McGRAW-HILL ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

8 GEO-HYS

www.MHEST.com

G

Geochemical prospecting — Gyrotron

Geochemical prospecting

Exploration for mineral deposits by chemically analyzing sampled rock, soil, vegetation, and other natural materials for trace amounts of the principal and associated elements. By now, most of the easily discovered mineral deposits have been found, making the search for new deposits ever more challenging. In spite of this, worldwide demands are rising for energy, metallic, and nonmetallic mineral resources. Search continues in regions of known mineralization and in areas where deposits are covered by sediment or obscured by vegetation and ice. Geochemical techniques have become an important part of almost all exploration programs. *See* MINERAL; PROSPECTING.

Types of mineral deposits. Mineral deposits occur in two broad categories: syngenetic and epigenetic. Syngenetic deposits form at the same time and by the same processes as the rocks in which they occur. They are a natural product of rock formation, and an understanding of how the rocks form is crucial to knowing where to look for important mineral concentrations. Epigenetic deposits are superimposed on rocks and may be completely unrelated to the processes that formed the surrounding rock (country rocks). Training and experience are important in evaluating both syngenetic and epigenetic mineralization. This is especially true if deposits have been exposed to surface weathering, or if they are covered by vegetation or the debris left behind when glaciers melted. *See* FORMATION; ORE AND MINERAL DEPOSITS.

Syngenetic and epigenetic mineral deposits often are exposed at the surface through prolonged physical and chemical weathering and erosion, and many deposits are completely destroyed as rock systems evolve naturally. The presence of syngenetic deposits can be deduced because the mineralization was part of a larger geologic setting, while epigenetic deposits may leave no clue to their former existence. *See* EROSION; WEATHERING PROCESSES.

Compared to the environment in which mineral deposits formed, the overall target size for explorationists is greater since they are looking not just for the mineralization but also for associated minerals and features that could be identified as important accompanying mineralization. Placer deposits are the exception to the rule. They are concentrations of chemically and mechanically durable, high-specific-gravity minerals that may not be associated in meaningful ways with the surrounding geology. Even in such settings, explorationists may be able to predict the presence of such deposits by looking at the regional setting of stream sediments, glacial deposits, and so on.

Once formed, both syngenetic and epigenetic deposits may be affected by tectonic adjustments (such as faulting, uplift, tilting) and by weathering and erosion as they are exposed at the surface. Thus, explorationists are forced to rely not only on their knowledge of the geologic setting of deposits but also on how mineralization may be affected by surface processes. Deposits may be covered by soil and vegetation or by residual weathering products in regions where soils are not well developed. The cover may be many meters thick, and both soil and residual material may be covered by surface debris and glacial deposits, and sometimes even by ice.

As the postmineralization history of deposits lengthens, the complexity and thickness of weathering products may increase, as does the chance that deposits will be covered by sediment and vegetation. It is increasingly difficult to “see through” the cover as the time of exposure increases, and the training, experience, and insight of the explorationist becomes more important. Augmented by technology, exploration in such areas can still be fruitful.

Brief history. The search for mineral deposits goes back to when humans began using resources. Just as today, early individuals likely noted the coloration of rocks and of soils near rocks that carried mineralization. In some cases, they may even have associated coloration with faults and fractures in the rocks, even to the point of correlating significant mineralization with intersections of linear features. Recognition of the relationship between leached and oxidized caprocks (gossans) and concealed mineralization at depth could not have been far behind.

References in very early literature to the use of stream water and water from springs and seeps as guides to ore attest to the curiosity and intelligence of explorationists as early as the sixteenth century. Most notable is *De Re Metallica* by Georgius Agricola, published in 1556.

By the latter half of the nineteenth century, workers had recognized the association of mineral deposits with centers of igneous activity. The regular zoning of metals with respect to pressure and temperature was becoming obvious. At about this time, wall rock alteration—changes in the mineral composition and appearance of rocks adjacent to mineralization—was recognized in bordering veins. Alteration later would include pervasive changes in the mineralogy and chemistry of large volumes of rock that completely encompass deposits. *See* IGNEOUS ROCKS.

The modern era of geochemical prospecting began in the early 1930s in Europe, specifically in the Soviet Union, and shortly thereafter in Scandinavia. The Soviet Union had a large geological survey for research on sampling protocols and sample survey design, which coincided with the development there of rapid multiple-element emission spectrographic analysis. Sample media included soil, rock, and vegetation. The analyses were semiquantitative, with an error of as much as 30% of the value. But in the regional context of many of the surveys, this would highlight broad chemical anomalies, even specific anomalies in many instances. *See* EMISSION SPECTROCHEMICAL ANALYSIS.

Area selection. Explorationists look first for direct evidence of mineralization, such as fragments of mineralization or rocks that are cut by mineralization. As the obvious deposits were mined out, searchers were forced to consider mineralization processes, including the important characteristics of deposits, metals associated with specific minerals, and types of rocks hosting deposits, as well as how the country rocks were affected by the mineralizing processes. Explorationists focus on genetically significant features, regardless of the type of mineralization. Characteristics that may be interesting but not particularly significant are eliminated, and efforts are concentrated on conceptual and quantitative information until broad targets, known as area selection, emerge. Careful selection of areas minimizes the cost of regional surveys by focusing on targets with the greatest potential.

Anomaly identification. As defined in the *Glossary of Geology* (1987), an anomaly is "a geological feature, especially in the subsurface, distinguished by geological, geophysical, or geochemical means, which is different from the general surroundings and is often of potential economic value." In exploration, especially regional exploration, investigators are looking for any hint that mineralization may be present, and always are alert to "good luck spikes" where samples come directly from zones of mineralization. Such spikes are not required if the sample program has been properly designed, and even with well-designed programs it is possible to miss mineralization.

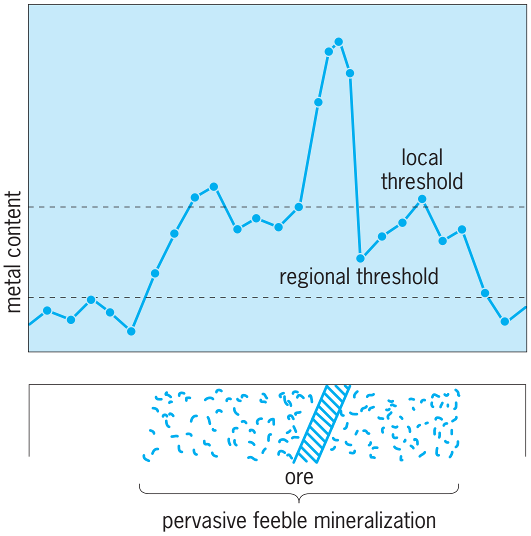


Fig. 1. Recognition of regional and local anomaly thresholds is the basis for geochemical exploration. (*From A. W. Rose et al., Geochemistry in Mineral Exploration, 2d ed., Academic Press, London, 1979*)

Figure 1 illustrates a fundamental characteristic of mineral deposits; that is, mineralization almost always lies within genetically related altered and mineralized rocks, ranging from centimeters thick for some veins to hundreds of meters around some deposits. Such zoning reflects the intensity, complexity, and duration of the mineralization processes. Knowledge of hydrothermal alteration and metal zoning can be used to great advantage, and information on structural orientation, host rock type, and episodes of hydrothermal activity can dramatically affect the size and shape of the target, or regional threshold (Fig. 1). Elements that are associated genetically with the principal target may be more mobile and thus more widely dispersed than the target within the hydrothermal environment (regional threshold). *See* HYDROTHERMAL ORE DEPOSITS.

In geochemical prospecting, indicator elements are the principal targets of exploration efforts, and the pathfinder elements are elements that occur with the indicator elements but often are present in greater quantities and/or exhibit greater mobility and thus greater dispersion within the host rocks. Pathfinder elements will be present in the regional threshold and perhaps within the local threshold (Fig. 1).

Pathfinder elements are likely to be more widely dispersed in weathering environments than indicator elements. They may be liberated more easily during weathering and more widely dispersed than indicator elements. Dispersion during weathering may distort the bedrock pattern considerably (for example, downslope and downstream), and the signature can become too diffuse to be recognizable, if the distance is too great.

If a regional threshold is discernible, the local threshold can be located with follow-up sampling (Fig. 1). Knowledge of the geology and the characteristics of the suspected target become critical as follow-up sampling is done. Sample frequency, perhaps even the configuration and type of samples being collected, may change at this point; for example, going from broadly spaced sediment samples at stream intersections to soil and bedrock samples in a region of iron-oxide-stained bedrock.

Vectoring in on targets. The goal of all exploration is to "zero in" on economic mineralization. Explorationists often call this process "vectoring in," in which geologic knowledge is used together with geochemical data and perhaps geophysical clues to point in the direction where additional data should be gathered, including downward when all factors have vectored in on specific locations.

Well-designed sample programs will include orientation information, by which a guess is made as to the type of deposit being sought, the samples that will have to be collected, and the spacing of samples (for example, grid size for soil samples). A lot of information is available in the literature and company files that can be used to design sampling and analytical protocols, and to help establish a background against which regional and threshold concentrations may be compared. The geology and mining history of a region often can be accessed in the library, to which may be added Landsat information and multispectral data from airborne and satellite platforms. Regional geophysical data on rock magnetism and density, plus electrical conductivity and radioactivity, also may be available. *See* GEOPHYSICAL EXPLORATION; REMOTE SENSING.

Sampling large areas ($\times 10^5$ km^2) usually involves collecting sediments from streams that drain the region, to which may be added data from mineral springs and seeps, even sediment from lakes that are oriented across suspected mineral trends. The confluences of streams are sampled as far upstream as practical. A sample is taken below the intersection or a suite of samples is collected, with one sample from each stream above an intersection and perhaps a third sample from below the intersection. The pH of the water may be recorded and the water may even be sampled, but the sediment is the most reliable medium to test for mineralization. Large fragments are screened out, and the remaining sediment is panned down to a heavy concentrate. The concentrate will provide an indication of mineralization in the drainage, but sieved fractions will allow the data to be refined. For example, iron oxide and manganese oxide coatings are known to scavenge heavy metals that concentrate in the fine fractions of stream sediments.

Land position. Early in the evaluation process, it is necessary to obtain a land position if the preliminary indications of mineralization are promising. Because of the capricious nature of mineral occurrences, it is always a good idea to secure enough land to account for heterogeneities in occurrence. It also may be a good idea to make sure the best available target is being leased or claimed, as well as to acquire every spot that exhibits the same characteristics as the one of initial interest, since mineral occurrences often are clustered together.

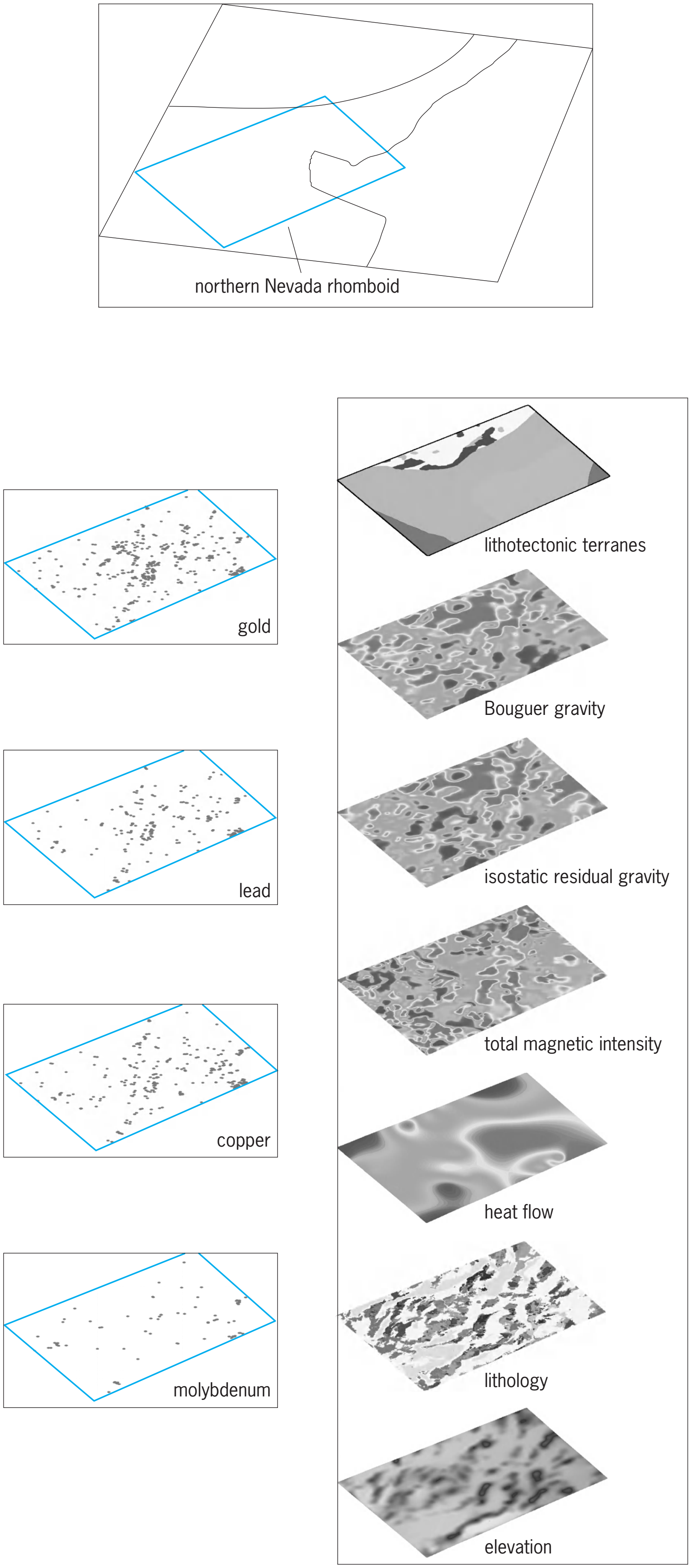
Target-specific exploration. If a reconnaissance study has identified a region of interest and a land position has been established, the explorationist will use techniques that are designed to highlight specific mineral targets. Geologic studies, including preliminary mapping, are appropriate if outcrops are available. In addition, low-level airborne geophysical and ground-based geophysical investigations may help vector in on targets. *See* GEOLOGY; GEOPHYSICS.

Once an area that may contain mineralization has been located, explorationists focus on the deposit type and metals of interest. Sample protocols may be modified at this time, and reconnaissance geology is done, mapping and sampling rocks, with samples in this case consisting of perhaps a kilogram of rock chips collected randomly within 2–3 m (6–10 ft) of outcrops.

Where rock outcrops are limited, small excavations can be made using a rock hammer or a shovel to turn up rock chips from the underlying bedrock. If the region is largely soil-covered, a soil-sample survey may be appropriate. In such cases, the spacing of the grid should be less than the maximum size of the suspected target. It would be a shame to be in the vicinity of mineralization but miss the target by choosing the wrong sample spacing. At a grid interval of 0.25 km, 25 samples collected from 1 km^2 would adequately cover a region if large, porphyry-scale mineralization was the target, but it likely would miss most vein occurrences. Interestingly, many of the historic mining districts of the world were developed in vein deposits, and porphyry mineralization often spawns such veins, so they cannot be ignored as sample plans are being developed. *See* PORPHYRY.

The types of soil being sampled must be well defined to ensure reliable results. The characteristics of soils and the mobility of metals within soils may vary considerably according to the parent rock, the topographic relief, and the climate under which the soil developed. These simple variables determine the depth, type, and stability of soils, as well as the residence time, Eh (oxidation-reduction potential) and pH characteristics of ground water, and the types and activity of soil bacteria. All of these variables influence the extent to which metals may be mobilized from source materials and the location within the soil profile where elements of interest may be concentrated. *See* SOIL; SOIL CHEMISTRY.

Detailed geochemical studies, perhaps accompanied by ground-based geophysical studies, cover less ground than reconnaissance studies, and are more costly but provide very good data, especially when combined with geology. High-resolution soil sampling uses a much closer spacing, and a biogeochemical evaluation may be useful at this time. Plants flux large quantities of water and dissolved components



through their root systems and through new leaves, stems, and reproductive cells. Indigenous plants in desert regions have very well developed root systems that may reach tens of meters below the surface to sample mineralized rocks. Organic debris concentrates in soil and soil litter, as well as in bogs and lake sediments. These media probably should be sampled and interpretation must be done with care, as false anomalies can occur in such settings. Soils contain a large number and variety of bacteria, and these now are studied by well-prepared explorationists. For example, N. L. Parduhn and J. R. Waterson have described a location in Alaska where *Bacillus cereus* were easier to culture from soils above a vein system containing gold and quartz. *See* BIOGEOCHEMISTRY.

Metals and metal compounds with high vapor pressures may "bleed" upward through soil and overburden to reveal anomalous occurrences. Mercury, for example, can be detected with "sniffers" consisting of an inverted cup containing a platinum wire coated with activated carbon. Buddingtonite ($NH_4AlSi_3O_8$), a type of feldspar that occurs with gold and silver deposits, releases ammonia, which can be detected with a sniffer similar in design to the mercury device. *See* FELDSPAR.

Vectoring in the third dimension. Augering, or digging, pits to bedrock, perhaps trench sampling with a bulldozer or backhoe, is expensive, but it provides significant data on rock structure and on hydrothermal alteration and mineralization. The best way to test for mineralization at depth is by drilling. Short of mining, drilling is the only reliable way to determine what is present at depth. Even at this point, the explorationist's work is not done. Chemical data, together with the characteristics of the country rocks and of alteration and mineralization, will define whether the vectoring process has been successful.

Samples of rock cuttings from augering and rotary drilling, and of drill-core samples, normally are analyzed for multiple elements which, together with the rock properties, provide explorationists an evolving picture of the mineral system with depth. Any interval can be chosen, but homogenized samples of 1.5–3-m (5–10-ft) core segments is a good place to start. As many as 75 elements can be determined from samples, with analytical techniques that range from fire assay and atomic absorption spectrometry to inductively coupled plasma spectrometry and mass spectrometry (ICP-MS). Major, minor, and rare-earth element concentrations provide important clues to the nature of the rock systems and to the events that have affected them, but the "big 11" elements provide the most reliable guides to mineralization. From

Fig. 2. Gold, lead, copper, and molybdenum with respect to a GIS for the northern Nevada rhomboid. GIS provides topography, geology, and geophysical background for the metal occurrences. The rhomboid is shown in all frames. Boundaries in the top panel are between igneous rocks with oceanic (NW) and continental (E) affinities. The rhomboid lies largely within a region of mixed oceanic and continental affinities. (*From C. M. Tremper and D. E. Pride, in Window to the World: Geological Society of Nevada Symposium Proceedings, vol. 2, 2005*).

low to high temperature (near surface to deep), they are (1) gold and (2) silver, plus (3) mercury, (4) arsenic, and (5) antimony, (6) lead, and (7) zinc, often with (8) copper, and copper and (9) molybdenum, sometimes accompanied by (10) tin and (11) tungsten. Anomalous concentrations of one or more of these elements signal success, and provide the basis upon which decisions are made that involve time, money, and reputation. *See* ANALYTICAL CHEMISTRY; DRILLING, GEOTECHNICAL.

Data manipulations. A large amount of data routinely is available to explorationists, in addition to mineralization models and personal experience. Data provide both a challenge and an opportunity, with tools available to accept the former and maximize the latter. The isolation of authentic anomalies (the goal in exploration) is accomplished most efficiently when various sources of information are integrated. Mathematical and statistical packages, together with geographic information systems (GIS) technology, provide the tools to combine, overlay, integrate, and manipulate data to highlight the best targets. *See* GEOGRAPHIC INFORMATION SYSTEMS; MATHEMATICAL SOFTWARE.

For grid soil surveys, and for rock surveys in which there is a reasonable distribution of samples, information may be gleaned by contouring element concentrations. The locations of important samples and clusters of samples are easily identified in this way. Since hundreds of samples and dozens of elements may be involved, the efficacy of doing this by computer is obvious.

