathbf{p}$ or its electron wave vector $\mathbf{k} = \mathbf{p}/\hbar$ (where $\hbar$ is Planck's constant h divided by 2π). If the electrons were free, the energy of an electron whose wave vector is $\mathbf{k}$ would be as shown in Eq. (1), where E_0 is the en-

$$E(\mathbf{k}) = E_0 + \frac{\hbar^2\mathbf{k}^2}{2m_0} \qquad (1)$$

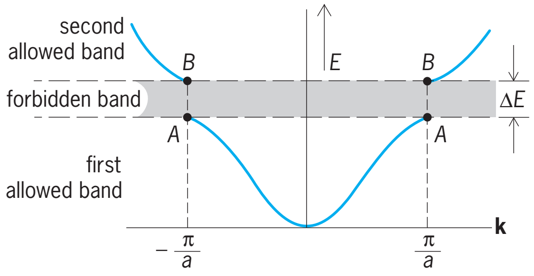
ergy of the lowest state of a valence electron and m_0 is the electron mass. In a crystal, although the electrons may move fairly independently, they are not free because of the effect of the crystal binding and the forces exerted on them by the atoms; consequently, the relation $E(\mathbf{k})$ between energy and wave vector is more complicated. The statement of this relationship constitutes the description of an energy band. $E(\mathbf{k})$ can be calculated as the eigenvalue of a one-electron Schrödinger equation. *See* BLOCH THEOREM.

A knowledge of the energy levels of electrons is of fundamental importance in computing electrical, magnetic, optical, or thermal properties of solids.

Allowed and forbidden bands. The ranges of possible electron energy levels in a solid are called allowed energy bands. It often happens that there are also ranges of energy which it is impossible, for all directions of motion, for an electron to have in a given crystal. Such ranges are called forbidden bands, or gaps. The allowed energy bands sometimes overlap and sometimes are separated by forbidden bands. The presence of a forbidden band, or energy gap, immediately above the occupied allowed states (such as the region A to B in the **illustration**) is the principal difference in the electronic structures of a semiconductor or insulator and a metal. In the first two substances there is a gap for all directions of motion, between the valence band or normally occupied states and the conduction band, which is normally unoccupied. In a metal there is no overall gap between occupied and unoccupied states.

The presence of a gap means that the electrons cannot easily be accelerated into higher energy states by an applied electric field. Thus, the substance cannot carry a current unless electrons are excited across the gap by thermal or optical means.

Effective mass. When an external electromagnetic field acts upon the electrons in a solid, the resultant motion is not what one would expect if the electrons were free. In fact, the electrons behave as though they were free but with a different mass, which is called the effective mass. This effective mass can be obtained from the dependence of electron energy on the wave vector, $E(\mathbf{k})$, in the following way.



Electron energy *E* versus wave vector k for a monatomic linear lattice of lattice constant *a*. (*After C. Kittel, Introduction to Solid State Physics, 6th ed., Wiley, 1986*)

Suppose there is a maximum or minimum of the function $E(\mathbf{k})$ at the point $\mathbf{k} = \mathbf{k}_0$. The function $E(\mathbf{k})$ can be expanded in a Taylor series about this point. For simplicity, assume that $E(\mathbf{k})$ is a function of the magnitude of $\mathbf{k}$ only, that is, is independent of the direction of $\mathbf{k}$. Then, by dropping terms higher than second order in the Taylor series, Eq. (2) results. By

$$E(\mathbf{k}) = E(\mathbf{k}_0) + {}^1/_2(\mathbf{k} - \mathbf{k}_0)^2 \left(\frac{d^2E}{dk^2}\right)_{\mathbf{k}_0} \qquad (2)$$

analogy with Eq. (1), a quantity m^* with the dimensions of a mass can be defined by relation (3).

$$\frac{\hbar^2}{m^*} = \left(\frac{d^2E}{dk^2}\right)_{\mathbf{k}_0} \qquad (3)$$

The quantity m^* is called the effective mass of electrons at $\mathbf{k}_0$. For many simple metals, the average effective mass is close to m_0, but smaller effective masses are not uncommon. In indium antimonide, a semiconductor, the effective mass of electrons in the conduction band is $0.013\ m_0$. In a semiclassical approximation, an electron in the solid responds to an external applied electric or magnetic field as though it were a particle of mass m^*. The equation of motion of an electron is shown in Eq. (4), where $\mathbf{v}$ is the

$$m^*\frac{d\mathbf{v}}{dt} = e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \qquad (4)$$

electron velocity, $\mathbf{E}$ the electric field, $\mathbf{B}$ the magnetic induction, and e the charge of the electron.

Quite frequently, the energy $E(\mathbf{k})$ does depend upon the direction of $\mathbf{k}$. In such a case, the effective mass is a tensor whose components are defined by Eq. (5). Equation (4) remains valid with a tensor

$$\frac{\hbar^2}{m_{ij}{}^*} = \left(\frac{\partial^2 E}{\partial k_i \partial k_j}\right)_{\mathbf{k}_0} \qquad (5)$$

m^*. Bismuth is an example of a metal in which the effective mass depends strongly on direction.

Transitions between states. Under external influences, such as irradiation, electrons can make transitions between states in the same band or in different bands. The interaction between the electrons and the vibrations of the crystal lattice can scatter the electrons in a given band with a substantial change in the electron momentum, but only a slight change in energy. This scattering is one of the principal causes of the electrical resistivity of metals. *See* ELECTRICAL RESISTIVITY.

An external electromagnetic field (for example, visible light) can cause transitions between different bands. In such a process, crystal periodicity leads to the requirement that the sum of the electron crystal momentum $\hbar\mathbf{k}$ and the photon momentum must be conserved. Because the momentum of a photon $h\nu/c$ (where ν is the frequency of the light and c its velocity) is quite small, the momentum of the electron before and after collision is nearly the same. Such a transition is called vertical in reference to an energy band diagram. Conservation of energy must also hold in the transition, so absorption of light is possible only if there is an unoccupied state of energy $h\nu$ available at the same **k** as the initial state. These transitions are responsible for much of the absorption of light by semiconductors in the visible and near-infrared region of the spectrum.

Energy band calculation. As is the case for any quantum-mechanical system, the energy levels of electrons in solids are determined in principle by the solution of the Schrödinger wave equation for the system. However, the enormous number of particles involved makes exact solution of this equation impossible. Approximate methods must be employed. The atoms are regarded as fixed in their equilibrium positions. Each electron is regarded as having an individual wave function $\psi_n(\mathbf{k},\mathbf{r})$, in which **k** is the wave vector and the index n designates a particular band. The wave function $\psi_n(\mathbf{k},\mathbf{r})$ is frequently called a Bloch function. The wave function of the many-electron system is written as a determinant of Bloch functions to satisfy the requirements of the Pauli exclusion principle. The general variational method of quantum mechanics may be employed to derive an equation for the individual Bloch functions. This equation, which is known as the Hartree-Fock equation, is similar to a one-electron Schrödinger equation in which each electron moves in the field of all the nuclei of the system and the average field of all the other electrons. An additional term, exchange, takes account of the reduction of the average electronic repulsion, because the probability of close encounters of electrons of parallel spin is reduced by the Pauli exclusion principle.

The Hartree-Fock equations are still quite complicated, and must be solved by the self-consistent field method. In this approach, some reasonable distribution of electrons is assumed to exist. An effective potential $V(\mathbf{r})$ can be calculated from this distribution, including the contribution from the nuclei. Usually it is an adequate approximation to include the exchange terms in this potential. Then the Bloch functions may be formed by solving Eq. (6), in which

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi_n(\mathbf{k},\mathbf{r})+V(\mathbf{r})\psi_n(\mathbf{k},\mathbf{r})=E_n(\mathbf{k})\psi_n(\mathbf{k},\mathbf{r})\qquad(6)$$

$V(\mathbf{r})$ is the potential described above and $E_n(\mathbf{k})$ is the energy of an electron in band n having wave vector **k**. The potential $V(\mathbf{r})$ is periodic in space with the periodicity of the crystal. The wave function $\psi_n(\mathbf{k},\mathbf{r})$ obtained from Eq. (6) yields a new electron distribution from which the potential $V(\mathbf{r})$ may be calculated again. The process is repeated until the potential used in the solution of Eq. (6) agrees with that obtained from the solution $\psi_n(\mathbf{k},\mathbf{r})$ to sufficient accuracy.

The local density approximation is frequently used to obtain this potential. In the simplest formula of this type, the Kohn-Sham potential, $V(\mathbf{r})$ is given by Eq. (7).

$$V(\mathbf{r})=-e^2\sum_{\mu}\frac{Z\mu}{|\mathbf{r}-\mathbf{R}_\mu|}+e^2\int\frac{p(\mathbf{r}')d^3r'}{|\mathbf{r}-\mathbf{r}'|}-2e^2\left(\frac{3\rho(\mathbf{r})}{8\pi}\right)^{1/3}\qquad(7)$$

The first term represents the potential energy of an electron in the field of all the nuclei of the system. These nuclei are located at sites $\mathbf{R}_\mu$ and have charge Z_μ. The second term contains the average electronic repulsion of the electron distribution, and $\pi(\mathbf{r})$ is the electron density. The third term approximately describes the exchange interaction, referred to above.

The self-consistent-field procedure is evidently quite complicated. Since it may not be practical to repeat the calculations to obtain a self-consistent solution, it is desirable to choose an initial $V(\mathbf{r})$ which approximates the actual physical situation as closely as possible. Choice of an adequate crystal potential is the chief physical problem in the calculation of energy levels of electrons in solids.

Several techniques are available for solving wave equation (6) with a given $V(\mathbf{r})$. Those in common use include the Green's function method, the augmented plane wave method, the orthogonalized plane wave method, and the linear combination of atomic orbitals method. It is also possible to use experimentally determined energy levels and effective masses to obtain a suitable potential for use in Eq. (6). This procedure, which bypasses many of the difficulties of the self-consistent-field approach, is known as the pseudopotential method and is now widely used. In simple metals (not transition metals) and wide-band semiconductors, such as germanium and silicon, the pseudopotential is rather weak, and many properties can be calculated by using perturbation theory.

Density of states. Many properties of solids, including electron specific heat and optical and x-ray absorption and emission, are related rather simply to a basic function known as the density of states. This function, here denoted $G(E)$, is defined so that the number of electronic energy states in the range between E and $E+dE$ is $G(E)dE$. It can be shown that $G(E)$ is given by Eq. (8), in which $E(\mathbf{k})$ is the energy

$$G(E)=\frac{\Omega}{4\pi^3}\int\delta[E-E(\mathbf{k})]\,d^3k=\frac{\Omega}{4\pi^3}\int\frac{dS_k(E)}{|\nabla_kE(\mathbf{k})|}\qquad(8)$$

band function, δ is the Dirac delta function, Ω is the volume of a unit cell of the solid, and the constant multiplying the integral has been chosen so that the number of states is two (one for each spin direction) for each energy band. The first integral in Eq. (8) is taken over the Brillouin zone, the second over a surface of constant energy E in the zone. The density of states will show structure in the neighborhood of energies corresponding to points where $|\nabla_kE(\mathbf{k})|$ vanishes. Such points, which are required to exist by reasons of crystal symmetry and periodicity, are known as Van Hove singularities. The energies of states at which Van Hove singularities occur can be determined from optical and (sometimes) x-ray measurements.

Experimental information. A considerable amount of experimental information has been obtained concerning the band structures of the common metals and semiconductors. In metals, experiments usually determine properties of the Fermi surface (the surface in **k** space bounding the occupied states at zero temperature), since only electrons in states near the top of the Fermi distribution are able to respond to electric and magnetic fields. For example, the magnetic susceptibility of a metal exhibits a component which is an oscillatory function of the reciprocal of the magnetic field strength when the field is strong and the temperature is quite low, typically a few kelvins (the de Haas–van Alphen effect). Measurement of the period of these oscillations determines the area of a cross section of the Fermi surface in a plane perpendicular to the direction of the magnetic field.

Other properties of the Fermi surface are obtained by studying the increase of the electrical resistance in a magnetic field, and from the magnetic field dependence of ultrasonic attenuation. The density of states (number of electron states per unit energy) and effective mass at the Fermi energy can be found from measurements of the electron contribution to the specific heat. Effective masses of electrons and holes can be determined by cyclotron resonance measurements in both semiconductors and metals.

Optical and x-ray measurements enable the determination of energy differences between states. The smallest width of the characteristic energy gap between valence and conduction bands in a semiconductor can be determined by measuring the wavelength at which the fundamental absorption begins. Application of a strong uniform magnetic field will produce an oscillatory energy dependence of this absorption, from which a precise determination of effective masses is possible.

In both semiconductors and metals, states removed in energy from the Fermi level can be studied optically. Measurements of the reflectivity will show structure at energies corresponding to transitions at symmetry points of the Brillouin zone where Van Hove singularities are likely to occur. To reveal such structure clearly against a background which frequently may be large, it is useful to subject the solid to a periodic perturbation, possibly an electric field, stress, or a change in temperature. This modifies the band structure slightly, particularly near Van Hove singularities, which tend to be sensitive to perturbations. As a result, the optical properties of the sample are slightly modulated, and this modulation is detected directly. It is possible to apply this technique in the soft x-ray wavelength region as well. Modulation of the soft x-ray emission by alternating stress (the piezo soft x-ray effect) enables the determination of the energies of transitions between band and deep core states.

Photoemission measurements can also yield information about energy bands below the Fermi energy. In these experiments, an electron is excited by a photon from a state of energy, E_i, below the Fermi energy to a state of high enough energy, E_f, such that the electron can escape from the crystal, usually from within a few tenths of a nanometer from the surface. The relation between the energies is $E_f = E_i + \hbar\omega$, where ω is the angular frequency associated with the light. The number of emitted electrons is studied as a function of the angle between the direction of the outgoing electrons and that of the incident photon (often chosen to be perpendicular to the surface). The polarization of the light can also be controlled. If the surface of a single crystal sample is plane, the component of the electron momentum parallel to the solid ($\mathbf{k}_\parallel$) is conserved. The measurement is then sensitive to the distribution of electrons in energy along a line in the Brillouin zone parallel to the surface normal and with the specified value of $\mathbf{k}_\parallel$. The results can be interpreted to give the positions of the energy band in the initial state. It is also possible to determine energies of electrons localized in the surface region.

Limitations of band theory. The results of energy-band calculations for ordinary metals usually predict Fermi surfaces and other properties that agree rather well with experiment. In addition, cohesive energies and values of the lattice constant in equilibrium can be obtained with reasonable accuracy, and, in the case of ferromagnetic metals, calculated magnetic moments agree with experiment. However, there are significant discrepancies between theoretical calculations and experiments for certain types of systems, as follows.

1. There are a number of metallic compounds containing rare-earth or actinide elements ($CeCu_6$ and UPt_3 are examples) in which the electron contribution to the specific heat is extremely large. These materials are known as heavy-fermion systems. If the specific heat is interpreted in terms of an electron effective mass, the values obtained may range from several hundred to more than a thousand times the free-electron mass. Band calculations have been successful in predicting the shapes of the observed Fermi surfaces, but cannot account for the large masses.

2. Band calculations for semiconductors such as silicon, germanium, and gallium arsenide (GaAs) predict values for the energy gap between valence and conduction bands in the range one-half to two-thirds of the measured values.

3. The occurrence of superconductivity at low temperatures in a metal indicates a breakdown of single-particle physics due to the formation of a collective or condensed state of electron pairs. It cannot be predicted, or even be adequately described phenomenologically within energy band theory. However, the single-particle states furnish a useful framework for the construction of a theory of electron pairing, and they are useful in describing the normal state above the transition temperature.

4. There are many systems, especially compounds of 3*d* transition elements, for which band calculations predict that some bands are only partially filled, and therefore should exhibit metallic behavior, but which in fact are insulators with antiferromagnetic structures. Cobaltous oxide (CoO) is an example. Materials of this type are called Mott-Hubbard insulators.

They are better described in terms of localized electron states on the atoms or ions in the system, rather than in terms of extended, band states. *See* SUPERCONDUCTIVITY.

In all the cases discussed, the failures of band theory are attributed to an inadequate treatment of strong electron-electron interactions. For example, in the Mott-Hubbard insulators, electron interactions are believed to lead to electron localization, and the occurrence of atomic magnetic moments. In conventional superconductors, the interaction of electrons with lattice vibrations is believed to lead to an effective attractive interaction between electrons, which in turn leads to the formation of pairs. The origin of the interactions leading to high-temperature superconductivity (as in $YBa_2Cu_3O_7$) is not yet understood. *See* BRILLOUIN ZONE; COHESION (PHYSICS); CRYSTAL STRUCTURE; EXCITON; FERROMAGNETISM; FREE-ELECTRON THEORY OF METALS; HOLE STATES IN SOLIDS; KRONIG-PENNEY MODEL; NONRELATIVISTIC QUANTUM THEORY; PHOTOEMISSION; SEMICONDUCTOR; SPECIFIC HEAT OF SOLIDS. Joseph Callaway

Bibliography. P. W. Anderson, *Concepts in Solids*, 1964; N. W. Ashcroft, *Solid State Physics*, 2d ed., 2000; R. H. Bube, *Electrons in Solids*: *An Introductory Survey*, 3d ed., 1992; G. Burns, *Solid State Physics*, 1985; J. Callaway, *Quantum Theory of the Solid State*, 2d ed., 1991; J. R. Hook and H. E. Hall, *Solid State Physics*, 2d ed., 1995; W. Jones and N. H. March, *Theoretical Solid State Physics*, 1973, reprint 1985; C. Kittel, *Quantum Theory of Solids*, 2d ed., 1987.

Banded iron formation

A sedimentary rock that was commonly deposited during the Precambrian. It was probably laid down as a colloidal iron-rich chemical precipitate, but in its present compacted form it consists typically of equal proportions of iron oxides (hematite or magnetite) and silica in the finely crystalline form of quartz known as chert. Its chemical composition is 50% silicon dioxide (SiO_2) and 50% iron oxides (Fe_2O_3 and Fe_3O_4), to give a total iron content of about 30%. Banding is produced by the concentration of these two chemical components into layers about 1–5 cm (1/2–2 in.) thick; typical banded iron formation consists of pale silica-rich cherty bands alternating with black to dark red iron-rich bands (**Fig. 1**). These contrasting layers are sharply defined, so that the rock has a striped appearance; banded iron formation is normally a hard, tough rock, highly resistant both to erosion and to breaking with a hammer.

Varieties and nomenclature. Historically, different names were applied to banded iron formations in different regions. The itabirites of Brazil, the ironstones of South Africa, and the banded hematite quartzites of India are now generally identified as banded iron formations. The iron formations of the Lake Superior area ranges (such as Marquette, Gogebic, Cuyuna, Mesabi, and Gunflint), which were the first to be systematically described, are a special case. In these the banding is usually coarser and less regular, and much of the constituent material consists of fine (1–2 mm) rounded grains of both silica and iron oxides, giving rise to the name granule or pelletal iron formation; the extensive banded iron formations of Labrador are also of this type. This rock is sometimes called taconite by mining engineers.

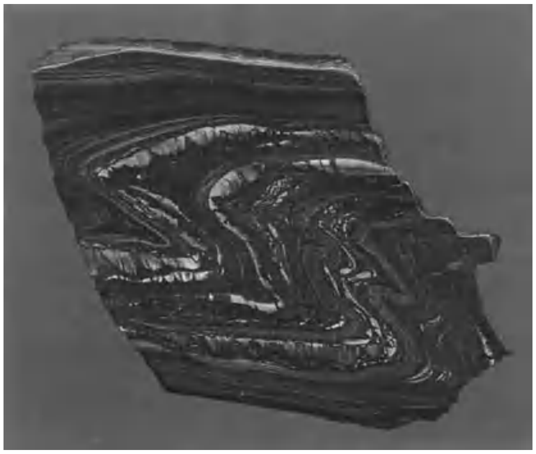


Fig. 1. Folded banded iron formation from the Ord Range, Western Australia. The distance between top and bottom of the polished face of the sample is about 15 cm (6 in.). Chert jasper bands alternate with dark magnetite-rich bands. The thin pale layers of irregular thickness are bands of asbestiform amphibole, now replaced by silica, to give the semiprecious material "tiger-eye." (*Photo courtesy of John Blockley*)

Global distribution and ages. Banded iron formation occurs in the Precambrian of all continents. The larger individual occurrences lie mainly in the southern continents and include the Carajas and Minas Gerais banded iron formations of Brazil, the Transvaal Supergroup banded iron formations of South Africa, and the Hamersley Group banded iron formation of northwestern Australia. All of these are individually larger than such Northern Hemisphere banded iron formations as those of North America (Labrador and the Lake Superior ranges), Ukraine (Krivoi Rog), west Africa (Mauritania and Liberia), India (Bababudan), and China (An Shan). While banded iron formations are generally restricted to the Precambrian, more specifically they range in age from about 3800 million years ago (Ma; Isua, Greenland) to less than 1000 Ma (Rapitan, western Canada; Damara Belt, Namibia; Urucum, Brazil). An exceptional banded iron formation also occurs in rocks of Devonian age (350 Ma) in the Altai area of Kazakhstan. Although banded iron formation deposition took place intermittently throughout this vast time range, banded iron formation was particularly abundant, and formed individually larger deposits, between about 2500 Ma and 2000 Ma.

Associated rocks, deposit dimensions, and band continuity. Banded iron formation may be stratigraphically associated in the same depositional basin with a variety of other sedimentary rock types, such as shale, sandstone, and carbonate (dolomite), as well as with volcanic rocks. However, in any one basin, banded iron formation tends to be concentrated in one or two stratigraphic levels, rather than being deposited repetitively throughout the life of the basin. Individual banded iron formations may be as thin as a few meters or over a kilometer thick, and may be traceable along their bedding from distances of one hundred to several hundred kilometers. The minimum depositional area of the Hamersley Group banded iron formations of northwestern Australia has been shown to be 150,000 km^2 (58,000 mi^2), and the banded iron formations have a total thickness of 700 m (2300 ft). Some thin (about 0.2–2.0 mm) repeated laminae, or microbands, of these banded iron formations have been shown to extend continuously over the whole area of the basin. It has been proposed that each microband, composed of an iron-rich and an iron-poor layer, represents a single year of deposition, and this hypothesis has now been shown to be consistent with the rate of banded iron formation deposition in the basin measured by isotope geochronology.

Origin. The mechanism of formation of most common sedimentary rocks, such as sandstone or limestone, is evident from observation of identical materials now being deposited. However, banded iron formation is not now occurring. A model for its origin must therefore be built up from the application of basic physical and chemical principles to detailed observation of its structure and composition. Early hypotheses mostly appealed to such concentrated sources of iron as local volcanic activity, or iron-rich river discharges resulting from an unusual Precambrian weathering regime. But recent studies suggest that the deposition of even very large banded iron formations could have been precipitated from ocean water containing as little as 10 parts per million iron, with about 30% of this precipitated each year from the water of local offshore basins. In the early Precambrian, when the oxygen content of the atmosphere and oceans was very low, upwelling could have provided an abundant supply of dissolved ferrous iron, leached from ocean-floor volcanic rocks. Seasonal oxidation by photosynthesizing algae in the near-surface water provides a credible precipitation mechanism, and the resultant insoluble ferric precipitate would have accumulated on the basin floor in periods of tectonic stability. This process may have provided a major sink for the oxygen produced by the earliest photosynthesizers, which would have lacked biochemical mechanisms to survive in an oxidizing environment, and whose biological productivity would have been restricted by the need to dispose of the oxygen they produced. The abundance of banded iron formation in the 2500–2000-Ma period may be related to the development of such mechanisms, while the later decrease in abundance probably reflects the fall of ocean water iron concentration to its present low level.

Economic importance. The world's iron and steel industry is based almost exclusively on iron ores associated with banded iron formation. Banded iron formation itself may be the primary ore, from which hematite or magnetite is concentrated after crushing. But the main ore now mined globally is high-grade (greater than 60% iron) material that formed within banded iron formation by natural leaching of its silica content.

Alec Francis Trendall

Bibliography. P. W. U. Appel and G. L. La Berge (eds.), *Precambrian Iron-Formations*, Theophrastus Publications, Athens, 1987; H. D. Holland and A. F. Trendall (eds.), *Patterns of Change in Earth Evolution*, Springer-Verlag, Berlin, 1984; A. F. Trendall and J. G. Blockley, Iron formations of the Precambrian Hamersley Group of Western Australia, with special reference to crocidolite, *Western Australia Geol. Surv. Bull.*, no. 119, 1970; A. F. Trendall and R. C. Morris (eds.), *Iron-Formation: Facts and Problems*, Developments in Precambrian Geology 6, Elsevier, Amsterdam, 1983.

Bandwidth requirements (communications)

The channel bandwidths needed to transmit various types of signals, using various processing schemes.

Propagation conditions for communication systems that transmit signals through the atmosphere or through free space are strongly dependent on the transmission frequency. Government regulations specify the modulation type, the bandwidth, and the type of information that can be transmitted over designated frequency bands. In the United States, the Federal Communications Commission (FCC) regulates radio, telephone, television, and satellite transmission systems. In North America, standards are developed by the Institute of Electrical and Electronics Engineers (IEEE), by Committee TI–Telecommunications (accredited by the American National Standards Institute, ANSI), and by the Telecommunications Industry Association (TIA). Internationally, frequency assignment and technical standards are developed by the International Telecommunications Union–Telecommunications Standardization Sector (ITU-T) and the International Telecommunications Union–Radiocommunication Sector (ITU-R), functioning under the auspices of the United Nations International Telecommunications Union (ITU). *See* RADIO SPECTRUM ALLOCATION.

Definition of bandwidth. Every signal observed in practice can be expressed as a sum (discrete or over a frequency continuum) of sinusoidal components of various frequencies. The plot of the amplitude versus frequency constitutes one feature of the frequency spectrum (the other being the phase versus frequency). The difference between the highest and the lowest frequencies of the frequency components of significant amplitudes in the spectrum is called the bandwidth of the signal expressed in the unit of frequency, hertz. Every communication medium (also called channel) is capable of transmitting a frequency band (spectrum of frequencies) with reasonable fidelity. Qualitatively speaking, the difference between the highest and the lowest frequencies of components in the band over which the channel gain remains reasonably constant (or within a specified variation) is called the channel bandwidth. *See* FOURIER SERIES AND TRANSFORMS; WAVEFORM.

The notions of significant amplitudes, reasonable fidelity, reasonably constant gain, and so forth, however, differ with different applications. Consequently, there are several quantitative definitions of bandwidths. The following bandwidth definitions are commonly observed in practice:

1. *Half-power bandwidth*. This is the difference between the two frequencies at which the amplitude spectrum is 3 dB below the peak value (which is the same as saying that the power spectrum is half of the peak value).

2. *Specified power containment bandwidth*. This is the band of frequencies that contains 100 $(1-\epsilon)\%$ of the total signal power. The value of ϵ is usually of the order of 0.01 to 0.1 depending on the application.

3. *Bounded-power spectral density bandwidth*. This is the width of the frequency band outside which the power spectral density must fall at least a given level (such as 40 or 60 dB) below its maximum value.

4. *Null-to-null bandwidth*. This is equal to the width of the main lobe of the power spectral density.

5. *Equivalent noise bandwidth*. This is the width of a rectangle whose height is equal to the maximum value of the power spectral density of the signal and whose area is equal to the area under the power spectral density of the signal. For modulated signals, this figure is multiplied by 0.5.

In analog signals and systems, half-power bandwidth is commonly specified, whereas in digital communication the specification of the null-to-null bandwidth is the usual practice.

Signal conditioning. Clearly, to transmit a signal with reasonable fidelity over a communication channel, the channel bandwidth must match and be at least equal to the signal bandwidth. Proper conditioning of a signal, such as modulation or coding, however, can increase or decrease the bandwidth of the processed signal. Thus, it is possible to transmit the information of a signal over a channel of bandwidth larger or smaller than that of the original signal.

Amplitude modulation (AM) with double sidebands (DSB), for example, doubles the signal bandwidth. If the audio signal to be transmitted has a bandwidth of 5 kHz, the resulting AM signal bandwidth using DSB is 10 kHz. Amplitude modulation with a single sideband (SSB), on the other hand, requires exactly the same bandwidth as that of the original signal. Amateur radio and commercial and military services are examples of this scheme. Vestigial sideband (VSB) modulation is a compromise between DSB and SSB. It requires a bandwidth slightly larger (typically 20–30%) than that of the message signal. Broadcast television is an example of this scheme. The television signal (video and audio) bandwidth is 4.5 MHz. It is transmitted by VSB modulation that requires a bandwidth of 6 MHz. *See* AMPLITUDE MODULATION; AMPLITUDE-MODULATION RADIO; SINGLE SIDEBAND; TELEVISION.

DSB, SSB, and VSB modulation are examples of amplitude-modulation schemes. In broadcast frequency modulation (FM), on the other hand, audio signal bandwidth is 15 kHz (for high fidelity), but the corresponding frequency-modulated signal bandwidth is 200 kHz. In telephony, the audio signal bandwidth is approximately 3 kHz and is transmitted over local lines of 3 kHz bandwidth, or over long distances (using SSB) requiring a transmission bandwidth of 4 kHz per telephone channel (3 kHz for the signal and a guard band of 1 kHz). On the other hand, a digitized telephone signal (processed by analog-to-digital conversion) requires a bandwidth of 32–64 kHz. Facsimile typically requires a bandwidth of 5 kHz for AM and 25 kHz for FM. In telegraphy, the bandwidth for transmission of 100 words per minute is 170–400 Hz. *See* ANALOG-TO-DIGITAL CONVERTER; FACSIMILE; FREQUENCY MODULATION; FREQUENCY-MODULATION RADIO; TELEGRAPHY.

In pulse communication systems, the rate at which pulses can be transmitted over a channel is directly proportional to the channel bandwidth. The sampling theorem leads to the conclusion that in the absence of noise a channel can transmit correctly at most two independent pieces of information per second per unit bandwidth (in hertz). In most pulse communication systems, for example, pulse amplitude modulation (PAM) or pulse code modulation (PCM), the information carried by a pulse resides in the pulse amplitude. Consequently, the information from one pulse is equivalent to one piece of information. Hence for PAM and PCM signals, up to $2B$ pulses per second can be transmitted over a bandwidth of B Hz. This, however, is an upper (theoretical) limit, which can be attained only by using ideal filters (or noncausal pulses of infinite duration); this limiting case is, of course, unrealizable. In practice, the transmission rate remains below $2B$ pulses per second.

The rates discussed above result from the inherent limitations imposed by the channel bandwidth. In practice, a specific coding scheme imposes further restrictions on the rate. For example, the so-called bipolar (or pseudo-ternary) scheme of binary signaling requires a transmission bandwidth of B Hz to transmit B binary pulses per second. This is only half the rate that would be allowed by the channel bandwidth restriction alone. On the other hand, a duobinary (or partial response coding) scheme, represented in digital telephony by the alternate mark inverted (AMI) line code, can transmit $2B$ pulses per second over a channel of bandwidth B.

Advances in coding and modulation theory, as well as in the design of electronic circuitry, have made possible the implementation of modulation schemes more complex than those described above. The more complex schemes sometimes carry many bits of information per pulse and can adapt themselves automatically to best utilize the capacities of channels with impaired frequency responses.

Such modulation schemes lie at the heart of the growing family of digital subscriber systems (xDSL, where x is a letter representing the system type). Such systems allow the transmission of data along copper cable pairs of the telephone network's subscriber loop plant at relatively high rates, extending the useful life of the loop plant. *See* MODULATION; PULSE MODULATION; RADIO.

Signal bandwidth restriction. It can be shown that a signal cannot be simultaneously time-limited and band-limited. Thus the spectrum of a finite-duration

-S

signal extends to infinity; consequently its bandwidth is infinite. Since all practical signals are of finite duration (time-limited), the bandwidth of any practical signal is infinite. The implication is that on any given channel, only one signal can be transmitted at any time; for example, on air waves, only one station can be broadcast at a time. Because of tremendous demand for communication in modern times, intolerable usage congestion would result. Fortunately, all practical signal spectra decay asymptotically with frequency and approach zero at infinite frequency. High-frequency components can be cut off, causing acceptable distortion in the signal waveform. Thus, every practical signal can be effectively band-limited to a certain essential bandwidth.

Effects of bandwidth restriction. Band-limiting a signal (cutting off high-frequency components) results in distortion of the signal. The effect of the distortion is often more serious for digital than analog signals. In analog systems, the distortion causes the signal quality to deteriorate somewhat and is not particularly significant. In digital communication systems, the signal is in the form of a sequence of pulses. It can be shown that the pulses are dispersed (spread out) when the channel bandwidth is restricted. Such a dispersion causes pulses to interfere with neighboring pulses. This interference (intersymbol interference) can be additive and leads to a higher probability of errors in pulse detection. In practice, equalizers (in the form of transversal filters) are used at the receiver to correct partly the damage to the signal resulting from channel bandwidth restrictions. *See* DISTORTION (ELECTRONIC CIRCUITS); ELECTRIC FILTER; EQUALIZER.

Shannon's equation. C. E. Shannon proved that over a channel of bandwith B the rate of information transmission, C, in bits/s (binary digits per second) is given by the equation below, where SNR is the

$$C = B \log_2(1 + \text{SNR}) \quad \text{bits/s}$$

signal-to-noise power ratio. This result assumes a white gaussian noise, which is the worst kind of noise from the point of view of interference. For any other noise the rate of information transmission will be at least as large as that indicated by Shannon's equation. *See* INFORMATION THEORY; SIGNAL-TO-NOISE RATIO.

It follows from Shannon's equation that a given information transmission rate C can be achieved by various combinations of B and SNR. It is thus possible to trade B for SNR, and vice versa. In a communication system, one of the two resources (B or SNR) may be more precious than the other, and the communication scheme would be optimized accordingly. A typical voice-grade telephone circuit has a limited bandwidth (3 kHz) but a lot of power is available, whereas in communications satellites large bandwidths are available but power is scarce. Hence the communication schemes required in the two cases are usually different. *See* SPACE COMMUNICATIONS.

Broadbanding. A corollary of Shannon's equation is that, if a signal is properly processed to increase its bandwidth, the processed signal becomes more immune to interference or noise over the channel. This means that an increase in transmission bandwidth (broadbanding) can suppress the noise in the received signal, resulting in a better-quality signal (increased SNR) at the receiver. Frequency modulation and pulse-code modulation are two examples of broadband schemes where the transmission bandwidth can be increased as desired to suppress noise. The former transmits information in analog form (an analog scheme), while the latter digitizes the analog information and transmits the resulting digital data (a digital scheme). In general, digital schemes use increased bandwidth more efficiently than do analog schemes in suppressing noise. Amplitude modulation, on the other hand, is not a broadband scheme because its transmission bandwidth is fixed and cannot be increased.

Broadbanding is also used to make communication less vulnerable to jamming and illicit reception by using the so-called spread spectrum signal. The jamming signal can be considered as noise, and by using the proper technique of broadbanding (spread spectrum) the jamming signal can be rendered less effective (suppressed). Spread spectrum may use direct-sequence, frequency, time-hopping, and chirp systems. In spread spectrum systems, some signal function other than the information being sent is employed to determine the transmitted signal bandwidth. For this reason, conventional FM, where the signal bandwidth is determined only by the message signal, cannot be considered a spread spectrum system. *See* ELECTRICAL COMMUNICATIONS; ELECTRONIC WARFARE; SPREAD SPECTRUM COMMUNICATION.

B. P. Lathi; Maynard Wright

Bibliography. R. L. Freeman, *Reference Manual for Telecommunication Engineering*, 2d ed., Wiley, 1994; B. P. Lathi, *Modern Digital and Analog Communication Systems*, 3d ed., Oxford, 1998; T. Starr, J. M. Cioffi, and P. Silverman, *Understanding Digital Subscriber Line Technology*, Prentice Hall, 1999; M. E. Van Valkenburg, *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, 9th ed., 2001.

Barbiturates

A group of drugs widely used for the suppression of anxiety, the induction of sleep, and the control of seizures. Some of them, when injected intravenously, produce a general anesthesia.

Barbituric acid, from which the various barbiturate congeners come, was synthesized in 1864 by Adolf von Baeyer and named after St. Barbara. It is a malonyl urea. In the years that followed, a dozen or more closely related compounds obtained by adding or substituting various radicals to the general formula have been synthesized. The names of many of them have become familiar; examples are Phenobarbital, Meberal, Seconal, Nembutal, Amytal, and Pentothal. They vary in speed and duration of action and in potency.

These drugs have been shown to act on the reticular formation of the upper brainstem and thalamus. Actually they suppress the excitability of all tissues; but all tissues are not equally sensitive. Low dosages induce drowsiness, and high dosages coma and death. The spread between the therapeutic and fatal doses varies with the different barbiturates.

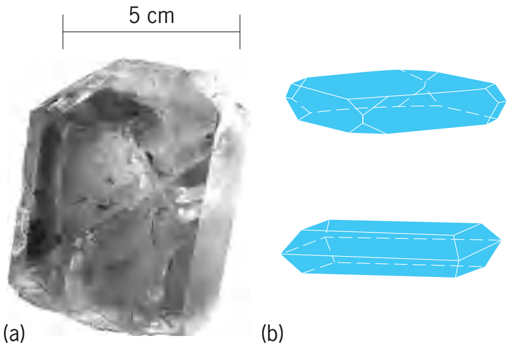
The prolonged use of these drugs results in habituation; and insomnia, agitation, confusional psychosis, and seizures may occur within 24 to 36 h of withdrawal. Overdose is one of the commonest means of suicide in Western countries, and life can be saved only by admission to a hospital where respiration can be maintained and cerebral anoxia prevented. *See* ADDICTIVE DISORDERS; CENTRAL NERVOUS SYSTEM.

Raymond D. Adams

Bibliography. R. Rosen (ed.), *Barbiturates and Other Depressants*, 1993.

Barite

A mineral with chemical composition $BaSO_4$. It is isostructural with celestite, $SrSO_4$, and anglesite, $PbSO_4$. Barite is orthorhombic, space group Pnma, $a = 0.887$ nanometer, $b = 0.545$ nm, $c = 0.714$ nm. The **illustration** shows typical crystals. Barite possesses one perfect cleavage, and two good cleavages as do the isostructural minerals. Anhydrite, $CaSO_4$, although also orthorhombic, possesses a different space group and a different structure. Solid solution in the series $BaSO_4$–$SrSO_4$ occasionally occurs in the natural environments. Solid solution in $BaSO_4$–$PbSO_4$ is uncommon, although up to 30 mole % $PbSO_4$ in barite has been reported. Infrequency of natural solid solution between these two end members may stem from the different origins of the two minerals, anglesite being an oxidation and weathering product of galena ores. Since Ba^{2+} has the largest ionic radius of the isostructural series with the ionic radius of $Ba^{2+} \approx Pb^{2+} > Sr^{2+} \gg Ca^{2+}$, separation of these species may be a consequence of this geometrical distinction. Barite can be distinguished from celestite and anhydrite by its superior specific gravity of approximately 4.5. The mineral is relatively soft, approximately 3 on Mohs scale. The color ranges through white to yellowish, gray, pale blue, or brown, and a thin section is colorless. Barite is often an accessory mineral in hydrothermal vein systems, but frequently occurs as concretions or cavity fillings in sedimentary limestones. It also occurs as a residual product of limestone weathering and in hot spring deposits. It occasionally occurs as extensive beds in evaporite deposits.



Barite. (*a*) Specimen from Dufton, Westmoreland, England (*American Museum of Natural History*). (*b*) Typical crystals; crystals may be very complex (*after C. S. Hurlbut, Jr., Dana's Manual of Mineralogy, 17th ed., John Wiley and Sons, Inc., 1959*).

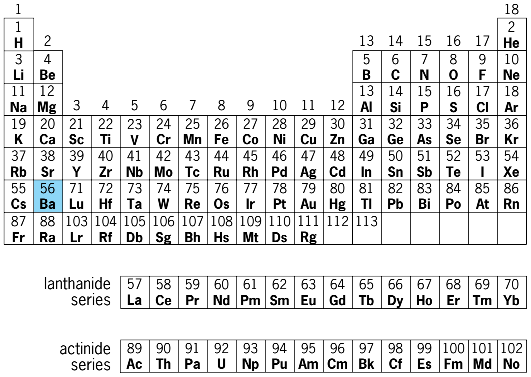
Since barite is dense and relatively soft, its principal use is as a weighting agent in rotary well-drilling fluids. It is the major ore of barium salts, used in glass manufacture, as a filler in paint, and, owing to the presence of a heavy metal and inertness, as an absorber of radiation in x-ray examination of the gastrointestinal tract.

Occurrences of barite are extensive. It is found as a vein mineral associated with zinc and lead ores in Derbyshire, England. Typical associated minerals are calcite, dolomite, pyrite, quartz, and fluorite. Large deposits occur at Andalusia, Spain. Commercial residual deposits occur in limestones throughout the Appalachian states such as Georgia, Tennessee, and Kentucky. It also occurred in substantial amounts in galena ore in Wisconsin and Missouri. *See* BARIUM.

Paul B. Moore

Barium

A chemical element, Ba, with atomic number 56 and atomic weight of 137.34. Barium is eighteenth in abundance in the Earth's crust, where it is found to the extent of 0.04%, making it intermediate in amount between calcium and strontium, the other alkaline-earth metals. Barium compounds are obtained from the mining and conversion of two barium minerals. Barite, barium sulfate, is the principal ore and contains 65.79% barium oxide. Witherite, sometimes called heavy spar, is barium carbonate and is 72% barium oxide. *See* BARITE; PERIODIC TABLE; WITHERITE.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

The metal was first isolated by Sir Humphry Davy in 1808 by electrolysis. Industrially, only small amounts are prepared by aluminum reduction of barium oxide in large retorts. These are used in barium-nickel alloys for spark-plug wire (the barium increases the emissivity of the alloy) and in frary metal, which is an alloy of lead, barium, and calcium used in place of babbitt metal because it can be cast.

Properties of barium

Property	Value
Atomic number	56
Atomic weight	137.34
Isotopes (stable)	130, 132, 134, 135, 136, 137, 138
Atomic volume	36.2 cm^3/g-atom
Crystal structure	Face-centered cubic
Electron configuration	2 8 18 18 8 2
Valence	2+
Ionic radius (A)	1.35
Boiling point, °C	1140(?)
Melting point, °C	850(?)
Density	3.75 g/cm^3 at 20°C
Latent heat of vaporization at boiling point, kj/g-atom	374

The metal reacts with water more readily than do strontium and calcium, but less readily than sodium; it oxidizes quickly in air to form a surface film that inhibits further reaction, but in moist air it may inflame. The metal is sufficiently active chemically to react with most nonmetals. Freshly cut pieces have a lustrous gray-white appearance, and the metal is both ductile and malleable. The physical properties of the elementary form are given in the **table**.

For the manufacture of barium compounds, soft (easily crushable) barite is preferred, but crystalline varieties may be used. Crude barite is crushed and then mixed with pulverized coal. The mixture is roasted in a rotary reduction furnace, and the barium sulfate is thus reduced to barium sulfide or black ash. Black ash is roughly 70% barium sulfide and is treated with hot water to make a solution used as the starting material for the manufacture of many compounds.

Lithopone, a white powder consisting of 20% barium sulfate, 30% zinc sulfide, and less than 3% zinc oxide, is widely used as a pigment in white paints. Blanc fixe is used in the manufacture of brilliant coloring compounds. It is the best grade of barium sulfate for paint pigments. Because of the large absorption of x-rays by barium, the sulfate is used to coat the alimentary tract for x-ray photographs in order to increase the contrast. Barium carbonate is useful in the ceramic industry to prevent efflorescence on claywares. It is used also as a pottery glaze, in optical glass, and in rat poisons. Barium chloride is used in purifying salt brines, in chlorine and sodium hydroxide manufacture, as a flux for magnesium alloys, as a water softener in boiler compounds, and in medicinal preparations. Barium nitrate, or the so-called baryta saltpeter, finds use in pyrotechnics and signal flares (to produce a green color), and to a small extent in medicinal preparations. Barium oxide, known as baryta or calcined baryta, finds use both as an industrial drying agent and in the case-hardening of steels. Barium peroxide is sometimes used as a bleaching agent. Barium chromate, lemon chrome or chrome yellow, is used in yellow pigments and safety matches. Barium chlorate finds use in the manufacture of pyrotechnics. Barium acetate and cyanide are used industrially as a chemical reagent and in metallurgy, respectively. Reel F. Riley

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; D. F. Shriver and P. W. Atkins, *Inorganic Chemistry*, 3d ed., 1999.

Bark

A word generally referring to the surface region of a stem or a root. The term has a variable meaning, and sometimes part or all of the bark is called rind. Occasionally the word is used as a substitute for periderm or for cork only. Most commonly, however, bark refers to all tissues external to the cambium (**Fig. 1**). If this definition is applied to stems having only primary tissues, it includes phloem, cortex, and epidermis; the bark of roots of corresponding age would contain cortex and epidermis. More generally, the term bark is restricted to woody plants with secondary growth.

In most roots with secondary growth, the bark consists of phloem and periderm since the cortex and epidermis are sloughed off with the first cork formation in the pericycle, that is, beneath the cortex. In stems, the first cork cambium may be formed in any living tissue outside the vascular cambium, and the young bark may include any or all of the cortex in addition to the phloem and periderm. The first cork cambium may persist throughout the life of the plant. The resulting periderm has a relatively smooth surface. In most plants, however, new periderms are formed in successively deeper tissues. The new layers of cork are often shell-like in form and enclose islands of dead phloem (**Fig. 2**). In some plants, the successive cork layers occur in wide bands or even in whole cylinders. The region composed of the successive layers of periderm and the enclosed dead tissues is called outer bark. The outer bark gradually splits, wears, and weathers and acquires a characteristic surface pattern in relation to the form of origin of the cork cambium and the composition of the dead

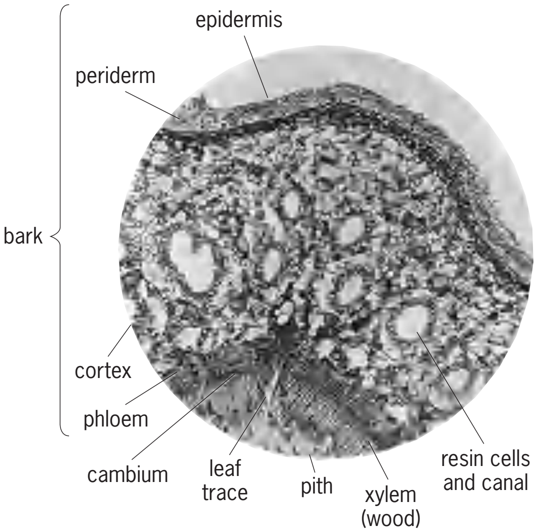


Fig. 1. Transverse section of young twig of balsam fir (*Abies balsamea*) showing tissues often included in the term bark. (*Forest Products Laboratory, USDA*)

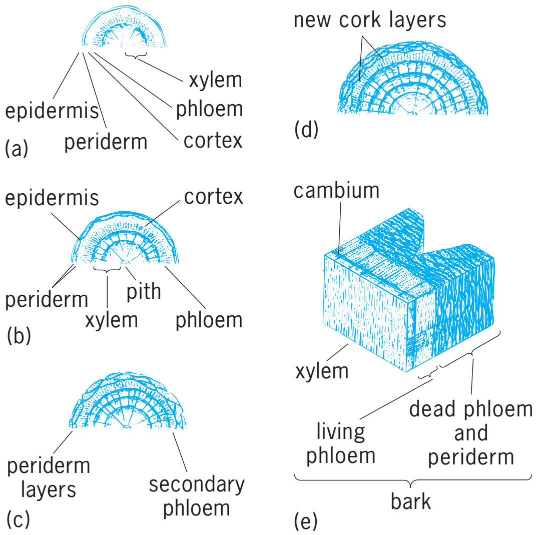


Fig. 2. Successive periderm layers in typical woody stem at (*a*) 1, (*b*) 2, (*c*) 3, (*d*) 4 years. (*e*) Outer tissues of old trunk. (*After A. J. Eames and L. H. MacDaniels, An Introduction to Plant Anatomy, 2d ed., McGraw-Hill, 1947*)

tissues. The surface may be fairly smooth or broken into strips conforming to the fibrous layers of the phloem, or into plates or strips conforming to the size and shape of the alternating zones of cork and phloem.

The outer bark composed of dead phloem alternating with bands of cork is called, technically, rhytidome. Both stems and roots may have rhytidome. The inner bark is living, and consists of phloem only. In this phloem only a narrow band next to the vascular cambium contains functioning sieve elements, but ray cells and parenchyma are alive throughout the inner bark. From these living cells, new cork cambia may develop. *See* CORTEX (PLANT); EPIDERMIS (PLANT); PARENCHYMA; PERICYCLE; PERIDERM; PHLOEM; ROOT (BOTANY); STEM.

H. Weston Blaser

Bibliography. J. D. Mauseth, *Plant Anatomy*, 1988.

Barkhausen effect

An effect, due to discontinuities in size or orientation of magnetic domains as a body of ferromagnetic material is magnetized, whereby the magnetization proceeds in a series of minute jumps. *See* FERROMAGNETISM; MAGNETIZATION.

Ferromagnetic materials are characterized by the presence of microscopic domains of some 10^{12} to 10^{15} atoms within which the magnetic moments of the spinning electrons are all parallel. In an unmagnetized specimen, there is random orientation of the various domains. When a magnetic field is applied to the specimen, the domains turn into an orientation parallel to the field, or if parallel to the field, the domains increase in size. During the steep part of the magnetization curve, whole domains suddenly change in size or orientation, giving a discontinuous increase in magnetization. If the specimen being magnetized is within a coil connected to an amplifier and loudspeaker, the sudden changes give rise to a series of clicks or, when there is a rapid change, a hissing sound. This is called the Barkhausen effect; it is an important piece of evidence in support of a domain theory of magnetism.

Kenneth V. Manning

Barley

A cereal grass plant whose seeds are useful to humans. It is grown in nearly all cultivated areas of the temperate parts of the world, and is an important crop in Europe, North and South America, North Africa, much of Asia, and Australia. Barley is the most dependable cereal crop where drought, summer frost, and alkali soils are encountered. In the United States, barley is grown to some extent in 49 states, with the most important production areas in North Dakota, Montana, and California. Principal uses of barley grain are as livestock feed, for the production of barley malt, and as human food. It is also used in some areas for hay and silage.

Origin and description. Barley is probably one of the most ancient cultivated cereal grains. Its origin is still debated. Archeological evidence in the form of heads, carbonized grains, and impressions of grains on pottery have been found in several areas of the world. Some in Egypt are dated over 5000 years ago. Barley was probably grown by the Swiss Lake Dwellers between 2000 and 3000 B.C., in China about 200 B.C., and in Greek and Roman areas about 300 B.C.

Taxonomically barley belongs to the family Gramineae, subfamily Festucoideae, tribe Hordeae, and genus *Hordeum*. Most of the modern cultivated barleys are *H. vulgare* (six-rowed) or *H. distichum* (two-rowed; **Fig. 1**). All cultivated barleys are annuals, are naturally self-pollinated, and have seven pairs of chromosomes. **Figure 2** shows a typical barley plant. Cultivated barleys can be readily crossed by emasculating before pollen is shed or through genetic male sterility. There are also two wild species having chromosomes homologous with the cultivated species which are crossable with each other and with the cultivated species. Both of these, *H. spontaneum* (two-rowed) and *H. agriocrithon* (six-rowed), have brittle rachis and readily reseed themselves in the wild state. There are also a number of wild grasslike species which belong to the genus *Hordeum*. The basic chromosome number of these species is 7, and $2n$ numbers of 14, 28, and 42 have been observed. These species do not cross readily with the cultivated species, although there has been marked advancement in artificial crosses with the use of colchicine techniques and in growing excised embryos on artificial media. Among this group of wild barleys are annuals including *H. murinum*, *H. gussoneanum*, *H. glaucum*, *H. leproinum*, *H. depressum*, and *H. geniculatum*. Perennial types are *H. jubatum*, *H. bulbosum*, and *H. montanese*. *See* CHROMOSOME; CYPERALES; GENETICS; POLLINATION; REPRODUCTION (PLANT).

Varieties. In the cultivated varieties, a wide diversity of morphological, physiological, and anatomical types are known. There are spring, facultative,

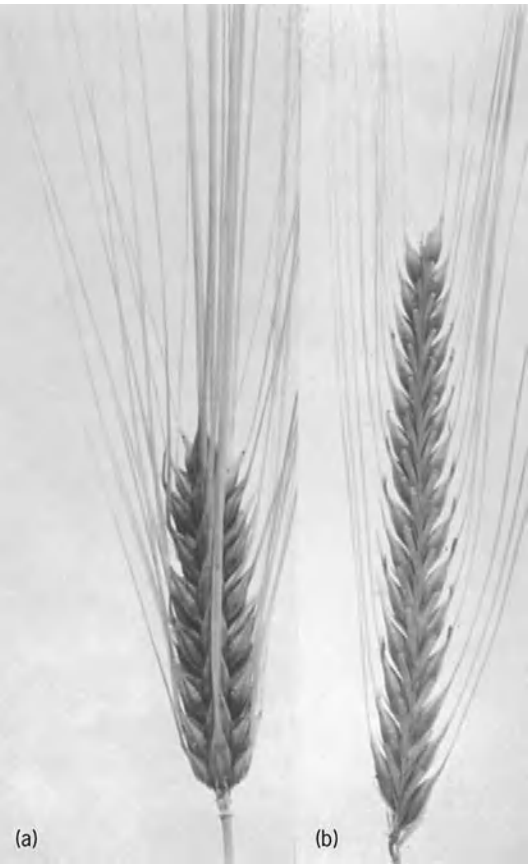


Fig. 1. Barley spikes: (*a*) six-rowed *Hordeurn vulgare*; (*b*) two-rowed *H. distichum* (*USDA*)

and winter growth habits; hulled and naked grain; awned, awnless, and hooded lemmas; black, purple, and white kernels; and also a wide range of plant heights, spike densities, and resistances to a wide range of diseases and insects. There are in excess of 150 cultivars presently commercially grown in the United States and Canada alone, and many additional cultivars are grown in other parts of the world. New and improved varieties produced by barley breeders are constantly replacing older varieties. Several barley collections are being maintained in different countries as germplasm sources for breeding and research. These include both collections made by direct exploration in many barley-growing areas of the world and lines from barley-breeding programs. Among the largest of these collections is one maintained by the U.S. Department of Agriculture, which includes more than 17,000 individual strains.

Cultural practices. The best soils for growing barley are well-drained fertile loams and clay loams. This grain is the most tolerant of the cereal grains to alkaline soils (pH 6.0 to 8.0), and the least tolerant to acid soils (below pH 5.0), especially if there is soluble aluminum in the soil solution. Liming may be necessary to correct acidity and to precipitate aluminum, and lime should be applied at the time of seedbed preparation. Soil tests will determine if phosphorus or potash is needed. Nitrogen will generally be needed for good production, depending on the previous crop and climatic conditions. Time and rate of seeding, seedbed preparation, and rotations vary widely with geographical area. In the United States, spring varieties are grown in the Midwest, northern Great Plains, and Pacific Northwest. Seeding should be done as early as the seedbed can be prepared. This will vary from about March 1 to 15 in the Kansas-Nebraska area, the Great Basin, and the Pacific Northwest to late April or early May in the northern Great Plains. Spring varieties are commonly sown in the late fall or winter in the southwestern states, the Atlantic seaboard, the southern Great Plains, and in the milder areas of the Pacific Northwest. Generally the planting date will be 1 to 2 weeks ahead of the earliest fall frost-free date. Common seeding rates are 4–8 pecks/acre (0.087–0.174 m^3/ha), depending on expected moisture conditions. Barley can be grown in a variety of rotations with other crops common to a particular region. In areas with long growing seasons, it is possible to double-crop (growing two crops in 12 consecutive months on the same land) with summer crops such as soybeans or cotton. Barley is ideally suited to this purpose because it is usually the earliest maturing of the small grains. The variety of the alternate crop should also be early-maturing. *See* GRAIN CROPS. David A. Reid

Diseases. Barley is one of the cereals most subject to diseases, especially in humid areas. Consequently, grain yields of varieties grown in humid areas have not been consistently high. Because of diseases and the grain quality needed for malting purposes, barley production has shifted westward from the Corn Belt to northwestern Minnesota, through North Dakota, Montana, northern Idaho, and California, where the relative humidity is lower.

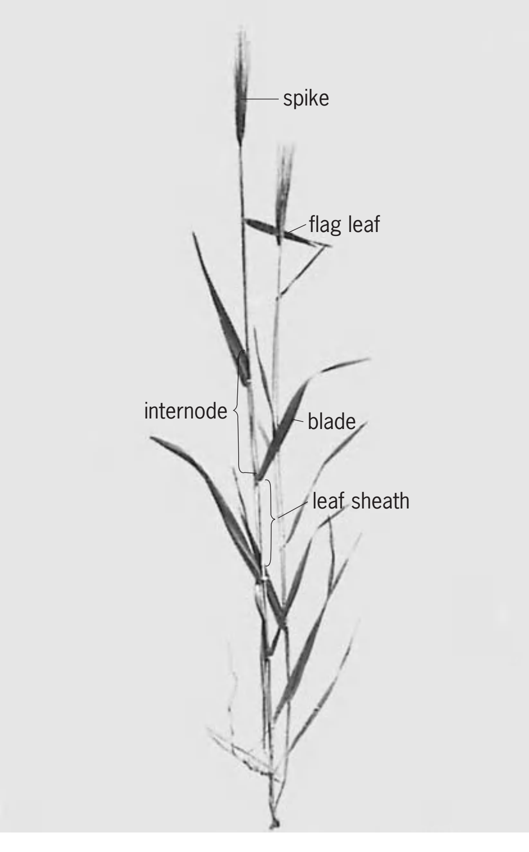


Fig. 2. Typical barley plant. (*USDA*)

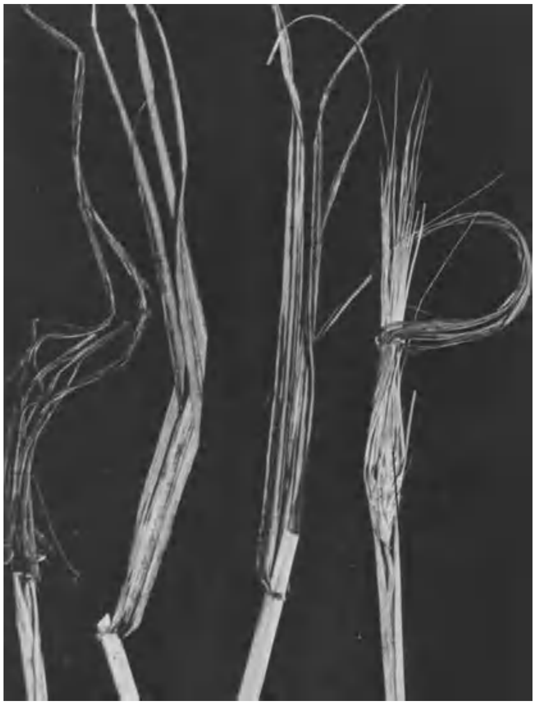


Fig. 3. Barley stripe, caused by *Helminthosporium gramineum*. (*Minnesota Agricultural Experiment Station*)

Yield losses from some diseases result from reductions in stand and in plant vigor. Other diseases cause the quality of the grain to be reduced through shriveling, discoloration, and toxin production in the grain itself.

Seed-borne pathogens. Fungi carried on or in the seed [*Ustilago hordei*, which causes covered smut; *U. nuda*, which causes loose smut; *Helminthosporium gramineum*, the cause of stripe (**Fig. 3**); and *H. sorokinianum* and *H. sativum*, which cause seedling blight and spot blotch (**Fig. 4**)] are controlled to a large extent by growing varieties with some degree of resistance and by fungus-free seed. Where clean seed is not available, treatment of the seed with a mixture of Carboxin and Thiram will give virtually 100% control. *See* FUNGISTAT AND FUNGICIDE.

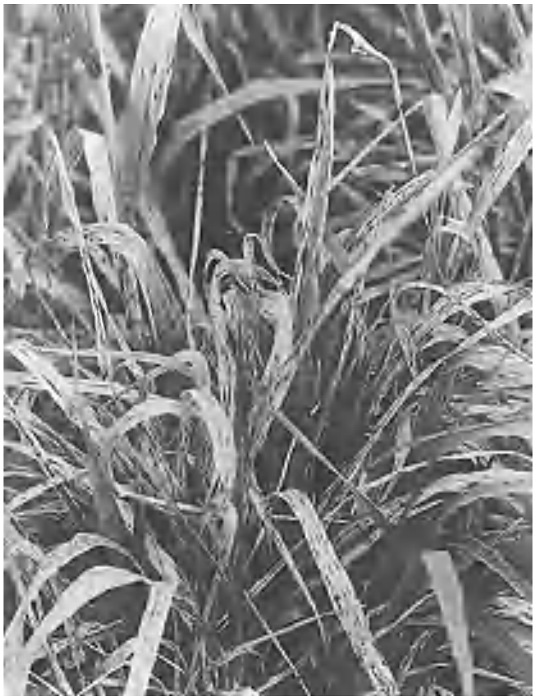


Fig. 4. Barley spot blotch, caused by *Helminthosporium sativum*. (*Minnesota Agricultural Experiment Station*)

Foliage diseases. Fungi that cause these diseases may be seed-borne, carried over on plant debris from the previous season, or blown in from distant sources. Varieties resistant to several of them have been developed. Seed treatment and crop rotation or disposal of debris (plowing-under) will also aid in their control. Examples of leaf diseases and their causes are spot blotch (*H. sorokinianum*), net blotch (*H. teres*), speckled leaf blotch (*Septoria passerinii*), and scald (*Rhynchosporium secalis*). Leaf rust (caused by *Puccinia hordei*, stem rust (caused by *P. graminis*), and powdery mildew (caused by *Erysiphe graminis*) are usually windborne from distant sources and are mainly controlled by the growing of resistant varieties. However, the occurrence of new pathogenic strains requires the development of new resistant varieties with different resistance genes.

Grain diseases. Kernel blights, caused by *H. sorokinianum*, *Gibberella zeae*, and *Alternaria* species (**Fig. 5**), shrivel and discolor the grain, thus reducing the marketing and particularly the malting quality. The scab disease, caused by *G. zeae*, also reduces the feeding value of barley, particularly for hogs, because of the toxin or toxins produced by infection. Because this fungus also infects corn as a stalk or ear rot, barley should not follow corn in a rotation. Saprophytic organisms on the kernels may cause deterioration of the grain in storage.

Seedling, root, and crown diseases. Several of the above-mentioned fungi cause rotting of various plant parts. These organisms may be seed- or debris-borne. Others, such as *Pythium* species, may be soil-borne. Pathogen-free seed, chemical seed treatment, sanitation, crop rotation, and cultural practices may be of value in their control.

Other diseases. Barley stripe mosaic and barley yellow dwarf diseases are caused by viruses. The first is seed-borne and is controlled by the use of virus-free seed and resistant varieties. The second is aphid-transmitted from perennial grasses, and is controlled by the use of resistant varieties. The oat cyst nematode (cereal root eelworm) *Heterodera avenae* has not been important in the United States. In Europe, it is a serious problem in barley as well as in oats. *See* PLANT PATHOLOGY; PLANT VIRUSES AND VIROIDS.

Deane C. Arny

Pot and pearled barley. The firmly adhering husk of the barley grain must be removed to facilitate use of barley in cooking. Pot and pearled barley are particulate products produced by abrasive removal of the hull or outer portion of the kernel. Thoroughly cleaned barley grain in sound condition is essential to production of good-quality products. The grain is exposed to the abrasive action of a series of rotating carborundum-coated disks to grind the hull from the kernel. Air aspiration is used to facilitate removal of fine particles and hull fragments. To produce pot

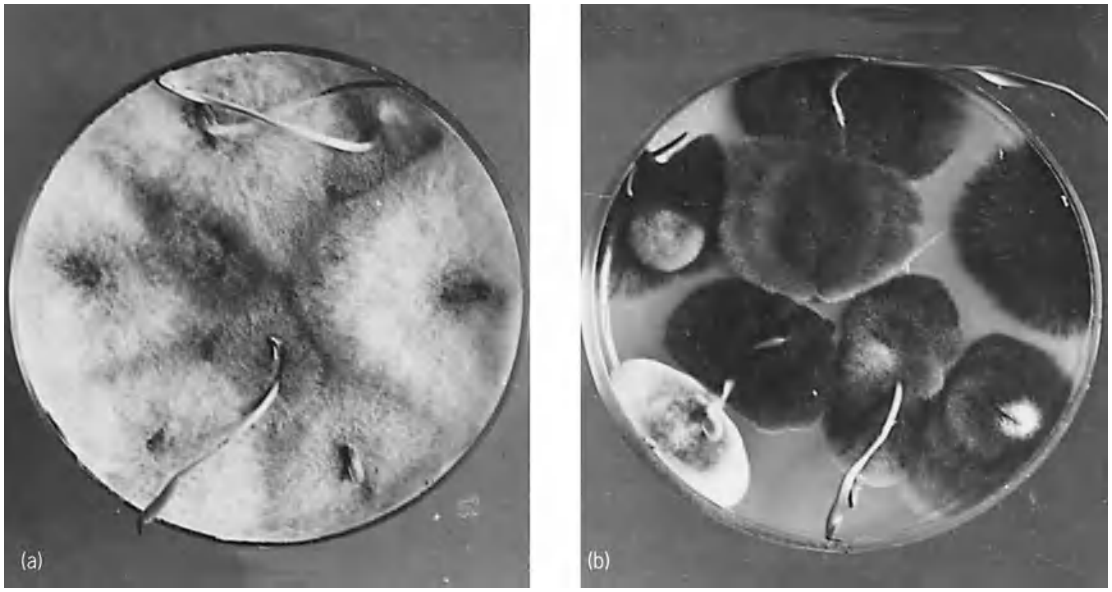


Fig. 5. Infected barley seed planted out on nutrient agar. (*a*) *Gibberella zeae*. (*b*) *Helminthosporium* and *Alternaria* sp. These kernel blights occur in more humid regions. (*Minnesota Agricultural Experiment Station*)

barley, the partially dehulled grain passes through a series of grinders until the kernel is mulled through the bran and aleurone layer. Removal of this outer portion results in about a 65% yield of clean pot barley suitable for cooking. Increasing the succession of passes through abrasive disks results in removal of greater portions of the kernel. White-pearled barley is obtained as a smooth and round kernel with a yield approximating 35%. The heating of grain during this attrition process must be controlled to avoid starch damage. Pot and pearled barley are cleaned, sized, and packaged for use in soups or porridges. Pearled barley may be further ground to produce high-quality grits or flour.

Malting. Malt is obtained from the controlled germination of barley grain. While malting procedures are an extremely varied art, precise scientific monitoring which utilizes numerous techniques or points of control is required. The characteristics of the resultant malt are influenced by the initial grain and the stages and conditions of the malting process. During the sprouting of cereal grains, numerous physiochemical changes occur, including increases of respiration (enzyme mobilization) and modifications of stored energy reserves (starch-sugars). The basic process for malting includes three stages: steeping, germination, and drying.

Steeping. Cleaned and sized barley is placed in large cylindrical steep tanks, covered with 50–60°F (10–16°C) water, aerated, and held there until the grain moisture content reaches 43–45%. The steeping time may vary from 30 to 70 h, during which time control of temperature, aeration rate, and steep-water quality are essential for maximizing germination potential.

Germination. Technically, germination is initiated during the steeping process; however, soaked grain is transferred to specialized germinating rooms or floors where the steeped grain is piled and turned by hand daily. Modern pneumatic malting equipment enables controlled mixing within drums or chambers. Grain possessing optimum moisture is maintained at 60–70°F (16–21°C) with adequate levels of aeration. Control of germinating temperature reduces losses to respiration and sprout growth and thus increases dry-matter yield. Germination time varies between 4 and 7 days depending on the type of barley and its intended use.

Drying or kilning. The germinated barley grain, known as green malt, is heat-inactivated at the appropriate stage of development. The green malt is heated to 110–130°F (43–54°C) and dried to about 6–12% moisture. This stabilized product will retain high levels of enzymatic activity and provide characteristic color and flavor. Kiln temperatures, ranging up to 195°F (90°C), may be used to develop highly flavored and colored malts with low enzyme activity. Barley sprouts are removed following drying, and malted grain is suitable for grinding and blending.

Use of malts. Malted barley is used in the brewing industry as an active source of starch-hydrolyzing enzymes (amylases) to improve the energy source, glucose, during fermentation and to impart the desired flavor and color to the brew. Food-grade malts used for selective enzyme activity or characteristic flavor and color are of increasing importance in the formulation of baked goods and breakfast cereals. *See* AMYLASE; FOOD MANUFACTURING; MALT BEVERAGE.

Mark A. Uebersax

Bibliography. D. E. Briggs, *Barley*, 1978; N. L. Kent, Barley: Processing, nutritional attributes, technological uses, *Technology of Cereals*, chap. 12, 1978; W. H. Leonard and J. H. Martin, *Cereal Crops*, 1963; D. E. Mathre (ed.), *Compendium of Barley Diseases*, no. 8, 1982; S. A. Matz, *Cereal Science*, 1969; Y. Pomeranz, Barley, *Elements of Food Technology*, pp. 173–178, 1977; Y. Pomeranz, Food uses of barley, *CRC Critical Reviews in Food Technology*, pp. 377–394, 1974; U.S. Department of Agriculture, *Barley*: *Origin, Botany, Culture, Winter Hardiness, Utilization, Pests*, 1979.

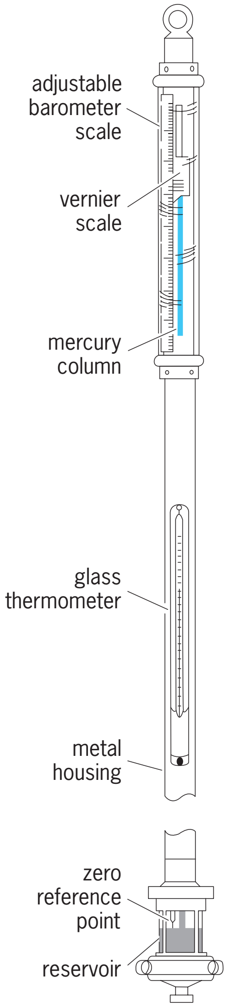
Baroclinic field

A distribution of atmospheric pressure and mass such that the specific volume, or density, of air is a function of both pressure and temperature, but not either alone. When the field is baroclinic, solenoids are present, there is a gradient of air temperature on a surface of constant pressure, and there is a vertical shear of the geostrophic wind. Significant development of cyclonic and anticyclonic wind circulations typically occurs only in strongly baroclinic fields. Fronts represent baroclinic fields which are locally very intense. *See* AIR PRESSURE; FRONT; GEOSTROPHIC WIND; SOLENOID (METEOROLOGY); STORM; WIND.
Frederick Sanders; Howard B. Bluestein

Barometer

An absolute pressure gage which is specifically designed to measure atmospheric pressure. This instrument is a type of manometer with one leg at zero pressure absolute. *See* MANOMETER.

The common meteorological barometer (see **illus.**) is a liquid-column gage with mercury. The top of the column is sealed, and the bottom is open and submerged below the surface of a reservoir of mercury. The atmospheric pressure on the reservoir keeps the mercury at a height proportional to that pressure. An adjustable scale, with a vernier scale, allows a reading of column height. Before each reading, the scale must be adjusted to correct for mercury level in the reservoir. Often a peg (the zero reference point) attached to the scale is adjusted by a micrometer to just touch the mercury surface. The apparent mercury height must be corrected for temperature and gravity, with standard conditions being 0°C and 980.665 cm/s^2.



Type of mercury barometer used in meteorology.

Typically, the barometer scale is in millimeters or inches of mercury, with sea-level pressures in the order of 760 mm or 29.9 in. of mercury.

Aneroid barometers using metallic diaphragm elements are usually less accurate, though often more sensitive, devices, and not only indicate pressure but may be used to record. *See* PRESSURE MEASUREMENT.
John H. Zifcak

Barotropic field

A distribution of atmospheric pressure and mass such that the specific volume, or density, of air is a function solely of pressure. When the field is barotropic, there are no solenoids, air temperature is constant on a surface of constant pressure, and there is no vertical shear of the geostrophic wind. Significant cyclonic and anticyclonic circulations typically do not develop in barotropic fields. Considerable success has been achieved, paradoxically, in prediction of the flow pattern at middle-tropospheric elevations by methods which are strictly applicable only to barotropic fields, despite the fact that the field in this region is definitely not barotropic. The subtropics, however, are to a large extent barotropic. *See* AIR PRESSURE; BAROCLINIC FIELD; GEOSTROPHIC WIND; SOLENOID (METEOROLOGY); WEATHER FORECASTING AND PREDICTION; WIND. Frederick Sanders; Howard B. Bluestein

Barracuda

Perciform fishes in a single genus, *Sphyraena*, in the family Sphyraenidae. They are characterized by an elongate, slightly compressed body; a long, pointed head with a large horizontal mouth with nonprotractile premaxillae and the lower jaw projecting beyond

Great barracuda (*Sphyraena barracuda*). (*Photo © John E. Randall*)

Names, distributions, habitats, and characteristics of 25 species of barracuda (*Sphyraena*)

Species	Common name	Distribution	Habitat	Max.length/weight
S. acutipinnis	Sharpfin barracuda	Indo-Pacific	Lagoons and seaward reefs	80 cm (32 in.)
S. afra	Guinean barracuda	Eastern Atlantic	Continental shelf to 75 m, enters lagoons and estuaries	205 cm (81 in.)/50 kg (110 lb)
S. argentea	Pacific barracuda	Eastern Pacific	Usually near shore in small schools	145 cm (57 in.)/12 kg (26 lb)
S. barracuda	Great barracuda	Indo-Pacific, Eastern and Western Atlantic	Estuaries, reefs, and open ocean to 100 m depth	200 cm (79 in.)/50 kg (110 lb)
S. borealis	Northern sennet	Western Atlantic from Massachusetts to Gulf of Mexico	Reefs	46 cm (18 in.)
S. chrysotaenia	Yellowstripe barracuda	Indo-Pacific		30 cm (12 in.)
S. ensis	Mexican barracuda	Eastern Pacific, Mexico to Ecuador	Continental shelf	127 cm (54 in.)/9.5 kg (21 lb)
S. flavicauda	Yellowtail barracuda	Indo-Pacific	Lagoons, sheltered seaward reefs	60 cm (24 in.)
S. forsteri	Bigeye barracuda	Indo-Pacific	Outer reefs	75 cm (30 in.)
S. guachancho	Guaguanche	Eastern and Western Atlantic	Turbid coastal waters, often in river estuaries, seaward to 100 m	20 cm (8 in.)/1.75 kg (3.9 lb)
S. helleri	Heller's barracuda	Indian Ocean and islands of the Pacific	Bays and coral reefs	80 cm (32 in.)
S. idiastes	Pelican barracuda	Southeast Pacific, Cocos and Galápagos islands	Reefs	91 cm (36 in.)
S. japonica	Japanese barracuda	Western Pacific, southern Japan, and South China Sea	Near shore	35 cm (14 in.)
S. jello	Pickhandle barracuda	Indo-West Pacific	Bays and seaward reefs, to 200 m depth	150 cm (59 in.)/11.5 kg (25 lb)
S. lucasana	Cortez barracuda	Eastern Central Pacific		70 cm (28 in.)
S. novaehollandiae	Australian barracuda	Indo-Pacific	Open channels and semiprotected areas	100 cm (39 in.)
S. obtusata	Obtuse barracuda	Indo-Pacific	Seagrass beds and rocky reefs in bays and estuaries, depth to 120 m	55 cm (22 in.)
S. picudilla	Southern sennet	Western Atlantic, Bermuda, Florida, Uruguay	Rocky and coral reefs of coastal waters	61 cm (33 in.)
S. pinguis	Red barracuda	Northwest Pacific	Near shore on mud or rock bottoms	30 cm (12 in.)
S. putnamae	Sawtooth barracuda	Indo-West Pacific	Bays, current-swept lagoons, and seaward reefs	90 cm (35 in.)
S. qenie	Blackfin barracuda	Indo-Pacific and Mexico and Panama in the Eastern Pacific		170 cm (67 in.)
S. sphyraena	European barracuda	Eastern Atlantic and Bermuda and Brazil in Western Atlantic	Coastal and offshore waters	165 cm (65 in.)
S. tome	No common name	Southwest Atlantic, Brazil, and Argentina	Depth to 83 m	45 cm (18 in.)
S. viridensis	Yellowmouth barracuda	East Central Atlantic, including the Azores	Probably coastal and offshore waters	128 cm (50 in.)
S. waitii	No common name	Indo-West Pacific		31 cm (12 in.)

the upper; strong, sharp, uneven conical teeth in the jaws as well as in the roof of the mouth (palatine teeth) and usually a large canine tooth near the tip of the lower jaw; two dorsal fins, both short, the first located above the pelvic fins with five strong spines, the second usually with one spine and nine soft rays, well removed from the first and a mirror image of the anal fin; short pectoral and pelvic fins; a forked caudal fin; cycloid scales that are smooth to the touch; and coloration that is usually gray to green or bluish above, with silvery reflections, lighter below, and sometimes with dark vertical bars or chevrons (see **illustration**).

Barracudas inhabit the tropical and subtropical Atlantic, Pacific, and Indian oceans, usually near the surface, but they may descend to 100 m (330 ft) or more. All are pelagic marine species; some frequently enter brackish waters, especially the juveniles. They are swift and powerful swimmers and voracious predators, feeding mostly on small fishes and to a lesser extent on shrimp and squid. For the most part barracudas are harmless; however, large individuals are known to attack humans, and in some parts of the world they are more feared than sharks. Although a food fish, some barracudas are known to contain ciguatera poison.

The known distribution, habitat, and characteristics of each of the 25 named species are given in the **table**. *See* PERCIFORMES.

Herbert Boschung

Bibliography. D. P. de Sylva, Sphyraenoidei: Development and relationships, in H. G. Moser et al., *Ontogeny and Systematics of Fishes*, Amer. Soc. Ichthy. Herp. Spec. Pub. no. 1, 1984.

Barretter

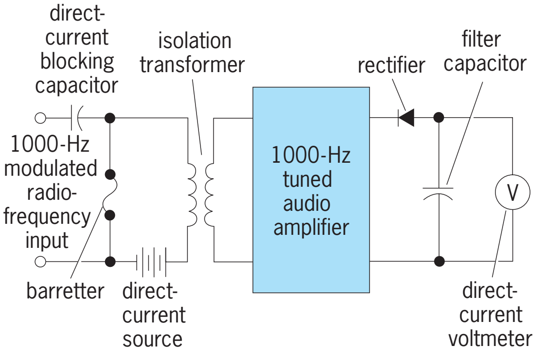
A bolometer element with a positive temperature coefficient of resistance, used to detect and measure power at radio, microwave, infrared, and optical frequencies. The temperature of the barretter increases when electromagnetic energy is absorbed. Barretters are made of metal; therefore, the electrical resistance increases when the temperature increases. The resulting resistance change of the barretter is measured by using direct-current or low-frequency instruments. Early radio demodulators and current regulator tubes using fine iron wires also have been called barretters.

Barretters for measuring power from less than 1 microwatt to several milliwatts are made from Wollaston wire. Wollaston wire is a platinum wire less than 100 microinches (2.5 micrometers) in diameter drawn inside a silver wire with a diameter about 10 times larger. The silver permits handling the wire for mounting and is then etched off to leave a bare platinum wire of the desired resistance.

The barretter resistance is selected to absorb most of the power when the barretter is mounted as a termination in a waveguide or coaxial transmission line. A barretter can be made to detect power at optical and infrared frequencies by using a very thin metal ribbon blackened to absorb light. *See* TRANSMISSION LINES; WAVEGUIDE.

Barretters with less sensitivity and accuracy for use at radio frequencies can be made by using low-current fuses made with fine wires. Fuses intended to protect instruments and rated for about 10 milliamperes work well. For higher powers, small lamp filaments can serve as barretters. *See* FUSE (ELECTRICITY); INCANDESCENT LAMP.

Measurement of modulated power. A meter can be made to measure high-frequency signal amplitudes using a barretter (see **illus.**). The temperature and hence the resistance of a barretter can change at audio-frequency rates, but the time constant of a barretter is too great for the resistance to vary at radio-frequency rates. A radio- or microwave-frequency current modulated at a low frequency will cause the barretter resistance to follow the low-frequency signal. If a direct-current voltage is applied to the barretter while the modulated radio-frequency current is also applied, the varying resistance will produce a current which follows the modulation. The low-frequency current can be coupled to the input of an audio amplifier tuned to the modulation frequency by using an audio transformer. The output of the audio amplifier may be rectified to drive a direct-current meter. The meter then indicates the relative amplitude of the radio-frequency or microwave signal. *See* AUDIO AMPLIFIER.


direct-current blocking capacitor
isolation transformer
rectifier
filter capacitor
1000-Hz modulated radio-frequency input
1000-Hz tuned audio amplifier
V
barretter
direct-current source
direct-current voltmeter

Circuit for measuring modulated radio-frequency power by using a barretter.

Bridge circuit. For precision measurements, the barretter can be used as one arm of a Wheatstone bridge. The dc power dissipated in the barretter and the barretter resistance can be calculated from the resistances of the other three arms of the bridge and the voltage applied to the bridge. If the barretter and the bridge resistors have the same temperature coefficient of resistance, the measuring circuit will be self-compensating for ambient temperature variations. *See* BRIDGE CIRCUIT; RESISTANCE MEASUREMENT; WHEATSTONE BRIDGE.

High-frequency effects. The current path and heating effects in a barretter are not exactly the same at low frequencies and high frequencies. The simple cylindrical geometry of a barretter, however, permits accurate calculation of the current and heat distribution. When direct currents heat the barretter, the center of the barretter is at a higher temperature than the surface because the power is dissipated uniformly but the heat is dissipated primarily at the surface. At high frequencies, however, the power is dissipated primarily at the surface because of skin effect. Direct currents dissipate power uniformly along the length of the barretter, but at higher frequencies, as the wavelength of the radio-frequency wave approaches the length of the barretter, the current approaches a sinusoidal distribution and the heating becomes nonuniform along the wire. The ratio of direct current to radio-frequency heating must be carefully calculated or, for critical work, calibrated against known standards. *See* BOLOMETER; ELECTROMAGNETIC RADIATION; SKIN EFFECT (ELECTRICITY).

Robert C. Powell

Bibliography. G. H. Bryant, *Principles of Microwave Measurements*, 1988; T. S. Laverghetta, *Modern Microwave Measurements and Techniques*, 1988.

Barrier islands

Elongate accumulations of sediment formed in the shallow coastal zone and separated from the mainland by some combination of coastal bays and their associated marshes and tidal flats. Barrier islands are typically several times longer than their width and are interrupted by tidal inlets. There are at least three possible explanations for their origin: (1) longshore spit development and subsequent cutting of inlets; (2) upward shoaling by wave action of subtidal longshore sand bars; and (3) drowning of coastal ridges. All three modes of origin may have occurred. The first two have been observed in recent times, the third is yet to be conclusively demonstrated.

Formation. Modern barrier islands extend along approximately 15% of the Earth's coasts. They are geologically quite young, having been formed about 7000 years ago or less, as sea level slowed or came to its present position during the Holocene transgression. Some are still in the process of forming and gaining definition. The primary requisites for their formation are a significant amount of available sediment, a place for it to accumulate, and significant wave-energy influence. Tectonic setting is an important control in the development of barrier islands, with most forming along the American-type trailing edge coasts of mature continental margins such as the east coasts of North and South America. The low-relief coastal plain and adjacent shelf are served by well-developed drainage systems with abundant sediment. Barriers are also well developed along marginal sea coasts such as the Gulf of Mexico. Both of these settings provide ideal conditions to produce long and continuous barrier island systems. This contrasts with leading-edge coasts on converging plate boundaries where there is high relief. This type of coasts tends to produce numerous but small barrier spits emanating from headlands, such as along the Pacific coast of the United States. *See* CONTINENTAL MARGIN; PLATE TECTONICS.

Barrier environments. It is not possible to discuss barrier islands without considering the adjacent and closely related environments within the coastal system. Beginning offshore and proceeding toward land, the sequence of environments comprises shoreface, including the offshore and nearshore; beach; dunes; washover fans, which comprise the back-island flats; marsh; tidal flat; and coastal bay, which may be a combination of estuarine and lagoonal environments leading to the mainland (**Fig. 1**). The barrier island proper includes the beach, dunes, and washover fans; however, the adjacent shoreface is well integrated with the barrier in terms of morphology, processes, and sediments.

Shoreface. The shoreface extends from the zone where storm waves influence bottom sediment transport, shoreward to the low-tide line. Typically it is subdivided into the outer shoreface, that area affected only by storm waves, and the inner shoreface or nearshore zone, where normal waves influence the bottom. The latter contains shallow longshore sand bars and troughs, varying in number from one coastal location to another. The depths of water marking the boundaries of the shoreface vary with location, depending upon wave climate.

Beach. The barrier island itself begins at the low-tide or outer limit of the beach. In many places the lower part of the beach is characterized by an ephemeral bar and trough generally called a ridge and runnel (Fig. 1). This small bar moves onto the upper beach under low-energy wave conditions, either repairing earlier storm erosion or adding to the beach, thereby causing the island to prograde. The upper intertidal or foreshore beach extends up to the berm, where the foreshore and backshore beach environments meet. The upper foreshore slopes toward the water and is the site of the uprush and backwash of waves—the swash zone. The backshore part of the beach is a dry, eolian environment except during storms. It tends to be horizontal or slightly landward-sloping. *See* EOLIAN LANDFORMS.

Dunes. Coastal dunes occupy a wide range of portions of the barrier island. The most common site of dune development is immediately landward of the beach in the form of shore-parallel ridges. These are called foredunes and may rise several meters above the beach. Barrier islands that receive abundant sediment and tend to prograde seaward may have numerous foredune ridges. *See* DUNE.

Low dunes or low areas between dunes may act as pathways for water and sediment during storm surges, producing washover fans on the landward, flat side of the barrier. This part of the barrier is generally only a meter or two above sea level and is covered with a floral community that extends from grasses in the high areas down to marsh grasses as the high

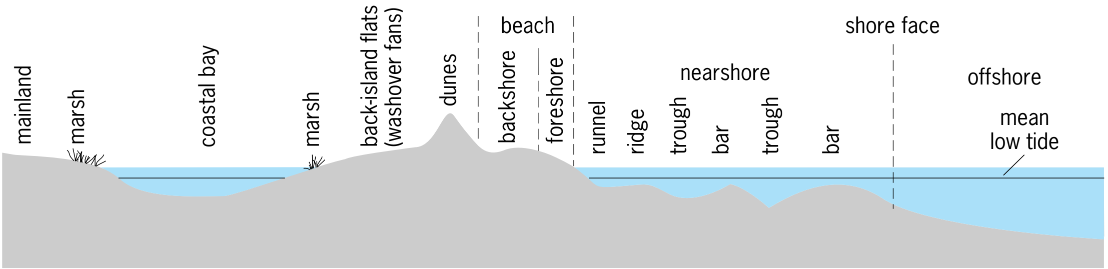


Fig. 1. Generalized profile across a barrier island complex showing various environments.

intertidal zone is reached. Sediment that supports this vegetation is delivered in the form of washover fans which coalesce to form a washover apron—the landward part of the barrier. An individual fan may be about a half meter in thickness and extend over several acres. Large storms may generate sufficient surge to produce washover fans that traverse the entire barrier and extend into the adjacent back-barrier bay.

Marshes and tidal flats. These are the landward continuation of the back-barrier environment. Their extent tends to be related to the combination of tidal range and morphology of the island; generally the higher the tidal range, the wider the back-barrier marsh and tidal-flat complex.