Moving-average calculations, and evaluations of regional trends in element concentrations, can identify broad anomalies with respect to background, and the "residuals," or samples that vary by more than can be accounted for mathematically, provide strong clues to the presence of anomalies. The data package itself can be "mined" statistically and mathematically by looking for multiple element correlations and ways elements may be related to each other or to other factors that are not obvious in the data. These methods are especially helpful as data packages become larger and contain larger numbers of elements. In such an environment, the ability to analyze and manipulate data becomes very important. Bias enters the picture only from the explorationist, whose experience and training make one set of data and geologic parameters somehow more interesting than a similar data package.

To make decisions, it often is necessary to overlay many types of information. Throughout history, this has been the time-honored approach to exploration, but today we are able to do this with geographic information systems (GIS) technology (**Fig. 2**). Using GIS, the land position, geology, topography, and geophysical parameters can be integrated with data layers of element concentrations, combinations of elements, perhaps ratios of element concentrations, to highlight areas with a high potential for discovery.

Douglas E. Pride

Bibliography. Agricola, Georgius, *De Re Metallica*, transl. by H. C. Hoover and L. H. Hoover, Dover, 1912; R. L. Bates and J. A. Jackson, *Glossary of Geology*, 3d ed., 1987; N. L. Parduhn and J. R. Waterson, *Preliminary Studies of Bacillus cereus Distribution near a Gold Vein and a Disseminated Gold Deposit*, U. S. Geol. Surv. OFR 84-509, 1984; A. W. Rose, H. E. Hawkes, and J. S. Webb, *Geochemistry in Mineral Exploration*, 2d ed., Academic Press, London, 1979; C. M. Tremper and D. E. Pride, Mineralization and the northern Nevada rhomboid, in H. N. Rhoden, R. C. Steininger, and P. G. Vikre (eds.), *Window to the World: Geological Society of Nevada Symposium Proceedings*, vol. 2, 2005.

Geochemistry

A field that encompasses the investigation of the chemical composition of the Earth, other planets, and indeed the solar system and universe as a whole, as well as the chemical processes that occur within them. The discipline is large and very important because basic knowledge about the chemical processes involved is critical for understanding subjects as diverse as the formation of economically valuable ore deposits, safe disposal of toxic wastes, and variations in the Earth's climate.

Geochemistry is rooted in geology, although many of its early contributors were chemists. Much of the initial work in geochemistry was descriptive, as practitioners attempted to determine the exact chemical composition of rocks, minerals, ocean and river water, the atmosphere, and meteorites. Current emphasis is on understanding geochemical processes. V. M. Goldschmidt is usually considered to be the father of modern geochemistry. His investigations focused on the basic "laws" underlying the observed distribution of the chemical elements in the Earth. He and his students conducted systematic element-by-element analytical studies of minerals and rocks in order to provide the basis for such understanding. "Goldschmidt's Rules" predict how the elements will behave, particularly in igneous rocks, based on their size and electrical charge. Like Goldschmidt's work, modern geochemistry is based on accurate analytical studies combined with investigations of the chemical processes that lead to the observed chemical state of the Earth, the solar system, and their component parts.

As in any broad interdisciplinary field, a number of specialties have become subdisciplines. However, their boundaries are by no means rigid, and substantial overlap occurs. A brief description of some of the more important subfields follows.

Isotope geochemistry. This subdiscipline is based on the fact that the isotopic compositions of various chemical elements may reveal information about the age, history, and origin of terrestrial and extraterrestrial materials. Isotopes of an element share the same chemical properties but have slightly different nuclear makeups and therefore different masses. During chemical reactions or processes such as diffusion or evaporation, mass differences between isotopes may lead to enrichment or depletion, and in such cases the isotopic compositions of the materials serve to identify and quantify the processes. Furthermore, the degree of enrichment or depletion is sometimes temperature-dependent, permitting interpretation of the isotopic composition of a sample in terms of past temperatures (paleothermometry).

Some naturally occurring isotopes are radioactive and decay at known rates to form daughter isotopes of another element; for example, radioactive uranium isotopes decay to stable isotopes of lead. Radioactive decay is the basis of geochronology, or age determination: the age of a sample can be found by measuring its content of the daughter isotope.

Both radioactive decay and the processes that enrich or deplete materials in certain isotopes cause different parts of the Earth and solar system to have different, characteristic isotopic compositions for some elements. These differences serve as fingerprints for tracing the origins of, and characterizing the interactions between, various geochemical reservoirs. *See* DATING METHODS; ELEMENTS, GEOCHEMICAL DISTRIBUTION OF; ISOTOPE; LEAD ISOTOPES (GEOCHEMISTRY).

Cosmochemistry. Cosmochemistry deals with nonearthly materials. Typically, cosmochemists use the same kinds of analytical and theoretical approaches as other geochemists but apply them to problems involving the origin and history of meteorites, the formation of the solar system, the chemical processes on other planets, and the ultimate origin of the elements themselves in stars. Cosmochemistry is important because the geochemical makeup and history of the Earth can only be properly understood within the framework of a general knowledge of the chemical composition and evolution of the solar system. Cosmochemistry has some unique aspects. For example, some of the radioactive isotopes created in the interiors of stars and spewed out into space during supernovae explosions have quite short half-lives and have now completely decayed away. Only in meteorites that have remained undisturbed since the time of formation of the solar system does evidence for these isotopes remain. *See* COSMOCHEMISTRY; METEORITE; SOLAR SYSTEM.

Organic geochemistry. Organic geochemistry deals with carbon-containing compounds, largely those produced by living organisms. These are widely dispersed in the outer part of the Earth—in the oceans, the atmosphere, soil, and sedimentary rocks. Organic geochemistry is important for understanding many of the chemical cycles that occur on Earth because biology often plays a major role. An example is the carbon cycle, during which carbon dioxide (CO_2) is consumed by plants, and carbon is stored by burial of plant and animal matter, transformed into coal and petroleum, then released again to the atmosphere as CO_2 when the fossil fuel is burned. Organic geochemists are also active in investigating such areas as the origin of life, the formation of some types of ore deposits that may be biologically mediated, and the origin of coal, petroleum, and natural gas. *See* BIOGEOCHEMISTRY; COAL; NATURAL GAS; ORGANIC GEOCHEMISTRY; PETROLEUM; PREBIOTIC ORGANIC SYNTHESIS.

Other areas. Low-temperature geochemistry is a broad field distinguished by its concern with processes that occur at approximately the ambient temperatures of the Earth's surface. For the most part, this branch of geochemistry deals with chemical reactions and processes involving the Earth's natural waters. Cycles of elements in the oceans, weathering processes on the continents, and sedimentation in lakes and seas are typical areas of interest. Marine chemistry and environmental geochemistry are aspects of low-temperature geochemistry that deal with the chemistry of the oceans and (for the most part) with the geochemical impact of humans on the Earth, respectively. Atmospheric geochemistry is also much concerned with the impact of humans, in this case on the atmosphere. *See* ATMOSPHERIC CHEMISTRY; HYDROSPHERE; MARINE SEDIMENTS; SEAWATER; WEATHERING PROCESSES.

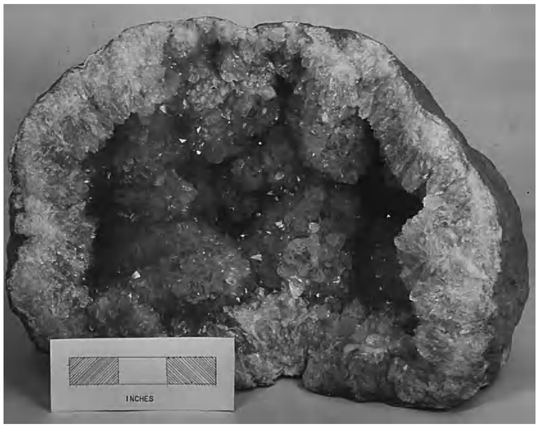
In recent years there has been widespread application of geochemical techniques to problems in paleoclimatology and paleoceanography. In this approach, ocean sediments, sedimentary rocks on land, ice cores, and other continuous records of the Earth's history are analyzed for fossil chemical evidence of past climates or seawater composition. As in most areas of geochemistry, precise and accurate analytical methods for determining the isotopic and elemental composition of the samples are critical. *See* EARTH SCIENCES; PALEOCEANOGRAPHY; PALEOCLIMATOLOGY.

J. D. Macdougall

Bibliography. A. H. Brownlow, *Geochemistry*, 2d ed., 1995; G. Faure, *Principles and Applications of Inorganic Geochemistry*, 2d ed., 1997; K. B. Krauskopf, *Introduction to Geochemistry*, 3d ed., 1994.

Geode

A roughly spheroidal hollow body, lined on the inside with inward-projecting small crystals (see **illus.**). Geodes are found most frequently in limestone beds but may occur in some shales. Typically, a geode consists of a thin outer shell of dense chalcedonic silica and an inner shell of quartz crystals, sometimes beautifully terminated, pointing toward the hollow



Geode, lined with quartz crystals, Keokuk, Iowa. 1 in. = 2.5 cm. (*Brooks Museum, University of Virginia*)

interior. Many geodes are filled with water; others, having been exposed for some time at the surface, are dry. Calcite or dolomite crystals line the interior of some geodes, and a host of other minerals are less commonly found. In some geodes there is an alternation of layers of silica and calcite, but almost all geodes show some banding suggestive of rhythmic precipitation. *See* CHALCEDONY.

The origin of geodes lies in the presence of an original cavity, in many cases a void within a fossil shell, from which the geode originally grew. The geode grows by expansion, the layer of chalcedonic silica being the hardened equivalent of an original silica gel. The expansion is due to osmotic pressure from original seawater trapped inside the silica gel shell and fresh water on the outside of the gel. The projecting quartz crystals are precipitated from later ground waters infiltrating the already hardened, hollow spheroid. *See* SEDIMENTARY ROCKS.

Raymond Siever

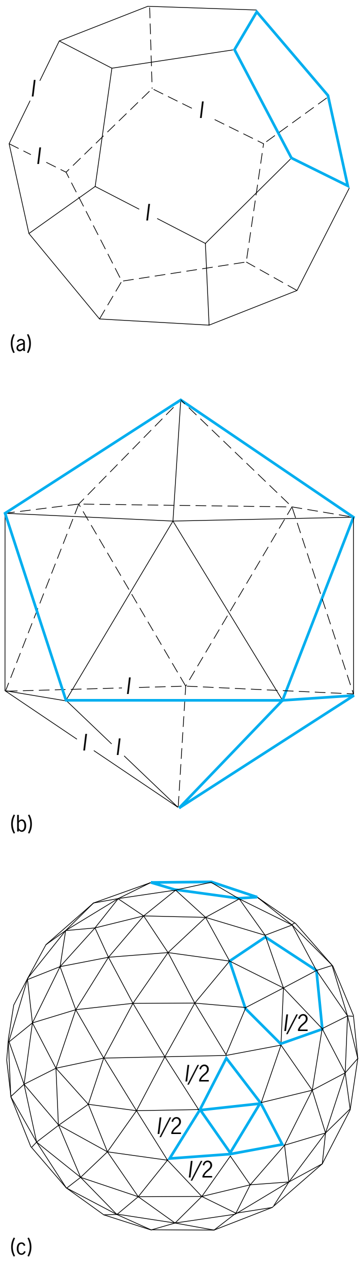
Geodesic dome

A curved lattice grid dome that utilizes the equilateral triangle as the basis of its surface grid geometry. Buckminster Fuller, the inventor and champion of the geodesic dome, obtained a patent in 1954 that described a method of dividing a spherical surface into equilateral triangles. The realization that a triangular modular grid could be imposed on a spherical surface had great impact on the design philosophies of architects in Fuller's time.

Fuller found that if a regular (equal-sided and -angled) pentagon was used as the base of a pentagonal pyramid and if the inclined sides of the pyramid were the same length as the pentagon's sides, the apex of the pyramid would be at the center of the pentagon and on the surface of a sphere, passing through all the vertices of the pentagon.

Based on this geometry, two geometric regular polyhedra (all of whose edges are the same length and all of whose faces are regular, identical polygons) are used in generation of the geodesic dome's grids. The two regular polyhedra that can be inscribed in a sphere are the dodecahedron (12 faces, each of which is a regular polygon; **illus.** *a*) and the more utilized icosahedron (20 faces, each of which is an equilateral triangle; illus. *b*). *See* POLYHEDRON.

The geodesic dome has been used for everything from great exhibition spaces and halls to outdoor tent supports and jungle gyms. Use of the equilateral triangle as the basic modular grid has led to use of the geodesic dome where prefabrication, speed of erection, and dismantling are major considerations. Geodesic domes have been built up to 143 m (469 ft) in diameter. The earliest example is the 30-m diameter (98-ft) dome designed by Fuller and built in 1952. Since 1955, the U.S. Army has used geodesic domes to enclose radar installations (radomes) throughout Canada and Alaska, because the design lends itself to prefabrication of units that can be easily erected.



Geometry of geodesic domes. (*a*) Dodecahedron: a regular pentagon is typical of each face; every point is an apex because all apexes are on the sphere; each strut (*l*) is the same length. (*b*) Icosahedron: an apex is above the center of each polygon and on the surface of the sphere; the equilateral triangle typical of each face is highlighted; each strut (*l*) is the same length. (*c*) Larger dome based on the icosahedron: subdivision is formed by connecting midpoints of struts (*l*) of equilateral triangles (each half strut is labeled *l*/2); the original pentagon is shown at the top, and a formed hexagon is also shown.

Prefabrication of the geodesic dome received impetus from the development of aluminum struts to interconnect the structure, and diamond-shaped aluminum sheets for the covering. The first aluminum dome was built in 1957 in Hawaii; it had a floor diameter of 44 m (145 ft) and a height of 15 m (49 ft). The geodesic dome that received the most publicity was erected in 2 weeks in Moscow in 1959 for the U.S. Technical Exhibition; it had a diameter of 61 m (200 ft).

By utilizing the icosahedron as the basic building block of the geodesic dome, larger domes are possible with additional triangular subdivisions. This subdivision is known as the frequency. The first frequency is to interconnect the projected midpoints of the struts of each equilateral triangle of the icosahedron as they will project on the spherical surface. The result is four almost equilateral triangles where there was one before. The resulting lattice has similar but not exactly equilateral triangles if the grid is to remain on the spherical surface. This subdivision process can continue. The resulting grids have both triangular and hexagonal grids as a by-product within the basic geodesic dome geometry, with pentagons around the apex of the basic underlying icosahedron framework (illus. *c*). I. Paul Lew

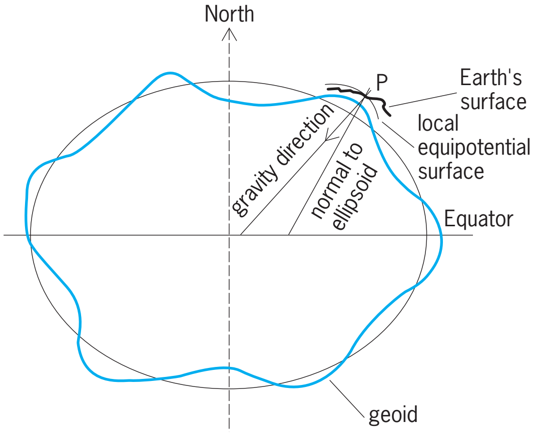


Fig. 1. Determination of geodetic and astronomic latitudes. P represents the location at which the latitude and longitude are to be defined.

Geodesy

The science of measuring the size, shape, and gravity field of the Earth. Geodesy supplies positioning information about locations on the Earth, and this information is used in a variety of applications, including civil engineering, boundary demarcations, navigation, resource management and exploration, and geophysical studies of the dynamics of the Earth. *See* EARTH.

The conventional measurement systems in geodesy are triangulation and trilateration for determining horizontal positions, and leveling for determining heights. These techniques depend on the Earth's gravity field, and so a major part of geodesy has been not only position determination but also the measurement of the Earth's gravity field.

Two major measurement systems were developed in the late 1970s and early 1980s: satellite laser ranging (SLR) systems, which could measure the distance from the ground to a satellite equipped with special corner-cube mirrors; and very long baseline interferometry (VLBI), which could measure the difference in arrival times between radio signals from extragalactic radio sources. With these systems it is possible to measure accurately (within a few millimeters) the distances between points located on different continents, making possible the creation of truly global coordinate systems. Both systems were deployed around the world to measure not only the positions of locations but also the changes in those positions; and thus it was confirmed that the Earth is not a static but a highly dynamic body, with much of this dynamism causing catastrophic events such as earthquakes and volcanic eruptions. The more recent Global Positioning System (GPS) offers much of the capability of SLR and VLBI. *See* RADIO ASTRONOMY.

Earth models. For much of the history of geodesy, measurements were made locally, and this is reflected in the ways that geodetic data are interpreted. When local measurements are made, the only direction that is known is the direction of the local gravity vector. A bubble in a liquid in a curved tube will be perpendicular to this direction, and a plumb line lies along this direction. When an astronomical latitude and longitude are determined, it is the direction of the gravity vector that is determined. When height differences between points are measured by using spirit leveling, it is the heights above an equipotential surface (that is, a surface that has a constant gravitational potential or that is always orthogonal to the local direction of gravity) that is determined.

The direction of gravity is affected by the mass distribution within the Earth; therefore coordinate systems based on gravity are difficult to use for precise determinations of the positions, because the coordinates do not vary in a geometrically predictable fashion. For horizontal positions such as latitude and longitude, it is possible to convert triangulation and trilateration measurements (the measurement of angles or a series of distances between points on the surface of the Earth) to geometric coordinates, that is, ellipsoidal coordinates. However, there is a difference between ellipsoidal and astronomical latitude (**Fig. 1**). For heights, conversion to geometric coordinates requires knowledge of the difference in height between an ellipsoidal shape and an equipotential surface known as a geoid. A geoid is the equipotential surface coinciding approximately with mean sea level, and deviates by up to 330 ft (100 m) from an ellipsoid shape. If mean sea level is assumed to form an equipotential surface, this definition can be used to globally define a height datum along all coastlines around the world. Spirit leveling can then be used to determine the heights at points away from the coast. Geodetic systems such as SLR and VLBI are able to determine the positions of points geometrically, and thus their natural coordinate system is an ellipsoidal one for both height and latitude and longitude. *See* LATITUDE AND LONGITUDE; SURVEYING; TOPOGRAPHIC SURVEYING AND MAPPING.

Techniques. Conventional geodesy uses the techniques of triangulation and trilateration to determine latitude and longitude. In triangulation, the angles in triangles are measured and converted to position by knowing the length of at least one side of the triangle. By linking the sides of the triangles, it is possible to determine all the lengths of all the sides of a linked set of triangles by measuring directly just a few distances in the set. Trilateration uses the direct measurements of the sides of the triangles, and became possible only in the 1950s after the development of radar. Distances are measured effectively by measuring the time it takes for a radio or light signal to travel along the side of a triangle. Individual triangles measured with these systems can be only as large as the eye can see. Linking networks of triangles together makes measurements across a whole country possible.

Height measurements are made using leveling in which, by measuring the height difference between nearby points, the height difference between distant points is obtained by summing the height differences of all linked pairs of points between them. The height differences are measured by setting a telescope perpendicular to the direction of gravity and reading the height difference from rulers (called staffs), each 7–10 ft (2–3 m) long, set vertically above the two points being measured. The telescope and two points need to be close together with separations of usually less than 164 ft (50 m). Long lines of these points form a leveling network that can run for thousands of kilometers and can take many years to measure. The absolute heights of the points in the network are found by having some of the points near the ocean, and the zero of the height scale is set to coincide with mean sea level measured on tide gauges over many years.

Modern geodetic measurements, referred to as a spaced-based system, use time-delay measurements that are far less affected by atmospheric refraction than angle measurement systems. Both SLR and VLBI allow measurement of positions of points separated by thousands of kilometers with accuracies of few millimeters. However, both systems are complex and expensive to operate. Satellite laser ranging uses pulsed laser beams, and the round-trip travel time to a satellite equipped with corner-cube reflectors can be measured to a few millimeters. Very long baseline interferometry records signals from extragalatic radio sources on high-density tape. When these recordings from different stations are cross-correlated, the difference in arrival times of the signals from the extragalatic radio source can be measured to a few millimeters of equivalent light-travel time. *See* LASER.

The use of SLR and VLBI has been restricted to about 200 locations around the world, and they have been accessible only to large government-run measurement programs. However, based on the experience with SLR and VLBI, a geodetic measurement system, the Global Positioning System, was developed, originally to allow world positioning to a few meters' accuracy in real-time, with lightweight portable equipment. The system consists of 24 GPS satellites in orbits that are about 12,000 mi (20,000 km) above the surface of the Earth. The satellites transmit encoded signals at two frequencies (1575.42 and 1227.60 MHz) in the L-band of the radio spectrum. The coding on the signals includes the time when a signal was transmitted and information about the location of the satellites. Each satellite transmits by using a different code sequence, called a pseudo random number sequence, which allows signals from different satellites to be separated in a GPS receiver, which receives signals from all visible satellites (usually about eight). A receiver, by measuring the time on its own clock when signals from different satellites are received and having determined where the GPS satellites are located, is able to determine its position in a global reference frame. The best GPS receivers can measure the time of arrival of the signals to about 1 nanosecond, which is equivalent to 12 in. (30 cm) of light travel time, thus allowing them to determine their positions to about 3 ft (1 m). The receiver is able to make this measurement in about one-fiftieth of a second, and therefore the receiver can continuously monitor its position at this level of accuracy. The GPS revolutionized navigation around the world, allowing aircraft, ships, and other moving vehicles to continuously monitor their positions. *See* ELECTROMAGNETIC RADIATION; SATELLITE NAVIGATION SYSTEMS.

Prior to January 2000, there were restrictions on access to the full accuracy of GPS signals: the transmission times of signals from the satellites were corrupted in a pseudorandom fashion, such that standalone nonauthorized receivers could be positioned only to 330 ft (100 m). This degradation of the positioning accuracy was called selective availability (SA) and could be circumvented by making measurements continuously at a known location and transmitting to all nearby GPS receivers the errors in the times of the GPS transmissions. This differential GPS (DGPS) system is still widely used for precise navigation in harbors and near airports where accuracy of better than 33 ft (10 m) is required. The use of DGPS greatly reduces positioning errors due to errors in modeling the locations of the GPS satellites and signal delays induced by the refraction of the Earth's atmosphere. The other known purposeful corruption of GPS signals, antispoofing (AS), is the addition of an extra code on the signal that denies access to the precise positioning codes that are written on the GPS signals at both frequencies. For very accurate applications, measurements at both GPS frequencies are needed to eliminate the error caused by the propagation of the signals through the Earth's ionosphere. With specialized receivers used in the highest-accuracy applications of GPS, it is possible to make measurements using both GPS frequencies, even in the presence of AS. The generation of GPS satellites launched starting in 2005 includes known codes on both GPS frequencies, making dual frequency measurement much easier.

The most accurate applications of GPS do not use directly the time information on the GPS signals, but the measurement of the phase of the radio signal carrier on which the GPS information is transmitted. The use of this phase information allows accurate measurements in the millimeter range to be made

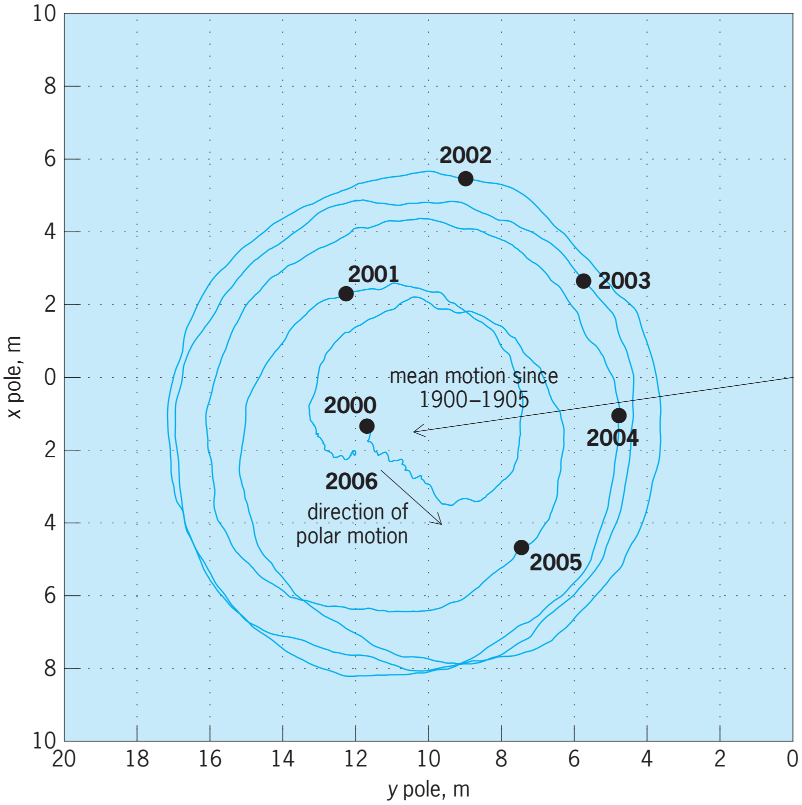


Fig. 2. Motion of the position of the Earth's rotation axis as determined by geodetic methods. The line shows daily measurements of the position of the pole with dots labeled for years 2000–2006. The *x* axis points along the Greenwich meridian, and the *y* axis points in the direction of longitude 90°W.

over thousands of kilometers. Such measurements have been used to study the dynamics of the Earth. For example, 4-in.-accuracy (10-cm) position measurements allow railroads to determine definitely on which track a train is located; vehicle collision avoidance systems are possible with systems possessing this accuracy; and such data can be used in large-scale construction projects.

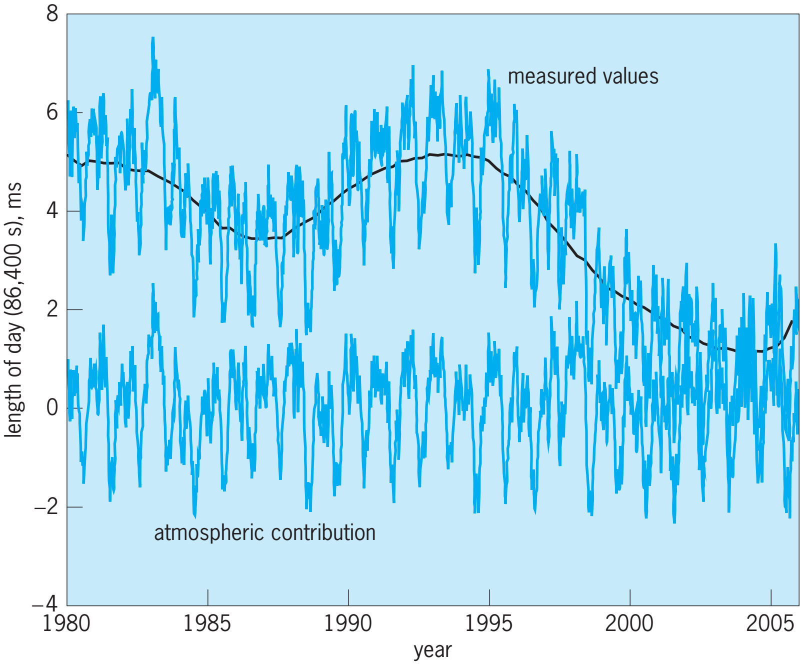


Fig. 3. Changes in the length of day for a 26-year interval, 1980–2006, measured with geodetic methods and inferred from variations in atmospheric angular momentum. The decadal time-scale differences are believed to arise from fluid motions in the Earth's core.

The most recent development in geodetic techniques is interferometric synthetic aperture radar (InSAR). This technique is used to measure heights of the topography or, if the topography is already known, the changes in the topography between two synthetic aperture radar (SAR) images. Heights measured with InSAR are far less accurate than normal geodetic height measurements, but since InSAR is an imaging system, large areas can be measured easily. If the InSAR instrument is on an orbiting spacecraft, global topography can be measured. The measurement of changes in topography with InSAR has been widely used to measure the surface displacements after earthquakes (by comparing before and after SAR images) and for monitoring volcanic deformations. *See* RADAR.

Geophysical applications. Some of the major impacts of modern geodetic measurements have been in the study of the dynamics of the Earth. The measurement systems enable the observation of many of the minute motions of the Earth, such as those associated with plate tectonics and other geophysical processes, and changes in the rotation of the Earth.

Earth deformations. Early evidence for plate tectonics was based on the geologic record, which showed that whole continents moved by tens to hundreds of kilometers per million years, but modern geodesy has allowed these motions to be observed on time scales of 10 years. The tectonic motions of thousands of locations around the world have been measured with VLBI, SLR, and GPS. The measurements clearly reveal the pattern of plate tectonic motions. These results show that plate tectonics is a continuously operating process when the sites are well away from regions of active volcanism and earthquakes. Even more remarkable is the agreement between the expected changes of this distance based on a million years of geology and the measurements made over a 10-year interval. In general, the motions of locations away from regions of active deformations agree within a few millimeters per year with geologic estimates. However, this is not always the case, and it is never the case in regions of active deformation. The study of regions where geology and geodesy disagree is one of the prime applications of modern geodesy. The western United States is one region that has been extensively measured with all forms of geodetic systems. A large motion as the west coast is approached is expected because of the motion of the Pacific tectonic plate relative to North America. However, the precise partitioning of a motion across California can be determined only with modern geodetic systems. The broad width of the motions, going all the way into the Mojave Desert, had not been expected before these measurements were made. *See* CONTINENTAL DRIFT; PLATE TECTONICS.

Earth motions. Measurements of changes in the rotation of the Earth have long played an important role in understanding the dynamics of the Earth.

Changes in the position of the Earth's rotation axis are of two types: changes in direction in inertial space (precession and nutation) and changes with respect to the crust of the Earth (polar motion). Precession is a secular motion of the rotation axis that causes the rotation axis to trace out a cone in space every 26,000 years. Nutations are the smaller-magnitude nodding of rotation axis about the precessional path of the rotation axis. Both precession and nutation are caused by the gravitational torque applied to the equatorial bulge of the Earth, and their values are a direct measure of the dynamic flattening of the Earth. The amplitudes of the nutations are also greatly influenced by the presence of the fluid core and the elastic properties of the Earth. Measurements of the nutations made with VLBI have yielded the most accurate estimate available of the flattening of the fluid core because of its influence on the nutations. Precession and nutations are determined from the apparent changes in the positions of stars and extragalactic radio sources. The major impact of VLBI has been that the extragalactic radio sources observed have very small motions, unlike those of stars. *See* EARTH ROTATION AND ORBITAL MOTION; NUTATION (ASTRONOMY AND MECHANICS); PRECESSION; TORQUE.

Polar motion is measured by apparent changes in the latitudes and longitudes of positions on the Earth's surface when these are determined by astronomical observations. The Earth has a natural period for the oscillations of the rotation axis about a mean position as determined by the direction of the maximum moment of inertia of the Earth; this oscillation mode is known as the Chandler wobble. It takes about 433 days for the axis to complete one cycle in this wobble mode. An example of the motion of the rotation axis can be shown as the position near the North Pole where the rotation axis pierces the Earth (**Fig. 2**). In one determination of this type, the conventional definition of the position of the North Pole was set to the mean position of the Earth rotation axis between 1900 and 1905; it was found that the average position of the rotation axis has changed by about 33 ft (10 m). Most of this motion is thought to arise from the relaxation of the Earth after the deformations caused by the Laurentide glaciation 10,000 years ago. Superimposed on the Chandler wobble is an annual motion of the rotation axis due to atmospheric mass redistribution between the Northern and Southern hemispheres between seasons. It is not known what process keeps the Chandler wobble excited to is current amplitude, but it is clear that changes in atmospheric wind and pressure and in ocean currents that seem to cause most of the smaller fluctuations are the primary mechanism. *See* ATMOSPHERIC GENERAL CIRCULATION.

The rotation rate of the Earth also varies on time scales of days to millions of years. A variety of processes contribute to changes in the rotation rate of the Earth. Some of the major ones are a slow decrease in the rotation rate due to energy loss to the lunar orbit through dissipation of tide energy and exchanges of angular momentum between the atmosphere and the fluid core and the solid Earth. For example, the dominant role of the atmosphere on short-period (less than several years) changes in the rotation rate of the Earth can be demonstrated by measuring changes in the length of day over a 2-year interval with modern geodetic systems, and predicting the changes based on calculation of the angular momentum of the atmosphere from wind speeds in the atmosphere and pressure measurements (**Fig. 3**). The variations are approximately 1 millisecond. The 2-ms mean value of the length of day is due to the slowing of the rotation since the average length of day was defined at the turn of the century. *See* EARTH, GRAVITY FIELD OF. Thomas Herring

Bibliography. G. Bomford, *Geodesy*, 4th ed., 1980; B. Hofmann-Wellenhof and H. Moritz, *Physical Geodesy*, 2005; K. Lambeck, *Geophysical Geodesy: The Slow Deformations of the Earth*, 1988; A. Leick, *GPS Satellite Surveying*, 3d ed., 2003.

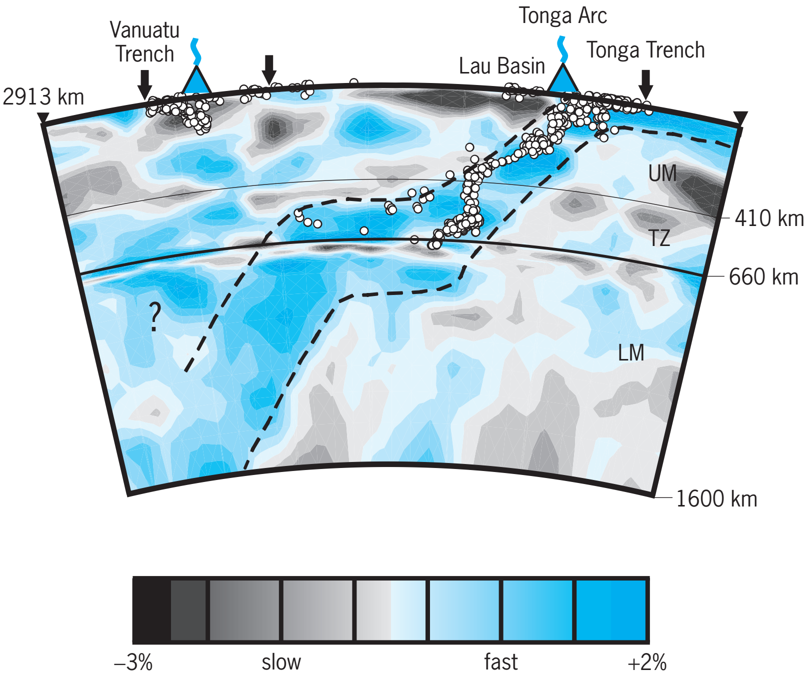
Geodynamics

The branch of geophysics that studies the processes leading to deformation of planetary mantle and crust and the related earthquakes and volcanism that shape the structure of the Earth and other planets. On the largest scale, these processes are a consequence of the transfer of heat out of planetary interiors due to cooling at their surfaces. Rock contracts as it cools, so that its density increases. The cool surface layer is heavier than the interior and has a tendency to sink into it. At the same time, cooling and solidification of the metallic core heats the deepest portion of the surrounding rocky mantle, causing it to become buoyant. The resulting flow of the mantle causes deformation at the surface. Volcanism arises from the partial melting of hot mantle that rises toward the surface from the deeper interior, in response either to buoyancy or to surface deformation. Surface deformation also results from external loads, such as the distribution of ice and water, tidal loads due to the gravitational attraction of nearby planetary bodies, and meteor impacts. *See* EARTH, CONVECTION IN; EARTH, HEAT FLOW IN; GEOPHYSICS; VOLCANO.

A planet's response to its internal heat flow depends largely on the rheology of deforming rock. At low temperatures, near the surface, rock behaves as a brittle-elastic material, allowing the propagation of seismic waves and the support of surface loads by elastic stresses. Deformation occurs by the formation of cracks or faults. On geologic time scales and at the higher temperatures of the deeper interior, thermally activated creep allows the solid, rocky mantle to flow like a viscous fluid. But even at these high temperatures, rock behaves elastically on short time scales so that elastic shear waves propagate through the slowly flowing mantle. In the case of the Earth in its current stage of evolution, plate tectonics describes how the surface behaves: large, cold, relatively rigid plates move laterally across the surface while the deeper mantle flows by creep. In the cold plates, deformation is largely confined to boundary faults between the plates. Faults slip with a stick-slip behavior, giving rise to large earthquakes that occur primarily on the plate boundaries. *See* EARTHQUAKE; PLATE TECTONICS; RHEOLOGY; ROCK MECHANICS.

Geodynamics deals with phenomena that occur on a wide range of time and length scales. In an earthquake, stored elastic strain energy is released in times on the order of seconds. In contrast, significant flow of the deeper mantle occurs on time scales of tens to hundreds of million years. Tidal and glacial loads occur on intermediate time scales ranging from hours to thousands of years. The time- and temperature-dependent rheology of rock is controlled by processes that occur on the millimeter scale of individual mineral grains and the boundaries between them. It is remarkable that processes occurring on this small scale control the behavior of the Earth and planets on scales of thousands of kilometers, comparable to the dimensions of the tectonic plates.

Given the difficulty of direct observation and the wide range of scales involved in phenomena of interest, multiple approaches are needed to understand geodynamic processes. Laboratory experiments on relatively small samples of rock are used to characterize the rock's physical properties, such as its rheology, at high pressures and temperatures. The rate of deformation due to creep in nature is much too slow to measure directly in the laboratory. Thus a theoretical understanding of deformation mechanisms is needed to extrapolate from the much higher deformation rates of laboratory measurements to naturally occurring conditions.



Seismic (P-wave) velocity in the mantle beneath the northern Tonga convergent plate boundary. The Pacific plate approaches the Tonga trench from the right and sinks into the mantle beneath the volcanic Tonga arc. Circles show hypocenters of earthquakes that occur in the sinking plate. The tint bar is the key to seismic velocity anomaly. Seismic velocities in the plate are higher than in the surrounding mantle. Although earthquakes in the sinking plate extend only to 660 km depth, the seismic velocity anomaly continues to greater depth. UM, upper mantle; TZ, transition zone; LM, lower mantle. (*Adapted from R. van der Hilst, Complex morphology of subducted lithosphere in the mantle beneath the Tonga trench, Nature, 374:154–157, Macmillan Magazines Ltd., 1995*)

Field studies of rocks once deep in the interior and brought to the surface by uplift and erosion provide evidence of the processes that have affected them. But the interior of the Earth where the processes of interest are actually occurring is not directly accessible for study. Thus geodynamicists must design large-scale observational experiments that allow them to create conceptual and physical images of the interior, using combinations of seismic, gravitational, electromagnetic, and heat-flow measurements. Variations in global gravity and corresponding surface topography can be remotely sensed from orbiting spacecraft. This is particularly important in the case of planets other than the Earth where other types of geophysical measurements have not yet been made. These data seldom have a unique interpretation, but numerical experiments with theoretical/computational models that describe relevant processes can determine which of these are consistent with available observations. Such analyses, in turn, suggest new observations that might further refine understanding. *See* EARTH CRUST; EARTH INTERIOR; GEODESY; GEOMAGNETISM; SEISMOLOGY.

One long-standing controversy has been whether plates that sink into the mantle at convergent plate boundaries descend into the deep interior or remain in the upper part of the mantle. This question of how deep the mantle circulation implied by plate tectonics extends has important implications for the thermal and chemical evolution of the Earth. A recent accomplishment of seismic tomography has been to image mantle elastic-wave speed variations associated with sinking plates. Tomographic images clearly suggest that plates do, at least sometimes, descend into the deep mantle (see **illus.**).