Coastal bay. The coastal bay that separates the barrier island from the mainland may take on various characteristics. In many areas it is an estuary, a bay that is diluted by fresh-water runoff from streams and that has good tidal circulation with the open ocean. Examples are found behind the Outer Banks barrier system of North Carolina and the barriers of the Florida Gulf Coast. The other extreme is a coastal bay that is a lagoon with no significant fresh-water input and very restricted tidal flux. Examples are Laguna Madre behind Padre Island on the southern Texas coast, and the Coorong lagoon in South Australia—both well over 100 km (60 mi) long. Some barrier systems include coastal bays that have important seasonal variation in their hydrography.

Inlets. The continuity of barriers is interrupted by inlets, which carry great quantities of water and sediment between the open marine environment and the coastal bays. Inlet spacing, size, and stability vary, depending upon the dominant processes controlling barrier island morphology. Inlets may be closely spaced such as along the Georgia and South Carolina coasts, or they may be widely spaced such as on the Outer Banks of North Carolina and on the Texas coast.

Processes. A variety of physical processes occur along the coast. These processes act to shape and maintain the barrier island system and also to enable the barrier to slowly migrate landward as sea level rises. The most important process in forming and maintaining barriers is the waves, which also produce longshore currents and rip currents. Waves and longshore currents dominate the outer portion of the barrier system, whereas tidal currents are dominant landward of the barrier, although small waves may also be present. Tidal currents also dominate inlet processes, although waves also may influence this environment. On the supratidal portion of the barrier island, the wind is the most dominant physical process.

Waves. Normal waves that cross the inner shelf and break in the surf zone do not cause significant sediment motion until they steepen and break in shallow water, typically over a longshore bar. Such breaking causes much sediment to be temporarily suspended and generates local currents that may carry sediment toward or away from the shore. Storm waves cause sediment motion in relatively deep water. They also tend to remove sediment from the beach and transport it seaward. The reverse situation exists during low-wave-energy conditions. Examples of this condition occur seasonally along much of the west coast of the United States. The winter storms produce large waves which remove sediment from the beaches and transport it offshore, where it is stored until the lower-wave-energy conditions of the spring. The sediment then returns and the beach is restored. This cycle of erosional and accretional beach is common globally and is related to stormy versus quiet wave conditions.

Waves rarely approach the coast with the crests parallel to the shore. Instead, they approach at some acute angle and are refracted, thereby generating longshore currents that flow away from the acute angle made by the wave and the coast. These currents are the primary mechanism for distribution of sediment along the barrier beach and surf zone. Longshore or littoral drift may cause a net transport of hundreds of thousands of cubic meters of sediment to pass a given point in a year's time. Waves and longshore currents, in concert with nearshore topography, cause formation of rip currents that may transport sediment seaward. These currents are produced by the temporary piling up of water between the inner longshore bar and the beach. This unstable condition causes water to flow seaward, and the water does so through low areas of the bar. These narrow currents are quite local but may be dangerous to swimmers. *See* OCEAN WAVES.

Tides. Tidal currents carry great quantities of water and sediment through the inlets and to a lesser extent in the estuaries landward of the barrier islands. Much sediment is deposited just landward of the inlet as a flood-tidal delta. The equivalent seaward accumulation, the ebb-tidal delta, varies greatly in size and shape, depending on wave and tidal interactions. Along many barriers, waves produce longshore currents that cause inlets to migrate or even close; these are wave-dominated inlets. By contrast, other inlets have strong tidal currents that cut deep and stable inlets which show little wave influence; these are tide-dominated inlets.

Tides are also important landward of the barriers in the estuaries, marshes, and tidal flats. Currents generated by tides are a primary mechanism for moving sediment both in suspension and as bedload. Flooding of low areas by tides provides a means for depositing sediment on the marsh or tidal flat. Typically the rate of tidal sediment flux is directly related to the tidal range. *See* OCEAN CIRCULATION.

Wind. The wind is also an important agent for sediment transport and therefore for molding barrier island morphology. Prevailing winds along the coast typically have some onshore component. The back-beach zone, or at low tide the entire beach, serves as a sediment source for onshore winds that accumulate this sediment behind the beach in the form of dunes. Wind also causes sediment to blow from the dunes back on the washover apron or marsh and also directly into the bay. Along parts of Padre Island in southern Texas, wind has blown so much sediment

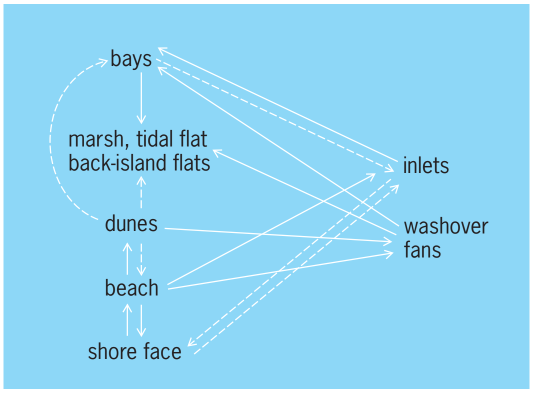


Fig. 2. Sediment budget for barrier island complex. Solid lines are major interactions; broken lines are relatively minor intersections.

into Laguna Madre that it is essentially filled. *See* NEARSHORE PROCESSES; WIND.

Sediment budget. The combined processes operating in the various sedimentary environments of the barrier island complex are best viewed through a sediment budget (**Fig. 2**). There are two distinct sources of sediment for this system (excluding biogenic or skeletal debris): the bays and the shoreface. Both of these environments receive feedback from other environments also. Shoreface sediment is transferred primarily to the beach, with small amounts going directly to the inlet. Beach sediment moves toward four sites: dunes, alongshore to inlets, over the barrier by storm washover, or seaward to the shoreface. Dunes also serve as a sediment source for washover fans, and they contribute as well to the back-island environments and bays through eolian transport of sediment. The marshes and other low-relief, back-island areas receive sediment from several sources, but do not provide sediment except in the case of severe storms, such as hurricanes, when even these low-energy environments are eroded.

The coastal bays show great variability in the sediment budget. Those bays that have little or no runoff from land or receive little tidal influence (lagoons) have little sediment input except from washover and blowover. Many bays receive much sediment runoff from the mainland and are subjected to pronounced tidal influence (estuaries). These bays provide sediment to marshes and tidal flats and some to inlets. They receive sediment from the same sources as lagoons, with the addition of inlets.

Morphodynamics. The response of the barrier island system to the above processes results in various sizes and shapes of barriers and their related inlets. These morphodynamics are dictated primarily by the interaction of wave- and tide-generated processes. Barrier islands are restricted to wave-dominated coasts and to coasts where there is a mixture of wave and tide influence. Tide-dominated coasts do not permit barrier islands to develop, due to the dominance of on- and offshore flux. Barrier islands on wave-dominated coasts are long and narrow with abundant and widespread washover fans (**Fig. 3***a*). Inlets tend to be widely spaced, small, and unstable due to the dominance of waves primarily through longshore currents.

Mixed-energy coasts produce barrier island systems that have short islands that are wide at one

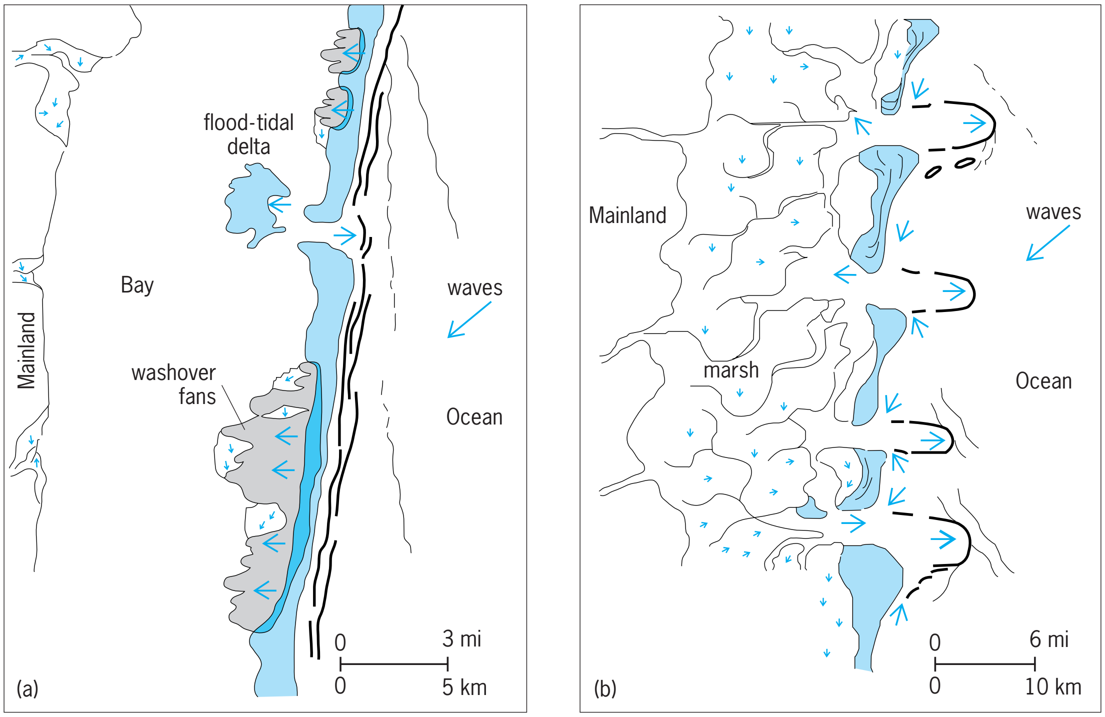


Fig. 3. Maps showing major coastal environments and directions, and directions of coastal processes of (*a*) wave-dominated barrier islands and (*b*) mixed-energy barrier islands. Large arrows indicate wave direction, and small arrows indicate direction of sediment transport. (*After S. P. Leatherman, ed., Barrier Islands from the Gulf of St. Lawrence to the Gulf of Mexico, Academic Press, 1979*)

end with large, fairly stable inlets resulting from the combination of wave and tide influence (Fig. 3*b*). Strong tidal currents in the inlets develop a substantial ebb-tidal delta that has an outer lobe smoothed by wave action. As waves are refracted around the ebb delta, there is a zone of longshore current reversal downdrift of the regional littoral drift direction. This condition causes sediment to become trapped and accumulate as prograding beach ridges. The remainder of the island is wave-dominated and is served by longshore currents moving in the direction of regional littoral drift. This part of the barrier island tends to be narrow and characterized by washover fans. Mixed-energy barriers have been termed drumstick barriers because of this shape.

Sediments and sedimentary structures. Although the barrier island system contains a broad spectrum of sediment types and sedimentary structures, each sedimentary environment can be characterized by its own suite. Knowledge of how these features formed and their spatial relationships with one another is important in recognizing and interpreting ancient barrier island systems preserved in the rock record.

The outer shoreface is typically composed of muddy sand or sandy mud with thin layers of sand. Bioturbation is nearly ubiquitous, and the surface may have inactive ripples or an irregular undulating surface from a previous storm. Discrete sand layers may show small-scale ripple cross-stratification and are related to storm events. The inner shoreface or nearshore zone displays great variety in sedimentary structures and has generally moderately to well sorted sand which may be a mixture of terrigenous and biogenic particles. Bedforms include symmetrical, wave-formed, two-dimensional ripples in the outer part, with three-dimensional megaripples on the seaward side of the longshore bars, where waves steepen producing combined flow conditions, and plane beds near the bar crests, where waves break. This sequence is repeated as each of the longshore bars is traversed by waves. The result is a complex but predictable pattern of bedforms and corresponding stratification types. Rip channels have seaward-oriented bedforms and cross-stratification due to the seaward transport of sediment in this localized environmnent.

The distinct zones within the beach also are readily distinguishable by their sediments and structures. Foreshore sediments are generally moderately to well sorted sand or gravel. Sorting values increase from worst to best from the base of the foreshore to the berm crest. Stratification is pronounced and slopes seaward at angles generally less that 7 or 8°. The great dynamics of this environment and the pronounced temporal changes in energy levels cause a range in the size of sediment particles deposited at any given time; some layers are sand, some are shell hash, and some are a combination. Erosion on the beach is typically marked by lag accumulations of dark, heavy minerals, resulting in thin layers of such minerals as zircon, garnet, magnetite, and rutile.

The backbeach is more uniformly finer-grained and better sorted that the adjacent foreshore. This is a response to eolian-dominated conditions, which may produce a thin surface deflation pavement of shell debris. Stratification is nearly horizontal and may be disrupted by burrows of various arthropods.

Dunes on the barrier island are characterized by fine, well-sorted sand which is dominantly quartz but may also contain carbonate shell debris. Dunes show the highest sorting values in the barrier system. There is a general trend of decreasing grain size and increasing sorting from the lower foreshore to the dunes. The surface of dunes is commonly covered with small wind-generated ripples. Internal stratification is dominated by large-scale cross-stratification formed as the active portion of the dune migrates. Dips range widely from about 10° to greater than 30°; generally a particular barrier is characterized by a given range in cross-stratification dips. Azimuths of the cross-stratification typically reflect prevailing and predominant wind directions.

The low-lying, back-island area is composed of moderately sorted sand dominated by horizontal, plane-bed stratification developed under sheet flow conditions of washover deposition. Roots and benthic invertebrates may destroy some of the stratification. Algal-mat-covered wind tidal flats may be widespread in the supratidal zone of some back-barrier areas. The blue-green algae stabilize the sediment under normal conditions but are torn up during storms, thereby permitting erosion by storm surge waters.

Marshes act as efficient sediment traps and contain a combination of mud, plant debris, and sand. Suspended sediment is transported onto the marsh as the tide rises, especially during storm surge conditions. Particles settle, and are trapped by the grass, resulting in slow accumulation of sediment. Washover fans may encroach upon the marsh, providing sand layers within the peaty muds.

Tidal flats receive sediment as the tides rise and fall, bringing both bedload sand and suspended mud. These sediments commonly accumulate in thin alternations, forming tidal bedding. The subtidal portion of the estuaries associated with barrier islands tends to be dominated by mud with variable amounts of terrigenous sand and shell gravel. Oyster shell debris is an important constituent of many low-energy estuaries, such as along the Atlantic and Gulf coasts of the United States. Some estuaries are influenced by strong tidal currents and have coarser sediment with bedforms, such as the Bay of Fundy in Nova Scotia, Canada.

Lagoons accumulate little terrigenous sediment because of the lack of mechanisms to transport it. Biogenic material and the precipitation of carbonate and evaporite minerals in some lagoons provide most of the sediment, with some contribution from both washover and blowover.

Inlets contain a wide variety of sediment texture and bedforms. Cross-stratification ranges from small-scale to large-scale, depending on the size of the bedforms present. A generally bimodal direction to this cross-stratification may be present, representing flooding and ebbing tidal currents.

Management. Continued pressures of growing populations have caused increased development of barrier islands, primarily for residential, commercial, and tourism purposes. This development along with the apparent increase in the rate of sea-level rise over the past several decades is causing severe problems. Natural barrier island processes include landward migration as washover occurs, and many tidal inlets tend to migrate along the coast. Construction of buildings, roads, and other infrastructure tends to lend some degree of permanence to barrier islands in that these structures are not designed to move as the island environments migrate.

Stabilization caused by structures commonly causes problems in various ways. Seawalls are designed to protect buildings or other structures from wave attack and coastal erosion. They also prohibit washover and landward migration, and commonly fail as wave action scours underneath them. Because of their nature, they prevent beaches from developing as barrier island retreat takes place. Structures that are perpendicular to the coast, such as groins and jetties, interrupt the littoral drift of sediment along the coast. These structures cause considerable sediment to pile up on their updrift side and result in erosion on the downdrift side because of lack of sediment supply. This situation is much like the sediment that is impounded behind a dam in a reservoir. Jetties are similar in their nature and are typically large because of their role in inlet stabilization. These concrete and rip-rap structures do a good job, but in the process they are also effective sediment traps on the updrift side, and the downdrift side of the inlet suffers serious erosion problems due to lack of sediment. Several locations around the United States now have sediment-bypassing systems to counteract this problem, but they are costly and commonly inoperable due to mechanical problems. New technology is needed to provide more efficient sand transfer across structured inlets.

There are federal, state, and local zoning restrictions on construction type and location, and other management programs that are in place. Most states have zoning that restricts construction on barriers through the implementation of some type of minimum setback from the active beach and dune environment. Some also prohibit seawalls and have limitations on rebuilding after loss due to storms. Extensive permitting procedures are involved in nearly all barrier island construction. New strict construction codes have resulted in tremendous reduction in structural damage due to severe storms such as hurricanes. Conditions associated with Hurricane Opal along the Florida panhandle coast in 1995 offer a good example. Of the nearly 2000 homes and other buildings along the open coast that were in the path of this storm, about 1000 incurred serious damage or were destroyed. Not one of the destroyed homes was built under recent construction codes, whereas homes built under these codes experienced no major damage.

Soft construction has become the primary method for protection of upland barrier properties. The most utilized approach is through beach nourishment, but various methods of constructing dunes and extensive vegetation programs are also common. All are designed to provide protection and stabilization to barrier environments while maintaining the esthetic appearance of the area and permitting natural processes to act on the islands.

Beach nourishment is expensive and temporary but has been effective in most places. This involves taking beach sand from a source location, generally offshore, and placing it on the beach in a specific design configuration. This solution typically costs a few million dollars per mile and lasts 5–10 years before additional nourishment is required. The cost of such projects is typically shared among federal, state, and local levels in decreasing proportions.

Although great strides have been made in coastal management over the past few decades, there is still a long way to go. Current regulations need to be enforced, and exceptions—which are too common—minimized. Like so many current environmental problems, the coastal zone is subjected to ever-increasing pressures of population and the related development. Barrier island systems are among the most affected. *See* COASTAL ENGINEERING.

Richard A. Davis, Jr.

Bibliography. R. W. G. Carter and C. D. Woodroffe (eds.), *Coastal Evolution*, 1994; R. A. Davis, Jr. (ed.), *Geology of Holocene Barrier Island Systems*, 1994; M. L. Schwartz (ed.), *The Encyclopedia of Beaches and Coastal Environments*, 1982.

Bartonellosis

An arthropod-borne infection caused by the bacterium *Bartonella bacilliformis*. The disease, also known as Carrión's disease, is confined to South America. It is characterized by two distinct clinical forms, Oroya fever and verruga peruana. Oroya fever causes a high temperature accompanied by the destruction of red blood cells; verruga peruana is a benign, nodular wart that develops on the surface of the skin. *See* ANEMIA.

Epidemiology. Bartonellosis occurs in an area that extends from 150 mi (240 km) north of the Equator to 975 mi (1570 km) south of it. It affects people on the western slopes of the Andes in Peru, Columbia, and Ecuador, at altitudes between 2000 and 8000 ft (600 and 2430 m). The geographic restriction of the disease is attributed to the limited habitats of *Phlebotomus*, its sandfly vector. *Phlebotomus verrucarum*, a night-biting sandfly with a short flight range, is the usual vector, but other species of sandfly have been implicated.

In the restricted geographic area, the disease is endemic. Ten to fifteen percent of the population have been shown to have *B. bacilliformis* in their peripheral blood. These asymptomatic cases, in addition to long-term carriers of the organisms, serve as reservoirs of infection. Animal reservoirs have not been identified.

Prevention of bartonellosis requires control of the vector, generally by spraying exteriors and interiors of buildings with insecticide. Insect repellents and bed netting provide significant personal protection.

Pathogenesis. Infection of nonimmune individuals caused by the bite of an infected sandfly results in the parasitization of erythrocytes. Large numbers of bacteria adhere to and penetrate erythrocytes, causing deep and long-lasting invaginations in erythrocyte membranes with resultant mechanical fragility. In individuals with Oroya fever, the organisms may be seen in Giemsa stains of the peripheral blood and in the cytoplasm of endothelial cells lining the blood vessels and lymphatics, where they occur in rounded masses. As many as 90% of the erythrocytes may be destroyed, leading to the severe anemia that is characteristic of the disease. *See* ANEMIA.

Oroya fever. The distinctive symptoms of the Oroya fever form of bartonellosis appear after an incubation period of 2–6 weeks. They include intermittent high fever, tender enlarged lymph nodes, diffuse muscle and joint pain, and the systemic symptoms associated with severe hemolytic anemia. Concurrent infections with salmonellosis, amebiasis, malaria, and tuberculosis are frequent and contribute to the severity and poor prognosis of Oroya fever. The mortality rate is 50% in untreated cases. Chloramphenicol is the recommended treatment, but the disease also responds well to penicillin, tetracyclines, and streptomycin.

Verruga peruana. The second stage of the disease, verruga peruana, may develop in survivors of Oroya fever after a variable period of time. This phase may also occur in individuals who have had a subclinical infection or who have never had a systemic *Bartonella* infection. The cutaneous phase is characterized by nodules that develop in successive crops over a period of 1–2 months, and may last for months to years. The verrugas (warts) are red to purple, diffuse or confluent, with individual lesions reaching 0.4–0.8 in. (1–2 cm) in diameter. They occur most frequently on exposed parts of the body but may also appear on mucous membranes and viscera. Fever and anemia are usually not present during this stage. *Bartonella* organisms are more difficult to identify in the blood during the cutaneous stage, but may be cultured from the verrugous lesions. The cutaneous lesions exhibit a variable response to antibiotic therapy. In verruga peruana the mortality rate is less than 5%.

Etiologic agent. The causative agent of bartonellosis is a motile, gram-negative, pleomorphic bacillus. In erythrocytes, organisms usually appear as short rods ranging from 1 to 3 micrometers in length and 0.25 to 0.5 μm in width. Forms vary from small coccoid and ring-shaped structures to slender bacilliform bodies. They stain poorly with the bacteriologic aniline dyes, but with the Giemsa stain they assume a reddish purple appearance and can be easily identified. In culture, the organisms possess unipolar flagella, but no flagella have been demonstrated in fresh preparations of clinical specimens.

Bartonella bacilliformis is aerobic, and may be cultured in the laboratory on semisolid agar containing fresh rabbit serum and rabbit hemoglobin. The organism grows over a temperature range of 28–37°C (82–99°F), but maximal longevity occurs at 28°C (82°F). *See* MEDICAL BACTERIOLOGY. **Hilda P. Willett**

Bibliography. J. C. Bartone, *Bartonella Infections: Index of Modern Information*, 1990; J. P. Kreier and M. Ristic, The biology of hemotrophic bacteria, *Annu. Rev. Microbiol.*, 35:325–338, 1981.

Barycentric calculus

The application of the properties of the centroid of a set of weighted points to geometry. For certain new operations such as the addition of point and point or of point and vector, it is possible to give their vector equivalents and work within the framework of vector algebra.

Theorems. A point P associated with a real number λ (for example, its mass or electric charge) is called a weight point and is denoted by λP. With a given origin O, the point λP has the companion vector $\lambda\overrightarrow{OP}$.

The sum $\mathbf{S} = \Sigma\lambda_i\overrightarrow{OP}_i$ of the companion vectors of n given points will be independent of the origin when, and only when, their total weight $\Sigma\lambda_i = 0$.

If $$\mathbf{S}' = \Sigma\lambda_i\overrightarrow{O'P}_i$$
then $$\mathbf{S} - \mathbf{S}' = \Sigma\lambda_i(\overrightarrow{OP}_i - \overrightarrow{O'P}_i) = (\Sigma\lambda_i)\overrightarrow{OO'}$$

Hence $\mathbf{S} = \mathbf{S}'$ implies $\Sigma\lambda_i = 0$ and conversely.

The properties of $\Sigma\lambda_i\overrightarrow{OP}_i$ thus depend on whether $\Sigma\lambda_i = 0$ (case 1) or $\Sigma\lambda_i \neq 0$ (case 2).

Case 1. $\Sigma\lambda_i = 0$. $\mathbf{S}$ may be computed by taking O successively at $P_1, P_2, \ldots, P_n$; then all n vectors $\Sigma_{i\neq j}\lambda_i\overrightarrow{PP}_i$ ($j = 1, 2, \ldots, n$) are equal to $\mathbf{S}$. This is one source of geometric theorems.

In particular, the vector $\mathbf{S}$ may be zero. When $n = 2, 3, 4$, the points are coincident, collinear, or coplanar, respectively. Writing $\overrightarrow{OA} = \mathbf{a}$, $\overrightarrow{OB} = \mathbf{b}, \ldots,$ gives the important theorems I and II.

Theorem I. Three distinct points A, B, C are collinear when, and only when, there are three nonzero constants α, β, γ such that Eqs. (1) hold.

$$\alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c} = \mathbf{0} \qquad \alpha + \beta + \gamma = 0 \tag{1}$$

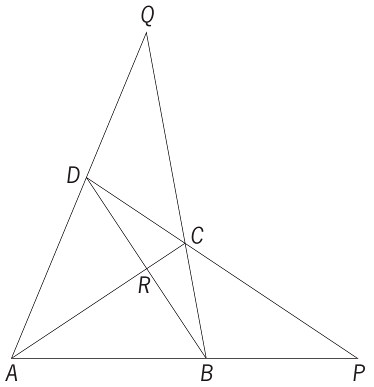
Theorem II. Four distinct points A, B, C, D, no three collinear, are coplanar when, and only when, there are four nonzero constants α, β, γ, δ such that Eqs. (2) hold.

$$\alpha\mathbf{a} = \beta\mathbf{b} = \gamma\mathbf{c} = \delta\mathbf{d} = \mathbf{0} \quad \alpha + \beta + \gamma + \delta = \mathbf{0} \tag{2}$$

The point C is said to divide the segment AB of a straight line in the ratio β/α if $\alpha\overrightarrow{AC} = \beta\overrightarrow{CB}$ or $\alpha(\mathbf{c} - \mathbf{a}) = \beta(\mathbf{b} - \mathbf{c})$; then Eq. (3) holds.

$$\mathbf{c} = \frac{\alpha\mathbf{a} + \beta\mathbf{b}}{\alpha + \beta} \tag{3}$$

Thus Eq. (1) implies that A, B, C divide BC, CA, AB in the ratios γ/β, α/γ, β/α.



Complete plane quadrilateral and its diagonal points.

Equation (2) implies that *ABCD* is a plane quadrilateral whose diagonal points *P, Q, R* (see **illus.**) are given by Eqs. (4). Metric and projective properties

$$\mathbf{p} = \frac{\alpha\mathbf{a} + \beta\mathbf{b}}{\alpha + \beta} \quad \mathbf{q} = \frac{\beta\mathbf{b} + \gamma\mathbf{c}}{\beta + \gamma} \quad \mathbf{r} = \frac{\gamma\mathbf{c} + \alpha\mathbf{a}}{\gamma + \alpha} \tag{4}$$

of the complete quadrilateral readily follow.

Case 2. $\Sigma\lambda_i = \lambda \neq 0$. The sum $\mathbf{S}_0 = \Sigma\lambda_i\overrightarrow{OP_i}$ now depends on O. But Eq. (5) defines the same point λP^*

$$\Sigma\lambda_i\overrightarrow{OP_i} - \lambda\overrightarrow{OP^*} = \mathbf{0} \tag{5}$$

for any choice of O; for the total weight of the points $\lambda_i P_i$ and $-\lambda P^*$ is zero. If O is taken at P^*, this gives the basic relation defining P^* according to Eq. (6). The

$$\Sigma\lambda_i\overrightarrow{P^*P_i} = \mathbf{0} \tag{6}$$

point λP^* is called the centroid of the points $\lambda_i P_i$. For an arbitrary origin the centroid is given by Eq. (7).

$$\lambda\mathbf{p}^* = \Sigma\lambda_i\mathbf{p}_i \tag{7}$$

On the right the sum extended over any subset of nonzero weight may be replaced by their centroid. This replacement principle is a second source of geometric theorems. It also reduces the problem of centroid location to that of two points.

If $\alpha = \beta \neq 0$, the centroid of αA and βB is the point $(\alpha + \beta)C$ given by $(\alpha + \beta)\mathbf{c} = \alpha\mathbf{a} + \beta\mathbf{b}$; and Eq. (3) shows that C divides AB in the ratio β/α. The centroid of A and B is $2\mathbf{c} = \mathbf{a} + \mathbf{b}$, and C is the midpoint of AB.

The centroid $3D$ of A, B, C is $3\mathbf{d} = \mathbf{a} + \mathbf{b} + \mathbf{c}$. Since Eq. (8) holds, the medians of a triangle meet in

$$3\mathbf{d} = \mathbf{a} + 2\frac{\mathbf{b} + \mathbf{c}}{2} = \mathbf{b} + \frac{\mathbf{c} + \mathbf{a}}{2} = \mathbf{c} + \frac{\mathbf{a} + \mathbf{b}}{2} \tag{8}$$

a point which divides each in the ratio 2:1.

The meaning of $\Sigma\lambda_i\mathbf{p}_i = \mathbf{0}$ when $\Sigma\lambda_i = 0$ is now apparent. If such a set is divided into two groups of nonzero weight by transposing some $\lambda_i\mathbf{p}_i$, the subsets have centroids at the same weighted point, whose weights differ only in sign.

Barycentric coordinates. If ABC is a triangle of reference, weights α, β, γ can be applied to A, B, C so that the centroid of αA, βB, γC is at any given point P in the plane of ABC; Eq. (9) holds; only the

$$(\alpha + \beta + \gamma)\mathbf{p} = \alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c} \tag{9}$$

ratios of the barycentric coordinates α, β, γ are determined; and Eq. (10) can be written, where PBC

$$\frac{\alpha}{PBC} = \frac{\beta}{PCA} = \frac{\gamma}{PAB} = \frac{\alpha + \beta + \gamma}{ABC} \tag{10}$$

is the area of the triangle PBC taken + or − depending on whether the circuit PBC is taken in the same sense as ABC or the opposite.

The equation $l\alpha + m\beta + n\gamma = 0$ represents a straight line; and $\alpha + \beta + \gamma = 0$ is the line at infinity.

In 3-space the weights α, β, γ, δ given to the vertices of a reference tetrahedron $ABCD$ are the barycentric coordinates of the centroid of αA, βB, γC, δD. $l\alpha + m\beta + n\gamma + k\delta = 0$ represents a plane; and $\alpha + \beta + \gamma + \delta = 0$ is the plane at infinity. The ratios of α, β, γ, δ may now be given in terms of oriented tetrahedra. *See* CENTROIDS (MATHEMATICS).

Louis Brand

Baryon

Strongly interacting composite particle that accounts for almost all of the visible mass of the universe. The proton and neutron that make up the atomic nucleus are examples of baryons. Strongly interacting particles (hadrons) are either baryons, which are fermions with half-integral spin, or mesons, which are bosons with integer spin. *See* FUNDAMENTAL INTERACTIONS; HADRON; MESON; NEUTRON; PROTON; QUANTUM STATISTICS.

Although baryons are very small (roughly 10^{-15} m in diameter), they consist of even smaller, elementary particles called quarks, together with the exchange particles of the strong interaction between the quarks, the gluons. This strong interaction is based on a three-valued “color” degree of freedom carried by the quarks. Modern particle theory has an as-yet-unproven postulate called confinement, which states that the only strongly interacting particles that can emerge from interactions and exist in isolation are those that carry no net color. This explains why quarks have not been seen in isolation. As the smallest number of colored quarks which can make up a colorless state is three, just as colorless (white) light is made from the three primary colors, baryons must contain at least three quarks. These are referred to as valence quarks. By contrast, mesons are made up of a minimum of one quark-antiquark pair, which can also be colorless. *See* COLOR (QUANTUM MECHANICS); GLUONS; QUARKS.

Because the interaction that binds quarks into a baryon is very strong, any number of quark-antiquark pairs, which are generically called sea quarks, can also be present in a baryon. Any antiquark which exists in a baryon must be a sea quark, but it may not be possible to distinguish between a sea quark and a valence quark with the same intrinsic properties.

Baryons are identified by properties referred to as

The six flavors of quarks, their charges, and their masses

Charge = $+2e/3$	Flavor: Mass, MeV/c^2:	up (u) 3	charm (c) 1250	top (t) 174,000
Charge = $-e/3$	Flavor: Mass, MeV/c^2:	down (d) 6	strange (s) 120	bottom (b) 4250

flavor, spin, and parity, all of which have discrete values. They also have measurable properties which are continuous, such as their mass and, for all but the proton which is the lightest baryon and so stable against decay into a lighter baryon plus other particles, their decay lifetime. The eventual product of all baryon decays is a proton. If more than one type of decay chain is possible, the fraction of decays which proceed through one mode or another may also be measured. A description of how the discrete properties of baryons are found from those of their constituent particles follows, and involves some interesting physical ideas.

Flavor. There are six different kinds, or flavors of quarks, in three families (see **table**). The lightest family contains the up (u) and down (d) quarks, with masses of roughly 3 and 6 in convenient units of MeV/c^2 (where MeV is a million electronvolts, the kinetic energy gained by an electron that is accelerated by a million-volt potential, and use is made of Einstein's relation $E = mc^2$ or $m = E/c^2$), and charges of $+^2/_3$ and $-^1/_3$ in units of the electron charge (e), respectively. These light quark masses are very small, on the scale of the mass of the proton, 938 MeV/c^2. *See* ELECTRONVOLT; FLAVOR.

The proton is made up of two up quarks and a down quark, labeled *uud*, and the neutron has flavor *udd*. The neutron is only slightly more massive than the proton, by about 1.3 MeV/c^2. Roughly 3 MeV/c^2 of mass is added by changing an up quark to a down quark, but this is largely compensated for by a loss in electromagnetic self-energy made by the same change. This is the energy required to collect the on-average positively charged quarks into the proton against the electrostatic repulsion. The resulting small size of the neutron-proton mass difference prevents neutrons from decaying when bound into most nuclei. If the down quark were significantly more massive than the up quark, the visible universe would be made up of only hydrogen atoms, as the neutrons in all nuclei would decay to protons, which cannot bind into nuclei by themselves.

Because the difference between the masses of the up and down quarks and the electromagnetic self-energy of quarks in a baryon are so small compared to baryon masses, the properties of baryons are to a very good approximation unchanged by replacing an up quark with a down quark. This is referred to as isospin symmetry. In this language the proton and neutron are referred to collectively as different charge states of the nucleon. *See* I-SPIN.

There is a third quark that is light on the scale of the proton mass, the strange quark s, with a mass of about 120 MeV/c^2 and charge $-e/3$. Baryons that differ from the proton and neutron by the conversion of an up or a down quark into a strange quark are denoted Λ or Σ. There are three Σ charge states: Σ^+ with quark content *uus*, Σ^0 with quark content *uds*, and Σ^- with quark content *dds*. The Λ^0 also has quark content *uds*, but differs from Σ^0 by its isospin symmetry. The flavor-dependent part of its quantum-mechanical wave function changes sign under switching the up and down quarks, while that of the Σ states does not. Baryons that contain one up or down quark and two strange quarks are called cascade or Ξ baryons, with $\Xi^0 = uss$ and $\Xi^- = dss$. Baryons that contain three strange quarks, called Ω^-, have also been observed. Baryons containing strange quarks are collectively known as hyperons. *See* HYPERON; STRANGE PARTICLES.

The remaining three flavors of quarks, the charm c, bottom b, and top t quarks, are heavy on the scale of the proton mass (see table). Charmed baryons made up of a pair of up or down quarks and a charmed quark are called Λ_c and Σ_c, in analogy to the equivalent strange baryons. Baryons with a charmed quark, a strange quark, and an up or down quark have been observed, and are denoted Ξ_c, and those with a charmed quark and two strange quarks are denoted Ω_c. Some evidence for a doubly charmed *ccu* or *ccd* baryon, denoted Ξ_{cc}, has been found. The naming scheme for baryons containing a bottom quark is similar, although so far only Λ_b has been definitively observed in nature. The top quark is very massive, and this causes it to decay so rapidly that it does not have time to form a hadron by binding strongly to other quarks or to an antiquark. *See* CHARM.

Spin. Quarks carry an intrinsic degree of freedom called spin. This is related to the angular momentum carried by a macroscopic object when it rotates, felt by everyone who has tried to stop a rotating merry-go-round or bicycle wheel. In the very short distance, quantum world, angular momentum is quantized in units of $\hbar = h/(2\pi)$, where h is Planck's constant. The smallest possible angular momentum is $\hbar/2$, and this is the spin of each of the quarks making up a baryon. Such particles are referred to as spin-$^1/_2$ particles, or spinors. To find the spin of baryons made up of quarks and gluons, first consider the rules for combining the spins of three spin-$^1/_2$ particles. Two spin-$^1/_2$ particles can have their spins combined to 0 (antialigned along an axis called the quantization axis) or to 1 (aligned along the quantization axis), in units of $\hbar$. With the addition of a third quark there are two ways to get to a spin of $^1/_2$, namely, combining spin 0 with spin $^1/_2$ or spin 1 with spin $^1/_2$; and one way to get a spin of $^3/_2$, namely, combining spin 1 with spin $^1/_2$. This means that the contribution to the spin of a baryon from the spin of the three valence quarks that must be present is either $^1/_2$ or $^3/_2$. *See*

ANGULAR MOMENTUM; SPIN (QUANTUM MECHANICS).

The gluons, which mediate the color force between the quarks, like the photons, which mediate the electromagnetic force between particles, carry a spin angular momentum of $\hbar$ and are referred to as spin-1 or vector particles. In addition to the spin angular momentum carried by the quarks and gluons, these particles also have orbital angular momentum as they move around their common center of mass. In the simplest possible (naïve quark model) picture, the spin angular momentum of a baryon is made up of the orbital angular momentum and intrinsic spin of the three valence quarks, and the contribution of the sea quarks and gluons is neglected.

Just as the lightest states of the hydrogen atom are spherically symmetric and have zero orbital angular momentum between the electron and proton, in this simple picture the lightest states of a baryon have zero orbital angular momentum between the quarks, and so their total spin is just that of the quarks, either $^1/_2$ or $^3/_2$. The proton and neutron do indeed have spin $^1/_2$, and the lightest excited baryon containing only the *u* and *d* quarks which make up the proton and neutron, called $\Delta(1232)$, has spin $^3/_2$. The mass of the Δ is given in brackets after its name, in MeV/c^2. It has four charge states, Δ^{++}, Δ^{+}, Δ^{0}, and Δ^{-}, consisting of *uuu*, *uud*, *ddu*, and *ddd* quarks. The mass difference between the Δ and the nucleon can be thought of as due to a kind of magnetic interaction between the colored, spinning quarks, analogous to the magnetic interaction between charged, spinning electrons. These interactions are repulsive when the quark spins are aligned; this raises the mass of the spin-$^3/_2$ Δ, with all three quark spins aligned, relative to that of the nucleon, with only two of the quark spins aligned. *See* DELTA RESONANCE; ELECTRON SPIN.

Baryons with the same quark content as the proton, neutron, and four Δ charge states exist with spin angular momentum higher than $^1/_2$ or $^3/_2$. In most cases this is because the quarks have orbital angular momentum around their common center of mass, but it is possible for the gluons and sea quarks present in baryons to carry angular momentum. Since orbital angular momentum comes in integral multiples of $\hbar$, which is also true of the intrinsic spin of a gluon or quark-antiquark pair, all baryons must have half-integral spin $J = {}^1/_2, {}^3/_2, {}^5/_2$, and so forth. Particles with half-integral spin are called fermions. Nucleon and Δ-flavored baryons containing only *u* and *d* valence quarks have been clearly observed with spins up to $^{11}/_2$, and states with spins up to $^7/_2$ have been observed for baryons containing strange quarks.

Exchange symmetry and color. The lightest baryons with a particular flavor content and isospin symmetry [like the nucleon and the $\Delta(1232)$, containing only up or down valence quarks] have either spin $^1/_2$ or spin $^3/_2$. Such states are called ground-state baryons. Since quarks have spin $^1/_2$, they are fermions, and the Pauli principle implies that the quantum-mechanical wave function of a baryon made up of quarks must be antisymmetric (change sign) under the exchange of any pair of identical quarks. When this idea is put together with the (isospin) symmetry of the strong interactions under the exchange of up and down quarks, the result is that the wave functions of the nucleon and Δ must be antisymmetric under exchange of any pair of valence quarks. *See* EXCLUSION PRINCIPLE.

In the case of the ground state Δ^{++}, this leads to a paradox. The flavor content is *uuu*, which is exchange symmetric. The three quark spins are maximally aligned to spin $^3/_2$, which is also symmetric. The $\Delta(1232)$ is the least massive Δ-flavored baryon and so should not have orbital angular momentum, which would significantly increase its mass. Orbital angular momentum zero states are known to be spherically symmetric and exchange symmetric. How can the wave function which describes the Δ^{++} be antisymmetric, as required by the Pauli principle?

The resolution of this paradox led to the postulate that the quarks carry an additional degree of freedom. The minimal assumption required for antisymmetry is that this degree of freedom is three-valued. This degree of freedom was called color. The modern theory of the strong interactions, called quantum chromodynamics (QCD), grew out of the conviction that the interaction between colored quarks and antiquarks is responsible for binding them into baryons and mesons, just as the interaction between charged nuclei and electrons is what binds them into atoms. *See* QUANTUM CHROMODYNAMICS.

Parity. Another interesting discrete property of a baryon is its parity. The strong interactions between colored quarks have the property that they do not change when viewed in a mirror. This is essentially the same as spatial inversion symmetry, invariance under the operation of reflecting all of the spatial coordinates through the origin. As a consequence, under the operation of inversion of the coordinates, the quantum-mechanical wave functions of states made up of strongly interacting particles either remain the same, called positive parity, or change sign, called negative parity. Also, if a system goes from one state to another via a strong interaction, the parity of the entire system cannot change. By examining the strong decays of excited baryons it is possible to determine if they have positive parity, the same as that of the proton, or negative. All ground-state baryons have positive parity.

After ground-state baryons, the next most massive excited states are those with negative parity, for all flavors of baryon. These states have the lowest possible orbital angular momentum, $\hbar$, which can result from the quarks orbiting around their common center of mass. The quantum-mechanical wave function of such states is known to change sign under the operation of spatial inversion. Negative-parity excited states have been discovered for nucleons and Δ-flavored baryons, as well as for Λ, Σ, and Ξ hyperons, and charmed baryons Λ_c and Σ_c. *See* PARITY (QUANTUM MECHANICS).

Strong decays. The $\Delta(1232)$ decays to a proton and a pion in a very short time (roughly 10^{-24} s) via the strong interaction. The same is true of the equivalent spin-$^3/_2$ $\Sigma(1385)$ baryon, which decays to the spin-$^1/_2$ Σ baryon. The pion is made up of a minimum of an up or down quark and an up or down antiquark, and is the lightest of the mesons, particles made up of a minimum of one quark-antiquark pair. Because the difference between the mass of the Δ and the nucleons is more than 140 MeV/c^2, the mass of the pion, it can decay very rapidly to a nucleon and a pion via the strong interaction. If the strong decay of any baryon into a less massive state consisting of a baryon and a meson is allowed by conservation of energy, then this will be how it decays.

In one simple picture of such strong decays, a quark-antiquark pair is created via the strong interaction, with the created quark ending up in the final baryon and the created antiquark in the final meson. The flavor of the created quark must be the same as that of the created antiquark, leading to no net change of flavor between initial and final states linked by a strong interaction. The strong decay of the Δ proceeds by the creation of an up or down quark-antiquark pair. More highly excited nucleon and Δ-flavored baryons can also decay via the creation of a strange-quark pair, creating a Λ or Σ hyperon and a *K* meson containing a strange antiquark, if their mass exceeds the sum of the masses of these final-state particles.

Electromagnetic and weak decays. Certain decays of a baryon would result in a state with the same quark content, but cannot proceed via the strong interaction because of conservation of energy. An example is the decay of the Σ^0 to the Λ^0, with masses that differ by less than a pion mass. Such decays proceed via the emission of an energetic particle of light, a gamma-ray photon, as the electromagnetic interaction between quarks, made possible by their electric charges, is the next strongest of the fundamental interactions between them. Electromagnetic decays are rapid; the lifetime of the Σ^0 is about 10^{-19} s. *See* GAMMA RAYS; PHOTON.

The baryon eventually resulting from a chain of strong and possibly electromagnetic decays will be a ground-state, spin-$^1/_2$ baryon whose quark content may differ from *uud*, that of the proton. In that case, to decay to the proton, this baryon must change the flavor of one of its quarks, but neither the strong nor the electromagnetic interactions can bring about such a change. Such decays proceed significantly more slowly via the weak interactions, with typical lifetimes of between 10^{-13} and 10^{-10} s, depending on the mass difference between the initial and final baryons involved. An exception is the weak decay of the down quark to the up quark, which allows an isolated neutron to decay to a proton, with an 886-s lifetime because of the tiny (1.3-MeV/c^2) neutron-proton mass difference.

Baryon decays that proceed via the weak interactions are mediated by the massive vector-boson particles W^+, W^-, or Z^0. Such decays can result in the creation of a charged lepton and a neutrino of the same family from a *W* boson, such as an electron and an electron antineutrino, or the electron's antiparticle (the positron) and an electron neutrino, and are known as semileptonic decays. An important example is the beta decay of the neutron to a proton, electron, and electron antineutrino. If allowed by conservation of energy, the more massive mu and tau leptons may be created. *See* ELECTRON; INTERMEDIATE VECTOR BOSON; LEPTON; NEUTRINO.

Alternatively, a quark-antiquark pair with mismatched flavors can be created from the decay of a weak vector boson, and the result is a flavor-changing decay involving a baryon and a meson in the final state. For example, the Σ^+ decays to a proton and a π^0, or a neutron and a π^+, removing a strange quark but maintaining the electric charge of the initial state, with roughly equal probability. Such decays are called nonleptonic weak decays. *See* WEAK NUCLEAR INTERACTIONS.

Baryon production. Since only protons and neutrons occur naturally in our environment, the knowledge that we have gained about other baryons has required their production in experiments involving collisions of particles. This production proceeds via the strongest of the mechanisms causing the decays of baryons, the strong and electromagnetic interactions. Nucleon and Δ-flavored baryons can be created by scattering a beam of relatively long-lived pions off nuclear targets which contain protons and neutrons, or by beams of energetic photons. Similarly, Λ- and Σ-flavored baryons can be made by scattering beams of *K* mesons, which contain the required strange quark, off nuclear targets. Baryons containing more strange quarks are produced by the strong decay of a Λ or Σ via the creation of a strange quark-antiquark pair. Those containing more massive quarks can be created in an electron-positron or proton-antiproton collider experiment, where these particles annihilate and form a charmed or bottom quark-antiquark pair. These states ultimately form a number of hadrons (with an equal number of baryons and antibaryons), which can include a heavy-quark baryon. *See* ELEMENTARY PARTICLE; PARTICLE ACCELERATOR.

Simon Capstick

Bibliography. F. Close, *The Cosmic Onion: Quarks and the Nature of the Universe*, Heinemann Educational Secondary Division, 1984; S. Eidelman et al. (Particle Data Group), Review of particle physics, *Phys. Lett.*, B592:1, 2004; N. Isgur and G. Karl, Hadron spectroscopy and quarks, *Phys. Today*, 36(11):36–42, 1983.

Basalt

An igneous rock characterized by small grain size (less than about 0.2 in. or 5 mm) and approximately equal proportions of calcium-rich plagioclase feldspar and calcium-rich pyroxene, with less than about 20% by volume of other minerals. Olivine, calcium-poor pyroxene, and iron-titanium oxide minerals are the most prevelant other minerals. Certain rare varieties of basalts are poor in SiO_2 and rich in melilite, olivine, and nepheline and may be poor in calcium-rich plagioclase and pyroxene. Most basalts are dark gray or black, but some are light gray. Various structures and textures of basalts are useful in inferring both their igneous origin and their environment of emplacement. Basalts are the predominant surficial igneous rocks on the Earth, Moon, and probably other bodies in the solar system. Several chemical-mineralogical types of basalts are recognized. The nature of basaltic rocks provides helpful clues about the composition and temperature within the Earth and Moon. The magnetic properties of basalts are responsible in large part for present knowledge of the past behavior of the Earth's magnetic field and of the rate of sea-floor spreading. *See* IGNEOUS ROCKS.

Extrusion of basalt. Basalt erupts out of fissures and cylindrical vents. Fissures commonly are greater than several feet wide. Cylindrical vents generally are many tens of feet in diameter and may pass downward into cylindrical shell-shaped or planar fissures. Repeated or continued extrusion of basalt from cylindrical vents generally builds up a volcano of accumulated lava and tephra around the vent. Fissure eruptions commonly do not build volcanoes, but small cones of tephra may accumulate along the fissure. *See* LAVA; VOLCANO; VOLCANOLOGY.

Form and structure. Basalts display a variety of structures mostly related to their environments of consolidation. On land, basalt flows form pahoehoe, aa, and block lava, while under water, pillow lava is formed. Basaltic lava flows on land have rugged to smooth surfaces. The rugged flows are called aa and have top surfaces made of spines and ridges up to several meters high. They are typically composed of three layers: a massive interior overlain and underlain by rubbly zones. Flows with smooth surfaces are called pahoehoe. Pahoehoe has been seen to form on fluid lava flows as the incandescent surface rapidly advances (3–33 ft/s or 1–10 m/s), cools below red heat, and stiffens into a rigid crust about an inch thick in about a minute. There are transitional types of surfaces, as well as a separate kind called block lava, which is similar to aa lava but has a surface made up of equidimensional blocks up to about 3 ft (1 m) in diameter.

Pillow lava is characterized by lens-shaped bodies (pillows) of basalt up to several feet in length which have upward convex tops and partially upward convex bottoms conforming to the convex tops of underlying pillows. Such structure is found in basalts dredged from the deep-sea floor, in basalts which extruded under ice caps, as in Iceland, and in basalts which have flowed into lakes, rivers, and the sea from vents on land.

Basalt occurs as pumice and bombs as well as lava flows. Commonly, basaltic pumice is called scoria to distinguish it from the lighter-colored, more siliceous rhyolitic pumice. Basaltic pumice can have more than 95% porosity. Bombs are relatively smooth, round bodies up to about 3 ft (1 m) in diameter whose surfaces commonly are cracked as a result of expansion of the effervescing interior which remains fluid longer than the quickly cooled surface. Bombs with cracked surfaces are called bread-crust bombs, the name attesting both to the fluid condition and to the gas content of the extruded material. Pele's hair and Pele's tears are terms applied to threads and drops of basaltic glass blown out of vents. Pasty material intermediate between pumice and bombs is called spatter and commonly forms agglutinate or vent breccia near vents of basalts. Accumulations of pumice, bombs, and spatter commonly develop around vents and form cinder cones, spatter cones, and spatter ramparts. Similar, but commonly much smaller, secondary vent deposits may develop on lava flows distant from the primary vent. Secondary vents are rooted within the flow and do not penetrate it; they are called hornitos. Littoral cones may form where a lava flow encounters abundant water and water vapor blows through the liquid lava. *See* VOLCANIC GLASS.

Basalt commonly contains tiny round voids called vesicles. In pumice, vesicles generally make up most of the volume. The vesicle porosity in flows is high at the top and bottom of flows, but it decreases dramatically a few meters away from cooling surfaces. In general, vesicles are smallest (0.04–0.4 in. or 1–10 mm in diameter) in the top and bottom couple of feet of flows, reach a maximum of a couple of inches in diameter several feet from the cooling surface, and then decrease in size and abundance toward the center of the flow. Vesicles in basalts on the sea floor (about 3 mi or 5 km depth) are very tiny (about 10 μm diameter), and the total porosity is commonly less than 1%. Many pillow basalts now on land lack vesicles or have only a few tiny vesicles and presumably formed in deep water. However, there are basalts on land which are vesicle-poor, and some sea-floor basalts have significant vesicles. In sum, the presence of vesicles indicates that basalts evolve gas in low-pressure environments, and the distribution of vesicles within flows provides a means of identifying the tops and bottoms of flows. Flows of basalt develop a characteristic kind of fracture called columnar joining, whereby the rock splits into polygonal columns oriented perpendicular to the cooling surfaces.

Physical properties. Many physical properties of basalt have been measured in the field and laboratory, making it one of the best-characterized rock types. The physical properties of basalts vary greatly because of the range of textures and compositions. Temperature ranges from 1830 to 2230°F (1000 to 1220°C) for flowing basalt. The viscosity for magma is 10^3–10^6 poise (10^2–10^5 N · s/m^2), increasing as the temperature falls and the crystal content of the magma increases. The surface tension for magma is 270–350 dynes/cm (0.27–0.35 N/m). For melt the density is 2.6–2.7 g/cm^3 (1.5–1.6 oz/in.^{3}) at 2190°F (1200°C), while for crystalline rock at standard temperature and pressure (STP), void-free (calculated from simple mixtures of pyroxene, plagioclase, and iron oxides consistent with the definition of basalt), density is 2.8–3.1 g/cm^3 (1.6–1.8 oz/in.^{3}). The velocity of sound is 1.4 mi/s (2.3 km/s) for a compressional wave in melt at 2190°F (1200°C), and 3.5 mi/s (5.7 km/s) for rock at 1290°F (700°C). Both isothermal and adiabatic compressibilities are closely equal to 7×10^{-12} cm^2/dyne for melt at 2190°F (1200°C), and 2×10^{-12} for rock at 1470°F (800°C). The crushing strength and cohesive strength of cold rock are 1700–2200 atm (170–220 megapascals) and 320–440 atm (32–44 MPa), respectively. The coefficient for volumetric thermal expansion of melt is about 3.8×10^{-5} per °F (2×10^{-5} per °C). The electrical conductivity is about 100 per ohm per centimeter for melt at 2190°F (1200°C). Thermal conductivity is about $1\text{–}3 \times 10^{-3}$ cal/cm s °C for melt at 2190°F (1200°C), and $4\text{–}6 \times 10^{-3}$ cal/cm s °C (1 cal = 4.19 joules) for rock at STP. Heating capacity at constant pressure is about 0.2–0.3 cal per °C per gram. About 100 cal/g is generally assumed for the heat of melting, but the figure is possibly uncertain by 30%. Theoretical and observed diffusion coefficients for Mg^{2+} in basaltic melt at 2140°F (1170°C) are 0.62×10^{-9} in.^{2}/s (4×10^{-9} cm^2/s) and 11×10^{-9} in.^{2}/s (72×10^{-9} cm^2/s).

The magnetic susceptibility of basalts is commonly about 10^{-4} to 4×10^{-3} emu/g. The natural remanent magnetization intensity of basalts depends upon the amount of magnetite, the composition of the magnetite, the grain size, the cooling history, and the paleointensity of the Earth's field. Observed values of natural remanent magnetization intensity are commonly between 10^{-4} and 10^{-3} emu/g for subaerial basalts, but they increase to 10^{-2} to 10^{-1} emu/g for basalts from the sea floor. Thus, the remanent magnetization of submarine basalts is stronger than their induced magnetization—a relationship that has made it possible for the sea floor to preserve a "tape-recorded" history of the reversals of the Earth's magnetic field. The natural magnetization of many basalts is sufficiently strong to be detected with a magnetic compass. *See* MAGNETOMETER; ROCK MAGNETISM.

Mineralogy and texture. The mineralogy and texture of basalts vary with cooling history and with chemical composition. With slow cooling (few degrees Fahrenheit per day 33 ft or 10 m deep in a flow), crystals grow to large sizes (millimeters) and acquire equant shapes. In basaltic melt, diffusion may limit the rate at which a given crystal-liquid boundary can move because the crystals differ in composition from the melt. If crystallization is forced at a rapid rate by rapid cooling (about 180°F or 100°C per minute 4 in. or 10 cm deep in the flow), either the crystals have high ratios of surface area to volume (needles, plates) or many seeds (nuclei) develop. Crystals which grow rapidly under strongly supercooled conditions have different compositions from those which form if slower cooling prevails. Consequently, the same basalt flow may contain different minerals in variously cooled parts of the flow.

The size of crystals in basalt reflects the rate of production of seed crystals (or nuclei) as well as the rate of growth of individual crystals. Crystals of pyroxene are able to grow at an average linear rate of about 8×10^{-10} in./s (2×10^{-5} μm/s), or 8×10^{-5} in./day (2 μm/day). At higher temperatures, in melts rich in magnesium and iron, crystals may grow as rapidly as 24×10^{-8} in./s (6×10^{-3} μm/s), or about 0.02 in./day (0.5 mm/day). The large crystals (phenocrysts) present in most basalts possibly formed in a few hours or days during the ascent of the magma through the crust. Alternatively, many large crystals possibly formed slowly in a body of stored magma. Such stored magma bodies, if thicker than about 300 ft (100 m), take many decades to cool, and their crystals may remain suspended or be picked up from the walls when magma erupts. *See* PHENOCRYST.

As basalt crystallizes, both the minerals and the residual melt change in composition because of differences between the composition of the melt and the crystals forming from it. In basalts, because of the rapid cooling, there is little chance for crystals to react with the residual melt after the crystals have formed. Early-formed crystals tend to become armored with later overgrowths. Because diffusion is sluggish in crystals even at the high temperatures at which basalts crystallize, the minerals generally retain their compositional zoning during cooling. Completely solid basalts generally preserve a record of their crystallization in the zoned crystals and residual glass. **Figures 1** and **2** illustrate the development of crystalline texture in basalt. **Figure 3** shows the sequence of crystallization in basaltic lava on Hawaii.

Basalts may alternatively yield a silica-saturated residual melt (rhyolite) or one deficient in silica to form nepheline, depending upon the initial bulk composition of the basalt. Basalts yielding nepheline upon crystallization are termed alkaline; those yielding rhyolite are called subalkaline (or tholeiitic).

Most basalts contain minor amounts of chromite, magnetite, ilmenite, apatite, and sulfides in addition to the minerals mentioned above. Chromite and the sulfides commonly form early in the sequence of crystallization, whereas magnetite, ilmenite, and apatite form late. Magnetite in basalts contains a history of the strength and orientation of the Earth's magnetic field at the time of cooling. Therefore, although magnetite is minor in amount, it is probably the most important mineral in terrestrial basalts, because it enables earth scientists to infer both the magnetic history of the Earth and the rate of production of basaltic ocean floor at the oceanic ridges.

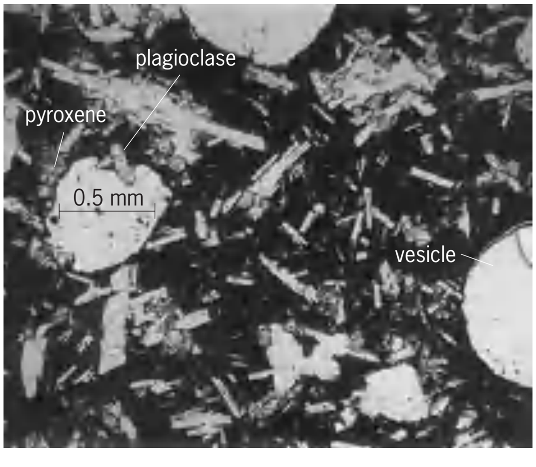


Fig. 1. Thin section of upper crust of lava about 1.2 in. (3 cm) from top; black is microcrystalline groundmass. 0.5 mm = 0.02 in.

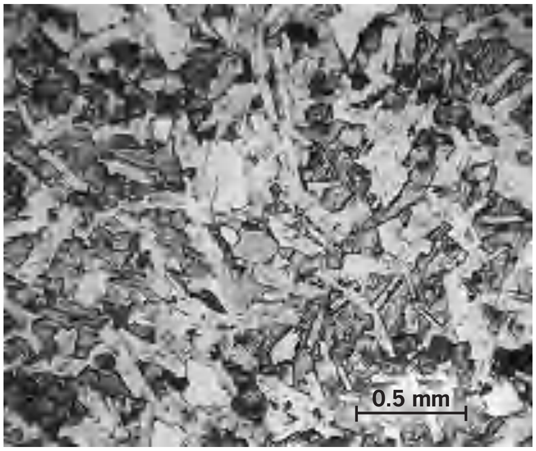


Fig. 2. Thin section of interior of lava shown in Fig. 1, but 26 ft (7.9 m) below the top of the flow. Clear plagioclase crystals are ophitically enclosed in large pyroxene grains up to 0.04 in. (1 mm) in diameter. Crystals of black magnetite are minor. A few irregular voids are present.

Chemical compositions of basalts, in weight percent*

	1†	2	3	4	5	6
SiO_2	49.92	49.20	49.56	51.5	45.90	45.5
TiO_2	1.51	2.32	1.53	1.1	1.80	2.97
Al_2O_3	17.24	11.45	17.48	17.1	15.36	9.69
Fe_2O_3	2.01	1.58	2.78	n.d.	1.22	0.00
Cr_2O_3	0.04	n.d.	n.d.	n.d.	n.d.	0.50
FeO	6.90	10.08	7.26	8.9	8.13	19.7
MnO	0.17	0.18	0.14	n.d.	.08	0.27
MgO	7.28	13.62	6.97	7.0	13.22	10.9
CaO	11.85	8.84	9.99	9.3	10.71	10.0
Na_2O	2.76	2.04	2.90	4.3	2.29	0.33
K_2O	0.16	0.46	0.73	0.80	0.67	0.06
P_2O_5	0.16	0.23	0.26	n.d.	0.62	0.10
Sum	100.00	100.00	100.00	100.00	100.00	100.02
H_2O	0.4	0.3	n.d.	2	n.d.	0.0
CO_2	0.02	0.01	n.d.	n.d.	n.d.	n.d.
F	0.02	0.03	n.d.	n.d.	n.d.	0.002
Cl	0.02	0.03	n.d.	0.09	0.01	0.0005
S	0.08	n.d.	n.d.	0.19	n.d.	0.07

*n.d. = not determined.

†Columns show (1) average of 10 basalts from oceanic ridges; (2) submarine basalt, Eastern Rift Zone, Kilauea Volcano, Hawaii; (3) average high-alumina basalt of Oregon Plateau; (4) initial melt of Pacaya Volcano, Guatemala; (5) alkali basalt, Hualalai Volcano, Hawaii; (6) average *Apollo 12* lunar basalt.

Chemical composition. Significant differences exist in the composition of basalt which relate to different tectonic environments. The **table** lists typical compositions of basalts from oceanic ridges, island arcs, continental platforms, and oceanic islands and of alkalic basalts which have no particular tectonic association. In addition to the common elements, the table gives the concentrations of H_2O, CO_2, F, Cl, and S typical of the various basalts. It is common practice to reduce chemical analyses of igneous rock to a volatile-free basis for comparison purposes, and this practice has been followed in the table.

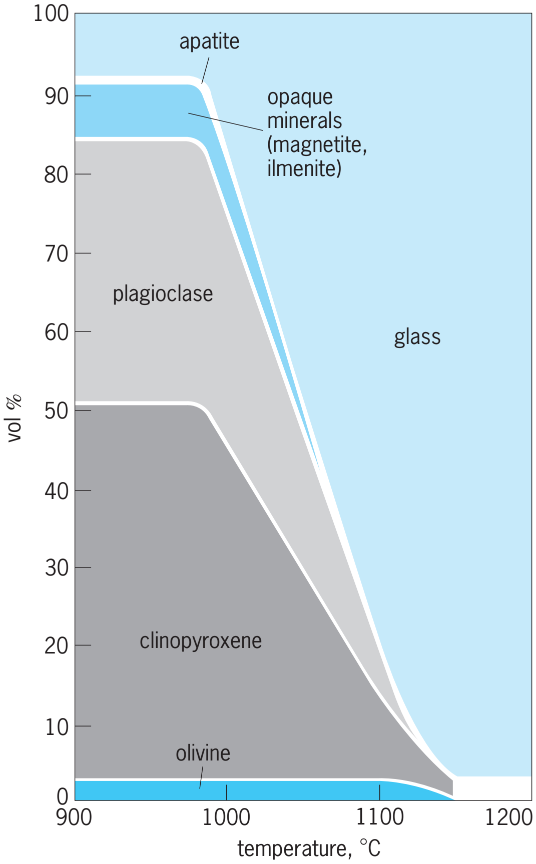


Fig. 3. Growth of crystals as a function of temperature in basalt in Hawaiian lava. °F = (°C × 1.8) + 32. (*After D. L. Peck, T. L. Wright, and J. G. Moore, Crystallization of tholeiitic basalt in Alae lava lake, Hawaii, Bull. Volcanol., 29:629–656, 1966*)

Chemical analyses of basalts are now used instead of, or together with, the textural criteria as a basis of classification. Geologists customarily recast the chemical analysis into a set of ideal minerals according to a set of rules. The result is called the norm of the rock. The principal normative minerals of basalts are the feldspars; diopside, $Ca(MgFe)Si_2O_6$; olivine, $(MgFe)_2SiO_4$; hypersthene, $(MgFe)SiO_3$; nepheline, $NaAlSiO_4$; and quartz, SiO_2. Because quartz reacts with olivine to form hypersthene and with nepheline to form feldspar (albite), basalts can be classified in order of increasing SiO_2 saturation by the presence of normative nepheline (lowest SiO_2), olivine, or quartz. Diagrams illustrating these relations and the particular names associated with individual combinations are shown in **Fig. 4**. In general, there is rather close correspondence between the normative minerals and the observed minerals in basaltic rocks.

Lunar basalts. Lunar basalts belong exclusively to the subalkaline class. They are similar to subalkaline basalts on Earth but have metallic iron rather than ferric-iron-bearing magnetite. Many lunar basalts are exceptionally rich in iron and titanium. They also differ compositionally from terrestrial basalts in being poor in alkali elements and siderophile (metal-seeking) elements such as nickel. *See* MOON.

Meteoritic basalts. Some meteorites are basaltic rocks. They differ significantly from lunar basalts and appear to have originated elsewhere in the solar system at a time close to the initial condensation of the solar nebula. Meteoritic basaltic rocks are called eucrites and howardites. *See* METEORITE.

Tectonic environment and origin. Basalts occur in four major tectonic environments: ridges in the sea floor, islands in ocean basins, island arcs and mountainous continental margins, and interiors of continents. The principal environment is the deep sea floor, where the rate of production of basalt is about 1.5×10^{16} g/year. Exploration of the Mid-Atlantic Ridge revealed that the ridge trough contains longitudinal depressions and mounds. It has been suggested that the floor of the ridge crest cracks open and is subsequently filled with a mound of pillow basalt on a time scale of 10,000–20,000 years. Basalts of the oceanic ridges typically are subalkaline in composition. Their chemical composition, tectonic environment, and the melting behavior of mantle rocks as a function of pressure suggest that ridge basalts evolve from primary melts that separate from partially melted mantle rocks at depths of a few tens of kilometers beneath the ridge (1 km = 0.6 mi).

Basaltic islands in the ocean basins contain diverse basalts. Major oceanic centers of volcanism (hot spots) such as Hawaii, Galápagos, and Réunion generally extrude subalkaline basalts. Although less abundant than subalkaline basalts, alkaline basalts are common on oceanic island volcanoes, and form mainly during the later stages of volcano development when rates of extrusion are less.

Island arc and continental margin basalts are predominantly subalkaline, but alkaline basalts are common away from the main volcanic front toward the continental interior. Island arc basalts differ compositionally from those at ocean ridges and islands. The differences, such as a higher concentration of alumina and more abundant feldspar phenocrysts, probably reflect the subduction and partial melting of siliceous and hydrous rocks mixed with altered oceanic crust and relatively little modified mantle rock. Water dissolved in magma modifies its crystallization behavior and stabilizes silica-poor, early-crystallizing minerals such as olivine and hornblende. *See* CONTINENTAL MARGIN; MARINE GEOLOGY; OCEANIC ISLANDS; PLATE TECTONICS.

Basalts in continental interiors fall into two classes: (1) major plateaus of lava flows extruded from giant fissures (Columbia River, Deccan Traps, Siberian Traps, Paranas Basin, Karoo, Tasmania), and (2) volcanoes (mostly small) of alkaline basalt commonly containing nodules of olivine-rich rocks apparently derived from the mantle.

The basalts of the major plateaus vary in chemical composition but are, with rare exception, subalkaline. Some of the plateau basalts are strongly similar to basalts from oceanic ridges, but are slightly richer in some of the elements normally found in high concentration in the granitic continental crust. Other plateau basalts are similar to the subalkaline basalts of oceanic islands. Still others are rather unique (Tasmania). The plateau basalts are unquestionably derived from the mantle, because they are erupted at too high a temperature to have been generated within the crust. The tectonic history and compositional aspects of plateau basalt provinces suggest that the provinces form in association with upwellings (plumes) of relatively pristine mantle rock from great depths.

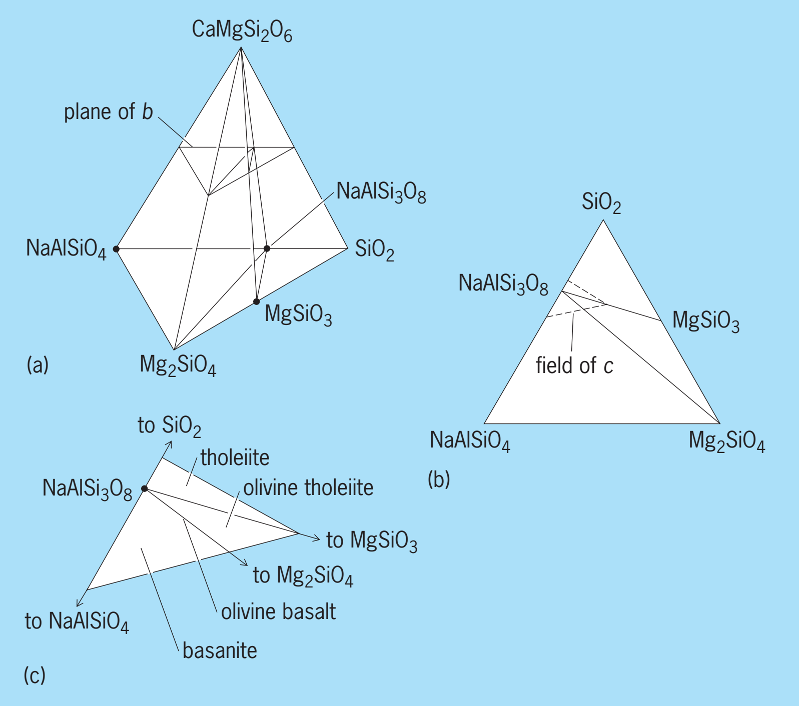


Fig. 4. Diagrams illustrating the compositions of basalts in terms of normative mineral formulas. Customarily, $CaAl_2Si_2O_8$ (anorthite) is grouped together with $NaAlSi_3O_8$ (albite), and Fe_2SiO_4, $FeSiO_3$, and $CaFeSi_2O_6$ are grouped with the corresponding Mg formulas. (*a*) The position of the basalt plane in a tetrahedron with $CaMgSi_2O_6$ (representing calcium-rich pyroxene) at the apex and plagioclase (represented by $NaAlSi_3O_8$) and other minerals indicates SiO_2 saturation on the base. (*b*) Diagram shows that natural basalts occupy only a small portion of the basalt plane. (*c*) Basalt plane from *b*.

Chemical and mineralogical variations. The variations which occur within individual basalt flows can mostly be explained by separation of crystals and their original associated melts. In some cases certain elements have been removed, gained, or redistributed by oxidizing and acid gases which permeate the cooling body of rock from surrounding air and the variably wet and vegetated ground.

In general, as the space and time frame of compared basalts is enlarged to include flows from other vents and eruptive episodes, explanation of chemical and mineralogical differences becomes more difficult. Two principal avenues of interpretation overlap: The first uses the crystallization process evident in individual lava flows broadened to include the effects of increasing pressure on the crystallizing (or melting) minerals as well as the complex flow patterns within large bodies of magma. The second uses a variable bulk-chemical composition of the parental material (heterogeneity of the mantle).

Because the dynamical and crystallization behaviors of basaltic magmas are incompletely predictable, only those compositional attributes insensitive to the melting and crystallization process can be taken as evidence for variation in the bulk composition of source. For the most part, these attributes have been limited to ratios of isotopes. As long as the bulk chemical compositions of sources are identical, it is expected that melting will yield isotopically identical primary basaltic liquids. Isotopically distinct liquids could form if there are major compositional differences or ages of source rocks.

Large ratios of $^3He/^4He$ in gases emitted near Yellowstone and elsewhere (larger than for ocean ridge basalts) suggest derivation of associated basaltic and other magmas from undifferentiated mantle source rocks. This is because helium is lost from the Earth's atmosphere and consequently is not available to be derived from an average collection of terrestrial source rocks. In such cases it is necessary to place some reliance on meteorites and other solar system data and even nucleosynthetic theory in order to justify an estimate of an initial undifferentiated composition. Therefore basalts eventually play an important role in concepts of the evolution of the solar system and its planets. *See* MAGMA. Alfred T. Anderson

Bibliography. P. Floyd (ed.), *Oceanic Basalts*, 1991; H. H. Hess (ed.), *Basalts*, 2 vols., 1968; A. R. McBirney, *Igneous Petrology*, 2d ed., 1992; H. Williams and A. R. McBirney, *Volcanology*, 1979, reprint 1982.

Base (chemistry)

In the Brönsted-Lowry classification, any chemical species, ionic or molecular, capable of accepting or receiving a proton (hydrogen ion) from another substance. The other substance acts as an acid in giving up the proton. A substance may act as a base, then, only in the presence of an acid. The greater the tendency to accept a proton, the stronger the base. The hydroxyl ion acts as a strong base. Substances that ionize in aqueous solutions to produce the hydroxyl ion (OH), such as potassium hydroxide (KOH) and barium hydroxide [$Ba(OH)_2$], are also conventionally called bases. Water (H_2O) acts as a weaker base in reaction (1), where H_3O^+ is the hydronium ion.

$$H_2O + H^+ \rightleftharpoons H_3O^+ \qquad (1)$$

Anions of weak acids, such as acetic and formic, act as bases in reacting with solvent water to form the molecular acid and hydroxyl ion, for example, the acetate ion (CH_3COO^-), as in reaction (2).

$$CH_3COO^- + H_2O \longrightarrow CH_3COOH + OH^- \qquad (2)$$

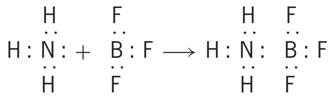
Ammonia (NH_3) and amines react similarly in aqueous solutions, as in reaction (3), where $NH_4{}^+$ is the ammonium ion.

$$NH_3 + H_2O \longrightarrow NH_4{}^+ + OH^- \qquad (3)$$

In these examples, the acetate ion and acetic acid (CH_3COOH) and NH_3 and the ammonium ion are conjugate base-acid pairs. The basicity constant, K_b, is the equilibrium constant for the proton transfer reaction, and it is a quantitative measure of base strength. For reaction (2), the basicity constant is given by expression (4), and its value at 298 K is 5.7×10^{-9}. For formate ion ($HCOO^-$), the corresponding K_b is 5.6×10^{-10}. Acetate is, therefore, the stronger base. While these concepts are most useful in aqueous solutions, they are equally valid in other solvent systems.

$$K_b = \frac{[CH_3COOH][OH^-]}{[CH_3COO^-]} \qquad (4)$$

The Lewis classification involves the concept of a base as a substance that donates an electron pair to an acid acceptor. In the gas phase [reaction (5)], NH_3 acts as a base in contributing an electron pair to the formation of a covalent bond with the boron trifluoride (BF_3) molecule. *See* ACID AND BASE.



$$\mathrm{H}:\underset{\mathrm{H}}{\overset{\mathrm{H}}{\overset{\cdot\cdot}{\underset{\cdot\cdot}{\mathrm{N}}}}}: + \underset{\mathrm{F}}{\overset{\mathrm{F}}{\overset{\cdot\cdot}{\underset{\cdot\cdot}{\mathrm{B}}}}}:\mathrm{F} \longrightarrow \mathrm{H}:\underset{\mathrm{H}}{\overset{\mathrm{H}}{\overset{\cdot\cdot}{\underset{\cdot\cdot}{\mathrm{N}}}}}:\underset{\mathrm{F}}{\overset{\mathrm{F}}{\overset{\cdot\cdot}{\underset{\cdot\cdot}{\mathrm{B}}}}}:\mathrm{F} \qquad (5)$$

Francis J. Johnston

Bibliography. B. W. Jensen, *The Lewis Acid-Base Concepts: An Overview*, 1979; D. R. Lide (ed.), *CRC Handbook of Chemistry and Physics*, 1993; K. W. Whitten, K. D. Gailey, and R. E. Davis, *General Chemistry*, 4th ed., 1992.

Basidiomycota

A phylum in the kingdom fungi; commonly known as basidiomycetes. Basidiomycetes traditionally included four artificial classes: Hymenomycetes, Gasteromycetes, Urediniomycetes, and Ustilaginomycetes. They are mostly filamentous fungi characterized by the production of basidia. These are microscopic, often club-shaped end cells in which nuclear fusion and meiosis usually take place prior to the maturation of external, typically haploid basidiospores, which are then disseminated. Common basidiomycetes are the rusts and smuts, which cause severe plant diseases, mushrooms (edible and poisonous), boletes, puffballs, stinkhorns, chanterelles, false truffles, jelly fungi, bird's-nest fungi, and conk or bracket fungi. Basidiomycetes are the most important decayers of wood, living or dead, in forests or buildings, causing either brown rot (for example, dry rot) or white rot. Many, especially mushrooms and boletes, are the primary fungal partners in symbiotic ectomycorrhizal associations with tree roots. Plant litter and soil are other major habitats. A few basidiomycetes are cultivated for food. Some are luminescent, hallucinogenic, lichenized, nematophagous, or aquatic. Some are cultivated by ants or termites, or are symbiotic with captured scale insects. Some can convert to a yeast (or single-cell) phase, one of which causes cryptococcosis in humans and animals. *See* MUSHROOM; MYCORRHIZAE; WOOD DEGRADATION.

In a typical life cycle, haploid basidiospores are forcibly discharged from projections on the basidia into the air, disperse, and germinate on a suitable substrate to form filaments (hyphae), which make up a vegetative stage (mycelium). This mycelium must anastomose with a compatible mycelium. Mating is controlled by one or two allelic genes resulting in two (bipolar) or four (tetrapolar) morphologically indistinguishable mating types. The mycelium resulting from mating has two compatible nuclei in each cell, and is said to be dikaryotic. This nuclear phase is unique to basidiomycetes and can last for years. Nuclear fusion eventually takes place in a basidium, either on the mycelium or on or in a fruit body (for example, mushroom or puffball), pustule, or sorus. Specialized sexual structures (pycnia, spermatia, protoaecia) are formed only by rusts. Many other basidiomycetes form characteristic short bridges (clamp connections) which act to bypass the wall (septum) between cells for one of the paired nuclei during mitosis. Many variations occur. Basidiospores may be dikaryotic when discharged. They may be passively released, formed in chains; dispersed dry or in slime by vectors such as insects or other agents such as water, either singly or in special organs; dispersed by special organs; or lacking entirely. The haploid phase may be yeastlike or lacking. The dikaryotic phase may be yeastlike. Nuclear fusion may occur in structures other than the basidia. The fungus may reproduce asexually by conidia or, if a rust, by aeciospores and urediospores; conidia may anastomose with mycelia to form dikaryons. *See* FUNGI; GASTEROMYCETES; HYMENOMYCETES; UREDINOMYCETES (RUST); USTILAGINOMYCETES (SMUT). Scott A. Redhead

Bibliography. G. C. Ainsworth, F. K. Sparrow, and A. S. Sussman, *The Fungi: An Advanced Treatise*, vol. 4B, 1973; D. L. Hawksworth et al., *Ainsworth & Bisby's Dictionary of the Fungi*, 8th ed., 1996; D. S. Hibbett et al., *Proc. Nat. Acad. Sci.*, 94:12002-12006, 1997; E. C. Swann and J. W. Taylor, *Can. J. Bot.*, Suppl. 1:S862-S868, 1995.

Basil

An annual herb (*Ocimum basilicum*) of the mint family (Labiatae), grown from seed for its highly scented leaves. It is sold fresh or dried, and its essential oil is used in pharmaceuticals and flavoring. The genus *Ocimum* has approximately 150 species, many of which are also grown as "basil." *See* LAMIALES.

Growth habit varies with type, but it is usually upright and branching from a square central stem up to 24 in. (61 cm) in height. Leaf shape and odor vary considerably from type to type, with "sweet basil," the most popular, having broad, flat, shiny green leaves to 3 in. (7.5 cm) long. Wild forms of basil are found in South America, Africa, India, and Europe. Cultivated basil differs little from its wild forms, although breeding and selection work have been performed on some of the varieties grown for essential oil production. Some of the more popular cultivated types are: sweet basil (used fresh or dehydrated), lemon-scented basil, opal basil (a purple-leaved type), and large- or lettuce-leaf basil. Many unusual types can be obtained which have a scent similar to camphor, lemon, licorice, or nutmeg.

Basil is grown commercially in many areas of the world. In the United States, the largest producer of dried leaf products is California, where the drier climate aids dehydration. Basil grown for oil production is centered in the eastern states, though some is also produced in Washington and Oregon. A considerable quantity of the basil consumed in the United States is imported from Europe.

All basil types have similar cultural requirements, which include high light intensity, 20–60% relative humidity, and a high level of soil fertility. Because of the large leaf suface, basil also requires a good supply of water. Deviation from these requirements results in either disease or low essential oil content. Since all propagation is by seed, basil is replanted annually. Large commercial farms utilize direct seeding into the field, planted in 3–6-in. (7.5–15-cm) rows, 24–42 in. (61–122 cm) apart.

Approximately 60 days after seeding, harvesting is begun, and may be repeated up to five times during the growing season. As is typical of nearly all herbs, regrowth occurs after trimming the plant. Harvesting is accomplished mechanically, utilizing machinery similar to that used in the mint industry. After transport from the field, either basil is dehydrated by forced air dryers or the oil is extracted with steam or solvent. Leaves are stripped from the stem after dehydration and then processed to obtain a uniform product.

Traditionally associated with use in Italian and southern European cuisines, basil is most often used in conjunction with other herbs for flavoring sauces, and is the primary ingredient in Italian pesto sauce. Basil oil is used in perfumery and as a source of eugenol. *See* SPICE AND FLAVORING. Seth Kirby

Bibliography. S. Arctander, *Perfume and Flavor Materials of Natural Origin*, 1960; L. H. Bailey, *Manual of Cultivated Plants*, rev. ed., 1975; J. C. Uphof, *Dictionary of Economic Plants*, 1968.

Basin

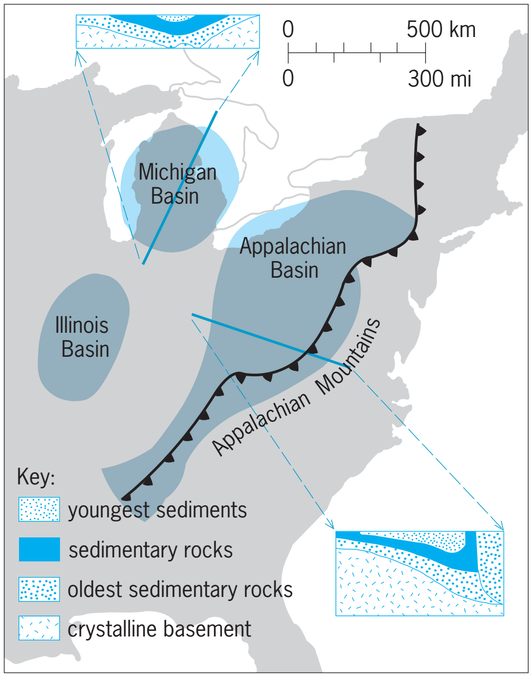
A low-lying area which is wholly or largely surrounded by higher land. An example is the Hudson Bay in northeastern Canada, which was formed by depression beneath the center of a continental ice sheet 18,000 years ago. Another example, the Qattara depression, is 150 mi (240 km) long and the largest of several wind-excavated basins of northern Egypt. Depressions in the ocean floor are also basins, as the Canary Basin, west of northern Africa, or the Argentine Basin, east of Argentina. These basins occur in regions where cold, dense oceanic crust lies between the topographically elevated ocean ridges and the continental margins. *See* CONTINENTAL MARGIN; MARINE GEOLOGY.

Drainage basins. A drainage basin is the entire area drained by a river and its tributaries. Thus, the

Mississippi Basin occupies most of the United States between the Rocky Mountains and the Appalachians. Interior drainage basins consist of depressions that drain entirely inward, without outlet to the sea. Examples may be quite small, such as the Salton Sea of southern California or the Dead Sea of central Asia. One of the most remarkable examples of an interior drainage basin is the Chad Basin in northern Africa, the center of which is occupied by Lake Chad. The fresh waters of the lake drain underground to feed oases in the lowlands 450 mi (720 km) to the northeast.

Geologic basins. In the geologic sense, a basin is an area in which the continental crust has subsided and the depression has been filled with sediments. Such basins were interior drainage basins at the time of sediment deposition but need not be so today. As these basins subside, the layers of sediment are tilted toward the axis of maximum subsidence. Consequently, when the sedimentary layers are observed in cross section, their geometry is a record of the subsidence of the basin through time and contains clues about the origin of the basin.

Characteristics. Many geologic basins are nearly circular in plan view and symmetrical in cross section, such as the Michigan Basin (see **illus.**) and the Paris Basin. The maximum subsidence in these basins occurred at a central point, producing a saucer-shaped depression. Other basins are elongate or trough-shaped in plan view, with lengths in some cases reaching thousands of miles. Some elongate basins have symmetrical cross sections, such as the rift basins of eastern Africa and the Newark Basins of eastern North America, while others are asymmetric in cross section and typically become deeper or thicker toward mountain ranges against which they abruptly terminate. Examples of elongate asymmetric basins are the Denver Basin, the Appalachian Basin (see illus.), and the Ganges Basin.



Paleozoic (approximately 250–570 million years old) basins of the eastern United States. The cross sections (insets) show the thickness and general pattern of deposition of sediments in the Michigan Basin and Appalachian Basin. The arrows show where that edge of the cross section is located on the plan-view map.

Origin. The different types of basins form in response to significantly different geologic processes. Elongate rift basins form where the Earth's lithosphere has been stretched in response to tensional forces, commonly, but not always, associated with the rifting and breakup of continents. Such basins are bounded by steeply dipping faults that were active as the basins subsided and filled with sediment. Many elongate asymmetrical basins form where the basin floor has been bent downward under the weight of slices of the Earth's crust that have been faulted, one on top of the other, along the edges of mountain ranges. Such basins, called foreland basins, commonly form where lithospheric plates have collided and the crust is under compression. *See* FAULT AND FAULT STRUCTURES; PLATE TECTONICS.

The most enigmatic basins are those that are symmetrical and have diameters approximately equal in all directions, such as the Michigan Basin. In contrast to the other types of basins, no geologic evidence of a mechanism has been found adjacent to or beneath these basins. They are thought to be the product of processes operating at a deep level within the Earth, but the exact nature of those processes is poorly understood.

The origin of geologic basins is a topic of continuing interest in both applied and basic geological studies. They contain most of the world's hydrocarbon reserves, and they are regarded as some of the best natural laboratories in which to understand the thermal and mechanical processes that operate deep in the interior of the Earth and that shape the Earth's surface. Gerard Bond; Michelle Kominz

Bibliography. P. A. Allen and J. R. Allen, *Basin Analysis: Principles and Applications*, 2d ed., 2005; M. Leeder, *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics*, 1999; A. D. Miall, *Principles of Sedimentary Basin Analysis*, 2d ed., 1990.

Basommatophora

A superorder of the molluscan subclass Pulmonata containing about 2500 species that are grouped into 11–15 families. Only a few members of the family Ellobiidae are terrestrial, with the other species today being tidal to supratidal or estuarine in habitat. The families Siphonariidae and Trimusculidae are marine limpets with caplike shells. The families Otinidae and Amphibolidae also are marine taxa. The remaining families inhabit a great variety of fresh-water situations and are quite varied in shell structure and shape. Because of this great variation in habitat and form, it is difficult to find structures that are common to all taxa and thus diagnostic of the group, which indeed may not have a common origin. The most easily observable characters are having the eyespots normally located at the base of two slender tentacles that are neither contractile nor retractile, and usually having two external genital orifices. Most features of the anatomy are shared with other pulmonate superorders, or are specializations related to major changes in habitat or body form.