The questions and problems that geodynamicists seek to answer encompass important concerns for humanity. Understanding where essential energy and material resources occur in the Earth, how they form, and the physical processes involved in their recovery is the basis of the current standard of living. For example, the flow of fluids through cracks in rocks is important in hydrocarbon extraction, nuclear waste isolation, and geothermal energy. Understanding the mechanisms that produce magma at depth, how magma is extracted and transported to the surface, and its episodic and sometimes spectacular eruptions is key to mitigating the hazards that volcanoes pose to human life and property. Geodynamicists are also concerned with environments, such as deep-sea hydrothermal systems, in which primitive forms of life originate and evolve on the Earth, and possibly on other planetary bodies. *See* DEEP-SEA FAUNA; MID-OCEANIC RIDGE; PETROLEUM GEOLOGY; VOLCANOLOGY.

E. M. Parmentier

Bibliography. C. M. R. Fowler, *The Solid Earth: An Introduction to Global Geophysics*, Cambridge University Press, 1990; R. van der Hilst, Complex morphology of subducted lithosphere in the mantle beneath the Tonga trench, *Nature*, 374:154–157, 1995.

Geodynamo

The mechanism thought to be responsible for the generation of the Earth's magnetic field through the convection of conducting fluids in the Earth's core.

Paleomagnetic measurements suggest that the Earth has possessed a magnetic field for at least 3.5 billion years. Geophysicists generally accept that the ambient magnetic field measured at the Earth's surface is due to electric currents flowing in its liquid iron core (**Fig. 1**). In the absence of electromotive forces, like those of chemical batteries, electric currents will decay as magnetic energy is converted to heat. Without some regenerative process to offset such natural ohmic dissipation in the Earth's core, any electric currents and the associated magnetic field would vanish in about 15,000 years. Regeneration of the field is necessary. In the Earth it is thought that the magnetic field is maintained by dynamo action, whereby the kinetic energy of convective motion in the Earth's liquid core is converted into magnetic energy. Since this process operates without an external energy source, the geodynamo is said to be self-sustaining. *See* GEOELECTRICITY; GEOMAGNETISM; PALEOMAGNETISM.

Many issues in geodynamo theory remain unresolved, since the nonlinear equations of magnetohydrodynamics are difficult to solve. These equations, describing the processes operating in the core, are akin to the equations of oceanography and meteorology, but with the additional complication presented by the magnetic field through the Lorentz force (magnetomotive force). In the laboratory, mechanical dynamos—moving machines consisting of particular arrangements of metal—have been made since the days of Michael Faraday. Still, it is not obvious how a simply connected conducting fluid body, like the Earth's core, functions as a dynamo without the induced currents simply short-circuiting and eliminating field generation. In fact, the electric current in a dynamo and the magnetic field that it sustains cannot be too simple; a theorem, due to T. G. Cowling, says that no axisymmetric, or even two-dimensional, dynamo magnetic field can exist. Although the magnetic north and south poles usually are nearly coincident with the geographic poles, indicating that the rotation arising from the Coriolis force plays an important role in the core's dynamics, it is no accident that the compass does not point toward true north everywhere on the Earth's surface—an inherent lack of symmetry. As a result, theoretical progress has been slow since scientists often take advantage of symmetry, should it be present, when solving mathematical equations. *See* CORIOLIS ACCELERATION; MAGNETOHYDRODYNAMICS.

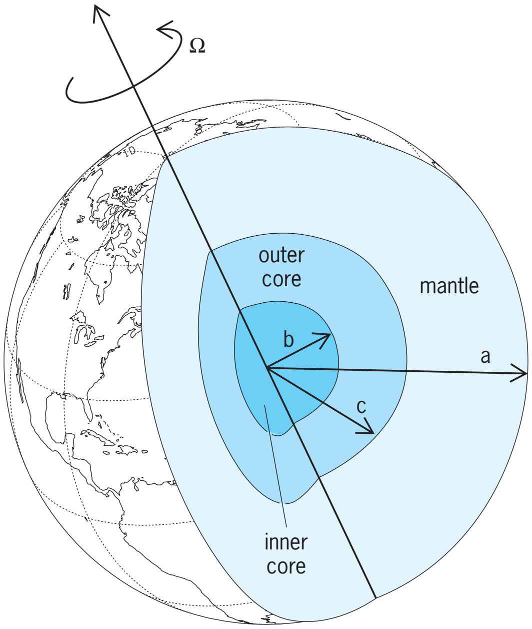


Fig. 1. Anatomy of the Earth. The rocky mantle has a radius *a* = 6371 km (3959 mi), the liquid iron outer core has a radius *c* = 3485 km (2165 mi), and the solid inner core has a radius *b* = 1215 km (755 mi). The Earth's rotational vector is Ω.

Geophysicists do, however, have a good qualitative understanding of how the geodynamo works. In the 1940s and 1950s, W. M. Elsasser and E. N. Parker first elucidated the so-called α-ω (alpha-omega) mechanism, by which core fluid motion can act as a dynamo if it consists of a combination of differential rotation and convective helical motion (**Fig. 2**). Since then it has been shown mathematically that dynamo regeneration can arise from the turbulent motion of a rotating fluid. Although the α-ω mechanism probably describes how the field is amplified, it is the dynamics that ultimately governs the strength of the field, which would grow until a rough balance between the Coriolis force and the Lorentz forces is attained.

With respect to the energy sustaining the fluid convection in the outer core, there are two possible sources of buoyancy—thermal and compositional. Thermal convection is perhaps most familiar, with heat sources, such as radioactive potassium, distributed over the volume of the outer core. With sufficient internal heating, the fluid is gravitationally unstable and, as a result, convection is sustained. Compositional convection is currently favored by most geophysicists as the energy source of the geodynamo. Although the core is primarily of iron, there are probably light impurities, such as sulfur. Due to the effects of pressure, as the Earth slowly cools, iron solidifies at the inner-core boundary. This causes the inner core to grow and leaves the lighter constituents behind in the fluid at the base of the outer core, supplying the buoyancy that drives the convection. *See* CONVECTION (HEAT).

Although dynamo action is an inherently asymmetric process, researchers are currently trying to resolve whether or not the the time-averaged morphology of the geomagnetic field is asymmetric, showing either persistent longitudinal variation or asymmetry under reflection through the equatorial plane. It is an issue of importance because persistent asymmetry in the geometric form of the field would indicate that the core and mantle are dynamically coupled. Such a possibility was first proposed by R. Hide, who was inspired by the Taylor-Proudman theorem of rotating

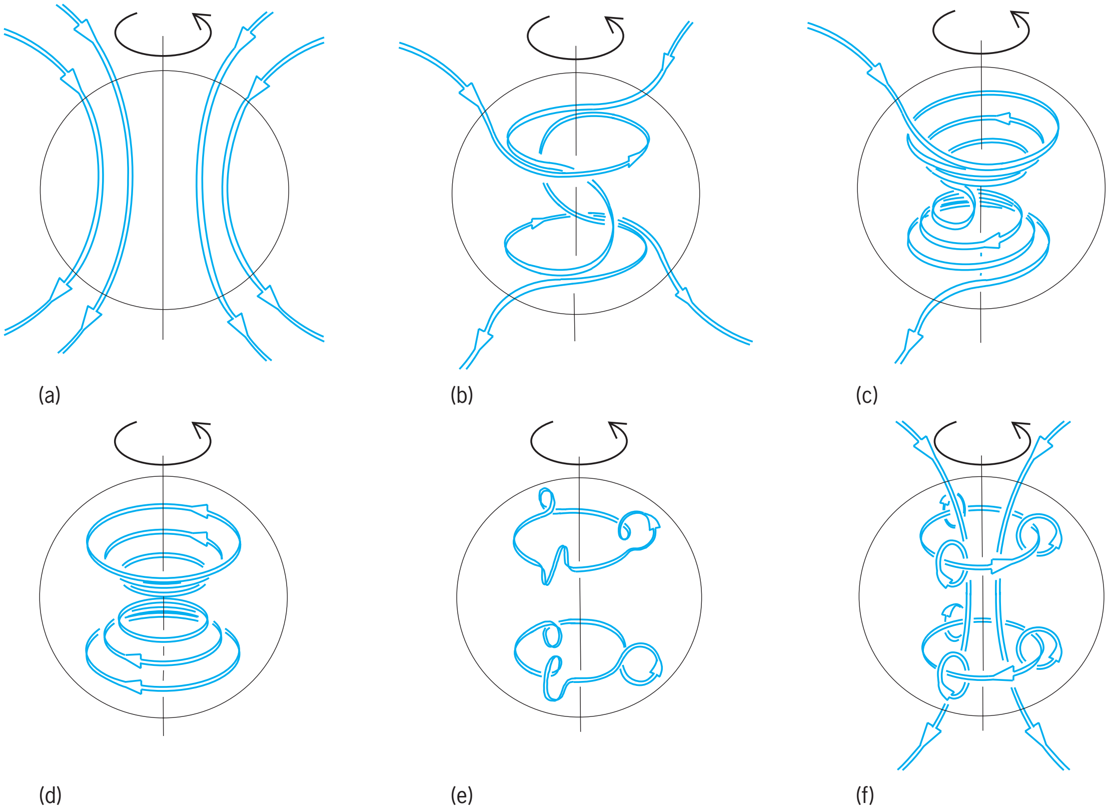


Fig. 2. The α-ω dynamo mechanism. Conventional geodynamo theory presupposes (*a*) an initial, primarily dipolar, poloidal magnetic field. The ω-effect consists of (*b*, *c*) differential rotation, wrapping the magnetic field around the rotational axis, thereby creating (*d*) a quadrupolar toroidal magnetic field. Symmetry is broken, and dynamo action maintained, by the α-effect, whereby (*e*) helical upwelling creates loops of magnetic field. (*f*) These loops coalesce to reinforce the original dipolar field, thus closing the dynamo cycle. (*After J. J. Love, Reversals and excursions of the geodynamo, Astron. Geophys., 40:6.14–6.19, 1999*)

fluid mechanics which states that small bumps on the boundary of a container of fluid can have a dramatic effect on the fluid's motion, even far from the boundary itself. Since the mantle convects very slowly compared to the core (their overturn times are approximately 10^8 and 10^3 years, respectively), the relatively steady nature of the conditions established by the mantle at the core-mantle boundary could also affect the pattern of core flow, thereby producing persistent asymmetric features in the magnetic field. Coupling between the core and mantle could arise from topography on the core-mantle boundary, or from thermal or electrical conductivity variations at the base of the mantle. *See* BOUNDARY LAYER FLOW; FLUID MECHANICS.

With an understanding of the boundary conditions applicable to the core, it should, in principle, be possible to solve the nonlinear magnetohydrodynamic equations that govern the geodynamo. The dynamo equations describe deterministic chaos, and data show that the magnetic field can exhibit highly aperiodic variation. Interestingly, the equations are symmetric under change in sign of the magnetic field, and thus there is no reason to expect that the Earth's field should have one polarity or the other. Moreover, a geomagnetic polarity reversal can occur with only a slight change in core's fluid motion. With respect to the cause of polarity transitions, mathematical models of hypothetical magneto-mechanical systems, in particular the Rikitake disc-dynamo system, exhibit irregular reversing behavior. So, it is not necessary to invoke external random factors such as ice ages or meteorite impacts to account for the irregular occurrence of polarity reversals.

The Sun is a familiar dynamo, and it reverses regularly almost every 11 years. So, why does the Earth's magnetic field not display such regularity? The difference is thought to be due to the presence in the Earth of a solid electrically conducting inner core, where the magnetic field can change only rather slowly by diffusion. Recent calculations suggest that because the inner core is electromagnetically coupled to the outer core, its presence acts to stabilize the magnetic field, so that only particularly large fluctuations of the field in the outer core are sufficient to overcome the damping effect of the inner core. *See* DIFFUSION; ELECTROMAGNETISM; GEOPHYSICS; MAGNETIC REVERSALS.

J. J. Love

Bibliography. F. H. Busse, Recent developments in the dynamo theory of planetary magnetism, *Annu. Rev. Earth Planet Sci.*, 11:241–268, 1983; J. A. Jacobs, *Deep Interior of the Earth*, Chapman & Hall, New York, 1994; J. J. Love, Reversals and excursions of the geodynamo, *Astron. Geophys.*, 40:6.14–6.19, 1999; E. N. Parker, Magnetic fields in the cosmos, *Sci. Amer.*, 249(2):44–54, 1983.

Geoelectricity

Electromagnetic phenomena and electric currents, mostly of natural origin, that are associated with the Earth. Geophysical methods utilize natural and artificial electric currents to explore the properties of the Earth's interior and to search for natural resources (for example, petroleum, water, and minerals). Geoelectricity is sometimes known as terrestrial electricity. All electric currents (natural or artificial, local or worldwide) in the solid Earth are characterized as earth currents. The term telluric currents is reserved for the natural, worldwide electric currents whose origins are almost entirely outside the atmosphere. Geoelectromagnetism is a more comprehensive term than geoelectricity. Time variations of any magnetic field are associated with an electric field that induces electric currents in conducting media such as the Earth.

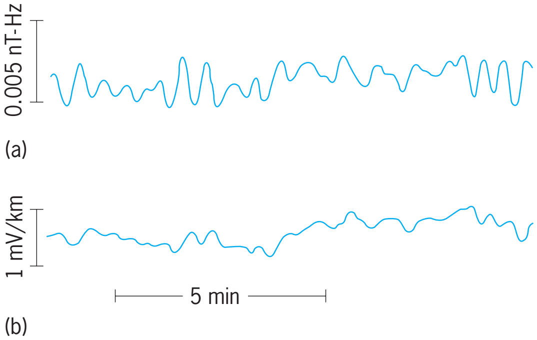


Fig. 1. Time variations of the horizontal orthogonal components of the natural (*a*) magnetic and (*b*) electric fields, simultaneously measured at one site at the surface.

Magnetic fields, electric fields, and electric currents are the constituents of electromagnetism, and are related by Maxwell's equations. For instance, **Fig. 1** shows the time variations of the natural magnetic and electric fields simultaneously measured at one location at the surface of the Earth. These two traces are related to each other, not only by Maxwell's equations but also by the physical properties of the subsurface rocks in the vicinity of the measuring site. Either one of the two traces may be computed synthetically from the other if the properties of the subsurface rocks are known. Conversely, the two traces together can yield geologic information; this is a form of geophysical exploration or prospecting. Thus, the terms geoelectricity, geomagnetism, and geoelectromagnetism are essentially interchangeable, although each one may have a somewhat different emphasis. For example, the term geomagnetism is sometimes used for the study of the Earth's quasi-stationary main magnetic field. *See* GEOMAGNETISM.

Measurements of electric and magnetic fields. A component of the electric field in a desired direction is measured by planting two electrodes (for example, metal stakes or special nonpolarizable electrodes) aligned in that direction. The electrodes are connected by an insulated wire, and voltage difference between them is measured with a voltmeter of high input impedance (for example, 10 megohms). The average electric field between the electrodes is expressed in units of volts per meter. Since this unit is very cumbersome for measuring the Earth's field, it is customary to use millivolts per kilometer. (Figure 1*b* and **Fig. 2** show the time variations of such components.) To obtain the total horizontal electric field, two orthogonal components, north–south (N–S) and east–west (E–W), are measured by means of an L-shaped electrode array. The trajectory of the head of the electric field vector is traced by feeding the two components into an oscilloscope or a paper X–Y recorder (**Fig. 3**). The magnetic field is measured by magnetometers. The cryogenic magnetometer has a resolution of better than 1 picotesla, 1 part in 50,000,000 of the Earth's total magnetic field. The nanotesla (nT) or gamma γ is used in practice. Figure 1*a* shows the time variations of one horizontal component of the Earth's natural magnetic field, measured with a coil-core magnetometer whose output is the time derivative of the magnetic field, with the scale given in terms of nanoteslas times frequency. Worldwide studies of natural electromagnetic phenomena are made by monitoring primarily the magnetic field rather than the electric field, which is much more affected by local geology.

Electric earth currents. These may be local or worldwide.

Local. Such currents can be natural or caused by human activities. The latter (called stray, industrial, or cultural currents) may be caused by electric trains, rural water pumps, and pipelines. Natural local

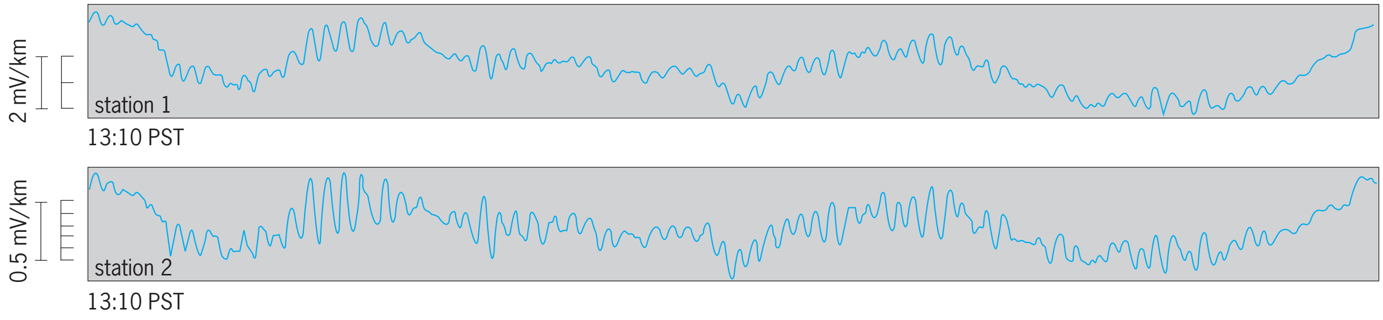


Fig. 2. Two tellurograms (stations 1 and 2, San Joaquin Valley, California) representing the time variations of the natural electric field (micropulsations), N60°E components, simultaneously recorded over a time interval of 30 min. The recording sites are separated by 27 mi (43 km) in the direction of the components. PST = Pacific Standard Time.

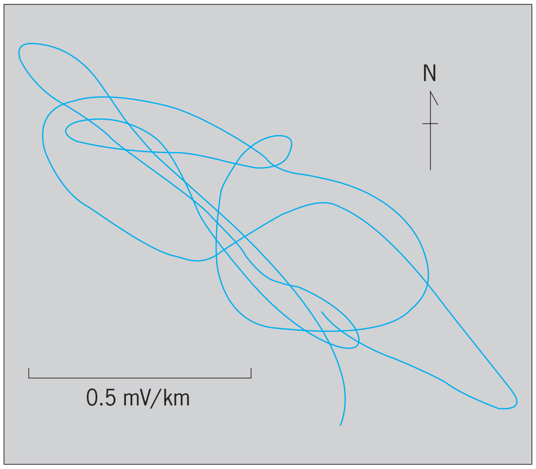


Fig. 3. A vectogram representing a few minutes of recording of the natural electric field vector. The band-pass filter peaked at the 20-s period (0.05 Hz).

currents represent the phenomena of spontaneous potentials or self-potentials. Some deposits in the Earth, such as certain metallic sulfides and graphite, constitute buried natural electric cells because of their high electrical conductivity and also because of oxidation and reduction processes associated with ground water. Thus, a hidden ore body, such as a copper ore deposit, can be discovered by measuring the electric field at the surface of the Earth, which may be as large as 1 V over a distance of 300 ft (100 m). Two other sources of spontaneous potentials are ground water movements and topographic elevation changes.

Worldwide. Telluric currents are of natural origin. There are various types, sources, and frequencies (or periods) of the worldwide natural electromagnetic fields which are associated with electric currents in the Earth (**Fig. 4**). The time variations of these electromagnetic fields are simply called variations.

Secular variations. The Earth's main magnetic field is thought to be caused by motions in the electrically conducting fluid core of the Earth, which acts as a kind of dynamo, creating electric currents which in turn create the magnetic field. This field is not stationary, but has time variations with periods ranging from about 30 to 300 years per cycle, which are the secular variations. Electric currents at the surface of the Earth associated with the main field and its secular variations have not been monitored effectively because of the difficulties involved in separating them from other effects, such as electrode potentials and tidal potentials. *See* GEOMAGNETIC VARIATIONS.

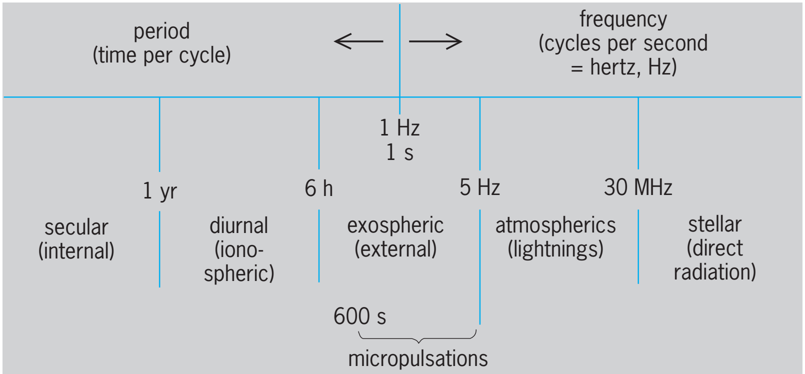


Fig. 4. Approximate and schematic frequencies and origins of the natural electromagnetic fields.

Diurnal (daily) variations. The air layers of the ionosphere, from a height of about 60–200 mi (100–300 km), are ionized by solar radiation, while air below the ionosphere is practically nonconducting. The ionization (electrical conductivity) in the ionosphere is renewed daily. Tidal oscillations of the ionosphere in the presence of the Earth's main magnetic field constitute an atmospheric dynamo, inducing electric currents in the Earth. They are thought to be driven primarily by the thermal effects of the Sun and partially by the attraction of the Moon (**Fig. 5**). *See* IONOSPHERE.

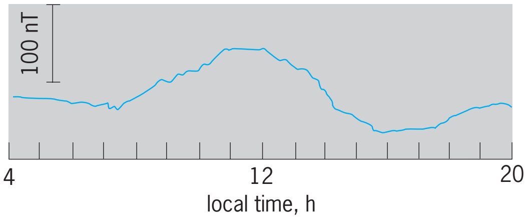


Fig. 5. Diurnal variations (solar plus lunar) of the horizontal component of the magnetic field, December 21, 1933, Huancayo, Peru. (*After S. Chapman and J. Bartels, Geomagnetism. 2 vols., Oxford University Press, 1962*)

Exospheric-origin variations, or micropulsations. Short time fluctuations of the Earth's magnetic field (micropulsations) that fall within the approximate period range of 0.2–600 s per cycle (5–0.0017 Hz) occur almost continuously as a background noise. Amplitudes depend on latitude, solar activity, frequency, local time, season, and local geology, with worldwide and long-term statistical amplitudes of the order of a few millivolts per kilometer and a few tenths of a nanotesla. While the mechanism of their generation is not completely understood, it appears that micropulsations are generated by the magnetohydrodynamic effect through the interaction of the solar wind with the main magnetic field and atmosphere of the Earth. Study of the exospheric-origin electromagnetic phenomena constitutes a branch of geophysics called aeronomy. Figure 1 is a record of micropulsations measured at stations located in a sedimentary basin. The magnetic field trace (Fig. 1*a*) is called a magnetogram; the electric field trace (Fig. 1*b*) a tellurogram. Figure 2 shows two tellurograms simultaneously measured at two stations 27 mi (43 km) apart and in the direction of station separation. These measurements represent normal, usual activity on a quiet day. **Figure 6** shows the amplitude spectra of the tellurograms shown in Fig. 2. The differences between the two tellurograms, and consequently between the two spectra,

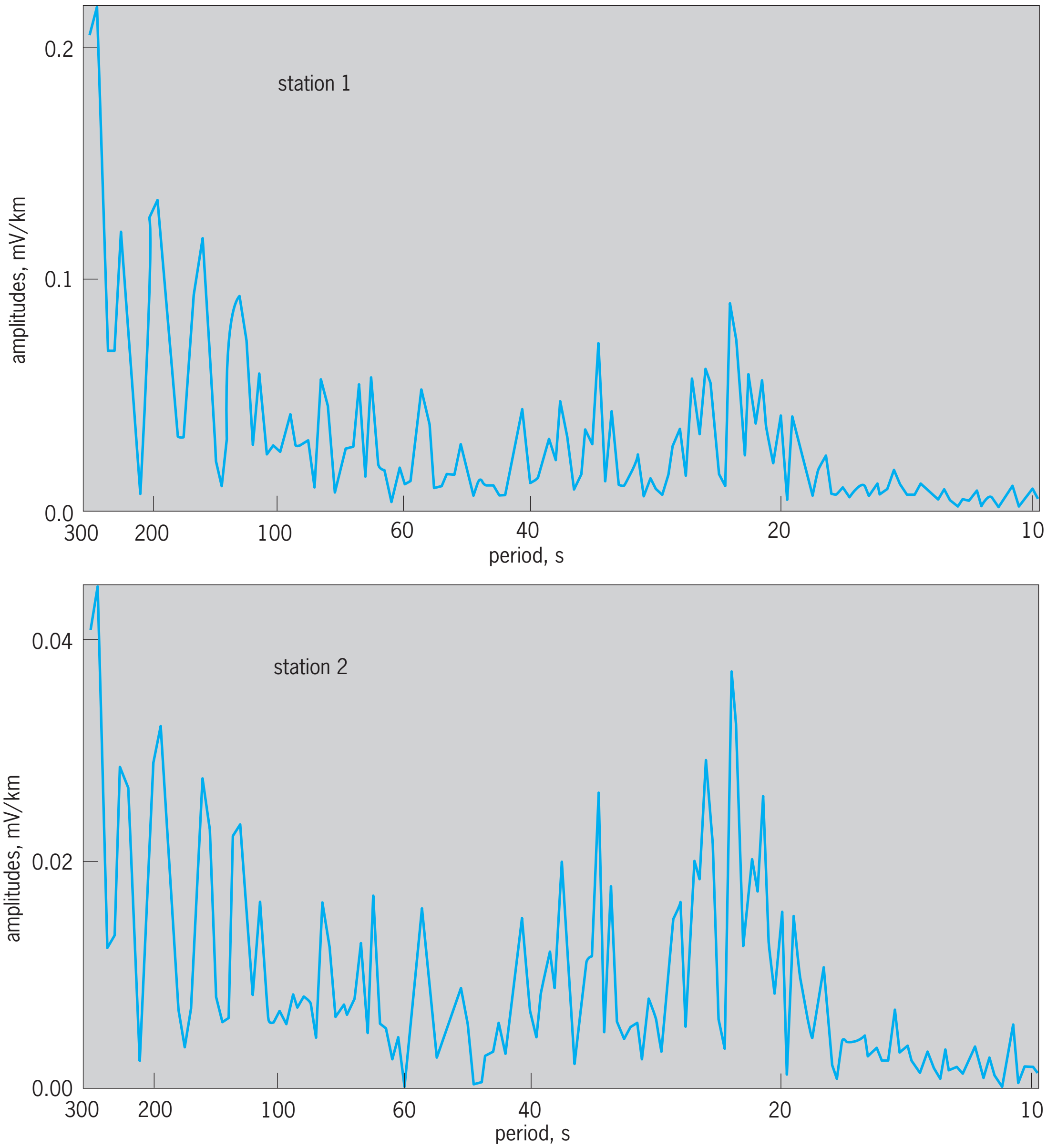


Fig. 6. Amplitude spectra of the tellurograms shown in Fig. 2.

are almost totally due to the differences in the geologic conditions at the two measuring sites (stations). Such measurements can be used for geologic exploration of the subsurface. *See* MAGNETOHYDRODYNAMICS; SEISMIC STRATIGRAPHY; SOLAR WIND; UPPER-ATMOSPHERE DYNAMICS.

Magnetic storms are very intense disturbances of long duration that occur about once a month on the average. Caused by large-scale bursts of solar wind associated with sunspots and solar flares, they usually commence suddenly and almost instantaneously (within about 0.5 min) throughout the world. Their amplitudes may reach hundreds of nanoteslas and hundreds of millivolts per kilometer, disrupting radio and telegraph communications. It is interesting to note that they cause fish to migrate into deeper waters. **Figure 7** shows the records of a magnetic storm. Magnetic storms are frequently associated with aurorae polares (northern or southern lights), which are seen as spectacular luminous formations at ionospheric heights. *See* AURORA.

Atmospherics. The major cause of the variations within the frequency range of about 5–10 kHz is the lightning occurring almost continuously in Central Africa and in the Amazon region. While audiofrequency variations are included in atmospherics, lightning itself is a concern of meteorology. *See* LIGHTNING; SFERICS.

Stellar variations. Above the frequency of 30 MHz, these originate predominantly from the direct radiation of electromagnetic waves propagated by the Sun.

Subsurface geophysical exploration. Electrical methods, more properly called electromagnetic methods, are used to explore the subsurface from depths of a few inches (for example, popular coin detectors

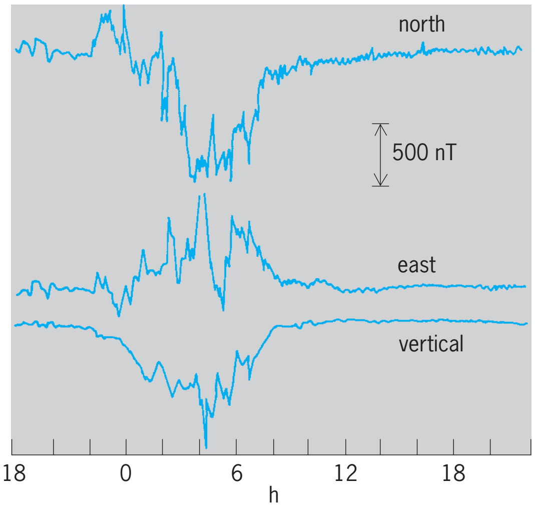


Fig. 7. Three components of the magnetic field of a magnetic storm of May 14, 1921, Potsdam, Germany. (*After S. Chapman and J. Bartels, Geomagnetism, 2 vols., Oxford University Press, 1962*)

or mine detectors) down to depths of hundreds of miles. In general, these methods require an input into the Earth, either an artificial direct or alternating electric current, or a natural electromagnetic field, such as micropulsations or diurnal variations. This input is a source signal coupled with the Earth, which behaves as a filter whose response is measured in terms of electric or magnetic fields. It is analogous to measuring the input and output of an electronic filter to determine its characteristics, which in this case is the geologic information sought. These methods supply only the electrical properties of the subsurface (mainly the electrical conductivity). Different rocks have, in general, different conductivities. For instance, limestones usually have much lower conductivities than clay-rich shales. A knowledge of the conductivity distribution in the subsurface, combined with other geologic information, allows interpretation of the rock-type distribution. Artificial direct-current methods involve feeding a current into the Earth with a pair of electrodes and measuring the resulting electric field with another pair of electrodes. The alternating-current methods use magnetometers to measure the magnetic field created by inducing currents in the Earth. The two most popular methods employing the natural electromagnetic fields are the magnetotelluric and telluric methods. The magnetotelluric method requires simultaneous measurements of the electric and magnetic fields at one site. Figure 1 represents such a data set. The telluric method requires only the measurements of the electric field, made simultaneously at two or more sites (Fig. 2). These electromagnetic (or electrical) methods are unlike the magnetic methods of geophysical exploration, which deal only with the magnetization of rocks due to the main magnetic field of the Earth. *See* GEOPHYSICAL EXPLORATION.

S. H. Yungul

Bibliography. G. V. Keller and F. C. Frischknecht, *Electrical Methods in Geophysical Prospecting*, 1966; S. H. Yungul, The telluric methods in the study of sedimentary structures: A survey, *Geoexploration*, 15:207–238, 1977; M. S. Zhdanov and G. V. Keller, *The Geoelectrical Methods in Geophysical Exploration*, 1993.

Geographic information systems

Computer-based technologies for the storage, manipulation, and analysis of geographically referenced information. Attribute information and spatial information are integrated in geographic information systems (GIS) through the notion of a data layer, which is realized in two basic data models—raster and vector (**Fig. 1**). The major categories of applications comprise urban and environmental inventory and management, policy decision support, and planning; public utility and business (including agribusiness) applications, engineering, and defense applications; and scientific analysis and modeling.

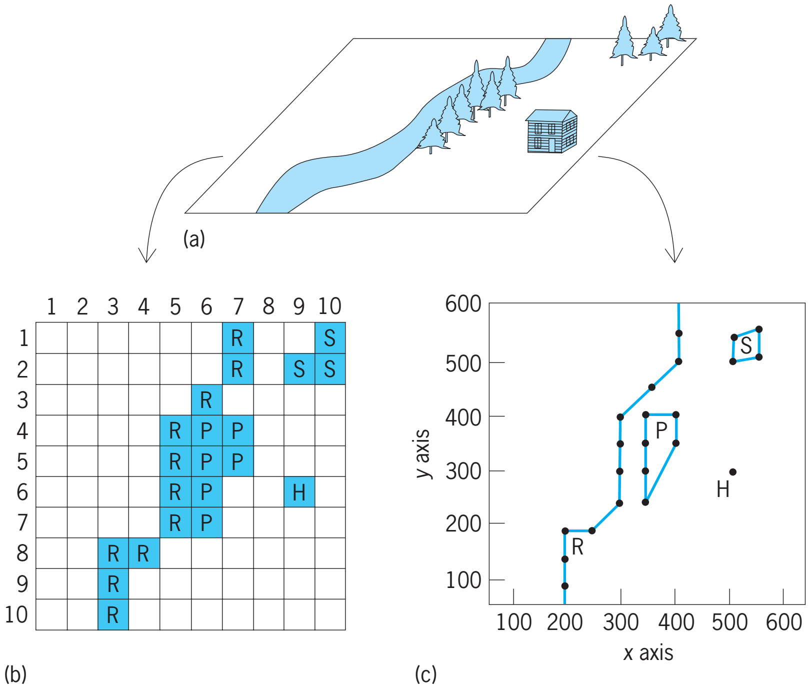
Characteristics. A geographic information system differs from other computerized information systems in two major respects. First, its information is geographically referenced (geocoded). Geocoding is usually achieved by recording the geographical coordinates of the objects of interest (which may be the corner points of a plot on a cadastral map, or the location of cities of more than 100,000 inhabitants at the global scale). Alternatively, the location of any point of interest within an area can be inferred on the basis of a number of reference points registered for that area. Second, a geographic information system has considerable capabilities for data analysis and scientific modeling, in addition to the usual data input, storage, retrieval, and output functions.

A geographic information system can answer arbitrarily complex queries about things on or near the surface of the Earth, their attributes, and the spatial relationships (such as distance, direction, adjacency, and inclusion) among them. Basic kinds of queries supported by this system include (1) location questions, to determine the attributes of a given place (for example, What tree types are found in a given forestry tract?); (2) condition questions, seeking to find locations fulfilling certain conditions (for example, Where are vacant lots larger than 5 acres and within 1 mile of a paved road?); (3) trend questions, seeking to determine changes in place attributes over time (for example, How much has the population of a specific census tract grown between 1980 and 1990?); (4) routing questions, especially useful in situations where vehicles need to be dispatched from a place of origin to a destination (for example, Which is the shortest, quickest, or safest route from an ambulance station to the site of an emergency?); and (5) pattern questions, whereby scientists or managers can investigate the spatial distribution of some phenomenon for diagnostic purposes or in the course of exploring some scientific hypothesis (for example, Is the density of diseased trees greater around campgrounds? Are pedestrian traffic casualties higher than expected in low-income neighborhoods?).

Elements. A geographic information system is composed of software, hardware, and data. Some authors also include the users and their institutional context. Originally used on mainframe computers, geographic information systems did not become popular until the 1980s, when software packages that could run on workstations and desktops became widely available. Some of the most powerful systems now run on minicomputers, but the field is dominated by packages developed for workstation platforms in the case of the larger systems and desktops for the rest. The trend in the late 1990s was toward flexible, portable, user-friendly systems, increasingly capable of interoperating, increasingly geared toward the Internet, and in many cases tailored to specific kinds of applications. Often these special-purpose geographic information systems are coupled with other digital systems or devices such as global positioning systems (GPS); intelligent vehicle highway systems (IVHS); navigational devices for vehicles, ships, or aircraft; or devices for the automatic delivery of agricultural fertilizers as a function of soil quality. *See* HIGHWAY ENGINEERING.

The notion of data layer (or coverage) and overlay operation lies at the heart of most software designed for geographic information systems. The landscape is viewed as a collection of superimposed elementary maps, each storing information pertinent to one aspect or attribute of the landscape: relief, soils, hydrology, vegetation, roads, land use, land ownership, and so forth. By combining the corresponding information on several layers, the composite properties of any object of interest can be deduced. (This approach is the electronic version of a traditional manual method involving transparent sheets.) Two fundamental data models, the vector and raster models, embody the overlay idea in geographic information systems (**Fig. 2**). In a vector geographic information system, the geometrical configuration of a coverage is stored in the form of points, arcs (line segments), and polygons, which constitute identifiable objects in the database. In a raster geographic information system, a layer is composed of an array of elementary cells or pixels, each holding an attribute value without explicit reference to the geographic feature of which the pixel is a part. Both models have strengths and drawbacks. *See* COMPUTER GRAPHICS; ELECTRONIC DISPLAY; IMAGE PROCESSING.

A data layer or coverage integrates two kinds of information: attribute and spatial (geographic). The functionality of a geographic information system consists of the ways in which that information may be captured, stored, manipulated, analyzed, and presented to the user. Spatial data capture (input) may be from primary sources such as remote sensing scanners, radar, or global positioning systems, or from scanning or digitizing images and maps derived from remote sensing. Output (whether as a display on a cathode-ray tube or as hard copy) is usually in map or graph form, accompanied by tables and reports linking spatial and attribute data. The critical data management and analysis functions fall into four categories: retrieval, classification, and measurement (for example, measurement of the area inside a polygon); overlay functions; neighborhood operations; and connectivity functions. *See* DATABASE MANAGEMENT SYSTEM; DIGITAL COMPUTER; REMOTE SENSING.