Some of the few marine taxa are phylogenetic relicts (Otinidae and Amphibolidae) that show unusual features such as retaining the operculum characteristic of prosobranch snails, are highly specialized in habitat, and are restricted in distribution. Only the Siphonariidae and Trimusculidae, intertidal limpets, are widely distributed and form a significant part of the littoral fauna. The Ellobiidae are scattered in distribution, occasionally abundant near the high-tide mark, as in *Melampus* in the Atlantic coast saltmarshes of North America. One genus, *Carychium*, is truly terrestrial and common in very wet areas along streams in Eurasia and North America.

The remaining families of the Basommatophora are fresh-water dwellers, usually highly visible, and ecologically significant, both in terms of serving as a food resource for vertebrates and in their browsing on the shallow-water encrusting organisms (algae, fungi, protozoa, and so forth) that grow on rocks and plants. Only a very few species are of medical significance, but these are part of very serious problems. The snails serve as intermediate hosts for trematode parasites of both humans and domestic animals. Some planorbids transmit schistosomiasis in Africa, the West Indies, and South America, while lymnaeids are involved in both sheep and cattle liver fluke life cycles. In North America, some bird schistosomes attempt to penetrate human skin, with only partial success, and then die. The rash produced by their decaying bodies produces the complaint known as swimmer's itch. The annual cost of diseases transmitted by these few snails is very great. Killing all the snails in a body of water has not proved effective, as many of the snails survive in nooks and crannies, and new specimens are carried in by birds and insects from neighboring bodies of water.

Indeed, one of the most important aspects of basommatophoran snails is the extent to which their biology and evolution have been shaped by the fact that most bodies of fresh water are small and isolated, and exist for only a few thousand years. When new ponds or lakes appear, they must be colonized by dispersal across land barriers. Basommatophorans appear in the geological record at the Jurassic-Cretaceous boundary, shortly after the origin of birds. Frequently pond snails are carried on the feet or feathers of birds from one body of water to another. So that one such successful colonization is enough to populate a new body of water, most species are capable of self-fertilization, although cross-fertilization is the normal mode of reproduction. The eggs are laid in gelatinous masses on rocks or plants. Generation time usually is very short, and enormous numbers of snails can build up very quickly during the favorable weather seasons. As a result of this ability to start a colony from one individual and to multiply rapidly, basommatophoran snails show a great deal of minor variation among populations. This makes identification of species very difficult and has led to much confusion as to the proper names for a species and in determining exactly what is a species in these families.

By far the vast majority of fresh-water basommatophorans live in water of less than 12 ft (4 m) depth, and only rarely have species been found living at relatively great depths. These few species must depend upon cutaneous respiration or an air bubble used as a physical gill for oxygen exchange, while those living in shallow water normally come to the surface at regular intervals to breathe.

Human activities, particularly polluting waters with industrial chemicals, excess fertilizers, and silt, have had profound effects upon the fresh-water snail faunas, greatly reducing both the diversity of species and numbers of individuals present.

The major families of fresh-water basommatophorans are readily separated. The South American Chilinidae have peculiarities in the nervous system and incomplete separation of the male and female ducts in the pallial region. The probably related Latiidae from New Zealand and Acroloxidae from Eurasia and the Rocky Mountain area of North America represent specializations toward limpet form. The nearly worldwide Lymnaeidae have completely separated pallial reproductive ducts, generally a high, spiral shell, rather flat tentacles, and several other minor anatomical differences. They are common in temporary ponds, lakes, and sluggish streams. The sinistrally coiled Physidae have peculiar radular teeth and a reduction in size of the jaw. They are common in temporary ponds and other isolated bodies of water in the Northern Hemisphere, and are thriving where introduced into Southern Hemisphere localities.

The Planorbidae and Ancyliade typically are readily separable by the limpet-shaped shell of the latter, but discoveries of intermediate taxa and studies of anatomy have demonstrated that the limpet shape was evolved several times, and some authors propose uniting them into one family. The Planorbidae are the most diverse in size and form of all basommatophoran groups. Both groups live in many types of fresh-water habitats. *See* PULMONATA.

G. Alan Solem

Bibliography. V. Fretter and J. Peake (eds.), *Pulmonates*, vol. 2A, 1978; L. Hyman, *The Invertebrates*, vol. 6: *Mollusca I*, 1967; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Basses

A common name applying to a large group of fishes, mostly in the order Perciformes, generally having the following characteristics: body about 2.5 or 3 times longer than its depth; two dorsal fins, the first composed of stiff spines; pectoral fins on the sides of the body just behind the head; pelvic fins thoracic in position, with one spine and five rays; ctenoid scales; a relatively large mouth; and a predatory lifestyle. Because "bass" is not a taxonomic unit, however, the term cannot be precisely described or defined. In the western North Atlantic, the eastern North Pacific, and freshwaters of Central and North America alone, the word "bass" appears in the vernacular of 53 species and six hybrids in 17 genera and six families. A generally accepted limited taxonomy of the basses follows. (A indicates a North Atlantic range; P, North Pacific; F, fresh waters of Central and North America.) Fishes called "basslet" (primarily family Grammatidae) are not included.

- Order Perciformes
  - Suborder Percoidei
    - Family Moronidae—temperate basses
      - *Morone americana*—white perch (A-F)
      - *Morone chrysops*—white bass (F)
      - *Morone mississippiensis*—yellow bass (F)
      - *Morone saxatilis*—striped bass (A-F)
      - ♂ *Morone americana* × ♀ *M. saxatilis* hybrid—Virginia bass (F)
      - ♀ *Morone americana* × ♂ *M. saxatilis* hybrid—Maryland bass
      - ♀ *Morone chrysops* × ♂ *M. saxatilis* hybrid—sunshine bass
      - ♂ *Morone chrysops* × ♀ *M. saxatilis* hybrid—palmetto bass
      - ♂ *Morone mississippiensis* × ♀ *M. saxatilis* hybrid—paradise bass
    - Family Acropomatidae—lanternbellies
      - *Synagrops bellus*—blackmouth bass (A)
      - *Synagrops spinosus*—keelcheek bass (A)
      - *Synagrops trispinosus*—threespine bass (A)
    - Family Polyprionidae—wreckfishes
      - *Stereolepis gigas*—giant sea bass (P)
    - Family Symphysanodontidae—slopefishes
      - *Symphysanodon berryi*—slope bass (A)
    - Family Serranidae—sea basses and groupers
      - *Anthias nicholii*—yellowfin bass (A)
      - *Anthias tenius*—threadnose bass (A)
      - *Anthias woodsi*—swallowtail bass (A)
      - *Bathyanthias mexacana*—yellowtail bass (A)
      - *Bullisichthys caribbaeus*—pugnose bass (A)
      - *Centropristis fuscula*—twospot sea bass (A)
      - *Centropristis ocyurus*—bank sea bass (A)
      - *Centropristis philadelphica*—rock sea bass (A)
      - *Centropristis striatus*—black sea bass (A)
      - *Dermatolepis dermatolepis*—leather bass (P)
      - *Hemanthias aureorubens*—streamer bass (A)
      - *Hemanthias leptus*—longtail bass (A)
      - *Hemanthias peruanus*—splittail bass (P)
      - *Hemanthias signifier*—hookthroat bass (P)
      - *Paralabrax auroguttatus*—goldspotted sand bass (A)
      - *Paralabrax clathratus*—kelp bass (P)
      - *Paralabrax loro*—parrot sand bass (P)
      - *Paralabrax maculatofasciatus*—spotted sand bass (P)
      - *Paralabrax nebulifer* barred sea bass (P)
      - *Paralabrax atrimanus*—bank bass (A)
      - *Plectranthias garrupellus*—apricot bass (A)
      - *Pronotogrammus eos*—bigeye bass (P)
      - *Pronotogrammus martinicensis*—roughtongue bass (A)
      - *Pronotogrammus multifasciatus*—threadfin bass (P)
      - *Pronotogrammus gregoryi*—reef bass (A)
      - *Pronotogrammus thaumasium*—Pacific reef bass (P)
      - *Serraniculus pumilo*—pygmy sea bass (A)
      - *Serraniculus annularis*—orangeback bass (A)
      - *Serraniculus atrobranchus*—blackear bass (A)
      - *Serraniculus baldwini*—lantern bass (A)
      - *Serraniculus chionaraia* snow bass (A)
      - *Serraniculus flaviventris*—twinspot bass (A)
      - *Serraniculus luciopercanus*—crosshatch bass (A)
      - *Serraniculus notospilus*—saddle bass (A)
      - *Serraniculus tigrinus*—harlequin bass (A)
      - *Serraniculus tortugarum*—chalk bass
    - Family Centrarchidae—sunfishes
      - *Ambloplites ariommus*—shadow bass (F)
      - *Ambloplites cavifrons*—Roanoke bass (F)
      - *Ambloplites constellatus*—Ozark bass (F)
      - *Ambloplites rupestris*—rock bass (F)
      - *Micropterus cataractae*—shoal bass (F)
      - *Micropterus coosae*—redeye bass (F)
      - *Micropterus dolomieu*—smallmouth bass (F)
      - *Micropterus notius*—Suwannee bass (F)
      - *Micropterus punctulatus*—spotted bass (F)
      - *Micropterus salmoides*—largemouth bass (F)
      - *Micropterus treculii*—Guadalupe bass (F)
      - *Lepomis macrochirus* × *Micropterus salmoides* hybrid—blue bass (F)

Acropomatidae. Acropomatidae is a family of about 40 species but only three bear the name bass. Adults of these fishes attain lengths of 11 to 23 cm (4 to 9 in.) and they are bathypelagic, occurring at depths to at least 910 m (3000 ft), in the Eastern and Western Atlantic and elsewhere.

Polyprionidae. Of the nine species in Polyprionidae only one is a bass, the giant sea bass, with a length of 250 cm (8 ft 2 in.) and weighing 256 kg (564 lb). This giant fish is known in the eastern Pacific from coastal Mexico and California, and in the western North Pacific from Japan. Juveniles live in and around kelp beds, whereas adults live near kelp in deeper water over rocky substrates. The giant sea bass is listed by the World Conservation Union as a critically endangered species.

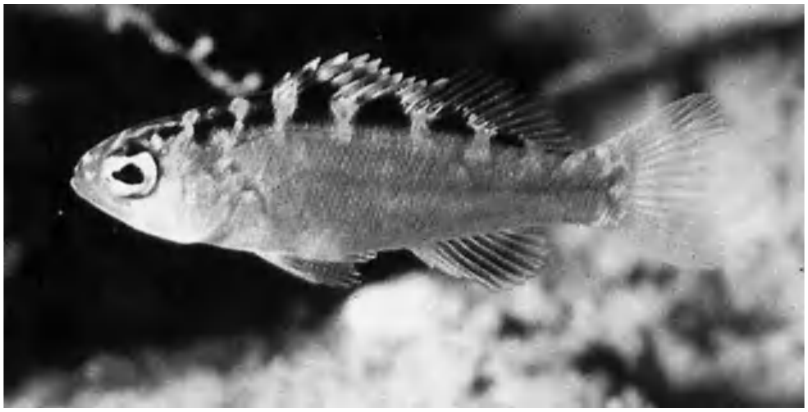
Symphysanodontidae. The slope bass is one of six species in the family Symphysanodontidae. It is a small bass of about 13 cm (5 in.) that inhabits rocky substrates at depth of 220 to 467 m (720–1500 ft) in the Bahamas and off the coasts of Central America and Venezuela.

Centrarchidae. Centrarchidae is a fresh-water family indigenous to North America east of the Continental Divide, except for the Sacramento perch in California. Centrarchids are actually sunfishes; only the species of genera *Ambloplites* and *Micropterus* are called basses. *See* SUNFISHES.

Moronidae. Members of the family Moronidae (temperate basses) are fresh-water or euryhaline (tolerating a wide range of salinities) fishes that spawn in fresh water. The family consists of six species in two genera, *Morone* in North America and *Dicentrarchus* off Europe and North Africa. The family Moronidae has also been, and still is by some authors, placed in the Percichthyidae, a family of fresh-water fishes limited to southern South America and Australia. Temperate basses differ from other basses in the following combination of characters: they have a dorsal fin in two distinct parts, the first with 8–10 spines and the second with one spine and 9–12 soft rays; an opercle with one or two flat spines; seven branchiostegal rays; a lateral line that is complete and extending onto the caudal fin; and a series of black horizontal stripes on the sides of the body, except in *Morone americana*, the white perch.

The white bass and yellow bass are small species, whose adult sizes rarely exceed 1 and 3 kg (2 and 6 lb), respectively. Both are popular food and game fishes, occupying clear lakes and reservoirs, as well as backwaters of large rivers. Most famous of the temperate basses is the striped bass, *Morone saxatilis*. The names *Roccus saxatilis* and *Roccus striatus* appear in the literature but they are synonyms of *Morone saxatilis*.

The native range of the striped bass is coastal waters and fresh-water streams from the St. Lawrence River south to the St. Johns River, Florida, disjunctly in Gulf tributaries from west Florida to Lake Pontchartrain, Louisiana. The species was introduced to the west coast in 1879 and now also occurs from southern California to Washington and possibly Alaska. Adults migrate up rivers in search of suitable spawning habitats. Their fertilized eggs, when water hardened, are about 3.5 mm in diameter and slightly denser than water. The eggs require several days in sunlit water to develop fry. Currents of about 60 cm/s keep the eggs suspended or nearly so. Should the eggs settle on a muddy substrate they will die; therefore, long stretches of clear, free-flowing water and hard substrates over which the eggs tumble are the essentials of a suitable habitat. Striped bass have also been introduced on the Atlantic coast, where they attain weights exceeding 35 kg (78 lb) and are a very popular target of surf fishers.



Chalk bass (*Serranus tortugarum*). (*Photo © John E. Randall*)

Serranidae. Certain species of the large family Serranidae are collectively known as the sea basses. The family is characterized by three flat spines on the opercle; generally ctenoid scales, cycloid in some species; a complete lateral line, not extending onto the caudal fin; a dorsal fin that is usually continuous and may be notched, usually with 10 or fewer spines; three anal fin spines; pelvic fins, each with one spine and five rays; usually seven branchiostegals; and usually 24 vertebrae (see **illustration**). Serranids are hermaphroditic, usually developing first as females and then becoming males; however, some (such as genus *Serranus*) are functional hermaphrodites (that is, individuals are both sexes simultaneously). Serranids vary in adult length from only 3 cm to 300 cm (1 in. to 10 ft) and attain weights of 455 kg (1003 lb). Sea basses occur worldwide in tropical and temperate seas, typically on reefs and objects that provide shelter. Depending on the species, their food may be plankton, crustaceans, or fishes. Many members of the family are important food and game fishes. Worldwide their commercial importance is enormous.

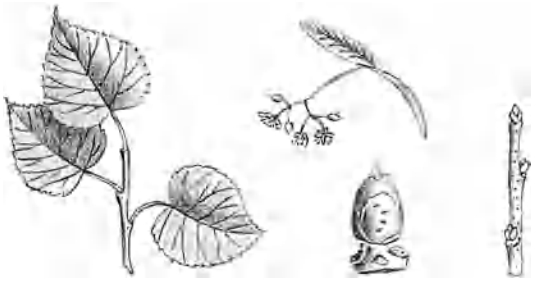
Of 62 genera and 449 species in the family, it is difficult to determine on a worldwide basis the number of serranid "basses." In the geographical area described above, 37 of 138 species of serranids bear the name bass. Some other serranid vernaculars are grouper, basslet, hamlet, soapfish, and sand perch.

Herbert Boschung

Bibliography. G. R. Allen and D. R. Robertson, *Fishes of the Tropical Eastern Pacific*, University of Hawaii Press, Honolulu, 1994; J. Baillie and B. Groombridge (compilers and editors), *1996 IUCN Red List of Threatened Animals*, IUCN, Gland, Switzerland, 1996; W. N. Eschmeyer, E. S. Herald, and H. Hammann, *A Field Guide to Pacific Coast Fishes of North America*, Houghton Mifflin, Boston, 1983; G. D. Johnson, Percoidei: Development and relationships, pp. 464–498, in H. G. Moser et al. (eds.), *Ontogeny and Systematics of Fishes*, Amer. Fisheries Soc. Spec. Publ. 1, Bethesda, MD, 1984; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994; J. S. Nelson et al., *Common and Scientific Names of Fishes from the United States, Canada, and Mexico*, Amer. Fisheries Soc. Spec. Publ. 29, Bethesda, MD, 2004.

Basswood

A member of the linden family (Tiliaceae) in the order Malvales. One species is known as the American linden or basswood (*Tilia americana*). It is a timber tree of the northeastern quarter of the United States and the adjacent area of Canada and grows to a height of 120 ft (36 m). The leaves are heart-shaped, coarsely toothed, long, pointed, and alternate (see **illus.**). *See* MALVALES.



American basswood (*Tilia americana*).

All species of *Tilia* can be recognized by the peculiar winter buds, which have a large outer scale that gives the bud a humped, asymmetrical appearance, and by the small, spherical, nutlike fruits borne in clusters. *Tilia* is also an ornamental tree. *Tilia europea*, on lime tree of Europe, is often cultivated along the streets. The lindens are also important as bee trees. A species of the southern United States, *T. heterophylla*, known as the bee tree linden, has leaves that are white beneath, or brownish with a dense, hairy coat.

There are about 30 species in the temperate regions of the Northern Hemisphere, in North America south to Mexico, but none in the western part. In Asia lindens grow south to central China and southern Japan. The wood of basswood, also known as whitewood, is white and soft and is used for boxes, venetian blinds, millwork, furniture, and woodenware. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

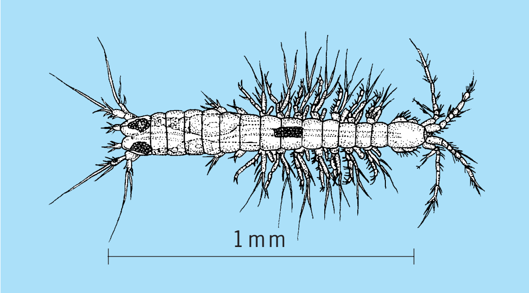
Batales

A small order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Dilleniidae of the class Magnoliopsida (dicotyledons). It consists of two small families and fewer than 20 species, of no economic significance. The plants are trees, shrubs, or subshrubs with simple, entire leaves and small, unisexual flowers. The perianth consists of small, separate or fused sepals with no petals. The stamens are four to numerous; the pistil has (1)2–many carpels, fused to form a plurilocular, superior ovary. The fruit is a dry, dehiscent nutlet or a drupe. The taxonomic disposition of the Batales has long been disputed, and it is often included in the Caryophyllidae. However, the presence of mustard oils as well as certain morphological features suggest that the order is best included near the Capparales in the Dilleniidae. *See* CAPPARALES; DILLENIIDAE; MAGNOLIOPHYTA; PLANT KINGDOM.

T. M. Barkley

Bathynellacea

An order of syncarid crustaceans found in subterranean waters in central Europe and England. The body is elongate and segmented. Organs and limbs are primitive. Each of the thoracic limbs has two epipodites which function as gills. The thorax and abdomen have a similar form (see **illus.**). These animals have no metamorphosis. *See* SYNCARIDA.



***Bathynella natans*, female.**

Hans Jakobi

Batoidea

One of two subdivisions of the subclass Elasmobranchii, consisting of four extant orders, 17 families, and about 456 species, known collectively as rays or batoids. The other subdivision, Selachii, includes the sharks. Some authors group the modern sharks and several families of fossil sharks plus the modern rays in a superorder, Euselachii, supposedly a monophyletic unit. Others group all Recent sharks and rays, each in a single subdivision, Selachii and Batoidei, as presented herein.

Class Chondrichthyes
- Subclass Elasmobranchii
  - Subdivision Selachii
    - Batoidea
      - Order Torpediformes (electric rays)
        - Pristiformes (sawfishes)
        - Rajiformes (skates)
        - Myliobatiformes (stingrays, round rays, butterfly rays, and eagle rays)

See separate articles on each order.

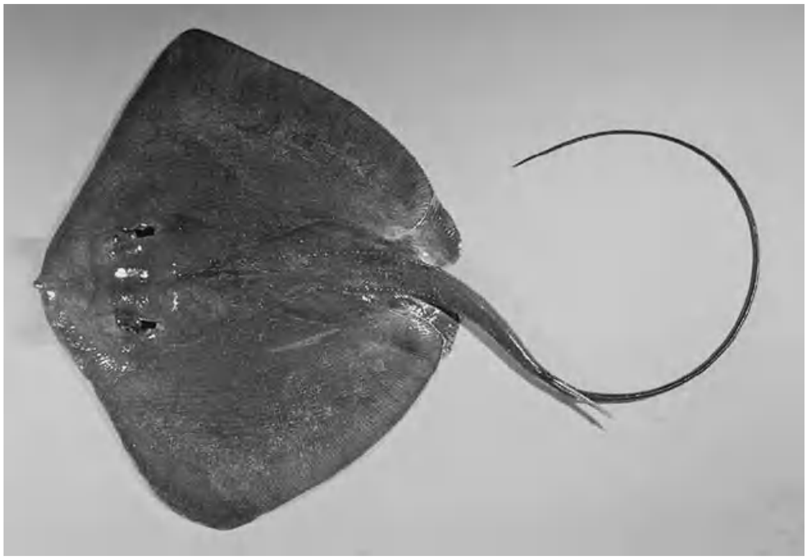
Most batoids are easily recognizable by the flattened head and trunk, with the tail more or less distinct from the body, and eyes and large spiracles on the dorsal surface of the head (see **illustration**). They differ from sharks in having ventral gill slits, the edge of the pectoral fins attached to the side of the head anterior to the gills, the eyeball attached to the upper margin of the orbit, and by lacking an anal fin (most sharks have one). Also, spiracular respiration is much more important to batoids than it is to sharks. A method of reproduction common to all batoids, internal fertilization is accomplished by means of intromittent organs called claspers, which are modified medial lobes of the pelvic fins. Most batoids are ovoviviparous, retaining the eggs in the oviducts until birth, but skates are oviparous, depositing eggs that are protected by horny cases. The early embryos are more sharklike than batoidlike, as the body is slender and the pectoral fins are posterior to the gill openings. Also in the early embryonic stage an anal fin is present as in sharks, but is soon lost. By full term the pectoral fins have migrated forward as well as rearward to form the characteristic disk. Batoids vary in size from less than 0.3 m (1 ft) in length as adults to a width of over 6 m (20 ft) in the giant mantas.

By wiggling their disk, the bottom-dwelling batoids will throw sand over their back, leaving only the eyes and spiracles exposed. All are carnivorous, feeding on a wide variety of marine worms, mollusks, crustaceans, and other invertebrates, as well as small fishes. They are represented in tropical to arctic latitudes of the Atlantic, Pacific, and Indian oceans of both hemispheres. Some are limited to fresh water (such as the subfamily Potamotrygoninae of South America) and a few marine species, especially stingrays and sawfishes, penetrate estuaries of rivers or even live for extended periods far upstream, but the vast majority are strictly marine species. Electric rays and skates may live at depths of 3000 and 9000 feet, respectively. However, members of both groups are more numerous in kind and numbers at depths between 120 to 600 feet. Typically the myliobatids are abundant in relatively shallow waters around islands. Although batoids are abundant in many seas, they are of minor commercial importance.

Where known, the body of extinct skates and rays appears to have resembled that of modern types in its depressed shape, expanded pectoral fins, and slender tail. In general, however, remains consist of teeth only. Many Paleozoic sharks appear to have been bottom dwelling, feeding on shelled invertebrates; these sharks became extinct, however, and the first, relatively rare, batoids are found in the Late Jurassic. With the exception of the eagle rays and torpedoes, all major groups now extant were present in the Upper Cretaceous and had assumed the prominent position that persist today. *See* ELASMOBRANCHII; SELACHII.

Reeve M. Bailey; Herbert Boschung

Bibliography. H. B. Bigelow and W. C. Schroeder, Sawfishes, guitarfishes, skates and rays, in *Fishes of the Western North Atlantic*, Mem. Sears Found. Mar. Res. Mem. 1(pt. 2), pp. 1–514, 1953; L. J. V. Compagno, Interrelationships of living elasmobranchs, pp. 15–61, in P. H. Greenwood et al. (eds.), *Interrelationships of Fishes*, Zool. J. Linn. Soc. 53, Suppl. 1, Academic Press, New York, 1973; L. J. V. Compagno, Checklist of living elasmobranchs, pp. 471–498, in W. C. Hamlett (ed.), *Sharks, Skates and Rays: The Biology of Elasmobranch Fishes*, John Hopkins University Press, 1999; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 2006.



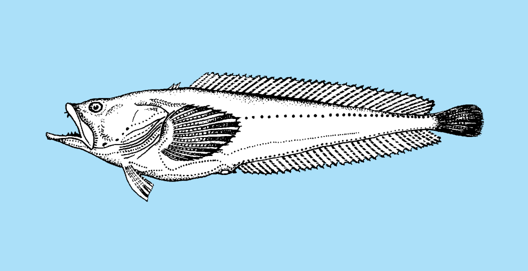
Brown stingray (*Dasyatis latus*). (*Photo © John E. Randall*)

Batrachoidiformes

The toadfishes, an order of actinopterygian (ray-finned) fishes consisting of a single family, Batrachoididae. Batrachoididae comprises three subfamilies, 22 genera, and 73 species. Also known as the Haplodoci, the toadfishes are most closely related to the Lophiiformes, and the two groups collectively are known as the Pediculati. These are rather robust fishes, tapering posteriorly, with a broad flattened head, often with barbels or a fleshy flap; undivided posttemporal bone; dorsal, skyward-looking eyes; a large mouth and powerful jaws equipped with blunt teeth (some species have canine teeth); only three gill arches and reduced gill openings; pelvic fins in the throat area (jugular), with one spine and two or three soft rays; fanlike pectoral fins; a short, spinous dorsal fin consisting of two or three strong sharp spines; long soft dorsal and anal fins; a usually scaleless body covered with mucus; one to several lateral lines on head and body; and mostly drab brown color variously marked with dark saddles and blotches.

Toadfishes occur in the Atlantic, Pacific, and Indian oceans, primarily as benthic coastal inhabitants; however, a few species enter brackish water, and several are limited to freshwater. These are sluggish fishes that ambush mollusks and crustacean prey, and they are known to bite when handled.

Members of the subfamily Batrachoidinae occur worldwide but have the greatest diversity in the New World, in the Western Atlantic and Eastern Pacific. They have three solid dorsal spines, an opercular spine, and no venom glands; a subopercular spine is present; canine teeth are absent; the body is scaleless or not; and there are no photophores. Most species occur in mud or sand substrates, but one



Atlantic midshipman (*Porichthys porosissimus*), about 8 in. (20 cm) long.

genus (*Sanopus*) inhabits coral reefs and is brightly colored. The Porichthyinae are New World species that have two solid dorsal spines, solid opercular spine, and no venom glands; a subopercular spine is present as are canine teeth; the body is scaleless with several lateral lines; and there are photophores in all but two species (*Aphos*). Members of the genus *Porichthys* (=*Nautopaedium*) have lines of bright yellow photophores (buttons) in the lower sides and belly, hence the common name midshipmen (see **illustration**). It is very unusual for a shallow-water fish to possess photophores. The Thalassophryninae are venomous toadfishes and also New World in distribution. They have a scaleless body; no subopercular spine; no photophores; no canine teeth; the lateral line is single or absent. The two dorsal spines and opercular spine are hollow, hypodermic-needle-like weapons that deliver a venom capable of producing excruciating wounds. *See* ACTINOPTERYGII; BIOLUMINESCENCE; TELEOSTEI. Herbert Boschung

Bibliography. B. B. Collette, A review of the venomous toadfishes, subfamily Thalassophryninae, *Copeia*, 1966(4):846–864, 1966; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994.

Battery

An electrochemical device that stores chemical energy which can be converted into electrical energy, thereby providing a direct-current voltage source. Although the term "battery" is properly applied to a group of two or more electrochemical cells connected together electrically, both single-cell and multicell devices are called battery. *See* ELECTROCHEMISTRY; ELECTROMOTIVE FORCE (CELLS).

Types. The two general types are the primary battery and the secondary battery. The primary battery delivers current as the result of a chemical reaction that is not efficiently reversible. Practically, this makes the primary battery nonrechargeable. Only one intermittent or continuous discharge can be obtained before the chemicals placed in it during manufacture are consumed. Then the discharged primary battery must be replaced. The secondary or storage battery is rechargeable because it delivers current as the result of a chemical reaction that is easily reversible. When a charging current flows through its terminals in the direction opposite to the current flow during discharge, the active materials in the secondary battery return to approximately their original charged condition.

Components. The cell is the basic electrochemical unit. It has three essential parts: (1) a negative electrode (the anode) and (2) a positive electrode (the cathode) that are in contact with (3) an electrolyte solution. The electrodes are metal rods, sheets, or plates that are used to receive electrical energy (in secondary cells), store electrical energy chemically, and deliver electrical energy as the result of the reactions that occur at the electrode-solution surfaces. Solid polymer or plastic active materials have been developed that can serve as the cathode in rechargeable batteries. The electrolyte is a chemical compound (salt, acid, or base) that when dissolved in a solvent forms a solution that becomes an ionic conductor of electricity, but essentially insulating toward electrons—properties that are prerequisites for any electrolyte. In the cell or battery, this electrolyte solution is the conducting medium in which the flow of electric current between electrodes takes place by the migration of ions. When water is the solvent, an aqueous solution is formed. Some cells have a nonaqueous electrolyte, for example, when alcohol is used as the solvent. Other cells have a solid electrolyte that when used with solid electrodes can form a leak-free solid-state cell or battery.

During charging of a secondary cell, the negative electrode becomes the cathode and the positive electrode becomes the anode. However, electrode designation as positive or negative is unaffected by the operating mode of the cell or battery. Two or more cells internally connected together electrically, in series or parallel, form a battery of a given voltage. Typical are the rectangular 9-V primary battery, which has six flat 1.5-V zinc-carbon or alkaline "dry" cells connected in series, and the 12-V automotive or secondary battery, which has six 2.1-V lead-acid "wet" cells connected in series.

Size. Both primary and secondary cells are manufactured in many sizes, shapes, and terminal arrangements, from the miniature coin- or button-shaped battery (which has a diameter greater than its height) and the small cylindrical penlight battery to the large submarine battery, where a single rectangular cell has weighed 1 ton (0.9 metric ton). For optimum performance, the battery must be constructed for its particular application consistent with cost, weight, space, and operational requirements. Automotive and aircraft batteries are secondary batteries that have relatively thin positive and negative plates with thin, porous, envelope separators to conserve space and weight and to provide high rates of current discharge at low temperatures. Standby batteries are secondary batteries that use thick plates and thick separators to provide long life. Solid-state batteries can be constructed with unusual features and in irregular sizes and shapes. Size and weight reductions in all types of batteries continue to be made through use of new materials and methods of construction.

Selection and applications. Batteries are probably the most reliable source of power known. Most critical electrical circuits are protected in some manner by battery power. Since a battery has no moving parts, tests, calculations, or comparisons are made to predict the conditions of the cells in some batteries. Growing battery usage reflects the increased demand for portable computers; mobile voice, data, and video communications; and new or redesigned/repowered products in the consumer, industrial, and transportation sectors. Further growth may result from significant increases in dc system operating voltage, such as from 12 V to 24 V or 48 V, which can provide much higher power generally with less weight, greatly broadening the range of potential battery applications.

For most applications, the basic choice in selection is whether to use either a primary (nonrechargeable) or a secondary (rechargeable) cell or battery. Electrical characteristics affecting selection include maximum and minimum voltage, current drain, and pulse current (if any), its duration and frequency of occurrence. Other factors such as performance in the specific application, operating environment, and final packaging of the cell or battery also must be considered.

Primary battery usage. Primary batteries are used as a source of dc power where the following requirements are important:

1. Electrical charging equipment or power is not readily available.

2. Convenience is of major importance, such as in a hand or pocket flashlight.

3. Standby power is desirable without cell deterioration during periods of nonuse for days or years. Reserve-electrolyte designs may be necessary, as in torpedo, guided missile, and some emergency light and power batteries.

4. The cost of a discharge is not of primary importance.

Secondary battery usage. Secondary batteries are used as a source of dc power where the following requirements are important:

1. The battery is the primary source of power and numerous discharge-recharge cycles are required, as in wheelchairs and golf carts, industrial hand and forklift trucks, electric cars and trucks, and boats and submarines.

2. The battery is used to supply large, short-time (or relatively small, longer-time), repetitive power requirements, as in automotive and aircraft batteries which provide power for starting internal combustion engines.

3. Standby power is required and the battery is continuously connected to a voltage-controlled dc circuit. The battery is said to "float" by drawing from the dc circuit only sufficient current to compensate automatically for the battery's own internal self-discharge. Computers and communications networks and emergency light and power batteries are in this category.

4. Long periods of low-current-rate discharge followed subsequently by recharge are required, as in marine buoys and lighthouses, and instrumentation for monitoring conditions such as earthquakes and other seismic disturbances.

5. The very large capacitance is beneficial to the circuit, as in telephone exchanges.

Ratings. Two key ratings of a cell are its voltage and ampere-hour (Ah) capacity. The voltage is determined by the chemical system created by the active materials used for the negative and positive electrodes (anode and cathode). The ampere-hour capacity is determined by the amount of the active materials contained in the cell. The product of these terms is the energy output or watthour capacity of the battery. In actual practice, only one-third to one-half of the theoretical capacity may be available. Battery performance varies with temperature, current drain, cutoff voltage, operating schedule, and storage conditions prior to use, as well as the particular design.

Many primary batteries are rated by average service capacity in milliampere-hours (mAh). This is the number of hours of discharge that can be obtained when discharging at a specified temperature through a specified fixed resistance to a specified final or cutoff voltage, which is either at the point of rapid voltage drop or at minimum usable voltage.

Secondary batteries, such as automotive batteries, have been rated by ampere-hour capacity. Typically, this is the amount of current that the battery can deliver for 20 h without the temperature-corrected cell voltages dropping below 1.75 V per cell. A battery capable of giving 2.5 A for 20 h is rated at 50 ampere-hours at the 20-h rate. This same battery may provide an engine-cranking current of 150 A for only 8 min at 80°F (27°C) or for 4 min at 0°F (−18°C), giving a service of only 20 and 10 ampere-hours, respectively. By multiplying ampere-hours by average voltage during discharge, a more practical watt-hour rating is obtained. Automotive batteries also are rated by reserve capacity (RC) and cold-cranking amps (CCA). Reserve capacity is the length of time that a fully charged battery at 80°F (27°C) can deliver 25 amperes before the voltage falls to 10.5 V. A typical rating is 125 min, which is the length of time the battery could carry a minimum electrical load after failure of the vehicle's charging system. Cold-cranking amps is a measure of the ability of a battery to crank an engine when the battery is cold. It is measured by the number of amperes that a 12 V battery can deliver for 30 s when it is at 0°F (−18°C) without the battery voltage falling below 7.2 V. A typical CCA rating for a battery with a reserve capacity of 125 min is 430 amperes. The CCA rating for most automobile batteries is between 300 and 600 amperes.

Life. The life expectancy of a cell or battery depends on its design and materials, as well as its application and operating conditions. Life expectancy is measured by shelf life and service life. Shelf life is the expected time that elapses before a stored battery becomes inoperative due to age or deterioration, or unusable due to its own internal self-discharge. Service life is the expected length of time or number of discharge-charge cycles through which a battery remains capable of delivering a specified percentage of its capacity after it has been put into service. This can vary from the one shot or single discharge

TABLE 1. Comparison of principal chemical systems of primary batteries

Principal chemical systems	General characteristics under optimum operating conditions	Configurations and rated capacity*	Nominal cell voltage, V
Zinc-carbon (Zn/MnO_2)	Popular, common, low-cost battery; moderate shelf life; sloping discharge characteristic	Cylindrical cells, to about 40 Ah	1.5
Zinc-alkaline-manganese dioxide (Zn/MnO_2)	Popular, general-purpose premium battery; good low-temperature and high-rate performance; good shelf life; sloping discharge characteristic	Cylindrical cells to 20 Ah; also available in small button cells	1.5
Zinc-mercuric oxide (Zn/HgO)	High volumetric capacity; good shelf life; flat discharge characteristic	Button and cylindrical cells from 40 mAh to 13 Ah; also available in special larger sizes	1.35 or 1.40
Zinc-silver oxide (Zn/Ag_2O)	High gravimetric capacity; good shelf life; flat discharge characteristic	Button cells to 180 mAh for specialized applications	
Zinc-air (Zn/O_2)	Highest energy density on continuous discharge; excellent shelf life (unactivated); limited rate capability and shelf life when activated; flat discharge characteristic	Button cells to 1150 mAh; larger cells to 6.5 Ah	1.4
Lithium-manganese dioxide (Li/MnO_2)	High energy density; good rate capability and low-temperature performance; excellent shelf life; relatively flat discharge characteristic	Coin cells to 500 mAh; spiral-wound cylindrical cells to 1.25 Ah; bobbin-type cells to 2.5 Ah	3.0
Lithium-sulfur dioxide (Li/SO_2)	High energy density; excellent high-rate and low-temperature performance; pressurized, hermetically sealed cell; excellent shelf life; flat discharge characteristic	Cylindrical cells from 700 mAh to 19 Ah	3.0

*Button cells are rated at 500–1000 h at 70°F (21°C), cylindrical cells at 50–100 h. The cutoff voltage is approximately 60% of the nominal voltage. Typical capacities can be higher.

obtainable from primary cells to 10,000 or more discharge-charge cycles obtainable from some secondary batteries. Automotive batteries may last as little as 18 months in hot climates and 10–12 years in cold climates, but typically they have an average life of 3 years and may last for 6000 cycles. Industrial batteries have a 10–20-year service life. Standby sizes may be expected to float across the dc bus 8–30 years. Generally the most costly, largest, and heaviest cells have the longest service life.

Many batteries experience an abrupt loss of voltage without warning when the active materials are depleted. However, in some batteries, open-circuit voltage serves as a state-of-charge indicator, decreasing slightly but continuously with loss of capacity. Many electronic devices now include a battery that will last for the life of the product. This eliminates the need for the consumer or user to replace the battery and risk causing damage to the device or the battery by inadvertently shorting the terminals, reversing the polarity, or installing a similar-size battery having the wrong chemistry.

To obtain the maximum life and ensure reliability of batteries, the manufacturer's recommendations for storage and maintenance must be followed. The stated shelf life and temperature of primary cells must not be exceeded. For dry reserve-electrolyte primary cells and secondary cells of the dry construction with charged plates, the cell or battery container must be protected against moisture, and storage must be within specified temperature limits. Wet, charged secondary batteries may require periodic charging and water addition, depending upon the construction.

Primary Batteries

A primary cell or battery is not intended to be recharged and is discarded when it has delivered all its electrical energy (**Fig. 1**). Several kinds of primary cell are widely used, particularly in portable devices and equipment, providing freedom from the dependence on alternating-current line power. They are convenient, lightweight, and usually relatively inexpensive sources of electrical energy that provide high energy density (long service life) at low-to-moderate or intermittent discharge rates, good shelf life, and ease of use while requiring little or no maintenance.

Primary cells are classified by their electrolyte, which may be described as aqueous, nonaqueous, aprotic, or solid. In most primary cells the electrolyte is immobilized by a gelling agent or mixed as a paste,

	Energy density †					Shelf life, % loss per year at 20° C (68° F)	Typical operating temperature,† °F (°C)
	Wh/lb	Wh/kg	Wh/in.^{3}	Wh/L	Typical applications		
Cylindrical:	34	75	2.3	140	General purpose for lighting, radio, novelties	10	23 to 113° (−5 to 45°)
Button: Cylindrical:	25 60	55 130	2.4 5.8	145 350	General purpose, such as lighting, photographic equipment, toys, radio and TV, tape recorders, watches and instruments; popular for high-drain applications	3	−4° to 130° (−20° to 54°)
Button: Cylindrical:	50 55	110 120	7.3 7.4	445 450	Hearing aids, medical and photographic equipment, communication devices, papers, detection and emergency equipment, and other applications requiring steady output voltage	4	15° to 130° (−9° to 54°)
Button:	60	130	8.2	500	Hearing aids, watches, photographic equipment, and special electronics requiring small, high-capacity batteries	6	−4° to 130° (−20° to 54°)
Button:	140	310	19.0	1150	Hearing aids, medical monitoring instruments, pagers, communication devices, emergency lighting, and other relatively short-term applications	2 (unactivated)	32° to 122° (0° to 50°)
Coin: Cylindrical: spiral-wound bobbin	80 105 135	175 230 300	8.3 8.3 10.1	505 505 61.5	General purpose in photographic and electronic applications requiring small, high-capacity batteries, such as watches, calculators, computers, clock/calendar, memory backup, photoflash, motor-driven cameras, and instruments	0.5−1	−4° to 140° (−20° to 60°) −4° to 140° (−20° to 60°) −4° to 160° (−20° to 71°)
Cylindrical:	275	125	440	7.1	Military and special industrial applications (such as process control equipment), requiring high-capacity, high rate, and extreme temperature operation	0.5−1	−40° to 160° (−40° to 71°)

†Batteries can perform over a wider temperature range, under certain discharge conditions.

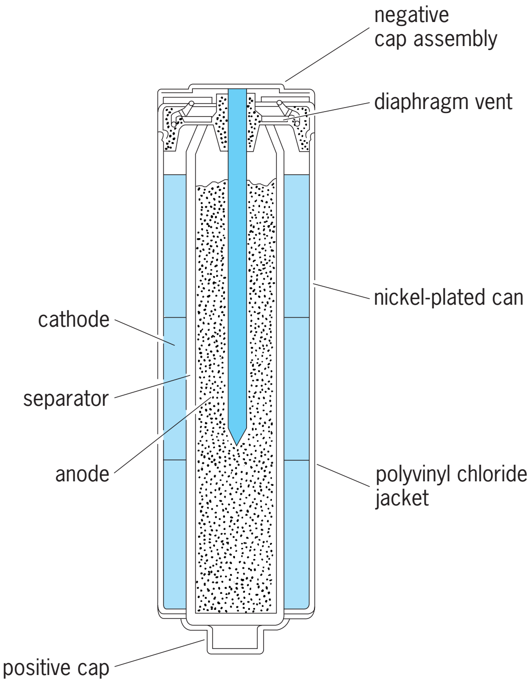


Fig. 1. Diagram of a zinc-alkaline-manganese dioxide cylindrical cell.

with the term “dry cell” commonly applied to the zinc-carbon Leclanche cell and sometimes to other types. An aqueous electrolyte or electrolyte system is used in zinc-carbon, magnesium, alkaline-manganese dioxide, mercuric oxide, silver oxide, and zinc-air cells. Nonaqueous electrolyte systems are used in lithium cells and batteries. *See* ELECTROLYTE.

Typical characteristics of different types of primary batteries and their applications are summarized in **Table 1.** Performance characteristics of primary batteries at various temperatures are shown in **Fig. 2.**

Zinc-carbon cells. The zinc-carbon or LeClanche dry cell was invented in 1866 and continues to be popular. It is made in sizes of varying diameter and height, and batteries are available in voltages ranging from 1.5 to 510 V. Common cell construction uses the cylindrical zinc container or can as the negative electrode and manganese dioxide (mixed with carbon black to increase conductivity and retain moisture) as the positive active material, with the electrical connection made through a center carbon electrode. The slightly acidic electrolyte is an aqueous solution of ammonium chloride and may also contain zinc chloride, immobilized in a paste or the paper separator. The electrochemical reaction between the cathode, anode, and electrolyte in a

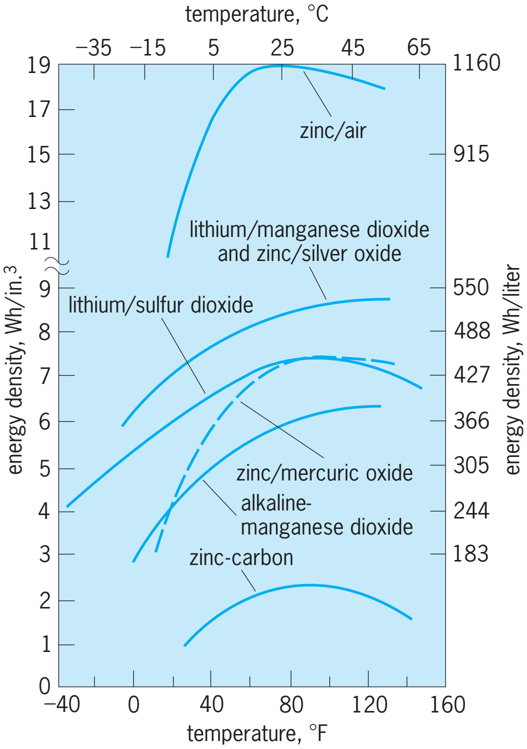


Fig. 2. Performance characteristics of primary batteries at various temperatures.

zinc-carbon LeClanche cell is shown in reaction (1).

$$2MnO_2 + 2NH_4Cl + Zn \rightarrow ZnCl_2 \cdot 2NH_3 + Mn_2O_3 + H_2O \qquad (1)$$

Typical open-circuit voltage of a fresh LeClanche cell is over 1.55 V. The closed-circuit voltage gradually declines as a function of the depth of discharge.

Significant improvements in capacity and shelf life of zinc-carbon batteries have been made through the use of new cell designs, such as the paper-lined cell, and new materials. Upgrading to the use of purer beneficiated manganese dioxide and an electrolyte that consists mainly of zinc chloride and water has created the heavy-duty or zinc chloride cell. Its construction is similar to the zinc-carbon cell, but having an anode of high-purity zinc alloy and typically a higher proportion of carbon to manganese dioxide and a greater volume of slightly more acidic electrolyte. The electrochemical reaction in a zinc chloride cell is shown in reaction (2). Open-circuit

$$8MnO_2 + 4Zn + ZnCl_2 + 9H_2O \rightarrow 8MnOOH + ZnCl_2 \cdot 4ZnO \cdot 5H_2O \qquad (2)$$

voltage of a fresh zinc chloride cell typically is over 1.60 V. As in the LeClanche cell, closed-circuit voltage gradually declines as a function of the depth of discharge.

Compared to zinc-carbon batteries, zinc chloride batteries have improved high-rate and low-temperature performance and are available in voltages ranging from 1.5 to 12 V. A common design is the flat cell, used in multicell batteries such as the 9-V battery, which offers better volume utilization and, in some designs, better high-rate performance. Characteristics of standard zinc-carbon and zinc chloride cells are given in **Table 2.**

Magnesium cells. The magnesium-carbon primary cell is basically a cylindrical zinc-carbon cell in which the can or container is made of magnesium or its alloy instead of zinc. Magnesium has a greater electrochemical potential than zinc, providing an open-circuit voltage of about 2.0 V. Developed for military use in radios and other equipment, the magnesium

TABLE 2. Characteristics of 1.5-V zinc-carbon unit cells

Cell designation*				Maximum dimensions, mm (in.)		Capacity: 2 h/day			
						Standard cell		Zinc chloride cell	
ANSI	IEC	NEDA	Weight, g (oz)	Diameter	Height	Starting drain, mA	Service hours	Starting drain, mA	Service hours†
N	R-1	910	6.2 (0.22)	12 (0.47)	30.2 (1.19)	1.5	320		
						7.5	60		
						15	20		
AAA	R-03	24	8.5 (0.30)	10.5 (0.41)	44.5 (1.75)	2	350		
						10	54		
						20	21		
AA	R-6	15	17 (0.60)	14.5 (0.57)	50.9 (1.99)	3	450		
						15	80	187.5	5.8
						30	32	375	1.9 (cont.)
C	R-14	14	40 (1.4)	26.2 (1.03)	50.0 (1.97)	5	520	187	17
						25	115	375	4.6 (cont.)
						50	53		
D	R-20	13	94 (3.3)	34.1 (1.34)	61.5 (2.42)	10	525	9	1100
						50	125	187	42
						100	57	667	5.3 (cont.)
F	R-25	60	160 (5.6)	34.1 (1.34)	91.4 (3.60)	15	630	187	10
						75	135	375	36
						150	60	667	18

*ANSI = American National Standards Institute: IEC = International Electrotechnical Commission: NEDA = National Electrical Distributors Association.
†cont. = continuous.

cell has twice the capacity or service life of a zinc-carbon cell of equivalent size, and longer shelf life, particularly at elevated temperatures. However, when the load is applied to a magnesium cell, a characteristic voltage drop occurs briefly before the cell recovers to its usable voltage. This may make the magnesium cell unacceptable for use in some applications. The magnesium cell has a vent for the escape of hydrogen gas, which forms during discharge.

Alkaline-manganese dioxide cells. These cells became the most popular battery during the 1980s. Available in voltages ranging from 1.5 to 12 V and in a variety of sizes and shapes, they have higher energy density, better low-temperature and high-rate performance, and longer shelf life than a zinc-carbon battery. Alkaline and zinc-carbon cells are chemically similar, both using zinc for the anode and manganese dioxide for the cathode active materials. However, the cells differ significantly in construction, electrolyte, and formulation of the active materials.

The container for the cylindrical alkaline cell is a steel can that does not participate in the electrochemical reaction. A gelled mixture of zinc powder and electrolyte is used for the anode. The highly alkaline electrolyte is an aqueous solution of potassium hydroxide, which is more conductive than the salt of the zinc-carbon cell. The cathode is a highly compacted, conductive mixture of high-purity manganese dioxide and graphite or carbon. To prevent contact between the anode and cathode, which would result in a very active chemical reaction, a barrier or separator is placed between the two. A typical separator is paper or fabric soaked with electrolyte. The separator prevents solid-particle migration, while the electrolyte promotes ionic conductivity. The electrochemical reaction in an alkaline-manganese dioxide cell is shown in reaction (3). Open-circuit voltage of a fresh cylindrical alkaline cell typically is 1.58 V. Closed-circuit voltage gradually declines as a function of the depth of discharge.

$$2Zn + 3MnO_2 \rightarrow 2ZnO + Mn_3O_4 \qquad (3)$$

Modifying the internal design of zinc-alkaline cells has resulted in cells that have an average of 20–50% higher amperehour capacity or service life. This newer cell design provides a significantly greater volume for the active materials, the factor most responsible for improved performance. The modified cell is manufactured with a nickel-plated steel can and an outside poly(vinyl chloride) jacket (Fig. 1). A diaphragm vent is incorporated into the assembly in case of pressure buildup. Alkaline-manganese dioxide cells also are made in miniature sizes with flat, circular pellet-type cathodes and homogeneous gelled anodes.

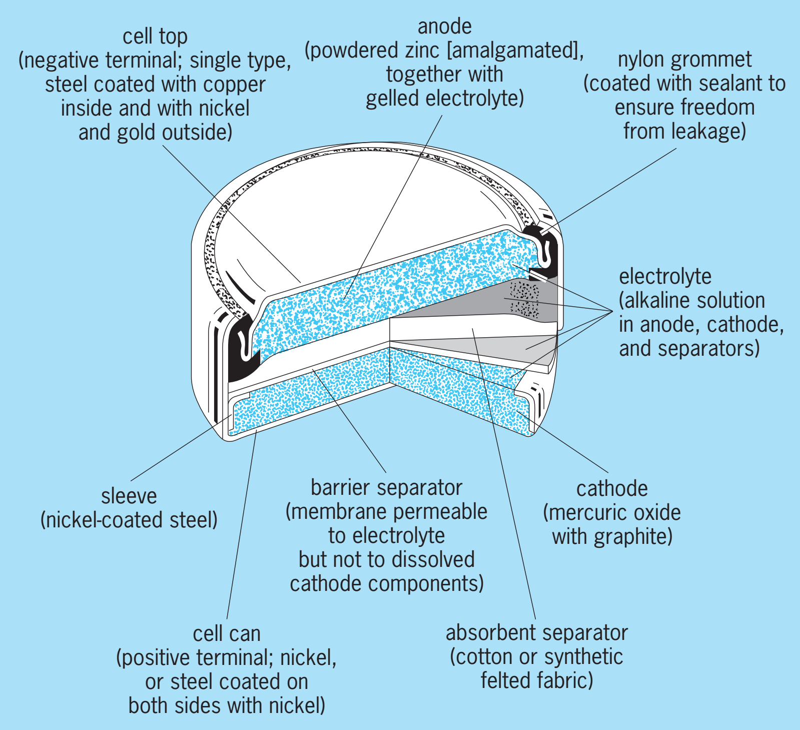

cell top
(negative terminal; single type,
steel coated with copper
inside and with nickel
and gold outside)
anode
(powdered zinc [amalgamated],
together with
gelled electrolyte)
nylon grommet
(coated with sealant to
ensure freedom
from leakage)
electrolyte
(alkaline solution
in anode, cathode,
and separators)
sleeve
(nickel-coated steel)
barrier separator
(membrane permeable
to electrolyte
but not to dissolved
cathode components)
cathode
(mercuric oxide
with graphite)
cell can
(positive terminal; nickel,
or steel coated on
both sides with nickel)
absorbent separator
(cotton or synthetic
felted fabric)

Fig. 3. Diagram showing the design of a mercury button cell.

Mercuric oxide cells. The zinc-mercuric oxide cell (**Fig. 3**) has high capacity per unit charge, relatively constant output voltage during discharge, and good storage qualities. It is constructed in a sealed but vented structure, with the active materials balanced to prevent the formation of hydrogen when discharged. Three basic structures are used: the wound anode, the flat pressed-powder anode, and the cylindrical pressed-powder electrode. Typically, the anode is zinc, the cathode is mercuric oxide, and the electrolyte is alkaline. The electrochemical reaction in a mercuric oxide cell is shown in reaction (4).

$$Zn + 2MnO_2 \rightarrow ZnO + Mn_2O_3 \qquad (4)$$

Open-circuit voltage of a mercuric oxide cell is 1.35 V, with some batteries delivering up to 97.2 V.

Mercuric oxide batteries have been made in various sizes and configurations, from the miniature button 16 mAh to the large 14 Ah cylindrical cell, and are suited for use at both normal and moderately high temperatures. However, the low-temperature performance is poor, particularly under heavy discharge loads. Use at temperatures below 32°F (0°C) is generally not recommended. Mercuric oxide batteries have been used for hearing aids, cameras, watches, calculators, implanted heart pacemakers and other medical applications, emergency radios, and military equipment. Today concerns about health hazards and environmental regulations related to disposal generally prevent manufacture and sale of mercuric oxide batteries. They are replaced with newer, less expensive designs such as zinc-air that do not contain mercury or its compounds.

The cadmium-mercuric oxide battery is similar to the zinc-mercuric oxide battery. The substitution of cadmium for the zinc anode lowers the cell voltage but offers a very stable system, with a shelf life of up to 10 years, and improved performance at low temperatures. Its watthour capacity, because of the lower voltage, is about 60% of the zinc-mercuric oxide battery. However, health concerns and environmental regulations related to disposal also limit the availability and usage of batteries containing cadmium.

Silver oxide cells. The zinc-silver oxide primary cell is similar to the zinc-mercuric oxide cell, but uses silver oxide in place of mercuric oxide (Fig. 3). This results in a higher cell voltage and energy. In the small button-cell configuration, this is a significant advantage for use in hearing aids, photographic applications, watches, and calculators. The silver oxide battery has a higher voltage than a mercury battery; a flatter discharge curve than an alkaline battery; good resistance to acceleration, shock, and vibration; an essentially constant and low internal resistance; and good performance characteristics at temperature extremes.

Silver oxide batteries generally have a flat circular cathode and a homogeneous gelled anode. The cathode is a mixture of silver oxide with a low percentage of manganese dioxide and graphite, and the anode is a mixture of amalgamated zinc powder. A highly alkaline electrolyte of either sodium hydroxide or potassium hydroxide is used, although potassium hydroxide makes the battery more difficult to seal. The separator is a material that prevents migration of any solid particles in the battery.

The chemical reaction in a silver oxide cell is shown in reaction (6). Open-circuit voltage of a sil-

$$\mathrm{Zn} + \mathrm{Ag_2O} \rightarrow \mathrm{ZnO} + 2\mathrm{Ag} \qquad (6)$$

ver oxide cell is 1.6 V, with batteries available having a nominal voltage of 6 V.

Zinc-air cells. Zinc-air cells (**Fig. 4**) have the highest energy density of the commercially available primary cells. They use atmospheric oxygen for the active cathode material, so there is no need to include the cathode material in the cell. This allows the cathode electrode to be very thin. The remaining space can be used for increasing the amount of zinc, which is the active anode material, resulting in a higher cell capacity (Fig. 4*a*).

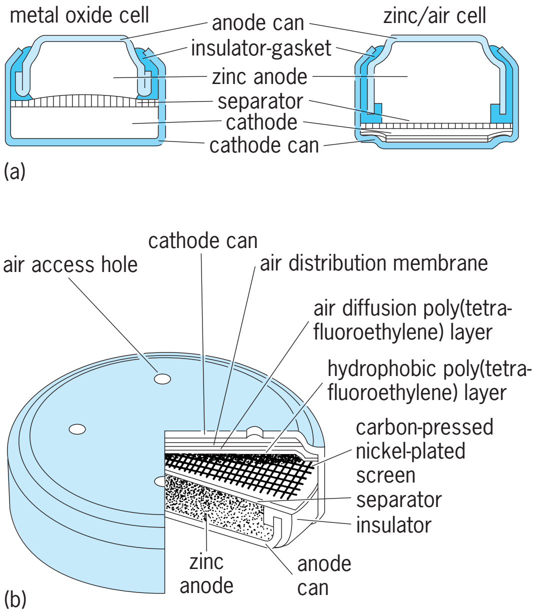


Fig. 4. Zinc-air cell. (*a*) Cross section compared with metal oxide cell. (*b*) Major components.

The zinc-air cell is usually constructed in a button configuration (Fig. 4*b*). The cell consists of two cans, isolated from each other by an annular insulator. One can contains the zinc anode and the other the air or oxygen cathode. The anode can is fabricated from a triclad metal; the external material is nickel for good electrical conductivity, the middle layer is stainless steel to provide strength, and the inner surface is copper, which is compatible with the cell components. The cathode can is fabricated of nickel-plated steel and contains holes that allow air to enter the cell. A layer of permeable polytetrafluoroethylene (Teflon) serves as a separator to assure the proper distribution of air and limit the entrance or exit of moisture. The anode is an amalgamated gelled zinc powder and electrolyte. Although the cathode uses atmospheric oxygen as the active material, a mixture of carbon, polytetrafluoroethylene, and manganese dioxide is impressed on a nickel-plated screen. The carbon and manganese dioxide serve as catalysts for the oxygen reaction. The electrolyte is an aqueous solution of potassium hydroxide with a small amount of zinc oxide.

The air holes of the zinc-air cell are sealed until the cell is used, to inhibit the entrance of air. In this condition, the cells can retain more than 95% of their rated capacity after 1 year of storage at room temperature. The cells are activated by removing the seal which permits the flow of air into the cell. The chemical reaction in a zinc-air cell is shown in reaction (7).

$$2\mathrm{Zn} + \mathrm{O_2} \rightarrow 2\mathrm{ZnO} \qquad (7)$$

The open-circuit voltage of a zinc-air cell is about 1.4 V, providing a flat discharge with the operating voltage between 1.35 and 1.1 V, depending on the discharge conditions. The cell is capable of low to moderately high discharge rates, and is used in applications requiring a relatively short operating time before replacement.

Zinc-air cells are manufactured in sizes from 50 to 6500 mAh. Multicell batteries are available in a wide range of voltage and capacity ratings, including a nominal 8.4-V zinc-air battery that is interchangeable in some applications with the zinc chloride and alkaline 9-V battery.

Lithium cells. Lithium is a silvery-white element that is the lightest metal, having a density only about half that of water, with which the alkaline lithium is highly reactive. Lithium has good conductivity, and its standard potential and electrochemical equivalence are higher than those of any other metal. It is widely used in various forms as battery anode material (negative electrode), especially in coin-size primary cells. There are several different types of lithium cells, primary as well as reserve and secondary, similar to the variety of cells using a zinc anode. Each type of lithium cell differs in the cathode material for the positive electrode, electrolyte, and cell chemistry as well as in physical, mechanical,

TABLE 3. Classification of lithium primary cells*

Cell classification	Typical electrolyte	Power capability	Size, Ah	Operating range, °F (°C)	Shelf life, years	Typical cathodes	Nominal cell voltage, V	Key characteristics
Soluble cathode (liquid or gas)	Organic or inorganic (w/solute)	Moderate to high power, W	0.5–20,000	−67 to 158 (−55 to 70)	5–10	Sulfur dioxide Thionylchloride Sulfuryl chloride	3.0 3.6 3.9	High energy output; high power output; low temperature operation; long shelf life
Solid cathode	Organic (w/solute)	Low to moderate power, mW	0.03–5	−40 to 122 (−40 to 50)	5–8	Silver chromate Manganese dioxide Carbon monofluoride Copper(I) sulfide Iron disulfide Iron sulfide Copper(II) oxide	3.1 3.0 2.6 1.7 1.6 1.5 1.5	High energy output for moderate power requirements; nonpressurized cells
Solid electrolyte	Solid state	Very low power, μW	0.003–0.5	32 to 212 (0 to 100)	10–25	I_2 poly-2-vinylpyridine	2.8	Excellent shelf life; solid-state—no leakage; long-term microampere discharge

*From D. Linden (ed.), *Handbook of Batteries and Fuel Cells*, McGraw-Hill, 2d ed., 1995.

and electrical features (**Table 3**). Performance advantages of lithium cells over other primary cells include high voltage (may be above 3 V), high energy density, operation over a wide temperature range, and long shelf life. Lithium cells are manufactured in different sizes and configurations, ranging in capacity from milliamperehours to over 20,000 Ah, with characteristics matched to the general application. *See* LITHIUM.

In a lithium cell, nonaqueous solvents must be used for the electrolyte because of the solubility and reactivity of lithium in aqueous solutions. Organic solvents, such as acetonitrile and propylene carbonate, and inorganic solvents, such as thionyl chloride, are typical. A compatible solute is added to provide the necessary electrolyte conductivity. Solid-cathode cells are generally manufactured as cylindrical cells in sizes up to about 30 Ah in both low- and high-rate constructions. The low-rate, or high-energy, designs use a bobbin-type construction to maximize the volume for active materials (**Fig. 5*a***). These cells have the highest energy density of cells of similar size. The high-rate cells use a jelly-roll (spiral-wound) construction, which provides the large electrode surface required for the high discharge rates (Fig. 5*b*). These cells deliver the highest rates of all primary batteries and are widely used in military applications. Early ambient-temperature lithium systems used lithium-sulfur dioxide chemistry, and found applications in medical implants, weapons systems, and spacecraft. These cells provide 3 V and have seven times the energy of a typical alkaline battery. However, other lithium chemistry has been developed. Four types of cells are lithium-carbon monofluoride, lithium-manganese dioxide, and lithium-iron disulfide, all of which have solid cathodes, and lithium-thionyl chloride, which has a liquid cathode.

Lithium-carbon monofluoride batteries have a solid carbon monofluoride cathode and characteristics that allow the battery to be used for the life of the product in which the battery is installed, such as computer real-time clock and memory backup applications, instead of having scheduled replacement intervals. The cell has a storage and operational temperature range of −40 to +85°C, and a long life made possible by a self-discharge rate that may be

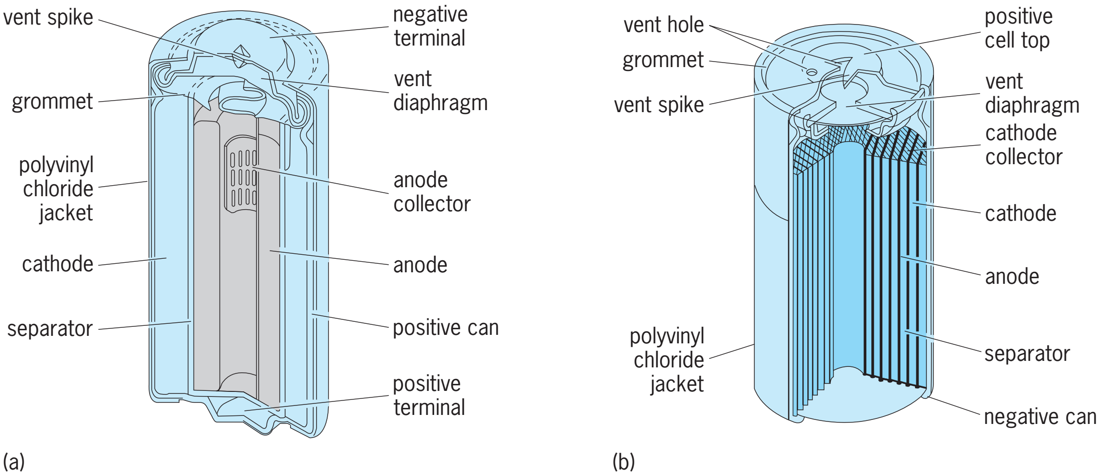


Fig. 5. Solid cathode cells. (*a*) Lithium-manganese dioxide bobbin cell. (*b*) Lithium-manganese dioxide spiral-wound cell. (*Duracell, Inc.*)

less than 0.2% of capacity loss per year. As a lithium-carbon monofluoride cell discharges while powering a specified load, internal resistance generally remains low and level while the closed-circuit voltage profile remains high, flat, and stable until the depth of discharge exceeds approximately 85%. When completely discharged, this type of lithium cell may be disposed of as nonhazardous waste. Batteries containing unreacted lithium metal are considered reactive hazardous waste and must be disposed of following the applicable local, state, and federal regulations.

Lithium-manganese dioxide cells have a solid manganese dioxide cathode and characteristics that make the preferred use of the cell to be where periodic replacement is routinely performed. The cell can supply both pulse loads and very small current drains, typical needs for microprocessor applications. As the cell discharges, its internal resistance increases because of the manganese dioxide, causing a tapered discharge profile and a declining closed circuit voltage. The manganese dioxide also limits the maximum temperature rating of the cell to 140°F (60°C), a temperature at which self-discharge may increase to a rate of over 8% per year. However, the cell costs less than a lithium-carbon monofluoride cell, and is widely used especially for applications requiring intermittent pulsing such as remote keyless entry systems. When completely discharged, this type of lithium cell may be disposed of as non-hazardous waste.

Lithium-iron disulfide cells have an anode of lithium foil in contact with a stainless-steel can and a cathode-material mixture of graphite and iron disulfide. These cells have an open-circuit voltage of 1.8 V, a nominal voltage of 1.5 V, and can be used in any application that uses other AA-size 1.5 V batteries. Each battery contains two safety devices: a thermal switch which acts as a current limiter if the battery begins to overheat, and a pressure-relief vent that opens at a predetermined temperature. This type of lithium cell may be disposed of using the same procedures approved for other cells.

Lithium-thionyl chloride cells were designed for the military and have been used for nuclear surety applications, which require long-life power sources for command and control systems. These cells have a voltage of about 3.6 V, making possible the use of fewer cells for the same application. They are classed as high-energy, high-current-drain-rate cells, with twice the energy density of lithium-sulfur dioxide cells, and will last twice as long. Lithium-thionyl chloride cells have a soluble cathode which is liquid, toxic, and corrosive. Each cell is hermetically sealed, with a designed-in fuse to avoid rupture or explosion if the battery is inadvertently charged or shorted. A lithium-thionyl chloride cell should be replaced only by a trained technician. Because thionyl chloride is highly toxic and corrosive, this battery is an environmental hazard which requires special handling and disposal.

Solid-electrolyte cells. These may be classified as (1) cells using a solid crystalline salt (such as lithium iodide) as the electrolyte and (2) cells using a solid-polymer electrolyte. With either type of electrolyte, the conductivity must be nearly 100% ionic as any electronic conductivity causes a discharge of the cell, which limits shelf life. Solid-polymer electrolytes are based on the characteristic that solutions of alkali metal salts dissolved in certain polymers, such as poly(ethylene oxide), form solid polymers that have reasonable ionic conductivity. The solid polymers can be fabricated in thin cross sections, are more stable than liquid electrolytes, and can serve as the separators. Most development work on thin-film polymer electrolyte batteries has been for rechargeable batteries, using a lithium anode and solid cathode materials. Energy densities of 5–10 times that of lead-acid cells have been calculated. The technology can also be applied to the design of primary cells and batteries. *See* ELECTROLYTIC CONDUCTANCE; IONIC CRYSTALS.

The most widely used solid-electrolyte cells have lithium iodide as the solid electrolyte. These batteries have a solid lithium foil anode and a cathode that is largely iodine. The iodine is made conductive by the addition of an organic compound, poly(2-vinylpyridine). This iodine-polyvinylpyridine cathode has a tarlike consistency in the fresh battery, then solidifies gradually as the battery is discharged. Lithium and iodine are consumed during the discharge, and the reaction product, lithium iodide, forms in place in the region between the two reactants where it serves as the cell separator. Because the electrolyte continues to form as the battery discharges, the overall resistance of the cell continually increases with discharge. This results in a drop in cell voltage for a given current drain. The nominal cell voltage is around 2.8 V. These batteries have a very low rate of self-discharge, a long storage life, and high reliability, but can be discharged only at low rates because of the low conductivity of the electrolyte. Applications for these batteries include memory backup and implanted cardiac pacemakers.

A typical implantable battery has a volume of 0.4 in.^{3} (6 cm^3), weighs 0.8 oz (23 g), and has a 2 Ah capacity. The lifetime of this battery is 5–10 years since pacemakers typically draw only 15-30 microamperes. The battery is considered to be discharged when the output voltage drops to 1.8 V. This may be detected by monitoring the patient's pulse rate. *See* MEDICAL CONTROL SYSTEMS.

Cells that use only solid components, including an ion-conducting solid electrolyte, are called solid-state cells or batteries. Some of these have electrodes made of fast-ion conductors, materials that are good conductors for both ions and electrons. Examples include the layered-structure disulfides of titanium and vanadium, TiS_2 and VS_2, which can be used as the cathode or sink for lithium, and aluminum-lithium alloys, which can be used as the anode or source of lithium in lithium batteries. These materials can sustain high reaction rates because of their unique crystalline structures which allow the incorporation of ions into their crystalline lattices without destruction of those lattices. *See* ENERGY STORAGE; INTERCALATION COMPOUNDS; SOLID-STATE CHEMISTRY.

In a solid-state cell, the polycrystalline pressed electrolyte is interspaced between a metallic anode and the solid cathode material. The electrodes are applied to the electrolyte by mechanically pressing the materials together, or in some cases the electrolyte is formed in place by reaction between the two electrodes. These cells are then stacked together to form a battery of the required voltage. A carbon current collector is often used on the cathode side, and this is frequently admixed with the cathode material. If the cathode material is sufficiently conductive, for example titanium disulfide, no carbon conductor is needed.

Reserve batteries. Most primary batteries are ready for use at the time of manufacture. However, high-energy primary batteries are often limited in their application by poor charge retention resulting from self-discharge between the electrolyte and the active electrode materials. The use of an automatically activated battery, or reserve battery, can overcome this problem, especially when the battery is required to produce high currents for relatively short periods of time (seconds or minutes), typical of military applications. A reserve battery is manufactured as a complete but inert battery that is stored in an inactive condition by keeping one of the critical cell components, such as the electrolyte, separated from the rest of the battery. The battery is activated just prior to use by adding this component manually or automatically. An important design consideration is to ensure that the electrolyte or other withheld component is delivered as quickly as possible at the time of activation while avoiding chemical short-circuiting of the cells.

Zinc-silver oxide reserve batteries. Automatically activated batteries have been used in missile and spacecraft applications, where the battery has to be activated from a remote source. The package contains a mechanism that drives the electrolyte out of the reservoir and into the cells of the battery. Most of these automatically activated batteries use zinc-silver oxide chemistry with a solution of potassium hydroxide for the electrolyte in order to achieve the required high discharge rates.

Magnesium water-activated batteries. Several types of water-activated (dunk type) batteries use magnesium anodes and silver chloride or cuprous chloride cathodes. The batteries are assembled dry, with the active elements separated by an absorbent. Activation occurs by pouring water into the battery container or by immersing the battery in water or seawater. Magnesium-silver chloride batteries have many marine applications, including buoys, beacons, flares, and safety lights. They also power the pingers which are attached to and help locate the cockpit

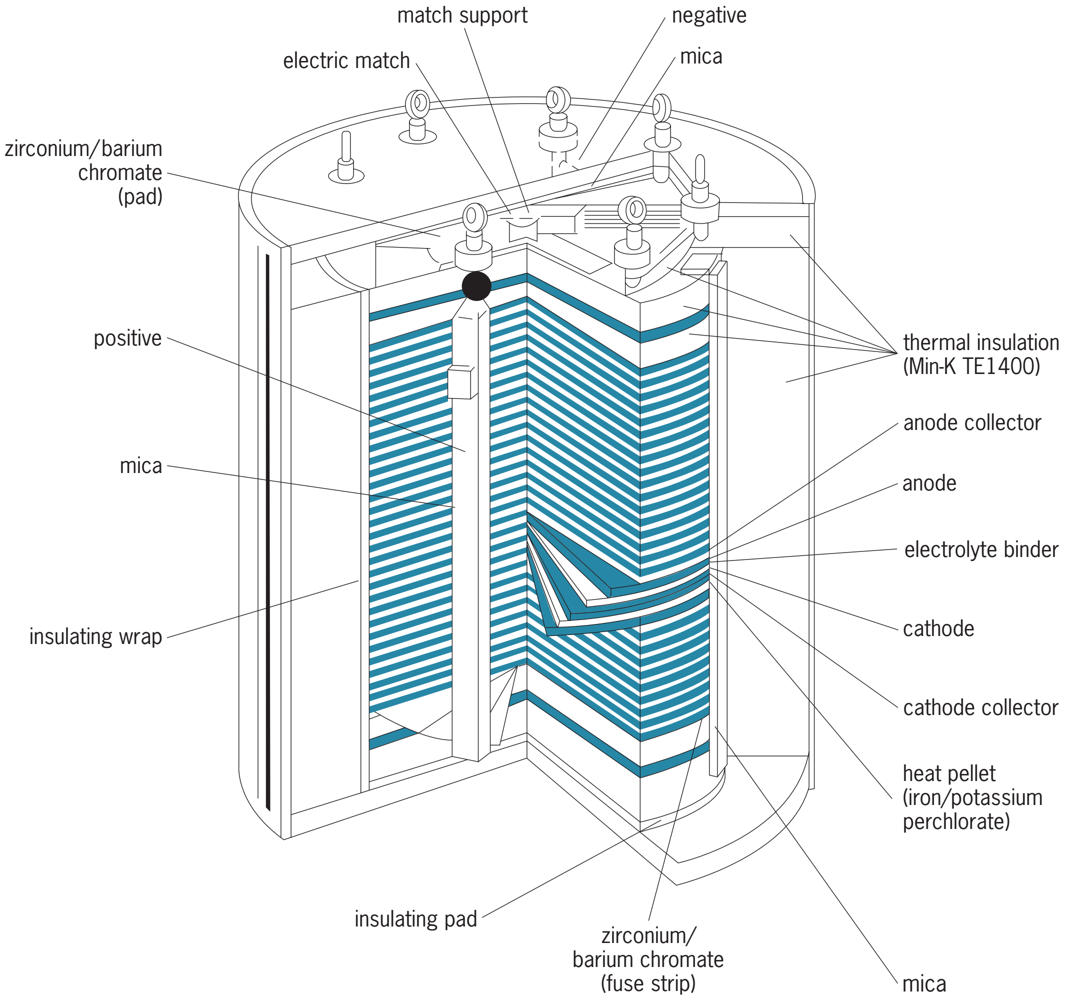

match support
negative
electric match
mica
zirconium/barium chromate (pad)
positive
mica
insulating wrap
thermal insulation (Min-K TE1400)
anode collector
anode
electrolyte binder
cathode
cathode collector
heat pellet (iron/potassium perchlorate)
insulating pad
zirconium/ barium chromate (fuse strip)
mica

Fig. 6. Lithium/iron disulfide thermal battery. (*Sandia Laboratories*)

-s

voice recorder and flight data recorder in downed aircraft, and emergency lighting for submarine escape hatches. Magnesium-copper iodide and magnesium-cuprous chloride batteries have similar capabilities.

Lithium-anode reserve batteries. Several types of batteries having a lithium anode are available as reserve batteries. These include batteries using lithium-sulfur dioxide, lithium-thionyl chloride, and lithium-vanadium pentoxide chemistry. Lithium-sulfur dioxide and lithium-thionyl chloride reserve batteries may be gas-activated, with activation triggered by electrical, mechanical, or pressure signals. The lithium-vanadium pentoxide battery is an electrolyte-activated battery. The electrolyte is contained in a glass ampule or vial, which is broken by mechanical shock, allowing the electrolyte to disperse within the cell.

Thermal batteries. A thermal battery, or fused-electrolyte battery, is a type of reserve battery that is activated by the application of heat. At room temperature, the electrolyte is a solid and has a very low conductivity, rendering the battery essentially inert. When the temperature is raised above the melting point of the electrolyte, the molten electrolyte becomes ionically conductive, and the battery is then capable of delivering electrical energy. Thermal batteries were introduced in 1955 to solve the wet-stand time limitation of aqueous systems, and subsequently were used as the main power source for nuclear weapon systems.

Early thermal batteries used calcium-calcium chromate chemistry. This was replaced in 1980 by thermal batteries that used a lithium anode, an iron sulfide cathode, and a lithium chloride-potassium chloride electrolyte. **Figure 6** shows a typical multicell lithium-iron disulfide thermal battery, which is activated by an electrical pulse to an integral electric match (squib). Other thermal batteries are activated mechanically using a percussion-type primer. The principal means for heating are pyrotechnic heat sources such as zirconium barium chromate heat paper or a heat pellet containing fine iron powder and potassium perchlorate. When the temperature rises to about 750°F (400°C), the electrolyte becomes molten and conductive. For other sources of electric energy known as batteries or cells *See* FUEL CELL; NUCLEAR BATTERY; SOLAR CELL. Donald L. Anglin

Secondary Batteries

Secondary batteries (also known as accumulators) are rechargeable. This means that the electrochemical reactions in the cell must be reversible so that if the load in the external circuit is replaced by a power supply, the reactions in the cell can be forced to run in reverse, thereby restoring the driving force for reaction and hence recharging the cell. In contrast, primary batteries cannot be recharged because the reactions that produce current cannot be made to run in reverse; instead, totally different reactions occur when the cell is forcibly charged, and in some instances the reaction products are dangerous, that is, explosive or toxic.

The paradigm of battery design is to identify a chemical reaction with a strong driving force and then to fashion a cell that requires the reaction to proceed by a mechanism involving electron transfer, thereby making electrons available to a load in the external circuit. The magnitude of the driving force will determine cell voltage; the kinetics of reaction will determine cell current.

There are many chemistries that will serve as the basis for secondary batteries. Distinctions can be made on the basis of the following metrics: voltage, current (maximum, steady-state, and peak), energy density (Wh/kg and Wh/L), power density (W/kg and W/L), service life (cycles to failure), and cost ($/kWh). **Table 4** summarizes performance characteristics of some of the secondary battery types. Such comparisons are made difficult by the fact that many performance characteristics are functions of battery size and service conditions. For example, service life is greatly affected by depth of discharge and discharge current. However, both of these operating parameters can be very different for two different battery types that were designed for very different applications.

Another classification scheme considers the state of aggregation of the principal cell components, that is, whether the electrolyte is a solid or liquid, and whether the electrode active material is a solid, liquid, or gas. Most batteries have solid electrodes and a liquid electrolyte. However, there are examples of batteries in which the anode and cathode are both liquid, and the electrolyte is solid.

What follows is a series of descriptions of the significant secondary battery chemistries.

Lead-acid batteries. The lead-acid battery is the dominant secondary battery, used in a wide variety of applications, including automotive SLI (starting, lighting, ignition), traction for industrial trucks, emergency power, and UPS (uninterruptible power supplies). The attributes of lead-acid batteries

TABLE 4. Comparison of performance characteristics of secondary batteries

Characteristics	Pb-acid	Ni-Cd	Ni-MH	Zn-Ag	Na-S	Zn-air[*]	Li-ion	Li-SPE[†]
Nominal voltage, V	2	1.2	1.2	1.5	1.9	1.5	3.6	3.6
Specific energy, Wh/kg	35	40	90	110	80	280	125	500
Specific energy, kJ/kg	126	144	324	396	288	1008	450	1800
Volumetric energy, Wh/L	70	100	245	220		385	440	900
Volumetric energy, kJ/L	252	360	882	792		1386	1584	3240

[*] Based upon commercial prototype notebook computer battery.
[†] Projections based upon thin-film microbattery test results in the laboratory.

include low cost, high discharge rate, and good performance at subambient temperatures. The first practical lead-acid cell was made by Planté in France in 1860.

The anode is metallic lead. The cathode active material is lead dioxide, which is incorporated into a composite electrode also containing lead sulfate and metallic lead. The electrolyte is an aqueous solution of sulfuric acid, 37% by weight when the battery is fully charged.

The half-cell reactions at each electrode are shown in reactions (8) and (9). The overall cell reaction is the formation of lead sulfate and water, shown in reaction (10). When the battery is discharging,

$$\text{Anode}: \mathrm{Pb} + \mathrm{SO_4^{2-}} \rightarrow \mathrm{PbSO_4} + 2e^- \quad (8)$$

$$\text{Cathode}: \mathrm{PbO_2} + \mathrm{H_2SO_4} + 2\mathrm{H^+} + 2e^- \rightarrow \mathrm{PbSO_4} + 2\mathrm{H_2O} \quad (9)$$

$$\text{Overall}: \mathrm{Pb} + \mathrm{PbO_2} + 2\mathrm{H_2SO_4} \rightarrow 2\mathrm{PbSO_4} + 2\mathrm{H_2O} \quad (10)$$

the reactions proceed from left to right; upon charging, the reactions proceed from right to left. The elementary electrochemical reactions are shown in reactions (11) and (12). On discharge, electrons are

$$\text{Anode}: \mathrm{Pb} \rightarrow \mathrm{Pb^{2+}} + 2e^- \quad (11)$$

$$\text{Cathode}: \mathrm{Pb^{4+}} + 2e^- \rightarrow \mathrm{Pb^{2+}} \quad (12)$$

produced at the anode as metallic lead is oxidized to Pb(II) present in lead sulfate; the complementary reaction at the cathode is the reduction of Pb(IV) present in lead dioxide to Pb(II) present in lead sulfate with the attendant consumption of electrons.

Sulfuric acid functions not only as the electrolyte but also as one of the reactants. Consequently, the acid concentration is a function of the instant state of charge and attains maximum value when the battery is fully charged.

Nominal voltage is 2 V, which decreases as the battery is discharged. The lower limit is governed by the depth of discharge. The theoretical specific energy is 170 Wh/kg. Commercially available lead-acid batteries deliver about 35 Wh/kg, which is about 20\% of theoretical. However, lead-acid batteries are capable of high specific power, have relatively flat discharge profiles, are comparatively insensitive to voltage changes with temperature, can be made for calendar lives exceeding 20 years, and are the least costly of all rechargeable technologies. Thus, when mobility is not a consideration and hence there is no penalty for the high density of lead, lead-acid technology can be found. The disadvantages of lead-acid technology include low specific energy and possibility of hydrogen evolution.

Nickel-cadmium batteries. The nickel-cadmium battery is the dominant alkaline secondary battery and is used in many of the same heavy industrial applications as are lead-acid batteries. At the same time, nickel-cadmium batteries are found in applications requiring portable power such as electronics and power tools. The first nickel-cadmium battery was made by Jungner in Sweden in 1900.

The anode is metallic cadmium. The cathode-active material is nickel oxide present as nickel oxyhydroxide (NiOOH). The active materials of both electrode are incorporated into porous sintered nickel plaques. The electrolyte is an aqueous solution of potassium hydroxide, 31% by weight when the battery is fully charged.

The half-cell reactions at each electrode are shown in reactions (13) and (14). The overall cell reaction is the formation of nickel hydroxide and cadmium hydroxide, shown in reaction (15). The elementary

$$\text{Anode}: \mathrm{Cd} + 2\mathrm{OH^-} \rightarrow \mathrm{Cd(OH)_2} + 2e^- \quad (13)$$

$$\text{Cathode}: 2\mathrm{NiOOH} + 2\mathrm{H^+} + 2e^- \rightarrow 2\mathrm{Ni(OH)_2} \quad (14)$$

$$\text{Overall}: \mathrm{Cd} + 2\mathrm{NiOOH} + 2\mathrm{H_2O} \rightarrow 2\mathrm{Ni(OH)_2} + \mathrm{Cd(OH)_2} \quad (15)$$

electrochemical reactions are shown in reactions (16) and (17). On discharge, electrons are produced

$$\text{Anode}: \mathrm{Cd} \rightarrow \mathrm{Cd^{2+}} + 2e^- \quad (16)$$

$$\text{Cathode}: 2\mathrm{Ni^{3+}} + 2e^- \rightarrow 2\mathrm{Ni^{2+}} \quad (17)$$

at the anode as metallic cadmium is oxidized to Cd(II) present in cadmium hydroxide; the complementary reaction at the cathode is the reduction of Ni(III) present in nickel oxyhydroxide to Ni(II) present in nickel hydroxide with the attendant consumption of electrons.

Nominal voltage is 1.2 V. The theoretical specific energy is 220 Wh/kg. Commercially available nickel-cadmium batteries deliver about 40 Wh/kg, although cells with plastic-bonded plates can achieve 55 Wh/kg. Nickel-cadmium batteries are capable of withstanding considerable mechanical abuse. They are also chemically quite stable owing to the fact that none of the cell components is attacked by the electrolyte. The disadvantages of nickel-cadmium technology include low specific energy, and the so-called memory effect in which the cell loses capacity as it seemingly conforms to a duty cycle that does not involve full discharge.

Nickel-metal hydride batteries. The nickel-metal hydride battery, commonly designated NiMH, is a comparatively new technology that has found uses in laptop computers, cellular telephones, and other portable devices where high specific energy is sought. Larger cells have been built for use in electric vehicles.

The electrochemistry shares some features with the nickel-cadmium battery. For example, the cathode active material is nickel oxyhydroxide. In contrast, the anode active material is hydrogen, which is present as the hydride of a proprietary metal alloy. The electrolyte is an aqueous solution of potassium hydroxide, roughly 30% by weight.

The half-cell reactions at each electrode are shown in reactions (18) and (19). The overall cell reaction is the formation of nickel hydroxide and extraction of the hydrogen storage alloy, shown in reaction (20).

$$\text{Anode}: \mathrm{MH} + \mathrm{OH}^- \rightarrow \mathrm{M} + \mathrm{H_2O} + \mathrm{e}^- \quad (18)$$

$$\text{Cathode}: \mathrm{NiOOH} + \mathrm{H_2O} + \mathrm{e}^- \rightarrow \mathrm{Ni(OH)_2} + \mathrm{OH}^- \quad (19)$$

$$\text{Overall}: \mathrm{MH} + \mathrm{NiOOH} \rightarrow \mathrm{Ni(OH)_2} + \mathrm{M} \quad (20)$$

Here M denotes the hydrogen storage alloy and MH is the hydride compound of same. The elementary electrochemical reactions are shown in reactions (21) and (22). On discharge, electrons are produced at the anode as hydrogen stored in the alloy is

$$\text{Anode}: \mathrm{H} \rightarrow \mathrm{H}^+ + \mathrm{e}^- \quad (21)$$

$$\text{Cathode}: \mathrm{Ni}^{3+} + \mathrm{e}^- \rightarrow \mathrm{Ni}^{2+} \quad (22)$$

oxidized to protons, which combine with hydroxyl ions in the electrolyte to form water; the complementary reaction at the cathode is the reduction of Ni(III) present in nickel oxyhydroxide to Ni(II) present in nickel hydroxide with the attendant consumption of electrons.

Nominal voltage is 1.2 V. The theoretical specific energy is 490 Wh/kg. Commercially available nickel-cadmium batteries deliver about 90 Wh/kg, which is substantially higher than nickel-cadmium or lead-acid batteries. However, materials costs associated with the hydrogen storage alloy are high, and thus the final cost per unit charge ($/kWh) is higher for NiMH than for the other two cited technologies.

Silver-zinc batteries. The silver-zinc battery boasts the highest specific energy and specific power of any secondary battery containing an aqueous electrolyte. However, the high cost of silver has restricted the battery's use to military and aerospace applications primarily.