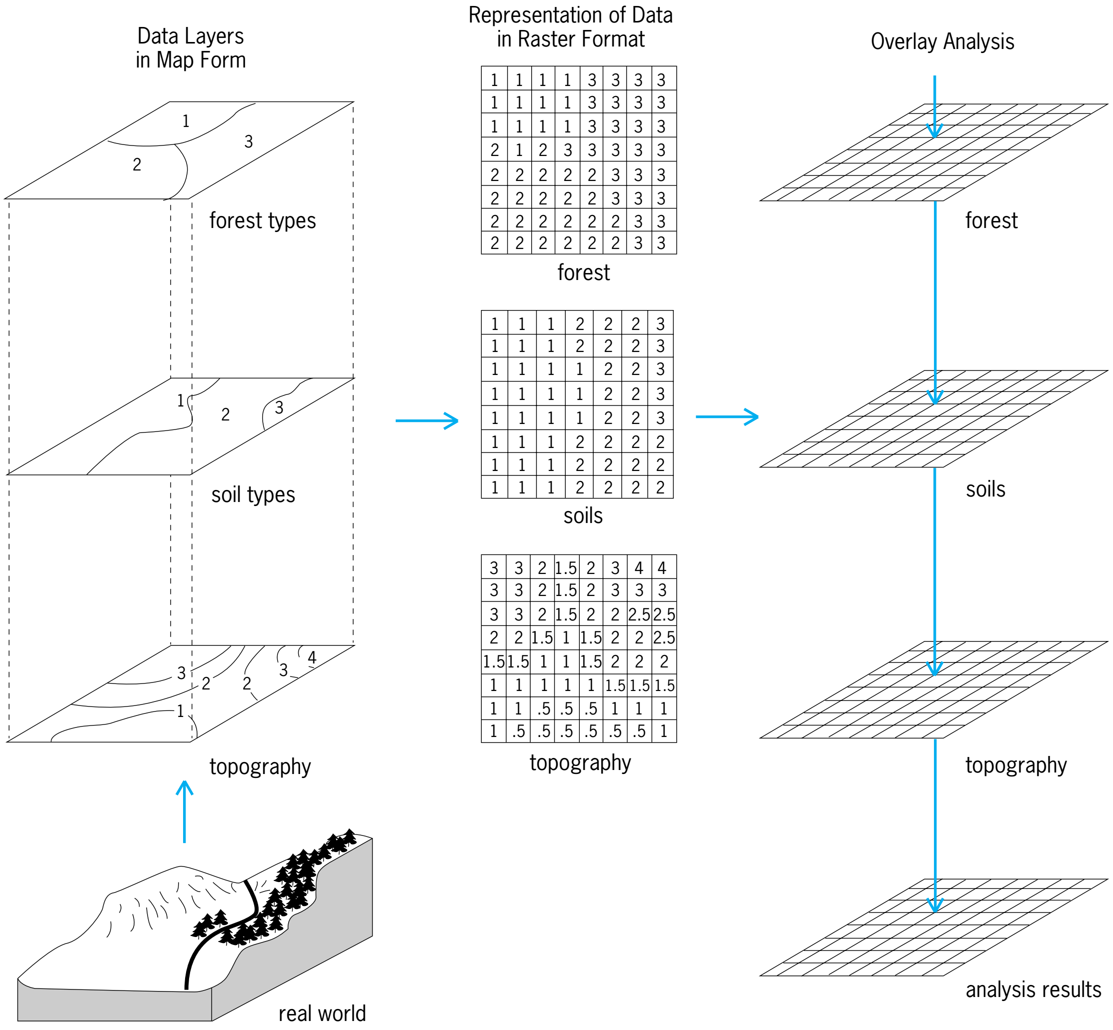


	1	2	3	4	5	6	7	8	9	10
1							R			S
2							R		S	S
3						R				
4					R	P	P			
5					R	P	P			
6					R	P			H	
7					R	P				
8			R	R						
9			R							
10			R							

Fig. 1. Comparison of the raster and vector models. The pine forest stand (P) and spruce forest stand (S) are area features. The river (R) is a line feature, and the house (H) is a point feature. (*a*) Landscape. (*b*) Raster representation. (c) Vector representation. (*After S. Aronoff, Geographic Information Systems: A Management Perspective, WDL Publications, 1989*)

Uses. Geographic information systems find application mainly in three areas: business, engineering, and defense; management, planning, and policy; and research in various disciplines.

Business, engineering, and defense. Business applications are increasingly widespread and include market analysis (for example, identifying a customer base in a given area), store location (for example, identifying the most competitive location for a new chain store outlet), and agribusiness (for example, determining the correct amount of fertilizers or pesticides needed at each point of a cultivated field). Engineers use geographic information systems when modeling terrain, building roads and bridges, maintaining cadastral maps, routing vehicles, drilling for water, determining what is visible from any point on the terrain, integrating intelligence information on enemy targets, and so forth. Such applications have been facilitated through the integration of geographic information systems with global positioning systems, involving the automatic determination of latitude and longitude coordinates and elevation based on the processing of signals from a network of satellites. Because global positioning systems allow the determination of position with unparalleled ease, speed, and accuracy, mobile units coupling geographic



1	1	1	1	3	3	3	3
1	1	1	1	3	3	3	3
1	1	1	1	3	3	3	3
2	1	2	3	3	3	3	3
2	2	2	2	2	3	3	3
2	2	2	2	2	3	3	3
2	2	2	2	2	2	3	3
2	2	2	2	2	2	3	3

1	1	1	2	2	2	3
1	1	1	2	2	2	3
1	1	1	1	2	2	3
1	1	1	1	2	2	3
1	1	1	1	2	2	3
1	1	1	2	2	2	2
1	1	1	2	2	2	2
1	1	1	2	2	2	2

3	3	2	1.5	2	3	4	4
3	3	2	1.5	2	3	3	3
3	3	2	1.5	2	2	2.5	2.5
2	2	1.5	1	1.5	2	2	2.5
1.5	1.5	1	1	1.5	2	2	2
1	1	1	1	1	1.5	1.5	1.5
1	1	.5	.5	.5	1	1	1
1	.5	.5	.5	.5	.5	.5	1

Fig. 2. Overlay analysis using raster data files. (*After S. Aronoff, Geographic information Systems: A Management Perspective, WDL Publications, 1989*)

information and global positioning systems have greatly increased the range of applications, especially for work in the field and navigation. *See* COORDINATE SYSTEMS; SATELLITE NAVIGATION SYSTEMS.

Management, planning, and policy. There are many management applications of geographic information systems. Among the earliest and still most widespread applications of the technology are land information and resource management systems (for example, forest and utility management). Other common uses of geographic information systems in an urban policy context include emergency planning, determination of optimal locations for fire stations and other public services, assistance in crime control and documentation, and electoral and school redistricting.

Nonurban applications include ecological area management (for example, estuarine or National Parks management), hazardous waste facility location and management, and biodiversity preservation projects. A significant development is the introduction of geographic information systems in third-world environmental management and planning. This development was made possible in the 1980s through the wide availability of inexpensive desktop computer platforms for geographic information systems, by the increasing ease of use of such computers, and by the diffusion of educational opportunities in geographic information. Major applications in developing countries include natural resource management, soil conservation and irrigation projects, and the siting of health and other services.

More sophisticated developments involve a new class of software known as spatial decision support systems, which can aid the policy-making and decision process whenever location is part of the problem context. A typical system of this type presents the different options, facilitates the comparison of costs and benefits, helps explore the consequences of changing the weights attributed to different decision criteria, and allows unanticipated alternatives (for example, solutions proposed by third parties) to be easily formulated and evaluated. The power of spatial decision support systems lies primarily in their flexible, interactive nature and in the visualization of the complex issues debated. *See* DECISION SUPPORT SYSTEM.

Other applications. Uses of geographic information systems have spread well beyond geography, the source discipline, and now involve most applied sciences, both social and physical, that deal with spatial data. These sciences include sociology, archeology and anthropology, urban studies, epidemiology, ecology, forestry, hydrology, and geology. The nature of the applications of geographic information systems in these areas ranges from simple thematic mapping for illustration purposes to complex statistical and mathematical modeling for the exploration of hypotheses or the representation of dynamic processes (such as the spread of fire on a vegetated landscape or the investigation of trends in crime patterns in an urban area). *See* ANTHROPOLOGY; ARCHEOLOGY; ECOLOGY; EPIDEMIOLOGY; FOREST AND FORESTRY; GEOLOGY; HYDROLOGY.

Conceptual and scientific issues. While the geographic information system is still considered largely a technology, its coming of age has been marked by an increasing emphasis on its conceptual and scientific aspects. There are a number of fundamental research questions concerning geographic information systems. They involve (1) the nature of spatial data, including questions of measurement, sampling, uncertainty, error, scale, and the nature of geographic space itself (Is it a container of geographical objects, or is it defined through the relations among objects?); (2) the digital representation of geographic space and phenomena, and in particular questions of database structures, models, and semantics; (3) functional issues, including the appropriate basic operations the geographic information system should support, the propagation of errors and uncertainties in spatial databases, and the appropriate query languages; (4) display issues, such as questions of visualization, user interfaces, and cartographic design; and (5) operational issues, dealing with quality controls and standards for both the systems themselves and their products, with legal and management issues, and with any other issues affecting the fitness of geographic information systems for their numerous kinds of current and potential applications. A major recent research theme involving all these issues concerns the interoperability of geographic information systems, or the possibility to move or share data, functions, and operations across systems and platforms. Helen Couclelis

Bibliography. J. C. Antenucci et al., *Geographic Information Systems: A Guide to the Technology*, 1991; T. Bernhardsen, *Geographic Information Systems: An Introduction*, 3d ed., 2002; P. A. Burrough and R. A. McDonnell, *Principles of Geographical Information Systems*, 2d ed., 1998; M. N. DeMers, *Fundamentals of Geographic Information Systems*, 3d ed., 2002; M. Fischer and P. Nijkamp (eds.), *Geographic Information Systems: Spatial Modeling and Policy Evaluation*, 1993; R. C. Frohn, *Remote Sensing for Landscape Ecology: New Metric Indicators for Monitoring, Modeling, and Assessment of Ecosystems*, 1998; P. Longley et al., *Geographical Information Systems: Principles and Applications*, 2d ed., 1998.

Geography

The study of physical and human landscapes, the processes that affect them, how and why they change over time, and how and why they vary spatially. Geographers determine the factors that influence landscapes or landscape features and seek to explain the interactions among them. This often involves explaining landscapes or environments from a spatial perspective by analyzing patterns or the lack of patterns. Geographers are also interested in the spatial arrangements of natural and human phenomena, as well as the relationships among these phenomena. Consequently, a key element of geography is areal differentiation, the process of defining regions with common unifying characteristics.

Geography is a broad-based discipline that integrates many aspects of the social and physical sciences. The discipline has been described as consisting of four traditions: Earth science, area studies, spatial analysis, and human–environment interactions. Earth science includes studies of the atmosphere, hydrosphere, lithosphere, and biosphere. Area studies refer to the detailed geographic knowledge and understanding of particular regions of the world. Spatial analysis deals with the measurement, analysis, prediction, and explanation of spatial organization, as well as the interactions, patterns, and factors that influence change in these landscape elements. Human–environment interaction is concerned with the relationships between humans and their environment. Generally, this includes the ways that humans modify and influence their physical environment to meet certain needs, as well as the ways that humans modify their behavior and surroundings to adapt to particular environmental circumstances. A majority of studies in geography concern at least one of these four traditions. *See* ATMOSPHERE; BIOSPHERE; HYDROSPHERE; LITHOSPHERE.

Geographers consider, to varying degrees, both natural and human influences on the landscape, although a common division separates physical and human geography. Physical geographers study landforms (geomorphology), water (hydrology), climate and meteorology (climatology), biotic environments (biogeography), and soils (pedology). Human geographers focus on cultural geography, economic geography, political geography, location analysis, and spatial analysis of ethnic or gender issues, as well as transportation, area studies, and urban, regional, or environmental planning. *See* BIOGEOGRAPHY; CLIMATOLOGY; GEOMORPHOLOGY; HYDROLOGY; PEDOLOGY; PHYSICAL GEOGRAPHY.

Most professional geographers specialize in one or two subdisciplines. In addition, many are involved in developing techniques and applications that support spatial analytical studies or the display of spatial information and data. Maps—printed, digital, or conceptual—are the basic tools of geography. Geographers are involved in map interpretation and use, as well as map production and design. Cartographers supervise the compilation, design, and development of maps, globes, and other graphic representations. *See* CARTOGRAPHY.

The techniques, tools, and methods used in geographic and spatial analytical work have undergone a revolution since the advent of relatively inexpensive computer graphics systems and technologies such as satellite imaging. Remote sensing specialists, often geographers, work to interpret and enhance digital and photographic images, particularly those gathered from aircraft or satellites. A geographic information system (GIS) combines the advantages of computer-assisted cartography with those of spatial database management, and has widespread applications. A geographic information system facilitates the storage, retrieval, and analysis of spatial information in the form of digital map "overlays," each representing a different landscape component such as terrain, hydrologic features, roads, vegetation, soil types, or any mappable factor. Each of these data layers can be fitted digitally in any combination to the same map scale and map projection, permitting the analysis of relationships among combinations of environmental variables. *See* COMPUTER GRAPHICS; GEOGRAPHIC INFORMATION SYSTEMS; MAP PROJECTIONS; REMOTE SENSING; SCIENTIFIC AND APPLICATIONS SATELLITES.

Many geographers are applied practitioners, solving problems using a variety of tools, particularly computer-assisted cartography, statistical methods, remotely sensed imagery, the Global Positioning System (GPS), and geographic information systems. Today, nearly all geographers, regardless of their focus, use some or all of these techniques in their professional endeavors. *See* SATELLITE NAVIGATION SYSTEMS. James F. Petersen

Bibliography. R. Abler, M. Marcus, and J. Olson, *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*, 1992; G. Gaile and C. Willmott (eds.), *Geography in America at the Dawn of the 21st Century*, 2005; Geography Education Standards Project, *Geography for Life: National Geography Standards*, 1994; P. Gould, *Becoming a Geographer*, 1999; National Research Council, *Rediscovering Geography: New Relevance for Science and Society*, 1997.

Geologic thermometry

The measurement or estimation of temperatures at which geologic processes take place. Methods used can be divided into two groups, nonisotopic and isotopic. The isotopic methods involve the determination of distribution of isotopes of the lighter elements between pairs of compounds in equilibrium at various temperatures, and application of these data to problems of the temperature at which these compounds (commonly minerals) form in nature.

Nonisotopic Methods

Earth temperatures can be measured directly by surface and near-surface features or indirectly from various properties of minerals and fossils.

Direct measurement. Temperatures can be measured directly in hot springs, fumaroles, flows of lava, and artificial openings such as mines, boreholes, and wells.

Hot springs. The temperatures of hot springs range from slightly above the mean annual temperature of the region in which they occur to the boiling point of water at the elevation of the outlets. In other words, in temperate regions and at moderate altitudes the temperatures of hot springs range from about 20 to 100°C (68 to 212°F).

Fumaroles. At temperatures above its boiling point, water issues as steam from vents called fumaroles: temperatures of these fumaroles have been measured up to 560°C (1040°F) near Vesuvius, and up to 645°C (1193°F) in the Valley of Ten Thousand Smokes (Alaska). Fumaroles in lavas may reach temperatures of 700–800°C (1290–1470°F).

Lava. Lava is molten rock flowing onto the Earth's surface. Its temperature on extrusion ranges from about 700 to 900°C (1290 to 1650°F) for andesitic and dacitic lava and to 1200°C (2190°F) for basaltic lava. The viscosity of lava increases with decreasing temperature, and basaltic lava ceases to flow when it cools to 700–800°C (1290–1470°F). Intrusive magmas of similar compositions are probably intruded at similar temperatures, as indicated by their effects on the coal beds into which they are intruded, and the forms and assemblages of the first minerals to crystallize. *See* LAVA.

The temperature to which rocks around an intrusion (country rocks) are heated depends on many factors: temperature of the magma; temperature, composition, and structure of the country rock; abundance and nature of solutions given off by the intrusion; and size of the intrusion. In general, the temperature of the country rocks at the contact will be much lower than that of the magma. For example, an intrusive sheet of dolerite at 1100°C (2010°F) may heat the contact rocks to 600–700°C (1110–1290°F). *See* MAGMA.

Mines, boreholes, and wells. Temperatures have been measured in enough artificial openings in the Earth's crust so that the temperature distribution is well known in many areas to depths of several thousand feet. Measurements of gradients range from about 12 to 52 m/°C (22 to 94 ft/°F) in nonvolcanic areas. The highest temperature yet encountered in such an area is 154°C (309°F) from a well 20,521 ft (6255 m) deep in Sublette County, Wyoming.

Gradients in volcanic areas are much higher near the surface (up to 0.4 m/°C or 0.7 ft/°F), but at depths of 750–1000 ft (230–300 m) a temperature of about 250°C (482°F) is commonly reached and persists to much greater depths (to at least 5500 ft or 1.7 km in Tuscany).

Calculations and extrapolations give greatly different pictures for temperature distribution from the zone of measurements to the center of the Earth, depending on the assumptions made. Estimates of the temperature at the center of the Earth range from 1600 to 76,000°C (2900 to 137,000°F).

Indirect methods. Some indirect nonisotopic methods appear to give estimates with the same accuracy as direct measurements; others place the temperature within a certain range; still others indicate only that a given process took place above or below a certain temperature.

As phase-equilibrium relationships in systems analogous to those in nature become more accurately known, it will be possible to make more exact estimates of geologic temperatures. The relationships commonly employed are listed below. *See* PHASE EQUILIBRIUM.

Melting points. The melting point of a mineral, corrected for the pressure under which it was formed, gives a maximum temperature of formation for the assemblage in which it grew because other substances, especially water, lower the crystallizing temperature for a mineral. For example, if realgar, AsS, occurs in a vein with other minerals in such a way that they must have crystallized simultaneously, then the whole assemblage must have formed at a temperature lower than 320°C (608°F), the melting point of realgar.

Transformation temperatures (inversions). Many minerals have two or more crystalline modifications which form, or exist, in different temperature ranges. For example, under certain conditions marcasite forms at temperatures below 300°C (570°F), but at about 450°C (840°F) it transforms to pyrite at an appreciable rate. Therefore, a coprecipitated mineral assemblage including marcasite was certainly formed below 450°C (840°F) and probably below 300°C (570°F).

Other pairs of minerals transform in either direction at a definite temperature; for example, low (α) quartz changes to high (β) quartz when it is heated to 573°C (1060°F), and high quartz changes to low when it is cooled below 573°C (1060°F). Therefore, phenocrysts of high quartz in lavas were formed above 573°C (1060°F), crystals of low quartz in veins, below 573°C (1060°F).

Dissociation and decomposition temperatures. Many minerals break up when they are heated. If one of the products is a gas, the temperature of decomposition changes rapidly with pressure; the pressure at the time of formation must be known or estimated before such a mineral can be used as a geologic thermometer. For example, under atmospheric pressure, calcite ($CaCO_3$) dissociates into lime and carbon dioxide (CO_2) at 885°C (1620°F), but under a sufficiently high pressure of CO_2 (1025 atm or 104 megapascals) it melts at 1339°C (2442°F) without decomposing.

When only solids and liquids are involved, however, pressure is relatively unimportant and can be neglected for processes that take place near the Earth's surface. For example, danburite decomposes to two liquids at about 1000°C (1830°F), so this mineral and others immediately associated with it in pegmatities must have formed at temperatures below 1000°C (1830°F).

Solid solutions and exsolution pairs. Many pairs of minerals with similar structures form homogeneous solid solutions at high temperatures, but on cooling separate (exsolve) into lamellae of the two minerals. This process commonly produces a characteristic texture that can be recognized. When two such minerals occur in a rock or ore in an exsolution relationship, this is evidence that the temperature of formation was above their temperature of homogenization. *See* SOLID SOLUTION.

Some common exsolution pairs and their temperatures of homogenization are magnetite-spinel, 1000°C (1830°F); ilmenite-hematite, 600–700°C (1110–1290°F); chalcopyrite-bornite, 500°C (930°F); chalcopyrite-cubanite, 450°C (840°F); bornite-tetrahedrite, 275°C (527°F); and bornite-chalcocite, 225°C (437°F). These temperatures depend upon the composition of the host phase, but the ones given here are for compositions commonly encountered in nature.

This method has been used to estimate the temperatures of formation of many ore deposits. Some examples are sulfide replacement in Gilman, Colorado, 150–300°C (300–570°F); sulfide mineralization at Pine Vale, Queensland, 475–500°C (807–930°F); scheelite veins, Australia, 350–600°C (660–1110°F); and sphalerite-stannite-chalcopyrite mineralization, Tasmania, about 600°C (1110°F).

In some mineral pairs there is a limited amount of solid solution: the amount depends upon the temperature of formation. For example, in the system iron sulfide- zinc sulfide (FeS-ZnS) the amount of FeS in sphalerite is a function of the temperature of formation, if there is an excess of FeS (pyrrhotite) present when the sphalerite crystallizes. Sphalerite formed at 200°C (390°F) under these conditions will contain about 7 mol % FeS; at 500°C (930°F), 18 mol %. Other associations that can be used to indicate temperature of formation in this way are scandium in biotite (in a given petrologic province), titanium dioxide (TiO_2) in magnetite (with coexisting ilmenite), iron and magnesium in coexisting olivines and pyroxenes, albite in potassium feldspar (with coexisting plagioclase), and magnesium carbonate ($MgCO_3$) in calcite (with coexisting dolomite).

By using the FeS-ZnS relationship, the temperature of formation of sphalerite in various settings has been estimated: in graphite schist near a granite contact (Norway), 440°C (820°F); replacement deposits, Gilman, Colorado, 380–600°C (710–1110°F) and Broken Hill, Australia, 600°C (1110°F); and in uranium deposits, Colorado Plateau, 138°C (280°F). TiO_2 in the magnetite of the northwestern Adirondacks indicates a temperature of formation of 475–600°C (807–1110°F). The composition of alkali feldspars in granite indicates a final consolidation temperature of about 600°C (1110°F). A metamorphic calcite with 10% $MgCO_3$ forms at about 650°C (1200°F); with 20%, at 830°C (1520°F).

Eutectics. When two or more minerals crystallize simultaneously at a eutectic, a so-called eutectic texture may be produced. However, it is difficult to be certain that a natural intergrowth of minerals was produced by eutectic crystallization. Therefore, this method has been little used thus far in geologic thermometry. It can be used in the same way that melting points can, that is, to indicate a temperature that cannot have been exceeded during crystallization of the assemblage. Some eutectic temperatures of common minerals are iron-nickel, 1435°C (2615°F); anorthite-diopside, 1270°C (2320°F); albite-nephelite, 1068°C (1954°F); orthoclase-albite, 1070°C (1958°F); orthoclase-silica, 990°C (1814°F); quartz-orthoclase-albite, 937°C (1719°F); silver-copper, 785°C (1445°F); chalcocite-galena-argenite, 400°C (750°F); and sulfur-selenium, 100°C (212°F). Some of these temperatures are affected markedly by volatile components of a magma. They may be lowered hundreds of degrees by 1000 atm (101 MPa) or more of water vapor pressure. *See* EUTECTICS.

Mineral assemblages. Certain types of mineral assemblages, such as eutectics and exsolution pairs, have been discussed. Other types that do not belong to one of these classes can also give indications of temperatures of formation. These indications may be based on syntheses, including hydrothermal experiments, known stability ranges of the individual minerals of the assemblage, and effect of pressure (where volatiles such as H_2O and CO_2 occur).

For example, hydrothermal experiments have shown that analcite forms at temperatures of about 100–380°C (212–716°F) under moderate water vapor pressures, but in runs of the same composition albite forms above 380°C (716°F). Therefore, in mineral assemblages that formed near the Earth's surface, that is, in cavities in volcanic rocks or shallow intrusions, the presence of analcite can be taken to indicate formation temperatures below about 400°C (750°F) and the presence of albite to indicate higher temperatures. Good crystals of potassium feldspar (variety adularia) have been formed hydrothermally at 245°C (473°F), and quartz crystals have been formed down to about 200°C (390°F). Well-formed crystals of these minerals, even in sedimentary rocks, indicate growth temperatures of at least 100°C (212°F) and probably 200°C (390°F) or higher.

Clay minerals can also be important indicators of temperatures of formation. Their stability ranges are affected by such factors as pH of the solutions, water vapor pressure, and concentrations of constituent cations in the solutions, but some useful generalizations can be made without quantitative evaluation of these variables. Kaolin forms in acid solutions up to about 350°C (660°F) if aluminum is high and potassium is low, but when the pH is significantly above 7, montmorillonite forms from the same compositions over the same temperature range. Coprecipitated gels of alumina and silica (neutral) produce kaolin up to a little over 300°C (570°F), dickite at about 345–360°C (650–680°F), and beidellite at 360–390°C (680–730°F). *See* CLAY MINERALS.

Sepiolite and attapulgite are decomposed hydrothermally at temperatures at least as low as 200°C (390°F); thus the presence of either of these in a mineral assemblage indicates temperature of formation not over 200°C (390°F) and possibly below 100°C (212°F). The lower limit of formation of these minerals is not known.

Similar experiments give analogous results for other minerals and mineral groups. Sericite, for example, forms at about 200–525°C (390–977°F) in slightly basic to somewhat more acid solutions if aluminum and potassium are both high. Pyrophyllite forms at about 300–500°C (570–1020°F) if aluminum and potassium are both low. Serpentine cannot form above about 500°C (930°F) even under very high pressures of water vapor, and most varieties of chlorite crystallize below 500°C (930°F). Muscovite is stable from about 400 to 800°C (750–1470°F) at pressures likely to be involved in rocks formed at depths small enough so they can be brought to the surface by erosion. Phlogopite forms from about 800 to 1100°C (1470 to 2010°F) in the same pressure range. Talc does not form above about 825°C (1510°F).

If two or more minerals have been formed in equilibrium, it may be possible to narrow the temperature range considerably. Though talc is stable at temperatures up to 825°C (1510°F), and in equilibrium with enstatite at moderate pressures, the assemblage is stable only from about 670 to 800°C (1230 to 1470°F). Likewise, serpentine can form in equilibrium with brucite only below 450°C (840°F) at moderate pressures or below 400°C (750°F) at low pressures.

So many mineral associations have been studied with respect to conditions of formation that only a few interesting examples can be mentioned in the brief review below. They range from very low temperature environments of clay minerals and carbonates through hydrothermal mineral deposits to very high temperature and pressure assemblages of kimberlites, peridotites, and other deep-seated rocks.

1. *Low-temperature processes*. (a) Montmorillonite loses water at about 100°C (212°F) and changes to a combination of mixed layer clay and illite. (b) Heating acanthite below 177°C (350°F temperature of inversion to the high-temperature form, argentite) produces twinning, which is not, therefore, evidence for inversion from an original high temperature form. (c) Mineral assemblages in the Green River evaporites, from invariant points and such in equilibrium diagrams, seem to have formed in the range of about 20–60°C (68–140°F). (d) Molar sodium (Na), potassium (K), and calcium (Ca) in natural waters indicate last equilibrium with rocks at 4–340°C (39–640°F). (e) Concentrations of noble gases in ground water indicate temperatures of equilibrium with the atmosphere when the water was on the surface; indicated temperatures are up to 63°C (145°F). Ratios of pairs of the gases [especially xenon/neon (Xe/Ne) and krypton/neon (Kr/Ne)] appear to give better results than concentrations of individual gases, of which Xe and Kr appear to be the best. (f) The glaucophane-lawsonite association indicates temperatures approximately 200°C (390°F) and pressures of approximately 8 kilobars (800 MPa).

2. *Intermediate temperatures*. The temperature of inversion of low quartz increases 3.6°C for each part per million (ppm) of Al accepted into the lattice.

Manganese:iron (Mn:Fe) ratios in wolframite indicate that higher Mn varieties were formed at higher temperatures and nearer the source of the mineralizing solutions. Magnesium (Mg) and Fe in coexisting garnet and biotite indicate temperatures of formation of 300–610°C (570–1130°F) in a series of rocks studied. Mg partition between staurolite and garnet in a series of metamorphic rocks suggests temperature of formation to be 515–570°C (960–1060°F). Compositions of coexisting muscovite and paragonite, interpreted from an experimental determination of the muscovite-paragonite solvus, suggest a temperature of formation of 550–570°C (1020–1060°F).

3. *High temperatures*. Numerous studies of mineral assemblages in kimberlites, lherzolites (diallage-bronzite peridotite), and other mantle-derived rocks and their included nodules have allowed workers to assign values to the temperatures and pressures (and therefore depths) of formation of these interesting assemblages.

(a) Among those that have been used are the following: Garnet-spinel (for example, 22 kilobars or 2.2 GPa at 1240°C or 2260°F). High chromium (Cr) spinels in kimberlite at 40–55 kilobars (4.0–5.5 GPa) and 950–1050°C (1740–1920°F). Volcanic spinels at 5–10 kilobars (0.5–1.0 GPa) and 800–900°C (1470–1650°F). Low-calcium clinopyroxenes (cpx) and calcic orthopyroxenes (opx) equilibrated at $T >$ 1300°C. Cr-poor cpx and opx; $T = 1025$–1310°C (1880–2390°F). Aluminum oxide (Al_2O_3) in equilibrium assemblage of cpx-pl-opx [for example, 9 kilobars or 0.9 GPa and 900°C (1650°F); pl = plagioclase]. Nickel (Ni) partitioning between ol, opx, and cpx in basalt; T (crystallization) = 1040°C (1900°F). Fe and Mg in garnet-cpx pairs at 10–20 kilobars (1.0–2.0 GPa); $T = 700$–1000°C (1290–1830°F). Garpy-ol at 37–43 kilobars (3.7–4.3 GPa; py = pyroxene); $T = 930$–1230°C (1700–2250°F). Assemblage ol-cpx-opx at 30 kilobars (3.0 GPa); $T = 1575$°C (2867°F). Assemblage opx-cpx in xenoliths in kimberlite at 32 kilobars (3.2 GPa), $T = 920$°C (1690°F); at 47 kilobars (4.7 GPa), $T = 1315$°C (2399°F). Fe and Mg in coexisting olivine pyroxene, up to 1440°C (2550°F).

(b) Feldspars have been much used in geologic thermometry. For example, the amount of the albite molecule in plagioclase and coexisting alkali feldspar has indicated temperature of extrusion of a quartz trachyte to be about 1000°C (1830°F); crystallization of quartz syenite, 100–700°C (1830–1290°F); and late granite crystallization, 700–650°C (1290–1200°F). Composition of plagioclase and coexisting magma (glass) gives temperatures that are close to those indicated by TiO_2 in coexisting iron oxide minerals, as does the two-feldspar geothermometer.

(c) Rare-earth-element partitioning among hydrous magma, amphibole, garnet, cpx, and opx has been used to estimate temperature of crystallization of a peridotite; also, the partitioning of the rare-earth elements and major elements in cpx has been used to estimate the temperature of crystallization of a spinel peridotite.

(d) Distribution of an ion between two inequivalent cation lattices can be used to indicate temperature of formation, if pressure can be estimated independently. With three mineral lattices, both temperature and pressure can be estimated. Epidote, binary pyroxene and amphibole solid solutions, aluminum silicate (Al_2SiO_5) polymorphs, and feldspars are among the minerals that should be most useful with this technique.

(e) Distribution of magnesium oxide (MgO) between phenocrysts of pyroxene and olivine in basalt, and glassy inclusions in the phenocrysts, has been used to estimate temperature of formation of the phenocrysts. Better results can be obtained if there are glassy inclusions in phenocrysts of feldspar that formed at the same time as the others.

Inclusions. When one mineral is included in another during crystallization and the association is cooled to room temperature, stresses develop in the mineral grains. By determining conditions under which relative compression between host and inclusion is eliminated, conditions of formation can be estimated.

Electrochemical methods. A cell is made using pairs of minerals that apparently formed in equilibrium and the temperature is varied until electromotive force (emf) = 0. This method has the advantage that it is not necessary to determine compositions or cell dimensions or to do other tedious experiments. *See* ELECTROMOTIVE FORCE (CELLS).

Fossil assemblages. By determining the temperatures of the water in which certain types of organisms grow, it is possible to infer the temperature of the water at the time that strata containing fossils of the same species, or perhaps closely similar species, were laid down. This method has been confined to differentiation between cold- and warm-water assemblages, but as more information is obtained about temperatures at which certain species lived, it should be possible to assign temperature ranges to the water in which the formations containing them were laid down.

Properties dissipated by heating. Properties such as thermoluminescence, radiation colors, and metamictization, which are exhibited by many minerals (and thermoluminescence by some rocks), are dissipated by heating. Most thermoluminescence is dissipated below 250°C (480°F), although a few materials have been reported which retain some capacity for thermoluminescence up to about 500°C (930°F). Likewise, most radiation colors in minerals are dissipated below 300°C (570°F), but in a few they persist up to 500°C (930°F) or higher.

Metamictization is the destruction of the regular internal structure of a mineral produced by emanations from contained radioactive elements. The metamictization of minerals can be dissipated and the original structure restored by heating them to about 450–900°C (840–1650°F), depending upon the mineral and rate and time of heating. *See* METAMICT STATE.

Fission tracks are another feature that can be annealed out by heating. This is important because of the effect it can have on fission track dating. For example, the fission track age of a rock from diopside and zircon was 4.5 × 10^8 years, but from sphene, hornblende, and apatite, it was 7 × 10^6 years. Original solidification and cooling took place at 4.5 × 10^8 years. There was a later reheating, 7 × 10^6 years ago, to a temperature greater than 600°C (1110°F) but less than 700°C (1290°F) which erased tracks in sphene (600°C or 1110°F) and hornblende and apatite (450°C or 840°F) but not in diopside (825°C or 1520°F) or zircon (700°C or 1290°F). Some minerals and temperatures at which tracks are erased are as follows: silica glass (650°C or 1200°F), tektites and pigeonite (525°C or 980°F), enstatite and hypersthene (475°C or 890°F), muscovite (450°C or 840°F), and phlogopite (375°C or 710°F). *See* FISSION TRACK DATING.

Possession of these properties by minerals does not mean that they were formed at temperatures lower than those at which the properties are dissipated, but that they have not been heated to higher temperatures since the properties were acquired. In deducing the thermal history of such materials, therefore, the problem of when the property was acquired becomes important.

Crystallography. The generalization has been made that crystals grown at relatively low temperatures are likely to be simple in habit; those grown at high temperatures, complex. The composition and pH of the solution, presence of impurities, rate of growth, and other factors can also affect crystal habit.

Potassium feldspar is a good example of the change of crystal habit with temperature of formation. Phenocrysts of potassium feldspar in porphyrites (800°C or 1470°F±) are dominated by base, clinopinacoid, and orthodomes, giving crystals elongate parallel to the a axis. Crystals in pegmatites (500°C or 930°F±) are elongate parallel to the c axis and are dominated by (010), (001), and (110). In high-temperature veins such as those of the Alps (350°C or 660°F±), the (110) form becomes more prominent and the crystals are simpler. In very low-temperature veins the potassium feldspar is adularia (150°C or 300°F±), and only the two forms (110) and (101) remain, so the crystals are as simple as possible. Other examples are quartz, calcite, and fluorite. *See* CRYSTAL STRUCTURE.

Liquid inclusions. Minerals crystallizing from aqueous solutions commonly have imperfections that retain samples of the solution as liquid inclusions in the final crystals. When the crystal cools, the solution contracts, and a vapor bubble appears in each liquid inclusion. By heating plates of the mineral on a heating stage on a microscope and determining the temperature at which the solution just fills the cavities, it is possible to estimate temperature of formation if the pressure at formation was essentially the same as the vapor pressure of the solution. For significantly higher pressures, it is necessary to estimate what the pressure was and apply a correction.

The following necessary assumptions appear to be justified in most, if not all, cases. (1) The inclusion cavities were just filled with fluid under the temperature and pressure prevailing during crystallization. (2) Change of volume of the mineral itself is not significant. (3) Changes in volume and concentration brought about by deposition of material during cooling are such as not to affect the result. (4) Primary and secondary liquid inclusions can be distinguished under the microscope. (5) There has been no leakage from or into the inclusions. (6) The liquid is an aqueous solution containing no carbon dioxide or other gas in large concentration. (7) Pressure-volume-temperature relations are near enough to those of pure water or chloride solutions that have been studied so that no serious errors are introduced by using the available data.

Temperature ranges that have been estimated by this method for some common vein and pegmatite minerals are as follows: calcite, 40–362°C (100–684°F); sphalerite, 75–275°C (170–527°F); fluorite, 83–350°C (181–660°F); vein quartz, 100–440°C (212–824°F); pegmatite quartz, 200–530°C (390–990°F); and topaz, 275–500°C (527–930°F).

Behavior of organic material. Organic compounds are generally more sensitive to thermal changes than are inorganic ones. Therefore, when they are heated in nature, even to moderate temperatures, characteristic changes take place that can be used to estimate the temperatures to which various kinds of organic matter have been exposed. Persistence of organic compounds indicates that the temperature of the environment to which they have been exposed has not been above the temperature at which they decompose at a rate significant with respect to duration of their burial.

Degree of condensation, loss of volatiles, thermal stability of individual compounds, and other factors have been used as indications of temperatures reached. For example, temperatures of condensation to asphalt, bernalite, carbonite, and so on have been estimated to range from about 30 to 370°C (90–700°F). In general, mostly water is lost from coal up to about 100°C (212°F); CO_2 and hydrogen sulfide (H_2S) are given off to about 350°C (660°F); methane (CH_4), ammonia (NH_3), and some hydrogen (H_2) to 800°C (1470°F); above 800°C (1470°F), mostly hydrogen. "Degree of coalification" (concomitant with loss of volatiles) and reflectance of vitrinite have also been used to estimate temperatures of formation of coal. *See* COAL.

Colors developed in conodonts indicate the degree of heating that the sediments containing them have undergone, ranging from pale yellow through brown to black with heating to 350°C± (660°F±). Higher temperatures dissipate the dark colors, and a temperature of 950°C (1740°F) for several hours makes the conodonts almost clear.

Logs on the Colorado Plateau have been coalified at about 125°C (247°F). In coal formed at 4000 m (2.5 mi) depth ($\simeq$165°C or 329°F), vitrinite reflectance approaches 4%, and the coal retains about 4% of volatiles. At 5% reflectance, volatile content approaches zero.

Carboxyl groups of porphyrins remain in petroleum at temperatures lower than 200°C (390°F). Amino acids can be used in the same way in some fossils. Melting points of bitumens indicate maximum temperatures for their formation. Fractionation series near dikes and veins give thermal gradients on a relative scale. Chemical processes such as hydrogenation, elimination of oxygen or nitrogen, carboxylation, and so forth, indicate exposure to increased temperature.

Most of the processes affecting organic materials are time-dependent as well as temperature-dependent, and some of them are pressure-dependent as well. Therefore, although relative temperatures can usually be indicated, it may be difficult or impossible to suggest absolute temperatures unless time or pressure or formation, or both, can be approximated. Earl Ingerson

Isotopic Methods

The stable isotopes of many light elements, including H, C, O, and S, are not uniformly distributed in natural substances. The $^{18}O/^{16}O$ ratio, for example, ranges from 0.00188 in Antarctic ice, through 0.00200 in ocean water, to 0.00206 in limestone. Relative to ocean water, the reference standard for oxygen, the limestone is 3% or 30‰ enriched in the heavy isotope. Such variations are caused by both equilibrium and kinetic processes. Equilibrium fractionations are temperature-dependent and can therefore be used as thermometers.

H. Urey suggested in 1947 that the partitioning of oxygen isotopes between calcite and water could be used to determine the temperature at which a marine carbonate fossil was precipitated in the geologic past. Largely as a result of this suggestion, several hundred laboratories throughout the world are now equipped to undertake research in stable isotope geochemistry. Spectacular achievements of this research have been made in the field of paleoclimatology, and understanding of igneous, metamorphic, sedimentary, and hydrothermal processes has also advanced enormously.

Calibration of isotopic thermometers. Isotopic fractionations among phases at equilibrium arise from the influence of isotopic mass on the vibrational frequencies of solids and on the vibrational-rotational frequencies of molecules. Fractionation of the isotopes among molecular species can be calculated from spectroscopic data, using methods of statistical mechanics, but the effects for crystals and liquids are difficult to calculate accurately. Experimentally determined fractionations, generally considered more reliable than those calculated from spectroscopic data, are obtained by analyzing the isotopic compositions of coexisting phases that have been thoroughly equilibrated at known temperatures in the laboratory. After this has been accomplished, the isotopic thermometer is said to be calibrated (**Figs. 1** and **2**). *See* STATISTICAL MECHANICS.

In most cases, isotopic fractionation between phases increases with decreasing temperature, and statistical mechanics predicts zero fractionation at infinite temperature. Pressure does not significantly affect the equilibrium distribution of isotopes among most substances, because the net volumetric effects of isotopic substitutions are typically negligible. Researchers have noticed that high pressures increase the rate of isotopic exchange in hydrothermal experiments and thereby facilitate the attainment of equilibrium between phases.