The anode is metallic zinc. The cathode active material is silver oxide. The electrolyte is an aqueous solution of potassium hydroxide, 45% by weight.

The discharge reaction occurs in two steps, shown in reactions (23) and (24). The overall cell reaction is

$$2\mathrm{AgO} + \mathrm{Zn} \rightarrow \mathrm{Ag_2O} + \mathrm{ZnO} \quad (23)$$

$$\mathrm{Ag_2O} + \mathrm{Zn} \rightarrow 2\mathrm{Ag} + \mathrm{ZnO} \quad (24)$$

the production of silver by metallothermic reduction of silver oxide by zinc. The elementary electrochemical reactions are shown in reactions (25), (26), and (27). On discharge, electrons are produced at the

$$\text{Anode}: \mathrm{Zn} \rightarrow \mathrm{Zn}^{2+} + 2\mathrm{e}^- \quad (25)$$

$$\text{Cathode}: 2\mathrm{Ag}^{2+} + 2\mathrm{e}^- \rightarrow 2\mathrm{Ag}^+ \quad (26)$$

$$2\mathrm{Ag}^+ + 2\mathrm{e}^- \rightarrow 2\mathrm{Ag} \quad (27)$$

anode as zinc is oxidized to Zn(II) present as zinc oxide; the complementary reactions at the cathode are the reduction of Ag(II) present in silver oxide first to Ag(I) present in $\mathrm{Ag_2O}$, followed by the reduction of Ag(I) to metallic silver, each step accompanied by the attendant consumption of electrons.

Nominal voltage is 1.5 V. The theoretical specific energy is 590 Wh/kg. Commercially available nickel-cadmium batteries deliver about 120 Wh/kg. However, silver-zinc suffers from the high costs associated with silver, loss of cathode active material through dissolution of silver oxide in the electrolyte, and cell shorting by zinc dendrites on charging, all of which lead to poor cycle life. However, when huge currents are needed (and cost is no object), silver-zinc is the battery of choice.

Sodium-sulfur batteries. The sodium-sulfur battery is unique in several ways. Most batteries have solid electrodes and a liquid electrolyte; the sodium-sulfur battery has molten electrodes separated by a solid membrane serving as the electrolyte. In order to keep the electrodes molten and the electrolyte sufficiently conductive, superambient temperatures are required; the sodium-sulfur battery operates at 300°C (572°F). Sodium-sulfur batteries were once prime candidates for electric vehicle power sources and electric utility energy storage for load leveling. However, in recent years, commercial production has ceased.

The electrolyte is a U-shaped tube made of the ceramic material aluminium oxide, specifically β''-$\mathrm{Al_2O_3}$ which is a sodium-ion conductor. The tube is filled with molten metallic sodium which is the anode active material. A metal rod immersed in the sodium serves as current collector. The anode/ electrolyte assembly in positioned inside a larger metal container. The space between the metal container and the ceramic electrolyte tube is occupied by carbon felt filled with elemental sulfur which is the cathode active material. The cathode current collector is the carbon felt which is in electrical contact with the outer metal container.

The half-cell reactions at each electrode are shown in reactions (28) and (29). The overall cell reaction is the formation of sodium sulfide, shown in reaction (30). The elementary electrochemical reactions

$$\text{Anode}: \ 2\mathrm{Na} \rightarrow 2\mathrm{Na}^+ + 2\mathrm{e}^- \quad (28)$$

$$\text{Cathode}: \ \mathrm{xS} + 2\mathrm{e}^- \rightarrow \mathrm{S_x^{2-}} \quad (29)$$

$$\text{Overall}: \ 2\mathrm{Na} + \mathrm{xS} \rightarrow \mathrm{Na_2S_x} \quad (30)$$

$$2.7 < x < 5$$

are identical to the above. The key to the cell's operation is the fact that β''-$\mathrm{Al_2O_3}$ acts as a separator to prevent direct contact of molten sodium and sulfur while allowing only Na^+ to pass from the anode chamber to the cathode chamber.

On discharge, electrons are produced at the anode as metallic sodium is oxidized to Na(I) present in the β''-$\mathrm{Al_2O_3}$ electrolyte as free sodium ions; the complementary reaction at the cathode is the reduction of elemental sulfur to S(II) present as polysulfide ion, S_x^{2-}, which reacts immediately to form sodium sulfide. The sodium ions are delivered to the cathode chamber by the solid electrolyte. Thus, as the cell discharges, the sulfur content of the cathode chamber decreases and the sodium sulfide concentration rises.

Nominal voltage is 1.9 V. The theoretical specific energy is 750 Wh/kg. The last commercially available sodium-sulfur batteries delivered about

80 Wh/kg. While sodium-sulfur batteries have performance characteristics comparable to the best batteries with aqueous electrolytes, thermal management requirements (keeping the cell at temperature) and the indeterminacy of the mode of failure, which in turn lead to safety concerns, have limited the successful deployment of this battery technology.

Zinc-air batteries. The combination of a metal anode and an air electrode results in a battery with an inexhaustible supply of cathode reactant. Recharging involves restoring the anode either electrochemically or mechanically, that is, by direct replacement. In principle, extremely high specific energies should be attainable as the theoretical energy density is 1310 Wh/kg.

The anode is metallic zinc. The cathode active material is oxygen, which is present as a component of air. The current collector is high-surface-area carbon. The electrolyte is an aqueous solution of potassium hydroxide 45% by weight, which may be gelled to prevent leakage.

The half-cell reactions at each electrode electrode are shown in reactions (31) and (32). The overall cell reaction is the oxidation of zinc to form tetrahydroxyzincate [$Zn(OH)_4{}^{2+}$] or zinc oxide, depending upon the instant composition of the electrolyte, shown in reaction (33).

$$\text{Anode}: Zn + 4OH^- \rightarrow Zn(OH)_4^{2-} + 2e^- \quad (31)$$

$$\text{Cathode}: {}^1/_2O_2 + H_2O + 2e^- \rightarrow 2OH^- \quad (32)$$

$$\text{Overall}: Zn + {}^1/_2O_2 + H_2O + 2OH^- \rightarrow Zn(OH)_4^{2-} \quad (33)$$

When the concentration of $Zn(OH)_4{}^{2-}$ exceeds the solubility limit, precipitation of zinc oxide occurs according to reaction (34).

$$\text{Precipitation}: Zn(OH)_4^{2-} \rightarrow ZnO + H_2O + 2OH^- \quad (34)$$

Under these conditions the overall reaction is effectively the oxidation of zinc according to reaction (35).

$$\text{Overall}: Zn + 1/_2O_2 \rightarrow ZnO \quad (35)$$

The elementary electrochemical reactions are shown in reactions (36) and (37).

$$\text{Anode}: Zn \rightarrow Zn^{2+} + 2e^- \quad (36)$$

$$\text{Cathode}: {}^1/_2O_2 + H_2O + 2e^- \rightarrow 2OH^- \quad (37)$$

On discharge, electrons are produced at the anode as zinc is oxidized to Zn(II) present as $Zn(OH)_4{}^{2-}$ or zinc oxide; the complementary reaction at the cathode is the reduction of oxygen present in air to O(II) present in hydroxide.

Nominal voltage is 1.5 V. For applications such as notebook computers demonstration cells have been built that deliver 280 Wh/kg. Other attributes include a relatively flat discharge curve, long shelf life thanks to water activation (battery is stored "dry"), and visual inspection of anode condition to determine state of charge. Among the disadvantages is the complexity of a system that comprises a solid anode, liquid electrolyte, and a gaseous cathode. Because the oxygen electrode does not stand up well to operation as an anode (which must happen on charging), a third electrode must be incorporated into the cell. Alternatively, provision must be made for easy disassembly for replacement of a fresh zinc anode and subsequent flawless reassembly of the "recharged" battery.

Lithium-ion batteries. The lithium-ion battery is fast becoming the dominant battery in applications requiring portable power such as electronics and power tools. As such, it is displacing nickel-cadmium and nickel-metal hydride from some of their traditional uses.

The term "lithium-ion" derives from the fact that there is no elemental lithium in the battery; instead, lithium shuttles between hosts, a behavior that has earned this battery the nickname "rocking chair." In the anode, the chemical potential of lithium is high, that is, desirably near that of pure lithium; in the cathode, the chemical potential of lithium is low. The anode is a carbonaceous material (graphite or coke) chosen for its electronic conductivity and its ability to intercalate lithium at potentials near that of pure lithium. An example is LiC_6. In commercial cells the cathode active material is a lithiated transition-metal oxide such as lithium cobalt oxide. Because lithium is more electropositive than hydrogen, the electrolyte must be nonaqueous and aprotic. A representative formulation is a solution (1:1 by volume) of ethylene carbonate and propylene carbonate containing a suitable lithium salt (at a concentration of about 1 *M*) such as lithium hexafluorophosphate, $LiPF_6$, which has been introduced in order to raise the conductivity of the electrolyte. For safety, a separator made of a polyolefin such as microporous polypropylene is placed between the electrodes. If the electrolyte temperature exceeds a certain value, the separator melts and current flow ceases.

The half-cell reactions at each electrode are shown in reactions (38) and (39).

$$\text{Anode}: xLi \rightarrow xLi^+ + xe^- \quad (38)$$

$$\text{Cathode}: xLi^+ + xe^- + LiCoO_2 \rightarrow Li_1 + xCoO_2 \quad (39)$$

The overall cell reaction is the formation of lithium cobalt oxide, shown in reaction (40). The elementary electrochemical reactions are shown in reactions (41) and (42).

$$\text{Overall}: xLi + LiCoO_2 \rightarrow Li_1 + xCoO_2 \quad (40)$$

$$\text{Anode}: xLi \rightarrow xLi^+ + xe^- \quad (41)$$

$$\text{Cathode}: xCo^{4+} + xe^- \rightarrow xCo^{3+} \quad (42)$$

On discharge, electrons are produced at the anode as elemental lithium is oxidized to Li(I) present in the electrolyte as free lithium ions; the complementary reaction at the cathode is the reduction of Co(IV) to Co(III) with the attendant consumption of electrons. Note that lithium intercalates into carbon as Li^0, a neutral species, whereas lithium intercalates into $LiCoO_2$ as Li+, a charged species. However, the presence of Li^+ in the cathode triggers a valence shift in the host itself: the compensating electron does not neutralize Li^+ but is localized on Co^{4+}, converting it to Co^{3+}.

Nominal voltage is 3.6 V. The theoretical specific energy of this cell fitted with a $LiCoO_2$ as cathode active material is 770 Wh/kg. Commercially available lithium-ion batteries deliver about 125 Wh/kg.

Lithium-solid polymer electrolyte (SPE) batteries. Since the 1970s it has been known that, upon addition of appropriate salts, polymers can be rendered lithium-ion conductors. Such materials can serve as electrolytes in lithium batteries. When full measure is given to the capacity for miniaturization of a fully solid-state battery, it becomes evident that, in principle, this has the potential to attain the highest specific energy and specific power of any rechargeable technology. In addition, a lithium-SPE battery offers other advantages: ease of manufacture, immunity from leakage, suppression of lithium dendrite formation, elimination of volatile organic liquids, and mechanical flexibility. It is this last attribute that makes lithium-SPE batteries most intriguing: a multilayer laminate of thin films of metal, polymer, and ceramic measuring some tens of micrometers in thickness, totally flexible, with a specific energy in excess of 500 Wh/kg and capable of delivering power at rates in excess of 1000 W/kg at room temperature, operable over a temperature range from $-30°C$ to $+120°C$, and costing less than $500/kWh. Realization awaits advances in research in materials science and battery engineering.

Outlook. What is in store for secondary power sources, remains speculative. Perhaps higher-capacity batteries for automobile traction will be developed. There may be a push for thin-film, all-solid-state microbatteries, which will enable distributed power sources in combination with their own charger and electrical appliance or load. An example of such integration might be an electrical device powered by a secondary battery connected to a photovoltaic charger. Donald R. Sadoway

Bibliography. J. O. Besenhard (ed.), *Handbook of Battery Materials*, Wiley-VCH, Weinheim, 1999; T. R. Crompton, *Battery Reference Book*, 2d ed., 1996; D. Linden, *Handbook of Batteries*, 2d ed., 1995; *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., Wiley, New York, 1992; *Ullmann's Encyclopedia of Industrial Chemistry*, 5th ed., VCH, Weinheim, 1985.

Bauxite

A rock mainly comprising minerals that are hydrous aluminum oxides. These minerals are gibbsite, boehmite, and diaspore. The major impurities in bauxite are clay minerals and iron oxides. Bauxite is a weathering product of aluminous rock that results from intense leaching in tropical and subtropical areas, a process called laterization. Bauxite deposits are generally found on plateaus in stable areas where they had sufficient geologic time to form and were protected from erosion. *See* ALUMINUM.

Bauxite is the primary ore of aluminum. The two types of bauxites that are used commercially as aluminum ores are laterite and karst. Lateritic bauxites constitute more than three-fourths of the world's bauxite resources. Karstic bauxites are formed on a carbonate terrain and are concentrated in sinkholes and solution depressions on the surface of carbonate rocks. *See* LATERITE.

Commercial sources. Bauxite is mined and processed in many tropical and subtropical areas. Lateritic deposits are mined mainly in Australia, Brazil, West Africa, India, China, Surinam, Guyana, and Indonesia. The largest producer is Australia, from which about one-third of the world's bauxite production, about 125 million metric tons annually, is mined. Karstic bauxites are mined in Jamaica, Dominican Republic, Haiti, Hungary, Greece, Yugoslavia, Rumania, and Turkey. All the mines are open pit, and the benefication involves washing and screening in order to remove some of the clay and iron minerals. The beneficiated bauxite is transported to locations where alumina is produced from the bauxite. Four to seven tons of bauxite, depending on the available alumina content, are required to produce 2 tons of alumina, and approximately 2 tons of alumina are used to produce 1 ton of aluminum.

Uses. Bauxite used to produce alumina is called metallurgical grade; approximately 85% of the world's production is for this purpose. Other major uses are in refractories, abrasives, chemicals, and aluminous cements. The compositional requirements are much more rigid for these uses. The alumina content must be higher, and the iron, silica, and titanium contents significantly lower, than for metallurgical-grade bauxite. World resources of bauxite are many tens of billions of tons, so an adequate supply is assured for hundreds of years. However, since no metallurgical-grade bauxite is being mined in the United States, the nation is completely dependent on foreign sources. Haydn H. Murray

Bibliography. G. Bardossy, *Karst Bauxites*, 1982; P. W. Harben, *The Industrial Minerals Handybook*, 1998; L. Jacob Jr. (ed.), *Bauxite*, 1984.

Bayesian statistics

An approach to statistics in which estimates are based on a synthesis of a prior distribution and current sample data. Bayesian statistics is not a branch of statistics in the way that, say, nonparametric statistics is. It is, in fact, a self-contained paradigm providing tools and techniques for all statistical problems. In the classical frequentist viewpoint of statistical theory, a statistical procedure is judged by averaging its performance over all possible data. However, the bayesian approach gives prime importance to how a given procedure performs for the actual data observed in a given situation. Further, in contrast to the classical procedures, the bayesian procedures formally utilize information available from sources other than the statistical investigation. Such information, available through expert judgment, past experience, or prior belief, is described by a probability distribution on the set of all possible values of the unknown parameter of the statistical model at hand.

This probability distribution is called the prior distribution. The crux of the bayesian approach is the synthesis of the prior distribution and the current sample data into a posterior probability distribution from which all decisions and inferences are made. This synthesis is achieved by using a theorem proved by Thomas Bayes in the eighteenth century.

Example. The basic principles of bayesian statistics may be understood by considering an example in which a quality control inspector regularly receives lots of a manufactured product, each lot consisting of 10 units. In the past, 70% of the lots have had no defective units, 20% have had one defective unit, and 10% have had two defective units. The quality control inspector has just received a lot of 10 units, but has no time for 100% inspection and must estimate the quality of the lot by testing five units randomly selected. The inspector finds one defective unit among the five tested, and must make an estimate of θ, the proportion of defective units in the lot. If the inspector ignores past experience, a reasonable estimate of θ is 1/5. However, the past experience about θ can be incorporated into the estimation process by basing the estimate on the posterior distribution of θ given the observed data x, namely, $\pi(\theta|x)$ defined by Eq. (1), where α indicates "is proportional to,"

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta) \tag{1}$$

$\pi(\theta)$ describes the prior belief in the value θ, and $f(x|\theta)$ assesses the chances of the actual observed x when θ is the actual value of the parameter. In the present case, $x = 1$ (the observed number of defectives), $\pi(\theta) = 0{\cdot}70$ if $\theta = 0$, $\pi(\theta) = 0{\cdot}20$ if $\theta = 1/10$, and $\pi(\theta) = 0{\cdot}10$ if $\theta = 2/10$. Furthermore, it can be argued that the chance of there being one defective unit among the five tested when θ is the proportion of defective units in the lot is given by Eq. (2), where

$$f(1|\theta) = \frac{\begin{bmatrix} 10\,\theta \\ 1 \end{bmatrix} \begin{bmatrix} 10\,(1-\theta) \\ 4 \end{bmatrix}}{\begin{bmatrix} 10 \\ 5 \end{bmatrix}} \tag{2}$$

the brackets indicate binomial coefficients. It follows from Eqs. (1) and (2), and the fact that the sum of the posterior probabilities must equal unity, that $\pi(\theta|1$ is given by $\pi(\theta|1) = 0$ for $\theta = 0$, $\pi(\theta|1) = 9/14$ for $\theta = 1/10$, and $\pi(\theta|1) = 5/14$ for $\theta = 2/10$. (In this case, the factor of proportionality in Eq. (1) equals 90/14.) According to the bayesian approach, it is the posterior distribution $\pi(\theta|1)$ which should form the basis for estimating θ. *See* BINOMIAL THEOREM.

The posterior distribution combines the prior information about θ with the information contained in the observed data to give a composite picture of the final judgments about θ. In order to arrive at a single number as the estimate of θ, it may be necessary to bring in the notion of the loss suffered by the decision maker as a result of estimating the true value θ by the number $\hat{\theta}$. Depending on the choice of the loss function, the mode, the mean, and the median of the posterior distribution $\pi(\theta|x)$ are all reasonable estimates of θ.

Conjugate priors. The choice of the prior distribution for the unknown parameter θ is of crucial importance in bayesian statistics. The selected prior distribution for θ must be at least a reasonable approximation to the true beliefs about θ. In addition, the prior distribution $\pi(\theta)$ must be such that the posterior distribution $\pi(\theta|x)$ is tractable. Frequently, conjugate priors provide the answer. If (1) $\Im$ is the class of probability density functions $f(x|\theta)$ [indexed by θ] chosen to model the data-generating process, and (2) $\wp$ is a class of prior probability density functions $\pi(\theta)$ with the property that the posterior probability density function $\pi(\theta|x)$ belongs to the class $\wp$ whenever f belongs to $\Im$ and π belongs to $\wp$, then $\wp$ is said to be a family of conjugate priors for $\Im$. Thus, a conjugate family is closed under sampling in the sense that if the sample data are used to update a prior distribution in a conjugate family by Bayes' theorem the resulting posterior distribution will also belong to the same family. A prior density $\pi(\theta)$ is called a natural conjugate to the likelihood function $\mathrm{L}(\theta|x) = f(x|\theta)$ if $\pi(\theta)$ and $\mathrm{L}(\theta|x)$ are proportional as functions of θ.

A simple example of the usefulness of conjugate priors is provided by random sampling from a normal distribution. If $x = (x_1, \ldots, x_n)$ is a random sample from a normal distribution with unknown mean θ and known standard deviation σ, and if the prior distribution of θ is normal with mean μ and standard deviation λ, then the posterior distribution of θ given x is normal with mean $\alpha\mu + (1-\alpha)\bar{x}$ and standard deviation $\lambda\alpha^{1/2}$, where α and $\bar{x}$ are given by Eqs. (3) and (4). Thus, the family of normal priors is a conjugate family for the normal distribution with unknown mean and known standard deviation. It updates the prior parameters μ and λ to $\alpha\mu + (1-\alpha)\bar{x}$ and $\lambda\alpha^{1/2}$, respectively, thereby making transparent the effect of prior and sample information. *See* DISTRIBUTION (PROBABILITY).

$$\alpha = \frac{\sigma^2}{\sigma^2 + n\lambda^2} \tag{3}$$

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} \tag{4}$$

Noninformative and improper priors. The bayesian approach remains applicable even when little or no prior information is available. Such situations can be handled by choosing a prior density giving equal weight to all possible values of θ. Priors that seemingly impart no prior preference, the so-called noninformative priors, also arise when the prior is required to be invariant under certain transformations. Frequently, the desire to treat all possible values of θ equitably leads to priors with infinite mass. Such noninformative priors are called improper priors.

An example of the usefulness of improper priors is that of a random quantity whose pattern of variation is describable by a normal distribution with unknown positive mean θ and unit standard deviation. If no specific prior knowledge is available, then the improper prior given by $\pi(\theta) = 1$ for all $\theta > 0$ may be used. Although this prior has infinite mass, it results in a perfectly legitimate posterior density given by Eq. (5). If the observed x equals zero and if

$$\pi(\theta|x) \propto \exp\left\{\frac{-(\theta - x)^2}{2}\right\} \qquad \theta > 0 \qquad (5)$$

the mean of the posterior distribution is used to estimate θ, then the estimate of θ can be easily shown to be $(2/\pi)^{1/2}$. The classical estimate of θ is zero, which is unsuitable because θ is given to be positive.

Bayesian estimation. Optimal estimation of θ by a single point $\hat{\theta}$ requires specification of $l\,(\hat{\theta}, \theta)$, the loss suffered when the actual value θ is estimated by $\hat{\theta}$. If θ is unidimensional and $l\,(\hat{\theta}, \theta)$ is the squared error loss function $(\hat{\theta} - \theta)^2$, then the mean of the posterior distribution is the optimal point estimate of θ. If $l\,(\hat{\theta}, \theta)$ is the absolute error loss function $|\hat{\theta} - \theta|$, then the median of the posterior distribution is used as the estimate of θ. Point estimates of θ are usually accompanied by a measure of the accuracy of the estimate. When θ is unidimensional and the loss function is the squared error, the measure of accuracy of the Bayes estimate of θ is the variance of the posterior distribution. In situations involving noninformative priors, the Bayes estimates and their measures of accuracy are sometimes numerically very similar to the corresponding non-bayesian quantities.

If it is desired to estimate θ by a range of values in a parametric space Ω rather than by a single point $\hat{\theta}$, the bayesian approach is to select a subset C of Ω such that the probability assigned to C by the posterior distribution is at least a prespecified number $1-\alpha$. Such a subset C is called a $100(1-\alpha)\%$ credible set for θ. It is the bayesian analog of a classical confidence set. Bayesian credible sets are generally easier to find than classical confidence sets. In bayesian practice, the credible set selected is the highest posterior density (HPD) credible set C defined to consist of all values of θ such that $\pi(\theta|x)$ is equal to or greater than $\kappa(\alpha)$, where $\kappa(\alpha)$ is the largest constant for which C is a $100(1-\alpha)\%$ credible set. Often, the posterior distribution can be approximated by a normal distribution which yields an approximate HPD credible set without any computational difficulties.

Bayesian hypothesis testing. When concerned with the problem of deciding between two hypotheses, H_0, stating that θ is a member of Ω_0, and H_1, stating that θ is a member of Ω_1, the bayesian approach is to simply calculate the posterior probabilities $\alpha_i = \text{Prob}(\Omega_i|x)$, $i = 0,1$, and choose between H_0 and H_1 accordingly. This approach is in sharp contrast to the classical non-bayesian approach in which a test procedure is evaluated in terms of the so-called type I and type II error probabilities. The bayesian approach to multiple hypothesis testing is equally simple, namely, to calculate the posterior probability of each hypothesis and decide in favor of the hypothesis having the largest posterior probability.

The posterior probability of a hypothesis H_0, stating that θ is a member of Ω_0, is zero when Ω_0 is a singleton and the prior distribution is a continuous distribution. It would appear that the bayesian approach would not work for testing such hypotheses. However, the problem can be resolved by assigning a positive probability π_0 to the simple hypothesis $H_0 : \theta = \theta_0$ and a density $(1 - \pi_0)\pi(\theta)$ to $\theta \neq \theta_0$. Here $\pi(\theta)$ is a proper prior for θ, and π_0 is chosen as the probability mass that would be assigned to the hypothesis $|\theta - \theta_0| < \delta$ for some positive δ determined by the practical problem at hand. The rationale for this approach is the fact that the hypothesis $\theta = \theta_0$ is rarely reasonable and that, for suitably chosen $\delta > 0$, all θ in $(\theta_0 - \delta, \theta_0 + \delta)$ are of equal practical importance.

Bayesian tests of hypotheses are frequently reported in terms of Bayes factors. The Bayes factor in favor of a hypothesis H_0 is defined by Eq. (6). A

$$B = \frac{\text{posterior odds ratio of } H_0 \text{ to } H_1}{\text{prior odds ratio of } H_0 \text{ to } H_1} \qquad (6)$$

person can determine his or her personal posterior odds by simply multiplying the reported Bayes factor by his or her personal prior odds.

Criticism. Proponents of the bayesian paradigm claim a number of significant and fairly convincing advantages over the classical non-bayesian approach. However, bayesian statistics is itself criticized for its apparent lack of objectivity. The use of a prior distribution introduces subjectivity into a bayesian analysis so that decision makers having different priors may come to different conclusions even when working on the same problem with the same observed data. This, however, is the price that must be paid in order to incorporate prior information or expert judgment. Moreover, the bayesian approach seems to be the most logical approach if the performance of a statistical procedure for the actual observed data is of interest, rather than its average performance over all possible data. *See* ESTIMATION THEORY; PROBABILITY; STATISTICS.

Syed N. U. A. Kirmani

Bibliography. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 3d ed., 1993; J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, 1994; G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, 1973, reprint 1992; P. M. Lee, *Bayesian Statistics: An Introduction*, 2d ed., 1997; D. V. Lindley, *Bayesian Statistics: A Review*, 1972, reprint 1984; J. S. Press, *Bayesian Statistics: Principles, Models, and Applications*, 1989.

Bdelloidea

A class of the phylum Rotifera characterized by paired gonads and obligately asexual reproduction, found in diverse freshwater habitats worldwide. Bdelloid rotifers are among the most common microinvertebrates, particularly in ephemerally aquatic environments such as rain gutters, bird baths, moss, lichen, and detritus. First observed by A. Leeuwenhoek more than 300 years ago, bdelloids have long been a favorite subject of amateur microscopists and can easily be isolated and observed with a low-power microscope.

Morphology. Bdelloids are small; adults of most species are between 0.1 and 0.3 mm long. Despite their minute size, they contain about 1000 nuclei (cell membranes are ill-defined and many cells are syncytial). The body contains a nervous system, complex musculature, and complete digestive system, along with the corona (anterior ring of cilia), mastax (muscular pharynx), toes, and syncytial intracytoplasmic lamina (a specialized integument) characteristic of rotifers (**Fig. 1**). Bdelloids propel themselves through water by rhythmic beating of their coronal cilia. When attached to a substrate by their toes, they may withdraw their corona and extend an apical structure called the rostrum. Bdelloids crawl in a leechlike fashion by extending the rostrum. *See* ROTIFERA.

Classification. Although the systematics of Rotifera and allied phyla have been undergoing revision, Bdelloidea is usually afforded the class status, and most morphological and molecular evidence indicates that it is most closely related to the rotifer group Monogononta. There are 370 described species classified in four easily distinguished families. Philodinidae and Habrotrochidae are the most species-rich families and contain the most common, morphologically typical species. Members of the family Adinetidae rarely swim and have a corona and mastax that are specialized for scraping food particles off the substrate. Philodinavidae is a poorly explored family that includes the only known predatory bdelloids. *See* MONOGONONTA.

Ecology and physiology. Bdelloids can be found in almost any freshwater ecosystem at temperatures ranging from 32 to 113°F (0 to 45°C) and pH levels from 2.5 to 10, and are important components of benthic and littoral ecosystems. Most eat bacteria, fungi, algae, and detritus, and are in turn eaten by cladocerans, copepods, insect larvae, and fish fry (newborn fish). Some species (particularly *Embata*) are epizoic, living on the bodies of freshwater arthropods. Bdelloids are particularly common in environments that are only temporarily hydrated and are capable of surviving desiccation at any stage of their life: as water evaporates from the local environment, a bdelloid exudes all free water, contracts into a form called a tun, and enters a metabolically quiescent state called anhydrobiosis (**Fig. 2**). In the anhydrobiotic state bdelloids can survive vacuum, high levels of radiation, and temperatures from −454 to 392°F (−270 to 200°C). When rehydrated, bdelloids resume activity with little or no decrease in lifespan or fecundity. There are apocryphal accounts of bdelloids revived after years of anhydrobioisis.

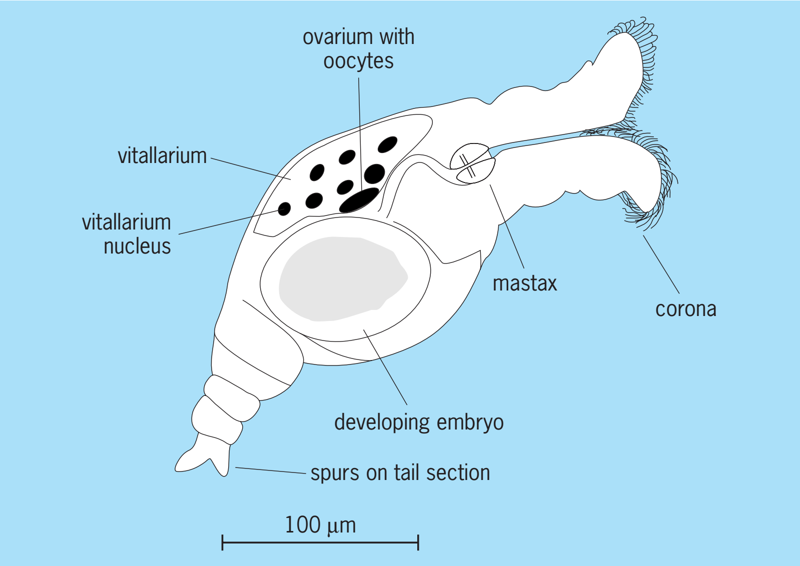


Fig. 1. Bdelloid rotifer *Macrotrachela quadricornifera*. The vitallarium provides nutrients to the developing egg.

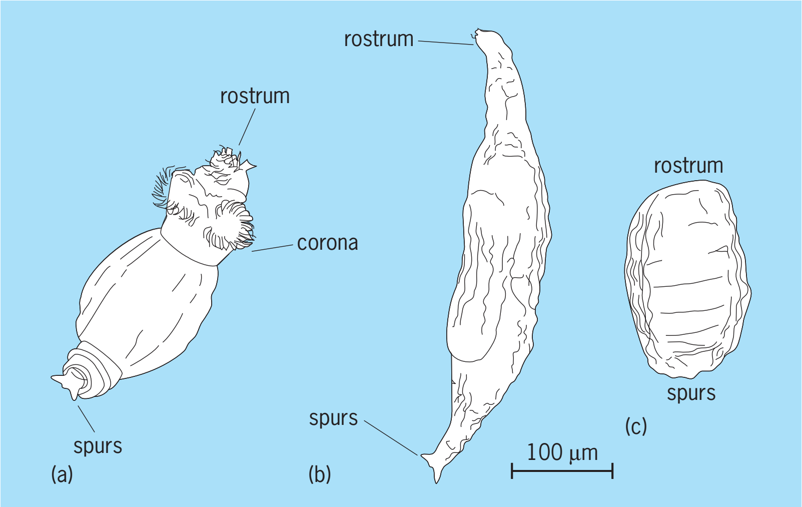


Fig. 2. Bdelloid rotifer *M. quadricornifera* contracts into an organized, compact tun as it undergoes desiccation and enters anhydrobiosis. (*a*) Active adult. (*b*) Adult contracting and entering anhydrobiosis. (*c*) Anhydrobiotic tun.

Reproduction. Males and hermaphrodites are unknown in Bdelloidea, and reproduction is parthenogenetic: females produce embryos from oocytes after two mitotic divisions without chromosome pairing or reduction (ameiotic, or apomictic, thelytoky). Bdelloid rotifers are the largest, oldest, most successful animal group for which observational, cytological, and molecular evidence are consistent with the long-term asexuality of the group. They are a model system for the study of the paradox of the maintenance of sexual reproduction.

David Mark Welch

Bibliography. J. Donner, *Ordnung Bdelloidea*, Akademie-Verlag, Berlin, 1965; J. Donner, *Rotifers*, Frederick Warne, London, 1966; L. H. Hyman, *The Invertebrates*, McGraw Hill, New York, 1951; J. Maynard Smith, Contemplating life without sex, *Nature*, 324:300–301, 1986; R. L. Wallace and T. W. Snell, Rotifera, pp. 187–248, in *Ecology and Classification of North American Freshwater Invertebrates*, ed. by J. H. Thorp and A. P. Covich, Academic, New York, 2001.

Bdellonemertini

An order of the class Enopla in the phylum Rhynchocoela. The proboscis is unarmed. The alimentary system comprises the mouth, papillate foregut, sinuous intestine without diverticula, and anus. Members are characterized by the presence of a posterior adhesive disk and dorsoventral flattening of the body, which lacks eyes, cephalic slits, and cerebral organs. The order contains the single genus *Malacobdella*, with three species ectocommensal in the mantle cavity of marine bivalves and one in the pulmonary sac of a fresh-water gastropod. *Malacobdella grossa*, in the bivalve *Zirfaea crispata*, abstracts particulate food such as planktonic organisms from the host's feeding current, collects these by a unique filtration system, and digests them in the intestine by a combination of extracellular and intracellular processes. Carbohydrases predominate in the digestive physiology, in contrast to other, carnivorous rhynchocoelans, where proteases are dominant. *See* ANOPLA; ENOPLA; HOPLONEMERTINI; RHYNCHOCOELA. J. B. Jennings

Beam

A structural member that is fabricated from metal, reinforced or prestressed concrete, wood, fiber- reinforced plastic, or other construction materials and that resists loads perpendicular to its longitudinal axis. Its length is usually much larger than its depth or width. Usually beams are of symmetric cross section; they are designed to bend in this plane of symmetry, which is also the plane of their greatest strength and stiffness. This plane coincides with the plane of the applied loads. Beams are used as primary load-carrying members in bridges and buildings.

Internal forces. The internal resisting forces developed in the plane of the cross section are known as shear forces, and those normal to the cross section are called bending moments. **Figure 1*a*** shows a steel beam cross section subjected to a bending moment. The resulting deformation is a linear rotation of the plane of the cross section from its original vertical position. The distance between a point on the original plane and the same point on the deformed plane is the strain (displacement/unit length). The inclination of the plane is the curvature of the beam. The corresponding stress (force/unit area) is related to the strain via the stress-strain relationship that is obtained by experiment for the material from which the beam is fabricated (Fig. 1*b*). Three stress distributions are shown in Fig. 1*a*: when the cross section is elastic, part of it has yielded, and the whole section is yielded. Yielding occurs in portion DBC in Fig. 1*b*. The relationship between the curvature and the internal bending moment is obtained by multiplying the stress at any point with its distance to the centroid of the cross section and integrating over the whole cross section. The resulting moment-curvature relationship (Fig. 1*c*) is nonlinear, and it attains its maximum at the plastic moment level. Under many practical conditions this is the capacity that can be utilized in design (point C). Sometimes it is necessary to curtail deformation to the elastic region, and then the maximum usable moment is the yield moment (point D). The moment-curvature relations for different materials and cross sections are obtainable in a similar manner. *See* SHEAR; STRESS AND STRAIN.

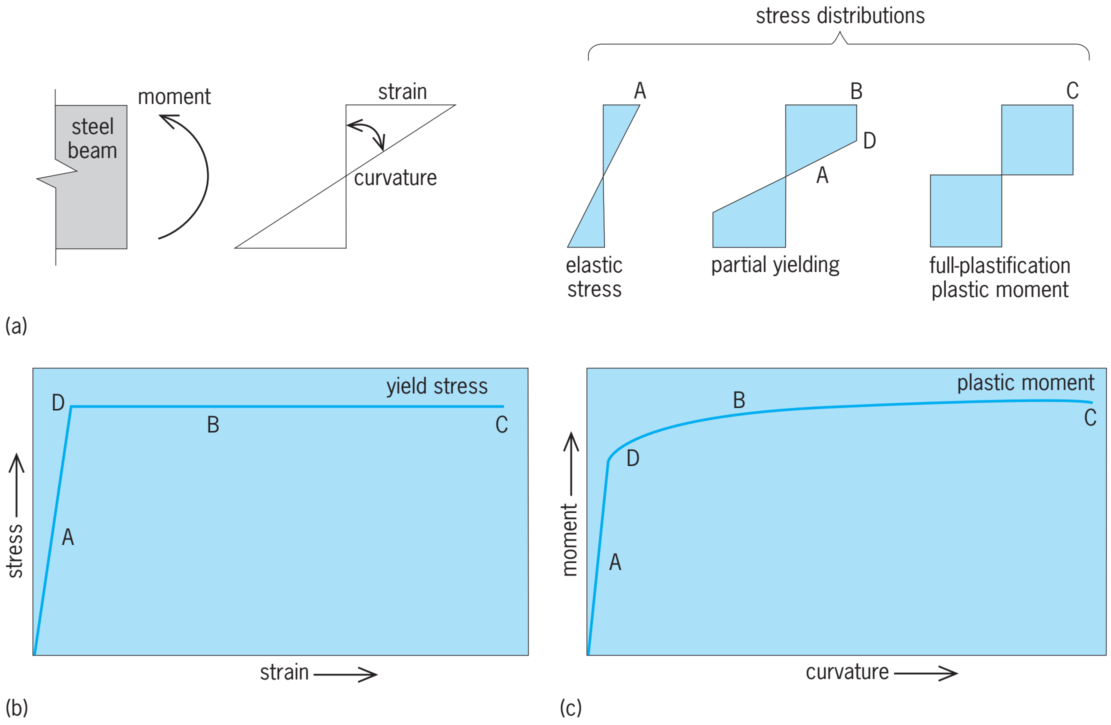


Fig. 1. Forces on a cross section of a beam. (*a*) Bending moment. (*b*) Stress-strain relationship. (*c*) Moment-curvature relationship.

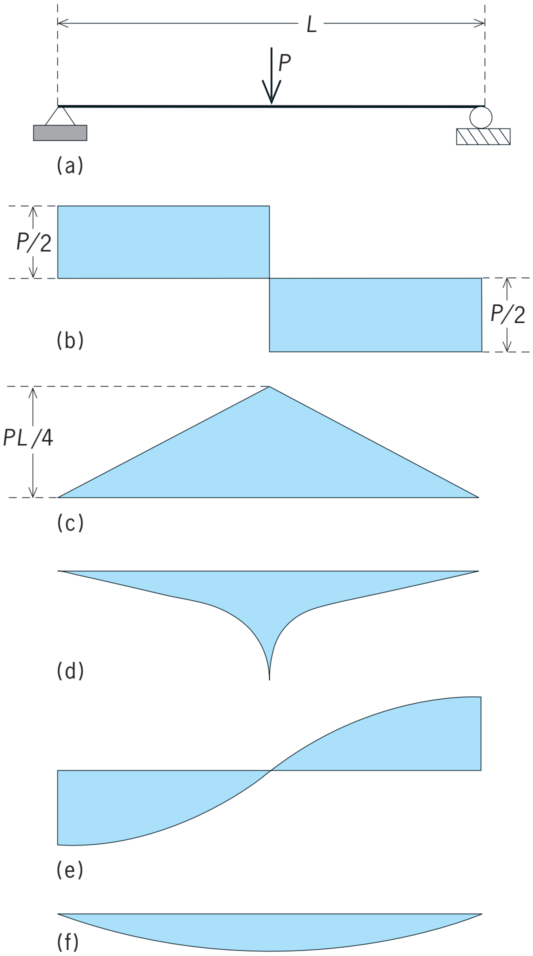


Fig. 2. Shear, moment, and deformation diagrams. (*a*) Loading. (*b*) Shear force. (*c*) Bending moment. (*d*) Curvature. (*e*) Slope. (*f*) Deflection.

Applied loads and resulting forces and deformations. In a simple beam of length *L*, a concentrated load *P* is applied at the center of the beam (**Fig. 2***a*). Such a beam has a support at one end that cannot move downward or sideways, but the beam end can freely rotate about the point of support. At the other end the beam may roll sideways also. There can be variations of the planar forces, or shears, along the length of the beam (Fig. 2*b*), or a bending moment distribution (Fig. 2*c*). The force diagrams are obtained by using the concepts of static equilibrium. Knowing the bending moment, the corresponding curvature can be obtained from the moment-curvature curve (Fig. 1*c*). The slope of the deflected beam at any location along its length (Fig. 2*e*) can then be determined from the curvature diagram by integration, since the curvature is the change in slope at any point along the beam. Similarly, the deflection (Fig. 2*f*) can be obtained by integrating the slope, since it is the change in deflection per unit length. These relationships can be found by integration downward or differentiation upward. Other load configurations can be handled in a similar manner. For beams that are not simply supported (clamped ends or multispan beams), additional conditions of compatibility of slopes and deflections at the supports must be introduced to obtain a solution. *See* LOADS, TRANSVERSE; STATICS; STRUCTURAL ANALYSIS.

Limiting conditions. Once the forces and deformations of the beam are known, they can be compared to the corresponding capacities of the beam. These are usually the shear strength, the bending strength, or the buckling strength. If these are close to the values that are imposed by the loading, the design is satisfactory. If the demand is much below the capacity or if the capacity is exceeded, a new beam must be tried and checked. It is most economical when the plastic moment can be used as the capacity. This can be achieved, in the case of metal or composite beams, if they are braced to prevent lateral buckling, or if stiffeners are provided to prevent local buckling of the plate components. The preferred failure of concrete beams is yielding of the reinforcing steel. To account for unavoidable uncertainties in the loads and capacities, load factors (larger than unity) and resistance factors (less than unity) are utilized.

Theodore V. Galambos

Bibliography. American Institute of Steel Construction, *Manual of Steel Construction: Load and Resistance Factor Design*, 2d ed., 1994; T. V. Galambos, F. J. Lin, and B. G. Johnston, *Basic Steel Design with LRFD*, Prentice Hall, 1996; R. C. Hibbeler, *Structural Analysis*, 4th ed., Prentice Hall, 1999; J. G. MacGregor, *Reinforced Concrete: Mechanics and Design*, 3d ed., Prentice Hall, 1997.

Beam column

A structural member that is subjected to axial compression and transverse bending at the same time. The combined compression and bending may be produced by an eccentrically applied axial load (**Fig. 1***a*), a concentrically applied axial load and an end moment (Fig. 1*b*), or an axial load and transverse load between the two ends. (Fig. 1*c*). A beam column differs from a column only by the presence of the

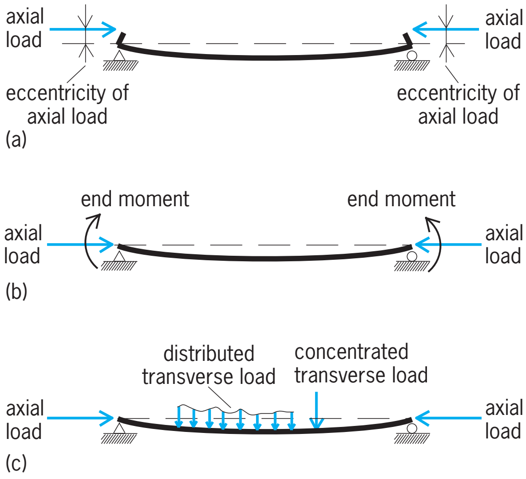

axial load
eccentricity of axial load
axial load
eccentricity of axial load
(a)
end moment
end moment
axial load
axial load
(b)
distributed transverse load
concentrated transverse load
axial load
axial load
(c)

Fig. 1. Loadings that create beam-column effects. (*a*) Eccentrically applied axial load. (*b*) Concentrically applied axial load and end moment. (*c*) Axial load and transverse load or loads between the two ends.

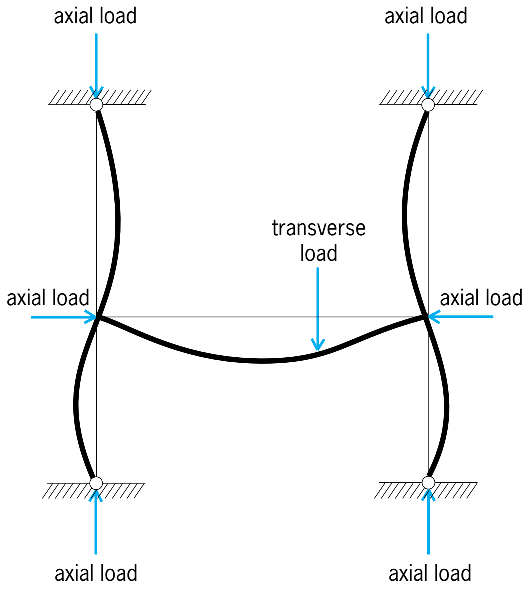


Fig. 2. Frame-type structure with one beam and four columns acting as beam columns.

eccentricity of the load application, end moment, or transverse load.

Beam columns are found in frame-type structures where the columns are subjected to other than pure concentric axial loads and axial deformations, and where the beams are subjected to axial loads in addition to transverse loads and flexural deformations (**Fig. 2**).

The other-than-axial loads create bending of the column immediately upon their application. Because of this initial bending and the consequent flexural stresses in addition to the pure axial stresses, a beam column fails at a lower axial force than a column of identical dimensions and end-support conditions.

The mathematical formulas representing bending deformation and variation of stress along the length of a beam column are generally more complex than those for a column. Structural design codes and specifications present simplified methods for the design of members under combined compression and bending. These methods take the forms of interaction formulas and interaction curves that limit the magnitudes of axial compression and transverse bending when the two are acting in tandem.

One form of the interaction formulas, which assures that the beam column's strength under axial load and bending is not exceeded, is given by the expression below, where P_u = ultimate axial load with-

$$\frac{P}{P_u} + \frac{M}{M_u} \leq 1$$

out the presence of bending, M_u = ultimate bending moment without the presence of the axial load, P = actual axial load when applied together with the bending moment, and M = actual bending moment when applied together with the axial load. The expressions for P_u and M_u can be quite complex, depending on the material, cross-sectional shape, slenderness, and types of end supports of the beam column. *See* BEAM; COLUMN; STRESS AND STRAIN; STRUCTURAL DESIGN.

Robert T. Ratay

Bibliography. E. H. Gaylord and C. N. Gaylord, *Design of Steel Structures*, 3d ed., 1992; A. H. Nilson and G. Winter, *Design of Concrete Structures*, 12th ed., 1997.

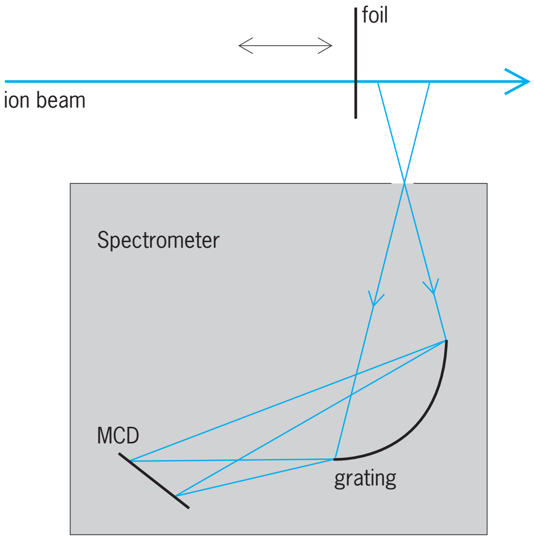
Beam-foil spectroscopy

A technique used in atomic physics to study the structure and dynamics of atomic ions of any element in any state of ionization. For this purpose, a beam of fast ions is sent through a very thin foil. The ion-foil interaction shakes up the electronic shells of the projectile and, after leaving the foil, the ions shed surplus energy by emitting photons, and sometimes electrons. The energies and intensities of these particles yield spectral information on the projectile.

Beam-foil spectroscopy is a late offspring of the canal-ray studies of the early twentieth century. The technique became possible only after the advent of accelerators. Such machines were originally developed to produce and accelerate charged atomic particles as projectiles in nuclear reactions. The target material is usually prepared as a thin foil—typically a few thousandths of the thickness of writing paper (a few hundred atomic layers thick)—through which the ions pass without significant energy loss. Most of the projectiles, however, do not undergo nuclear reactions, but collide only with the many target electrons (and therefore largely maintain their original speed and direction) and suffer a shake-up of their own electron shells. Excited states can be studied by monitoring the photons and electrons that the fast projectiles emit in deexcitation after their passage through the foil. *See* NUCLEAR REACTION; PARTICLE ACCELERATOR.

Exit beam composition. The multitude of collisions inside the foil changes the complement of electrons that travel with the projectile ion; some are ejected and others are captured from the target atoms. The ion beam therefore has a different charge-state composition after passage through the foil. (Higher exit charge states are produced at higher incident beam energies.) The beam-foil interaction efficiently populates atomic levels with a high degree of excitation such as multiply excited and core-excited levels. Such levels are rare in conventional light sources. The richness of the resulting spectra yields a great deal of information on atoms and ions, but it also causes problems. For example, it is often difficult to resolve the details of such line-rich spectra that reflect the complexity of multiply excited systems.

Timing measurements. The ion beam travels in a high vacuum before and after transiting the target foil. This environment minimizes collisional perturbation of the ions. The sudden termination of the ion-foil interaction provides an inherently good time resolution to beam-foil spectroscopy. This property of the source permits measurements of lifetimes of



Schematic of a beam-foil spectrometer equipped with a multichannel detector (MCD). Light emitted by the ions after traversing the foil is incident on the grating and MCD. The spectrometer permits simultaneous spectral coverage of a selected wavelength range.

atomic levels populated during the beam-foil interaction as well as the observation of coherent-excitation phenomena such as quantum beats. Because the ion velocity is constant and measurable, it is sufficient to trace the change in intensity of the fluorescence from the ion beam as a function of distance from the foil in order to determine atomic level lifetimes (see **illus.**). At an ion energy of 0.5 MeV/nucleon, the ions travel 1 cm in 1 nanosecond (10^{-9} s). An easily measured mechanical displacement of 10 micrometers then corresponds to a time interval of only 1 picosecond (10^{-12} s). This "magnification" allows short time intervals to be measured without any need for fast electronic timing. The typical size of a high-vacuum experimental chamber sets a practical upper limit of about 100 ns. An extension of the beam-foil technique, however, uses a heavy-ion storage ring to measure atomic lifetimes in the millisecond range. *See* FLUORESCENCE.

Intensity decay curves are single-exponential if a single level is excited. A laser, which is highly selective, can excite a single level. Laser excitation is, however, not generally applicable to highly charged ions, since excitation energies are large. For many levels within the higher ionization stages, nonselective beam-foil excitation is one of the few methods available. In nonselective excitation, often a level of interest may be repopulated by the decay from one or more higher-lying excited levels that have also been populated. However, if all of these repopulating cascades are measured as well, the time constant of the decay of the level of interest can be obtained with precision (1% or better) by a correlated analysis. If the cascade information is incomplete, the precision of the desired atomic lifetime can be 10% or more. *See* LASER.

Metastable states. The properties of highly charged ions can be predicted by scaling the known properties of low-charge ions with the same number of electrons (that is, members of the same isoelectronic sequence). With increasing ion charge, the wavelengths of transitions and lifetimes of excited states become shorter. In most highly charged ions the level lifetimes become so short that the postfoil fluorescence cannot be measured with sufficient time resolution. An exception is the decay of the much longer lived metastable states. For example, transitions involving a change of spin or higher-order multipole radiation only decay at a much slower rate. These unusual transitions, however, are of fundamental interest in understanding atomic structure, and are often used in the diagnostics of terrestrial plasmas, such as those in fusion research, or in astrophysical observations. For example, the ultraviolet part of the spectrum of the solar corona shows many lines that, for a long time, had not been observed on Earth. In sources such as laser-produced and tokamak plasmas, where they might be expected to appear, they were probably collisionally quenched. A number of such lines were eventually identified, however, by beam-foil spectroscopy. The spectra were studied at a distance of a few centimeters from the foil (or after delay times of a few nanoseconds). After such a delay, the fast decays have died out and the weak decays associated with long-lived or metastable levels become apparent. *See* ISOELECTRONIC SEQUENCE; METASTABLE STATE; PLASMA (PHYSICS); PLASMA DIAGNOSTICS; SOLAR CORONA; SUN.

Fast-beam spectroscopy. Beam-foil spectroscopy has developed into many variants which now go under the name of fast-beam spectroscopy. For example, a gas target may be used, a laser, a combination of gas or foil and laser, or a target of free electrons in a heavy-ion storage ring. The ion-foil interaction is capable of producing all ionization stages of all elements from negative ions to U^{91+}. The production of the highest ionization stages, however, requires a beam energy of about 500 MeV/nucleon, which can be reached only at the most energetic accelerators. However, since only the relative motion of electrons and ions is important, the same degree of ionization can nowadays be reached by use of 250-keV electrons in an electron-beam ion trap (EBIT). The device offers easier ways to attain high spectroscopic precision because the ions are practically at rest. In beam-foil spectroscopy the ions are rapidly moving, which shifts and broadens the spectral lines. This, in turn, causes problems in wavelength calibration and spectral resolution. However, the inherent time resolution of the foil-excited fast-ion-beam source is unique and remains a great asset in time-resolved spectroscopic measurements. *See* ATOMIC STRUCTURE AND SPECTRA; ION SOURCES; SPECTROSCOPY.

Elmar Träbert

Bibliography. G. W. F. Drake (ed.), *Atomic, Molecular, and Optical Physics Reference Book*, AIP Press, Woodbury, 1996; S. M. Shafroth and J. C. Austin (eds.), *Accelerator-Based Atomic Physics: Techniques and Applications*, AIP Press, Washington, DC, 1997.

Bean

Any of several leguminous plants, or their seeds, long utilized as food by humans or livestock. Some 14 genera of the legume family contain species producing seeds termed beans and useful to humans. Twenty-eight species in 7 genera produce beans of commercial importance, implying that the bean can be found in trade at the village level or up to and including transoceanic commerce.

Varieties. The principal Asiatic beans include the edible soybeans, *Glycine* sp., and several species of the genus *Vigna*, such as the cowpea (*V. unguiculata*) and mung, gram, rice, and adzuki beans. The broad bean (*Vicia faba*) is found in Europe, the Middle East, and Mediterranean region, including the North African fringe. Farther south in Africa occur *Phaseolus* beans, of the *vulgaris* (common bean) and *coccineus* (scarlet runner) species. Some *Phaseolus* beans occur in Europe also. The cowpea, used as a dry bean, is also found abundantly in Nigeria. *See* COWPEA; SOYBEAN.

In the Americas, *P. vulgaris* and *P. lunatus* (lima bean) are the principal edible beans, although the blackeye cowpea, mung bean, and chick pea or garbanzo (*Cicer arietinum*) are grown to some extent. *Phaseolus coccineus* is often grown in higher elevations in Central and South America, as is *Vicia faba*. The tepary bean (*P. acutifolius*) is found in the drier southwestern United States and northern Mexico. *See* ROSALES.

Characteristics. Bean plants may be either bush or vining types, early or late in maturity, with white, yellow, red, or purple flowers. The seed itself is the most differentiating characteristic of bean plants. It may be white, yellow, black, red, tan, cream-colored, or mottled, and range in weight from 0.0045 to over 0.025 oz (125 to over 700 mg) per seed. Seeds are grown in straight or curved pods (fruit), with 2–3 seeds per pod in *Glycine* to 18–20 in some *Vigna*. Pods vary in fiber content; the immature pods of low-fiber types are often used as a green vegetable. Beans are normally self-pollinated, although a small amount of natural crossing may occur. Scarlet runner bean is normally highly cross-pollinated.

Production. Most navy beans in the United States are produced in Michigan; great northern beans in Nebraska, Wyoming, and Idaho; pinto beans in Colorado, North Dakota, and Idaho; red kidneys in New York, Michigan, and California; and large and baby limas in California. Such regional distribution of bean classes is partly historical and partly due to climatic and disease factors.

Culture. Most beans are a temperate-season crop, with major producing areas in the United States having a mean August temperature of 70°F (21°C) or less and a mean July temperature of between 72°F (22°C) and 74°F (23°C), except for the central valley of California. With higher temperatures there is extensive flower and pod abscission, except for those species such as *P. acutifolius* and *Vigna* spp. that require the higher temperature for seed development. Flower initiation in beans may be affected by the length of day, there being long-day-, day-neutral-, and short-day-requiring types. Many types suffer a delay in flowering with lower mean temperature. *See* PHOTOPERIODISM.

Beans are grown on a wide variety of soil types, the most productive being the well-drained loams, silt loams, and clay loams abundantly supplied with organic matter. It is important that the soils not be compacted. Sandy soils are satisfactory where rainfall is sufficient and well distributed throughout the season. Beans grow best in soils of pH 6.5–7.2 that are also high in phosphorus and potassium. Zinc and manganese are minor elements that may be the first to become limiting in the lake bed or outwash soils of the United States. Except where a special nutrient deficiency exists, beans do not respond well to added fertilizer. *See* FERTILIZING.

Beans are normally planted after danger of frost is past and when soil temperature has reached 65°F (18°C). Inoculation with nitrogen-fixing bacteria (*Rhizobium* spp.) is generally unnecessary since most bean soils are naturally supplied with nodulating bacteria. Beans under irrigated or humid conditions are usually grown in rows 20 to 28 in. (50 to 70 cm) apart, with plants 2–4 in. (5–10 cm) apart in the row. Under dryland conditions the spacing is greater. In much of the tropics a considerable portion of the production is from beans grown in association with another crop, such as maize.

Weed control. One of the main cultural requirements for successful bean growing is weed control. Cultivation by hand or by mechanical means is the time-honored method, but selective chemical herbicides are available which, under favorable temperature and soil moisture conditions, are capable of giving almost 100% control of most annual grasses and broadleaved weeds. Combinations of a preplant herbicide with a preemergence herbicide are capable of excellent long season weed control with a high safety factor for the bean plant. However, perennial weeds are still a serious problem in bean fields. Plant diseases continue to be a major hazard in production, although progress toward disease control by developing resistant varieties is being made. *See* HERBICIDE.

Harvesting. Most bean crops will mature in 80 to 120 days; in higher elevations or cooler seasons they may take up to 150 days. Beans are ready for harvest when the pods and upper portion of the plants have changed color from green to tan or brown and seeds have dried to under 20% moisture. Bean seeds split or break easily when threshed mechanically at under 14% seed moisture. The ripened plants are pulled from the soil by hand or by a machine called a puller which has two sets of heavy steel knives set at an angle to the row and positioned to intercept the roots about 2 in. (5 cm) below the soil surface. The pulled plants are formed into windrows and cured for a few hours in bright clear weather before threshing with a grain combine or special bean harvester. In some countries all the harvesting operations are carried out by hand, and winnowing of the seed is still practiced. A product of high physical quality usually results from this process. In the United States field-run beans require screening, and, for highest quality, picking, that is, the removal of seed-sized foreign matter and off-colored, stained or damaged seeds. This was formerly done by hand but presently is accomplished by electric-eye machines in the cleaning plants or elevators. *See* AGRICULTURAL MACHINERY.

Utilization. Beans are consumed as food in several forms. Lima beans and snap beans are used by canning or freezing. Limas are also used as a dry bean. Mung beans are utilized as sprouts.

Usage of dry beans (*P. vulgaris*) for food is highly dependent upon seed size, shape, color, and flavor characteristics, and is often associated with particular social or ethnic groups. Popular usage includes soups, mixed-bean salads, pork and beans, beans boiled with meat or other vegetables or cereals, baked beans, precooked beans, and powder, and, in the Peruvian Andes, parched or roasted beans. Per capita consumption of dry beans in the United States is 6.7 lb/year (3 kg/year).

Dry beans range in protein content from 17 to 32%. Commercial types average around 23–25%. The sulfur-containing amino acids methionine and cystine are the first limiting factors in bean protein, from a nutritional standpoint. The dry beans have less than 2% oil (except soybeans) and some 61–63% carbohydrate, and are rich sources of fiber, calcium, and phosphorus. Because of the presence of heat-labile growth retardants, beans must be thoroughly cooked before being used as food. Newer developments are in the areas of cooking time, the manufacture of precooked dry powders that can be readily rehydrated to make a palatable soup, and new research impetus to identify the flatulence factors in the dry beans. *See* AGRICULTURAL SOIL AND CROP PRACTICES; BREEDING (PLANT); LEGUME. M. W. Adams

Origin and evolution. Molecular techniques that allow analysis of specific compounds such as seed storage proteins, enzymes, or deoxyribonucleic acid sequences have helped to clarify the origin of various bean species. Specific variants of these molecules are identified in the cultivars, and by tracing these variants among the existing wild populations of the crop species the actual progenitor populations of the crop can be identified and a domestication area can be proposed. It has been determined, for example, that the common bean (*Phaseolus*) was domesticated in at least three areas, Mexico, Central America, and the Andes mountains in South America. These domestications gave rise to two groups of cultivars. A group with small or medium-sized seeds originated in Middle America (Mexico and Central America) and is represented by navy, pinto, great northern, and black beans. The other group, with large seeds, came from the Andes and includes the kidney, yelloweye, cranberry, and snap beans.

Subsequent to the arrival of Columbus in the Americas, both cultivar groups were dispersed worldwide. The Middle American group was introduced into the southwestern and Great Plains areas of the United States. The Andean group was introduced into Europe, from where it was further disseminated to North America and possibly Africa. The lima bean resulted from a similar domestication pattern, leading to a small-seeded, Middle American and a large-seeded, Andean cultivated group. The cowpea was domesticated in Africa but was introduced as early as 1000 B.C. into the Indian subcontinent, where additional cultivar types were selected. During the slave trade the cowpea was introduced into the United States and Brazil. Soybean originated in northern China, and spread to Japan, Korea, and southeastern Asia. It is only in the last century that the cowpea has actively been grown outside Asia, mainly because it is an important source of oil and protein.

Molecular analyses show that domestication has induced a reduction in genetic diversity in beans. Wild progenitors can therefore be a source of useful traits. For example, the arcelin seed protein was identified in wild common bean from Mexico. This protein confers resistance to seed weevils and has now been introduced into bean cultivars. P. Gepts

Diseases. Beans are affected by nonparasitic and parasitic diseases. The latter are more widespread and serious and are caused by fungi, bacteria, and viruses. The former result from unfavorable environment, including improper soil-mineral relationships and impure air. Although there are more than 40 known diseases which affect field-grown beans, only half of them occur consistently and of these about 9 are considered important on a worldwide basis. Some common diseases kill young seedlings, others cause injury or death of older plants, and some produce spots on pods and seeds that make the product unsalable.

Bacterial diseases. Bacteria-induced bean diseases are especially important where rain falls frequently during the growing season, or where overhead irrigation is compulsory. The blight bacteria, *Xanthomonas phaseoli* and *Pseudomonas phaseolicola*, are seedborne. Symptoms of blight appear most conspicuously on leaves as water-soaked spots which may

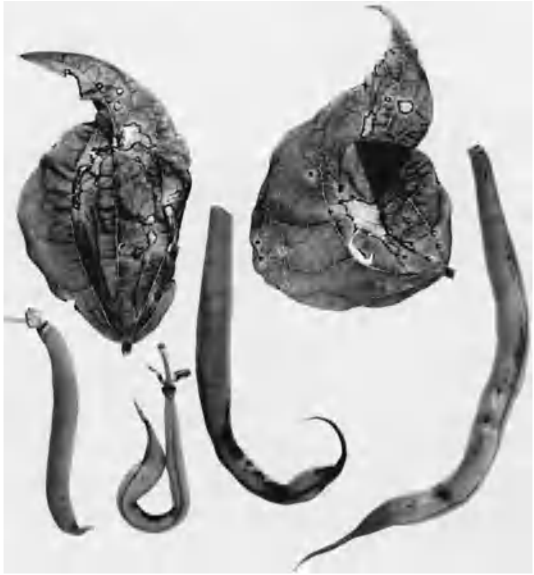


Fig. 1. Bacterial brown spot on leaves and pods of bean.

later merge to form brown, dead areas. The diseased areas caused by *X. phaseoli* often have narrow yellow margins; those caused by *P. phaseolicola* have large yellow halos. Diseased pods also develop water-soaked spots which later turn brown and form sunken lesions with reddish-brown margins. Bacterial brown spot is caused by *P. syringae*, which incites small, brown lesions on both leaves and pods (**Fig. 1**). The centers of leaf lesions fall out, producing the "shot-hole" symptom. Diseased pods are often strikingly misshapen. The bacterial wilt pathogen, *Corynebacterium flaccumfaciens*, causes yellowing, wilting, and shriveling of infected plants. Seedlings produced by infected seeds may be stunted and may not emerge from the soil.

Fungal diseases. The most important fungal diseases of beans are rust, root rots, and anthracnose. Rust, caused by *Uromyces phaseoli* f. sp. *typica*, primarily attacks leaves and produces numerous small, rust-colored pustules or blisters which may lead to premature leaf drying and defoliation. The spores formed within the blisters may be spread by wind for hundreds of miles. Root rots are caused by several fungi, including *Fusarium solani* f. sp. *phaseoli, Pythium* ssp., and *Rhizoctonia solani*. All produce reddish-brown lesions which often coalesce to form large rotted areas on the roots, resulting in stunting of diseased plants. Extensive root decay (**Fig. 2**) causes death of the plant. Lesions caused by *R. solani* are often confined in size but are characteristically sunken. If moisture and nutrition are adequate, root-rot–diseased plants may respond by production of many new fibrous roots just beneath the soil; these new roots maintain the weakened plant. Anthracnose, induced by *Colletotrichum lindemuthianum*, causes dark, sunken cankers on pods (**Fig. 3**) and angular dead spots on leaves. The pathogen is spread by infected seed and by wind, rain, and insects.

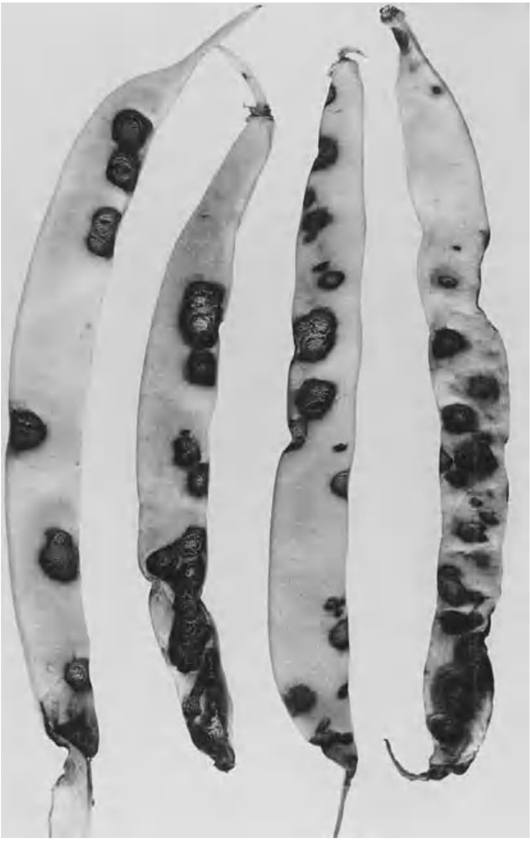


Fig. 3. Anthracnose on bean pods (fruits).

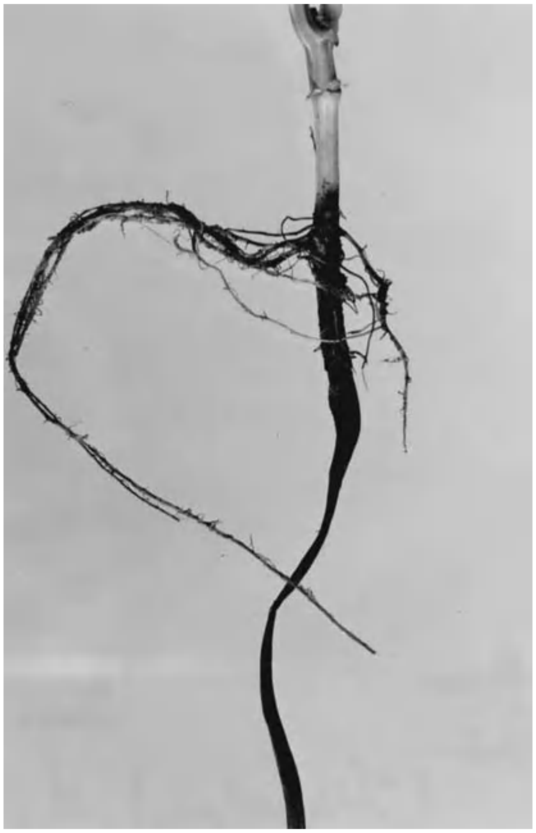


Fig. 2. Root rot of a bean plant.

Viral diseases. Bean diseases caused by viruses are troublesome throughout the world and are often serious locally. They are manifested by leaf mottling and distortion and plant stunt, malformation, and sometimes death. The important bean common mosaic virus is aphid- and seed-borne. Yellow mosaic virus is aphid-transmitted. The curly top virus is carried by leafhoppers. The golden mosaic virus of Latin America is spread by white flies. *See* PLANT VIRUSES AND VIROIDS.

Disease control. The most desirable control for bean diseases is the use of resistant cultivars. This will be possible for almost all of the diseases mentioned above in the near future. At present, there are many bean cultivars resistant to rust and to common bean mosaic. Use of disease-free seed produced in the semiarid western states is another inexpensive and practical control measure of major significance in the United States. Crop rotation and field sanitation should also be practiced. Treating seeds with chemical protectants and proper spraying of plants with safe, effective chemicals are also very useful control measures. *See* PLANT PATHOLOGY. D. J. Hagedorn

Bibliography. R. Hall (ed.), *Compendium of Bean Diseases*, 1991; R. J. Summerfield and E. H. Roberts (eds.), *Grain Legume Crops*, 1985.

Bear

The bear family (Ursidae) contains the largest terrestrial carnivores. Most taxonomists agree that there are eight extant species of bears in three subfamilies: Ursinae (American black, Asiatic black, brown, polar, Malayan Sun, and sloth bears); Tremarctinae (Andean or spectacled bear); and Ailuropodinae (giant panda). Bears currently inhabit North America, South America, Europe, and Asia. Habitats range from polar ice floes to tropical forests. They are absent from Australia, Africa, and Antarctica.

General characteristics. The fur of bears is long, shaggy, and generally unicolored brown, black, or white, although some genera possess a prominent crescent-shaped white mark on their chests. *Tremarctos*, the spectacled bear of South America, typically has a patch of white hairs encircling each eye. *Ailuropoda*, the giant panda, has a striking black-and-white color pattern. The head and body length of adults ranges between 1000 and 2800 mm (39–109 in.) with a tail length between 65 and 280 mm (3–8 in.). Adults weigh between 27 kg (60 lb) (Malayan sun bear, *Ursus malayanus*) and 775 kg (1700 lb) (Alaska brown bear, *Ursus arctos*). Males average about 20% larger than females.

Bears have a heavily built body with short, powerful limbs and a short tail. The head is large, the eyes are small, and the ears are small, rounded, and erect. Five toes are present on each foot, and each has a long, strong, moderately curved, nonretractile claw that is used for tearing and digging. The soles of the feet are hairy in species that are mainly terrestrial but naked in species such as the sun bear, which spends a great deal of time climbing. Locomotion is plantigrade (both heel and toe contact the ground). Most bears possess 42 teeth (I 3/3, C 1/1, PM 4/4, M 2/3 × 2). The sloth bear, as an adaptation for sucking up termites, has only two upper incisors and a total of 40 teeth. *See* DENTITION.

Bears do not have particularly good eyesight, but their senses of hearing and smell are excellent. They rely on their keen sense of smell, color vision, and a good memory to locate food sources. Most bears are omnivorous with the exception of the polar bear, which feeds mainly on seals and fish. Bears are among the few mammals able to survive for half a year or more without eating, drinking, urinating, or defecating. In many regions, bears are known to enter dormancy when cold weather approaches. They become fat during autumn, but with the approach of cold weather they cease eating, go into a den in a protected location, and sleep through the winter, living mainly off stored fat reserves. Dens may be in hollow trees, in excavated depressions in the ground, or in caves. Their bones do not deteriorate despite complete inactivity, and fluids and amino acids are derived by recycling products that are normally excreted as urine. Although the heart rate and respiration rate are lower than when they are active, their metabolic processes do not drop as dramatically as in a true hibernating mammal such as a woodchuck. Therefore, there is disagreement as to whether bears should be classified as true hibernators. They can be easily aroused during the winter and may awaken on their own during periods of mild weather. *See* HIBERNATION AND ESTIVATION.

Bears are normally solitary animals, with the exception of females accompanied by their young. Litters of 1 to 6 young are produced at intervals of 1 to 4 years. Births occur while the mother is in her den during the winter. Although mating occurs in late spring or summer, the fertilized egg does not implant in the uterus immediately (delayed implantation). Thus, the period of pregnancy usually extends 6 to 9 months. The young are relatively tiny at birth, ranging from 225 to 680 g (8–24 oz). They are smaller in proportion to the size of the mother (less than 1%) than is the case with any other placental mammal. They remain with their mother at least through the first autumn. They become sexually mature between 2.5 and 6 years of age and normally live 15–30 years in the wild. Populations of many species are affected by excessive killing, habitat loss, and population fragmentation from road building.

American black bear (Ursus americanus). The most numerous of all bears, this inhabits virtually all of the forested areas of North America from northern Canada to central Mexico. This species is hunted as a game animal in many areas, with over 40,000 being killed annually in North America. Black bears are extremely adaptable, a characteristic which has aided them in becoming tolerant of human presence. Although most have long, shaggy, black fur, many are brown (**Fig. 1**). Black bears are distinguished from grizzly bears by the straight (versus dished) profile of their face and by their lack of a shoulder hump. They also have smaller claws, which are adapted more to climbing than digging. Adults are 1.2–1.9 m (4–6.2 ft) in length and stand 0.7–1.0 m (2.3–3.3 ft) at the shoulder. Males usually weigh 60–225 kg (130–500 lb), females 40–150 kg (90–330 lb). Mating usually occurs from May to July with young being born in January. Vegetable matter such as fruits, berries, nuts, acorns, grass, and roots constitutes at least 75% of the diet. Other foods may include insects, fish, rodents, carrion, and occasionally, young ungulates.

Asiatic black bear (Ursus thibetanus). This bear is found from Iran and the Himalayas to Japan, China, and eastern Siberia. The coat is black with a white V-shaped marking on the chest. The chin is white. A ruff of longer hair is present on the neck. Adults are 1.2–1.9 m (3.9–6.2 ft) in total length and stand 0.7–1.0 m (2.3–3.3 ft) at the shoulder. Males weigh 60–200 kg (130–440 lbs), females 40–140 kg (90–310 lb). This species is mainly herbivorous. Asiatic bears have been severely affected by the commercial trade in bear parts; The meat, fat, and paws are popular foods in these countries. The bones, brain, blood, spinal cord, and gallbladder are sought for medicinal uses in Chinese medicine.

-s

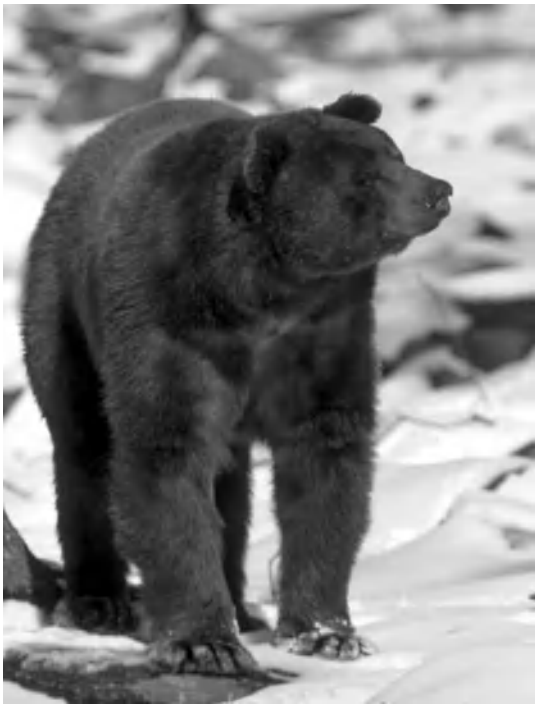


Fig. 1. Black bear (*Ursus americanus*). (*Photo by Tom Brakefield; © 1999 California Academy of Sciences*)

Brown bear (Ursus arctos). Also known as the grizzly or Kodiak bear, this species is both the most widespread bear and the largest living carnivore. It ranges from northwestern North America and Scandinavia through Russia to Japan. Scattered populations also exist in parts of Europe, the middle east, the Himalayas, China, and Mongolia. These bears are highly adaptable and can live in forests, tundra, and deserts. The pelage (coat, including all fur and hairs) is uniform brown or blonde to near black. Adults are 1.5–2.8 m (5.0–9.2 ft) in total length and stand 0.9–1.5 m (3–5.0 ft) at the shoulder. Males weigh 135–545 kg (300–1200 lb), females 80–250 kg (175–550 lb). Many populations are very small and in danger of extirpation. For example, in the United States outside of Alaska, only five isolated populations survive, totaling about 1000 animals. These populations are on the Endangered Species list.

Polar bear (Ursus maritimus). Polar bears are among the largest of the world's bear species. Despite their different appearance, polar bears and brown bears are closely related. The polar bear's body is stocky but lacks the shoulder hump of the brown bear. The neck is longer in relation to the rest of the body than in other bears, and the large, oarlike forepaws are adapted for swimming (**Fig. 2**). The ears are small, the fur is whitish to yellowish, and the skin beneath the coat is black. Polar bears mainly inhabit arctic regions around the North Pole and the coastlines of continents and islands where there are high densities of ringed seals, their primary prey. Adults are 180–250 cm (5.8–8.3 ft) in total length. Males weigh 400–600 kg (880–1320 lb), females 200–350 kg (440–770 lb). Polar bears are not endangered, but are threatened by increasing pollution and global warming. The Agreement on the Conservation of Polar Bears (1976) requires that this species be managed according to "sound conservation practices," and mandates protection of their habitat.

Sloth bear (Ursus ursinus). This bear inhabits India, Nepal, Bhutan, Sri Lanka, and possibly Bangladesh. The coat is long, shaggy, and black with a white U on the chest. Termites are the primary food, but other insects, honey, eggs, carrion, and berries are also eaten. Adaptations for its unusual diet include long, slightly curved claws; large, protrusible lips; a broad, hollowed palate for sucking; and the absence of two front upper incisors. Sloth bears are largely nocturnal, and females frequently carry their young on their back, two behaviors that are unique among bears. Sloth bears are 1.4–1.9 m (4.6–6.2 ft) in total length and stand 60–90 cm (2–3 ft) at the shoulder. Adult males weigh 80–145 kg (176–320 lb), adult females 55–95 kg (120–210 lb).

Malayan sun bear (Ursus malayanus). These are the smallest of the bears and have the shortest hair. They inhabit the dense forests of southeast Asia including Sumatra, Borneo, eastern India, and southern China. The coat is black with an ocher or white circle or semicircle on the chest. Both their snout and ears are short. These nocturnal bears are highly arboreal and often sleep in nests in trees. They are omnivorous, feeding primarily on insects and fruit. Adults are 1.1–1.5 m (3.6–4.9 ft) in total length and stand approximately 70 cm (2.3 ft) at the shoulder. Adults weigh 27–65 kg (60–143 lb).

Andean bear (Tremarctos ornatus). Also known as the spectacled bear because of the light markings that often encircle the eyes, these bears are the only South American ursid and are the second largest land mammal in South America. They inhabit the

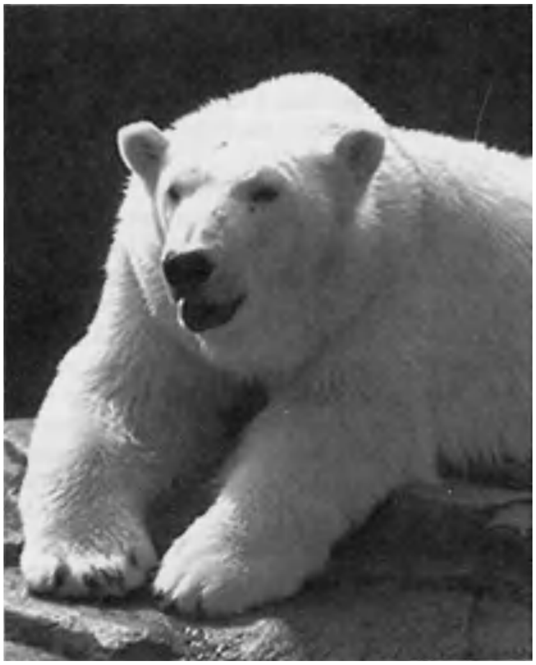


Fig. 2. Polar bear (*Ursus maritimus*).

mountainous regions from eastern Panama to northern Argentina. They may be found from high altitudes to coastal lowlands and are found in a wide variety of habitats including cloud forests, rainforests, high-altitude grasslands, and scrub deserts. Together with Malayan sun bears, these are the only bears that live on both sides of the equator. The coat is black with a creamy white, biblike marking on the chin, neck, and/or chest. These relatively slim and agile bears build tree nests for sleeping and feed primarily on plant matter, especially bromeliads. Adults are 1.3–2.0 m (4.3–6.6 ft) in total length and 70–90 cm (2.3–3.0 ft) at the shoulder. Males weigh 100–175 kg (220–385 lb), females 60–80 kg (132–176 lb).

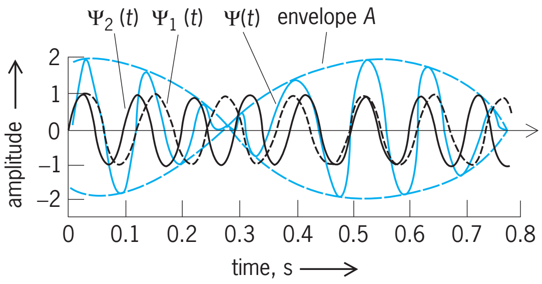
Giant panda (Ailuropoda melanoleuca). The current range of this species has been restricted to six isolated mountain ranges in central and western China. Giant pandas live in cool, damp bamboo forests at altitudes of 1500–3400 m (4900–11,000 ft). It is one of the most endangered mammals in the world; only about 1000 exist in the wild. Although formerly classified with the Procyonidae (raccoons and their relatives), recent genetic studies show that the giant panda represents an early divergence from the bear family. It differs from other bears in several ways: it does not hibernate, one of its wrist bones has evolved into a "pseudothumb"; and a newborn is remarkably small, weighing just 100–200 g, about 0.001% of its mother's weight. The ears, eye patches, legs, and shoulders are black; the remainder of the body is white. Giant pandas are 1325–1625 mm (57–70 in.) in total length, including a 120–130 mm (4–5 in.) tail. They stand 700–800 cm (2.7–3.2 ft) at the shoulder and weigh 100–150 kg (220–330 lb). They feed almost exclusively on bamboo. *See* PANDA.

Donald W. Linzey

Bibliography. G. A. Feldhamer, B. C. Thompson, and J. A. Chapman (eds.), *Wild Mammals of North America: Biology, Management, and Conservation*, 2d ed., Johns Hopkins University Press, 2003; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., John's Hopkins University Press, 1999; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

Beat

A variation in the intensity of a composite wave which is formed from two distinct waves with different frequencies. Beats were first observed in sound waves, such as those produced by two tuning forks with different frequencies. Beats also can be produced by other waves. They can occur in the motion of two pendulums of different lengths and have been observed among the different-frequency phonons in a crystal lattice. *See* CRYSTAL STRUCTURE; PENDULUM.



Production of beats. Two waves, $\psi_1(t)$ with frequency $\omega_1 =$ 8 Hz and $\omega_2(t)$ with frequency $\omega_2 =$ 10 Hz, produce composite wave $\psi(t)$ with amplitude A, an envelope that encloses ψ.

As a simple example, consider two waves of equal amplitudes A_1 and different frequencies ω_1 and ω_2 in hertz at the same spatial point given by $\psi_1(t) = A_1 \sin(\omega_1 t)$ and $\psi_2(t) = A_1 \sin(\omega_2 t)$. The sum of these waves at this point at an instantaneous time is given by Eq. (1), assuming that they are coherent. The last

$$\psi = \psi_1(t) + \psi_2(t) = A_1(\sin \omega_1 t + \sin \omega_2 t)$$

$$= \{2A_1 \cos[{}^1/_2(\omega_1 - \omega_2)t]\} \sin[{}^1/_2(\omega_1 + \omega_2)t] \quad (1)$$

form of the wave can be written as $\psi(t) = A \sin(\omega t)$, providing that the amplitude of the composite wave $\psi(t)$ is given by $A = 2A_1 \cos[\frac{1}{2}(\omega_1 - \omega_2)t]$ and the frequency is $\omega = \frac{1}{2}(\omega_1 + \omega_2)$. The **illustration** shows the case $A_1 = 1$, $\omega_1 = 8$ Hz, and $\omega_2 = 10$ Hz. The amplitude A of ψ is the envelope enclosing the curve of the total wave. The beat pulsations occur because of this low-frequency envelope.

There is a simple and interesting approximation to Eq. (1) when $\omega_1 - \omega_2 = \epsilon$ is small and short times are considered. The amplitude of $\psi(t)$ is approximately $B \simeq 2A_1 - \frac{1}{2}A_1\epsilon^2 t^2$, and the composite wave is given by Eq. (2) for sufficiently short times. Of course, the

$$\psi_{1f}(t) \approx (2A - {}^1/_2 A_1 \epsilon^2 t^2) \sin[{}^1/_2(\omega_1 + \omega_2)t] \quad (2)$$

beat phenomenon is lost in this limit. Similarly, the wave in Eq. (1) would reduce to Eq. (3), and only the

$$\psi_{bf} \approx A_1 \cos[{}^1/_2(\omega_1 - \omega_2)t] \quad (3)$$

beat frequency would appear if the sum frequency $\frac{1}{2}(\omega_1 + \omega_2)$ was too high to be observed.

The individual waves $\psi_1(t)$ and $\psi_2(t)$ could represent a model for monochromatic sound waves, the simple harmonic motion of two pendulums, two acoustic phonons in a crystal, or two electromagnetic waves at a point in space. For sound waves, the beats produce a throbbing sound at a frequency $(\omega_1 - \omega_2)/2$. In the example given in the illustration, $\omega_1 - \omega_2 = 2$ Hz, and one beat per second is produced; this is the envelope. If the difference in frequencies is greater than 12–16 Hz, the beats are difficult to distinguish from one another unless the difference frequency is audible, and, in this case, a difference tone is produced. It is also possible to produce audible summation tones. *See* ELECTROMAGNETIC RADIATION; HARMONIC MOTION; LATTICE VIBRATIONS; SOUND.

One application of beat phenomena is to use one object with accurately known frequency ω_1 to determine the unknown frequency ω_2 of another such object. The beat-frequency or heterodyne oscillator also operates by producing beats from two frequencies. *See* OSCILLATOR; WAVE MOTION. Brian De Facio

Beaver

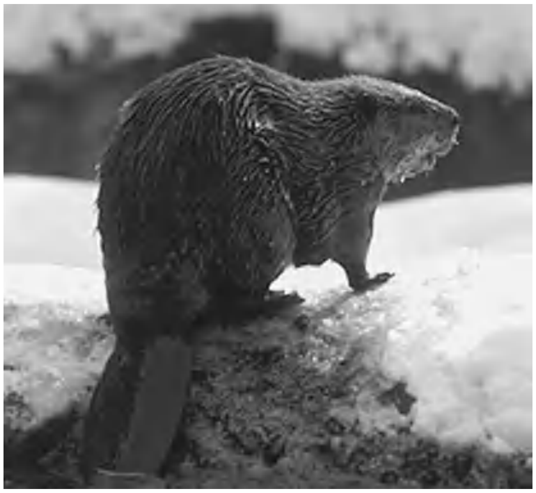
The common name for the most primitive living rodent, which has an ancestry that goes back to the time when modern mammals evolved. Two families and three species of beavers exist in the world. The family Aplodontidae contains a single species, the mountain beaver (*Aplodontia rufa*). The family Castoridae contains two species: the American beaver (*Castor canadensis*; see **illustration**) and the Eurasian beaver (*Castor fiber*).

Mountain beaver. This species is limited to the coastal conferous forests from southern British Columbia to northern California. It has no close living relatives. Its name is a misnomer as it is not particularly montane (being more often found in lowland forests, although it does occur at all elevations from sea level to the treeline), nor does it swim like American beavers.

Mountain beavers, also known as sewellels or boomers, are medium-sized rodents with blackish to reddish-brown fur on their dorsal surface and tawny hair on their ventral surface. The head is broad and flattened, the ears and eyes are small, and the vibrissae are long, stiff, and whitish. A distinctive white spot is present beneath each ear. The furred tail is short and vestigial. The senses of sight and hearing are poor, but these mammals have excellent olfactory and tactile sensations. The teeth are rootless and grow throughout life. The dental formula is I 1/1, C 0/0, PM 2/1, M 3/3 × 2 for a total of 22 teeth. Adults are 238–470 mm (9–18 in.) in total length, including a 19–55 mm (0.7–2 in.) tail. They weigh 806–1.325 g (28–46 oz). *See* DENTITION.

Mountain beavers inhabit moist forests and forest openings with a dense understory. These nocturnal, mostly solitary mammals spend most of their time in a series of underground burrows with interconnecting feeding and nesting chambers. They do not hibernate but may remain in their burrow for extended periods during cold weather. They are strictly vegetarians and eat a variety of plant foods including ferns, grasses, and roots. If necessary, they are capable of climbing trees in search of food. Mating occurs in late winter and a single litter of 2 or 3 young are born in the spring after a gestation of about 1 month. Sexual maturity is attained in the second year.

American beaver. This is the largest rodent in the Northern Hemisphere and one of the largest in the world, second only to the capybara of South America. Beavers are stocky mammals with scaly, nearly naked, horizontally flattened, paddle-shaped tails and webbed hind feet (see illustration). The dense pelage consists of fine, short underfur overlaid with long, coarse, shiny guard hairs. The coat is brown to blackish-brown, becoming slightly paler on the underparts. The tail and feet are black. The ears are short and black and are set far back on the broad, rounded head. The small ears, like the nose, are equipped with valves that shut when the animal is underwater. Although a beaver's eyesight is not exceptional, its senses of hearing, smell, and touch are excellent. Each foot has five clawed toes. The claws of the two inner toes on each hind foot have split nails and are modified as combs for grooming. Both sexes possess paired anal scent glands known as "castors." These glands are largest in males and secrete a bitter, orange-brown, pungent oil known as castoreum. This secretion is deposited on scent mounds usually near the edge of an individual's territory and appear to serve as a population regulation mechanism to warn other beavers away from occupied territory. Adult beavers are 900–1200 mm (35–47 in.) in total length, including a 225–300 mm (9–12 in.) tail. Adults weigh 16–30 kg (35–65 lb).



American beaver (*Castor canadensis*).

Beavers formerly ranged along streams and lakes throughout most of the forested regions of North America. They have been extirpated from many areas primarily by overtrapping, but with regulation of trapping and translocations, populations have increased dramatically. They have been introduced into South America, Europe, and Asia.

Beaver signs include the familiar dams, lodges, canals, and tree cuttings. Dams are constructed across slow, meandering streams that pass through flat, moist, wooded valleys. With their powerful, chiselllike incisor teeth, beavers can fell a tree 10 cm (4 in.) thick in 15 min. Lodges are dome-shaped and are formed by a haphazard piling and interlacing of thick sticks and heavier poles that are plastered together with mud. A small air hole is left at the top for ventilation. Beaver lodges usually rise 0.9 –1.5 m (3–5 ft) above the surface of the water and may be 12 m (40 ft) in diameter at the base. Entrances are located beneath the water level and a crude platform in the center of the pile is used for a place to rest. Bank burrows, also with entrances below the surface of the water, may be used in some areas.

Beavers, which are mainly nocturnal, feed primarily on the bark and adjacent outer layers (cambium) of deciduous trees such as aspen, cottonwood, birch, willow, and alder. In areas of freezing winter temperatures, beavers cache branches in the mud at the bottom of the pond as a winter food supply. Their fur-lined lips can be closed and sealed behind the incisors so that they can carry branches underwater without taking water into their lungs.

Beavers are monogamous and normally mate for life. Mating usually occurs in January and February. After a gestation of approximately 106 days, a single litter of 3 or 4 kits is born. Young beavers mature slowly. Even though they learn to swim when a month old and are weaned after 6 weeks, they remain under their parents' care for about 2 years. Thus, a typical beaver colony may consist of the parents, kits, and yearlings. Sexual maturity is usually attained during the beaver's second winter. Life span in the wild is 10–12 years, although some wild individuals have been known to live over 20 years.

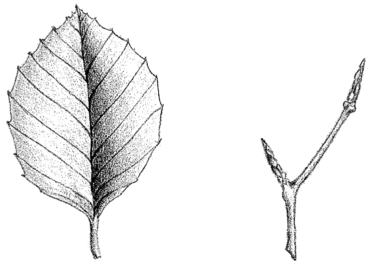
Eurasian beaver. Once found throughout Europe and Asia, this species was reduced to only eight relic populations by the early 1900s. Populations have grown and the range has expanded due to reintroductions and translocations. The description and ecology of this species is similar to *Castor canadensis.* Donald W. Linzey

Bibliography. G. A. Feldhamer, B. C. Thompson, and J. A. Chapman (eds.), *Wild Mammals of North America: Biology, Management, and Conservation*, 2d ed., Johns Hopkins University Press, 2003; D. W. Linzey, *The Mammals of Virginia* McDonald & Woodward, 1998; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

Beech

A genus, *Fagus*, of deciduous trees of the beech family, Fagaceae, order Fagales. They can best be distinguished by their long (often more than 1 in. or 2.5 cm), slender, scaly winter buds; their thin, gray bark, smooth even in old trees; and their simple, toothed, ovate or ovate-oblong, deciduous leaves. *See* FAGALES.

The American beech (*F. grandifolia*) grows to a height of 120 ft (36 m) and is native in the United States east of the Mississippi River and in the lower Mississippi Valley. The hard, strong wood is used for furniture, handles, woodenware, cooperage, and veneer. The small, edible, three-sided nuts, called beechnuts, are valuable food for wildlife. They are borne in pairs, surrounded by a spiny involucre, the homolog of the spiny covering in the chestnut bur and the cup of the acorn, which also belong to the beech family. *See* FRUIT.



European beech (*Fagus sylvatica*).

The European beech (*F. sylvatica*; see **illus.**) is more popular as an ornamental tree than the American species. Its leaves are smaller, with 5–9 pairs of primary veins compared with 9–14 pairs in the American beech. The leaf teeth are also shorter. Important ornamental varieties are *F. sylvatica purpurea*, the copper or purple beech; var. *incisa*, the cut-leaved or fern-leaved beech; and *F. pendula*, the weeping European beech. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Beef cattle production

Raising of cattle for meat. The muscle from cattle over 6 months of age is beef and from younger cattle, veal. Beef is a highly nutritious food, containing approximately 20% protein; it is rich in essential amino acids and minerals, and is a good source of the B vitamins (B_6, B_{12}, niacin, and riboflavin). Beef also supplies a generous portion of the daily dietary iron requirements.

Cattle, members of the family Bovidae, are ruminants because they have a four-compartment stomach filled with microorganisms which allow digestion of forages (grasses) and other fibrous food sources. Most domestic breeds common to the temperate zone originated from European *Bos taurus* ancestry, whereas Zebu or Brahman originated from Asian *B. indicus*, indigenous to the tropics. *See* ARTIODACTYLA; ZEBU.

World beef production. Beef cattle numbers have either stayed the same or increased slightly in almost all areas of the world since 1970. As countries improve their standards of living and become increasingly industrialized and as populations grow, there will be increased demand for beef cattle worldwide. Competition from pork and poultry products has resulted in decreased beef consumption in many developed countries. A slow decrease of beef cattle numbers is expected to continue in the United States due to a stable population growth and cost competitiveness of other food commodities.

The United States is the major beef producer. Australia, New Zealand, Argentina, and Brazil are leading beef exporters, followed by France and Denmark. Leading importing countries include the United States, Great Britain, Germany, and Japan. China has also increased beef imports markedly.

World beef production is based on the use of grass and harvested forages. Many land areas are unsuitable for growing crops but are suitable for forage production. Cattle can graze these areas and convert vegetation into high-quality protein food. The major phases of beef production involve the use of cow herds for raising calves and then calves grown out and fattened for market. Calves weigh 65–100 lb (30–45 kg) at birth and nurse until they are 6–7 months of age, when they are weaned. Under good management, weaned calves can weigh 500 lb (225 kg) at

7 months. In some countries, weaned calves are grown on grass or forages until becoming yearlings, at which time they are fattened on grain until marketed at the age of 15–24 months. The pampas of Argentina provide year-round grazing and produce slaughter beef directly from grass at 2 years of age. Many New Zealand, Australian, and Argentinean cattle are marketed from pasture alone, while many cattle in the United States, Canada, or England are grain-fed. Grain feeding allows cattle to be marketed at young ages and with a more palatable flavor. Cattle could be marketed successfully from forages alone, though it would take longer to get cattle to acceptable market weights.

United States beef production. Approximately 3% of United States beef cattle are in purebred registered herds. These cattle are identified and selected by using extensive record systems in order to take advantage of genetic improvements in cattle production. The superior purebred bulls are used by commercial producers to increase market acceptability of the calves. The rank of the six leading purebred breeds are Angus (**Fig. 1**), Hereford (**Fig. 2**), Limousin Simmental (**Fig. 3**), Polled Hereford, and Charolais (**Fig. 4**). In the southern region of the United States, Brahman (**Fig. 5**) or breeds with Brahman blood such as Santa Gertrudis, Brangus, and Beefmaster are more tolerant of humidity, heat, and insects. A systematic cross of two or three breeds is popular with producers because the hybrid vigor of mating different breeds results in lower mortality, improved growth, and reproductive efficiency. Purebred producers use measures at birth, at weaning (205 days), and at yearling to project the growth rate of cattle; they also measure milk production and fertility in producing cows. National sire summaries which statistically calculate expected progeny differences for purebred cattle provide genetic predictions of cattle performance and are used to improve accurate matings of animals by purebred producers. *See* BREEDING (ANIMAL).

Beginning in the late 1960s, many commercial producers seeking greater size and other improvements included one or more of the European breeds of Charolais, Simmental, Limousin, Maine Anjou, Chianina, Beef Fresian, and others in their crossing systems. As a result, over 40 breeds of cattle are used by producers in the United States. The climate, market demands, and management capabilities determine which breeds or crosses are best adapted for production. Due to the rough topography and limited rainfall of the Rocky Mountains and sections of the central and western plains, these areas are adapted to the production of feeder calves, yearlings, and some 2-year-olds. In the range areas of the Nebraska Sandhills, Kansas, and Oklahoma Flint Hills, approximately 8–15 acres (3–6 ha) of grass are required annually for a cow and calf, while in more arid areas of Arizona and New Mexico 150–200 acres (60–80 ha) are needed. In the Midwest, a cow and calf can be maintained on 2–4 acres (1–2 ha).

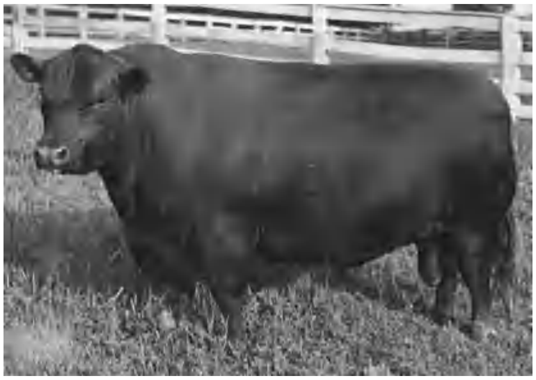


Fig. 1. Angus bull. (*American Angus Ass.*)

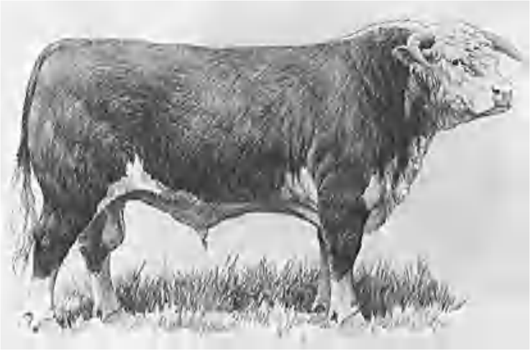


Fig. 2. Hereford bull. (*American Hereford Ass.*)

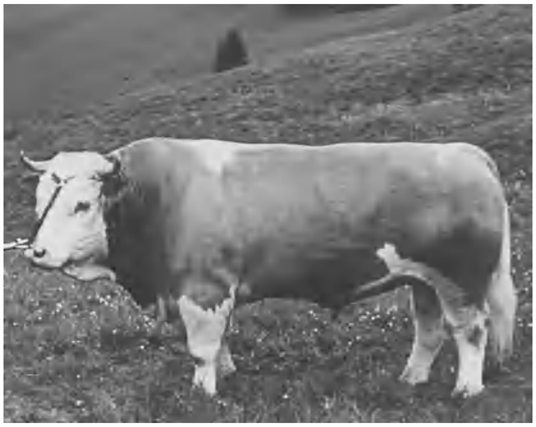


Fig. 3. Simmental bull. (*American Simmental Ass.*)

Feeding beef cattle. Feed is the major cost item in beef production. It is estimated that roughages furnish 80% of the total feed for cattle. From birth to a market weight of 1000 lb (450 kg), it takes less than 3 lb of grain for every pound of live weight gain. Most grain is fed in the final stages before marketing to shorten the feeding period and ensure a juicy, highly palatable meat product.

The mature cow is physiologically adapted to utilize, in addition to quality feeds, roughages such as straws, corn stover, cottonseed hulls, and byproducts of the sugar and canning industries. The rumen compartment of a mature bovine stomach can hold 40–60 gal (150–230 liters) and has microorganisms, such as bacteria, yeasts, and protozoa, to aid roughage digestion. The rumination processes of

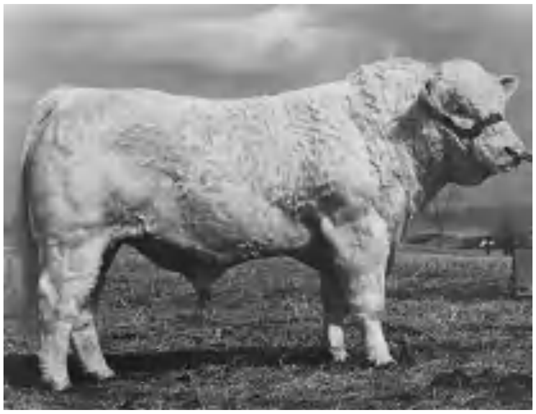


Fig. 4. Charolais bull. (*American International Charolais Ass.*)

cud chewing, regurgitation, and rumen microbial fermentation aid in the release of digestible energy in the form of glucose and acetic and propionic acids from the fibrous cellulose stems of roughage. Cattle have an advantage over monogastric swine or poultry because they can utilize forages and synthesize their necessary B vitamins and essential amino acids. The nutrient requirements of cattle are water, proteins, carbohydrates, fats, minerals, and vitamins. Energy from carbohydrates and fats along with proteins is a major dietary and cost consideration. Quality forages and good pastures, except in soil-deficient areas, can furnish adequate nutrients for maintenance, growth, and reproduction.

With population increases and greater human consumption of concentrate grains, research efforts have been directed to maximizing the use of corn and milo stovers as well as by-product straws of grain harvest in beef rations. Improved storage techniques and the processing of low-quality forages with nitrogen (protein) additives allow producers to get feed from waste materials. A popular feed for cattle is corn gluten feed, which is a high-moisture by-product of the distilling industry. Mature cattle eat about 2 lb daily for each 100 lb of weight; younger cattle eat 2.5–3.0 lb of air-dry feed per 100 lb (cured forages and grains of approximately 10–15% moisture).

The National Research Council gives the maintenance requirement of cattle as 0.75 lb of total digestible nutrient (TDN) energy per 100 lb liveweight. The TDN of a feed is defined as the sum of all digestible organic nutrients, protein, fiber, carbohydrate, plus 2.25 times the fat content. The council also expresses energy requirements as digestible energy, with 1 lb of TDN equivalent to 2000 kilocalories of digestible energy. One pound of weight gain requires 3.53 lb of TDN above maintenance. Air-dry grains, such as oats, barley, corn, and milo, vary in energy but average 72–80% TDN, and cured roughages, such as straw and hays, average 40–50% TDN. Balanced grain rations have more digestible energy and produce larger daily weight gains. Energy utilization of grains is improved by steaming and cracking, rolling, or flaking. The average daily gain of calves (over 6 months), yearlings, and 2-year-old steers fed a full feed of 1.75 lb grain per 100 lb are about 2.10, 2.25, and 2.50 lb, respectively, Compared to steers, weight gains are 5–10% slower for heifers and 5–10% faster for bulls. Efficiency of gains (energy per pound of gain) favors calves over yearlings and yearlings over 2-year-olds. As an animal grows, more fat is deposited and the animal becomes less efficient, requiring more feed to produce a pound of weight gain.

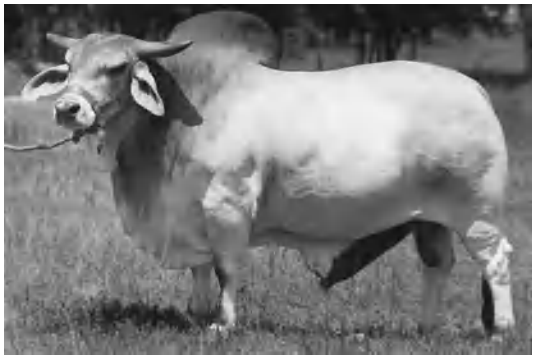


Fig. 5. Brahman, tolerant of humidity, heat, and insects. (*American Brahman Breeders Ass.*)

The recommended crude protein percent of fattening rations is approximately 12 and 11%, respectively, for calves and yearlings, and is somewhat greater for breeding bulls, lactating cows, and very young calves. Digestible protein needs are 75% of crude protein. Grains and many hays and straws are low in protein and are often supplemented with soybean and linseed and cottonseed oilmeal. Green pastures and legume hays are good sources of protein.

Nonprotein nitrogen sources, such as urea, biuret, and starea (supplements synthetically crystallized from nitrogen in the air), have found wide use as cheap protein substitutes for oilmeals. The rumen microorganisms combine the nonprotein nitrogen with carbohydrate and mineral elements to provide essential amino acids for use by cattle.

Ensiling is the method of preserving grass forages with 40–50% moisture (haylage), or grain forages (silage) with 60–70% moisture, in airtight structures. The exclusion of air with consequent fermentation of carbohydrates in the forages produces lactic and acetic acids, which act as preservatives. Silages are palatable and are widely used not only because they reduce curing and harvest losses but also because they provide more nutrients per acre. Certain hormonelike feed additives (such as ionophores) can alter rumen digestion and improve feed efficiency, while other hormonelike compounds, fed or implanted under the skin, can increase gains and efficiency 8–15%. All feed-additives use is extensively regulated to prevent any residue from remaining in the meat. Low-level feeding of antibiotics, such as aureomycin and terramycin, improves roughage utilization and helps minimize disease. *See* ANIMAL FEEDS; ANTIBIOTIC.

Marketing and grading. Cattle of varying weights, ages, and sexes are slaughtered annually. For clarity of market reporting, cattle are classified as steers, heifers, cows, bulls, stags, veals, and slaughter calves.

In 1973 a bullock market class was included for slaughter bulls under 2 years of age.

Consumers prefer beef from young grain-fed cattle averaging 900–1200 lb (400–540 kg) alive. Steer (castrated males) beef at a young age is generally the highest priced. Young fat heifers, if not pregnant, yield carcasses comparable to steers; however, heifers mature and fatten sooner and must be sold at lighter weights. Pregnancy reduces carcass yield and so heifers undersell steers. Cows (mature females) yield less muscular carcasses and less tender beef. Bulls and stags (late-castrated males) have coarser-textured meat. A veal is a fleshy young calf averaging 150–300 lb (65–135 kg), while older and heavier calves—up to 550 lb (248 kg)—are slaughter calves. The meat of young cattle has more moisture and a lighter-colored lean.

Slaughter cattle are purchased primarily on the quality of the carcass, probable dressing percent, and lean yield. They are graded according to government standards in descending value as prime, choice, select, standard, commercial, utility, and cutter. The age of the carcass plus the quality of the lean determines the final quality grade. The yield grade is determined by the relation of the amount of muscle to the amount of fat in the carcass. Dressing percent is the carcass yield based upon liveweight (the ratio of liveweight to carcass weight) and varies from 65% for prime cattle to 40% for cutter cattle. Fat increases dressing percent, whereas water, feed fill remaining in the paunch, coarse bone, and heavy hides reduce dressing percent.

Marbling, an interspersion of fat with lean, contributes to flavor, whereas texture (diameter of muscle fibers) and age determine tenderness. Young carcasses have softer, red-colored bones with cartilage tips on the vertebrae as contrasted to flinty white bone of older beef. Meat from older cattle is less tender because of intermuscular collagen and connective tissue. In the United States, cattle are marketed at young ages (12–20 months), and as the age of market cattle has been reduced, cattle are marketed with less fat waste. Old producing cows and bulls are used primarily as ground beef. Because of a vast increase in fast-food restaurants, ground beef makes up over half of the beef consumed.

Aging of beef (connective tissue breakdown) occurs naturally in beef for the first 7 days after slaughter. Electrical stimulation applied to the carcass shortly after slaughter can also improve tenderness. Provisioners have inproved efficiency and reduced their transportation costs by boning and trimming carcasses at the slaughter sites. A majority of beef is trimmed, packaged, and shipped in boxes to improve efficiency and standardize packages. *See* FOOD MANUFACTURING.

Beef economics. Production and marketing of cattle in the United States changes periodically. In the mid-1800s, slaughter plants were established in the East and were supplied with western grass cattle. In the 1920s and through the early 1940s, the corn belt states were important feeding areas. Many farmers used their corn to fatten 50–100 head annually and marketed them at major slaughter plants in Chicago, Omaha, and Kansas City. Annual per capita beef consumption increased to over 100 lb (45 kg) after World War II, and the feeding and slaughtering centers moved westward to the High Plains. Feedlots with 1000 to 10,000 head were established in Texas, Oklahoma, Colorado, and Nebraska, close to feeder cattle supplies.

Beef numbers, because of supplies and prices, have fluctuated in cycles every 10–12 years. Low cattle numbers cause an increase in prices, which in turn encourages herd increases until an oversupply occurs. A surplus in turn results in lower prices. To minimize risks in cattle feeding, operators are using computers to formulate least-cost diets and are using future contracts to hedge purchases and marketings. Since imports affect supplies and prices, producers have promoted legislation to control imports in proportion to domestic production.

Higher living standards made steak a prestige restaurant meat. Changes in life styles popularized hamburger food chains throughout the world. Shortages of cow slaughter and of grass cattle for hamburger have increased blendings of soybean protein with beef. Westernization and industrialization created greater demands for fed beef in Europe, in Japan, and in the Mideast and other developing areas.

Slaughter. In slaughtering, carcasses are split through the backbone into sides. For wholesale trade, sides are divided between the twelfth and thirteenth ribs into fore- and hindquarters and then retailed primarily as steaks and roasts. The major forequarter cuts are chuck and rib; the hindquarter cuts are loin and round.

Approximately 75% of fresh beef is merchandised by large grocery chains and is mostly choice and select grade. The emphasis for lean, marbled beef has resulted in further scoring carcasses within grades for cutability with 1 as top yield and 5 as the poorest. Cutability or lean yield is determined by measuring the loin area and fat thickness over the longissimus rib-eye muscle exposed between the twelfth and thirteenth ribs, plus accounting for kidney fat and carcass weight.

An attempt is made to utilize all offal of slaughter. Edible offal, such as heart, liver, and tongue, are called meat specialties. The hide, used as leather, represents about 8% of liveweight. Use is made of inedible offal for production of bone meal, blood meal, and meat scraps for livestock feeds; organs and glands for pharmaceutical products; hooves and horns for glue and fertilizer; and intestines for sausage casings.

Reproduction. Cattle are polyestrous, or breed at all seasons. Heifers reach puberty or sexual maturing, depending upon nutrition, at about 5–12 months of age. The estrous period, or acceptance of the male, averages about 18 h, and the estrous cycle occurs at intervals of 18–24 days. Well-grown heifers can be bred to calve at 2 years of age, but some cattle raisers prefer to calve later at $2^1/_2$–3 years. Average gestation or pregnancy is about 283 days. Twin births occur about 1 in 200, and other multiple births are less common. A heifer born twin to a bull is usually infertile and is called a freemartin. Young bulls are virile at about 10 months and should be used sparingly until 16 months. In range areas one active bull over 2 years is kept for 20–30 cows. Artificial insemination (AI) is not used as extensively with beef as with dairy cattle. Synchronizing or controlling estrus by feeding or injecting progesterone hormones is used. Estrous synchronization permits breeding of many cows (mostly by AI) close together and results in more uniform age calves. Paralleling AI use, the techniques of fertilized ova transplants and hormonal superovulation have proved fairly successful. These techniques permit annual multiple-offspring production from superior genetic cows through surgical transfer of fertilized ova to compatible recipient mothers. *See* ESTRUS.

Progressive cattle ranchers use performance tests to select cows that will wean 500–600-lb (225–270-kg) calves at 7 months and feeders that will gain $2^1/_2$–3 lb (1.1–1.4 kg) daily and weigh 1000–1100 lb (450–495 kg) at 1 year. Rate of gain and efficiency are highly correlated and heritable, meaning that fast-gaining parents should produce fast- and efficient-gaining offspring. Genetic engineering has attempted to change the characteristics of the beef animal by altering or rearranging its deoxyribonucleic acid (DNA). Mapping of the beef genome would allow producers to identify beef cattle that produce more efficiently, yield a more nutritious product, and develop a greater resistance to disease.

Diseases. The most universally destructive diseases of cattle are tuberculosis, brucellosis or abortion, foot-and-mouth disease, anthrax, and blackleg. Competent veterinarian diagnosis and treatment are important.

Tuberculosis is highly contagious and is transmissible to humans through milk. It is caused by a specific bacterium that may attack all parts of the body but particularly the lymphatic glands. There is no specific treatment, but the identification by an annual tuberculin skin test and disposition of reactors has markedly reduced tuberculosis. *See* TUBERCULOSIS.

Contagious abortion, brucellosis or Bang's disease, is a serious reproductive disease caused by *Brucella abortus*. Bang's disease causes birth of premature or dead calves. It has been reduced by blood-testing females (agglutination test) annually and slaughtering carriers. Calfhood vaccination of heifers between 4 and 8 months can provide immunity. *See* BRUCELLOSIS.

Other reproductive diseases, such as leptospirosis, vibriosis, and trichomoniasis, are prevalent, but they can be minimized by testing, immunization, and good health management.

Foot-and-mouth disease is caused by a virus and produces blister lesions on the feet and in and on the mouth. The disease is more debilitating than fatal and causes great economic losses due to weight reductions when cattle cannot or refuse to eat. A 1947 outbreak in Mexico was eradicated through inspection, vaccination, quarantine, and slaughter. Outbreaks in Europe and particularly in England in 1967 caused banning of European cattle imports into the United States. Outbreaks of foot-and-mouth disease in Argentina and other countries have limited imports to only canned beef from South America.

In 1984, a new vaccine for foot-and-mouth disease was developed by using biotechnology. The vaccine involves a live bacterin and is not used in the United States since it is free of foot-and-mouth disease. *See* FOOT-AND-MOUTH DISEASE; VIRUS.

Anthrax (splenic fever) is very infectious and usually fatal. The anthrax organism, a bacillus, can live many years and is transferable to other species and to humans. Diagnosis can be made by microscopic examination of a blood sample for the specific bacillus. *See* ANTHRAX.

Blackleg is an acute disease of young cattle. Annual calfhood vaccination is practiced in areas where the disease prevails. When outbreaks occur, the best preventive measure for blackleg and anthrax are burning or deep burial of infected carcasses and prompt vaccination with antiserum. *See* BLACKLEG.

Bloat is a digestive disturbance which causes a loss of millions of dollars annually. Bloat is recognized by distention of the animal's left side due to excessive gas formation in the rumen. Pasture forages such as alfalfa, clover, and other legumes eaten wet oftentimes produce heavy fermentation. A gradual change from dry roughage to legume pasture, as well as feeding hay on pasture, may help minimize bloat. The use of miscible oils in the drinking water and the use of poloxalene in mineral mixes and in protein blocks help reduce pasture bloat. BSE (bovine spongiform encephalopathy, or mad cow disease) is a progressive neurological disorder of cattle that results from infection by an unconventional transmissible agent termed a prion; it is an invariably fatal brain disease with an unusually long incubation period (measured in years). *See* PRION DISEASE.

Many other diseases affect the beef cow herd, though most can be minimized by a good health management program. Diseases include anaplasmoisis, which is transmitted by insect vectors and causes anemia, weakness, and dehydration; bovine respiratory disease, which is caused by several viruses and bacteria and produces respiratory problems; bovine virus diarrhea, which is caused by direct contact with infected animals and results in respiratory and alimentary problems; and bovine rhinotracheitis (rednose), which is a viral disease resulting in respiratory problems, abortion, conjunctivitis, enteritis, and encephalitis.

Feedlot health. To reduce shipping fever (hemorrhagic septicemia flu), incoming feedlot cattle are fed for 2 weeks after arrival daily high levels (300–500 mg) of antibiotics. New feeder cattle are usually wormed, sprayed for lice and grubs, and vaccinated for leptospirosis, infectious bovine rhinotracheitis, and bovine virus diarrhea.

There are two basic types of vaccines, killed (inactivated) and live. Killed products are more stable to use in feedlot situations.

There are programs of preconditioning feeder cattle on ranches for parasites and feedlot disease immunity. Foot rot or foul foot, common in wet feedlots, is controlled by low levels of antibiotics and iodine compounds in rations or in salt-mineral mixtures. Vitamin A deficiencies in feedlot cattle are prevented through intramuscular injections of vitamins

A, D, and E or by supplemental mixes in the rations. *See* AGRICULTURAL SCIENCE (ANIMAL); DAIRY CATTLE PRODUCTION. Douglas F. Parrett

Bibliography. B. A. Anderson and I. M. Hoke, *Composition of Foods: Beef Products*, Agric. Handb. 8-13, 1990; H. M. Briggs, *Modern Breeds of Livestock*, 1980; National Research Council, *Recommended Nutrient Allowances for Beef Cattle*, 1990; R. E. Taylor, *Beef Production*, 1992; U.S. Department of Agriculture, *Official Standards for Grades of Carcass Beef*, 1973.

Beekeeping

The management and maintenance of colonies of honeybees. Beekeeping is an ancient art. Although the commonly known honeybee species is native to Europe and Africa only, humans have transported them to other continents, and in most places they have flourished. The natural home for a honeybee colony is a hollow tree (**Fig. 1**), log, or cave. European strains of the honeybee build a nest only in locations which are dry and protected from the wind and sunlight (**Fig. 2**). African bees are less selective and may nest in hollowed-out termite mounds, rock piles, and locations which are less well protected. It is not known when humans first started to husband bees. However, there is much archeological and historical evidence that there were beekeepers among all of the ancient European and African civilizations.

Bees are important plant pollinators, but early beekeepers were not concerned with pollination and cross pollination of agricultural crops since their fields were small and they conducted diversified farming. For them, the importance and value of honeybees lay in the various hive products, especially honey, which was used as a sweetener, in making a fermented beverage, in medicine, and in conducting rituals. Beeswax was civilization's first plastic, and was used as a polish, in the preparation of molds for metal casting, for waterproofing, and in many other ways. Some ancient peoples ate honeybee brood, the developing larvae and pupae in the hive. Brood eating has never been widely practiced, but even today the habit is found among some African and Asian peoples.

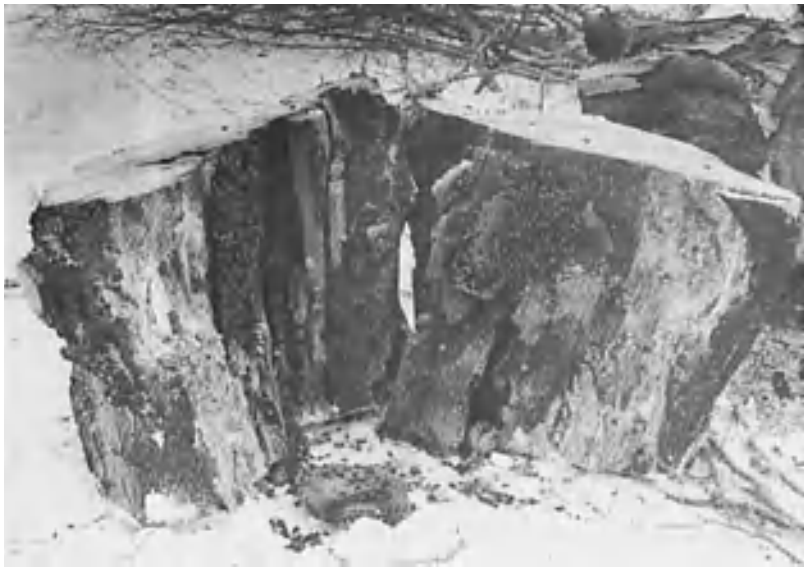


Fig. 1. A hollow tree, the original, natural home for honeybees; this tree trunk has been split apart to expose the nest. The bees used today are not domesticated and can easily revert to their old ways. Beekeeping is possible because of present-day understanding of bee biology. (*New York State College of Agriculture, Cornell University, Ithaca, New York*)

Honey. Honey was a precious commodity in early Roman, Greek, and Egyptian civilizations. For people 2000 to 4000 years ago, it was the main source of sugar. Thus, in the ancient Mediterranean world, honey and beekeepers were held in extremely high esteem. Today, honey is still valued as a sweetener and is used to some extent in medicine.

Production. The honey which beekeepers harvest is made from nectar, a sweet sap or sugar syrup produced by special glands in flowers, collected from both wild and cultivated plants. Nectar, the honeybees' source of sugar or carbohydrate, and pollen, their source of protein and fat, make up their entire diet.

Nectar contains 50–90% water, 10–50% sugar (predominantly sucrose), and 1–4% aromatic substances, coloring material, and minerals. To transform nectar into honey, bees reduce its moisture content, so that the final honey produced contains between 14 and 19% water, and also add two enzymes which they produce in their bodies. One enzyme inverts the sucrose, a sugar with 12 carbon atoms, into two 6-carbon sugars, levulose and dextrose. The second enzyme added by bees is glucose oxidase, which is inactive when honey has a normal moisture content (14–19%). However, when the honey is diluted, as when it is fed to honeybee larvae, the enzyme becomes active, resulting in the breakdown of dextrose into hydrogen peroxide and gluconic acid. Hydrogen peroxide prevents or slows the growth of bacteria and thus protects the diluted honey from attack by microbes. Glucose oxidase is also active while the bees are in the process of reducing the nectar's moisture content, a process which may take 24 h or more. The gluconic acid produced during this time, along with other acids present in honey, gives honey a pH of 3.5 to 4.0 which, among food products, is quite acid and makes honey an inhospitable medium for bacterial growth. Also, since it is a supersaturated sugar solution, honey has a high osmotic pressure. Therefore, any microorganisms present in honey are plasmolyzed and die.

Use in medicine. Honey has been used in medicines for thousands of years. Since honey is about half levulose, the sweetest of the common sugars, it is sweeter than cane sugar or corn syrup. Even today, honey is used in certain medicinal compounds because the sweet levulose covers up harsh, bitter flavors better than other sugars. The most popular honey for use by drug manufacturers is tupelo honey, which is produced only along the banks of the Apalachicola River in west Florida. Tupelo honey is the sweetest of the commercially produced honeys in the United States. Honeys high in levulose granulate much more slowly than do honeys high in dextrose; this is an added advantage for the drug manufacturer, who does not want the product to granulate or contain sugar crystals.

Before the discovery of modern salves and creams, honey was widely used as a wound dressing. It was an ingredient in over half of the medicinal compounds and cures made by Egyptian physicians 2000 to 5000 years ago. Honey is an effective dressing, not only because bacteria cannot live in it, but also because honey on a wound creates a physical barrier through which bacteria cannot pass. Some of the ancient Egyptian salves called for a mixture of honey and grease. Lint was sometimes added to help bind the mixture. If the honey was diluted with water, the glucose oxidase became active, producing hydrogen peroxide and again giving the wound some protection. Honey also prevented dried blood from adhering to a bandage and kept the bandage soft. Thus, dressings could be removed from a wound periodically, and the wound treated with more salve and rebandaged.

Use in baking. About half the honey produced in the world is sold as table honey and is used as a sweetener on bread, cereal, or desserts. The remaining half is used in the baking trade. Honey is hygroscopic, that is, it absorbs moisture. When honey is added to a baked product some of this property is retained, helping to keep the bread or cake moist. Bakers are also concerned with the browning qualities of their product, and the sugar levulose in honey has the desired browning qualities. Because the price of honey is many times that of sugar, it is too expensive to use in all baked products. Thus, honey is added only in specialty products.

Biology of the honeybee. Honeybees are social animals. A minimum of about 200 worker bees and a queen must live together for social order to exist; however, such a small number of bees could not survive for long without human assistance.

A colony of bees in an artificial hive, or living wild, contains about 10,000 worker bees, all females, in midwinter. The same colony would have a peak population of 30,000 to 60,000 workers in early summer. In addition to the workers, a normal colony has a single queen and 0 to 3000 drones (males). The sole function of the drones is to mate, and their presence in the hive is not tolerated by the workers during nonmating seasons. In the fall, the drones are deprived of food by the workers and are driven from the hive. The rearing of new drones in the spring starts well after bees have initiated the rearing of workers.

It is generally accepted that, in most situations, all the bees in a colony develop from eggs laid by the single queen. The size and composition of the colony population is regulated by how many eggs are reared during each season. In the Northern Hemisphere, bees rear the least number of young, called brood, in October and November. The least number of adult bees is found in the hive about 3 to 4 months later. Pollen first becomes available in the spring, February in the southern states and April in the northern states. The availability of pollen controls brood rearing, for without large quantities only a small number of young are reared. The natural seasonal population fluctuation which takes place in a colony is referred to as the cycle of the year.

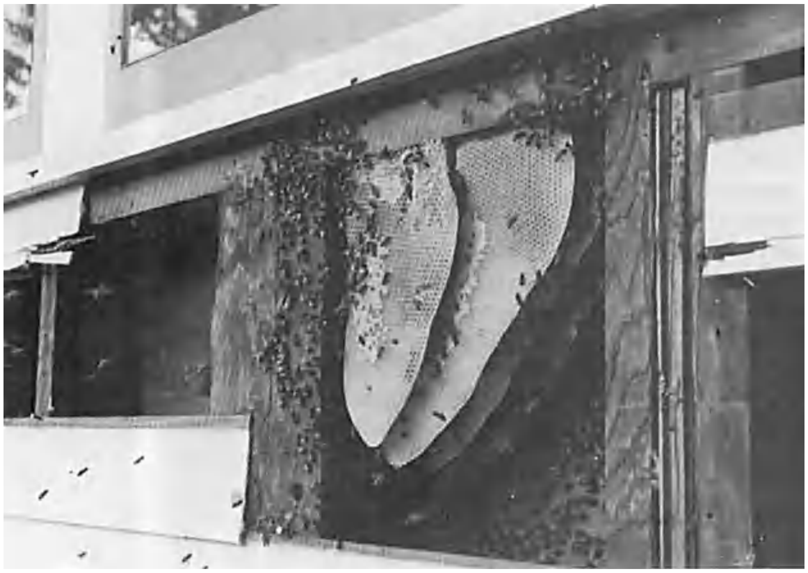


Fig. 2. A swarm which escaped from a beekeeper's hive and took up residence in the side of a house. The house boards have been removed to photograph the swarm. A swarm can live successfully in the side of a house, though the bees may be a nuisance.

Honeybees separate their young and their food, which is one of the factors which makes beekeeping possible. Honey and pollen are usually stored above the brood-rearing area, which is called the brood nest. The brood nest, which contains the eggs, larvae, and pupae, is a compact area more or less the shape of a ball. Brood are reared in many combs at the same time, but always in combs which are side by side. Honeybees maintain a constant brood-rearing temperature of about 92–95°F (24–36°C). This allows them to rear their young in a set period of time (see **table**) regardless of seasonal and climatic temperature variations. A few other species of social insects exercise some control over brood-rearing temperature, but none so precisely as the honeybee.

When cool weather prevails, the bees form an insulating shell of live bees around the outside of the brood nest. Bees in the center of the brood-rearing area make physical movements to generate heat. If the temperature drops, more bees generate heat and the shell of bees around the area becomes more compact. In this manner the heat is retained within the cluster; bees do not heat the inside of the whole hive. Bees can survive severe winters, even in Alaska, if they have sufficient honey. However, because so much honey is required for energy, usually in excess of 100 lb (45 kg) per colony in Alaska, it is considered

Development time for honeybee castes, in days

Stage	Queen	Worker	Drone
Egg	3	3	3
Larva	$5\frac{1}{2}$	6	$6\frac{1}{2}$
Pupa	$7\frac{1}{2}$	12	$14\frac{1}{2}$
	—	—	—
Total	16	21	24

uneconomical to keep bees there. In the summer when the temperature is high, bees collect water, deposit it in droplets around the hive, and fan their wings, causing masses of air to pass through the hive, evaporating the water and preventing excessive temperatures in the brood nest. Colonies of bees have survived in the Arizona desert, where temperatures may reach 120°F (49°C). In such cases, a colony may need to collect and evaporate as much as 1 or 2 gal (3.8 or 7.6 liters) of water per day.

Remarkably, the largest, most important individual in the colony, the queen, is reared in the shortest period of time. Evolution has favored this fast development. If a queen dies, a replacement must be produced rapidly or the colony will perish. Queens and workers develop from identical eggs. In fact, a queen can be made from a larva that is from 24 to about 48 h old. Larvae destined to become queens are fed a special food, royal jelly, which is produced by glands in the heads of worker bees; larvae destined to become worker bees receive less of this glandular secretion and more honey and pollen in late larval life. Thus, in a colony that loses its queen but has young larvae present, a queen may be produced in 11 or 12 days. *See* SEX DETERMINATION.

Pollination. Pollination is the process by which pollen, comprising the male germ cells, is transferred from the male part of a flower (anthers) to the female part (stigma) on the same plant. The process of transferring the pollen from the flower of one plant to that of another is called cross pollination. Pollen may be carried by wind, water, or animals. While only a small number of plants are water-pollinated, wind pollination is a fairly common occurrence.

Many plants, however, require animal agents for cross pollination. Certain birds, especially hummingbirds, are especially adapted for this purpose, but insects, especially bees, are the most efficient pollinators. Most bees are solitary insects, living and making their nests in hollow twigs, under bark, in stone crevices, or in tunnels in the ground. These insects pollinate many of the flowers which produce seeds, nuts, and fruit. Some species of social bees, however, including bumblebees and honeybees, are important pollinators. In many areas in the United States the solitary twig- and ground-nesting bees and the bumblebees outnumber honeybees.

The honeybee is especially well adapted to gathering large quantities of pollen and nectar and, in the process, bringing about efficient cross pollination. A honeybee is a very hairy creature, the hairs on its body being plumose, or branched, so that pollen grains are easily caught. When a honeybee collects pollen, it actually wallows in the flower parts, dusting itself thoroughly with pollen. It then hovers above the flower and uses its legs to clean its body of pollen; the pollen is raked into areas of longer hairs, called pollen baskets, one on each of its hindlegs. Bees pack pollen into the baskets until they become large, round balls; bright colored pollen balls may be seen with the naked eye on foraging bees, both as they collect more pollen in the field and as they later enter their hive.

When the pollen balls on the two hindlegs are large enough, the bee carries the pollen back to the hive, where it is deposited in cells and becomes food for adult bees and larvae. A bee never cleans its body so thoroughly that all the pollen grains are removed, and does not clean itself after visiting every flower. Thus, in the process of moving from one flower to another, large numbers of pollen grains are transferred. *See* FLOWER; POLLINATION.

Modern beekeeping. Scientific beekeeping started in 1851 when an American, L. L. Langstroth, discovered bee space and the movable frame hive. Bee space is the open space which is about 0.4 in. (1 cm) wide and maintained around and between the combs in any hive or natural nest and in which the bees walk. If this space is smaller or larger than 0.4 in. (1 cm), the bees will join the combs. When the combs are stuck together, the hive is not movable, and it is not possible for beekeepers to manipulate a colony or to examine a brood nest. Prior to 1851, beekeeping methods were crude and the quantity of honey harvested per hive was probably one-quarter of that obtained today.

It was found, in 1857, that bees could be forced to build a straight comb in a wooden frame by giving them a piece of wax, called foundation, on which the bases of the cells were aready embossed. Bees use these bases to build honeycomb, the cells of which are used for both rearing brood and for storing honey. When a hive of bees is given a frame of foundation, they are forced to build the comb where the beekeeper wants it and not where they might otherwise be inclined to build it.

Another discovery, made in 1865, was that honey can be removed from the comb by placing a comb full of honey in a centrifugal force machine, called an extractor. Approximately 16 lb (7 kg) of honey must be consumed by the bees in order for them to make 1 lb (0.45 kg) of wax comb. If the beekeeper can return an intact comb to a hive after removing the honey from it, the bees are saved the time and trouble of building a new comb, and the honey harvest is increased.

The next discovery, in 1873, was the modern smoker. When bees are smoked, they engorge with honey and become gentle. Without smoke to calm a hive, normal manipulation of the frames would not be possible. While smoke has been used to calm bees for centuries, the modern smoker has a bellows and a nozzle which allows a beekeeper to use smoke more effectively by directing the smoke where it is needed.

Within the short span of 22 years, these four discoveries laid the basis for the modern beekeeping industry. Prior to 1851, only a few individuals in the United States owned a hundred or more colonies of bees. By 1880, several people owned a thousand or more colonies each, and honey, which had once been a scarce commodity, became abundant.

Hives. All beehives used in the industry today are made of wooden boxes (**Fig. 3**) with removable frames of comb, and 90% of these are 10-frame Langstroth hives. Beekeepers have standardized their

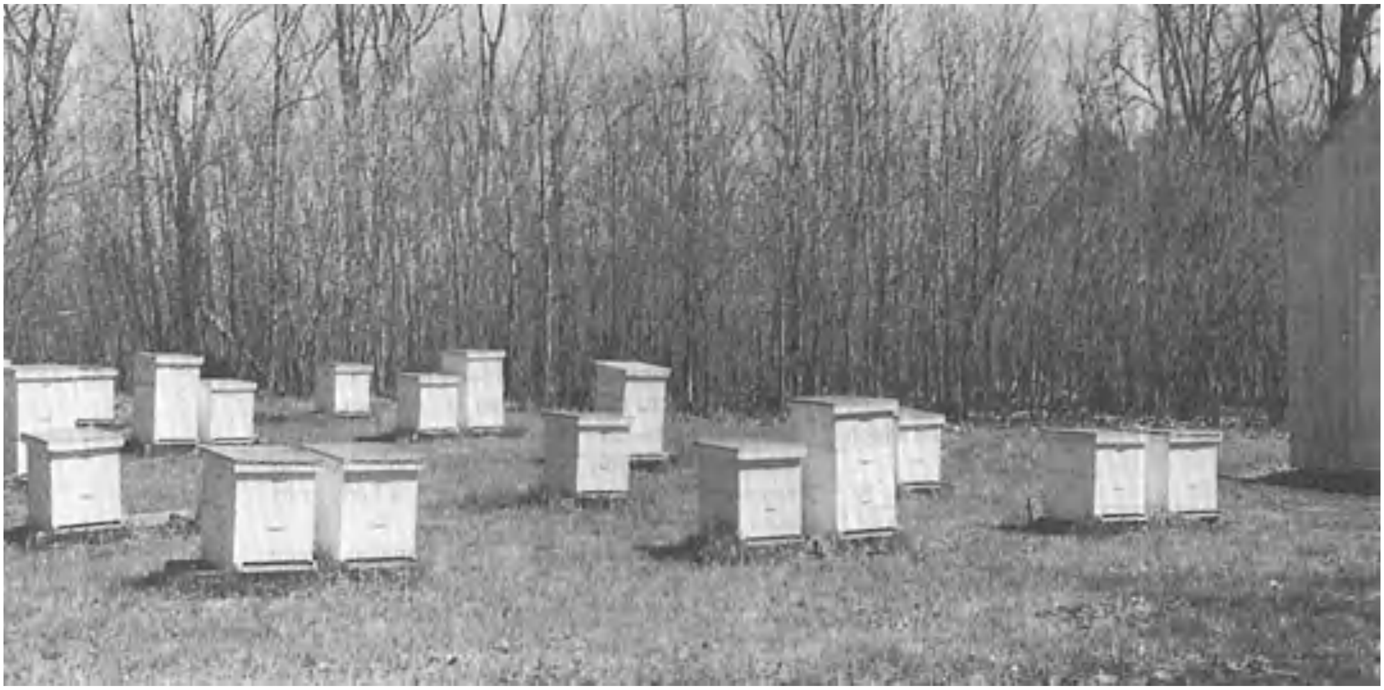


Fig. 3. A typical modern-day commercial apiary. Beekeepers often select apiary sites surrounded by woods so as to protect bees from damage by wind; hidden apiaries are also less likely to be vandalized.

equipment so that parts will be interchangeable within an apiary and from one commercial operation to another. The dimensions of the modern-day hive (**Fig. 4**) were determined in 1851 by Langstroth, who made a box based on the size of lumber available to him and on what he thought was an easy unit to lift and handle. The honeybee is remarkably adaptable. While it may show a preference for a nest of about 1.4 ft^3 (40 liters) in volume, it will live almost anywhere it can keep its combs dry and protected.

Commercial practice. To be successful in commercial beekeeping, beekeepers must move to those areas where nectar-producing plants abound. The best-known honey in the United States is clover honey. Clovers grow in almost every state and, where abundant, they may produce large quantities of nectar. Alfalfa is also a good nectar-producing plant and it too is a major source of honey. Oranges grown in Florida, Arizona, and California are also known for the high quantity and quality of nectar which they produce. Just as each flower is different in its color, shape, and design, so each flower produces a unique nectar. Thus, orange honey does not taste like clover honey, and clover honey is different from the honey which is produced by buckwheat, sage, gallberry, and so on. Bees collect nectar from hundreds of kinds of flowers, and thus there are a great variety of honey flavors.

In agricultural regions where land use is intense, large areas of land are cleared, removing hedgerows and woodlots in which many insects nest and grow. Large acreages of such monocrops as apples, oranges, alfalfa, almonds, cantaloupes, and watermelons are planted, decreasing the diversity of the vegetation and eliminating many indigenous insect pollinators. In these areas there are few solitary bees.

Thousands of colonies of honeybees are rented each year by growers of crops needing cross pollination. When the plants are coming into bloom the beekeeper moves the bees, usually at night when all the bees are in the hive, and places them in groups in groves, orchards, and fields. While beekeepers make most of their living producing honey, the real importance of their bees in the agricultural economy

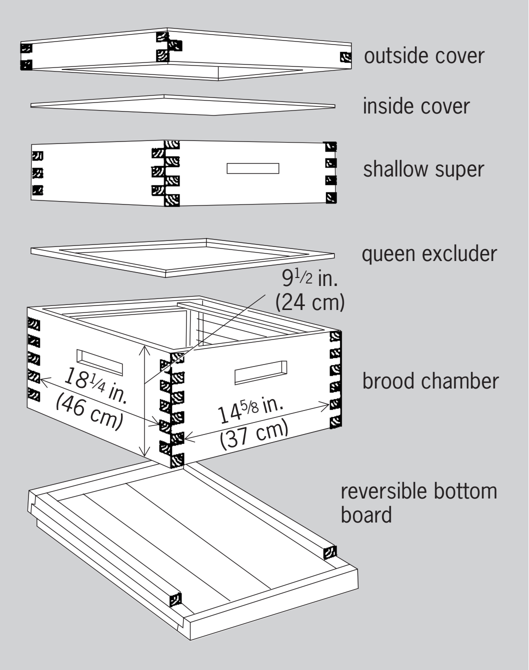


Fig. 4. Construction and dimensions of a 10-frame beehive. (*Apiculture Research Branch, Plant Industry Station, Agricultural Reseach Service, USDA, Beltsville, Maryland*);

is as cross pollinators. Without cross pollination the abundance and variety of food and flowers would not exist.

It is estimated that there are about 2000 commercial, full-time beekeepers and 200,000 persons for whom beekeeping is a part-time job or hobby in the United States. As mentioned, to be successful in beekeeping one must move to areas where nectar-producing plants are numerous. One may plant some flowering trees, bushes, and ground plants on which bees may forage, but a colony needs so much foraging area during the course of a year that a beekeeper cannot plant all that is needed. Most commercial beekeepers agree they need to harvest an average of 60 to 100 lb (27 to 45 kg) of honey per colony per year to be successful. In addition 30 to 60 colonies are needed in an apiary to be economically efficient. Not all locations have sufficient nectar forage to support that number of colonies. Commercial beekeepers feel they need to own 500 to 2000 colonies to make a living, depending on how they market their honey. Some beekeepers sell their honey on the wholesale market only, while others pack their honey themselves and devote a great deal of time to sales. *See* AGRICULTURAL SCIENCE (ANIMAL). Roger A. Morse

Beet

The red or garden beet (*Beta vulgaris*), a cool-season biennial of Mediterranean origin belonging to the plant order Caryophyllales (Centrospermales). This beet is grown primarily for its fleshy root, but also for its leaves, both of which are cooked fresh or canned as a vegetable (see **illus.**). The so-called seed is a fruit containing two to six true seeds. Detroit Dark Red strains predominate. Beets are only slightly tolerant of acid soils and have a high boron requirement. Cool weather favors high yields of dark red roots. Exposure to temperatures of 40–50°F (4–10°C) for 2 weeks or more encourages undesirable seed-stalk development. Roots are harvested when they are 1–3 in. (2.5–7.5 cm) in diameter, generally 60–80 days after planting. Wisconsin, New York, and Texas are important beet producing states. *See* CARYOPHYLLALES; SUGARBEET. H. John Carew

Fleshy roots and leaves of beet plants. (*Asgrow Seed Co., Subsidiary of the Upjohn Co.*)

Behavior genetics

The hereditary factors of behavior may be studied in animals and humans. Charles Darwin, who originated the theory that natural selection is the basis of biological evolution, was persuaded by Francis Galton that the principles of natural selection applied to behavior as well as physical characteristics. Members of a species vary in the expression of certain behaviors because of variations in their genes, and these behaviors have survival value in some environments. One example of such a behavior is curiosity—some organisms are more curious than others, and in some settings curiosity is advantageous for survival. Therefore, in those environments more of the organisms that express curiosity survive to reproduce than do those that lack this trait.

The science of behavior genetics is an extension of these ideas and seeks (1) to determine to what extent the variation of a trait in a population (the extent of individual differences) is due to genetic processes, to what extent it is due to environmental variation, and to what extent it is due to joint functions of these factors (heredity-environment interactions and correlations); and (2) to identify the genetic architecture (genotypes) that underlies behavior.

Traditionally, some of the clearest and most indisputable evidence for a hereditary influence on behavior comes from selective-breeding experiments with animals. Behavior genetic research has utilized bacteria, paramecia, nematodes, fruit flies, moths, houseflies, mosquitoes, wasps, bees, crickets, fishes, geese, cows, dogs, and numerous other organisms. Breeding of these organisms allows genetically useful types of relationships, such as half-sibs, to be produced easily. Although not all of this work involves breeding experiments, enough of it does to demonstrate conclusively the importance of genetic processes on behavior in every species examined. Artificial selection (selective breeding) can be used to obtain a population that scores high or low on specific traits. Inbred strains of many animals (populations that are made up of individuals which are nearly identical genetically as a result of inbreeding), particularly rodents, are readily available, and the study of various types of crosses among them can provide a wealth of information. An experimental design using the recombinant inbred-strain method shows great promise for isolating single-gene effects. This procedure derives several inbred strains from the F_2 generation (grandchildren) produced by a cross between two initial inbred strains. Since it is possible to exert a great deal of control over the rearing environments, the experimenter can manipulate both heredity and environment, a feat virtually impossible in human studies but required for precise answers to behavior genetic questions.

Other work has focused on the effects of the environment and genotype-environment interactions.

For example, experiments with mice have shown that, with respect to several learning tasks, early environmental-enrichment effects and maternal effects were quite small, relative to the amount of normal genetic variation found in the strains of mice tested. Only a few genotype-environment interactions were found. Still other work has shown that early experiences affect later behavior patterns for some strains but not others (a genotype-environment interaction).

An increasing role for animals in genetic research is to provide models of human genetic diseases, many of which have behavioral features. Such animal models may occur naturally or may be engineered in the laboratory. For example, a mouse model was produced which is deficient in the enzyme HPRT. Deficiency of this enzyme in humans is due to a recessive X-linked gene and leads to a disorder known as Lesch-Nyhan syndrome, which is characterized by self-injurious behavior and mental retardation. Animal models are available for many other neurobehavioral disorders, including narcolepsy, various epilepsies, and alcoholism. The availability of animal models allows researchers to obtain information about the development of genetic disorders and the effects of different environments on this development, as well as to explore treatment options. While it is not always prudent or desirable to generalize from animal results to humans, it is assumed that basic genetic systems work in similar ways across organisms, and it is likely that these types of animal studies will play a key role in elucidating the ways in which environment influences phenotypic variation. With advances in genetic technology, it is possible to observe genetic variation more directly by locating, identifying, and characterizing genes themselves.

Chromosomes and Genes

In most organisms, chromosomes are linear structures comprising strands of deoxyribonucleic acid (DNA) and accompanying proteins. A gene is a specific segment of DNA representing the genetic information for making a needed protein or for regulating other genes. Each gene may come in several variations, called alleles, that may produce somewhat different phenotypic outcomes. An organism's phenotype, or complement of observable properties, is due to the combined effects of its genetic information and the environmental conditions it experiences. *See* CHROMOSOME; GENE.

Each human inherits 23 paternal chromosomes and 23 maternal chromosomes at fertilization. Genes occupy (fairly) specific positions on these chromosomes. Located in the nucleus of each cell, the 46 chromosomes can be thought of as comprising 23 pairs, with one member of each pair coming from each parent; for example, the maternal chromosome 1 pairs with the paternal chromosome 1, the maternal chromosome 15 pairs with the paternal chromosome 15, and so on. These are called homologous pairs. Thus, at the same location on each member of the pair, there is a gene for a particular kind of genetic information, or two genes of that kind per individual. Since each chromosome in the pair comes from one parent, each gene in the pair comes from one parent. This is true for the first 22 pairs of chromosomes, called the autosomes. The twenty-third pair is the sex chromosomes, and they may or may not be the same. A female has two X chromosomes (designated 46, XX), but a male has one X chromosome and one Y chromosome (46, XY). The Y chromosome contains crucial genetic information that causes an embryo to develop as a male. Thus, a mother can contribute only an X chromosome to her offspring, but a father can contribute either an X or a Y chromosome. It is, therefore, the father who determines the sex of the offspring. *See* HUMAN GENETICS; SEX DETERMINATION.

The Y chromosome is smaller than the X chromosome and contains less genetic material. Consequently, few genes that reside on the X and Y chromosomes are paired. Traits influenced by genes carried only on the X chromosome are called X-linked, while genes carried only on the Y chromosome are called holandric genes. Red-green color blindness is under the control of a recessive variant (allele) of a gene on the X chromosome. This allele is called recessive, and the gene for normal color vision is called dominant because, in females, the gene for red-green color blindness must be present on both X chromosomes in order for the female to be red-green color-blind. One dominant normal color vision gene can override one defective gene. In the male, however, the situation is different, because this X-chromosome gene does not have a matching gene on the Y chromosome. If the X chromosome carries the defective version of the gene and is matched with a Y chromosome, that individual is red-green color-blind. Females can only be red-green color-blind if both genes at that location are recessive, so this behavioral trait is more frequent among males than females. *See* DOMINANCE; SEX-LINKED INHERITANCE.

Chromosome abnormalities. There are a number of abnormalities involving entire chromosomes or pieces of chromosomes, rather than isolated genes, that influence behavior, and many important ones involve the sex chromosomes. The reason is that most abnormalities of the autosomes are lethal during fetal development. It is estimated that as many as 50% of all conceptuses have gross chromosomal abnormalities, while these defects are found in only 0.5% of newborns. It seems that an excess of chromosomal material is better tolerated than a deficit; even so, when a newborn is found to have an extra autosome, it is usually one of the smaller chromosomes, such as 13, 18, or 21. *See* CHROMOSOME ABERRATION; CONGENITAL ANOMALIES.

Autosomal chromosome anomalies. The abnormal presence or absence of autosomes is usually lethal in the early stages of fetal development. There are, however, several exceptions. The best known is Down syndrome, or trisomy 21, in which there is an extra twenty-first chromosome (47, +21), occurring in about 1 in 700 live births. The physical features of these individuals are striking: round face with broad head, small medial epicanthal fold, flattened bridge of the nose, small ears and nose, and protruding tongue. Skeletal and congenital heart defects are common, as is increased susceptibility to respiratory infections. Down children have a reduced life expectancy, although modern antibiotics and heart surgery have improved this somewhat. Although the IQ of these individuals ranges from less than 25 to about 75, with a mean of 40–50, many can be taught to read and write and to care for themselves.

Other kinds of chromosomal abnormalities exist, including duplications, rearrangements, and deletions of various amounts of chromosomal material. In most cases, the more chromosomal material involved, the more deleterious the outcome. However, it is becoming apparent that even very small deletions and duplications can lead to serious disorders, although these usually are considered single gene disorders rather than chromosomal abnormalities. For example, the abnormal repetition of three of the units (base pairs) that make up the gene's DNA has been implicated in, among others, Fragile X, Huntington's disease, a progressive neurological disorder, and myotonic dystrophy, characterized by muscle spasms and muscle wasting, cataract, and other features.

Sex chromosome anomalies. Most sex chromosome anomalies involve the presence of one or more extra X or Y chromosomes, such as XXX, XXY, XXXX, or XYY. An exception is Turner syndrome, in which only one complete X chromosome is present (45, X). [In some cases, part of a second X may be present. In a few individuals, some cells have two X chromosomes while others have only one; these individuals are referred to as mosaics.] Each type of sex chromosome anomaly has specific features or abnormalities associated with it. For example, most individuals with Turner syndrome have only one sex chromosome, an X, but are phenotypically female. Physical features commonly found in these individuals include webbed neck, short stature (usually less than 5 ft or 150 cm), streak gonads, infantile genitalia, shield chest, and widely spaced nipples. Pubic and underarm hair may or may not be present. Since the ovaries do not develop, the women do not menstruate and are sterile. However, not all of these features are present in all individuals with Turner syndrome. There are certain behavioral characteristics that have been associated with women having Turner syndrome, such as a lack of aggressive tendencies, reduced libido, and strong inhibitions; today, however, there is some question about whether these characteristics are due in part to the way these short, immature-looking women are treated by others. Sex hormone therapy can help to increase height somewhat and to induce physical maturation; such changes may have psychological benefits as well. Intelligence is usually in the normal range in Turner syndrome, although the individuals may have deficits in space-form recognition and directional sense. *See* DOWN SYNDROME.

Another type of sex chromosome abnormality leads to the Fragile X syndrome. The name derives from a fragile site on the long arm of the X chromosome which can be seen when cells are grown under special conditions. Affected males usually have some characteristic facial features, enlarged testes, and mental retardation. The Fragile X syndrome is the most frequent known cause of mental retardation in males. Females with one Fragile X chromosome and one normal X chromosome may have a lesser degree of mental retardation.

Variations in genes. Far more common than chromosome abnormalities are variations in the genes themselves. Most of these variations lead to differences among individuals which are considered normal, for example, variations in eye color. Some gene variations, however, lead to differences which are not considered normal, such as Huntington's disease, albinism (lack of pigment in hair, skin, and eyes), or hemophilia (lack of blood clotting factors). It is now known that some differences among individuals are due to variations in single genes, while others are due to the combined effects of two or more genes (polygenic or continuous characters). It must also be remembered that the environment is always involved in the expression of a trait.

Single-gene effects. The effects of a single gene on behavior have been most extensively studied in the domain of mental retardation. Research has shown that there are a large number of metabolic pathways (series of chemical reactions used by cells to make or break down molecules) which have defects due to a single gene. Over 100 of these defects influence mental ability. *See* MENTAL RETARDATION.

One such single-gene defect is classic phenylketonuria (PKU), an autosomal recessive disorder, which also illustrates the role that environment can play in the expression of a trait. Individuals who are homozygous (having two copies of the PKU allele) are unable to make the enzyme phenylalanine hydroxylase, which converts the essential amino acid phenylalanine to tyrosine, a nonessential amino acid. Instead, the excess phenylalanine builds up in the blood and is converted into phenylpyruvic acid, which is toxic to the developing nervous system in large amounts. The main effect of untreated PKU is severe mental retardation, along with a distinctive odor, light pigmentation, unusual gait and posture, and seizures. Many untreated individuals with PKU show fearfulness, irritability, and violent outbursts of temper. The treatment for PKU consists of a diet that severely restricts phenylalanine intake. If the diet is carefully maintained from the first few weeks of life through about 8–10 years of age, PKU individuals have a good chance of developing normal or near-normal intelligence. Most states in the United States now require that newborns be tested for PKU before leaving the hospital. *See* PHENYLKETONURIA.

Polygenic effects. Most of the traits of interest to behavioral scientists are continuous (ranging from low to high, such as intelligence) rather than discontinuous (for example, affected or unaffected with PKU). They are continuous because they are under the control of many genes, each having a small effect (polygenic). Thus, a trait governed by a single gene usually shows two or three discrete phenotypic categories. A polygenic trait shows a distribution of phenotypes that approximates a normal curve. For some types of behavior, the question of whether it is caused by a single gene or whether it is a polygenic trait has not been settled, and the role played by the environment is at issue.

Nature or Nurture Question

Every organism develops in a particular environment, and both genes and environment control development. It is, therefore, not possible to state that a particular behavioral trait is either genetic or environmental in origin. It is possible, however, to investigate the relative contributions of heredity and environment to the variation among individuals in a population. With humans, it is possible to obtain approximate results by measuring the similarity among relatives on the trait of interest.

The most widely used measure of similarity is the correlation coefficient. This measure varies from −1 to +1. If a large number of pairs of people were selected at random from a population and then compared, each pair's scores on some measure (for example, a physical variable such as height, or a psychological variable, such as extroversion), the degree of similarity among them, as measured by a correlation coefficient, would be zero. The reason is that, on the average, they would be genetically unrelated and would have been reared in different environments, and therefore would be unlikely to be similar on that measure, regardless of the influence of heredity or environment. In order to obtain information about these influences, scientists compare individuals who are related or who have shared environments. Because related individuals usually share some environments as well as genes, it is useful, but not absolutely necessary, to compare the results with those from the relatives that have been reared apart.

Twin and adoption studies. Twins are often used in behavior genetic studies. One method compares the similarity within pairs of both identical (monozygotic or one-egg) twins and fraternal (dizygotic or two-egg) twins reared together. Identical twins have all their genes in common by descent, since they arise from a single fertilized egg. Fraternal twins arise from two fertilized eggs and so, like any pair of nontwin siblings, share on average one-half of their genes. (Parents and offspring share exactly half of their genes, as described earlier.) If it is assumed that the effects of the shared environments of the two types of twins are equal (a testable assumption), greater resemblance between identical twins than fraternal twins should reflect the proportion of genes they share, and the difference between the correlations of the two twin types should represent about one-half the genetic effect.

A second type of twin study compares not only twins reared together but twins who have been reared apart. The degree of similarity between identical twins reared in the same home would reflect the fact that all their genes are identical and that they share a common family environment. On the other hand, if identical twins can be located who had been adopted by different families chosen at random (an unlikely event, since adopted children tend to be selectively placed), a measure of their degree of similarity would reflect only the effect of their common genes. If it were true that an individual's level on a measure (for example, height or extroversion score) is determined in large part by the characteristics of his or her family and the opportunities that the family makes available to him or her, reared-apart identical twins should be no more alike than pairs of individuals chosen at random. If they do exhibit some degree of similarity, it would reflect genetic effects alone. The existence of even very large genetic effects, however, would in no way imply that the environment was unimportant in the development of the trait; it would simply imply that environment was less important than genes in determining the variation among individuals on the trait in question at the time of measurement. That is, the individuals would differ more because of the genes they carry than because of the particular environments to which they were exposed. In another range of environments, the results might be different.

Another method of determining the relative importance of genes and environment is to compare the degree of similarity of unrelated individuals raised in the same family: adopted children and their adoptive parents and siblings. Any similarity between these individuals would be due entirely to common family environment and not to genes (barring placement in similar families). Family, twin, and adoption studies have their respective advantages and limitations; when results from different kinds of studies are similar regarding a particular trait, they can be viewed with increased confidence.

In addition, biometric and other quantitative methods allow information from many types of family relationships to be processed simultaneously. These approaches estimate the different genetic and environmental components of variance for a trait. They do this by estimating the phenotypic variances and covariances of measurements from different types of biological and adoptive relationships, for example, biological parent-offspring, adoptive parent-offspring, biological sib-sib, identical twins, fraternal twins, and so on. These estimates then allow the fitting or testing of models which predict certain amounts of genetic and environmental variance. Finding the best-fitting model leads to a better understanding of the underlying genetic (and environmental) architecture of a trait. Finally, modern genetic technology is being applied to single-gene conditions with behavioral components. One approach analyzes DNA from populations of affected and unaffected individuals to determine if a known, or candidate, gene (chosen because its function conceivably could contribute to the observed phenotype) could be the gene in question. Abnormal alleles of genes affecting neurons in the brain, such as those coding for neurotransmitters and their receptors and for ion channels, among others, are being explored as candidate genes for certain behavioral disorders. Several candidate genes for schizophrenia have been ruled out. *See* BIOMETRICS; TWINS (HUMAN).

Gene mapping. Gene mapping or linkage analysis is a technology in which a gene is localized to a specific position on a chromosome. The DNA from affected and unaffected members of families is analyzed to find an association between the disease state and the presence of a gene or a piece of DNA whose position is known, called a marker. The assumption is that if the association is very high it is due to the fact that the gene involved in the trait under investigation and the marker DNA are physically close (linked) on the chromosome. Various techniques can then be used to identify the DNA that makes up the gene itself. One of the very first genes to be localized to a particular chromosome by this method was that for the progressive neurological disorder Huntington's disease (located on chromosome 4). Once a gene has been identified, its function can be determined and studied and treatment options considered. Linkage analysis is applied most often to the study of genetic disorders, also can be used to learn more about genes contributing to normal variation.

Gene mapping is a powerful tool that is widely applied. However, it cannot always provide answers to genetic questions. Its success depends upon several factors, including understanding of the mode of transmission of the gene (for example, whether the allele of interest is dominant or recessive, on an autosome or the X chromosome), having the correct diagnosis, and assuming that the trait in question is due to the same genes in all the individuals who have the trait (genetic homogeneity). Alzheimer's disease provides an illustration of genetic heterogeneity, being found in familial and nonfamilial as well as early- and late-onset forms. Three chromosomal locations have been identified among groups of families with inherited Alzheimer's disease—chromosome 19 for late-onset, and 14 and 21 for early-onset types. In addition, gene mapping requires that markers for the appropriate chromosomal locations are available and that the families being studied show some variation in allelic forms of these markers. *See* ALZHEIMER'S DISEASE.

Behavioral traits often present special challenges for this type of analysis. The mode of inheritance is often not known with confidence. It is widely believed that many, if not most, behavioral traits are polygenic in nature. If each gene involved has only a small effect, it will be difficult to identify the genes by mapping alone. However, biometric techniques developed for analysis of polygenic traits in plants (quantitative trait loci, or QTL, methods) can be used in animals, sometimes in conjunction with recombinant inbred strains. Another potential difficulty with behavioral traits is making an accurate, or at least consistent, diagnosis. Inclusion of too many or too few characteristics in a diagnosis will lead to incorrect categorization of some individuals in each family studied and will produce spurious results. Finally, it is thought that genetic heterogeneity may be common in behavioral traits; if it is not identified, results of linkage analyses may be inconclusive.

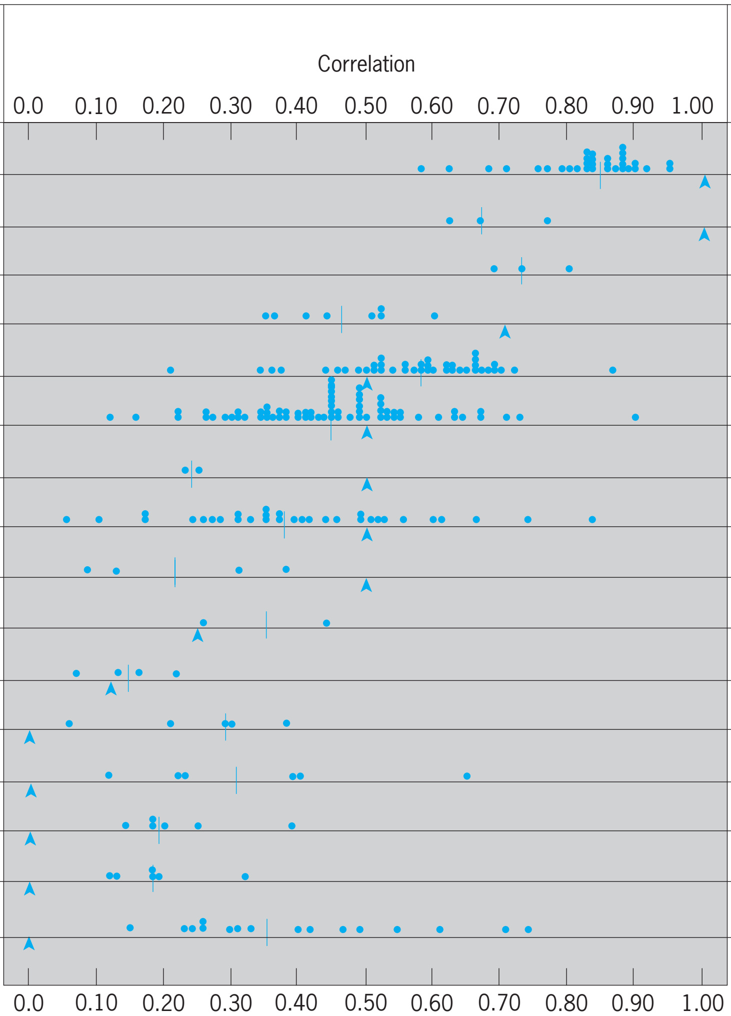
Environmental variables. Developmental psychologists are finding that differences in children's behavioral phenotypes are due more to their different genotypes than to their different rearing environments, as long as those environments are within a normal range of experiences. Identifying environmental variables from this normal range that have an important effect on the behavioral phenotype may be even more difficult than identifying contributing genes. Advances in theory and new technologies, combined with information from more traditional methodologies, will continue to provide insight into the contributions of genes and environment to behavior.

Intelligence. The results of over 50 years of research on the similarity in intelligence between relatives (as estimated by intelligence, or IQ, scores) indicate that identical twins reared apart are remarkably similar relative to adopted individuals reared in the same home (see **illus.**). Identical twins reared together are much more similar than fraternal twins reared together, and dramatically more similar than unrelated individuals reared together. This general evidence clearly suggests that heredity is an important determinant of population variation in intelligence. It has been determined that approximately40–70% of the variation in intelligence can be explained by hereditary factors. *See* INTELLIGENCE.

Special mental abilities. Intelligence (IQ) tests, while controversial, are felt by many to be a measure of general intelligence. Other special mental abilities exist, such as spatial orientation (the ability to rotate three-dimensional figures mentally) and perceptual speed and accuracy (the ability to detect errors in text quickly). It is possible to ask if profiles of abilities (patterns of strengths and weaknesses) are influenced by heredity.

Scores on the various subtests of the Wechsler intelligence scales can be thought of as reflecting measures of some of these special mental abilities. Two studies examined the correlations of the patterns of subtest scores in young twins. The first used the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) to measure 70 pairs of identical and 46 pairs of fraternal 6-year-old twins. The correlations are 0.43 and 0.27, respectively. The second study measured 69 identical and 35 fraternal pairs, with a mean age of 8 years, on the Wechsler Intelligence Scale for Children (WISC-R) and obtained correlations of 0.45 and 0.24, respectively. Data from the Minnesota Study of Twins Reared Apart, using the Wechsler Adult Intelligence Scale (WAIS) score from 32 pairs of identical twins reared apart and 13 pairs of fraternal twins reared apart, provide correlations of 0.52 and 0.12. The finding that the fraternal twins are approximately half as similar as the identical twins, and the similarity of results across age groups, despite the use of different instruments, strongly support the conclusion of a significant genetic influence on the patterning of mental abilities.

Developmental genetics and behavior. The Louisville Twin Study followed nearly 500 pairs



	Number of correlations	Number of pairings	Median correlation	Weighted average
Monozygotic twins reared together	34	4672	0.85	0.86
Monozygotic twins reared apart	3	65	0.67	0.72
Midparent-midoffspring reared together	3	410	0.73	0.72
Midparent-offspring reared together	8	992	0.475	0.50
Dizygotic twins reared together	41	5546	0.58	0.60
Siblings reared together	69	26,473	0.45	0.47
Siblings reared apart	2	203	0.24	0.24
Single parent–offspring reared together	32	8433	0.385	0.42
Single parent–offspring reared apart	4	814	0.22	0.22
Half-siblings	2	200	0.35	0.31
Cousins	4	1,176	0.145	0.15
Nonbiological sibling pairs (adopted/natural pairings)	5	345	0.29	0.29
Nonbiological sibling pairs (adopted/adopted pairings)	6	369	0.31	0.34
Adopting midparent–offspring	6	758	0.19	0.24
Adopting parent–offspring	6	1397	0.18	0.19
Assortative mating	16	3817	0.365	0.33

Familial correlations for IQ. The vertical bar in each distribution indicates the median correlation; the arrowhead, the correlation predicted by a simple polygenic model. (*After T. J. Bouchard, Jr., and M. McGue, Familial studies of intelligence: A review, Science, 212(4498):1055–1059, 1981*)

of twins since infancy as part of a longitudinal study of mental development. The results showed that developmental profiles—the spurts and lags in mental development over time, as measured by several age-appropriate tests—are more similar in identical twins than in fraternal twins. This evidence suggests that genes turn on and off during the course of mental development and influence behavior much the same way as they influence physical development (for example, children also show spurts and lags in growth). The Louisville data also show that the degree of similarity between identical twins, and between parents and offspring, increases over time; the degree of similarity between fraternal twins, while high very early in life, decreases. The correlations between both types of twins at age 15 are very close to those expected under a simple polygenic model.

These synchronous patterns of change show that mental development is under the control of genetic processes that unfold throughout childhood and adolescence. While the environment provided by parents is important in allowing children to maximize their developmental potential, the pattern of growth and the ultimate degree of similarity between relatives strongly reflect the effects of their shared genes. Identical twins reared in different families from a very young age are almost as similar on some traits as identical twins reared together.

Personality. Personality inventories measure such characteristics as introversion and extroversion, well-being, aggression, risk-taking and traditionalism. Many psychologists, while accepting the possibility of genetic influences on intelligence and mental abilities, believe that common family environmental factors are the principal determinants of personality traits. The evidence is, however, quite to the contrary. A number of adoption studies have demonstrated only modest degrees of similarity between genetically unrelated individuals who are reared together. A study comparing reared-apart and reared-together identical twins on a number of personality, psychological interest, and social attitude scales found that the correlations on each measure were very similar for both groups of twins (see **table**). These findings suggest that genetic differences

Intraclass correlations for Minnesota identical twins reared apart and together[*]

	Reared apart		Reared together		
Variables	Ratio	Number of pairs	Ratio	Number of pairs	Reared apart/ reared together ratio[†]
Personality variables					
Mean of 11 Multidimensional Personality Questionnaire scales	0.50	44	0.49	217	1.02
Mean of 18 California Psychological Inventory scales	0.48	38	0.49	99	0.98
Psychological interests					
Mean of 23 Strong Campbell Interest Inventory scales	0.39	52	0.48	116[‡]	0.81
Mean of 17 Minnesota Occupational Interest scales	0.40	40	0.49	376	0.82
Social attitudes					
Mean of 2 religiosity scales	0.49	31	0.51	458	0.96
Mean of 14 nonreligious social attitude items	0.34	42	0.28	421	1.21
Traditionalism scale	0.53	44	0.50	217	1.06

[*] After T. J. Bouchard, Jr., et al., Sources of human psychological differences: The Minnesota Study of Twins Reared Apart, *Science*, 250:223–228, 1990.

[†] Ratios of < 1.0 suggest some effect of shared environment.

[‡] This value is for 116 studies, not pairs.

account for about half the group variability for these characteristics and that common family environment did not make the reared-together twins any more similar than the reared-apart twins. The remainder of the variability (and the reason the identical twins do not have a correlation of 1.0) must be due largely to the unique or unshared environments experienced by each twin. *See* PERSONALITY THEORY.

Psychopathology. While twin, family, and adoption studies support the role of the genes in the development of schizophrenia and the affective disorders (depression and manic-depression), the environment plays an important role: among identical twin pairs in which one twin is affected, the cotwin is often unaffected. However, if the offspring of members of such discordant twin pairs are examined, the risk of being affected is the same in the offspring of the unaffected twin as in those of the affected twin, demonstrating that the unaffected twin did indeed carry the disease alleles. Early excitement generated by reports of mapping schizophrenia to chromosome 5 and manic-depressive illness (or bipolar disorder) to chromosomes 11 and X has been curbed by failure to replicate findings in other affected families or even in the same families at a later time. While reports of linkage of manic-depressive illness to the X chromosome have seemed more robust than linkage of the disorder to chromosome 11, a reanalysis of several large families finds weakened support for a gene on the X chromosome in two of the three families. Reasons for the revised findings in these studies of psychopathologies include the possibility of incorrect diagnoses in previous studies; individuals becoming affected who were not previously, since many behavioral disorders often do not show up until adulthood; and the use of DNA markers for the later analyses (earlier ones were often based on phenotypic markers). Availability of informative DNA markers for all regions of all the chromosomes will eventually be realized, in large part through the Human Genome Initiative. Researchers will continue to look for genes that contribute to these disorders. *See* AFFECTIVE DISORDERS; SCHIZOPHRENIA.

Alcoholism. As in the psychopathologies, data accumulated over many years from twin, family, and adoption studies point to genetic influences on alcoholism. This view is also supported by the ability to breed selectively for sensitivity to ethanol and for ethanol-related behaviors in animals such as rats and mice. In humans, certain alleles of genes that code for the alcohol-metabolizing enzymes, alcohol dehydrogenase and aldehyde dehydrogenase, are found in different frequencies in alcoholic and nonalcoholic Chinese men. In spite of these findings, there has been little evidence for linkage of alcoholism to a particular gene or genes. Alcoholism is a very complex disorder, and it is expected that many genes, some conferring primary risks and others that interact with them, as well as numerous environmental variables and gene–environment interactions, are contributory. *See* ALCOHOLISM; PSYCHOLOGY.

Kimerly J. Wilcox; Thomas J. Bouchard, Jr.

Bibliography. E. L. Gershon and R. O. Reider, Major disorders of mind and brain, *Sci. Amer.*, 267(3):126–133, 1992; R. Plomin, J. Owen, and P. McGuffin, The genetic basis of complex human behavior, *Science*, 264:1733–1739, June 1994; S. Scarr, Developmental theories for the 1990s: Development and individual differences, *Child Dev.*, 63:1–19, 1992; R. S. Wilson, The Louisville twin study: Developmental synchronies in behavior, *Child Dev.*, 54(2):298–316, 1983.

Behavioral ecology

The branch of ecology that focuses on the evolutionary causes of variation in behavior among populations and species. Thus it is concerned with the adaptiveness of behavior, the ultimate questions of why animals behave as they do, rather than the proximate questions of how they behave. The principles of natural selection are applied to behavior with the underlying assumption that, within the constraints of their evolutionary histories, animals behave optimally by maximizing their genetic contribution to future generations. Several aspects of both individual and social behavior have been analyzed from this perspective as the following examples illustrate.

Behavioral homeostasis. Animals must maintain their internal physiological conditions within certain limits in order to function properly, and often they do this by behavior. Small organisms may avoid desiccation by living under logs or by burrowing. Many insects must raise body temperatures to 86–95°F (30–35°C) for effective flight, and achieve this by muscular activity, such as the shivering of butterflies in the early morning, or by orienting to the Sun. Butterflies spread their wings to bask in the sun and close them when warmed up. Dragonflies, grasshoppers, and lizards orient themselves perpendicular to the Sun's rays to warm up and in parallel to cool down. Such behavior may raise body temperatures 18–27°F (10–15°C) above the surrounding air. Desert and grassland mammals burrow to avoid high or low temperatures. The elaborate burrow system of prairie dogs provides an efficient cooling system. Some burrow mouths are in mounds so that as warm air rises from them, cool air is sucked in via lower entrances a short distance away. Honeybees regulate hive temperature by fanning at the entrance or fetching water for evaporative cooling.

Habitat selection. Animals can be very selective about where they live. In England, for example, the peppered moth occurs in two forms; a black form lives in sooty industrial areas and a mottled or "peppered" form in pollution-free woodlands. The black form prefers to rest on dark tree trunks and the peppered form on gray patches of lichens. By making these choices, each matches its background and is protected from predators. The mechanisms of habitat choice are often genetically influenced, but may be modified by experience. Cases in point are prairie and woodland deermice, which prefer their own habitats, but less strongly so if each is reared in the habitat of the other. *See* PROTECTIVE COLORATION.

An extreme case of habitat selection is migration, which evolves largely in impermanent habitats, allowing animals to escape deteriorating conditions, colonize new areas, and exploit seasonal breeding sites. Insects, such as water striders, common in ponds and streams, have evolved winged and wingless forms. The former occur in temporary ponds or intermittent streams, while the latter are found on permanent ponds or lakes. Some species have a wingless summer generation which channels energy to reproduction rather than flight, and a winged fall generation which flies to protected overwintering sites. Monarch butterflies that breed in eastern North America fly thousands of miles to overwinter in the mountains of Mexico. Insect migrants are adapted for colonizing by their ability to fly and to reproduce rapidly.

The migrations of birds may also have evolved to allow escape from impermanent habitats. Birds overwintering in the tropics usually occupy marginal or stressed environments in their winter ranges. They also exploit temperate zone habitats only seasonally available. The latter behavior is also true of fish. The breeding sites of oceanic herring, for example, occur in seasonal shallow-water habitats, and the fish migrate to nursery and feeding areas. The streams to which salmon migrate are also seasonally limited. The explosive and once-only reproduction of Pacific salmon may result because repeated migrations up long steep rivers are too costly. *See* MIGRATORY BEHAVIOR.

Foraging. All animals must find food to survive and reproduce. Caterpillars graze on plants, predatory starfish pull mollusks from their shells, and chimpanzees use sticks as tools to draw ants from their nests. Despite the diversity of foods and feeding behavior, ecological theory predicts that natural selection should favor animals that forage efficiently. Those who get the most food energy in the least time will have the most time and energy available for other activities.

Animals are constantly faced with choices about what to eat and where and how to forage. Northwestern crows, for example, feed on shellfish during low ocean tides. Crows have been observed picking up whelks, flying 15 ft (5 m) above the ground, and dropping their prey on the rocks to break open the shells and expose the meat. Crows pick only the largest whelks, which break more easily and provide more food than small whelks. Moreover, 15 ft (5 m) is the minimum height to ensure that whelks will break on the first drop. Northwestern crows, then, apparently choose a foraging strategy that provides them with the maximum energy for the minimum effort.

Nutrient requirements and predation pressures may alter an animal's feeding strategy. Moose, for example, need sodium in their diet. Sodium is most abundant in energy-poor aquatic plants; moose divide their foraging time between aquatic plants and energy-rich terrestrial plants. Seed-eating dark-eyed juncos must constantly stop foraging and scan the skies for predatory hawks. The time that juncos divide between feeding and watching reflects the trade-off between getting necessary food energy and ensuring that they will survive.

Groups. Animals of all sizes and from many taxa form groups such as herds, troops, schools, and pods (of whales). A major theory as to why they do this is the "selfish herd" hypothesis; that is, animals will group in order to keep another individual between themselves and a predator. It has been found that guppies in Trinidad streams are influenced by predation. Fish from streams with predators respond more quickly to a strange object by schooling and school more than fish from streams without predators. The behavior is at least partly genetic since it persists in laboratory descendants of the original guppies, implying that the schooling is a product of natural selection. Among birds, single pigeons are more likely to be caught by a hawk than are pigeons in a flock even though the single birds spend more time on the alert. *See* POPULATION ECOLOGY.

Animals may also feed more efficiently in groups as several individuals searching for food increases the likelihood of finding some. Predators catch larger prey by cooperative hunting. Solitary leopards take prey about half their own size, while the highly social Cape hunting dog can bring down wildebeest and zebra 10 times larger than themselves. *See* ECOLOGICAL COMMUNITIES.

Territoriality. When resources, such as food and mates, are limited, members of a species will compete for access to them. One form of competition is when animals defend resource-rich areas, called territories, from conspecifics. Territorial behavior is very widespread. For example, female poplar aphids defend tiny territories at the base of poplar leaves. These are prime sites for establishing galls (in which offspring are produced), and there are many more females than available leaves. Similarly, male songbirds form spring territories in food-rich habitats. Females are attracted to these territories because abundant food supply enables them to raise offspring. Not all males manage to hold territories, and most females mate only with territorial males.

Territory holders gain exclusive access to valuable resources, but they also incur costs. Territorial defense requires time and energy, as exemplified by the calls and chases that male songbirds use to repel intruders. The conspicuous behavior of defenders may also increase predation risks. Animals should be expected to form territories only when the benefits of territoriality exceed the costs. Sandpipers, for example, are shorebirds which form feeding territories when their prey (small crustaceans) are at moderate densities but not when the prey are at very low or very high densities. At low densities, prey are too scarce to be worth defending, and at very high densities there are so many sandpipers trying to feed that territorial defense may be too costly. *See* TERRITORIALITY.

Reproduction. Males produce many small gametes (sperm), while females produce fewer, large gametes (eggs). As a result, males and females differ in their strategies of reproduction. Males can potentially produce many offspring by mating with several females, while females may benefit by ensuring that their relatively few offspring are of high quality. Males, then, generally compete for females. Selection for male mating ability, termed sexual selection, is a powerful evolutionary force.

Sexual selection underlies the elaborate courtship displays seen in many species. Male sage grouse, for example, have evolved bright plumage which they display in an effort to attract females. Sexual selection may favor females which choose high-quality males. Male hangingflies, for example, present gifts of prey (small insects) to prospective mates. It has been found that the prey vary in size and nutritional quality and that females preferentially copulate with males that bring the most valuable prey.

Animals have a variety of mating systems. In monogomous animals, males and females mate with one partner per breeding season, and parental care is often provided by both parents. Monogamous animals are frequently found in habitats with scattered resources. Many species are polygynous; that is, males mate with several females per breeding season; females mate with one male; and parental care is usually provided by females. Polygynous animals are frequently found in habitats with rich patches of resources. In polyandrous animals, females mate with several males per breeding season, while males mate with one female; parental care is usually provided by males. Polyandrous species, although generally rare, are most common in habitats with rich resources and high predation on offspring. In promiscuous animals, males and females mate with several partners per breeding season, and parental care is provided by males or females. Promiscuous animals are found both in habitats with scattered resources and highly synchronized female breeding and in habitats with unpredictable resources and high predation on offspring.

Most mammals are polygynous. A few males outcompete their rivals by defending groups of females or the resources which attract females to territories. The dominant males mate frequently, most males do not mate at all, and each female mates only once. Ecological constraints, however, often limit the ability of a few males to monopolize females. In most bird species, for example, females breed synchronously, and males have little opportunity to mate more than once. Young birds often require feeding by both parents, so the best male strategy is to stay with one female throughout the breeding season.

Parental care is common in birds and mammals, but examples can be found in almost all taxa. Parental care may include both provision of food and protection from predators. Mother lace bugs, for example, guard their nymphs against predators and greatly increase the chance that the nymphs will survive to adulthood. In general, females more frequently provide care than do males. Male parental care is most common when males are certain of their paternity. Thus, in bony fishes, species with male parental care tend to have external fertilization, and those with female parental care have internal fertilization.

In several species, nonbreeding animals help raise the offspring of their conspecifics. Breeding pairs of Florida scrub jays, for example, are aided by up to six helpers. Helpers are usually older siblings of the nestlings and are unable to establish their own breeding territories. They may benefit by gaining breeding experience and by ensuring the success of their relatives. The theory of kin selection, an extension of natural selection, predicts that animals should preferentially help kin, because close relatives share in common a large proportion of their genes. *See* BEHAVIOR GENETICS; ETHOLOGY; REPRODUCTIVE BEHAVIOR.

Hugh Dingle; Peter Frumhoff

Bibliography. J. Alcock, *Animal Behavior: An Evolutionary Approach*, 6th ed., 1997; J. R. Krebs and N. B. Davies, *An Introduction to Behavioral Ecology*, 1981; D. H. Morse, *Behavioral Mechanisms in Ecology*, 1980.

Behavioral psychophysics

The use of behavioral methods to measure the sensory capacities of animals and nonverbal humans such as infants. The observations are analogous to those obtained from adult human subjects who are asked to observe appropriate stimuli and report what they see. The behavioral methods differ primarily in that stimulus-response associations are established by means other than verbal instructions, either as unlearned reflexes or through conditioning.

Measures. Any sense or species may be studied, but most work has been done on the vision and hearing of primates, cats, pigeons, and rats. Typical investigations determine (1) the absolute threshold (the minimum intensity needed to elicit a standard response); (2) the difference threshold (the minimum change in a stimulus needed to elicit a standard response); and (3) points of apparent equality (values of stimuli that elicit no response because a change in one aspect compensates for a change in another). A few investigations have determined stimulus scales that express quantitative relations between the physical stimulus and the perceptual effect over a range of stimulus values. These various measures provide a picture of the sensory function, such as visual acuity, color sensitivity, loudness or pitch perception, and odor discrimination.

Methods. Efficiency and sensitivity to the sensory function of interest are the major factors that govern the choice of method in behavioral psychophysics.

Reflex methods. These methods are the most convenient since no training is required. An example is the use of the optokinetic reflex to estimate visual acuity. Here, the subject is exposed to a large field of moving stripes. Such a field elicits in most species a reflex turning of the head and eyes to follow the motion of the stripes. The stripes are made finer and finer until the turning response ceases, defining an acuity threshold. Another example is the preferential looking response in infants. Without training, infants spend more time looking at patterned stimuli than at blank fields. Two stimulus fields are placed before the infant, and the relative time spent looking at each is determined. A preference for one pattern according to this measure indicates that the infant can detect a difference between the two.

Unconditioned reflex methods are limited to sensory functions for which appropriate reflexes may be found; also, they usually impose severe limits on the specific stimulus conditions that may be used. Conditioning methods add considerable flexibility. In Pavlovian conditioning, the stimulus of interest becomes the conditioned stimulus through its association with a stimulus that elicits a clear-cut reflexive response (the unconditioned stimulus). For example, the sensitivity of an animal to air-pressure changes may be determined by causing pressure changes to precede the application of a mild electric shock. Since they signal shock, the pressure changes come to elicit heart-rate acceleration as the indicator response. The absolute threshold is that pressure change just strong enough to elicit a heart-rate acceleration.

Operant conditioning. This method offers still more flexibility. Typically the subject is rewarded for making an indicator response in the presence of one stimulus value; reward is withheld (or another response rewarded) in the presence of other stimulus values. The responses are selected to suit the species and the stimulus under study. For example, sucking and head turning have been employed in visual experiments with infants; bats have been taught to fly to sound sources; rats have pushed panels with their nose or jumped to platforms marked by the stimulus value in question.

A specific experiment indicates general aspects of behavioral psychophysics as well as the capabilities of the operant method. The purpose of this experiment was to determine the absolute sensitivity of the pigeon's eye continuously through time. The **illustration** diagrams the procedure. The pigeon was trained to work in the dark, positioning its head through a hole. Two response keys, A and B, sensed the pigeon's pecks. When the stimulus light was turned off, the bird was rewarded with food for pecking key B. Pecks on key A were rewarded indirectly: they periodically turned the stimulus off by closing a shutter, thus producing the condition under which pecking key B produced food. In this way, the pigeon learned to peck A only when the light was visible and B only when the light was invisible.

In addition to producing reward, pecks to A and B were made to control the intensity of the stimulus light through an automatic circuit. Between scheduled shutter closures, pecks on key A reduced stimulus intensity by moving a variable density optical wedge. Thus, when the bird could see the light, it pecked key A, reducing the light's intensity until this fell below the bird's absolute threshold. This disappearance of the light could not be discriminated from total darkness (shutter closure), so the bird

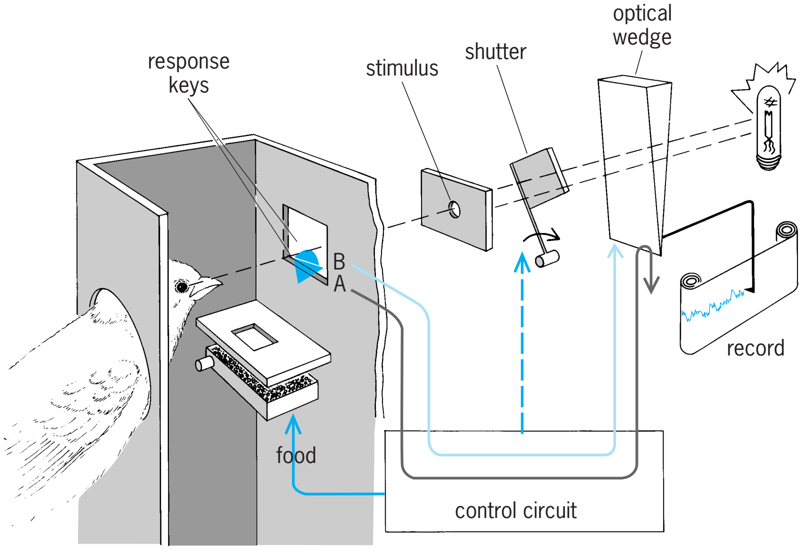


Diagram of the apparatus for measuring the pigeon's absolute visual threshold. (*After D S. Blough, A method for obtaining psychophysical thresholds from the pigeon, J. Exp. Anal. Behav., 1:31–43, 1958*)

now pecked key B, raising the stimulus to visibility once again. The pigeon's pecks thus caused the stimulus intensity to oscillate up and down across the bird's absolute threshold. Periodic shutter closures and feeding maintained the appropriate response pattern. This means of tracking the threshold through time enables one to study the changes in visual sensitivity, such as those that are induced by exposure to bright light or to certain drugs.

Significance. Full understanding of a sensory system can be achieved only when the relevant structures (such as the eye) and physiological mechanisms (such as neural activity in the visual system) are related to the overall functioning of the system as expressed in the behavior of the whole animal (such as visually guided behavior). In most instances, structure and physiology may best be studied in a nonhuman subject. Behavioral psychophysics in this species then provides coordinate information on the operation of the system as a whole. By far the largest number of studies using the methods described here have been motivated by this desire to understand a sensory system at several levels in a single species.

Interest in the development of sensory function has spurred the use of behavioral methods in infants and young children, while the effects of controlled sensory input during development have been extensively monitored in nonhuman subjects. Applications to the prevention and control of sensory disorders also are increasing. For example, a number of toxicants, drugs, and environmental stressors have been related to sensory disorders; the methods of behavioral psychophysics are used to follow the development of these disorders under controlled conditions, and to uncover potential preventive and therapeutic measures. *See* PSYCHOLOGY; SENSATION.

Donald Blough

Bibliography. H. Geissler et al., *Cognition, Information Processing and Psychophysics: Basic Issues*, 1989.

Behavioral toxicology

The study of behavioral abnormalities induced by exogenous agents such as drugs, chemicals in the general environment, and chemicals encountered in the workplace. Just as some substances are hazardous to the skin or liver, some are hazardous to the function of the nervous system. In the case of permanent effects, changes in sensation, mood, intellectual function, or motor coordination would obviously be undesirable, but even transient alterations of behavior are considered toxic in some situations. For example, operating room personnel accidentally exposed to very small doses of anesthetic do not exhibit altered performance on an intelligence test, a dexterity test, or a vigilance task. However, significant decrements in performance occur in recognizing and recording visual displays, detecting changes in audiovisual displays, and recalling series of digits.

By comparing the behavior of exposed subjects and control subjects, behavioral toxicologists seek to identify agents capable of altering behavior and to determine the level of exposure at which undesirable effects occur. When the agent under study is one in common use, and there is no evidence of its being hazardous to health, experiments may be carried out on human volunteers, or comparisons may be made from epidemiologic data. More frequently, safety considerations dictate the use of laboratory animals in toxicology research. Primates are favored for their wide behavioral repertoire, trainability, and long lifespan, but they are extremely expensive to purchase and maintain. Rodents can be produced and maintained more economically, and their short gestation period and rapid development are advantageous in some types of research. Scientists often investigate toxic effects in several species to take advantage of the useful characteristics of each and to increase the generality of their conclusions.

Humans. Perhaps the best-known example of toxicity in humans is methyl mercury poisoning (Minimata disease), which occurred in epidemic proportions in a Japanese coastal town where the inhabitants ate fish contaminated with mercury from industrial pollution. Although mercury affects a variety of behaviors, the most obvious symptoms are tremors and involuntary movements.

A different set of functional problems is exemplified by the effects of ethyl alcohol, a single agent with direct and indirect, short- and long-term consequences. The short-term, low-dose effects of alcohol include sensory disturbances, motor problems, and difficulties with processing information. Neurologically, alcohol is usually described as a central nervous system depressant which is general, in the sense that it disrupts many functions.

In some individuals, large quantities of alcohol consumed over a long period lead to permanent damage to the nervous system. Behaviorally, individuals with Korsakoff's syndrome exhibit severe memory deficits. Anatomically, their brains are found to have degenerative changes in the thalamus. This syndrome is not just an extension of the short-term effects. In fact, it is thought to arise from alcohol-induced malnutrition rather than as a direct effect of the alcohol itself.

Lasting injuries to the nervous system have been reported to occur in children exposed to alcohol before birth. The behavioral problems associated with fetal alcohol syndrome do not appear to be related to either Korsakoff's syndrome or the immediate effects of alcohol. Rather, they constitute a third set of effects, including learning deficits, problems with inhibiting inappropriate behavior, and fine motor dysfunction, along with some visible physical abnormalities. While malnutrition may play a role in this congenital syndrome, the mechanism and locus of damage are not known. *See* ALCOHOLISM; CONGENITAL ANOMALIES; FETAL ALCOHOL SYNDROME.

Mature animals. In the controlled environment of the laboratory, researchers can deliver carefully measured doses of test agents and manage the subjects' physical condition and experiential history. In addition, many aspects of behavior can be quantified. For example, whereas reports of family members can be used to judge the activity level of a child at home, a laboratory rat can be housed in a cage designed to record the animal's movements, giving the experimenter an objective measure of activity. However, the specificity of laboratory studies creates a major problem for the behavioral toxicologist, for it is not possible to test all behaviors of an animal. Therefore, the researcher must decide which behaviors are to be evaluated. The problem of test selection is serious, because alterations of function can be very selective, leaving many behaviors normal while a few are grossly abnormal. A common strategy is to sample a variety of functions that appear to be unrelated, in the hope of covering as many functional systems as possible. For example, a series that included tests of vision, hearing, coordination, and strength would be more likely to detect toxic effects than a series of four visual tasks.

Another approach to the problem of test selection is to include some tests that require the integrity of many systems. For example, scientists at the National Institute of Occupational Safety and Health employ an elaborate operant task for primates. The animal sits in a chamber from which it can see a visual display panel. Lights signal that a trial is about to begin. To receive a food reward, the animal must grasp a spring-mounted lever, move it to a specified position, and hold it there for several seconds. A series of lights signals when the lever is in the proper position and tells the subject how long it must wait for the reward. While this task was designed as a very sensitive measure of motor control and tremor, abnormalities of attention, time perception, memory, or vision could also interfere with performance. A task like this requires a long period of training, but the trained animal can be used over and over, as long as it sustains no permanent injury from exposures to suspected toxic agents. When exposure is likely to cause permanent alterations of behavior, it is more economical to employ tests which require less investment of time and less expensive animals.

Developing animals. The study of congenital behavioral deficits is known as behavioral teratology. In these experiments, animals are exposed before birth or soon after birth and are later tested for behavioral abnormalities. As in the example of fetal alcohol syndrome, it is often impossible to predict the effects of an agent on the developing brain from its effects on the adult brain. Embryos and neonates survive some insults, such as anoxia, better than adults. Conversely, some agents that are well tolerated by mature animals are disastrous to developing ones. For example, agents that interfere with cell proliferation are not hazardous to the mature brain because neurons do not multiply in adults. But in the developing brain, interference with proliferation, as by irradiation with x-rays, can permanently alter the number of neurons and consequently the adult behavior of exposed animals. Regulations regarding exposure to radiation reflect this difference, limiting the exposure of pregnant women. Some agents which have reversible effects in adults may have permanent effects when administered to young animals. For example, several psychoactive drugs are believed to alter the developing nervous system of a young animal and lead to lasting behavioral changes.

Because an animal can be treated only once in studies of developmental toxicity, the typical experimental paradigm in this field employs many animals, often rodents. A behavioral test battery might begin with tests of reflex development in neonates. Other popular tests include open-field measures of emotionality and general activity, maze-learning problems, avoidance tasks, and sensory measures.

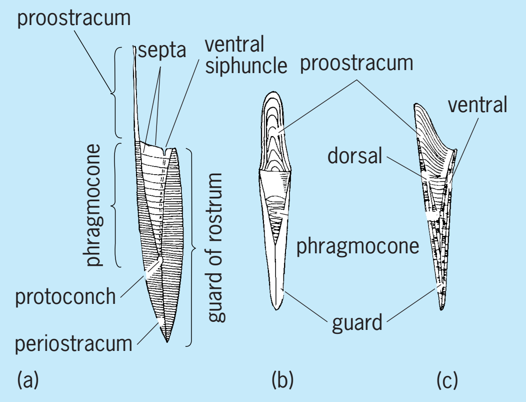
Significance of behavioral effects. When toxicity is considered only in terms of direct risk to survival, behavioral toxicity may seem to be of minor importance. However, survival is not the only criterion of good health. In a complex society that places heavy demands on an individual's educability, alertness, and emotional stability, even small deviations in behavior are potentially hazardous. Severe disabilities, as in the gross motor malfunctions of Minimata disease, have drawn attention to behavioral toxicology. Such incidents represent failures of control of toxic substances. Successes are difficult to measure, for they can be seen only in reduction of risk—the ultimate goal of toxicology. *See* TOXICOLOGY.

Patricia M. Rodier

Bibliography. Z. Annau (ed.), *Neurobehavioral Toxicology*, 1987; O. Hutzinger (ed.), *Reactions and Processes*, 1984; W. R. Notton et al. (eds.), *Health Surveillance of Individual Workers Exposed to Chemical Agents*, 1988; G. Zbinden et al. (eds.), *Application of Behavioral Pharmacology in Toxicology*, 1983.

Belemnoidea

An order of extinct dibranchiate cephalopods. These mollusks ranged from the Upper Mississippian to the upper Eocene. The oldest known representative is highly specialized, although the belemnoids are considered to be the most primitive of the dibranchiates. Belemnoids have a chambered



Belemnoidea. (*a*) Diagram of a belemnitid shell. (*b*) Ventral aspect and (*c*) lateral aspect of a belemnite. (*After R. R. Shrock and W. H. Twenhofel, Principles of Invertebrate Paleontology, 2d ed., McGraw-Hill, 1953*)

shell or phragmocone which fits into a guard or rostrum, the apical portion of the shell (see **illus.**). *Belemnites* is an important index fossil. *See* CEPHALOPODA; INDEX FOSSIL. Charles B. Curtin

Bell

A hollow metallic cylinder closed at one end and flared at the other. A bell is used as a fixed-pitch musical instrument or a signaling device, frequently in clocks. A bell is set vibrating by a clapper or tongue which strikes the lip. The bell has two tones, the strike note or key, and the hum tone, which is a major sixth below the strike note. Both tones are rich in harmonics.

The finer bells are made for carillons. Small carillons have two full chromatic octaves and large ones have four or more chromatic octaves, the bell with the lowest tone being called the bourdon. The Riverside Church in New York has 72 bells, the largest or bourdon weighing 18.25 tons (16.6 metric tons). All bells of a carillon are shaped for homogeneity of timbre.

Bell metal is approximately four parts copper and one part tin, although zinc, lead, or silver may be used. The shape of a bell—its curves and full trumpet mouth—evolved from experience. The bow or edge has a thickness of one-twelfth to one-fifteenth its diameter. The contour of a new bell is formed in the molds by rotating strickle boards, one for the core or inner mold and one for outer surface (**Fig. 1**). A dummy or false bell is made of brittle clay over the core; any required inscription is placed in relief with wax on the dummy. The cope or outer mold is then built up on the dummy, first with a thin soup clay mixture to fit the inscription, then with thicker mixtures.

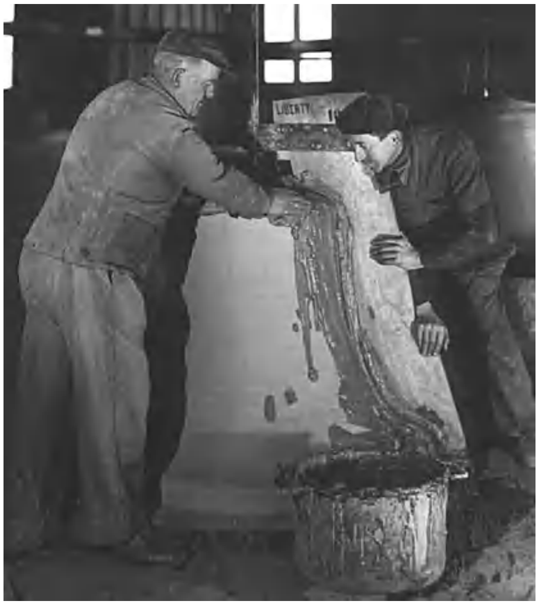


Fig. 1. Internal core of bell, which is shaped by revolving a strickle board. (*Les Fils de G. Paccard*)

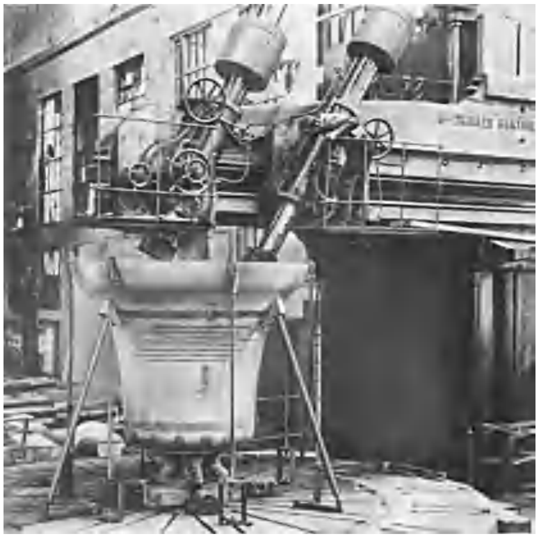


Fig. 2. A 10-ton (9-metric-ton) bell being milled after the molding process. (*Les Fils de G. Paccard*)

Molds and dummy are dried and baked to melt out the wax and harden the core and cope. The dummy is removed, the mold is locked in place, and the melted bronze is poured in and allowed to cool. An 18-ton (16-metric-ton) bourdon may take a quarter hour to pour and 36 to cool. The mold is removed and the bell trimmed, sandblasted, and possibly polished.

A bell can be pitched slightly. The tone is raised by grinding the outer surface, effectively decreasing the diameter, or lowered by grinding the inner surface (**Fig. 2**). Too much tuning of a well-shaped bell is apt to degrade its voice or timbre. Frank H. Rockett

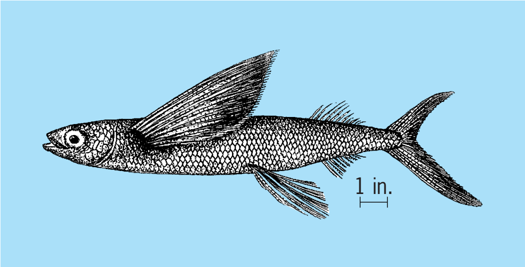
Belladonna

The drug and also the plant known as the deadly nightshade, *Atropa belladonna*, which belongs to the nightshade family (Solanaceae). This is a coarse, perennial herb native to the Mediterranean regions of Europe and Asia Minor, but now grown extensively in the United States, Europe, and India. During the blooming period, the leaves, flowering tops, and roots are collected and dried for use. The plant contains several important medicinal alkaloids, the chief one being atropine, which is much used to dilate the pupils of the eye. *See* ATROPINE; SOLANALES.

Perry D. Strausbaugh; Earl L. Core

Beloniformes

An order of actinopterygian fishes that comprise, with the Atheriniformes and Cyprinodontiformes, the series Artherinomorpha. These fishes are identified by the interarcural cartilage that connects the first and second gill arches; small second and third epibranchials; absence of interhyal; a nonprotrusible upper jaw; far posterior placement of the dorsal



California flyingfish (*Cypselurus californicus*). 1 in. = 2.5 cm. (*After G. B. Goode, Fishery Industries of the U.S., 1884*)

and anal fins; abdominal pelvic fins; and the lower caudal fin lobed with more principal rays than the upper lobe and significantly longer in needlefishes and flyingfishes. The order consists of 191 species in 38 genera and 5 families. Each family has distinctive features that make identification simple.

Adrianichthyidae. Adrianichthyids (ricefishes) are small fishes that inhabit fresh and brackish waters from India to Japan and to the Indo-Australian archipelago. Adults vary from 3 to 20 cm (1.2 to 8 in.) in length. They are best described by skeletal structures that they have lost. Absent are certain head bones, such as vomer, meta- and ectopterygoids, rostral cartilage, supracleithrum, and lateral lines. In one species, the smallest, the maxillae are absent as well as the right pelvic fin of females, and the mouth is very small, whereas in several of the larger species the jaws are greatly enlarged and the mouth functions like a scoop shovel.

Exocoetidae. This family, the flyingfishes, is familiar to seafarers in the tropical and subtropical Atlantic, Indian, and Pacific oceans. It is readily recognized by having unusually large pectoral fins that can be used for gliding above the ocean surface. Some species gain additional gliding ability from greatly enlarged pelvic fins. A flyingfish does not flap its "wings"; instead it glides on them after being launched into the air by powerful thrusts of the tail. Sustained "flight" is accomplished by repeated contacts of its elongate lower caudal fin lobe with the water. Worldwide there are 8 genera and about 52 species, 26 of which occur in blue waters of North America, including all of Mexico.

Hemiramphidae. Known as halfbeaks, these fishes are readily distinguished from other beloniforms by having the upper jaw much shorter than the lower; premaxillae pointed anteriorly, forming a triangular upper jaw; short pectoral and pelvic fins; and the tip of the long lower jaw bright red or orange in most species. Of the 13 genera and more than 100 species, most occur in the warm waters of the Atlantic, Indian, and Pacific oceans, but several genera, some of which are viviparous, live in freshwaters of Southeast Asia. Four genera and 12 species are known from the marine waters of Mexico and the United States, and a single freshwater species occurs in Mexico.

Belonidae. The family Belonidae (needlefishes) differs from the preceding three families in having both upper and lower jaw elongated (very rarely the upper jaw is short) and equipped with numerous sharp teeth. It differs from the Scomberesocidae in lacking dorsal and anal finlets. At hatching, both jaws are short and of equal length (as in ricefishes and adult flyingfishes). As development progresses, the lower jaw elongates first (halfbeak stage, as in adult halfbeaks), followed by elongation of the upper jaw (needlenose stage, typical of adult belonids). Some of the larger belonids, such as the hound fish, *Ablennes hians*, can skitter over the water surface at great speeds, potentially posing a danger to waders, especially at night when the fish are attracted by lights or when they find themselves trapped by a net. There are 10 genera represented by 21 epipelagic species in tropical and temperate seas, and 11 species restricted to freshwater. Ten species in 4 genera occur in North American waters and only one, the Maya needlefish of Mexico (*Strongylura hubbsi*), is restricted to freshwater. Another, the Atlantic needlefish (*S. marina*), ascends rivers of the Atlantic and Gulf coasts to the Fall Line, including rivers of Mobile Basin, where it reproduces.

Scomberesocidae. This family, called sauries, consists of only four monotypic genera. They are closely related to the needlefishes, but finlets following the dorsal and anal fins easily distinguish them from other beloniforms. The jaws are delicate and vary from long and slender to relatively short. Sauries are epipelagic in tropical and temperate seas. *See* ACTINOPTERYGII; ATHERINIFORMES; CYPRINODONTIFORMES; TELEOSTEI. Herbert Boschung

Bibliography. B. B. Collette, *Family Belonidae Bonaparte 1832—Needlefishes*, Calif. Acad. Sci. Annotated Checklists of Fishes No. 16, 2003; B. B. Collette et al., Beloniformes: Development and relationships, pp. 334–354, in H. G. Moser et al., (eds.), *Ontogeny and Systematics of Fishes*, Amer. Soc. Ich. Herp. Spec. Publ. No. 1, 1984; L. R. Parenti, *Relationships of Atherinomorph Fishes* (*Teleostei*), *Bull. Mar. Sci.*, 52(1):170–196, 1993; D. E. Rosen and L. R. Parenti, *Relationships of Oryzias, and the Groups of Atherinomorph Fishes*, Amer. Mus. Novit. 2719, 1981.

Belt drive

The lowest-cost means for transmitting power between shafts that are not necessarily parallel. Belts run smoothly and quietly, and they cushion motor and bearings against load fluctuations. Belts typically are not as strong or durable as gears or chains. However, improvements in belt construction and materials are making it possible to use belts where formerly only chains or gears would do.

Advantages of belt drive are: They are simple. They are economical. Parallel shafts are not required. Overload and jam protection are provided. Noise and vibration are damped out. Machinery life is prolonged because load fluctuations are cushioned (shock-absorbed). They are lubrication free. They require only low maintenance. They are highly efficient (90–98%, usually 95%). Some misalignment is tolerable. They are very economical when shafts are separated by large distances. Clutch action may be obtained by relieving belt tension. Variable speeds may be economically obtained by step or tapered pulleys.

Disadvantages include: The angular-velocity ratio is not necessarily constant or equal to the ratio of pulley diameters, because of belt slip and stretch. Heat buildup occurs. Speed is limited to usually 7000 ft/min (35 m/s). Power transmission is limited to 500 hp (370 kW). Operating temperatures are usually restricted to −31 to 185°F (−35 to 85°C). Adjustment of center distance or use of an idler pulley is necessary for wear and stretch compensation. A means of disassembly is provided to install endless belts.

Belt types and general use. There are four general types of belts, each with its own special characteristics, limitations, advantages, and special-purpose variations for different applications.

Flat belts. Flat belts, in the form of leather belting, served as the basic belt drive from the beginning of the Industrial Revolution. They can transmit large amounts of power at high speeds—up to 500 hp (370 kW) and 10,000 ft/min (50 m/s)—if wide belts are used over large pulleys. Such drives are cumbersome; they require high belt tensions, which in turn produce large bearing loads. For these reasons V-belts have generally replaced flat belts in power delivery applications. Flat belts find their widest application where high-speed motion, rather than power, is the main concern. Flat belts are very useful where large center distances and small pulleys are involved. They can engage pulleys on both inside and outside surfaces, and both endless and jointed construction is available.

V-belt. V-belts are the basic power-transmission belt. They provide the best combination of traction, operating speed, bearing load, and service life. The belts are typically endless, with a trapezoidal cross section which runs in a pulley with a V-shaped groove (**Fig. 1**). The wedging action of the belt in the pulley groove allows V-belts to transmit higher torque at less width and tension than flat belts. V-belts are far superior to flat belts at small center distances and high reduction ratios. Center distances greater than the largest pulley diameter and less than three times the sum of the two pulley diameters are preferred. Optimum speed range is 1000–7000 ft/min (5–35 m/s). V-belts require larger pulleys than flat belts because of their greater thickness. Several individual belts running on the same pulley in separate grooves are often used when the power to be transmitted exceeds that of a single belt. These are called multiple-belt drives.

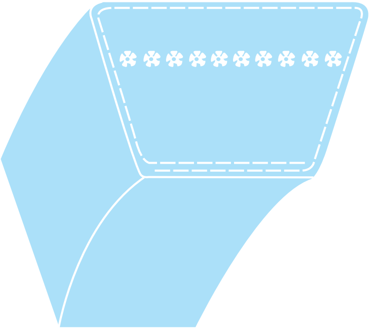


Fig. 1. V-belt in cross section.

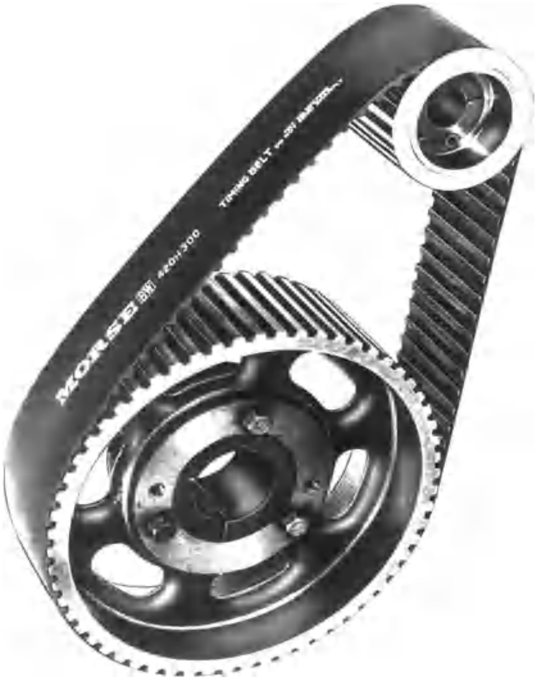


Fig. 2. Timing belt and pulleys. (*Morse Chain Division of Borg-Warner Corp.*)

Jointed and link V-belts are available when the use of endless belts is impractical. Jointed belts are not as strong and durable as endless and are limited to speeds of 4000 ft/min (20 m/s). A link V-belt is composed of a large number of rubberized-fabric links jointed by metal fasteners. The belt may be disassembled at any point and adjusted to length by removing links.

Film belts. Film belts are often classified as a variety of flat belt, but actually they are a separate type. Film belts consist of a very thin (0.5–15 mils or 100–4000 micrometers) strip of material, usually plastic but sometimes rubber. They are well suited to low-power (up to 10 hp or 7 kW), high-speed application where they provide high efficiency (as high as 98%) and long life. One of their greatest advantages is very low heat buildup. Business machines, tape recorders, and other light-duty service provide the widest application of film belts.

Timing belts. Timing belts have evenly spaced teeth on their bottom side which mesh with grooves cut on the periphery of the pulleys to produce a positive, no-slip, constant-speed drive, similar to chain drives (**Fig. 2**). They are often used to replace chains or gears, reducing noise and avoiding the lubrication bath or oiling system requirement. They have also found widespread application both in miniature timing applications and in automobiles. Timing belts, known also as synchronous or cogged belts, require

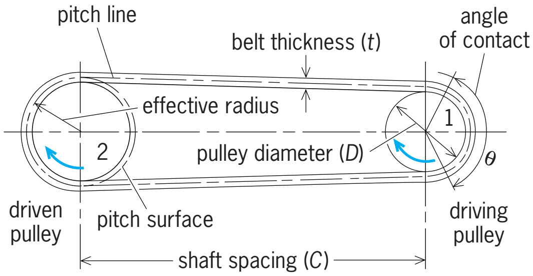


Fig. 3. Basic elements of open belt drive.

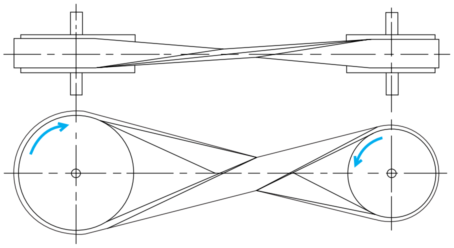


Fig. 4. Two views of crossed belt drive.

the least tension of all belt drives and are among the most efficient. They can transmit up to 200 hp (150 kW) at speeds up to 16,000 ft/min (80 m/s). There is no lower limit on speed. Disadvantages are high first cost, grooving the pulleys, less overload and jam protection, no clutch action, and backlash.

Belt-pulley arrangements. The most common arrangement, by far, is the open belt drive (**Fig. 3**). Here both shafts are parallel and rotate in the same direction. The cross-belt drive of **Fig. 4** shows parallel shafts rotating in opposite directions. Timing and standard V-belts are not suitable for cross-belt drives because the pulleys contact both the inside and outside belt surfaces. Nonparallel shafts can be connected if the belt's center line, while approaching the pulley, is in the pulley's central plane (**Fig. 5**). *See* ROLLING CONTACT.

Belt materials. Industrial belts are usually reinforced rubber or leather, the rubber type being predominant. Nonreinforced types, other than leather, are limited to light-duty applications. General belt material properties are given in the **table**.

Speed relationship between pulleys. Accurate calculation of the speed relationship between pulleys is based on pitch diameter, the effective pulley diameter. As a belt bends around a pulley, the outer fibers are in tension, the inner in compression. In between is the pitch line, which is subject to neither stress nor change in length (Fig. 3). For film and flat belts, the pitch line is midway through the belt surface. For timing and V-belts, the pitch line depends on cross-section shape and size. The factors to convert pulley diameters to pitch diameters are typically listed in engineering handbooks and manufacturer's literature. The speed relationship between pulleys is given by Eq. (1), where N_1 is the angular speed of pulley 1,

$$\frac{N_1}{N_2} = \frac{PD_2}{PD_1} \tag{1}$$

N_2 is the angular speed of pulley 2, PD_1 is the pitch diameter of pulley 1, and PD_2 is the pitch diameter of pulley 2.

The actual pulley speeds are often 0.5–1% less than predicted by Eq. (1), because of belt slip and stretch. The exact speed relationship for timing belts is also the inverse ratio of the number of teeth on the pulleys.

The speed in m/s of the belt itself is given by Eq. (2), where PD is the pitch diameter of the given pulley in rpm.

$$V = \pi(PD)(N) \tag{2}$$

Selection criteria. To design a belt drive, the following information is required: speeds of drive and driven unit; power transferred between drive and driven unit: power (kW) = torque (newton-meters) × rpm × 1.05×10^{-4}; desired distance between shafts; and operating conditions.

The drive horsepower is corrected for various conditions such as speed ratio; distance between shafts (long or short); type of drive unit (electric motor, internal combustion engine); service environment (oily, wet, dusty); driven unit loads (jerky, shock, reversed); and pulley-belt arrangement (open, crossed, turned). The correction factors are listed in engineering handbooks and manufacturer's literature. The corrected horsepower is compared to rated horsepowers of standard belt cross sections at specific belt speeds to find a number of combinations that will do

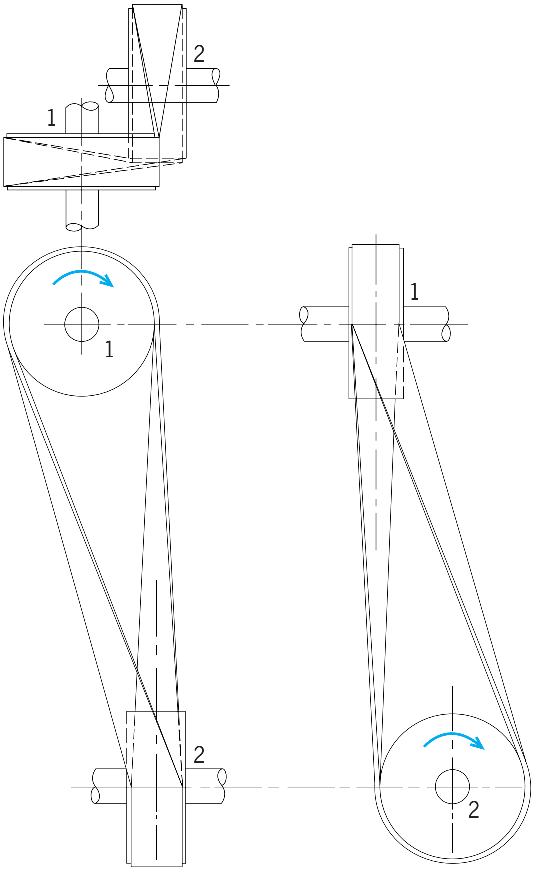


Fig. 5. Quarter-turn belt.

Belt material properties

Material	Properties	Uses
Nonreinforced rubber	Low-power and low-speed applications; very useful for fixed-center drives	Flat and film belts, vulcanized joints
Nonreinforced plastics	Low-power and low-speed application	Flat, film, and miniature timing belts, molded endless
Nonreinforced leather	High power; long service life; high cost; must be cleaned and dressed; prone to stretching and shrinking; limited to low and moderate speeds	Flat belts only; often laminated, with glued, riveted, or wire-laced joints
Fabric	Ability to track uniformly; low friction is a disadvantage	Endless flat belts constructed from cotton or synthetic fabric, with or without chemical or rubber coatings
Steel	Very high-speed, low-power applications; high tension required	Thin steel bands with riveted or silver-soldered joints
Reinforced rubber, leather, and plastics	Nylon, polyamide, glass fiber, or steel wire tensile members covered with rubber, leather, or plastic; speed and power capabilities are greatly improved by the reinforcement	All but film belts
Rubberized fabric	Consists of plies of cotton or synthetic fabric impregnated with rubber; low cost; good abrasion resistance, but has a lower power-transmitting ability and service life than leather; limited to speeds below 6000 ft/min (30 m/s)	All but film belts
Rubberized cord	Constructed with rubber-impregnated cords running lengthwise; carries 50% more power than rubberized fabric belts; fair shock-absorption capacity	Timing, flat, and V-belts, endless only
Rubberized cord and fabric	Combines strength of cord with abrasion resistance of fabric; well suited to minimum-stretch, high-speed operations	All but film belts, endless only

the job. At this point, the pulley diameters are chosen. A trade-off must be made between larger pulley diameters or larger belt cross sections, since larger belts transmit this same power at low belt speeds as smaller belts do at high speeds. To keep the drive unit as small as possible, minimum-diameter pulleys are desired. Minimum pulley diameters are limited by the elongation of the belt's outer fibers as the belt bends around the pulleys. Small pulleys increase this elongation, greatly reducing belt life. Minimum pulley diameters are often listed with each cross section and speed, or listed separately by belt cross section. Once the most economical pulley diameters and belt section are chosen, the belt length is calculated. If endless belts are used, the desired shaft spacing may need adjusting to accommodate standard length belts. It is often more economical to use two or more V-belts side by side, rather than one larger belt.

If a large speed ratio or a small distance between shafts is used, the angle of contact between the belt and pulley may be less than 180°. If this is the case, the drive power must be further increased, according to manufacturer's tables, and the selection process repeated. The drive power must be increased because belt power capacities are specified assuming 180° contact angles. Smaller contact angles mean less area for the belt to obtain traction, and thus the belt carries less power.

Belt specifications. To fully specify a belt, the material, length, and cross-section size and shape are required. Timing belts, in addition, require that the size of the teeth be given.

For all belt types, and for open and crossed belt drives, an accurate approximation of belt length is Eq. (3). (Symbol definitions are given in Fig. 3.) The

$$L = 2C + 1.57(D_2 + D_1) + \frac{(D_2 \pm D_1)^2}{4C} \qquad (3)$$

sign of the second D_1 is plus for crossed belt drives and minus for open belt drives.

The angle of contact in radians between belt and pulley (Fig. 3) is given by Eq. (4); the plus sign is

$$\theta = \pi \pm 2 \arcsin\left(\frac{D_2 - D_1}{2C}\right) \qquad (4)$$

for the larger pulley, the minus sign for the smaller pulley.

For the crossed belt drive, both angles of contact are calculated by Eq. (5).

$$\theta = \pi + 2 \arcsin\left(\frac{D_2 + D_1}{2C}\right) \qquad (5)$$

When correcting for angle of contact, the smallest angle calculated should always be used.

Belt tension. Belts need tension to transmit power. The tighter the belt, the more power transmitted. However, the loads (stress) imposed on the belt and supporting bearings and shafts are increased. The best belt tension is the lowest under which the belt will not slip with the highest load condition. Belt manufacturers recommend operating belt tensions according to belt type, size, speed, and pulley diameters. Belt tension is determined by measuring the

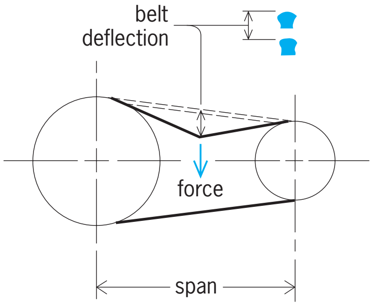


Fig. 6. Tension measurement by deflection.

force required to deflect the belt a specific distance per inch of span length (**Fig. 6**). Timing belts require minimum tension, just enough to keep the belt in contact with the pulleys.

Belt wear. Belts probably fail by fatigue more often than by abrasion. The fatigue is caused by the cyclic stress applied to the belt as it bends around the pulleys. Belt failure is accelerated when the following conditions are present: high belt tension; excessive slippage; adverse environmental conditions; and momentary overloads caused by shock, vibration, or belt slapping. *See* CHAIN DRIVE; PULLEY; WEAR.

Arthur Erdman; Raymond Giese

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; W. S. Miller (ed.), Drive systems, *Mach. Des.*, 50 (15):26–33, June 29, 1978; C. R. Mischke and J. E. Shigley, *Mechanical Engineering Design*, 6th ed., 2000; P. B. Schulbert (ed.), *Machinery's Handbook*, 21st ed., 1979.

Bentonite

The term first applied to a particular, highly colloidal plastic clay found near Fort Benton in the Cretaceous beds of Wyoming. This clay swells to several times its original volume when placed in water and forms thixotropic gels when small amounts are added to water. Later investigations showed that this clay was formed by the alteration of volcanic ash in place; thus, the term bentonite was redefined by geologists to limit it to highly colloidal and plastic clay materials composed largely of montmorillonite clay minerals, and produced by the alteration of volcanic ash in place. Many mineralogists and geologists now use the term bentonite without reference to the physical properties of the clay. On the other hand, the term has been used commercially for any plastic, colloidal, and swelling clays without reference to a particular mode of origin. *See* CLAY; GEL; MONTMORILLONITE.

Origin. The occurrence of shard structures of ash as pseudomorphs in the clay often indicates the volcanic ash parent material of bentonites. The presence of nonclay minerals characteristic of igneous material also provides evidence for the origin from ash.

Evidence strongly indicates that the transformation of volcanic glass to montmorillonite takes place either during or soon after accumulation of the ash. The alteration process is essentially a devitrification of the natural glass of the ash and the crystallization of montmorillonite. The ash often contains excess silica which may remain in the bentonite as cristobalite. *See* VOLCANIC GLASS.

Occurrence. Bentonites have been found in almost all countries and in rocks of a wide variety of ages. They appear to be most abundant in rocks of Cretaceous age and younger. In such material in older rocks the montmorillonite is often almost completely collapsed by compaction and metamorphism. This altered bentonite swells very little and does not have the usual high colloidal properties of bentonite; it is also called metabentonite.

Beds of bentonite show a variable thickness and can range from a fraction of an inch up to as much as 50 ft (15 m) thick. The color of bentonite is also variable, but perhaps most often it is yellow or yellowish-green. When observed in the weathered outcrop, bentonite tends to develop a characteristic cellular structure. It shrinks upon drying and often shows a characteristic jigsaw-puzzle set of fractures. Not all beds of ash in the geologic column have altered to bentonite. When present, bentonite is most frequently found in marine beds; thus, it seems certain that the alteration from ash to bentonite is favored in sea water.

In the United States, bentonites are mined extensively in Wyoming, Arizona, and Mississippi. England, Germany, former Yugoslavia, Russia, Algeria, Japan, and Argentina also produce large tonnages of bentonite.

Uses. Bentonites are of great commercial value. They are used in decolorizing oils, in bonding molding sands, in the manufacture of catalysts, in the preparation of oil well drilling muds, and in numerous other relatively minor ways. Since the properties of bentonites vary widely, they are not all suitable for commercial use. The properties of a particular bentonite determine its economic use. For example, the bentonite found in Wyoming is excellent for drilling muds and for foundry use, but it is not useful for the making of catalysts or for oil decolorizing. *See* CLAY, COMMERCIAL.

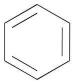
Ralph E. Grim; Floyd M. Wahl

Benzene

A colorless, liquid, inflammable, aromatic hydrocarbon of chemical formula C_6H_6 which boils at 80.1°C (176°F) and freezes at 5.4–5.5°C (41.7–41.9°F). In the older American and British technical literature benzene is designated by the German name benzol. In current usage the term benzol is commonly reserved for the less pure grades of benzene.

Benzene is used as a solvent and particularly in Europe as a constituent of motor fuel. In the United States the largest uses of benzene are for the manufacture of styrene and phenol. Other important outlets are in the production of dodecylbenzene, aniline, maleic anhydride, chlorinated benzenes (used in making DDT and as moth flakes), and benzene hexachloride, an insecticide.

X-ray and electron diffraction studies indicate that the six carbon atoms of benzene, each with a hydrogen atom attached, are arranged symmetrically in a plane, forming a regular hexagon. The hexagon symbol (**1**), commonly used to represent the structural formula for benzene, implies the presence of a carbon atom at each of the six angles and, unless substituents are attached, a hydrogen at each carbon atom. Whereas the three double bonds usually included in formula (**1**) are convenient in accounting for the addition reactions of benzene, present evidence is that all the carbon-to-carbon bonds are identical.



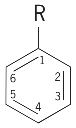
(**1**)

Prior to World War II most of the benzene commercially produced was a by-product of the steel industry. High-temperature pyrolysis of coal to produce metallurgical coke for the steel industry yields gases and tars from which benzene may be recovered.

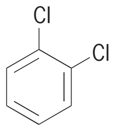
In the present era nearly all commercial benzene is a product of petroleum technology. The gasoline fractions obtained by reforming or steam cracking of feedstocks from petroleum contain benzene and toluene, which can be separated economically. Unfortunately the presence of benzene and toluene are important for the high octane number of these same gasoline fractions, with the result that the prices of benzene and toluene are affected by that of lead-free gasoline. Benzene may also be produced by the dealkylation of toluene. *See* PETROLEUM PROCESSING AND REFINING.

Despite its low hydrogen-to-carbon ratio, benzene reacts chiefly by substitution. Because of the symmetry of the molecule, a given reagent will react to afford only one monosubstituted benzene. Some typical substitution products are included in the **table**. *See* SUBSTITUTION REACTION.

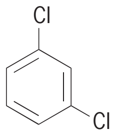
A large number of disubstitution products of benzene are known. For any two substituents there are only three possible arrangements. The groups may be adjacent to each other (ortho, *o*-). They may be located on carbons separated from each other by a single CH, that is, at positions 1 and 3 of formula (**2**) (meta, *m*-), or they may be at any two opposite corners of the hexagon (para, *p*-). The three possible dichlorobenzenes are shown below.



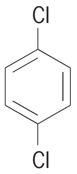
(**2**)



o-Dichloro-benzene



m-Dichloro-benzene



p-Dichloro-benzene

Substitution products of benzene

Reagents	Formula (**2**), R =	Name of product
H_2SO_4	$—SO_3H$	Benzenesulfonic acid
$HNO_3 + H_2SO_4$	$—NO_2$	Nitrobenzene
Cl_2 + Fe	—Cl	Chlorobenzene
$C_2H_4 + AlCl_3$	$—C_2H_5$	Ethylbenzene

The position of substituents in trisubstituted or polysubstituted benzenes is usually indicated by numbers. For instance, the trichlorobenzene in which all three chlorines are adjacent is called 1,2,3-trichlorobenzene.

Benzene may also react by addition. Hydrogenation in the presence of a catalyst yields cyclohexane, C_6H_{12}, an intermediate in the manufacture of fibers, while addition of chlorine (usually in the presence of ultraviolet light) yields benzene hexachloride, $C_6H_6Cl_6$, an insecticide.

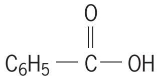
The largest commercial uses for benzene are in the manufacture of ethylbenzene (a styrene intermediate), cumene (an intermediate in the manufacture of phenol), and cyclohexane (an intermediate in the manufacture of fibers).

Benzene is a toxic substance, and prolonged exposure to concentrations in excess of 35–100 parts per million in air may lead to symptoms ranging from nausea and excess fatigue to anemia and leukopenia. In addition, the Occupational Safety and Health Administration has listed benzene as a potential carcinogen. *See* AROMATIC HYDROCARBON. Charles K. Bradsher

Bibliography. T. W. G. Solomons, *Organic Chemistry*, 7th ed., 1999; K. Weissermel and H. J. Arpe, *Industrial Organic Chemistry*, 3d ed., 1999.

Benzoic acid

An organic acid, also known as benzene carboxylic acid, with the formula below. Melting point is 250.2°F (121.2°C), and the acid sublimes at 212°F (100°C). Benzoic acid is only slightly soluble in water but is soluble in most organic solvents, and reacts with bases to form the corresponding benzoate salts. Benzoic acid was first obtained by sublimation from gum benzoin. It occurs both free and combined in nature, being found in many berries (cranberries, prunes, cloves) and as the end product of phenylalanine metabolism.



$$C_6H_5—\overset{\overset{\Large O}{\|}}{C}—OH$$

Preparation. Benzoic acid is prepared in the laboratory by the Grignard reaction, hydrolysis of benzonitrile (C_6H_5CN), or prolonged oxidation of alkyl benzenes with potassium permanganate regardless of the length of the alkyl group. Commercially it was previously prepared by the chlorination of toluene ($C_6H_5CH_3$) with the subsequent hydrolysis of the benzotrichloride ($C_6H_5CCl_3$), and by the monodecarboxylation of phthalic anhydride (from naphthalene). Modern preparation is by the catalytic oxidation of toluene at elevated temperatures with a cobalt catalyst, and purification by sublimation. *See* GRIGNARD REACTION.

Reactions. Benzoic acid is stable to mild oxidizing agents, but alkyl derivatives of benzene are readily oxidized to the acid. If *o*-phenyl fatty acids contain an odd number of carbon atoms in the chain, benzoic acid is formed, but if the chain contains an even number of carbon atoms, then phenylacetic acid is formed, proving degradation occurs two carbon atoms at a time (beta oxidation).

Benzoic acid undergoes the normal reactions of the aromatic ring (nitration, sulfonation, halogenation, alkylation). Groups are inserted in the meta position due to the directive influence of the carboxyl group. Substitution occurs less readily than with ortho- or para-directing groups due to the deactivating effect of the meta-directing group. Ortho or para derivatives can be obtained with some starting materials other than the acid. Benzoic acid also undergoes the usual reactions of the carboxyl group, forming acyl halides, anhydrides, amides, esters, and salts. *See* HALOGENATION; NITRATION; SUBSTITUTION REACTION; SULFONATION AND SULFATION.

Salts. Sodium benzoate is the only salt of importance. It is water-soluble, has antipyretic and antiseptic properties, is useful as a corrosion inhibitor with sodium nitrite if used for iron, and is also used to modify alkyd resins by increasing hardness, adhesion, and gloss. It is nearly tasteless, only slightly toxic, and is widely used as one of the few materials permitted by law (maximum 0.10%) as a food preservative, for which it is used particularly with acid foods, such as pickles, cider, catsup, and sodas. It is most effective at pH 2.5–4.5, with its effectiveness decreasing as the solution becomes more basic. It is removed from the body (detoxified) in the liver, where it reacts to form benzoyl coenzyme A, which then combines with glycine and is excreted as hippuric acid. Large doses (0.2–0.3 oz/day or 6–8 g/day) of sodium benzoate may cause nausea and vomiting; however, some people are not affected by 0.9 oz/day (25 g/day). In this way, it acts similarly to salicylates. *See* PH.

Esters. Esters of benzoic acid are also found in nature. They are almost universally fragrant. Methyl benzoate is the fragrant principle in tuberose. Some esters of benzoic acid are used in the perfume industry, for example, benzyl ester as a fixative. The butyl ester is used as a dye carrier because of its desirable biodegradable properties; and glycol esters are used as plasticizers. *See* ESTER.

Derivatives. Benzoic acid is a weak acid ($K_a = 6.3 \times 10^{-5}$), being slightly stronger than simple aliphatic acids. However, electrophilic groups substituted on the aromatic ring will increase the acidity; for example, *o*-chlorobenzoic acid has a $K_a = 130 \times 10^{-5}$ and *p*-nitro derivative has a $K_a = 620 \times 10^{-5}$. Hydroxybenzoic acid (salicylic acid) is used to prepare aspirin and salol, the latter as an enteric coating for pills which allows the medicine to pass through the acid stomach essentially unchanged but to be released in the alkaline intestine. *p*-Aminobenzoic acid (PABA) is included in the vitamin B complex and makes up a portion of the folic acid molecule. *o*-Aminobenzoic acid (anthranilic acid) is often used in the manufacture of aniline dyes. Benzoic acid reacts with benzotrichloride to produce benzoyl chloride, from which the peroxide can be made, this being a catalyst for free-radical reactions. *See* PARA-AMINOBENZOIC ACID; ASPIRIN; FREE RADICAL.

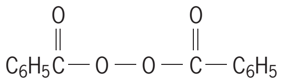
Uses. General class of uses for both benzoic acid and its derivatives include the pharmaceuticals and synthetic polymers. Benzoic acid is used in preservatives and many cosmetics. The derivatives are used in the dyeing industry, with some applications in the cosmetic industry. Pure benzoic acid is a standard for bomb calorimetry because of its ease of purification by sublimation. *See* CALORIMETRY; CARBOXYLIC ACID.

Elbert H. Hadley

Bibliography. R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998; *Kirk-Othmer Encyclopedia of Chemical Technology*, vol. 3, 4th ed., 1992; *Merck Index*, Merck and Co., 11th ed., 1989; J. D. Morrison and R. N. Boyd, *Organic Chemistry*, 5th ed., 1987.

Benzoyl peroxide

A chemical compound, sometimes called dibenzoyl peroxide, with the formula shown. It is a colorless, crystalline solid, melting point 106–108°C (223–226°F), that is virtually insoluble in water but very soluble in most organic solvents. Benzoyl peroxide is an initiator, a type of material that decomposes at a controlled rate at moderate temperatures to give free radicals. *See* AUTOXIDATION; FREE RADICAL.



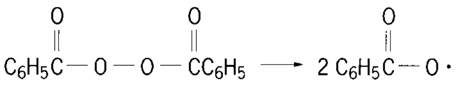
$$C_6H_5\overset{\overset{O}{\|}}{C}-O-O-\overset{\overset{O}{\|}}{C}-C_6H_5$$

Production. Benzoyl peroxide is synthesized in about 70% yield from the reaction of benzoyl chloride with hydrogen peroxide in alkaline aqueous solution or with sodium peroxide in aqueous acetone. It can be purified by precipitation from cold chloroform by addition of methanol. Recrystallization from hot solvents is hazardous; peroxides should be heated only in dilute solution.

Chemical uses. Benzoyl peroxide has a half-life (10 h at 73°C or 163°F in benzene) for decomposition that is very convenient for many laboratory and commercial processes. This, plus its relative stability among peroxides, makes it one of the most frequently used initiators. Its primary use is as an initiator of vinyl polymerization, used in amounts of 0.1 mole % of the vinyl monomer. It also is the preferred bleaching agent for flour; is used to bleach many commercial waxes and oils; and is the active ingredient in commercial acne preparations. *See* BLEACHING; POLYMERIZATION.

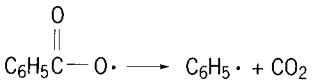
Hazards. Benzoyl peroxide itself is neither a carcinogen nor a mutagen. However, when coapplied with carcinogens to mouse skin it is a potent promoter of tumor development. All peroxides should be treated as potentially explosive, but benzoyl peroxide is one of the least prone to detonate. However, it should be stored in paper or plastic containers and not in rigid, screw-top containers since unscrewing a cap that contains benzoyl peroxide crystals in its threads can result in detonation. *See* MUTAGENS AND CARCINOGENS.

Reactions. Like all peroxides, benzoyl peroxide decomposes by rupture of its oxygen-oxygen peroxide bond, reaction (1), to give two benzoyloxyl



$$C_6H_5\overset{O}{\overset{\|}{C}}-O-O-\overset{O}{\overset{\|}{C}}C_6H_5 \longrightarrow 2\,C_6H_5\overset{O}{\overset{\|}{C}}-O\cdot \qquad (1)$$

radicals. (The odd electron in the free radicals produced in this process is usually represented by a dot.) The benzoyloxyl radical rapidly decarboxylates (loses CO_2) to give the phenyl radical (C_6H_5), reaction (2). Thus, benzoyl peroxide provides a source



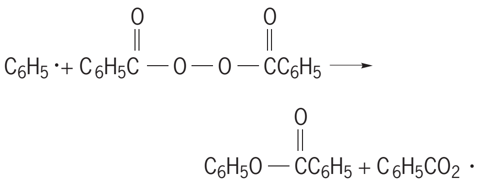
$$C_6H_5\overset{O}{\overset{\|}{C}}-O\cdot \longrightarrow C_6H_5\cdot + CO_2 \qquad (2)$$

of phenyl and benzoyloxyl radicals that can perform chemical reactions. For example, the phenyl radicals can add to the double bond of vinyl chloride (C_2H_2Cl) to form a new radical and thus initiate the polymerization of the vinyl monomer, reaction (3).

$$C_6H_5\cdot + CH_2=CHCl \longrightarrow C_6H_5-CH_2-\dot{C}HCl \qquad (3)$$

Accelerated and induced decomposition. Many materials cause benzoyl peroxide to decompose at an accelerated rate over that observed in pure benzene. For example, solvents with high reactivity toward free radicals (such as diethyl ether) cause a greatly accelerated decomposition. For this reason, concentrated solutions of benzoyl peroxide should be prepared only by experienced workers. Compounds that can act as nucleophiles, reducing agents, or electron donors also can greatly accelerate the decomposition; examples are amines and sulfides. Transition-metal ions such as iron or copper also cause an increased rate of decomposition; this can be utilized in commercial practice to achieve a lower temperature or more rapid decomposition.

All free radicals, including radicals from the peroxide itself, attack the peroxide at the oxygen-oxygen bond and cause it to decompose in a radical-induced decomposition. For example, phenyl radicals cause the induced decomposition shown in reaction (4).



$$C_6H_5\cdot + C_6H_5\overset{O}{\overset{\|}{C}}-O-O-\overset{O}{\overset{\|}{C}}C_6H_5 \longrightarrow C_6H_5O-\overset{O}{\overset{\|}{C}}C_6H_5 + C_6H_5CO_2\cdot \qquad (4)$$

See ORGANIC REACTION MECHANISM. William A. Pryor

Bibliography. W. Ando, *Organic Peroxides*, 1992; D. C. Nonhebel and J. C. Walton, *Free Radical Chemistry*, 1974; S. Patai, *The Chemistry of Peroxides*, 1984; W. A. Pryor, *Free Radicals*, 1966.

Bering Sea

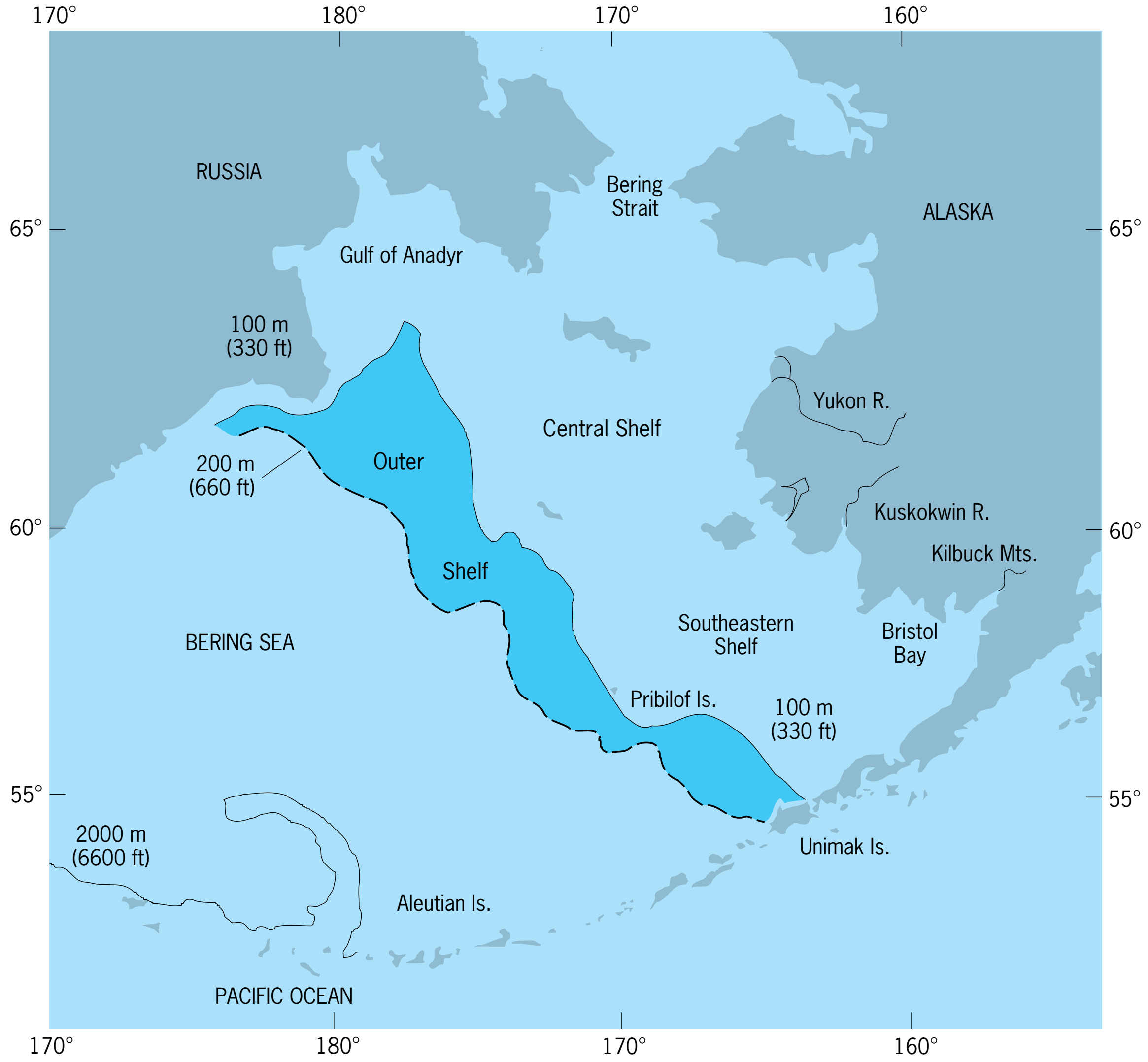
A water body north of the Pacific Ocean, 875,900 mi^2 (2,268,000 km^2) in area, bounded by the Siberian and Alaskan landmasses to the west and east respectively, and the Aleutian-Komandorskiye island arc to the south. The Bering Sea is very productive biologically, and the ecosystem supports large populations of fish, shellfish, and marine birds and mammals.

The Bering Sea consists of a large, deep basin in the southwest portion, with depths as great as 10,000 ft (3000 m). To the north and east, an extremely wide, shallow continental shelf extends north to the Bering Strait. These two major regions are separated by a shelf break, the position of which coincides with the southernmost extent of sea ice in a cold season (see **illus.**). Ice is a prominent feature over the Bering Sea shelf during the winter months. Coastal ice begins to form in late October, and by February coastal ice may be found as far south as the Aleutians, such that the ice edge in the eastern Bering Sea advances and retreats seasonally over a distance as great as 600 mi (1000 km). Ice-free conditions can be expected throughout the entire region by early July. Extraordinary interannual variability in ice cover, air, sea-surface temperature, and winds have been observed in the region over the past decades.

The main water connections with the Pacific Ocean are in the west part of the Aleutian Islands: the 6000-ft-deep (2000-m) pass between Attu and the Komandorskiye Islands and the 14,000-ft-deep (4400-m) pass between the Komandorskiyes and Kamchatka. The Bering Sea connection with the Arctic Ocean (Chukchi Sea) is the Bering Strait, 53 mi (85 km) wide and 140 ft (45 m) deep. This is the only avenue of exchange for ice, water, heat, and nutrients between the Pacific and Arctic oceans. *See* ARCTIC AND SUBARCTIC ISLANDS; ARCTIC OCEAN; PACIFIC OCEAN.

Temperature and salinity. Over the continental shelf, the waters receive an excess of precipitation over evaporation, and about 1% of their volume is provided annually by river runoff. The hydrographic structure over the shelf is determined principally by the mechanical stirring of tidal action and wind mixing, and there is a negligible mean flow of water. Large seasonal variations in temperature occur in this area. In winter, the water temperatures are extremely cold, and under the ice they approach the freezing point (29°F or −1.7°C). Salinities range between 31 and 33 parts per thousand (‰). In summer, surface temperature rise to 43–50°F (6–10°C), and occasionally higher near the coast.

The shelf waters are separated from the adjacent oceanic waters by semipermanent fronts. These fronts are detectable as a relatively strong horizontal gradient in upper-layer salinity. The location of these


170°
180°
170°
160°
65°
60°
55°
RUSSIA
Bering Strait
ALASKA
Gulf of Anadyr
100 m (330 ft)
Yukon R.
Central Shelf
Outer
200 m (660 ft)
Kuskokwin R.
Kilbuck Mts.
Shelf
Southeastern Shelf
Bristol Bay
BERING SEA
Pribilof Is.
100 m (330 ft)
2000 m (6600 ft)
Unimak Is.
Aleutian Is.
PACIFIC OCEAN

Map of the Bering Sea showing the position of the shelf and shelf break.

fronts appears to be tied to the continental-shelf break. They are centered over about 480 ft (150 m) depth, with a typical width of about 31 mi (50 km). These are areas of extremely high phytoplankton production and are critical to the southeast Bering Sea ecosystem. The oceanic water abutting the front on the seaward side is relatively invariant in temperature and salinity compared with the waters over the shelf, although the surface waters show some seasonal variation. A temperature maximum is found between 600 and 1300 ft (200 and 400 m), below which the temperature decreases to bottom values of 34.7–34.8°F (1.5–1.6°C). Between 300 and 600 ft (100 and 200 m) there is a halocline where salinity increases from the relatively low upper-layer values to over 34‰. The deep water exhibits a gradual increase in salinity to bottom values of 34.6–34.7‰. At the top of the halocline, about 30 ft (10 m), is a temperature minimum in summer, possibly in part due to water cooled locally during the previous winter. In winter, water of this density is also produced in large quantities on the shelf and its embayments, notably the Gulf of Anadyr, and is advected over the basin beneath the surface layer. *See* CONTINENTAL MARGIN; SEAWATER.

Tides and currents. Tides in the Bering Sea are semidiurnal, with a strong diurnal inequality typical of North Pacific tides. The mean range varies from about 4.8 ft (1.5 m) in the Aleutians to 12 in. (30 cm) in the Bering Strait, but are much larger in Bristol Bay (greater than 13 ft or 4 m). Three water masses are associated with the Bering Sea—Western Subarctic, Bering Sea, and the Alaska Stream. The general circulation of the Bering Sea is counterclockwise, with many small eddies superimposed on the large-scale pattern. The currents in the Bering Sea are generally an inch (a few centimeters) per second except along the continental slope, along the coast of Kamchatka, and in certain eddies, where somewhat higher values have been found. *See* OCEAN CIRCULATION; TIDE.

The direction of the overall flow over the Bering Sea shelf is to the north, ultimately through the Bering Strait into the Chukchi Sea. This flow is driven by a mean sea-level slope that results from the difference in sea level between the Atlantic and Pacific oceans. St. Lawrence Island forms two straits: the

Anadyr Strait to the west and Shpanberg Strait to the east. The most rapid flow over the Bering Sea shelf takes place through the Anadyr Strait, where the flow rate can reach 6 in. (15 cm) per second. This water originates as a northward offshoot of the Bering Slope Current, which flows along the Bering Sea outer shelf from east to west, and bifurcates when it reaches the Siberian coast. This northward-flowing Anadyr water is very rich in nutrients and is responsible for the high productivity of the northern Bering Sea, whereas the water farther to the east over the shelf is relatively unproductive. Three water masses have been identified in the north Bering Sea: Alaska Coastal Water, Bering Shelf Water, and Anadyr Water.

Biological production. Primary production in the Bering Sea is highly variable seasonally and spatially. High primary production in the form of a spring bloom is found at the retreating ice edge and may be more sustained along the frontal areas along the shelf. The highest sustained primary production rates occur in the Anadyr Water in the Gulf of Anadyr, the northern Bering Sea, and the Bering Strait. A few sites on the shelf have been identified as particularly effective in translating the high biological productivity into higher trophic levels. For example, the Chirikov Basin to the north of St. Lawrence Island is a major deposition site for biological material, as the Anadyr Current flows through the region. As a result, the basin is a major summer feeding ground for gray whales, which feed on the abundant tube-dwelling amphipods that live in the bottom. In the southeast portion of the shelf, the sediment-dwelling fauna (infauna) is dominated by polychaete worms, small clams, and brittle stars. The sediment-surface-dwelling animals are dominated by echinoderms in the northern Bering Sea and by commercially important crabs in the southeast. Total annual primary productivity varies between 0.48 and 1.3 oz carbon/ft^2 (150 and 400 g C/m^2), depending on the specific location. This input, along with input of organic matter from the river systems and from the coastal ecosystems (especially sea-grass beds in the Aleutians), supports a large population of organisms at higher trophic levels. Benthic (bottom-living) organisms vary in abundance, depending on the area.

A large proportion of southeastern Bering Sea bottom-living organisms are of direct use as food for humans. King crabs, snow crabs, and snails of the genus *Neptunea* are among the most abundant invertebrates, and the area may also support clam resources of commercial magnitude. Shrimp, especially pink shrimp, are also of potential commercial use. The most abundant fish species are walleye pollock and yellowfin sole. Pacific cod, sablefish, Pacific salmon, perch, herring, capelin, sculpin, plaice, halibut, other sole species, and flounder also occur.

Marine mammals, numbering about a million on the ice pack alone, include seals, whales, walruses, sea otters, and polar bears. A variety of marine birds are also associated with sea ice in the Bering Sea, including three species of gulls, old squaws, king eiders, kittiwakes, murres, and guillemots. Shearwaters and auklets are also abundant in the Bering Sea. *See* BIOLOGICAL PRODUCTIVITY; MARINE ECOLOGY.

Fisheries. The major groundfish fishery of the Bering Sea is pollock (*Theragra chalcogramma*). With a yearly harvest approaching 2.2 million tons (2 million metric tons), pollock constitutes 60% of the fishery. There is great concern about the impacts of this pollock fishery on other species in the Bering Sea ecosystem, especially the Steller sea lion (*Eumetopias jubata*). International issues have come into play in connection with the exploitation of the stocks. Fishing activity in the so-called Donut Hole, a region of international waters surrounded by United States and Russian waters, has been of particular concern, and a treaty was negotiated to conserve and manage pollock in the central Bering Sea. The pollock resource is capable of very rapid increases in abundance as well as decreases, related to variations in year-class abundance. Salmon also represents a major segment of the southeast Bering Sea fishery, with major runs in the Bristol Bay area. The herring fishery is small and targets roe for export to Japan.

Most of the fishing over the shelf is confined to the southern portion. The commercially important crabs dominate the southeast portion of the surface-sediment-dwelling fauna (epifauna). The decline of the king crab population and its associated fishery between 1980 and the mid-1990s was dramatic. The red king crab (*Paralithoides camschatica*) is most abundant in the middle shelf area, and in the Bering Sea one species of Tanner crab, *Chionoecetes bairdii*, is found only in the southeastern portion. The crab population of the northern Bering Sea is dominated by a second species of Tanner crab, *C. opilio*. *See* MARINE FISHERIES.

Environmental impact. Major changes in the Bering Sea environment over recent decades have included a long-term decline in the Steller sea lion population; their listing as an endangered species has resulted in regulatory impacts on the pollock fishery. In the late 1960s, a climatic regime shift produced increased temperatures and reduced sea ice. The most dramatic biological changes took place in 1997 and 1998, with a major die-off of marine birds, drastically reduced salmon runs in western Alaska, declines in sea otters, and a dense bloom of coccolithophorids which persisted over the Bering Sea shelf throughout the entire summer season (coccolithophorids had not previously been reported in large numbers in the Bering Sea). While the causes and interrelationships for these biological events remain obscure, physical factors and a reduced nutrient availability are implicated. *See* ENDANGERED SPECIES; FISHERIES ECOLOGY.

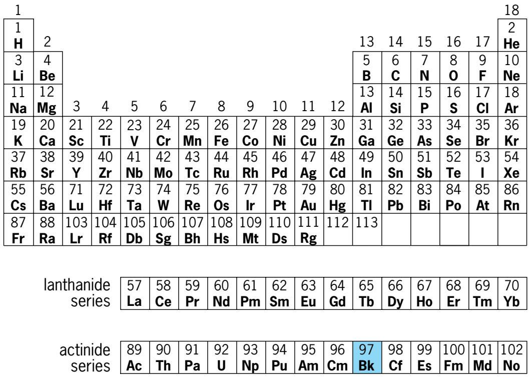
Vera Alexander

Bibliography. L. K. Coachman, Circulation, water masses, and fluxes on the southeastern Bering Sea shelf, *Continental Shelf Res.*, 5:23–108, 1986; L. K. Coachman and D. A. Hansell (eds.), ISHTAR: Inner Shelf Transfer and Recycling in the Bering and Chukchi Sea, *Continental Shelf Res.* (spec. iss.), 13:473–704, 1993; National Research Council, Committee on the Bering Sea Ecosystem, *The Bering Sea Ecosystem*, National Academy of Sciences

Press, Washington, DC, 1996; H. J. Niebauer and V. Alexander, Oceanographic frontal structure and biological production at an ice edge, *Continental Shelf Res.*, 4:367–388, 1985; North Pacific Marine Science Organization, *Dynamics of the Bering Sea*, ed. by T. Loughlin and K. Ohtani, University of Alaska Sea Grant, AK-SG-99-03, Fairbanks, 1999; J. D. Schumacher and P. J. Stabeno, The continental shelf of the Bering Sea, in K. H. Brink and A. R. Robinson (eds.), *The Sea*, vol. 11, Wiley, New York, 1998; P. J. Stabeno et al., Under-ice observations of water column temperature, salinity and spring phytoplankton dynamics: Eastern Bering Sea shelf, *J. Mar. Res.*, 56:239–255, 1998; J. J. Walsh et al., The role of the Bering Strait in the carbon/nitrogen fluxes of polar marine ecosystems, *6th Conference of the Comité Arctique International*, Fairbanks, Alaska, 1989; J. J. Walsh and C. P. McRoy, Ecosystem analysis in the southeastern Bering Sea, *Continental Shelf Res.*, 5:259–288, 1986.

Berkelium

Element number 97, symbol Bk, the eighth member of the actinide series of elements. In this series the 5*f* electron shell is being filled, just as the 4*f* shell is being filled in the lanthanide (rare-earth) elements. These two series of elements are very similar in their chemical properties, and berkelium, aside from small differences in ionic radius, is especially similar to its homolog terbium. *See* PERIODIC TABLE; RARE-EARTH ELEMENTS; TERBIUM.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 **H**																	2 **He**
3 **Li**	4 **Be**											5 **B**	6 **C**	7 **N**	8 **O**	9 **F**	10 **Ne**
11 **Na**	12 **Mg**											13 **Al**	14 **Si**	15 **P**	16 **S**	17 **Cl**	18 **Ar**
19 **K**	20 **Ca**	21 **Sc**	22 **Ti**	23 **V**	24 **Cr**	25 **Mn**	26 **Fe**	27 **Co**	28 **Ni**	29 **Cu**	30 **Zn**	31 **Ga**	32 **Ge**	33 **As**	34 **Se**	35 **Br**	36 **Kr**
37 **Rb**	38 **Sr**	39 **Y**	40 **Zr**	41 **Nb**	42 **Mo**	43 **Tc**	44 **Ru**	45 **Rh**	46 **Pd**	47 **Ag**	48 **Cd**	49 **In**	50 **Sn**	51 **Sb**	52 **Te**	53 **I**	54 **Xe**
55 **Cs**	56 **Ba**	71 **Lu**	72 **Hf**	73 **Ta**	74 **W**	75 **Re**	76 **Os**	77 **Ir**	78 **Pt**	79 **Au**	80 **Hg**	81 **Tl**	82 **Pb**	83 **Bi**	84 **Po**	85 **At**	86 **Rn**
87 **Fr**	88 **Ra**	103 **Lr**	104 **Rf**	105 **Db**	106 **Sg**	107 **Bh**	108 **Hs**	109 **Mt**	110 **Ds**	111 **Rg**	112	113					

lanthanide series	57 **La**	58 **Ce**	59 **Pr**	60 **Nd**	61 **Pm**	62 **Sm**	63 **Eu**	64 **Gd**	65 **Tb**	66 **Dy**	67 **Ho**	68 **Er**	69 **Tm**	70 **Yb**
actinide series	89 **Ac**	90 **Th**	91 **Pa**	92 **U**	93 **Np**	94 **Pu**	95 **Am**	96 **Cm**	97 **Bk**	98 **Cf**	99 **Es**	100 **Fm**	101 **Md**	102 **No**

Berkelium does not occur in the Earth's crust because it has no stable isotopes. It must be prepared by means of nuclear reactions using more abundant target elements. These reactions usually involve bombardments with charged particles, irradiations with neutrons from high-flux reactors, or production in a thermonuclear device.

Berkelium metal is chemically reactive, exists in two crystal modifications, and melts at 986°C (1806°F). Berkelium was discovered in 1949 by S. G. Thompson, A. Ghiorso, and G. T. Seaborg at the University of California in Berkeley and was named in honor of that city. Nine isotopes of berkelium are known, ranging in mass from 243 to 251 and in half-life from 1 hour to 1380 years. The most easily produced isotope is ^{249}Bk, which undergoes beta decay with a half-life of 314 days and is therefore a valuable source for the preparation of the isotope ^{249}Cf. The berkelium isotope with the longest half-life is ^{247}Bk (1380 years), but it is difficult to produce in sufficient amounts to be applied to berkelium chemistry studies. *See* ACTINIDE ELEMENTS; TRANSURANIUM ELEMENTS.

Glenn T. Seaborg

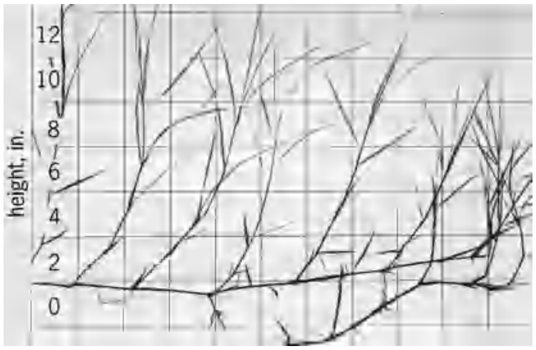
Bibliography. S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; G. T. Seaborg (ed.), *Transuranium Elements: Products of Modern Alchemy*, 1978; G. T. Seaborg and W. D. Loveland, *The Elements Beyond Uranium*, 1990.

Bermuda grass

A long-lived perennial (*Cynodon* spp.) that originated in Africa. It is believed that in the early 1500s Spanish explorers unintentionally brought the first common Bermuda grass seeds to the Western Hemisphere with the hay that fed their horses. The weedy common type, *C. dactylon* (the type species), can now be found throughout the tropical and subtropical regions of the world. *See* CYPERALES.

Cynodon dactylon is an extensively creeping grass with both stolons and rhizomes. It has short leaves borne on 8–12-in. (20–30-cm) upright culms with several slender seed-bearing spikes digitate at their summit (see **illus.**). It is propagated by planting seed, stolons, rhizomes, or green stems cut at an advanced hay stage. It is well adapted to a wide range of soil types, tolerates a pH range from 4.0 to over 8.0, prefers well-drained soils, is very drought-tolerant, and responds well to fertilization. It grows best during high summer temperatures, grows little at temperatures below 65°F (18°C), and goes dormant when temperatures drop below 30°F (−1°C). Although a weed in cultivated fields, it controls erosion, makes excellent turf, and supplies good grazing for all classes of livestock. *See* LAWN AND TURF GRASSES.

Bermuda grass is a highly variable, heterozygous, cross-pollinated species. Breeding work, begun in



Coastal Bermuda grass, showing its surface runners, characteristic seedheads, and short leaves. (*From G. W. Burton, Coastal Bermuda grass, Ga. Agr. Exp. Sta. Bull., N.S. 2, 1954*)

1936, has produced a number of superior F_1 hybrids that are sterile and must be propagated vegetatively. Vegetative propagation is facilitated by the rapid spread of the aboveground stolons of the new hybrids. Plants set on 6 × 6 ft (2 × 2 m) centers with adequate fertilization, water, and weed control cover a pasture in two or three summer months and produce up to a ton of hay per acre (2.5 metric tons per hectare).

In developing countries, the improved Bermuda grass hybrids can be established by pushing the stolons into moist soil with a pointed stick. In developed countries, tractor-drawn machines are usually used to dig and plant the hybrids.

The first hybrid to be released, in 1943, was the best of 5000 hybrids between a very good common Bermuda grass and one introduced from South Africa. It was named Coastal for the Georgia Coastal Plain Station where it was bred, and it has been planted on more than 10 million acres (4 million hectares) across the South. Compared with common Bermuda grass, Coastal yields twice as much, is more disease-resistant and drought-tolerant, uses fertilizer and water more efficiently, and grows tall enough to be cut for hay. Coastal Bermuda has retained its superiority for 40 years and has been an excellent parent to produce improved varieties. Coastal crossed with a highly digestible, cold-susceptible *C. nlemfuensis* plant from Kenya produced the sterile F_1 hybrid Coastcross-1, which is 12% more digestible than Coastal, and when grazed by cattle gives 30–40% better daily gains and gains per acre. Coastcross-1 is less cold-tolerant than Coastal. Tifton 44 is the best of several thousand F_1 hybrids between Coastal and a winter hardy common Bermuda grass from Berlin, Germany. It is more cold-tolerant, more digestible, and when grazed gives up to 19% better daily gains than Coastal. Because it is male-sterile and equal or superior to Coastal in other important characteristics, Tifton 44 has replaced Coastal as the female for new forage hybrids being developed at Tifton, Georgia.

The top turf Bermuda grasses, Tifgreen, Tifway, and Tifdwarf, are the best of a number of F_1 hybrids between selected plants of *C. dactylon* and *C. transvaalensis*. The parents have 36 and 18 chromosomes as their respective 2*n* number and the Tif hybrids have 27. This chromosome number indicates sterility, making vegetative propagation necessary. Tifdwarf, tolerant of daily mowing at $^3/_{16}$ in. (0.5 cm), is unsurpassed for use as top-quality turf for golf greens and bowling greens. Tifgreen, also bred for golf greens, must be mowed at a higher cut. Tifway, more resistant to pests, weeds, and frost, is the best Bermuda grass for golf fairways and tees, lawns, and athletic fields. Tifway II and Tifgreen II are improved mutants of Tifway and Tifgreen created by exposing dormant rhizome sections to gamma radiation. Most of the mutants so created were inferior to their parents. The sterile turf hybrids are planted by machines as sprigs or sod. When sod of these hybrids is stripped from nurseries, enough rhizomes (underground stems) are left in the soil to reestablish the sod.

People suffering from asthma and hay fever find Bermuda grass pollen to be one of the worst offenders. The sterile turf Bermuda grass hybrids solve this problem by producing no pollen. *See* ASTHMA; GRASS CROPS; POLLEN.

Glenn W. Burton

Bernoulli's theorem

An idealized algebraic relation between pressure, velocity, and elevation for flow of an inviscid fluid. Its most commonly used form is for steady flow of an incompressible fluid, and is given by the equation below, where p is pressure, ρ is fluid density

$$\frac{p}{\rho} + \frac{V^2}{2} + gz = \text{constant}$$

(assumed constant), V is flow velocity, g is the acceleration of gravity, and z is the elevation of the fluid particle. The relation applies along any particular streamline of the flow. The constant may vary across streamlines unless it can be further shown that the fluid has zero local angular velocity. *See* KELVIN'S CIRCULATION THEOREM.

The above equation may be extended to steady compressible flow (where changes in ρ are important) by adding the internal energy per unit mass, e, to the left-hand side. *See* COMPRESSIBLE FLOW.

The equation is limited to inviscid flows with no heat transfer, shaft work, or shear work. Although no real fluid truly meets these conditions, the relation is quite accurate in free-flow or "core" regions away from solid boundaries or wavy interfaces, especially for gases and light liquids. Thus Bernoulli's theorem is commonly used to analyze flow outside the boundary layer, flow in supersonic nozzles, flow over airfoils, and many other practical problems. *See* AERODYNAMICS; BOUNDARY-LAYER FLOW.

According to the above equation, if velocity rises at a given elevation, pressure must fall. This principle is used in Bernoulli-type flow meters, such as orifice plates, venturi throats, and pitot tubes. A flow obstruction is deliberately created to cause a local pressure drop which can be calibrated with the flow rate. *See* FLOW MEASUREMENT.

By invoking additional restrictive assumptions, the Bernoulli theorem can be extended to unsteady flow with zero or constant angular velocity, flow in rotating machinery, and piping systems with frictional losses. *See* PIPE FLOW; TURBINE.

It is named after Daniel Bernoulli, who hinted at a "proportionality" between pressure and velocity in a 1738 hydrodynamics textbook. But the above equation was actually derived by Leonhard Euler in a series of papers in 1755. *See* FLUID FLOW.

Frank M. White

Bibliography. V. L. Streeter et al., *Fluid Mechanics*, 9th ed., 1998; F. M. White, *Fluid Mechanics*, 4th ed., 1998; C. S. Yih, *Fluid Mechanics*, 2d ed., rev. reprint, 1988.

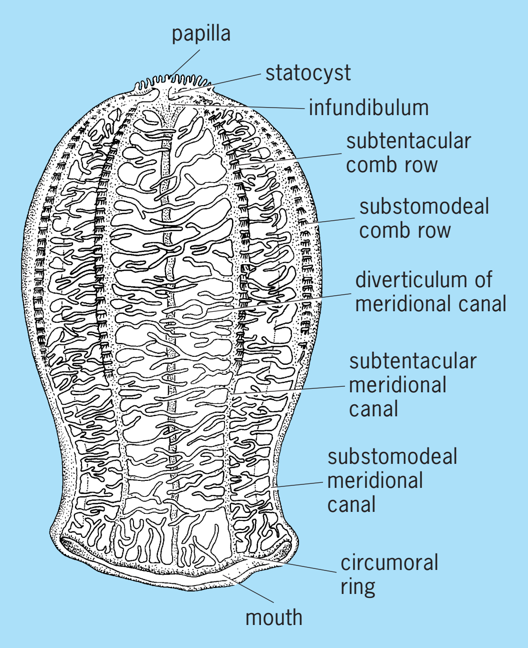
Beroida

An order of the phylum Ctenophora (comb jellies) comprising two genera, *Beroë* and *Neis*. The single species of *Neis (N. cordigera)* is known only from Australia, but many species of *Beroë* have been described from polar to tropical waters all over the world. Beroids are predators on other ctenophores and occur in all parts of the water column where ctenophores are found.

Beroids are the only ctenophores that lack tentacles throughout their life, and their phylogenetic relationship to other ctenophores is uncertain. The conical or cylindrical body is flattened in the tentacular plane (see **illus.**) and ranges in size from a few millimeters to about 16 in. (40 cm). Most of the body is occupied by the stomodeum. *Neis* differs from *Beroë* in having prominent aboral keels. The body is mostly transparent or translucent but may have bluish, orange, pink, or red coloration of the meridional canals or stomodeum. Pigmentation can vary with age, diet, or environment and is probably not a stable species characteristic. Like other ctenophores, beroids are strongly bioluminescent, producing blue-green light along the meridional canals when disturbed.

Locomotion is produced by the well-developed comb (ctene) rows, which extend along the entire body length in some species. Beroids generally swim mouth first and can turn or reverse quickly. The meridional canals have side branches, which in some species anastomose with each other or with branches of the paragastric canals to form a complex gastrovascular network in the mesoglea. On either side of the aboral statocyst are papillae, which may have a sensory function but do not occur in other ctenophores. Beroids are simultaneous hermaphrodites and have ovaries and testes along the meridional canals. Eggs and sperm are shed through numerous gonopores directly to the outside. A fertilized egg develops directly into a miniature adult, with no trace of the cydippid larva stage characteristic of other ctenophores. Beroids are capable of rapid exponential increase in their local populations.

The mouth is large and extensile, capable of engulfing prey larger than the beroid itself. The lips of some *Beroë* are kept closed by special zipperlike strips of tissue while the animal swims. Upon contact with prey, the mouth unzips instantly to swallow it. The inner lips are lined with macrocilia, bundles of ciliary shafts (axonemes) up to 60 micrometers long surrounded by a common membrane. The tips of the macrocilia are formed into pointed teeth, which vary in shape among species. Prey are gripped by these teeth and drawn into the stomodeum by the metachronal beating of the macrocilia. Generally, whole prey are engulfed on contact, but pieces of large prey can be bitten off with the macrocilia. There is morphological and behavioral evidence that chemoreception is involved in detecting and recognizing prey organisms.


papilla
statocyst
infundibulum
subtentacular comb row
substomodeal comb row
diverticulum of meridional canal
subtentacular meridional canal
substomodeal meridional canal
circumoral ring
mouth

Beroid, *Beroë cucumis*.

Beroids feed mainly on other ctenophores, and in some situations may control the population of species such as *Pleurobrachia*. Some beroids are highly specific, consuming only a single prey species, but other are more general, sometimes eating siphonophores and salps as well as ctenophores. Predators on beroids include medusae, heteropod mollusks, hyperiid amphipods, and various fishes. *See* CTENOPHORA.

L. P. Madin

Bibliography. G. R. Harbison, L. P. Madin, and N. R. Swanberg, On the natural history and distribution of oceanic ctenophores, *Deep-sea Res.*, 25:233–256, 1978; F. W. Harrison and J. A. Westfall (eds.), *Microscopic Anatomy of Invertebrates*, vol. 2, 1991; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Beryciformes

An order of somewhat intermediate position among actinopterygian fishes. It shares with the perciforms fin spines and ctenoid scales, an upper jaw bordered by the premaxillae, a ductless swim bladder, and absence of a mesocoracoid. An orbitosphenoid bone is present, as is typical of many lower teleosts. The pelvic fins are thoracic or subabdominal in position, and each has a spine and 3 to 13 (usually more than 5) soft rays; the pelvic girdle is most often attached to the cleithra; and there are 18 or 19 principal caudal rays. *See* PERCIFORMES; SWIM BLADDER.

Classification. Beryciforms are classified in 7 families as shown below, 28 genera, and 123 species.

Anoplogastridae (fangtooths). Body short and deep; numerous long fanglike teeth; small eyes; small or minute scales; lateral line an open groove; fins

Yellowfin soldierfish (*Myripistis chryseres*). (*Photo © John E. Randall*)

without spines; maximum length 16 cm (6.3 in.). Bathypelagic in the Atlantic, Pacific, and Indian oceans to depths of 5000 m (16,500 ft).

Diretmidae (spinyfins). No lateral line; dorsal and anal spines absent, pelvic fin with laminar spine; abdomen with sharp edge formed by ventral scutes; maximum length 37 cm (15.6 in.). Found in the Atlantic, Pacific, and Indian oceans.

Anomalopidae (flashlight fishes or lanternfishes). Light organ located beneath the eye with a shutter mechanism for controlling the emission of light, which is produced by symbiotic bacteria. Maximum length 27 cm (10.6 in.). Primarily found in tropical waters of the Indo-Pacific, from the surface to 300 m (1000 ft) depth.

Monocentridae (pinecone fishes). Body covered with large, heavy, platelike scales; organs on the lower jaw to emit light produced by luminous bacteria; maximum size 21 cm (8.3 in.). Found in tropical and subtropical inhabitats in Indo-Pacific ocean at depths of 30 to 300 m (100 to 1000 ft).

Trachichthyidae (roughies or slimeheads). Angle of the preopercle and posttemporal bone, each possessing distinct spine; median ridge of scutes on abdomen; maximum length 55 cm (22 in.). Found in the Atlantic, Indian, and Pacific oceans at depths of 100 to 1500 m (330 to 5000 ft).

Berycidae (alfonsinos). Compressed body; spiny scales; very large eyes; a large and oblique mouth; a single, relatively short dorsal fin, without a notch; maximum length has not been determined, but western Atlantic species probably do not exceed 60 cm (24 in.). Found in the Atlantic, Indian, western and central Pacific oceans at depths to 200 to 600 m (660 to 2000 ft).

Holocentridae (squirrelfishes and soldierfishes). Largest family of beryciform fishes (65 species), more perchlike than other families of the order (see **illustration**). Body compressed; opercle with spines, large eyes; the ctenoid scales large and very rough; long dorsal fin long, divided by a notch, with 10 to 13 sharp spines and 11 to 17 rays, reddish in color; a maximum body length of 61 cm (24 in.). Found in tropical to temperate waters of the Atlantic, Indian, and Pacific oceans, mostly from shore to a depth of 100 m (330 ft), rarely over 200 m (660 ft).

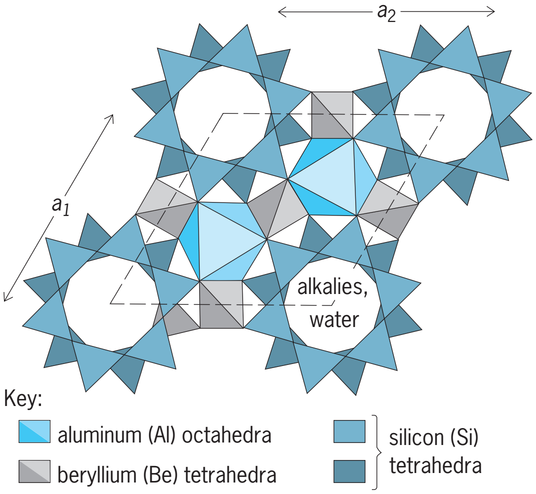
Squirrelfishes feed on benthic invertebrates and small fishes, soldierfishes on large zooplankton; both are nocturnal. These are important food fishes in many parts of the world. Fifteen species are found in Atlantic and Pacific waters of Mexico and North America. *See* ACTINOPTERYGII; OSTEICHTHYES; TELEOSTEI.

Reeve M. Bailey; Herbert Boschung

Bibliography. G. D. Johnson and R. H. Rosenblatt, Mechanism of light organ occlusion in flashlight fishes, family Anomalopidae (Teleostei: Beryciformes), and the evolution of the group, *Zool. J. Linn. Soc.*, 94:65–96, 1988; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994; L. P. Woods and P. M. Sonoda, Order Berycomorphi (Beryciformes): Fishes of the western North Atlantic, *Sears Found. Mar. Res. Mem.*, no. 1, (6):263–396, 1973.

Beryl

The most common beryllium mineral. Beryl, $Al_2[Be_3Si_6O_{18}]$, crystallizes in the hexagonal system with space group P6/mcc. The crystal structure consists of six-membered rings of corner-sharing silicon-oxygen (SiO_4) tetrahedra cross-linked by corner-sharing beryllium-oxygen (BeO_4) tetrahedra to make a three-dimensional honeycomb structure; aluminum-oxygen (AlO_6) octahedra lie between the Si_6O_{18} rings (see **illus.**). Cordierite, $(Mg,Fe)_2[Al_4Si_5O_{18}]$, has a similar structure. Open channels parallel to the sixfold axis contain variable amounts of alkalies (up to 5–8 wt %), such as lithium (Li), sodium (Na), potassium (K), rubidium (Rb), and cesium (Cs), and water molecules (up to $\sim$2.5 wt %). Minor amounts of iron (Fe), chromium (Cr), manganese (Mn), and other metals substitute for aluminum, contributing to the various colors of natural beryl. Beryl has a vitreous luster and is



Crystal structure of beryl looking parallel to the sixfold axis. (*Modified from C. Klein, Manual of Mineralogy, 21st ed., John Wiley and Sons, Inc., 1993*)

typically white to bluish- or yellowish-green, but it can also be shades of yellow, blue, and pink. Its hardness is 7.5–8 on Mohs scale; it has an imperfect basal cleavage and a specific gravity of 2.7–2.9 (increasing with alkali content). Weakly colored varieties can be confused with quartz or apatite. *See* CRYSTAL STRUCTURE; HARDNESS SCALES.

Beryl is stable over a wide range of conditions from room conditions to more than 2200°F (1200°C) and more than 8 kilobars (800 megapascals). At high temperatures, beryl melts to silica-rich liquid, phenakite (Be_2SiO_4), and chrysoberyl ($BeAl_2O_4$). Under hydrous conditions at low temperatures (less than about 550°F or 300°C), beryl breaks down to a variety of minerals, including bertrandite, phenakite, kaolinite, and muscovite. Beryl has been synthesized from oxides and carbonates by dry and hydrothermal methods over a large range of temperature and pressure.

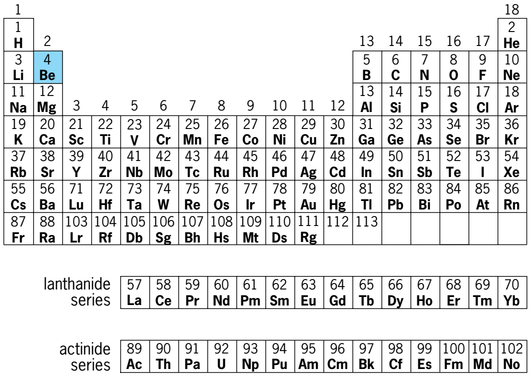
Beryl is a minor accessory mineral in many natural environments, most commonly in granites and associated hydrothermally altered rocks. Granitic pegmatites constitute the major source of beryl (used for beryllium and gemstones); rarely, single crystals weigh many tons. Alkali-rich beryl occurs in complex pegmatites which contain abundant rare-element minerals such as spodumene, lepidolite, and tourmaline. Alkali-poor beryl occurs in mineralogically simple pegmatites, tin and tungsten deposits, and hydrothermal veins. The gem varieties of beryl, aquamarine (blue), emerald (deep green), and morganite (pink to red), are produced from pegmatites (aquamarine, morganite, some emerald), veins (some aquamarine, most emerald), and, rarely, rhyolites (ruby-red morganite). Important pegmatite provinces occur in Brazil, Russia, Madagascar, New England, and California. Beryl is recovered from mineralized granites in Russia and is widespread in granite-related ore deposits from all continents. Emerald occurs in mica schists in Egypt, Austria, the Urals, and North Carolina; the largest, most important emerald deposits are albite-bearing quartz veins in bituminous limestone in Colombia. *See* BERYLLIUM; BERYLLIUM MINERALS; PEGMATITE; SILICATE MINERALS.

Mark D. Barton

Bibliography. J. Sinkankas, *Emerald and Other Beryls*, 1981, reprint 1989.

Beryllium

A chemical element, Be, atomic number 4, with an atomic weight of 9.0122. Beryllium, a rare metal, is one of the lightest structural metals, having a density about one-third that of aluminum. Some of the important physical and chemical properties of beryllium are given in the **table**. Beryllium has a number of unusual and even unique properties. *See* PERIODIC TABLE.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Physical and chemical properties of beryllium

Property	Value
Atomic and mass properties	
Mass number of stable isotopes	9
Atomic number	4
Outer electronic configuration	$1s^22s^2$
Atomic weight	9.0122
Atomic diameter	0.221 nm
Atomic volume	4.96 cm^3/mole
Crystal structure	Hexagonal close-packed
Lattice parameters	$a = 0.2285$ nm $c = 0.3583$ nm
Axial ratio	$c/a = 1.568$
Field of cation (charge/radius^2)	17
Density*, 25°, x-ray (theoretical)	1.8477 ± 0.0007 g/cm^3
Density, 1000°, x-ray	1.756 g/cm^3
Radius of atom (Be^0)	0.111 nm
Radius of ion, Be^{2+}	0.034 nm
Ionization energy ($Be^0 \rightarrow Be^{2+}$)	27.4 eV
Thermal properties	
Melting point	1285°C (2345°F)
Boiling point†	2970°C (5378°F)
Vapor pressure ($T = K°$)	$\log P\ (\text{atm}) = 6.186 + 1.454 \times 10^{-4}T - (16{,}700/T)$
Heat of fusion	250–275 cal/g
Heat of vaporization	53,490 cal/mole
Specific heat (20–100°)	0.43–0.52 cal/(g)(°C)
Thermal conductivity (20°)	0.355 cal/(cm^2)(cm)(s)(°C) (42% of copper)
Heat of oxidation	140.15 cal
Electrical properties	
Electrical conductivity	40–44% of copper
Electrical resistivity	4 microhms/cm (0°C) 6 microhms/cm (100°C)
Electrolytic solution potential, Be/Be†	$E^0 = -1.69$ volts
Electrochemical equivalent	0.04674 mg/coulomb

*Measured values vary from 1.79 to 1.86, depending on purity and method of fabrication.

†Obtained by extrapolation of vapor pressure data, not considered very reliable.

The largest volume uses of beryllium metal are in the manufacture of beryllium-copper alloys and in the development of beryllium-containing moderator and reflector materials for nuclear reactors. Addition of 2% beryllium to copper forms a nonmagnetic alloy which is six times stronger than copper. These beryllium-copper alloys find numerous applications in industry as nonsparking tools, as critical moving parts in aircraft engines, and in the key components of precision instruments, mechanical computers, electrical relays, and camera shutters. Beryllium-copper hammers, wrenches, and other tools are employed in petroleum refineries and other plants in which a spark from steel against steel might lead to an explosion or fire. *See* ALKALINE-EARTH METALS; BERYLLIUM ALLOYS.

Beryllium has found many special uses in nuclear energy because it is one of the most efficient materials for slowing down the speed of neutrons and acting as a neutron reflector. Consequently, much beryllium is used in the construction of nuclear reactors as a moderator and as a support or alloy with the fuel elements. *See* NUCLEAR REACTOR.

The following list shows some of the principal compounds of beryllium. Many of the compounds listed are useful as intermediates in the processes for the preparation of ceramics, beryllium oxide, and beryllium metal. Other compounds are useful in analysis and organic synthesis.

Acetylacetonate
Ammonium beryllium fluoride
Aurintricarboxylate
Basic acetate, $BeO \cdot Be_3(CH_3COO)_6$
Basic beryllium carbonate
Beryllate, $BeO_2{}^{2-}$
Beryllium ammonium phosphate
Bromide, BeB_2
Carbide, Be_2C
Chloride, $BeCl_2$
Dimethyl, $Be(CH_3)_2$
Fluoride, BeF_2
Hydroxide
Nitrate, $Be(NO_3)_2 \cdot 4H_2O$
Nitride, Be_3N_2
Perchlorate, $Be(ClO_4)_2 \cdot 4H_2O$
Oxide, BeO
Plutonium-beryllium, $PuBe_{13}$
Salicylate
Silicates (emerald)
Sulfate, $BeSO_4 \cdot 4H_2O$
Uranium-beryllium, UBe_{13}

Beryllium is surprisingly rare for a light element, constituting about 0.005% of the Earth's crust. It is about thirty-second in order of abundance, occurring in concentrations approximating those of cesium, scandium, and arsenic. Actually, the abundances of beryllium and its neighbors, lithium and boron, are about 10^{-5} times those of the next heavier elements, carbon, nitrogen, and oxygen. At least 50 different beryllium-bearing minerals are known, but in only about 30 is beryllium a regular constituent. The only beryllium-bearing mineral of industrial importance is beryl. Bertrandite substitutes for beryl as a domestic source in the United States. *See* BERYL.

Jack Schubert

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; D. F. Shriver and P. W. Atkins, *Inorganic Chemistry*, 3d ed., 1999.

Beryllium alloys

Dilute alloys of base metals which contain a few percent of beryllium in a precipitation-hardening system. They are the principal useful beryllium alloys manufactured today. Although beryllium has some solid solubility in copper, silver, gold, nickel, cobalt, platinum, palladium, and iron and forms precipitation-hardening alloys with these metals, the copper-beryllium system and, to a considerably lesser degree, the nickel-beryllium alloys are the only ones used commercially.

TABLE 1. Composition of principal beryllium-copper alloys

Alloy grade	Be, %	Other, %
25	1.80–2.05	0.20–0.35 Co
165	1.60–1.79	0.20–0.35 Co
10	0.40–0.70	2.35–2.70 Co
50	0.25–0.50	1.40–1.70 Co 0.90–1.10 Ag
275 C	2.60–2.85	0.35–0.65 Co
245 C	2.30–2.55	0.35–0.65 Co
20 C	2.00–2.25	0.35–0.65 Co
165 C	1.60–1.85	0.20–0.65 Co
10 C	0.55–0.75	2.35–2.70 Co
50 C	0.40–0.65	1.40–1.70 Co 1.00–1.15 Ag

Other than the precipitation-hardening systems, small amounts of beryllium are used in alloys of the dispersion type wherein there is little solid solubility (Al and Mg). Various amounts of beryllium combine with most elements to form intermetallic compounds. Development of beryllium-rich alloys has been chiefly confined to the ductile matrix Be-Al, Be-Cu solid solution alloy with up to 4% Cu, and dispersed-phase type alloys having relatively small amounts of compounds (0.25–6%), chiefly as BeO or intermetallics, for dimensional stability, elevated temperature strength, and elastic limit control. *See* ALLOY.

Beryllium-copper alloys. Commercial alloys of the beryllium-copper group are divided into cast or wrought types, usually in ternary combination such as Cu-Be-Co (**Table 1**). Alloys with higher beryllium content have high strength while those with low beryllium have high electrical and thermal conductivity.

Age-hardenable Cu-Be-Co alloys offer a wide range of properties, because they are extremely ductile and workable in the annealed condition and are strong and hard after precipitation or aging treatment. Cobalt is added to inhibit grain growth and provide uniform heat-treatment response. These alloys also have inherent characteristics of substantial electrical and thermal conductivity and resistance to corrosion and wear, being protected by BeO films which impart this property to all materials containing Be. These age-hardenable alloys resist dimensional change and fatigue.

Properties of beryllium-copper alloys, because of composition and thermal treatment, may vary greatly to provide applications with a wide range of useful characteristics, as shown in **Table 2**.

Applications. Primary applications are found in the electronics, automotive, appliance, instrument, and temperature-control industries for electric current-carrying springs, diaphragms, electrical switch blades, contacts, connectors, terminals, fuse clips, and bellows (foil, strip, and wire), as well as resistance-welding dies, electrodes, clutch rings, brake drums, and switch gear (rod, plate, and forgings). With 1.5% Be or more, the melting point of

TABLE 2. Range of properties of beryllium-copper alloys

Property	Variation possible in commercial alloy
Tensile strength	35,000–215,000 lb/in.^{2} (240–1480 MPa)
Yield strength	10,000–200,000 lb/in.^{2} (70–1400 MPa)
Elongation	0.5–60%
Hardness, Rockwell B and C scale	B_{10}–C_{47}
Thermal conductivity	600–1600 Btu-in./h-ft^2-°F
Electrical conductivity	15–60% IACS
Melting point	1570–2040°F (845–1115°C)

copper is severely depressed and a high degree of fluidity is encountered, allowing casting of intricate shapes with very fine detail. This characteristic is important for plastic injection molds.

For special applications specific alloys have been developed. Free machining and nonmagnetic alloys have been made, as well as high-purity materials. A precipitation-hardening beryllium monel for oceanographic application, containing about 30% Ni, 0.5% Be, 0.5% Si, and the remainder copper, illustrates one of a series of alloys having strength, corrosion resistance to seawater, and antifouling characteristics.

New applications in structural, aerospace, and nuclear fields are submarine repeater cable housings for transoceanic cable systems, wind tunnel throats, liners for magnetohydrodynamic generators for gas ionization, and scavenger tanks for propane-freon bubble chambers in high-energy physics research. Important developing applications for beryllium-copper are trunnions and pivot bearing sleeves for the landing gear of heavy, cargo-carrying aircraft, because these alloys allow the highest stress of any sleeve bearing material.

Production. Beryllium-copper master alloys are produced by direct reduction of beryllium oxide with carbon in the presence of copper in an arc furnace. Because beryllium carbide forms readily, the master alloy is usually limited to about 4.0–4.25% Be. The master alloy is remelted with additional copper and other elements to form commercial alloys. Rolling billets up to 1500 lb (680 kg) have been made by continuous-casting techniques.

Beryllium-copper alloys can be fabricated by all the industrial metalworking techniques to produce principally strip, foil, bar, wire, and forgings. They can be readily machined and can be joined by brazing, soldering, and welding. Annealing, to provide high plasticity at room temperature, is accomplished by heat-treating from 1450 to 1475°F (790 to 802°C) for the high-Be alloys (1650–1700°F or 900–930°C for the low-Be alloys) and water quenching. Precipitation hardening is accomplished by heating to 750–900°F (400–480°C; low Be) and 550–650°F (290–340°C; high Be). *See* COPPER ALLOYS.

Alloys with nickel and iron. Nickel containing 2% Be can be heat-treated to develop a tensile strength of 240,000 lb/in.^{2} (1700 megapascals) with 3–4% elongation. Little commercial use is made of the hard nickel alloys although they have been employed, principally as precision castings, for aircraft fuel pumps, surgical instruments, and matrices in diamond drill bits. Another nickel alloy (Brush M 220C), having 2.0–2.3% Be, 0.5–0.75% C, and the balance of nickel and refractory metals, has been used for mold components and forming tools for glass pressware of optical and container quality. Thermal conductivity, wear resistance, and strength, coupled with unusual machinability for a Ni-Be alloy, make this alloy particularly advantageous for glassworking tooling.

Attempts have been made to add beryllium to a number of ferrous alloys. Small amounts are used to refine grains and deoxidize ferritic steels (Japan), while promising properties have been developed for austenitic and martensitic steels. Stainless steels (Fe-Ni-Cr) may be made maraging by adding 0.15–0.9% Be, developing strengths as high as 260,000 lb/in.^{2} (1800 MPa) as cast (330,000 lb/in.^{2} or 2300 MPa as rolled) while retaining their oxidation- and corrosion-resistant characteristics. Amounts of 0.04–1.5% Be have been added to various Fe-base alloys containing Ni, Cr, Co, Mo, and W for special applications such as watch springs. *See* IRON ALLOYS; NICKEL ALLOYS.

Additions to nonferrous alloys. In materials for nuclear reactors small amounts of beryllium have been added to Mg, Al, Cr, and Zr in Russia for grain refinement and for decreasing surface reactivity. In the United Kingdom a family of reactors (Haddon Hall) used a Mg-Th-Be alloy for fuel element cans.

About 0.1–0.5% Be in Al-Mg alloys refines grains, promotes fluidity, and premits better workability and increased strength. Protection of magnesium from oxidation, burning, and dross formation during casting is achieved by adding as little as 0.005% Be.

Beryllium-base alloys. Three types of beryllium-base alloys are of interest. These consist of dispersed-phase types containing up to 4% BeO; ductile-phase or duplex alloys of Be and Al, particularly 38% Al in Be; and solid solution alloys of up to 4% Cu in Be.

Dispersed-phase alloys containing oxides, carbides, nitrides, borides, and intermetallic compounds in a beryllium matrix are chiefly characterized by increased strength and resistance to creep at elevated temperature. Major commerical alloys in this series are of the fine-grain, high-BeO (4.25–6%), hot-pressed types such as materials used for inertial guidance instruments characterized by high dimensional stability, high-precision elastic limit, 8000–15,000 lb/in.^{2} (55–103 MPa), and good machinability (Brush I-400).

62 Be-38 Al (Lockalloy) was developed as a structural aerospace alloy to combine high modulus and low density with the machining characteristics of the commoner magnesium-base alloys. This alloy in sheet form has at room temperature about 50,000-lb/in.^{2} (344-MPa) ultimate strength, 47,000-lb/in.^{2} (324-MPa) yield strength, and about 8% elongation. It is also produced as extrusions. It has a duplex-type microstructure characterized

TABLE 3. High-temperature, oxidation-resistant beryllides

General category	Example	Wt % of Be	Melting point, °F (°C)	X-ray density, g/cm^3 (oz/in.^{3})
Nb-Be	$NbBe_{12}$	53.8	3070 (1688)	2.92 (1.69)
	Nb_2Be_{19}	48.0	~3100 (1704)	3.17 (1.83)
	Nb_2Be_{17}	45.2	3100 (1704)	3.28 (1.90)
Ta-Be	$TaBe_{12}$	37.4	3360 (1849)	4.18 (2.42)
	Ta_2Be_{17}	29.8	3610 (1988)	5.05 (2.92)
Mo-Be	$MoBe_{12}$	53.2	~3100 (1704)	3.03 (1.75)
Ti-Be	$TiBe_{12}$	69.3	2900 (1593)	2.26 (1.31)
	Ti_2Be_{17}	61.5	2970 (1632)	2.46 (1.42)
Zr-Be	$ZrBe_{13}$	56.2	3500 (1927)	2.72 (1.57)
	Zr_2Be_{17}	45.7	3600 (1982)	3.08 (1.78)

by a semicontinuous aluminum phase. Other alloys of the Be-Al type have been reported, but the 62% Be–38% Al alloy is the most used.

Solid solution alloys containing up to 4% Cu have been intensively investigated because of higher indicated fracture strength, and toughness compared with beryllium.

Intermetallic compounds. Beryllium, combined with most other elements, forms intermetallic compounds having high strength at high temperature—up to 80,000-lb/in.^{2} (552-MPa) modulus of rupture at 2300°F (1260°C), good thermal conductivity, high specific heat, and good oxidation resistance. The BeO film formed by surface oxidation is protective to volatile oxides (Mo) and to elements of high reactivity (Zr and Ti) at temperatures 2000–2700°F (1100–1500°C) and for short times up to 3000°F (1650°C). Some of the lighter refractory beryllides are compared in **Table 3**, showing their composition and low density. Beryllides are also formed with actinide and rare metals, as well as with the transition metals.

The beryllides are of interest to the nuclear field, to power generation, and to aerospace applications. Evaluation of the intermetallics as refractory coatings, reactor hardware, fuel elements, turbine buckets, and high-temperature bearings has been carried out. Wallace W. Beaver; W. Dean Trautman

Bibliography. G. E. Darwin and J. H. Buddery, *Beryllium*, 1960; H. H. Hausner, *Beryllium: Its Metallurgy and Properties*, 1965; O. Kubaschewski (ed.), *Beryllium: Physiochemical Properties of Its Compounds and Alloys*, 1973; H. Okamoto and L. E. Tanner (eds.), *Phase Diagrams of Binary Beryllium Alloys*, 1987, reprint 1990; D. Webster et al. (eds.), *Beryllium Science and Technology*, 2 vols., 1979; C. G. Wilber, *Beryllium: A Potential Contaminant*, 1980.

Beryllium metallurgy

Production of beryllium and its compounds uses 4000–10,000 tons (3600–9000 metric tons) per year of the mineral beryl. The variant amounts of raw materials reflect wide differences in production rates for the beryllium industry. Beryl, a beryllium-aluminum silicate containing (in commercial grades) about 11–13% beryllium oxide (BeO), 17–19% aluminum oxide (Al_2O_3), 64–70% silicon dioxide (SiO_2), and 1–2% alkali metal oxides, is the only mineral used as a raw material for beryllium. Although beryl ore is widely dispersed throughout the world, it is not concentrated in deposits sufficient to justify direct mining. *See* BERYL.

Large crystals of beryl found in pegmatites are a primary source of raw material; beryl is also found as a by-product of the mining of other minerals, such as mica, feldspar, and spodumene. Although concentration processes involving flotation have been tried, none has provided a means of producing quantities of commercial importance. The major producers of beryl ore are Mozambique and Brazil. *See* PEGMATITE.

Little beryl is produced in the United States, although large reserves of low-grade ores appear to be available. The discovery in Utah of the mineral bertrandite dispersed in clay deposits provided a source of domestic raw material. Bertrandite, $4BeO \cdot 2SiO_2 \cdot H_2O$, is dispersed in montmorillonite clays containing less than 0.5% Be. Processes have been developed to recover the beryllium content as BeO of a quality comparable to that resulting from beryl.

Two processes are used to extract beryllium oxide or hydroxide from beryl ore. There are also two methods employed to reduce BeO to beryllium metal. The two extraction methods are based on dissolving beryl as either a fluoride or a sulfate. Reduction is accomplished thermally by means of magnesium with beryllium fluoride, and electrolytically with beryllium chloride. Over 90% of the metal is made from the thermal process. *See* PYROMETALLURGY, NONFERROUS.

Extraction of oxide or hydroxide. In fluoride extraction finely powdered beryl is mixed with sodium fluoferrate (Na_3FeF_6), briquetted, and heated to 1380°F (750°C). The chemical reaction is given below.

$$3BeO \cdot Al_2O_3 \cdot 6SiO_2 + 2Na_3FeF_6 \rightarrow 6SiO_2 + Fe_2O_3 + Al_2O_3 + 3Na_2BeF_4$$

Water-soluble sodium beryllium fluoride (Na_2BeF_4) is removed from the insoluble oxides by filtration after leaching. The filtrate containing Na_2BeF_4 is treated with caustic soda or ammonium hydroxide to precipitate beryllium hydroxide. The resulting hydroxide is decomposed to BeO by firing at 1470°F (800°C). After removal of the beryllium hydroxide

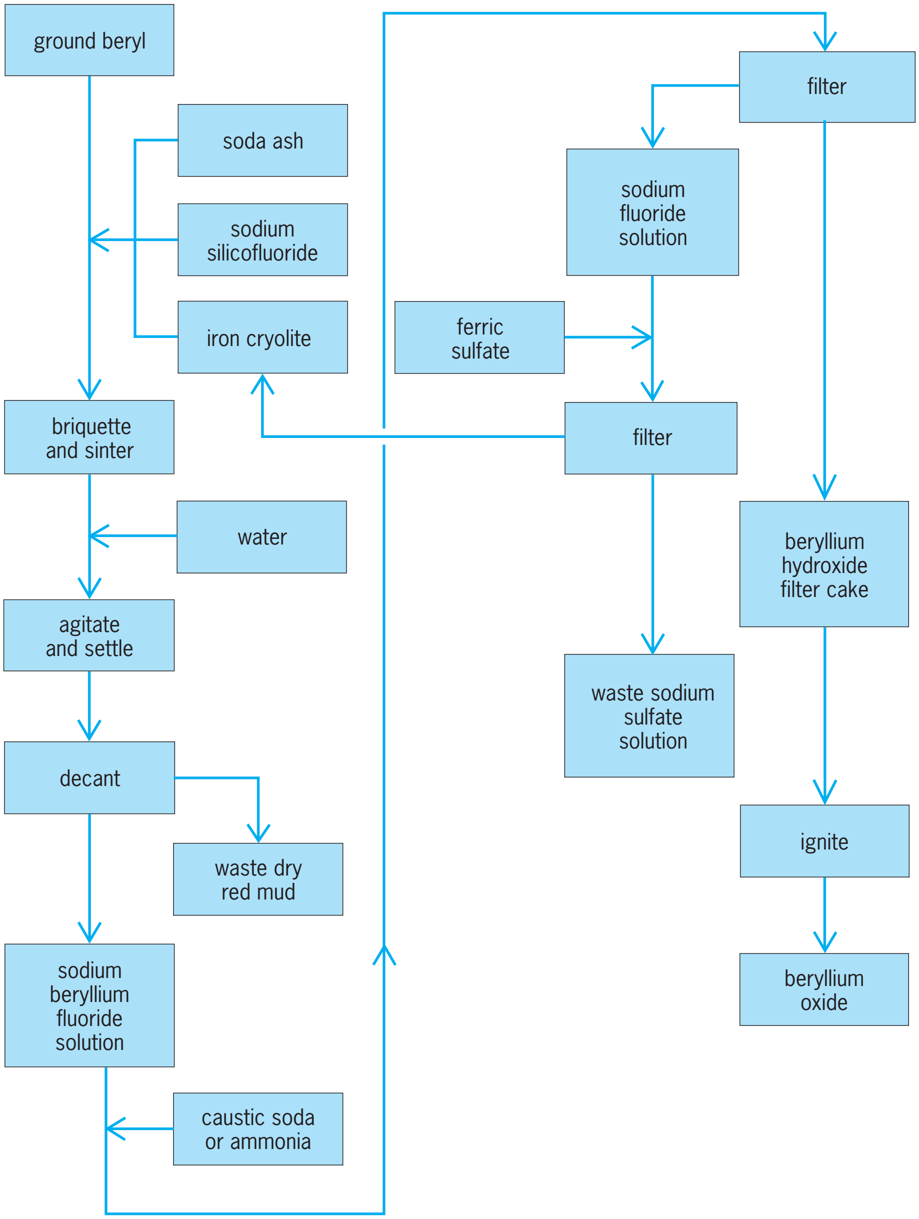

ground beryl
soda ash
sodium silicofluoride
iron cryolite
briquette and sinter
water
agitate and settle
decant
waste dry red mud
sodium beryllium fluoride solution
caustic soda or ammonia
filter
sodium fluoride solution
ferric sulfate
filter
waste sodium sulfate solution
beryllium hydroxide filter cake
ignite
beryllium oxide

Fig. 1. **Flow sheet for the extraction of BeO from beryl by means of the fluoride process. (*Beryllium Corp.*)**

[$Be(OH)_2$], the dilute filtrate containing sodium fluoride (NaF) is treated with ferric sulfate [$Fe_2(SO_4)_3$] to produce Na_3FeF_6 for further beryl extraction (**Fig. 1**).

In sulfate extraction beryl ore is melted at 2957°F (1625°C) and quenched in water to destroy the original crystal structure, enhancing reactivity with sulfuric acid. This beryl glass is heat-treated further to precipitate the BeO out of solid solution and thus to increase leaching yield. The heat-treated frit is then ground and treated with sulfuric acid at 390–570°F (200–300°C). The soluble beryllium and aluminum sulfates are leached out with water and removed from the insoluble oxides. Aluminum is separated by converting it to ammonium aluminum sulfate through addition of NH_4OH and crystallizing. Chelating compounds are added to hold iron and nickel in solution. Sufficient sodium hydroxide is then added to convert the beryllium to sodium beryllate, which is hydrolyzed to precipitate beryllium hydroxide.

The use of solvent extraction for the purification of beryllium-containing solutions that result from solubilizing beryl and other beryllium minerals is also used. Liquid-liquid extraction is especially attractive for purifying highly impure liquors with beryllium contents of 0.13 oz/gal (0.2–0.4 g/liter). When this process is used with beryl, the steps between Be-Al sulfate solution and hydrolysis are replaced (**Fig. 2**). *See* EXTRACTION; HYDROMETALLURGY; LEACHING.

Beryllium oxide. In the manufacture of ceramic-grade beryllium oxide, impure beryllium hydroxide is generally converted to either the sufate, nitrate, oxalate, basic carbonate, or basic acetate and is purified by recrystallization. The purified salt is then decomposed thermally to yield beryllium oxide. The sinterability of the resulting oxide is chiefly a function of temperature, which should be sufficiently high

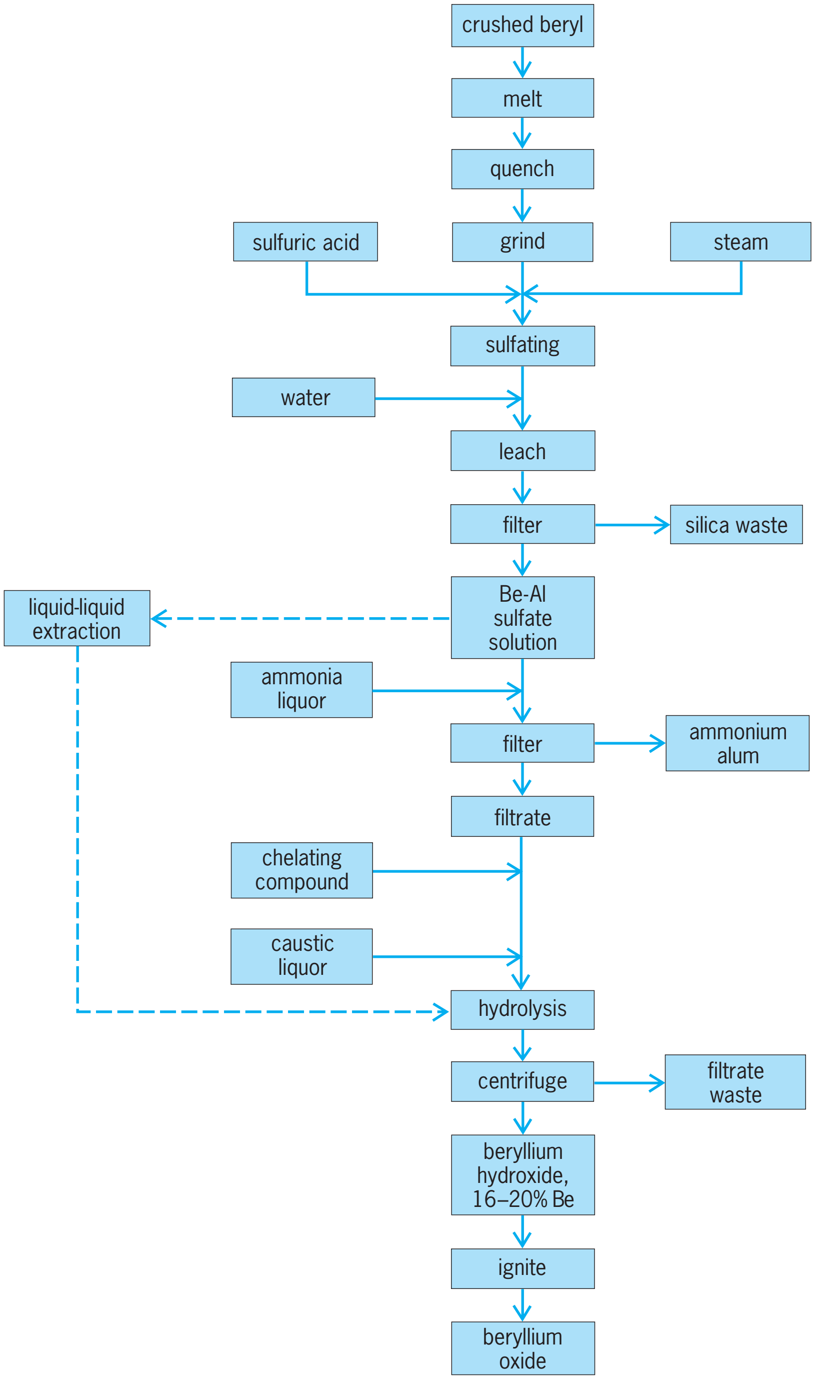


Fig. 2. **Flow sheet for the extraction of beryllium oxide, BeO, from beryl by the sulfate process. (*Brush Beryllium Co.*)**

to complete decomposition, but also sufficiently low to produce active surfaces. Generally, calcining at 1650–1830°F (900–100°C) gives sinterable-grade oxide, whereas higher temperatures produce a more refractory oxide. Beryllium oxide is primarily employed as a refractory for the melting of reactive metals, for high-temperature nuclear applications, and as a thermally conducting, electrically resisting ceramic.

Beryllium metal. In thermal reduction beryllium hydroxide or oxide is treated with ammonium bifluoride to form $(NH_4)_2BeF_4$. This solution is purified by adding lead dioxide to remove manganese and chromium, calcium carbonate to precipitate aluminum, and after removal of these impurities, ammonium polysulfide to precipitate remaining heavy elements. The purified solution is evaporated and crystallized to salt, which is decomposed to NH_4F and BeF_2 at temperatures of 1650–2000°F (900–1100°C).

Beryllium fluoride is mixed with an amount of magnesium corresponding to about 75% of that stoichiometrically required, and is reduced at 1650°F (900°C) in a graphite crucible employing a high-frequency electric furnace. After the reaction is complete, the temperature is raised to about 2550°F (1400°C) to allow the beryllium to collect. Excess BeF_2 is leached from the slag and the beryllium pebble and MgF_2 are separated. The resulting pebbles, containing slag and some magnesium, are vacuum-melted for purification in a BeO crucible. The metal is poured into graphite molds to produce castings of 120 lb (54.5 kg) each.

For electrolysis beryllium chloride can be made by passing chlorine gas over BeO and carbon at 1830°F (1000°C). $BeCl_2$ is collected as a sublimate in a condenser maintained below 750°F (400°C). The electrolytic cell may consist of a nickel or stainless-steel pot serving as a cathode and of a graphite electrode serving as an anode. A cover is provided to control atmospheric conditions in the pot.

The electrolyte consists of beryllium chloride mixed with sodium chloride. Electrolysis is carried out at about 700°F (370°C). Beryllium flakes, resulting from electrolysis, are leached in water and dried in alcohol. As in the case of pebbles thermally reduced by magnesium from BeF_2, the electrolytic flake is generally vacuum-cast to remove chlorides. *See* ELECTROLYSIS.

Fabrication. Vacuum-cast beryllium ingots either are produced as raw material for the powder-metal process, which accounts for over 90% of the material being used, or are cast directly for fabrication. In the latter process square billets are usually cast by bottom pouring, using a chill plate at the bottom of the mold and a hot top for directional cooling. This process produces sound ingots.

For powder metallurgy round vacuum-cast ingots are converted into chips, or swarf, by multiple turning and cutting operations. The chips are attritioned between beryllium plates in a controlled atmosphere. Powders from 200 mesh down to 15 micrometers maximum particle size are produced for various applications. For finer powders the attritioned particles are comminuted by ball milling.

The major portion of the powdered metal is hot-pressed at 100–2000 lb/in.^{2} (0.69–13.8 megapascals) in steel or graphite dies under vacuum at temperatures of 1920–2100°F (1050–1150°C). Sintering of slip-cast, isopressed, or axially pressed powders can be accomplished at 2057–2190°F (1125–1200°C), and sintering of loose powder employs temperatures up to 2255°F (1235°C). Porous bodies are made by sintering after mixing with a blowing agent for low densities; by temperature, time, and particle-size control for high densities; and with fugitive binders for intermediate densities.

The sintered product is characterized by thin films of BeO of varying amounts defining the grain boundaries and by fine grains (5–20 micrometer average) with little tendency for gain growth at elevated temperatures. Ingot beryllium has a low oxide content, but tends to have coarser grain structure (greater than 50 μm) with high grain-growth potential after working. Consequently, machined parts are usually made from powder-metal products, while ingot beryllium is usually rolled or otherwise worked.

Rolling, forging, extrusion, and drawing are performed between the temperature of maximum ductility (750°F or 400°C) and the average recrystallization temperature (1470°F or 800°C). The bulk of rolling, forging, and extrusion is carried out at 1110–1470°F (600–800°C), while drawing is carried out at the lower end of the fabrication range. Because of the lower impurity level in ingot, greater reduction or lower temperatures of fabrication are possible than with the finer-grained sintered metal. Usually, fabrication processes are carried out by using protective lubricants or metal covers to prevent damage to the beryllium or work tooling surfaces. Hot rolling may be carried out in steel frames and jackets for support. Above 1470°F (800°C) it is mandatory to fully jacket to provide support and prevent oxidation; only a few extrusion operations use temperatures above 1470°F (800°C). Beryllium sheet can be formed, shaped, spun, and punched. Adequate surface protection and slow forming speeds are advantageous.

Beryllium can be welded, soldered, brazed, or plastically bonded, with brazing and plastic bonding, as well as mechanical fasteners, being the usual production methods. Surface protection of finished beryllium surfaces can be provided by anodizing, plating, optical polishing, or using conversion coatings. Chemical machining and chemical milling are used to provide patterned and low-damage surfaces.

Physical metallurgy. Extensive research has been conducted on the close-packed hexagonal beryllium crystal. Failure takes place by basal plane cleavage at relatively low stresses, but it can be increased with solid solution alloying elements. The critical resolved shear stress in beryllium crystals is largely a function of purity in an inverse relation to the glide strain slip. Prism slip occurring at room temperature is much greater than basal slip. When basal slip is increased by extensive purification, prism slip is increased even more, promoting severe crystal anisotropy even with extensive basal and prism slip available. *See* BERYLLIUM; BERYLLIUM ALLOYS.

In polycrystalline aggregates the plasticity is low at temperatures up to about 480°F (250°C) at which other slip planes become active. The metal becomes quite plastic up to temperatures where grain boundary failure becomes predominant, this type of failure defining the elevated temperature plastic working range. Strength is increased by smaller grain sizes and the presence of solid solution and dispersion hardening elements. Ductility is favored by preferred crystal alignment and finer grains. On working polycrystalline beryllium, severe structural anisotropy occurs, providing little plasticity in the direction of sheet thickness, but as much as 30% elongation in the plane of the sheet. Beryllium has poor notch sensitivity and tends to fracture under high rates of loading.

For design purposes beryllium has excellent compressive and shear properties with good tensile strength, especially at elevated temperatures. Its modulus of 42,000,000 lb/in.^{2} (290 gigapascals) in tension gives the highest stiffness-to-weight ratio of any metallic element. Fatigue properties are high. *See* METAL, MECHANICAL PROPERTIES OF.

Applications. Because of its ability to reflect, moderate, and react (n, α; n, $2n$ reactions) with neutrons, beryllium has had a long history of use in atomic energy as neutron sources, reflectors, and moderators in thermal and intermediate reactors. About all high-power engineering and test reactors utilize beryllium, as do package airborne power systems, such as SNAP and other thermionic-type reactors, and direct thermal systems, such as Phoebus and Nerva. Much auxiliary nuclear equipment also is made of beryllium, including photoneutron sensitizers, monochromatic neutron spectroscopes, and supporting hardware. A number of experimental power systems (France, United Kingdom, Germany, and United States) have employed beryllium as a fuel jacket cover.

Application of beryllium to aerospace usually provides weight saving up to 60% of that of a competing material on the same design basis. Other applications of beryllium are based on properties such as thermal stability (infrared mirrors), creep resistance and thermal capacity (Mercury heat shields and Mercury and Gemini cabin sections), and thermal conductivity (C5-A and other heavy aircraft brakes). *See* ATMOSPHERIC ENTRY.

Extensive use of beryllium in reentry structures (Ranger, Mariner, Polaris, and Poseidon) depends on high-temperature strength and thermal capacity. The metal's transparency to radiation (as used in x-ray windows) also is important to missile structures, which must withstand electromagnetic pulse conditions created by antimissile tactics. Finally, ability to machine to high tolerance coupled with dimensional stability has created almost exclusive employment of beryllium in inertial guidance navigation and control devices. *See* ACOUSTIC RADIOMETER; ELECTROMAGNETIC RADIATION.

Inasmuch as beryllium powder has one of the highest heats of reaction of any element, it may be mixed with oxidizers and binders in liquid, solid, and hybrid types of high-energy propellants. Solid-propellant motors containing up to 11,000 lb (5000 kg) of fuel mixtures of beryllium have been made.

Experimental evaluation of beryllium metal has been conducted for jet engine compressor blades, disks, shafts, and related hardware; actuators, tube supports, and fittings for space vehicles; vertical and horizontal stabilizers; skins for elevator control surfaces and trailing edges; and internal support hardware for aircraft. Applications are in high-speed

-s

process machinery, for stiffening other metals or plastics in composites or as honeycomb, and as a cryogenic conductor for transmission of electrical energy. *See* SUPERCONDUCTIVITY.

Wallace W. Beaver; W. Dean Trautman

Beryllium minerals

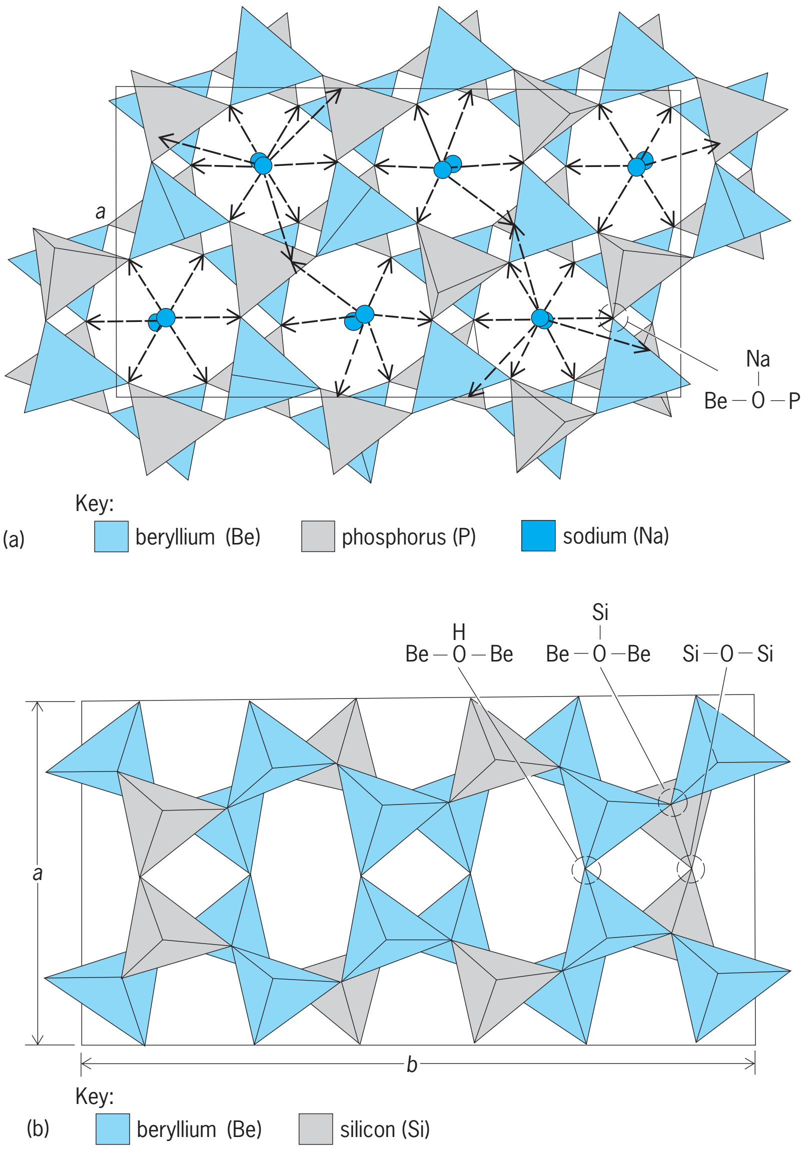
Minerals containing beryllium as an essential component. Over 50 beryllium minerals have been identified, even though beryllium is a scarce element in the Earth's crust. The unusual combination of low charge (+2) and small ionic radius (0.035 nanometer) of the beryllium ion accounts for this diverse group of minerals and their occurrence in many natural environments. *See* BERYLLIUM.

Principal minerals. The **table** lists many of the geologically and mineralogically significant beryllium minerals. Nearly all beryllium minerals can be included in one of three groups: compositionally simple oxides and silicates with or without aluminum; sodium- and calcium-bearing silicates; and phosphates and borates. The first group is by far the most abundant; it contains beryl, the most common beryllium mineral, plus the common minerals phenakite, bertrandite, chrysoberyl, and euclase. Of this group, only beryl shows a wide compositional variation; it can have percent quantities of the alkali elements, volatile components, and transition metals. The other minerals, although widely distributed, rarely constitute more than a fraction of a percent of any rock. *See* BERYL.

Some significant beryllium minerals

Name	Formula*
Barylite	$Ba[Be_2Si_2O_7]$
Behoite	$Be(OH)_2$
Bertrandite	$Be_4Si_2O_7(OH)_2$
Beryl	$Be_3Al_2Si_6O_{18}\cdot nH_2O$
Beryllonite	$Na[BePO_4]$
Bityite	$CaLiAl_2[AlBeSi_2O_{10}](OH)_2$
Bromellite	BeO
Chkalovite	$Na_2[BeSi_2O_6]$
Chrysoberyl	$BeAl_2O_4$
Euclase	$BeAlSiO_4(OH)$
Eudidymite	$Na[BeSi_3O_7](OH)$
Gadolinite	$Y_2FeBe_2Si_2O_{10}$
Gugiaite	$Ca_2BeSi_2O_7$
Hambergite	$Be_2BO_3(OH)$
Helvite group	$(Mn,Fe,Zn)_4[BeSiO_4]_3S$
Herderite	$CaBe(PO_4)(F,OH)$
Hurlbutite	$Ca[BePO_4]_2$
Phenakite	Be_2SiO_4

* Brackets indicate framework component of formula.



Arrangements of tetrahedra in beryllium minerals. (*a*) Beryllonite; oxygen shared by two tetrahedra. (*b*) Bertrandite; oxygen shared by three tetrahedra. (*After M. Ross, Crystal chemistry of beryllium, USGS Prof. Pap. 468, 1964*)

Crystal structures. The beryllium minerals have many structural characteristics similar to the major rock-forming silicate minerals, but are distinguished by containing large quantities of tetrahedrally coordinated beryllium ion (Be^{2+}) in place of, or in addition to, tetrahedrally coordinated aluminum ion (Al^{3+}) and silicon ion (Si^{4+}). In order to maintain electrical neutrality in the crystal structures, the lower charge on the beryllium ion requires either simultaneous substitution of other elements with higher charge or modification of the structural arrangement. Nearly all beryllium minerals have extensive or complete sharing of corners of beryllium-oxygen (BeO_4^{6-}) tetrahedra with other BeO_4^{6-}, and/or SiO_4^{4-}, AlO_4^{5-}, or phosphorus-oxygen (PO_4^{3-}) tetrahedra to form continuous two- or three-dimensional sheet or framework structures. In many minerals, the apical oxygen ions are shared by only two tetrahedra, in which case the excess charge on the oxygen is balanced by nontetrahedral cations (see **illus.**). These structures can closely resemble those of common silicates (for example, beryl = cordierite; beryllonite = nepheline; helvite = sodalite; bityite = margarite; gugiaite = melilite; hurlbutite = feldspar). In other minerals (for example, phenakite, bertrandite, and euclase), the apical oxygen ions are shared by two beryllium and one silicon tetrahedra, leading to local charge balance and also leading to structural linkages distinct from the common silicates (see illus.). A few beryllium minerals do not fall in these categories but have other structures, including chrysoberyl (olivine structure) and hambergite. *See* CHRYSOBERYL; SILICATE MINERALS.

Stabilities. Only minerals in the BeO-Al_2O_3-SiO_2-H_2O system have well-determined stability relationships. They are typified by the progressive replacement of hydrous phases by anhydrous equivalents with increasing temperature. Bertrandite dehydrates to phenakite at 460–520°F (240–260°C). Euclase is stable to 750–1050°F (400–600°C), depending on the water pressure. Beryl has a wide stability range, which is enhanced by the incorporation of alkalies and water into its structure. Chrysoberyl and phenakite are stable to more than 2500°F (1400°C) and pressures of more than 20 kilobars (2 gigapascals). Equilibria in more complex systems are conveniently cast in terms of the activities (thermodynamic concentrations) of silica (SiO_2) and alumina (Al_2O_3). In so doing, it is seen that the minerals order bromelite $<$ chrysoberyl $<$ euclase $<$ bertrandite $=$ phenakite $=$ beryl for compatibility with silica, and bromellite $<$ bertrandite $=$ phenakite $<$ beryl $<$ euclase $<$ chrysoberyl for compatibility with alumina. In natural environments, the more abundant, rock-forming minerals govern the activities of silica and alumina and thus which beryllium mineral is stable.

The relative stabilities of other beryllium minerals can be inferred from their natural occurrences; no experimental studies have been done. The phosphates and borates (for example, beryllonite, hambergite, and herderite) are stable only at geologically high activities of phosphate and borate. The aluminum-free sodium- and calcium-bearing silicates (for example, chkalovite, eudidymite, and gugiaite) are stable only with high alkali activities and low alumina activities. The helvite group requires moderately, but not strongly, sulfidizing conditions.

Geological occurrence. Beryllium minerals occur in many geological environments, where they are generally associated with felsic (abundant feldspar ± quartz) igneous rocks and related, metasomatically altered rocks. Minerals from the BeO-Al_2O_3-SiO_2-H_2O system occur in virtually all environments where separate beryllium phases exist. The minerals bearing sodium (Na), calcium (Ca), sulfur (S), phosphorus (P), boron (B), manganese (Mn), and fluorine (F) have more restricted occurrences, reflecting the special conditions required for their stabilization. Quartz-bearing (silica-saturated) settings include the feldspar- and mica-bearing environments typical of granites, pegmatites and many ore deposits, and the alkali-deficient vein and greisen environments. Beryl, phenakite, and bertrandite occur in all settings; restricted to particular types are euclase (alkali-deficient settings, because it is incompatible with feldspar), chrysoberyl (high-temperature only, because it is incompatible with quartz at low temperature), and the phosphates and helvite-group minerals (in pegmatites and ore deposits, where phosphorus and sulfur are concentrated by magmatic or hydrothermal processes). Quartz-absent (silica-undersaturated) environments include skarns (silicate replacement of carbonate rocks), desilicated pegmatites, and some low-temperature volcanogenic deposits. Chrysoberyl, phenakite, and bertrandite are typical of the silica-undersaturated environments; less commonly beryl, bromellite, behoite, and helvite-group minerals occur. Alkali-excess igneous environments, including many syenites, commonly contain phenakite, bertrandite, and the alkali-beryllium silicates (for example, chkalovite and eudidymite).

Uses. Beryl and bertrandite, mined from granitic pegmatites and altered volcanic rocks, are the principal ores of beryllium; deposits of chrysoberyl and phenakite may become economically significant in the future. The colored varieties of beryl (emerald, aquamarine, morganite) are valued gemstones; chrysoberyl, phenakite, and a few of the other minerals are less common gemstones. Synthetic bromellite has specialized applications in the ceramic and nuclear industries. *See* EMERALD; GEM. **Mark D. Barton**

Bibliography. M. D. Barton, Phase equilibria and thermodynamic properties of minerals in the BeO-Al_2O_3-SiO_2-H_2O system, with petrologic applications, *Amer. Mineralog.*, 71:277–300, 1986; P. Cerny (ed.), *Granitic Pegmatites in Science and Industry*, Mineralogical Association of Canada Short Course Notes, pp. 329–346, 1982; M. Ross, *Crystal Chemistry of Beryllium*, USGS Prof. Pap. 468, 1964.

Bessel functions

By definition the solutions of Bessel's differential equation, Eq. (1). Bessel functions, also called cylinder functions, are examples of special functions which are introduced by a differential equation.

$$z^2 d^2y/dz^2 + z dy/dz + (z^2 - \nu^2)y = 0 \qquad (1)$$

Bessel functions are of great interest in purely mathematical concepts and in mathematical physics. They constitute additional functions which, like the elementary functions z^n, $\sin z$, or e^z, can be used to express physical phenomena.

Applications of Bessel functions are found in such representative problems as heat conduction or diffusion in circular cylinders, oscillatory motion of a sphere in a viscous fluid, oscillations of a stretched circular membrane, diffraction of waves by a circular cylinder of infinite length or by a sphere, acoustic or electromagnetic oscillations in a circular cylinder of finite length or in a sphere, electromagnetic wave propagation in the waveguides of circular cross section, in coaxial cables, or along straight wires, and in skin effect in conducting wires of circular cross section. In these problems Bessel functions are used to represent such quantities as the temperature, the concentration, the displacements, the electric and magnetic field strengths, and the current density as function of space coordinates. The Bessel functions enter into all these problems because boundary values on circles (two-dimensional problems), on circular cylinders, or on spheres are prescribed, and the solutions of the respective problems are sought either inside or outside the boundary, or both.

Definition of Bessel functions. The independent variable z in Bessel's differential equation may in applications assume real or complex values. The parameter ν is, in general, also complex. Its value is called the order of the Bessel function. Since there are two linearly independent solutions of a linear differential equation of second order, there are two independent solutions of Bessel's differential equation. They cannot be expressed in finite form in terms of elementary functions such as z^n, $\sin z$, or e^z unless the parameter is one-half of an odd integer. They can, however, be expressed as power series with an exception for integer values of ν. The function defined by Eq. (2)

$$J_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{l=0}^{\infty} \frac{(-z^2/4)^l}{l!\,\Gamma(\nu+l+1)} \tag{2}$$

is designated as Bessel's function of the first kind, or simply the Bessel function of order ν. $\Gamma(\nu+l+1)$ is the gamma function. The infinite series in Eq. (2) converges absolutely for all finite values, real or complex, of z. In particular, Eqs. (3) may be expressed. Along

$$\begin{aligned} J_0(z) &= 1 - \frac{1}{1!1!}\left(\frac{z}{2}\right)^2 + \frac{1}{2!2!}\left(\frac{z}{2}\right)^4 \\ &\qquad - \frac{1}{3!3!}\left(\frac{z}{2}\right)^6 + \cdots \\ J_1(z) &= \frac{z}{2} - \frac{1}{1!2!}\left(\frac{z}{2}\right)^3 + \frac{1}{2!3!}\left(\frac{z}{2}\right)^5 - \cdots \end{aligned} \tag{3}$$

with $J_\nu(z)$, there is a second $J_{-\nu}(z)$. It is linearly independent of $J_\nu(z)$ unless ν is an integer of n. In this case $J_{-n}(z) = (-1)^n J_n(z)$. *See* GAMMA FUNCTION.

The Bessel function of the second kind, also called Neumann function, is defined by Eq. (4).

$$Y_\nu(z) = \frac{\cos \nu\pi J_\nu(z) - J_{-\nu}(z)}{\sin \nu\pi} \tag{4}$$

If $\nu = n$, this expression is indeterminate, and the limit of the right member is to be taken. In particular, Eq. (5) may be expressed, with $\ln \gamma = 0.577215\ldots$

$$\begin{aligned} Y_0(z) &= \frac{2}{\pi} J_0(z) \ln\left(\frac{1}{2}\gamma z\right) \\ &\quad + \frac{2}{\pi}\left[\frac{1}{1!1!}\left(\frac{z}{2}\right)^2 = -\frac{1}{2!2!}\left(1+\frac{1}{2}\right)\left(\frac{z}{2}\right)^4\right. \\ &\quad \left. + \frac{1}{3!3!}\left(1+\frac{1}{2}+\frac{1}{3}\right)\left(\frac{z}{2}\right)^6 - \cdots\right] \end{aligned} \tag{5}$$

(Euler's constant). There are two Bessel functions of the third kind, designated as first and second Hankel functions. They are defined as Eqs. (6).

$$\begin{aligned} H_\nu^{(1)}(z) &= J_\nu(z) + iY_\nu(z) \\ H_\nu^{(2)}(z) &= J_\nu(z) - iY_\nu(z) \end{aligned} \tag{6}$$

Elementary properties. For $z = 0$, then $J_0(0) = 1$, $J_n(0) = 0$ if $n = 1, 2, \ldots$, while $Y_n^{(1)}$ and therefore also $H_n^{(1)}$ and $H_n^{(1)}$ infinite for all n. The behavior of Bessel functions for small values of z is readily seen from the power series expansion in Eq. (2), if ν is not an integer. For integer $\nu = n = 0$, one has for small z the relations in notation (7) (the sign $\approx$

$$\begin{aligned} J_n(z) &\approx \left(\frac{z}{2}\right)^n \Big/ n! \\ Y_0(z) &\approx \frac{2}{\pi} \ln \frac{\gamma z}{2} \\ Y_n(z) &\approx -\frac{1}{\pi}\left(\frac{z}{2}\right)^{-n} (n-1)! \\ &\quad (n = 1, 2, \ldots) \end{aligned} \tag{7}$$

means approximately equal). Here $n! = 1 \cdot 2 \cdot 3 \ldots n$, $0! = 1$. The behavior of the functions for large values of $z(|z|\Gamma|\nu|^2)$ is given for all values of ν by the formulas in notation (8). They hold not only for real

$$\begin{aligned} J_\nu(z) &\approx \left(\frac{2}{\pi z}\right)^{1/2} \cos\left(z - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) \\ Y_\nu(z) &\approx \left(\frac{2}{\pi z}\right)^{1/2} \sin\left(z - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) \\ H_\nu^{(1)}(z) &\approx \left(\frac{2}{\pi z}\right)^{1/2} \exp\left[i\left(z - \frac{\nu\pi}{2} - \frac{\pi}{4}\right)\right] \\ H_\nu^{(2)}(z) &\approx \left(\frac{2}{\pi z}\right)^{1/2} \exp\left[-i\left(z - \frac{\nu\pi}{2} - \frac{\pi}{4}\right)\right] \end{aligned} \tag{8}$$

positive values of z but also for complex values in the angular domain $-\pi < \arg z < \pi$. These formulas are said to be asymptotic. This means that the difference of the left and right member, multiplied by a certain power of z, which is here $z^{3/2}$, is bounded as z goes to infinity along a straight line in the angular domain mentioned. The asymptotic formulas can be used to obtain approximate numerical values of the functions for near large $|z|$. They reveal that the Bessel functions are closely related to the functions $\cos z$, $\sin z$, $\exp(iz)$, $\exp(-iz)$, respectively.

The general behavior of Bessel functions can to a certain extent be inferred from the behavior near $z = 0$ and from the asymptotic behavior. The functions $J_\nu(z)$ and $Y_\nu(z)$ of real positive argument z and real order ν are oscillatory in z with a period that is not quite a constant but approaches 2π as $z \to \infty$. The oscillations have, however, an amplitude which decreases in accordance with $(2/\pi z)^{1/2}$ as $z \to \infty$. They are used in applications to represent standing cylindrical and spherical waves. The behavior of Hankel functions follows then from Eqs. (6). They are used to represent progressing cylindrical and spherical waves.

Figures 1 and **2** give the Bessel and Neumann functions for $\nu = 0, 1, 2$ and positive values of z.

The large zeros $z_{\nu,s}$ of the function $J_\nu(z) \cos \alpha + Y_\nu(z) \sin \alpha$ with α a constant are given by the asymptotic expression $z_{\nu,s}^2 \approx [(s + \nu/2 - 1/4)\pi + \alpha]^2 + 1/4 - \nu^2$, in which s assumes large integer values. Even the first positive zero of $J_0(z)$ is very accurately given by this formula, namely, $z_{0,1} \approx 2.409$, as compared with the accurate value $z_{0,1} = 2.4048\ldots$

There exist two relations between Bessel functions whose orders differ by one. If $C_\nu(z)$ stands for any one of the four kinds of functions $J_\nu(z)$, $Y_\nu(z)$, $H_\nu^{(1)}(z)$, $H_\nu^{(1)}(z)$, then Eqs. (9) may be written, and

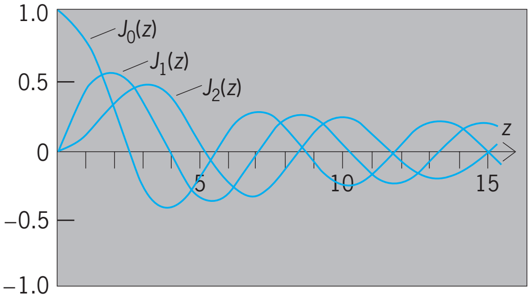


Fig. 1. Bessel functions $J_0(z)$, $J_1(z)$, $J_2(z)$ for $\nu = 0, 1, 2$ and positive values of z.

hence Eqs. (10) hold.

$$\begin{aligned}(2\nu/z)C_\nu(z) &= C_{\nu-1}(z) + C_{\nu+1}(z)\\ 2dC_\nu(z)/dz &= C_{\nu-1}(z) - C_{\nu+1}(z)\end{aligned} \qquad (9)$$

$$\begin{aligned}z^\nu C_{\nu-1}(z) &= d[z^\nu C_\nu(z)]/dz\\ z^{-\nu} C_{\nu+1}(z) &= -d[z^{-\nu} C_\nu(z)]/dz\end{aligned} \qquad (10)$$

Further properties. Some differential equations can be solved in terms of Bessel functions. The differential equation (11) has the solutions $y = z^{\beta\nu-\alpha}C_\nu(\gamma z^\beta)$, while $d^2y/dz^2 + (e^{2z} - \nu^2)y = 0$

$$z^2\frac{d^2y}{dz^2} + (2\alpha - 2\nu\beta + 1)z\frac{dy}{dz} + [\beta^2\gamma^2 z^{2\beta} + \alpha(\alpha - 2\nu\beta)]y = 0 \qquad (11)$$

has the solutions $y = C\nu(e^2)$. There are series with terms containing Bessel functions, whose sum can be expressed by elementary functions, for example, Eqs. (12). In the last formula the functions $P_n(u)$ are

$$\begin{aligned}\exp(iz\cos\theta) &= J_0(z) + 2\sum_{n=1}^{\infty} i^n J_n(z)\cos n\theta\\ 1 &= J_0^2(z) + 2\sum_{n=1}^{\infty} J_n{}^2(z)\\ e^{iuz} &= (\pi/2z)^{1/2}\sum_{n=0}^{\infty} i^n(2n+1)J_{n+1/2}(z)P_n(u)\end{aligned} \qquad (12)$$

the Legendre polynomials. There are also addition theorems of the type in Eq. (13).

$$J_0(x \pm y) = J_0(x)J_0(y) + 2\sum_{n=1}^{\infty}(\mp 1)^n J_n(x)J_n(y) \qquad (13)$$

Bessel functions are also useful because many definite integrals can be expressed in terms of Bessel functions. Examples are shown as Eqs. (14). The

$$\pi J_n(z) = i^{-n}\int_0^\pi e^{iz\cos\alpha}\cos n\alpha\, d\alpha \qquad (n = \text{integer})$$

$$\frac{\pi}{2}x^{1/2}\left[J_{1/3}(2x^{3/2}) + J_{-1/3}(2x^{3/2})\right] = \int_0^\infty \cos(t^3 - 3tx)\, dt \qquad (14)$$

last integral is called Airy's integral. Definite integrals containing Bessel functions can often be expressed by elementary functions, for example, Eq. (15),

$$\int_0^\infty e^{-t}J_0(at)\, dt = (1 + a^2)^{-1/2} \qquad (15)$$

when the imaginary part of a is between −1 and +1.

Expansions in terms of Bessel functions. In applications there is often the problem of expanding a function in terms of Bessel functions or expressing it as an integral with Bessel functions in the integrand.

The Fourier-Bessel series is a generalization of the Fourier series. Let $z_1, z_2, z_3, \ldots$ be the positive zeros of $J_\nu(z)$ for some real $\nu = -\frac{1}{2}$ arranged in increasing order of magnitude. Then a continuous function $f(x)$ defined in the interval $0 \leqq x \leqq 1$ can be expanded into the Fourier-Bessel series, Eq. (16), where the coefficients a_m are given by Eq. (17).

$$f(x) = \sum_{m=1}^{\infty} a_m J_\nu(z_m x) \qquad (0 < x < 1) \qquad (16)$$

$$\left[J_{\nu+1}(z_m)\right]^2 a_m = \int_0^1 tf(t)J_\nu(z_m t)\, dt \qquad (17)$$

The Fourier-Bessel integral is a generalization of the Fourier integral. Let $f(x)$ be a continuous function, such that the integral

$$\int_0^\infty f(x)\, dx$$

exists and is absolutely convergent. Furthermore, let ν be a real number $= -{}^1\!/_2$. Then the function $f(x)$ can be represented by an integral containing Bessel functions, the Fourier-Bessel integral, Eq. (18), where

$$f(x) = \int_0^\infty t^{1/2}J_\nu(xt)g(t)\, dt \qquad (18)$$

$g(t)$, the Fourier-Bessel transform of $f(x)$, is given by Eq. (19). Both representations hold also if $f(x)$ has a

$$g(t) = \int_0^\infty x^{1/2}J_\nu(xt)f(x)\, dx \qquad (19)$$

finite number of discontinuities, except at points of discontinuity themselves.

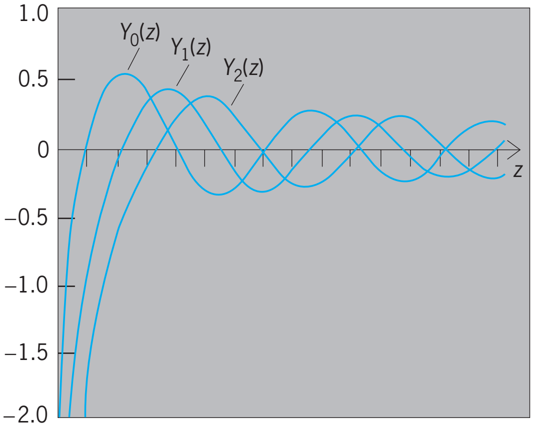


Fig. 2. Neumann functions $Y_0(z)$, $Y_1(z)$, $Y_2(z)$ for $v = 0, 1, 2$ and for positive values of z.

Functions related to Bessel functions. In some applications, Bessel functions occur with a fixed value of $\arg z \neq 0$, for instance with $\arg z = \pi/2$, that is, positive imaginary z. For convenience, special notations are used in such cases. For positive x the functions in Eqs. (20) are defined. These functions are called

$$\begin{aligned} I_\nu(x) &= \exp(-i\nu\pi/2)J_\nu(ix) \\ K_\nu(x) &= (i\pi/2)\exp(i\nu\pi/2)H_\nu^{(1)}(ix) \end{aligned} \tag{20}$$

modified Bessel function and modified Hankel function or MacDonald functions of order ν. For real values of $\nu = 0$, the functions $I_\nu(z)$ are monotonically increasing to infinity while the functions $K_\nu(z)$ are monotonically decreasing from infinity to zero as z goes through real values from 0 to ∞.

Particularly in problems for which the use of spherical coordinates is appropriate, there occur Bessel functions with a factor $z^{-1/2}$. They are called spherical Bessel, Neumann, and Hankel functions and defined by Eqs. (21). These functions can be expressed

$$\begin{aligned} \psi_\nu{}^{(1)}(z) &= (\pi/2z)^{1/2}J_{\nu+1/2}(z) \\ \psi_\nu{}^{(2)}(z) &= (\pi/2z)^{1/2}Y_{\nu+1/2}(z) \\ \psi_\nu{}^{(3)}(z) &= (\pi/2z)^{1/2}H^{(1)}{}_{\nu+1/2}(z) \\ \psi_\nu{}^{(4)}(z) &= (\pi/2z)^{1/2}H^{(2)}{}_{\nu+1/2}(z) \end{aligned} \tag{21}$$

by elementary functions in finite form, if ν is an integer. In particular, Eqs. (22) may be written.

$$\begin{aligned} \psi_0{}^{(1)}(z) &= \frac{\sin z}{z} \qquad \psi_{-1}{}^{(1)}(z) = \frac{\cos z}{z} \\ \psi_1{}^{(1)}(z) &= \frac{-\cos z}{z} + \frac{\sin z}{z^2} \\ \psi_{-2}{}^{(1)}(z) &= \frac{-\sin z}{z} - \frac{\cos z}{z^2} \end{aligned} \tag{22}$$

See FOURIER SERIES AND TRANSFORMS. Josef Meixner

Bibliography. M. Abramowitz and I. A. Stegun (eds.), *Handbook of Mathematical Functions*, 10th ed., 1972; L. C. Andrews, *Special Functions of Mathematics for Engineers*, 2d ed., 1992; A. Erdelyi (ed.), *Higher Transcendental Functions*, 3 vols., 1953–1955, reprint 1981; A. Erdelyi (ed.), *Tables of Integral Transforms*, 2 vols., 1954; A. N. Kiforov and V. Uvarov, *Special Functions of Mathematical Physics*, 1988; Z. X. Wang and D. R. Guo, *Special Functions*, 1989; G. N. Watson, *Treatise on the Theory of Bessel Functions*, 2d ed., 1944.

Beta particles

The name first applied in 1897 by Ernest Rutherford to one of the forms of radiation emitted by radioactive nuclei. Beta particles can occur with either negative or positive charge (denoted β^- or β^+) and are now known to be either electrons or positrons, respectively. Electrons and positrons are now referred to as beta particles only if they are known to have originated from nuclear beta decay. Their observed kinetic energies range from zero up to about 5 MeV in the case of naturally occurring radioactive isotopes, but can reach values well over 10 MeV for some artificially produced isotopes. *See* ALPHA PARTICLES; ELECTRON; GAMMA RAYS; POSITRON; RADIOACTIVITY.

Properties. When a nucleus beta-decays, it emits two particles at the same time: One is a beta particle; the other, a neutrino or antineutrino. With this emission, the nucleus itself undergoes a transformation, changing from one element to another. In the case of isotopes that β^+-decay, each decaying nucleus emits a positron and a neutrino, simultaneously reducing its atomic number by one unit; for those isotopes that β^--decay, each nucleus emits an electron and an antineutrino while increasing its atomic number by one. In both classes of decay, the energy released by the nuclear transformation is shared between the two emitted particles. Though the energy released by a particular nuclear transformation is always the same, the fraction of this energy carried away by the beta particle is different for each individual decaying nucleus. (The neutrino always carries away the remainder, thus conserving energy overall.) When observed collectively, the decaying nuclei of a simple radioactive source emit their beta particles with a continuous distribution of kinetic energies covering the range from zero up to the total nuclear decay energy available (**Fig. 1**).

Radioactive samples often contain several radioactive isotopes. Since each isotope has its own decay energy and beta-particle energy distribution, the energy spectrum of beta particles observed from such a sample would be the sum of a number of distributions like the one in Fig. 1, each with a different end-point energy. Indeed, many isotopes, especially those artificially produced with accelerators, can themselves beta-decay by additional paths that also release part of the energy in the form of gamma radiation. In general then, observed beta-particle spectra are more complex than that shown in Fig. 1, but they are always smooth distributions over a range of energy.

Under certain conditions, nuclear decay can result in the emission of monoenergetic electrons. These are called conversion electrons and are atomic

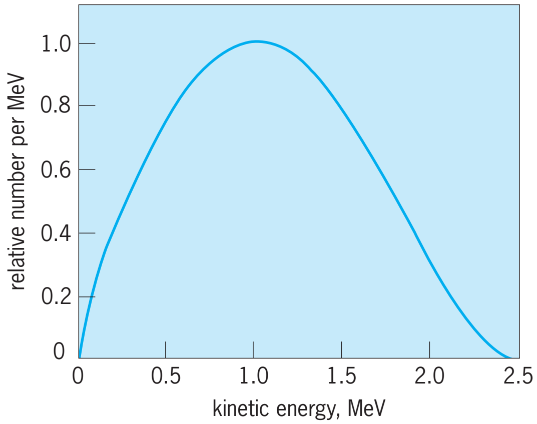


Fig. 1. Spectrum of energies of beta particles emitted from a radioactive source. The distribution shown corresponds to a light nucleus that releases a maximum kinetic energy of 2.5 MeV.

electrons ejected from their orbit by the energy released from a nucleus that does not change its atomic number. In the early 1900s, they were confused with the electrons emitted in beta decay, but their appearance is actually quite independent of the beta-decay process.

Interactions with matter. As a beta particle penetrates matter, it loses its energy in collisions with the constituent atoms. Two processes are involved. First, the beta particle can transfer a small fraction of its energy to the struck atom. Second, the beta particle is deflected from its original path by each collision and, since any change in the velocity of a charged particle leads to the emission of electromagnetic radiation, some of its energy is lost in the form of low-energy x-rays (called, in this context, bremsstrahlung). Though the energy lost by a beta particle in a single collision is very small, many collisions occur as the particle traverses matter, causing it to follow a zigzag path as it slows down. *See* BREMSSTRAHLUNG.

Because of their random zigzag paths, even monoenergetic electrons will not all travel through the same thickness of material before stopping. However, the thickness that is just sufficient to stop all the beta particles of a particular energy is called the range of those particles. For the continuous energy distribution normally associated with a source of beta particles, the effective range is the one that corresponds to the highest energy in the primary spectrum. That thickness of material stops all of the beta particles from the source. The range depends strongly on the electron energy and the density of the absorbing material (**Fig. 2**).

The slowing-down processes have the same effect on both β^- and β^+ particles. However, as antimatter, the positron (β^+) cannot exist for long in the presence of matter. It soon combines with an atomic electron, with which it annihilates, the masses of both particles being replaced by electromagnetic energy. Usually this annihilation occurs after the positron has come to rest and formed a positronium atom, a bound but short-lived positron-electron system. In that case, the electromagnetic energy that is emitted from the annihilation takes the form of two 511-keV gamma rays that are emitted in opposite directions to conserve momentum. *See* POSITRONIUM.

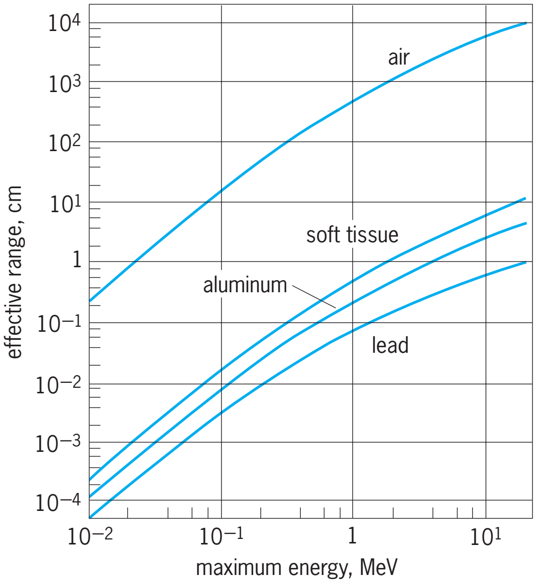


Fig. 2. Ranges of beta particles in representative substances, plotted as a function of their maximum energy. (*After L. Pages et al., Atomic Data, vol. 4, pp. 1–127, 1972*)

Detection. Beta particles are detected through their interaction with matter, and a wide variety of detection techniques are now available. One class of detectors employs gas as the detection medium. Ionization chambers, proportional counters, and Geiger-Müller counters are of this class. In these detectors, after entering through a thin window, the beta particles produce positive ions and free electrons as they collide with atoms of the gas in the process of their slowing down. An electric field applied across the volume of gas causes these ions and electrons to drift along the field lines, causing an ionization current that is then processed in external electronic devices. Depending on the application, the current can be integrated over time or, instead, the output can be recorded pulse by pulse to give information—energy and time of arrival, for example—on each individual beta particle. *See* GEIGER-MÜLLER COUNTER; IONIZATION CHAMBER; PARTICLE DETECTOR.

More precise energy information can be achieved with scintillation detectors. In certain substances, the ion-electron pairs produced by the passage of a charged particle result in the emission of a pulse of visible or near-ultraviolet light. If a clear plastic scintillator is used, it can be mounted on a photomultiplier tube, which converts the transmitted light into a measurable electrical current pulse whose amplitude is proportional to the energy deposited by the incident beta particle. *See* SCINTILLATION COUNTER.

Even better energy information comes from semiconductor detectors, which are effectively solid-state ionization chambers. When a beta particle enters the detector, it causes struck electrons to be raised into the conduction band, leaving holes behind in the valence band. The electrons and holes move under the influence of an imposed electric field, causing a pulse of current to flow. Such detectors are compact and easy to use, but have the disadvantage that they are relatively small and have a lower detection efficiency. They are useful mainly for low-energy beta particles. *See* JUNCTION DETECTOR.

If the ultimate in energy information is required, albeit at the further expense of efficiency, any one of these detectors can be combined with a magnetic spectrometer. Beta particles, like any charged particles, follow curved paths in a perpendicular magnetic field, their radius of curvature being proportional to the square of their energy. Their detected position on exiting the magnetic field can be precisely related to their energy. The best current measurement of the electron antineutrino mass comes from a spectrometer measurement of the tritium beta-decay spectrum. *See* NEUTRINO. John Hardy

Bibliography. K. Heyde, *From Nucleons to the Atomic Nucleus*, 1998; N. A. Jelley, *Fundamentals of Nuclear Physics*, 1990; W. S. C. Williams, *Nuclear and Particle Physics*, 1991.

Betatron

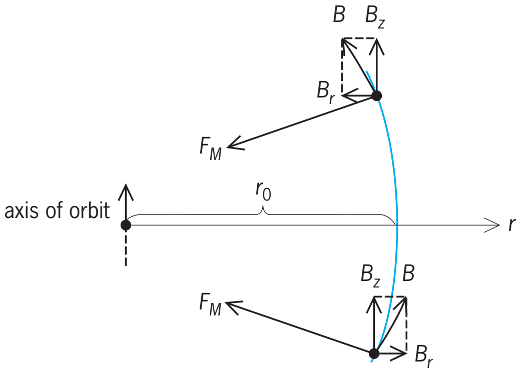
A device for accelerating charged particles in an orbit by means of the electric field E from a slowly changing magnetic flux Φ. The electric field is given by $E = -(^1/_{2\pi r_o})\, d\Phi/dt$ (in SI or mks units), where r_o is the orbit radius. The name was chosen because the method was first applied to electrons. In the usual betatron both the accelerating core flux and a guiding magnetic field rise with similar time dependence, with the result that the orbit is circular. However, the orbit can have a changing radius as acceleration progresses. For the long path (usually more than 100 km), variations of axial and radial magnetic field components provide focusing forces, while space charge and space current forces due to the particle beam itself also contribute to the resulting betatron oscillations about the equilibrium orbit. In many other instances of particle beams, the term betatron oscillation is used for the particle oscillations about a beam's path.

Although there had been a number of earlier attempts to make a betatron, the complete transient theory guided the construction of a successful 2.3×10^6 eV accelerator, producing x-rays equivalent to the gamma rays from 2 g of radium, at the University of Illinois in 1940.

It can be shown that the flux change within an orbit of fixed radius must be twice that which would exist if the magnetic field were uniform. The flux can be biased backward by a fixed amount ϕ_0 to avoid iron saturation, as it was in the University of Illinois 320-MeV betatron.

The beam must be focused in both the radial direction and the axial direction (the direction perpendicular to the plane of the orbit). For radial focusing, the centripetal force F_o required to hold a particle in a circular orbit of radius r must be supplied by the magnetic field B exerting a force F_M. At an equilibrium radius r_o, $F_C = F_M$, and if F_M decreases less rapidly than $1/r$ with increasing radius, there will be a net focusing force directed back toward r_o when $r \neq r_o$. This focusing force will cause the radius r to undergo a simple harmonic oscillation about r_o. If the axial magnetic field B_z falls off less rapidly than $1/r$, say $B_z \sim 1/r^n$, then it can be shown that this oscillation has a frequency $\omega_r = \sqrt{1-n}\ \Omega$, where Ω is the angular velocity in the particle's orbit. Thus $n < 1$ is necessary for a stable radial oscillation giving a frequency less than the frequency of rotation.

In addition, axial focusing results from requiring the magnetic field to decrease with radius so that the lines of force are curved, giving a component of F_M on a displaced particle back toward the orbital plane, as shown in the **illustration**. It can be shown that this focusing force gives rise to an axial oscillation with frequency $W_z = \Omega\sqrt{n}$. Thus $0 < n < 1$ is necessary for complete focusing. This is the so-called weak focusing conditions for betatron oscillation in accelerators.



Magnetic fields and their resulting forces in the axial plane of a betatron orbit, which result in axial focusing.

By dividing the focusing structure into sectors of large and alternating gradients, $+n \gg 1$, so-called strong focusing results. This allows a small volume for the beam path, or direct-current focusing magnets with a finite change in orbit radius as the particle gains great energy. Such betatrons have been constricted.

Injection into an orbital oscillation is possible because the restoring force is proportional to B^2, which rises relatively rapidly at injection time. This increasing focusing force causes the oscillation amplitude to decrease as $1/\sqrt{B}$ and enables the particles to escape subsequent collision with a properly designed injection structure.

Collective effects from self-fields of the beam have been found important and helpful in injecting. Circulating currents of about 3 A are contained in the numerous industrial and therapeutic betatrons, although the average currents are below 10^{-7} A. Such beams have been extracted by using a magnetic shunt channel where $n = {^3/_4}$ so that $\omega/\Omega = {^1/_2}$. This field bump excites a half-integral resonance which throws the particles into the channel. *See* PARTICLE ACCELERATOR.

Donald W. Kerst

Betel nut

The dried, ripe seed of the palm tree *Areca catechu* (Palmae), a native of Sri Lanka and Malaya, which attains a height of 30–50 ft (9–15 m). The nuts, slightly larger than a chestnut, when broken open have a faint odor and a somewhat acrid taste. Betel nuts are chewed by the natives together with the leaves of the betel pepper, *Piper betle* (Piperaceae), and lime. The mixture discolors and eventually destroys the teeth. Frequently, habitual chewers of the nut are toothless by the age of 25. The seeds contain a narcotic that produces some stimulation and a sense of well-being. *See* NARCOTIC; PIPERALES.

Perry D. Strausbaugh; Earl L. Core

Betelgeuse

A cool and highly luminous star, prominently located in the right shoulder of the constellation Orion and noticeably red in color. Betelgeuse, or α Orionis, is a supergiant star about 130 parsecs (430 light-years) from the Sun. Its spectral type of M2 indicates an effective temperature of approximately 3500 K (5800°F). This temperature would result in a low overall luminosity were it not for the enormous diameter of the star, about 1100 times that of the Sun. Replacing the Sun, Betelgeuse would fill the solar system to beyond the orbit of Mars. Betelgeuse is a supergiant star with approximately 150,000 times the Sun's luminosity. Betelgeuse varies in brightness with a period of 6.4 years. Its mean density is extremely low, less than one-millionth the density of water, and the low surface gravity of the rarefied outer layers results in a continual loss of matter, presently at the rate of 1 solar mass every 250,000 years. Observations at infrared and radio wavelengths show a complex system of dust and gas shells extending out as far as a thousand stellar radii. If this material were visible to the eye, Betelgeuse would appear as large as Mars through a telescope. These shells appear to originate from occasional outbursts in which material is ejected and eventually merges with the interstellar medium. *See* SPECTRAL TYPE; STELLAR EVOLUTION; SUPERGIANT STAR; VARIABLE STAR.

Because of its enormous size, Betelgeuse is one of the few stars resolvable with special techniques, such as speckle interferometry, at large optical telescopes, where a diameter of about 0.055 arc-second has been measured. These techniques have also produced the first images of the disk of a star and show asymmetric brightness variations over the surface that change with time and are possibly due to large-scale convection of material within the star's atmosphere. Betelgeuse has also been imaged by the Hubble Space Telescope at ultraviolet wavelengths showing the asymetric inner atmosphere of the star and a central hot spot. In a million years or so, the star's central core will collapse in a supernova explosion, leaving behind a neutron star or a white dwarf star remnant. *See* SPECKLE; STAR; SUPERNOVA.

Harold A. McAlister

Bibliography. A. Frankoi, D. Morrison, and S. C. Wolff, *Voyages Through the Universe*, 3d ed., Brooks Cole, 2004; R. L. Gilliland and A. K. Dupree, First image of the surface of a star with the Hubble Space Telescope, *Astrophys. J.*, 463:L29–L32, 1996; J. B. Kaler, *The Hundred Greatest Stars*, 2002; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millennium*, 3d ed., Brooks Cole, 2007.

Bias (electronics)

The establishment of an operating point on the transistor volt-ampere characteristics by means of direct voltages and currents.

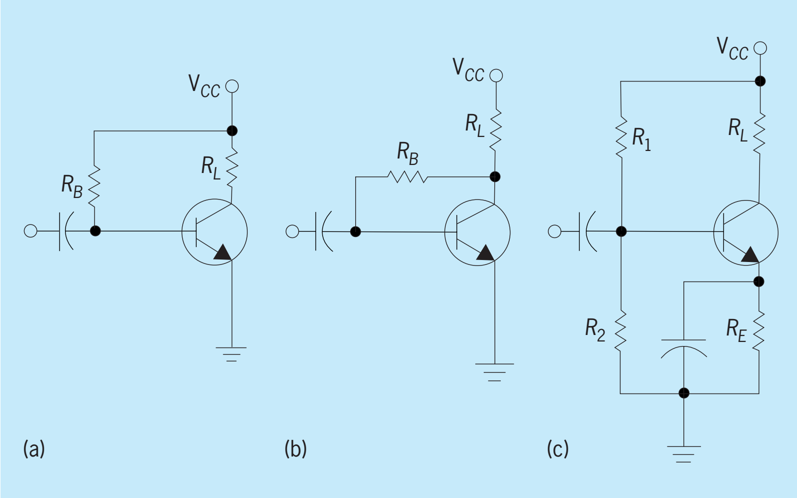


Fig. 1. Translator circuits. (*a*) Fixed-bias. (*b*) Collector-to-base bias. (*c*) Self-bias.

Biasing biopolar transistors. Since the transistor is a three-terminal device, any one of the three terminals may be used as a common terminal to both input and output. In most transistor circuits the emitter is used as the common terminal, and this common emitter, or grounded emitter, is indicated in **Fig. 1***a*. If the transistor is to used as a linear device, such as an audio amplifier, it must be biased to operate in the active region. In this region the collector is biased in the reverse direction and the emitter in the forward direction. The area in the common-emitter transistor characteristics to the right of the ordinate $V_{CE} = 0$ and above $I_C = 0$ is the active region. Two more biasing regions are of special interest for those cases in which the transistor is intended to operate as a switch. These are the saturation and cutoff regions. The saturation region may be defined as the region where the collector current is independent of base current for given values of V_{CC} and R_L. Thus, the onset of saturation can be considered to take place at the knee of the common-emitter transistor curves. *See* AMPLIFIER.

In saturation the transistor current I_C is nominally V_{CC}/R_L. Since R_L is small, it may be necessary to keep V_{CC} correspondingly small in order to stay within the limitations imposed by the transistor on maximum-current and collector-power dissipation. In the cutoff region it is required that the emitter current I_E be zero, and to accomplish this it is necessary to reverse-bias the emitter junction so that the collector current is approximately equal to the reverse saturation current I_{CO}. A reverse-biasing voltage of the order of 0.1 V across the emitter junction will ordinarily be adequate to cut off either a germanium or silicon transistor.

The particular method to be used in establishing an operating point on the transistor characteristics depends on whether the transistor is to operate in the active, saturation or cutoff regions; on the application under consideration; on the thermal stability of the circuit; and on other factors.

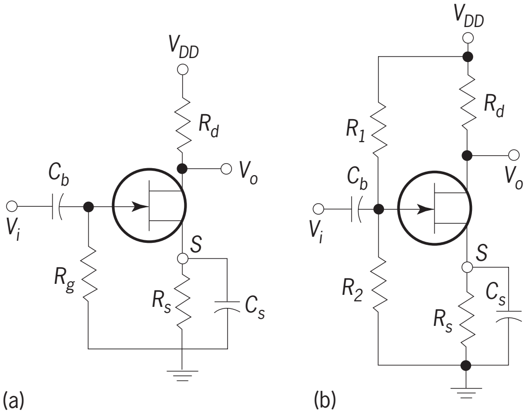


Fig. 2. Biasing circuits for FETs and MOSFETs. (*a*) Source self-bias circuit for an FET and a depletion-type MOSFET. (*b*) Biasing circuit for an FET and an enhancement-type MOSFET. (*After J. Millman Microelectronics McGraw-Hill, 1979*)

Fixed-bias circuit. The operating point for the circuit of Fig. 1*a* can be established by noting that the required current I_B is constant, independent of the quiescent collector current I_C, which is why this circuit is called the fixed-bias circuit. Transistor biasing circuits are frequently compared in terms of the value of the stability factor $S = \partial I_C/\partial I_{CO}$, which is the rate of change of collector current with respect to reverse saturation current. The smaller the value of S, the less likely the circuit will exhibit thermal runaway. S, as defined here, cannot be smaller than unity. Other stability factors are defined in terms of dc current gain h_{FE} as $\partial I_C/\partial h_{FE}$, and in terms of base-to-emitter voltage as $\partial I_C/\partial V_{BE}$. However, bias circuits with small values of S will also perform satisfactorily for transistors that have large variations of h_{FE} and V_{BE}. For the fixed-bias circuit it can be shown that $S = h_{FE} + 1$, and if $h_{FE} = 50$, then $S = 51$. Such a large value of S makes thermal runaway a definite possibility with this circuit.

Collector-to-base bias. An improvement in stability is obtained if the resistor R_B in Fig. 1*a* is returned to the collector junction rather than to the battery terminal. Such a connection is shown in Fig. 1*b*. In this bias circuit, if I_C tends to increase (either because of a rise in temperature or because the transistor has been replaced by another), then V_{CE} decreases. Hence I_B also decreases and, as a consequence of this lowered bias current, the collector current is not allowed to increase as much as it would if fixed bias were used. The stability factor S is shown in Eq. (1). This value

$$S = \frac{h_{FE} + 1}{1 + h_{FE}R_L/(R_L + R_B)} \qquad (1)$$

is smaller than $h_{FE} + 1$, which is the value obtained for the fixed-bias case.

Self-bias. If the load resistance R_L is very small, as in a transformer-coupled circuit, then the previous expression for S shows that there would be no improvement in the stabilization in the collector-to-base bias circuit over the fixed-bias circuit. A circuit that can be used even if there is zero dc resistance in series with the collector terminal is the self-biasing configuration of Fig. 1*c*. The current in the resistance R_E in the emitter lead causes a voltage drop which is in the direction to reverse-bias the emitter junction. Since this junction must be forward-biased (for active region bias), the bleeder R_1-R_2 has been added to the circuit.

If I_C tends to increase, the current in R_E increases. As a consequence of the increase in voltage drop across R_E, the base current is decreased. Hence I_C will increase less than it would have had there been no self-biasing resistor R_E. The stabilization factor for the self-bias circuit is shown by Eq. (2), where

$$S = (1 + h_{FE})\frac{1 + R_B/R_E}{1 + h_{FE} + R_B/R_E} \qquad (2)$$

$R_B = R_1R_2/(R_1 + R_2)$. The smaller the value of R_B, the better the stabilization. Even if R_B approaches zero, the value of S cannot be reduced below unity.

In order to avoid the loss of signal gain because of the degeneration caused by R_E, this resistor is often bypassed by a very large capacitance, so that its reactance at the frequencies under consideration is very small.

Biasing the FET and MOSFET. The selection of an appropriate operating point (I_D, V_{GS}, V_{DS}) for a field-effect transistor (FET) amplifier stage is determined by considerations similar to those given to transistors, as discussed previously. These considerations are output-voltage swing, distortion, power dissipation, voltage gain, and drift of drain current. In most cases it is not possible to satisfy all desired specifications simultaneously.

Source self-bias. The configuration shown in **Fig. 2*a*** can be used to bias junction FET devices or depletion-mode metal-oxide-semiconductor FETs (MOSFETs). For a specified drain current I_D, the corresponding gate-to-source voltage V_{GS} can be obtained analytically or by using the plotted drain or transfer characteristics. Since the gate current (and, hence, the voltage drop across R_g) is negligible, the source resistance R_s can be found as the ratio of V_{GS} to the desired I_D.

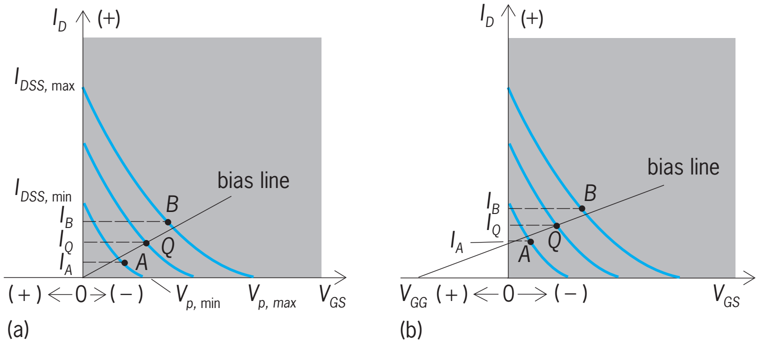


Fig. 3. Maximum and minimum transfer curves for biasing an *n*-channel FET. (*a*) Bias line passes through origin. (*b*) Bias line does not pass through origin. (*After J. Millman, Microelectronics, McGraw-Hill, 1979*)

Biasing against device variation. FET manufacturers usually supply information on the maximum and minimum values of I_{DSS} and V_P at room temperature, where I_{DSS} is the value of the drain current I_{DS} with $V_{GS} = 0$, and V_P is the pinch-off voltage. They also supply data to correct these quantities for temperature variations. The transfer characteristics for a given type of *n*-channel FET may appear as in **Fig. 3*a*** where the top and bottom curves are for extreme values of temperature and device variation. Assume that it is necessary to bias the device at a drain current which will not drift outside $I_D = I_A$ and $I_D = I_B$. Then the bias line $V_{GS} = -I_D R_s$ must intersect the transfer characteristics between the points *A* and *B*, as indicated in Fig. 3*a*. The slope of the bias line is determined by the source resistance R_s. For any transfer characteristic between the two extremes indicated, the current I_Q is such that $I_A < I_Q < I_B$, as desired.

Consider the situation indicated in Fig. 3*b*, where a line drawn to pass between points *A* and *B* does not pass through the origin. This bias line satisfies Eq. (3).

$$V_{GS} = V_{GG} - I_D R_s \qquad (3)$$

Such a bias relationship may be obtained by adding a fixed bias to the gate in addition to the source self-bias, as indicated in Fig. 2*b*. This circuit requires only one power supply and satisfies Eqs. (4), where it

$$V_{GG} = \frac{R_2 V_{DD}}{R_1 + R_2} \qquad R_g = \frac{R_1 R_2}{R_1 + R_2} \qquad (4)$$

is assumed that the gate current is negligible. It is also possible for V_{GG} to fall in the reverse-biased region, so that the line in Fig. 3*b* intersects the axis of abscissa to the right of the origin. Under these circumstances, two separate supply voltages must be used.

Biasing the enhancement MOSFET. The self-bias technique of Fig. 2*a* cannot be used to establish an operating point for the enhancement-type MOSFET because the voltage drop across R_s is in a direction to reverse-bias the gate, and a forward gate bias is required. The circuit of **Fig. 4*a*** can be used, and for this case $V_{GS} = V_{DS}$, since no current flows through R_f. If, for reasons of linearity in device operation or maximum output voltage, it is desired that $V_{GS} \neq V_{DS}$, then the circuit of Fig. 4*b* is suitable. In this case, $V_{GS} = [R_1/(R_1 + R_f)] \cdot V_{DS}$. Both circuits discussed here offer the advantages of dc stabilization through the feedback introduced with R_f.

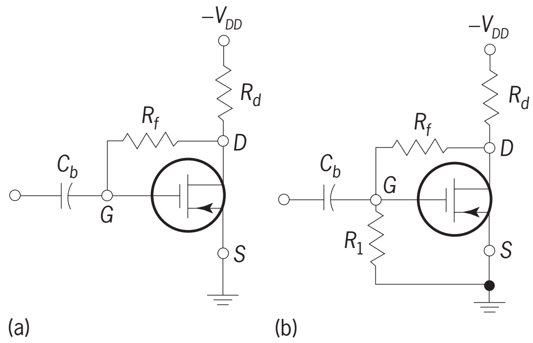


Fig. 4. Drain-to-gate bias circuits for enhancement-type MOSFET. (*a*) Circuit with $V_{GS} = V_{DS}$. (*b*) Improved version, with $V_{GS} \neq V_{DS}$. (*After J. Millman, Microelectronics, McGraw-Hill, 1979*)

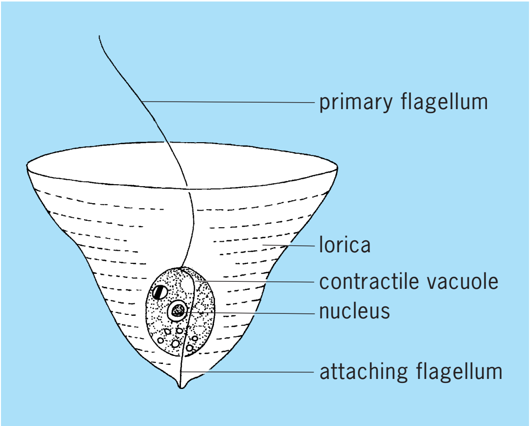
The circuit of Fig. 2*b* is often used with the enhancement MOSFET. The dc stability introduced in Fig. 4 through the feedback resistor R_f is then missing, and is replaced by the dc feedback through R_s. *See* TRANSISTOR. Christos C. Halkias

Bibliography. P. Horowitz and W. Hill, *The Art of Electronics*, 2d ed., 1989; J. Millman, *Microelectronics*, 2d ed., 1987; J. Singh, *Semiconductor Devices: An Introduction*, 1994.

Bicosoecida

An order of Zoomastigophorea (Protozoa). They are colorless, free-living cells, each with two flagella, one of which is used for attaching the organism to its exoskeleton (lorica). Although the attachment is normally from the front end, the flagellum emerges alongside the primary or vibrating one (see **illus.**). The anterior end in many species (*Codomonas annulata, Poteriodendron petiolatum*, and *Stephanocodon stellatum*) is ameboid, or at times appears to be formed into a lip which can turn in and engulf a bacterium. *Stenocodon* has the same sort of lip but neither of the two flagella is used for attachment. These Bicosoecida are common in fresh water, often seen attached to desmids or other algae. *Bicosocca mediterranea* is common at times in salt water, where assemblages of it are found on diatoms. The order has very few genera and species, and individuals easily escape observation because of small size and transparency. *See* CILIA AND FLAGELLA; PROTOZOA; ZOOMASTIGOPHOREA.

James B. Lackey



Bicosoecid, *Codomonas annulata.*

Big bang theory

The theory that the universe began in a state of extremely high density and has been expanding since some particular instant that marked the origin of the universe. The big bang is the generally accepted cosmological theory; the incorporation of developments in elementary particle theory has led to the inflationary universe version. The predictions of the inflationary universe and older big bang theories are the same after the first 10^{-35} s. *See* INFLATIONARY UNIVERSE COSMOLOGY.

Two observations are at the base of observational big bang cosmology. First, the universe is expanding uniformly, with objects at greater distances receding at a greater velocity. Second, the Earth is bathed in the cosmic background radiation, an isotropic glow of radiation that has the characteristics expected from the remnant of a hot primeval fireball. Since the last years of the twentieth century, a third observation has come to the fore: the expansion of the universe is accelerating.

Cosmological theory in general and the big bang theory in particular are based on the theory of gravitation advanced by A. Einstein in 1916 and known as the general theory of relativity. Though the predictions of this theory have little effect in the limited sphere of the Earth, they dominate on as large a scale as the universe, and have been well tested in such sources as the binary pulsar. *See* RELATIVITY.

Expansion of the universe. In the 1920s it became clear that the "spiral nebulae," clouds of gas with arms spiraling outward from a core, were galaxies on the scale of the Milky Way Galaxy. This was established in 1925, when observations of variable stars in several galaxies by E. Hubble enabled the distance to these galaxies to be determined with some accuracy.

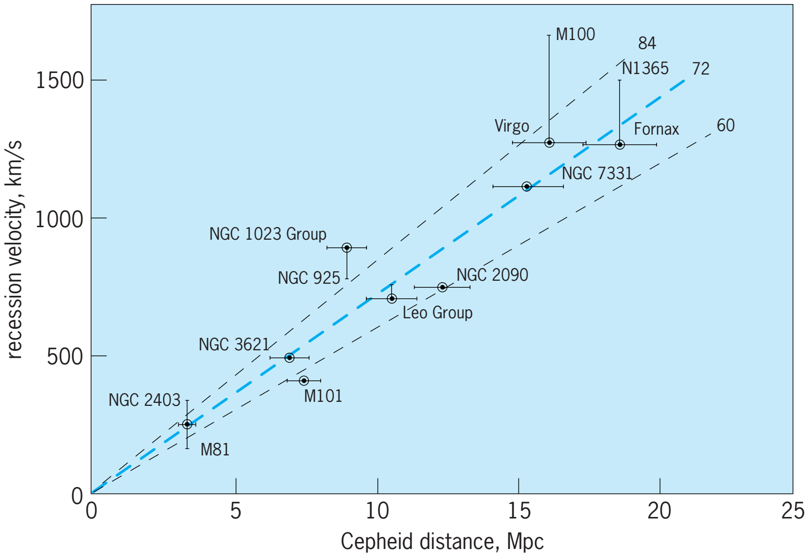


Fig. 1. Hubble diagram. Numbers next to the straight lines passing through the origin indicate corresponding values of Hubble's constant in km · s^{-1} · Mpc^{-1}. Data on this diagram give $H_0 = 72$ km · s^{-1} · Mpc^{-1}, with an uncertainty of ±4 km · s^{-1} · Mpc^{-1} and a scatter of ±17 k · s^{-1} · Mpc^{-1}. (*W. L. Freedman, B. Madore, and R. C. Kennicutt; NASA*)

Starting in 1912, the spectra of many of these spiral nebulae were found to have large redshifts. According to the Doppler effect, these large redshifts correspond to large velocities of recession from the Earth. Estimates of distances to these objects by Hubble and colleagues established a direct relation between the distance to a galaxy and its velocity of recession. It was soon interpreted as the result of an expanding universe. *See* DOPPLER EFFECT; REDSHIFT.

The relation known as Hubble's law is $v = H_0 d$, where v is the velocity of recession, d is the distance to the galaxy, and H_0 is a constant known as Hubble's constant. Determining Hubble's constant requires the independent measurement of the distances to galaxies; the redshift can easily be measured on spectra.

In the Hubble diagram (**Fig. 1**), the horizontal axis is the distance as derived from observations of the periods of Cepheid variable stars in distant galaxies and the vertical axis is the velocity of recession, calculated from the redshift $z = \Delta\lambda/\lambda$, where $\Delta\lambda$ is the shift in wavelength in radiation of wavelength λ measured on the spectra of those galaxies. The Hubble law is expressed by the fact that the data points in the diagram lie near a straight line. *See* CEPHEIDS.

If velocity is expressed in kilometers per second, and distance is expressed in megaparsecs (where 1 parsec is the distance from which the radius of the Earth's orbit would subtend 1 second of arc, and is equivalent to 3.26 light-years or 3.09×10^{13} km or 1.92×10^{13} mi), then Hubble's constant H_0 is given in km · s^{-1} · Mpc^{-1}. The determination of Hubble's constant was the aim of one of the three Key Projects of the *Hubble Space Telescope*, especially through studies of Cepheid variable stars in galaxies sufficiently distant to be beyond internal gravity effects in the Local Group, the cluster of galaxies to which the Milky Way belongs. The result is 72 km · s^{-1} · Mpc^{-1}. Tying the *Hubble Space Telescope* results to distances derived from supernovae extends Hubble's constant to still greater distances. Other methods involving planetary nebulae, the spread of rotational velocities of galaxies, the statistical fluctuations of each pixel of galaxy images, and supernovae, for example, also gave values close to 75 km · s^{-1} · Mpc^{-1}. The result from the National Aeronautics and Space Administration's (NASA) *Wilkinson Microwave Anisotropy Probe (WMAP)* closely confirms the Hubble Key Project value. *See* GALAXY, EXTERNAL; HUBBLE CONSTANT; LOCAL GROUP; SUPERNOVA.

Reversing the expansion rate, tracing the expansion of the universe back in time shows that the universe would have been compressed to infinite density approximately 8–16 × 10^9 years ago (for $H_0 = 72$ km · s^{-1} · Mpc^{-1}). The *WMAP* pinned this value down more precisely, to 13.7 × 10^9 years ago. In the big bang theory, the universe began at that time as a so-called big bang began the expansion. The big bang was the origin of space and time.

In 1917, Einstein found a solution to his own set of equations from his general theory of relativity that predicted the nature of the universe. His universe, though, was unstable: it could only be expanding or contracting. This seemed unsatisfactory at the time, for the expansion had not yet been discovered, so Einstein arbitrarily introduced a special term—the cosmological constant—into his equations to make the universe static. The need for the cosmological constant seemed to disappear with Hubble's discovery of the expansion, though the cosmological constant has subsequently reappeared in some models. *See* COSMOLOGICAL CONSTANT.

Further solutions to Einstein's equations, worked out in the 1920s, are at the basis of the cosmological models that are now generally accepted. These solutions indicate that the original "cosmic egg" from which the universe was expanding was hot and dense. This is the origin of the current view that the universe was indeed very hot in its early stages.

Early universe. Modern theoretical work has been able to trace the universe back to the first instants in time. In the big bang theory and in related theories that also propose a hot, dense early universe, the universe may have been filled in the earliest instants with exotic elementary particles of the types now being studied by physicists with large accelerators. Individual quarks may also have been present. By 1 microsecond after the universe's origin, the exotic particles and the quarks had been incorporated in other fundamental particles. *See* ELEMENTARY PARTICLE; QUARKS.

Work in the early 1980s incorporated the effect of elementary particles in cosmological models. The research seems to indicate that the universe underwent a period of extremely rapid expansion in which it inflated by a factor of billions in a very short time. This inflationary universe model provides an explanation for why the universe is so homogeneous: Before the expansion, regions that now seem too separated to have been in contact were close enough to interact. After the inflationary stage, the universe was in a hot stage and was still dense; the models match the big bang models thereafter.

In the inflationary universe models, the universe need not have arisen from a single big bang. Rather, matter could have appeared as fluctuations in the vacuum.

It is not definitely known why there is an apparent excess of matter over antimatter, though attempts in elementary particle physics to unify the electromagnetic, the weak, and the strong forces show promise in explaining the origin of the matter–antimatter asymmetry. The asymmetry seems to have arisen before the first millisecond. The asymmetry in the decay of certain mesons may provide a clue to resolving this question. *See* ANTIMATTER; FUNDAMENTAL INTERACTIONS.

By 5 s after the origin of the universe, the temperature had cooled to 10^9 K (2×10^9 °F), and only electrons, positrons, neutrinos, antineutrinos, and photons were important. A few protons and neutrons were mixed in, and they grew relatively more important as the temperature continued to drop. The universe was so dense that photons traveled only a short way before being reabsorbed. By the time 1 min had gone by, nuclei of the light elements had started to form.

After about 400,000 years, when the universe cooled to 3000 K (5000°F) and the density dropped sufficiently, the protons and electrons suddenly combined to make hydrogen atoms, a process called recombination. Since hydrogen's spectrum absorbs preferentially at the wavelengths of sets of spectral lines rather than continuously across the spectrum, and since there were no longer free electrons to interact with photons, the universe became transparent at that instant. The average path traveled by a photon—its mean free path—became very large. The blackbody radiation of the gas at the time of recombination was thus released and has been traveling through space ever since. As the universe expands, the spectrum of this radiation retains its blackbody shape though its characteristic temperature drops. *See* BLACKBODY; HEAT RADIATION.

Background radiation. Between 1963 and 1965, observations with a well-calibrated horn antenna revealed an isotropic source of radio noise whose strength was independent of time of day and of season, and whose intensity at the observing wavelength of 7 cm (2.8 in.) was equivalent to that which would be emitted by a blackbody—an idealized radiating substance—at a temperature of about 3 K (−454 °F). Prior calculations that indicated that the remnant radiation from the big bang might be detectable led to the interpretation of the newly discovered radiation in terms of fossil radiation from the big bang. In 1975, final confirmation of the blackbody nature of the radiation was provided by measurements of the spectrum of the background radiation in a frequency range that included the peak of the blackbody curve and extended into the infrared.

The observations of the spectrum from the *Cosmic Background Explorer* (*COBE*) spacecraft during 1989–1993 were in complete agreement with a blackbody curve at 2.728 K (−454.7°F). The *WMAP*'s value, released in 2003, is in close agreement, and updates from *WMAP* will be released during and after its four-year lifespan. The radiation is very isotropic, once the dipole anisotropy caused by the Earth's motion is subtracted. Observations from another experiment on *COBE* discovered temperature fluctuations at the level of 5 parts per million. *WMAP* observed them in still finer detail. These fluctuations, largely statistical in nature, are revealing the seeds from which galaxies formed. Inflationary big bang models seem best able to explain how widely separated parts of the universe can be as similar as the *COBE* and *WMAP* observations have revealed them to be. Confirmations of these ripples in space have come from cosmic background telescopes on balloons and in Canada and Antarctica. Spacecraft missions with improved resolution and sensitivity, including the European Space Agency's *Planck* mission (2007), are planned. *See* COSMIC BACKGROUND RADIATION.

Nucleosynthesis. As the early universe cooled, the temperatures became sufficiently low for element

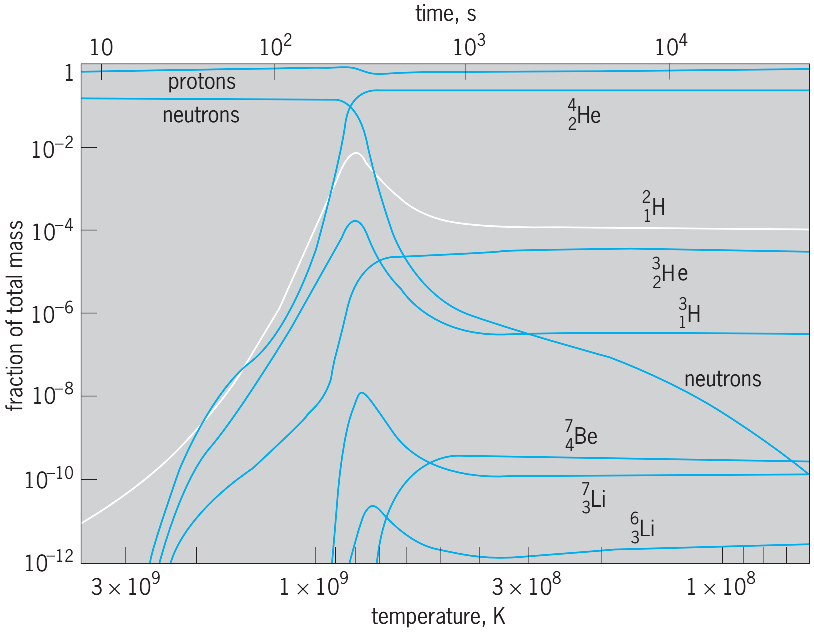


Fig. 2. **Relative abundance of isotopes of the light elements in the first few minutes after the big bang according to the model of R. V. Wagoner, for a certain assumed density at some early time. (*After J. M. Pasachoff, Astronomy: From the Earth to the Universe, 6th ed., Brooks/Cole, 2002*)**

formation to begin. By about 100 s, deuterium (comprising one proton plus one neutron) formed. When joined by another neutron to form tritium, the amalgam soon decayed to form an isotope of helium. Ordinary helium, with still another neutron, also resulted.

The relative abundances of isotopes of the light elements in the first few minutes have been calculated (**Fig. 2**). The calculations depend on a certain assumed density at some early time. They show that within minutes the temperature drops to 10^9 K (2 × 10^9 °F), too low for most nuclear reactions to continue. Most models give a resulting abundance of about 25% of the mass in the form of helium, regardless of the density of the universe. The helium abundance is hard to determine observationally. Current results are in rough agreement with the theoretical value.

The abundances of others of the light elements are more sensitive to parameters of matter in the early universe. Such abundances can be calculated as a function of the present-day density of matter (**Fig. 3**). From knowledge of the approximate rate of expansion of the universe, the density of matter in the early universe can be deduced from the current density, and abundances can then be calculated. In particular, the deuterium abundance is especially sensitive to the cosmic density at the time of deuterium formation, because the rate at which deuterium is "cooked" into tritium increases rapidly with increasing density.

Big bang nucleosynthesis, although at first thought to be a method of forming all the elements, foundered for the heavy elements at mass numbers 5 and 8. Isotopes of these mass numbers are too unstable to form heavier elements quickly enough. The gap is bridged only in stars, through processes worked out in 1957. Thus the lightest elements were formed as a direct result of the big bang (Fig. 3), while the heavier elements as well as additional quantities of most of the lighter elements were formed later in stars or supernovae. Measurements of the relative abundances of the light isotopes and elements have been used to set a limit of three on the number of neutrino types and thus on the number of quark-lepton families that exist. *See* LEPTON; NEUTRINO; NUCLEOSYNTHESIS.

Open versus closed universe. The two extreme possibilities for the future of the universe are that the universe will continue to expand forever, or that it will cease its expansion and begin to contract. It can be shown that the case where the universe will expand forever, in the absence of dark energy or other accelerating mechanisms, corresponds to an infinite universe. The term applied is the open universe. The case where the universe will begin to contract corresponds to a finite universe. The term applied is the closed universe. The inflationary universe scenario has the universe on the boundary between open and closed, as signified by the parameter Ω taking the value of 1. Such a universe would expand forever but at an ever-decreasing rate in the absence of accelerating mechanisms. However, according to more recent conclusions, discussed below, the expansion of the universe is now accelerating, having passed a transition point where dark energy has begun to dominate.

Deceleration parameter. One basic way to test whether the universe is open or closed is to determine the rate at which the expansion of the universe is slowing, that is, the rate at which it is deviating from Hubble's law. The slowing is measured with a term

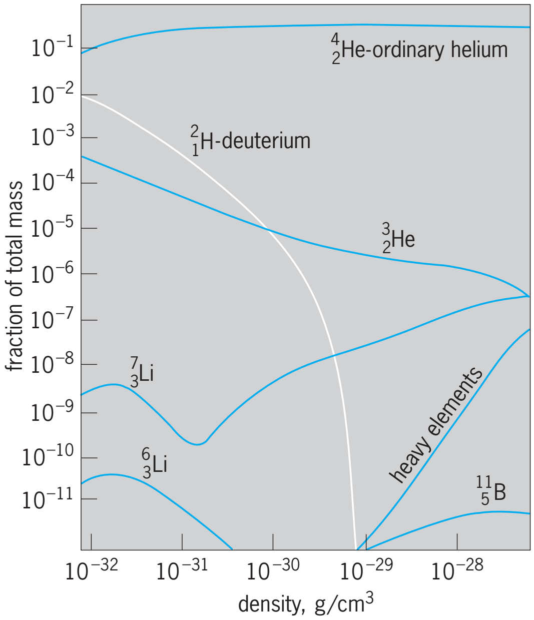


Fig. 3. **Relative abundances of isotopes and elements as a function of present-day density of matter, according to the model of R. V. Wagoner. (*After J. M. Pasachoff, Astronomy: From the Earth to the Universe, 6th ed., Brooks/Cole, 2002*)**

called q_0, the deceleration parameter. The dividing line is marked by $q_0 = {}^1/_2$, with smaller values corresponding to an open universe and larger values corresponding to a closed universe. The value $q_0 = {}^1/_2$ corresponds to $\Omega = 1$, where Ω is the ratio of the total density of matter in the universe to the critical density that divides open from closed universes.

The most obvious way to measure q_0 is by looking at the most distant galaxies or clusters of galaxies, measuring their distances independently from Hubble's law, and plotting Hubble's law in a search for deviations. But all such observations are subject to the considerable uncertainty introduced by the fact that observations deep into space also see far back into time. The galaxies then were surely different than they are now, so evolutionary effects must be taken into account. It is not even known for certain whether galaxies were brighter or dimmer in the distant past, and thus the method may reveal more about the evolution of galaxies than about cosmology.

Accelerating universe. The deceleration parameter is being studied from very distant supernovae, which are discovered in quantity through automated searches. Astonishingly, evidence seems to show that the expansion of the universe is accelerating, the opposite of what was expected. These data must be scrutinized for systematic errors before they can be definitely accepted, but enough independent confirmation has appeared (as of 2005) from different kinds of observations that doubt of this astonishing result is vanishing.

Observations of extremely distant supernovae have given astronomers views far enough back in time to see the era before the acceleration started. In that era, the universe's expansion was slowing. These discoveries gave confidence to astronomers that the conclusion that the universe is accelerating is correct. Results in 2004 from observations by the *Chandra X-ray Observatory* of extremely distant clusters of galaxies have confirmed that the universe's expansion is accelerating.

The results from *WMAP,* endorsing earlier results from studying ripples in the cosmic background radiation, have shown that the universe's content is exactly at the critical value between open and closed. At the same time, they show (endorsing nucleosynthesis observations summarized below) that the amount of ordinary matter (composed of atomic particles called baryons, like protons and neutrons) is only 4% of this critical value. Other studies, such as those of the gravitational attractions of objects in clusters of galaxies, show the existence of dark matter—with gravity but without emitting electromagnetic radiation—accounting for another approximately 30% of the critical value. The remaining two-thirds of the content of the universe remains unknown, and has been given the name "dark energy." Scientists have no definite idea of what the dark energy is composed, though Einstein's old "cosmological constant"—a seemingly arbitrary term in his equations—is one of the possibilities being considered. This dark energy now dominates the universe, causing the acceleration of the expansion. When the cosmological constant Λ is taken into account, $q_0 = {}^1/_2\,\Omega_m - \Omega_\Lambda$ (where Ω_m is associated with the matter in the universe and Ω_Λ is associated with the cosmological constant), so it is now negative. *See* ACCELERATING UNIVERSE; BARYON; DARK ENERGY; DARK MATTER.

Great Attractor. From studies of elliptical and spiral galaxies, a bulk streaming motion has been found toward a region called the Great Attractor. The region would be a concentration in the direction of the constellation Centaurus, 20 times more populous than the Virgo Cluster. Its distortion of the Hubble flow may have led to misevaluation of the Hubble constant. An even greater attractor farther out has also been reported. The presence of the Great Attractor and related regions of space mean that the measurements of velocities we made in local areas of the universe can be affected by the Great Attractor's mass, where it gave misleading results about the Hubble flow before the *WMAP* and supernova determinations of Hubble's constant. *See* VIRGO CLUSTER.

Cosmic deuterium abundance. Because deuterium is a sensitive indicator of the cosmic density at the time of its formation soon after the big bang, and because deuterium can only be destroyed and not formed in stars, the study of the cosmic deuterium abundance is one of the best methods for determining the future of the universe. Basically, it involves assessing whether there is enough gravity in the universe to halt the expansion.

Counts can be made of all the stars, galaxies, quasars, interstellar matter, and so forth, and the sum gives a density much too low to close the universe. But this method does not assess the amount of invisible matter, which could be in the form of intergalactic gas, black holes, and so forth. Indeed, studies of the motions of galaxies in clusters of galaxies often indicate that much more mass is present inside galaxy clusters than is visible. The amount of this missing mass (actually the mass is present and it is the light that is missing) may be 50 times the amount of ordinary mass (known as baryons). As previously noted, it is known as dark matter.

Assessing the density of the universe through studies of the deuterium abundance is independent of whether the ordinary matter is visible or invisible (detectable, for example, from its radio or x-ray emissions). Although deuterium is present even on Earth as a trace isotope, with an abundance 1/6600 that of normal hydrogen in seawater, it is difficult to detect in interstellar space. However, since 1972 a number of determinations of interstellar deuterium abundance have been carried out, through radio observations and studies in the ultraviolet made with telescopes in space, the most precise with the *Hubble Space Telescope*. Determinations of deuterium in the atmospheres of the planets have also been carried out. Though there are some discrepancies remaining between abundances determined in different locations and in different fashions, the deuterium observations seem to indicate clearly that the density of the universe is very low and hence that the universe is open as far as baryons are concerned. This finding is supported by observations with one of the 10-m (400-in.) Keck telescopes of deuterium in a distant quasar, redshifted into the visible part of the spectrum. This result that the baryon density is low, only about 4% of the critical value that divides open and closed universes, matches the conclusions from *WMAP's* observations of the ripples in the cosmic background radiation. *See* DEUTERIUM.

Hot and cold dark matter. Observations of a diffuse background of x-rays had led to the suspicion that a lot of hot material, previously undiscovered, may have been present between the galaxies. However, subsequent observations revealed that at least some of the x-ray background came from faint quasars, which appeared on long exposures of even fields that had been thought to be blank. Continued mapping and pointed observations have detected extensive halos around certain galaxies and even around some sparse clusters of galaxies. These observations indicated the presence of extensive dark matter.

Since neutrinos are known to be plentiful in the universe, if they each had enough mass they could dominate the universe. Since there are about 100 neutrinos in each cubic centimeter (1500 in each cubic inch) of the universe, much mass would be in that form if neutrinos had even a small rest mass. Indeed, observations of neutrinos with the Super-Kamiokande detector show that neutrinos change in type, which can occur only if they have mass. But the mass itself is not determined from these observations. Observations of the neutrinos from Supernova 1987A placed a sensitive limit on the amount of mass that neutrinos could have, since the arrival times of massive neutrinos would have spread out. Since neutrinos move so fast, they are known as hot dark matter. But observations from *WMAP* showed that the total mass of neutrinos does not dominate the universe.

Since hot dark matter would not explain satisfactorily the quickness with which matter clumped to form galaxies, most models include at least some so-called cold dark matter. The cold dark matter could be in the form of elementary particles of types yet unknown, such as axions or weakly interacting massive particles (WIMPs). Or it could be in the form of dim stars, brown dwarfs, or other macroscopic bodies known collectively as massive astrophysical compact halo objects (MACHOs). Supercomputer calculations involving millions of points interacting in a three-dimensional grid have computed the evolution of the universe for models with different values of Ω and for different combinations of cold and hot dark matter. *See* BROWN DWARF; WEAKLY INTERACTING MASSIVE PARTICLE (WIMP).

Inflationary scenarios. Although several lines of evidence, including studies of differential velocities of nearby galaxies caused by density perturbations and of the cosmic abundance of deuterium, indicate that the universe is open, it was not clear why the universe was so close to the dividing line between being open or closed that the subject was much in doubt. The inflationary universe model provides a natural explanation for the universe being on this dividing line. After expansion slows down at the close of the inflationary stage (thus causing a phase change, much like supercooled liquid water turning abruptly into ice when perturbed, a more extreme example of a phase change than the household boiling of water into steam), the universe necessarily approaches this line. Thus whether the universe was open or closed was the wrong question to ask. Further work on inflationary scenarios continues to assess certain problems, such as the inflationary model's predictions of the density fluctuations that lead to the coalescence of galaxies, now measured by the *WMAP* spacecraft. Though the simplest inflationary scenarios have been ruled out, more complex inflationary theories are viable and are widely accepted by cosmologists. *See* COSMOLOGY; UNIVERSE.

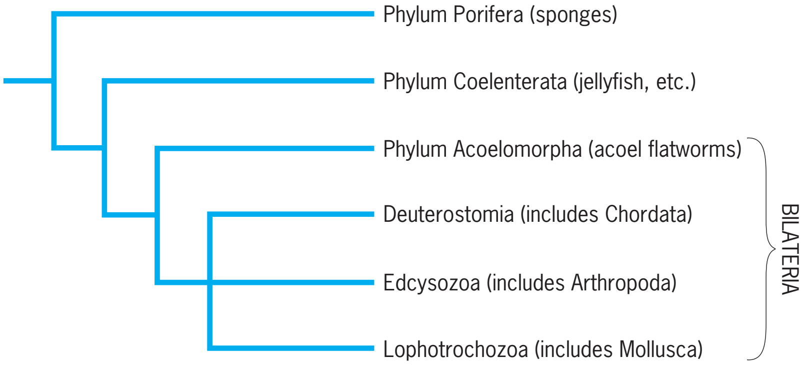
Jay M. Pasachoff

Bibliography. T. Ferris, *The Whole Shebang: A State of the Universe(s) Report*, Simon & Schuster, 1997; S. W. Hawking, *A Brief History of Time: The Updated and Expanded Tenth Anniversary Edition,* Bantam, 1998; J. Silk, *A Short History of the Universe*, Freeman, 1997; A. H. Guth and A. P. Lightman, *The Inflationary Universe: The Quest for a New Theory of Cosmic Origins*, Perseus, 1998; R. Kirshner, *The Extravagant Universe*, Princeton University Press, 2002; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millennium*, 2d ed., Brooks/Cole, 2004, 3d ed., 2007; M. Rees, *Our Cosmic Habitat*, Princeton University Press, 2001; S. Weinberg, *The First Three Minutes*, 2d ed., BasicBooks, 1994.

Bilateria

A major division of the animal kingdom comprising all forms with a bilaterally symmetrical body plan (that is, their left and right sides are mirror images, or nearly so) and organ systems that develop from three tissue layers. The animal kingdom, Metazoa, is subdivided into about 30 major animal groups, or phyla, each constructed on a different architectural plan. Most of these phyla are grouped as the Bilateria, comprising a major early branch on the family tree of animals that may have originated nearly 600 million years ago. All bilaterians possess some features that are not perfectly symmetrical (for example, the human heart and liver), but all have descended from an essentially bilateral ancestor. The nonbilaterian phyla include sponges (which lack muscles and nerves), jellyfish, and sea anemones and their allies, which have quasiradial symmetry and organ systems that develop from only two tissue layers (see **illustration**). *See* ANIMAL KINGDOM; ANIMAL SYMMETRY; METAZOA.

Classification. Within Bilateria three main branches are usually recognized (see illustration). The first branch, Deuterostomia, includes the phyla Echinodermata (for example, sea stars) and Chordata (including vertebrates). A second branch, Ecdysozoa, includes Arthropoda (for example,


Phylum Porifera (sponges)
Phylum Coelenterata (jellyfish, etc.)
Phylum Acoelomorpha (acoel flatworms)
Deuterostomia (includes Chordata)
Edcysozoa (includes Arthropoda)
Lophotrochozoa (includes Mollusca)
BILATERIA

Position of Bilateria and of its principal subdivisions within the animal kingdom (Metazoa).

insects and crustaceans) and allied forms. The third branch, Lophotrochozoa, includes Mollusca (for example, clams and snails), Annelida (for example, earthworms), and allied forms. *See* ANNELIDA; ARTHROPODA; CHORDATA; ECHINODERMATA; MOLLUSCA.

Fossil record. The first metazoan fossils date back about 580 million years ago; they may include bilaterians, though their identities are not certain. The earliest evidence of bilaterian activity comes from 560- to 555-million-year-old trace fossils left by organisms crawling on muddy sea floors, leaving a trail in the sediment. Those organisms were small (less than a millimeter in width), and their precise body architecture is uncertain. Although the fossil record between 560 and 530 million years ago contains a few enigmatic forms that may be larger bilaterians, the recognizable ancestors of many living bilaterian phyla, especially those with durable skeletons, did not appear until about 530 to 520 million years ago. These appearances, collectively known as the Cambrian explosion, are our first glimpse of the wonderfully disparate bilaterian architectures known today. *See* FOSSIL; TRACE FOSSILS.

Living bilaterians. The earliest branch of living bilaterians comprises small simple worms called acoels. These worms probably represent the general architectural plan of the trace-makers of 560 million years ago. Acoels lack true body cavities, and instead of a brain they have a nerve plexus. Circulatory and excretory systems are also absent; since they are minute, diffusion is adequate for exchanging metabolic materials with the environment.

Simple worms are composed of approximately 15 different types of cells. In comparison, humans have a few hundred cell types and are vastly more complex, yet they have only about twice the number of genes as simple worms. Many of the same genes are present in most bilaterians, which have achieved their diversity and complexity by the use of different patterns of gene expression in increasingly complex networks to produce different body architectures in diverging lineages adapting to distinctive ways of life. The rise of this extensive morphological diversity and complexity is much of the story of animal evolution. *See* ANIMAL EVOLUTION; GENE.

James W. Valentine

Bibliography. R. C. Brusca and G. J. Brusca, *Invertebrates*, 2d ed., Sinauer Associates, Sunderland, MA, 2002; J. Cracraft and M. J. Donoghue (eds.), *Assembling the Tree of Life*, Oxford University Press, 2004; C. Nielsen, *Animal Evolution*, 2d ed., Oxford University Press, 2001; J. W. Valentine, *On the Origin of Phyla*, University of Chicago Press, 2004.

Bilirubin

The predominant orange pigment of bile. It is the major metabolic breakdown product of heme, the prosthetic group of hemoglobin in red blood cells, and other chromoproteins such as myoglobin, cytochrome, and catalase. The breakdown of hemoglobin from the old red cells takes place at a rapid rate in the reticuloendothelial cells of the liver, spleen, and bone marrow. The steps in this breakdown process include denaturation and removal of the protein globin, oxidation and opening of the tetrapyrole ring, and the removal of iron to form the green pigment biliverdin, which is then reduced to bilirubin by the addition of hydrogen. The formed bilirubin is transported to the liver, probably bound to albumin, where it is conjugated into water-soluble mono- and diglucuronides and to a lesser extent with sulfate. *See* LIVER.

In mammalian bile essentially all of the bilirubin is present as a glucuronide conjugate. The conjugation is catalyzed by the liver microsomal enzyme uridyl diphosphoglucuronyltransferase. Water-soluble bilirubin (conjugated) forms a purple substance when it is reacted with diazotized sulfanilic acid (van den Bergh reaction). Free bilirubin (unconjugated) must be made water-soluble, usually by the addition of alcohol, before the reaction will take place. Plasma bilirubin may be increased in diseases which result in increased hemoglobin destruction or diminished hepatic function.

Bilirubin glucuronide is passed through the liver cells into the bile caniculi and then into the intestine. The bacterial flora further reduce the bilirubin to colorless urobilinogen. Most of the urobilinogen is either reduced to stercobilinogen or oxidized to urobilin. These two compounds are then converted to stercobilin, which is excreted in the feces and gives the stool its brown color. A small amount of urobilinogen is absorbed into the bloodstream through the portal circulation. Some is reconverted into bilirubin and a small portion is excreted unchanged into the urine. On standing, the urinary urobilinogen is oxidized to urobilin. The principal yellow pigment of urine, urochrome, is a compound of urobilin, urobilinogen, and a peptide.

The conjugation of bilirubin is reduced at birth and gradually becomes effective in a period of several days to 2 weeks. A genetic defect in an enzyme mechanism can result in excessive formation of unconjugated bilirubin. *See* HEMOGLOBIN.

Morton K. Schwartz

Bibliography. F. B. Armstrong and T. P. Bennet, *Biochemistry*, 1979; D. Bergsma (ed.), *Bilirubin Metabolism in the Newborn*, vol. 2, 1976; G. J. Dutton (ed.), *Glucuronic Acid, Free and Combined*, 1966.

Binary asteroid

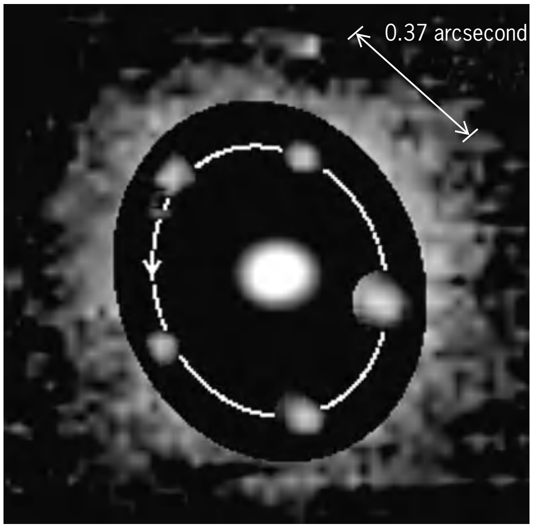
A system composed of two almost equal-size asteroids orbiting around their center of mass, or one moonlet revolving around a larger primary body. The discovery of binary asteroids will aid the work of observers and theorists alike, since they help to determine the nature of these remnants of the solar system formation.

After the *Galileo* spacecraft discovered the first asteroid companion in 1993 [Dactyl, a 1.6-km (1.0-mi) companion of main-belt asteroid 243 Ida], it was realized that satellites might in fact be common around main-belt asteroids. A team led by W. J. Merline reported in 1999 the first direct detection of a satellite, Petit-Prince, of asteroid 45 Eugenia, using the Canada-Hawaii-France Telescope. From a study of the companion's orbit and via Kepler's third law, they derived a low bulk density for Eugenia and concluded that its structure either is that of a "rubble pile" or is composed of primitive, icy materials.

Occurrence. These fundamental findings have motivated several groups to conduct systematic searches for possible satellites of main-belt asteroids using adaptive optics imaging systems, and discoveries are accumulating (see **illus.**). At the begining of 2004, over 20 main-belt binary asteroids had been observed, and this number was growing rapidly. Based upon these surveys, it is estimated that about 2% of all main-belt asteroids are double. Studies suggest that less than 4% of the Trojan asteroids are binary. *See* TROJAN ASTEROIDS.

Evidence is also growing that near-Earth asteroids have companions. The binary nature of several near-Earth asteroids was established by radar observations. The complex visible-light curves of other near-Earth asteroids and the presence of doublet craters on the surface of terrestrial planets indicate that about 10–20% of near-Earth asteroids are binary systems.

After the binary trans-Neptunian object system consisting of Pluto and Charon, additional double trans-Neptunian objects were discovered, beginning in 2001. Hubble Space Telescope observations revealed that most of them are composed of similarly sized companions separated by more than 22,000 km (13,700 mi). *See* HUBBLE SPACE TELESCOPE; KUIPER BELT.



Asteroid 121 Hermione and its moonlet companion observed with the adaptive optics system mounted on the European Southern Observatory's Very Large Telescope [consisting of four 8.2-m (25-ft) telescopes] at six different times. The satellite revolves at a distance of about 790 km (490 mi) from the primary in 1.65 days. The angle of 0.37 arcsecond corresponds to a length of 750 km (466 mi) at the distance of the asteroid. (*After F. Marchis et al., On the mass and density of 121 Hermione from the analysis of its companion orbit, Icarus, in revision, 2004*)

Knowledge from observations. Until recently, little was known about the internal composition and structure of the asteroids. Binary asteroids provide unique information about the internal properties (mass, density, and porosity) of their components, as well as about the formation, history, and evolution of the solar system. This information otherwise may only be speculated upon from spacecraft mission flybys, which are extremely uncommon. In addition, the study of the companion orbits is a unique experiment to improve the knowledge of dynamical interactions. For example, both the irregular shape of the primary and strength of the solar tide influence the orbital distance and the system's stability. Also, the angular momentum distribution within the system (that is, the senses of rotation of the components, the separation of the components, the inclination of the orbiting system plane, and so forth) helps to distinguish between formation scenarios such as fast-rotating fission, fragment ejection, impact disruption, gravitational capture, or tidal force disruption. *See* ASTEROID; SOLAR SYSTEM.

Franck Marchis

Bibliography. W. J. Merline et al., Asteroids *do* have satellites, in W. F. Bottke et al. (eds.), *Asteroids III*, pp. 289–312, University of Arizona Press, 2003; S. J. Weidenschilling, P. Paolicchi, and V. Zappala, Do asteroids have satellites?, in R. P. Binzel, T. Gehrels, and M. S. Matthews (eds.), *Asteroids II*, pp. 643–658, University of Arizona Press, 1989.

Binary star

Two stars held together by their mutual gravitational attraction in a permanent (or long-term) association. The stellar universe is hierarchical. Stars exist singly, in binary pairs, in multiple systems, and in clusters. On large scales, roughly 10^5 light-years, astronomical objects congregate into galaxies. In fact, most stars are in binary systems. The Sun, with its collection of planets, is an exception. Stars in binaries revolve around their common center of mass (which itself moves through space). This star-star gravitational interaction makes possible the measurement of stellar masses and other basic properties. Stellar evolution in some binary systems can lead to spectacularly energetic activity. *See* GALAXY, EXTERNAL; MILKY WAY GALAXY; SOLAR SYSTEM; STAR CLUSTERS.

Types of binary systems. Binary systems are observed in various ways. They may be classified as visual binaries, spectroscopic binaries, or spectroscopic-eclipsing binaries, according to the means of observation.

Visual binaries. In these systems the stars are widely separated compared to their sizes and appear as separate points of light in the telescope. William Herschel first detected the curved path of one star relative to the other, proving their physical association (**Fig. 1**). After years of observation, he found that these motions obey Kepler's law of areas, showing that the law of gravitation applies in the stellar regime. Orbital periods run from years to hundreds of years, and orbits are usually noncircular. Binary stars provide the only proof that has been found so far of gravitation's validity beyond the solar system. (This proof has been extended to general relativity by observations of slow orbital decays in binary pulsars.) *See* CELESTIAL MECHANICS; GRAVITATION; KEPLER'S LAWS.

More detailed observations reveal the apparent orbit of each star relative to the center of mass, leading directly to the stellar mass ratio of the two components. If the binary's parallax (distance) can be measured, individual stellar masses can be found. Without a knowledge of masses, stellar evolution cannot be understood.

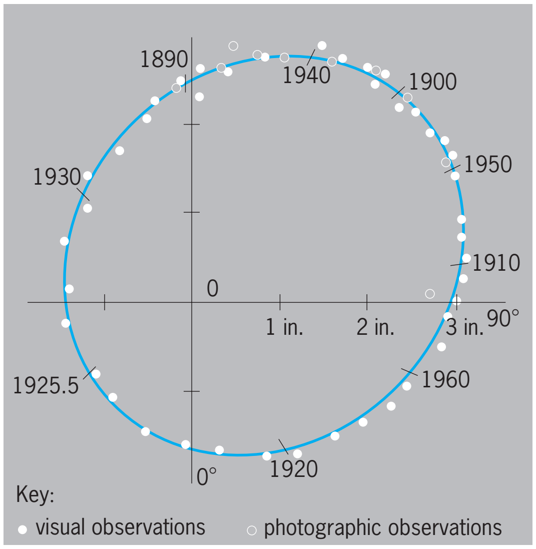


Fig. 1. Apparent orbit of visual binary Krüger 60.

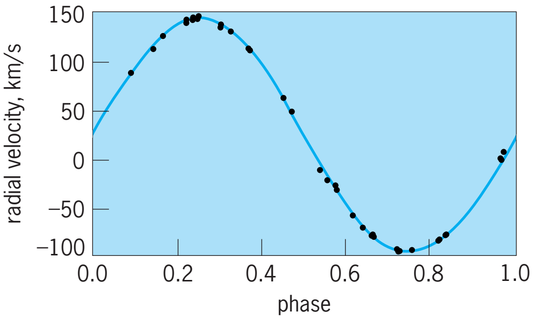


Fig. 2. Radial velocity curve of the loser in the eclipsing binary RX Geminorum. Phases 0.0 and 1.0 correspond to mideclipse of the hotter star. (*E. C. Olson and P. B. Etzel, Mount Laguna Observatory*)

Fewer than 1000 visual binaries have known orbits. Few professional astronomers now make these observations, and amateurs have been encouraged to take up this work. In some cases, orbital periods are so long that binary stars are recognized only by their parallel motions through space.

Atmospheric seeing and telescopic resolution prevent the measurement of visual binary orbits where stars are less than about 0.5 arc-second apart. Interferometric observations have made possible the measurement of more closely spaced binary pairs, opening a new vista on these objects.

Stars in visual binary systems comprise main sequence dwarfs and more evolved giant and supergiant stars. Observations with the Hubble Space Telescope reveal visual binaries composed of brown dwarf pairs, whose masses are too low to ignite hydrogen thermonuclear fusion in their cores. *See* BROWN DWARF; DWARF STAR; GIANT STAR; SUPERGIANT STAR.

Spectroscopic binaries. Large stellar separations in visual binaries result in very low orbital speeds. There is a large class of nonvisual binaries in which the stars are separated by only a few stellar diameters. They cannot be resolved in telescopes, and are referred to as close binaries. The closeness of the stars profoundly affects their evolution. It also creates high orbital speeds (tens to hundreds of kilometers per second), rendering them easily detectable through Doppler shifts of their spectral lines. During the orbital period, which may range from a fraction of a day to hundreds of days, the radial (sight-line) velocity varies. If the orbits are circular, as they often are in close binaries, then the velocity-time curve is sinusoidal (**Fig. 2**).

Radial velocity curve shapes for elliptical orbits are readily calculated. If the two stars are of comparable luminosity, both stars contribute equally to their combined spectra and stellar lines are double over much of the orbital cycle. These are double-line spectroscopic binaries. The inverse ratio of their velocity amplitudes is the stellar mass ratio. Some 1500 spectroscopic binaries are known, though only a small fraction have well-determined velocity curves.

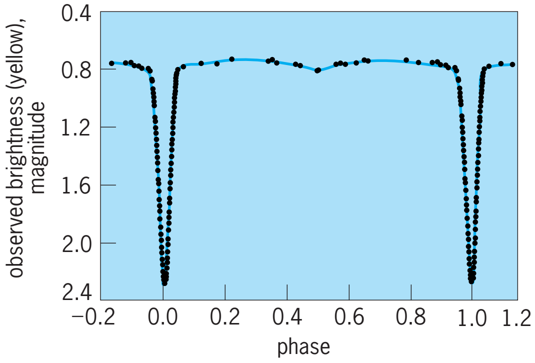


Fig. 3. Yellow light curve of RX Geminorum. Observed brightness is in magnitudes, where an increase of 1.0 magnitude corresponds to a brightness decrease by a factor of 2.5. (*E. C. Olson and P. B. Etzel, Mount Laguna Observatory*)

If the orbit plane is normal to the sight line, there is no radial motion and no Doppler shift. Where shifts are detected, the orbital inclination is still indeterminate. Thus, the information from spectroscopic binaries is limited. *See* ASTRONOMICAL SPECTROSCOPY; DOPPLER EFFECT.

Spectroscopic-eclipsing binaries. Here the orbital plane of a close binary is nearly in the sight line. Mutual stellar eclipses occur, yielding a brightness-time variation called the light curve (**Fig. 3**). The shape of this curve depends strongly on the inclination and relative stellar sizes and surface temperatures. That is, these quantities determine there are eclipses, and the nature of such eclipses (total, partial, or annular). Analysis of the light curve yields the inclination which, coupled with radial velocity curves of both stars, leads to both stellar masses and stellar sizes. (Radial velocities scale the orbital size, as parallaxes do in visual binaries.) Fairly good stellar parameters have been found for a few hundred eclipsing binaries. Interferometric techniques, lunar occultations, and such binary observations are the only source of stellar sizes (apart from the Sun). *See* OCCULTATION.

Modern eclipsing-binary light curves, usually obtained in various colors, reveal higher-order effects. These include mutual gravitational and rotational distortions of the stars, mutual heating of their facing hemispheres, limb darkening of the stellar disks, gravity brightening (a radiative transfer effect that brightens high-gravity parts of the surface), effects of star spots, and orbital eccentricity (though most orbits are nearly circular). Where orbits are noncircular, the major axis of the orbit slowly precesses, and in a few cases provides a critical test of general relativity. *See* ECLIPSING VARIABLE STARS; RELATIVITY.

Stellar evolution. To understand the evolution of binary systems, single stars will first be considered. Stars form from contracting clouds of hydrogen-rich interstellar gas and dust. Contraction releases gravitational energy which heats the forming star. When the core temperature reaches 10^7 K, thermonuclear fusion of helium nuclei from hydrogen nuclei (protons) begins. Fusion occurs only in the stellar core (about 10% of its total mass), and the star stops contracting. Stars spend most of their lives as main sequence dwarfs. Low-mass stars remain longest on the main sequence. The Sun, of 4.5×10^9 years age, is about halfway through this hydrogen-burning phase.

Red giants and white dwarfs. At the end of its main sequence life, the now nearly pure helium core is temporarily out of nuclear fuel. It reverts to gravitational contraction, bringing hydrogen-rich matter inward and heating it. Evolution now accelerates. A thin shell of hydrogen ignites around the contracting core, and the outer layers respond to the new energy source by expanding. A single star becomes a red giant (or supergiant, if massive enough). Its radius increases by 10–100 times and the surface cools (hence, its red color). In some famous visual binaries, the brighter component is a red giant and the fainter a hot main sequence star. Both stars formed together and are the same age. The red giant, of larger mass, reached the end of its main sequence phase first and subsequently expanded to its current state.

Contraction continues to heat the core of a red giant, until helium ignites in the so-called triple-alpha process, to fuse carbon (in repeated explosive flashes if the core has become electron-degenerate). Additional nuclear fusions may occur. The evolving star now has a highly compressed core and a very extended, rarefied envelope. If the star is not too massive, the envelope is eventually driven off to form a planetary nebula, leaving the hot, dense exposed core to cool and become a white dwarf of roughly terrestrial size. *See* PLANETARY NEBULA; WHITE DWARF STAR.

Supernova explosions. Alternatively, stars more massive than about 8 solar masses end their lives in supernova explosions, leaving a neutron-star remnant of about 10-km (6-mi) radius, or a black hole. Neutron stars are radio (or even optical and x-ray) pulsars by virtue of their highly magnetic, rapidly rotating state. The best known is the Crab Nebula pulsar, a remnant of the supernova explosion observed by the Chinese and others in 1054. The pulse interval is the neutron-star spin period. The Crab pulsar rotates about 30 times per second. As single neutron stars radiate photons and particles, they gradually slow down. *See* BLACK HOLE; CRAB NEBULA; NEUTRON STAR; PULSAR; SUPERNOVA.

End products. Thus, following post-main-sequence envelope expansion and loss, and possibly a supernova explosion, stars reach their graves as white dwarfs, neutron stars, or black holes. Electron-degenerate matter in white dwarfs is roughly 10^6 times as dense as water, while neutron-degenerate matter in neutron stars is at least 10^7 times denser, approaching the interior density of an atomic nucleus. White dwarfs and neutron stars have been described as gigantic nuclei.

Evolution of close binaries. Stellar evolution in close binaries is drastically changed by the second star. The system is already a binary in its pre-main-sequence formative stage. Initial conditions determine whether the system will be a visual binary or a close binary. In the latter, the more massive star will

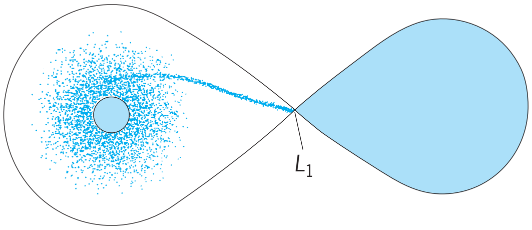


Fig. 4. Roche lobes surrounding the stars in a binary system. The contact point is L_1. The less massive star on the right already fills its lobe. A mass-transferring stream supplies an accretion disk around the other star. (*Matthew Senay*)

exhaust its core hydrogen first and begin its evolution to a red giant. Its expansion is halted when the star's surface reaches the so-called inner lagrangian surface. This is a teardrop-shaped equipotential surface, usually called the Roche lobe, that touches the lagrangian surface of the less massive companion at the inner lagrangian point L_1.

Instead of expanding through this surface, matter flows along the Roche lobe to L_1, evaporates from L_1, and forms a stream that accelerates hypersonically toward the companion (**Fig. 4**). In other words, the post-main-sequence star transfers matter to its companion. One star loses mass (the loser) while the other gains mass (the gainer). These surfaces and motions are determined by the gravitation of the stars and the fictitious centrifugal force in the coordinate frame that rotates with the orbital motion.

A visual binary, in which both stars are far from their Roche lobes, is referred to as detached. A system in which one star has evolved to fill its Roche lobe is called semidetached (Fig. 4). W Ursa Majoris stars are a class of main sequence binaries where both stars fill their lobes and share in effect a common envelope. Their periods are usually less than a day, and they are called contact binaries.

Algol eclipsing binaries illustrate evolution in close binaries. They contain a fairly massive main sequence gainer and a less massive, evolved subgiant (or giant) loser that fills its Roche lobe. Their existence posed the Algol paradox: how could the less massive star be the more evolved? The more evolved star must have originally been the more massive; mass transfer through L_1 to the companion has reversed the mass ratio. Transfer rates in Algols are of order 10^{-7} solar mass per year. Thus, evolution in these binaries is some 100 times more rapid than is solar evolution on the main sequence. RX Geminorum (Figs. 2 and 3) is a long-period Algol.

The Roche lobe prematurely removes the outer envelope of the post-main-sequence loser, precluding the planetary nebula phase of single star evolution. The loser may already have an electron-degenerate core. At some point the gainer will itself evolve from the main sequence. Its envelope expansion may well engulf the original loser, whose orbit will then decay because of friction with the envelope gas (the so-called common-envelope phase). This shortens the orbital period. If conditions are right, the envelope will eventually disperse, and the final product will be a binary white-dwarf system. Such binaries are known. Plausible evolutionary sequences that can explain a large variety of high- and low-mass binary star configurations are now known. *See* STELLAR EVOLUTION; SYMBIOTIC STAR.

Cataclysmic binaries. Stellar evolution with a common-envelope stage may produce cataclysmic variables, in which a white dwarf is paired with a Roche-lobe-filling main-sequence loser. Orbital periods are typically less than a day. The mass-transferring stream from the loser supplies an accretion disk rotating around the white dwarf. Matter in the stream is rapidly accelerated toward the white dwarf, producing significant shock heating on impact with the disk. Disk viscosity, necessary to transfer angular momentum outward to allow matter to fall onto the white dwarf, helps to maintain the disk temperature. Here the energy source is also gravitational.

The cataclysmic binary zoo includes novae, dwarf novae, recurrent novae, and novalike variables. Much of their light comes from the thick accretion disk. Thermal instabilities in the disk may trigger mass dumps onto the white dwarf, producing eruptive brightenings. Amateur astronomers have been enlisted to search for such events. In some cases, hydrogen-rich matter from the loser slowly accumulates on the high-gravity surface of the white dwarf. At a critical density this matter becomes electron-degenerate, igniting a thermonuclear runaway that is a nova outburst. The explosion does not destroy the star, and outbursts may recur after many centuries of quiescence. *See* CATACLYSMIC VARIABLE; NOVA.

High-energy close binaries. The most energetic binary systems are those that contain a neutron star (or rarely, a black hole) remnant. This can be seen from a comparison of the available energy sources.

Stellar energy sources. Stellar evolution taps both gravitational and thermonuclear energy sources. Nuclear sources are usually the most potent available. For example, gravitational energy release accompanying the formation of the Sun supplied its luminosity for a scant tens of millions of years. Core hydrogen burning yields a main sequence solar lifetime of some 10^{10} years. Thus, thermonuclear energy exceeds gravitational energy by about 3000 times in the Sun.

For this discussion a unit is chosen to be the amount of energy released when a kilogram of hydrogen is fused into helium in thermonuclear reactions. A kilogram of matter falling from a large distance onto the Sun's surface releases about 3×10^{-4} unit of gravitational energy. Thus, thermonuclear production is much larger than gravitational. (This ratio is consistent with the time scales just quoted, since only about 10% of the Sun's mass fuses hydrogen.)

The remnants of stellar evolution are not much larger than a solar mass, but are compressed to Earth size in white dwarfs and to a radius of about 10 km (6 mi) in neutron stars. Matter falling onto such compact objects releases much more gravitational energy. For a white dwarf it is about 3×10^{-2} unit per kilogram (nuclear energy still dominates), but for a neutron star it is roughly 50 units. By a wide margin, gravitation wins the energy contest when neutron stars (or black holes) are involved.

Single neutron stars. Single neutron stars are rapidly rotating and highly magnetic, with typical polar fields of 10^8 tesla. (The field has been amplified by the collapse that formed the neutron star.) Spin periods run from fractions of a second to a few seconds, and slowly increase as rotational energy is lost by radiation. Most of the radiation originates near the magnetic poles. Magnetic and rotational axes are not quite aligned (the case even with the Earth). The magnetic pole facing the Earth is visible for only half the rotational period, producing the lighthouselike pulsing of the observed radiation (which is usually at radio wavelengths). This radiation is further beamed by the radiation mechanism.

X-ray binaries. Neutron stars in close binary systems, where the companion fills its Roche lobe and transfers mass, can tap the enormous gravitation energy of in-falling matter. The strong magnetic field interferes with the formation of an accretion disk and funnels transferred mass onto the magnetic poles of the neutron star. This matter, enormously accelerated, strikes the surface at about half the speed of light and heats the poles to 10^8 K. Thermal radiation at such enormous temperatures (hotter than the solar core) is mainly x-rays. An x-ray binary results. Her X-1 is a well-known x-ray binary. Its pulse period is about 1 second, and the binary orbital period is just under 2 days.

High- and low-mass x-ray binaries are now recognized. The former are born with stars of comparable mass, totaling more than 20 solar masses. The more massive component explodes as a supernova and leaves a neutron star (or a black hole) remnant, but the binary is not disrupted. As the second star evolves and expands, it feeds mass to the neutron star. The resulting x-radiation may reach 10^5 solar luminosities, making these high-mass binaries among the most luminous objects in the Milky Way Galaxy.

Low-mass x-ray binaries form with stars of differing mass (typically 10 and less than 2 solar masses). The more massive star suffers a supernova explosion and leaves a neutron star remnant. Mass accreted from the other star may spin up the neutron star to make a so-called millisecond pulsar. The enormous radiation from the neutron star may literally strip the companion to its degenerate helium core, as seems to be happening in the Black Widow Pulsar.

If the unseen x-ray source in a close binary has a mass significantly larger than about 3 solar masses, it cannot exist as a neutron star. According to current theory, it must have collapsed into a stellar-mass black hole. A handful of such candidates exist in the Milky Way Galaxy.

Type Ia supernovae. When the evolution of a moderate-mass binary results in two carbon-oxygen white dwarfs orbiting each other inside a common envelope, a remarkable event may occur. Viscous friction with the envelope causes the stars to spiral inward. If they merge into a total mass larger the about 1.4 solar masses (the Chandrasekhar limit), extremely intense thermonuclear reactions produce iron-peak elements, the structure abruptly runs out of thermonuclear energy, and an implosion-explosion occurs, producing a type Ia supernova. These highly luminous objects are seen in distant galaxies. They behave as standard candles (distance indicators), and are used to study the long-term evolution of the expansion rate of the universe. *See* COSMOLOGY; NUCLEOSYNTHESIS; UNIVERSE.

Gamma-ray bursts. The most energetic events in the universe are gamma-ray bursts, now suspected to occur in distant galaxies. Their origin is currently unknown, but one hypothesis suggests the merger of a pair of neutron stars to form a black hole. There are several binary pulsars (containing pairs of neutron stars) in the Milky Way Galaxy. Relativistic effects cause them to radiate their orbital energy, leading to possible coalescence. A gamma-ray burst in the Milky Way Galaxy could pose a serious hazard to life on Earth.

Frequency and origin of multiple systems. Observations show that the ratios of single:binary:triple: quadruple systems are about 58:35:6:1. That is, a randomly selected stellar system has about a 42% chance of being nonsingle, and most stars are in binary or multiple systems. The fraction of binaries may be even higher among forming, gravitationally contracting stars (known as T Tauri stars). Theoretical calculations suggest that fragmentation of contracting, rotating clouds may produce binary and multiple stars, if cloud rotation is sufficiently rapid. *See* PROTOSTAR; T TAURI STAR.

Extrasolar planetary systems. More than a dozen such systems are known. While not binary stars, most have been found using radial velocity observations like those used for spectroscopic binaries. However, while radial velocities of kilometers per second are seen in binary stars, the detection of planetary systems requires velocity measurements of meters per second.

Edward C. Olson

Bibliography. A. H. Batten, *Binary and Multiple Systems of Stars*, Pergamon Press, Oxford, 1973; I. Iben and A. Tutukov, The lives of binary stars, *Sky Telesc.*, 95(1):42–48, January 1998; F. H. Shu, *The Physical Universe*, University Science Books, Mill Valley, CA, 1981.

Binaural sound system

A sound-reproducing system in which sound is recorded or transmitted by using two microphones mounted at the ears of a dummy human head. To preserve the binaural effect, the sound must be monitored by a listener wearing a set of earphones identically spaced. In an ideal binaural transmission system, both the amplitude and phase of the sound waves incident on the dummy's ears are duplicated at the listener's ears.

Binaural systems have been made so perfect that the listener is unable to distinguish the monitored sound from the real sound. For example, when a person walks around the dummy head, the listener, upon hearing the footsteps, has the compelling illusion of someone walking around him. No other sound system thus far devised can even approximate such an effect. *See* SOUND-REPRODUCING SYSTEMS; STEREOPHONIC SOUND; VIRTUAL ACOUSTICS.

Harry F. Olson

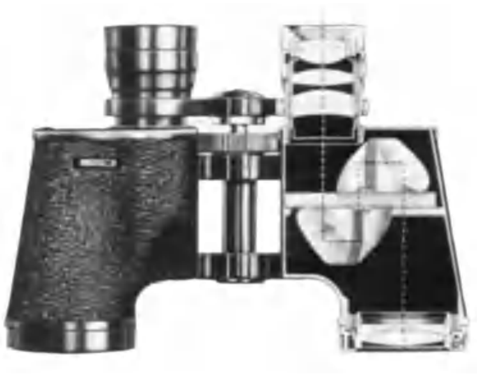
Bibliography. G. Alkin, *Sound Recording and Reproduction*, 3d ed., 1997; J. Eargle, *Handbook of Sound System Design*, 1989; F. A. Everst and R. Streicher, *The New Stereo Soundbook*, 1992.

Binoculars

Optical instruments consisting of a parallel pair of matched telescopes, used to extend the range of stereoscopic vision to far distances. Each half of a binocular consists of a telescope objective, a prism for inverting the image, and an eyepiece (see **illus.**). *See* EYEPIECE; STEREOSCOPY; VISION.

Characteristics. The characteristics of a binocular are stated by using a pair of numbers, such as 7×50 (seven by fifty), where the first number is the magnifying power of the binocular and the second is the diameter of the objective of the binocular in millimeters. The apparent field of view which the user sees is given by multiplying the angle in object space by the magnifying power. For example, an object appearing to subtend 5° by the unaided eye will appear to be enlarged to about 35° when viewed through the binocular. *See* MAGNIFICATION.

Image inversion. Since a lens forms an inverted image, the straight-through combination of an objective and eyepiece would provide an inverted field of view to the eye, as in an astronomical telescope. Almost all binoculars use prisms with an odd number of reflecting surfaces to invert the image correctly. The choice of prisms must also provide for the adjustment in pupil separation by having the optical axes on either side of the prism displaced but parallel. The most frequently used prism is a Porro prism, which is a combination of two 45°–90° prisms. These lead to a bulky mechanical construction, so that many modern binoculars use straight-through prisms. *See* MIRROR OPTICS; OPTICAL PRISM.



Modern prism binocular. (*Bausch and Lomb Optical Co.*)

Focusing and variable magnification. The magnification produced in binoculars reduces the apparent depth of focus because of the enlarged image. Therefore, binoculars require some ability to change the separation between the eyepiece and the objective to provide focusing to accommodate different object distances and possible refractive errors in the eye. Most binoculars provide for a joint focus adjustment of both tubes with one eye having an additional focus range to compensate for users who have differing refractive errors in each eye. The ability to adjust for focus becomes more important with the age of the user, since the eye's accommodative range is diminished as one grows older.

Another optical feature often available in binoculars is a variable magnification or zoom system. This can be accomplished by moving elements in either the eyepiece or the objective, although the smaller elements of the eyepiece are usually chosen for practical reasons. *See* ZOOM LENS.

Selection. Selection of binoculars should be made with some consideration of the intended use. A larger objective will permit use of the binoculars at lower light levels. However, binoculars with larger-diameter objectives and higher powers are heavier and less convenient to use. The jitter produced while holding the binoculars will be magnified by the binocular, so that very high power binoculars usually require a stable support, such as a tripod, to be used properly. A modest power such as 6 is usually more comfortable than a high power such as 10. The wider the field, the more complex and heavier the binocular, as well as the more costly. In addition, the eye relief, the distance that the eye may be placed behind the rear surface of the eyepiece, is important to user comfort. Since the full field of the binocular can be viewed only if the eye location coincides with this exit relief distance, long eye relief is important to users who wear corrective lenses.

The mechanical characteristics of binoculars are also of importance. Errors in alignment of the optical axes of the two halves of the binocular cause visual fatigue when the instrument is used for long periods. The instrument should be rugged, and the images through the eyepieces of a distant object should properly overlap, and not be rotated with respect to each other. There are, necessarily, some aberrations and distortion present in binoculars, especially compact and wide-field binoculars. A reasonable portion of the field of view should be free of obvious blurring or color fringes, and the distortion should be similar in each half of the binocular. *See* ABERRATION (OPTICS).

Opera glasses. Opera glasses are a type of low-power binoculars which use simpler optics. The use of a negative lens as an eyepiece, as in a Galilean telescope, limits the power and the field of view but permits a lighter and less expensive instrument. *See* TELESCOPE.

Robert R. Shannon

Bibliography. H. Paul, *Binoculars and All Purpose Telescopes*, 1980; L. J. Robinson, *Outdoor Optics*, 1989.

Binomial theorem

One of the most important algebraic identities, with many applications in a variety of fields. The binomial theorem, discovered by Isaac Newton, expresses the result of multiplying a binomial by itself any number of times, as given in Eq. (1), where the coefficients c_1, $c_2, c_3, \ldots, c_r, \ldots$, are given by Eqs. (2). The standard notation for these coefficients is given by Eqs. (3).

$$(a+b)^n = a^n + c_1a^{n-1}b + c_2a^{n-2}b^2 + c_3a^{n-3}b^3 + \cdots + c_ra^{n-r}b^r + \cdots + b^n \qquad (1)$$

$$\begin{aligned} c_1 &= \frac{n}{1} \\ c_2 &= \frac{n(n-1)}{1\cdot 2} \\ c_3 &= \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3} \\ c_r &= \frac{n(n-1)(n-2)\cdots(n-r+1)}{1\cdot 2\cdot 3\cdots r} \end{aligned} \qquad (2)$$

$$c_1 = \binom{n}{1},\ c_2 = \binom{n}{2},\ \ldots,\ c_r = \binom{n}{r} \qquad (3)$$

In this notation

$$\binom{n}{r}$$

is the coefficient of the term containing b^r in the expansion of $(a+b)^n$. It is a fraction with the numerator and denominator each containing r factors; those in the denominator begin with 1 and increase by 1; those in the numerator begin with n and decrease by 1. It is easily shown that Eq. (4) is valid.

$$\binom{n}{r} = \binom{n}{n-r} \qquad (4)$$

Some examples of the binomial theorem are as follows:

1. $\binom{17}{3} = \frac{17\cdot 16\cdot 15}{1\cdot 2\cdot 3} = 680$

2. $\binom{12}{5} = \frac{12\cdot 11\cdot 10\cdot 9\cdot 8}{1\cdot 2\cdot 3\cdot 4\cdot 5} = 792$

3. $$\begin{aligned}(a+b)^6 &= a^6 + \binom{6}{1}a^5b + \binom{6}{2}a^4b^2 + \binom{6}{3}a^3b^3 + \binom{6}{4}a^2b^4 + \binom{6}{5}ab^5 + b^6 \\ &= a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6\end{aligned}$$

4. $$\begin{aligned}(2x-3y)^4 &= [(2x)+(-3y)]^4 \\ &= (2x)^4 + \binom{4}{1}(2x)^3(-3y) + \binom{4}{2}(2x)^2(-3y)^2 + \binom{4}{3}(2x)(-3y)^3 + (-3y)^4 \\ &= 16x^4 - 96x^3y + 216x^2y^2 - 216xy^3 + 81y^4\end{aligned}$$

Under suitable conditions the binomial formula is valid when n is not a positive integer. In this case the formula does not terminate, but generates an infinite series.

Much of the utility of the binomial theorem stems from the properties of the coefficients. In particular, the coefficient

$$\binom{n}{r}$$

gives the number of combinations of n distinct objects taken r at a time. The set of coefficients for any value of n form a distribution that has fundamental importance in the study of probability and statistics. *See* ALGEBRA; COMBINATORIAL THEORY; DISTRIBUTION (PROBABILITY); SERIES.

Hollis R. Cooley