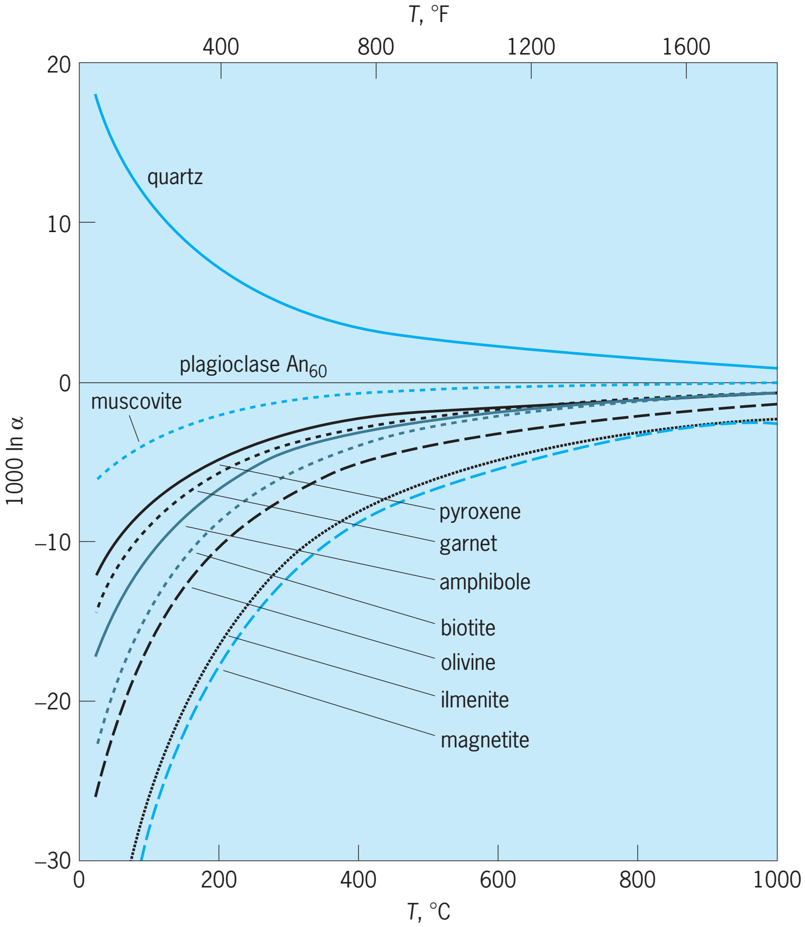


Fig. 1. Equilibrium oxygen isotope fractionations between several important rock-forming minerals and plagioclase feldspar (An_{60}). Minerals with the highest values have the strongest tendency to concentrate ^{18}O. The fractionations between different minerals are largest at low temperatures. (*Modified from I. Friedman and J. R. O'Neil, Compilation of stable isotope fractionation factors of geochemical interest, U.S. Geol. Surv. Prof. Pap. 440-KK, 1977*)

The requirements for a reliable temperature determination are that isotopic equilibrium was originally attained in the natural system at the time of crystallization or formation; that a suitable isotopic thermometer has been calibrated under known laboratory conditions; and that the original isotopic composition of the natural material has been preserved throughout its subsequent history. Unfortunately, the more readily equilibrium is attained in the laboratory, the less likely it is to be "frozen in" and subsequently preserved in natural materials. In fact, naturally occurring mineral assemblages probably could not retain their initial isotopic compositions over geologic time, were it not for their tendency to be coarser-grained and to be more dehydrated than samples equilibrated in hydrothermal laboratory experiments.

Terminology. The δ value is defined as the per mille (‰) difference in isotopic ratios between a sample

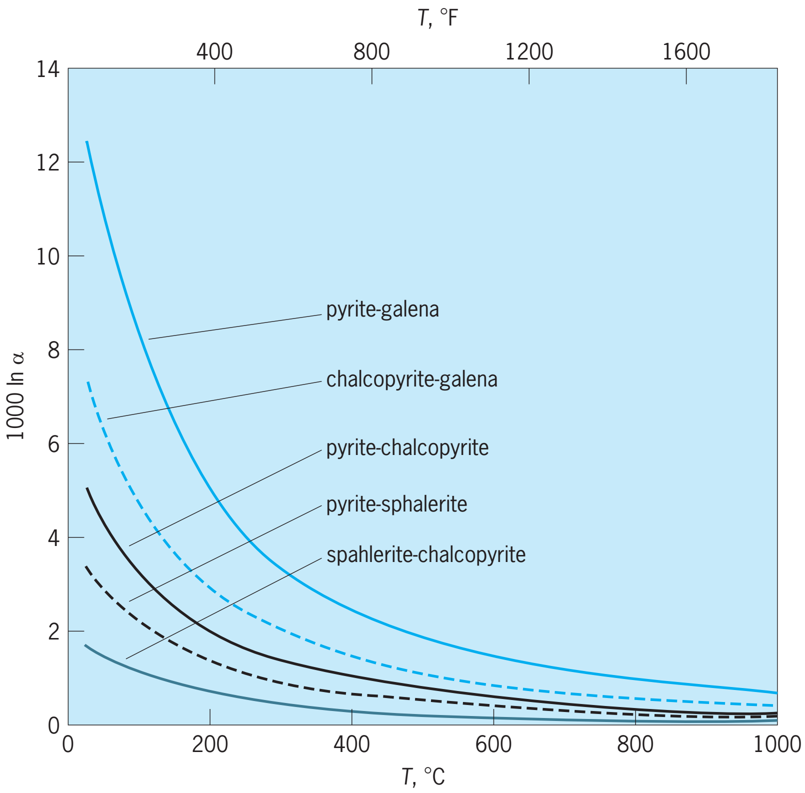


Fig. 2. Equilibrium sulfur isotope fractionations between different sulfide minerals. (*Modified from I. Friedman and J. R. O'Neil, Compilation of stable isotope fractionation factors of geochemical interest, U.S. Geol. Surv Prof. Pap. 440-KK, 1977*)

and a standard as shown in Eq. (1), where R_A is

$$\delta_A = \left(\frac{R_A - R_{\text{std}}}{R_{\text{std}}} \right) \times 10^3 \qquad (1)$$

the ratio of deuterium to hydrogen (D/H), carbon-13 to carbon-12 ($^{13}C/^{12}C$), oxygen-18 to oxygen-16 ($^{18}O/^{16}O$), or sulfur-34 to sulfur-32 ($^{34}S/^{32}S$) in substance A. The standard for hydrogen and oxygen isotopes is standard mean ocean water (SMOW). Calcite from a fossil belemnite known as PDB is used as a standard for both oxygen and carbon isotopes. Troilite from the Canyon Diablo meteorite is the standard for sulfur isotopes.

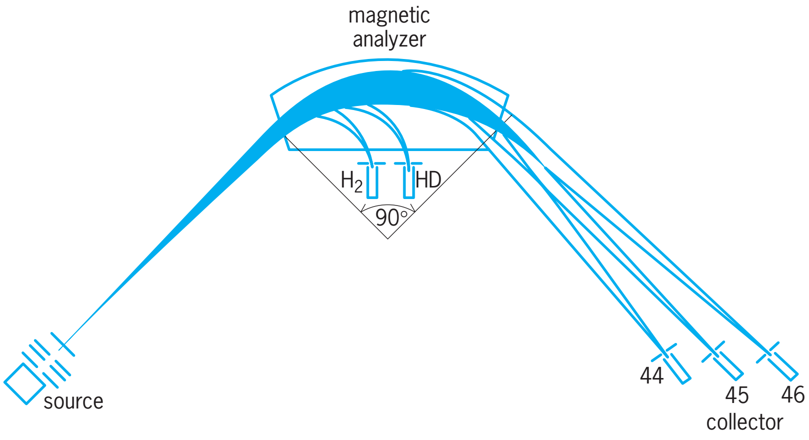


Fig. 3. A modern isotope-ratio mass spectrometer can analyze several different gases, and some can analyze samples as small as 100 micrograms or less. Modern mass spectrometers can be optimized for many different analytical purposes, but practically all rely on the traditional use of a magnetic field to effect the separation of ions having different charge to mass ratio. This MAT 252 spectrometer is set up to ionize and accelerate either hydrogen gas or carbon dioxide gas, separate the gas into distinct beams representing the different isotopic species (H_2 or HD; or masses 44, 45 and 46 for CO_2), and then precisely compare the intensities of the different mass beams. (*After R. E. Criss, Principles of Stable Isotope Distribution, Oxford University Press, 1999*)

The isotopic fractionation factor between substances A and B is defined as $\alpha_{AB} = R_A/R_B$, a quantity that is related to the equilibrium constant of an isotopic exchange reaction. An example is D/H exchange between hydrogen gas and water [reaction (2)].

$$HD + H_2O = H_2 + HDO \qquad (2)$$

The per mille fractionation, 10^3 ln α_{AB}, is approximately equal to $\delta_A - \delta_B$ or Δ_{A-B}, which for reaction (2) is the difference between the δD values of the water and the hydrogen gas.

Analytical techniques. The element to be isotopically analyzed is quantitatively extracted from the sample and converted to a gas such as hydrogen (H_2), carbon dioxide (CO_2), or sulfur dioxide (SO_2). The gas is then introduced into a special mass spectrometer designed to simultaneously collect two or more ion beams differing in mass per unit charge, such as $[HD]^+$ and $[H_2]^+$, or $[^{12}C^{16}O^{18}O]^+$, $[^{13}C^{16}O^{16}O]^+$, and $[^{12}C^{16}O^{16}O]^+$. Electronic comparison of the ion-beam currents yields the desired isotopic ratio to a precision of ±0.5‰ for H, or ±0.05‰ for carbon (C), oxygen (O), and sulfur (S) [**Fig. 3**].

Many techniques can be used to obtain a gas from a sample for analysis by mass spectrometry. Examples are hydrogen extracted from water by reaction with hot zinc, chromium, or uranium metal; carbon dioxide obtained from carbonates by reaction with phosphoric acid (H_3PO_4) at a fixed temperature; organic compounds combusted to CO_2; oxygen liberated from silicates and oxides by reaction with hot fluorine (F_2) or bromine pentafluoride (BrF_5); and sulfides ground with copper(I) oxide (Cu_2O) and heated to produce SO_2. More recent developments include the ability to ablate and analyze microscopic samples using laser or ion microprobes, the analysis of tiny gas samples that are introduced as pulses into a stream of helium that continuously flows into the mass spectrometer, or using mass spectrometers that are directly interfaced with gas chromatographs or other analytical equipment to analyze isotopes of various compounds in complex mixtures. *See* MASS SPECTROMETRY.

Igneous rocks. Partitioning of oxygen isotopes among minerals at igneous temperatures can be studied (**Fig. 4**). The tendency of many minerals to concentrate the heavy isotope can be represented by B, the coefficient in Eq. (3), where α_{MP} is the equilibrium fractionation factor between the mineral and plagioclase (An_{60}), and T is absolute temperature.

$$1000 \ln \alpha_{MP} = \frac{B \times 10^6}{T^2} \qquad (3)$$

The temperatures at which igneous rocks crystallize can be estimated from experimental petrology

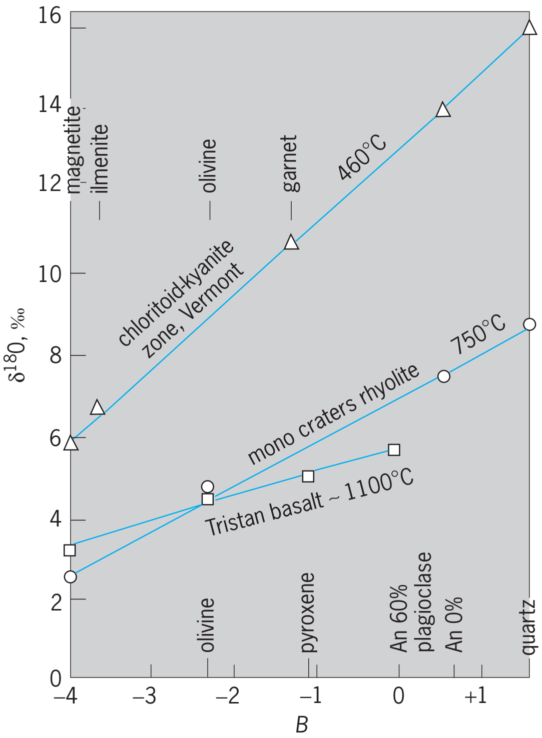


Fig. 4. Oxygen isotope compositions of minerals in a basalt, a rhyolite, and a metamorphic rock. The horizontal scale represents the tendency of each mineral to concentrate ^{18}O relative to plagioclase. The slopes decrease with increasing temperature of formation. °F = (°C × 1.8) + 32. (*After R. W. Fairbridge, ed., Encyclopedia of Geochemical and Environmental Sciences, Van Nostrand Reinhold, 1972*)

and direct observation, so isotopic thermometry is not especially valuable for this particular application. Crystal-melt fractionations are generally small and produce only small isotopic variations during igneous differentiation, unless crystallization is accompanied by assimilation of the wall rock.

Subsolidus exchange. Surprisingly, numerous departures from oxygen isotopic equilibrium have been noted in igneous rocks, and these provide far more interesting information about their genesis than does the isotopic thermometry of their crystallization. Plutonic rocks commonly show evidence of subsolidus isotopic exchange between coexisting minerals, produced during slow cooling. A good example is presented by the plagioclase-ilmenite fractionations of about 5‰ in rocks from the Labrieville anorthosite, indicating a temperature of 600°C (1110°F) that is much lower than the crystallization temperature of 1100°C (2010°F). In plutonic rocks with many minerals, slow cooling will typically result in disequilibrium fractionations that would indicate inconsistent isotopic "temperatures" for various mineral pairs, which are all significantly lower than the crystallization temperature.

Rocks that have been altered with infiltrating fluids typically display even more profound departures from isotopic equilibrium. In contrast with slow cooling, which merely causes the redistribution of oxygen isotopes within the rock, fluid infiltration conveys exotic oxygen into the rock and thereby can change the overall heavy isotope concentration. Thus, slow cooling can produce large, disequilibrium, fractionation factors as some minerals gain oxygen-18 at the expense of other minerals that lose oxygen-18, but fluid infiltration can cause all the minerals in the rock to become either richer or poorer in oxygen-18, mostly depending on the composition of the fluid. Proof that such fluid–rock interactions have occurred is therefore recorded in steep, positive slopes on plots where the delta values of one mineral in a rock are plotted directly against those of another mineral (**Fig. 5**).

Isotopic studies of several igneous provinces have revealed widespread depletion of ^{18}O and deuterium (D) in the upper crust due to interaction with circulating geothermal waters of meteoric origin. Such investigations have shown that the order of increasing resistance to hydrothermal ^{18}O exchange is feldspar-biotite-pyroxene-magnetite-quartz. This hierarchy helps explain the steep slopes of the altered rock suites in Fig. 5, and indicates that the frequently used quartz-magnetite isotopic thermometer is relatively resistant to hydrothermal alteration. *See* IGNEOUS ROCKS.

Large oxygen isotope variations discovered in mantle-derived eclogitic xenoliths in kimberlite pipes were originally attributed to crystal-melt fractionation at high pressure. However, it is now realized that the eclogites represent ancient oceanic crust that experienced isotopic exchange with convecting seawater at a spreading ridge. Oceanic crust has thus descended to depths of over 60 mi (100 km) and eventually has returned to the surface as fragments in volcanic eruptions. Some of the associated diamonds may represent organic carbon. *See* ECLOGITE.

Metamorphic rocks. Regularities in the oxygen isotope fractionations among coexisting quartz, plagioclase, pyroxene, garnet, rutile, ilmenite, and magnetite in regionally metamorphosed rocks suggest that these minerals commonly reached equilibrium at the maximum metamorphic temperatures attained, and preserved their isotopic compositions during subsequent retrograde cooling. Isotopic analyses of these minerals have yielded temperatures of metamorphism that are consistent with petrologic estimates. However, confidence in isotopic geothermometry is justified only when several mineral pairs provide concordant temperatures.

Isotopic thermometry indicates that lizardite-rich serpentinites formed at temperatures below 100°C (212°F), antigorite serpentinites at 200–400°C (370°–1020°F), glaucophane schists at 200–550°C (390–1020°F), garnet-grade schists at 450–500°C (390–750°F), and sillimanite schists and eclogites at 500–600°C (930–1110°F).

In addition, carbon and sulfur isotope fractionations have been applied to metamorphic thermometry. The concentration of ^{13}C into calcite relative to graphite is useful in marbles; the partitioning of ^{34}S among sulfide minerals is useful in metamorphosed ore deposits. *See* METAMORPHIC ROCKS.

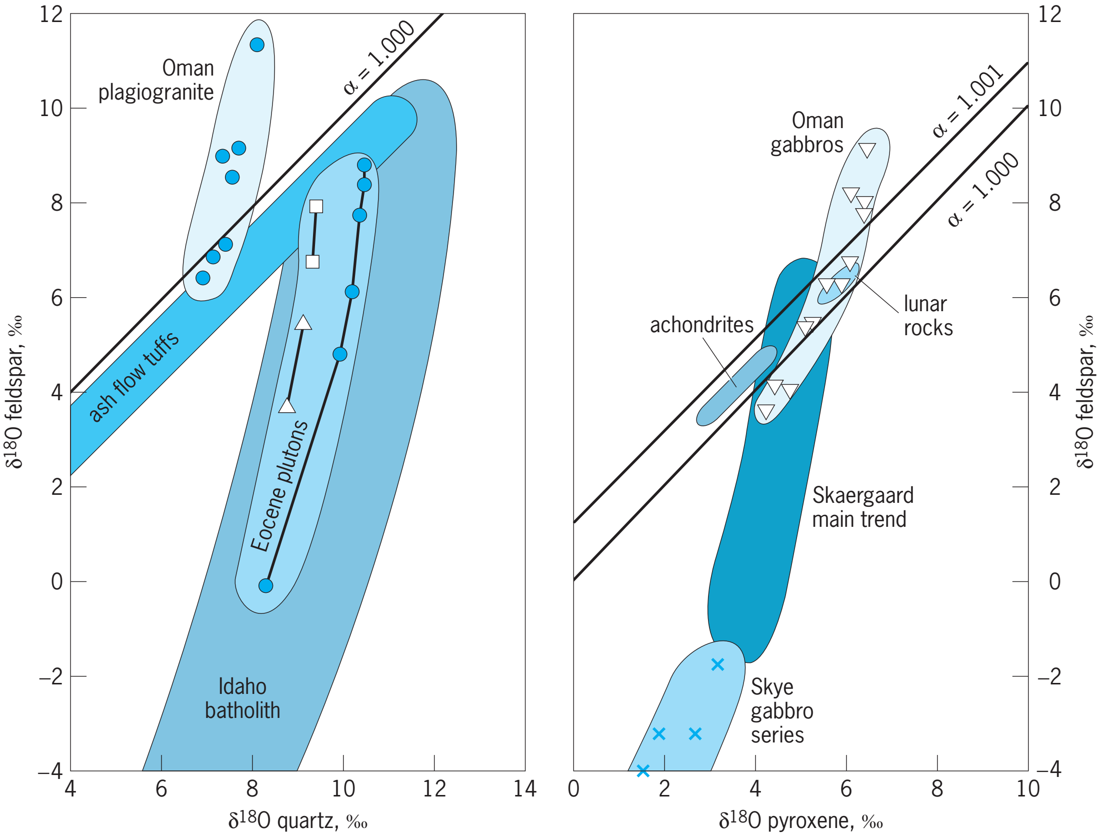


Fig. 5. Oxygen isotope compositions of coexisting quartz and feldspar, or pyroxene and plagioclase, in several igneous rock suites. Lunar rocks, achondritic meteorites, and many ash-flow tuffs cooled quickly and preserve small fractionations indicative of their magmatic crystallization temperatures; these plot along high-temperature isotherms which have unit slopes and small *y*-intercepts on this diagram. In contrast, many plutonic rock suites such as the Idaho batholith and the Skaergaard intrusion have been altered by fluids; such sample suites define steep positive slopes with anomalously large, variable, disequilibrium fractionation factors. The samples from Idaho, Skaergaard, and Syke interacted with meteoric waters which are low in ^{18}O, whereas those from Oman were infiltrated by sea water which is relatively high in ^{18}O. (*Modified from R. T. Gregory, R. E. Criss, and H. P. Taylor, Jr., Oxygen isotope exchange kinetics of mineral pairs in closed and open systems: Applications to problems of hydrothermal alteration of igneous rocks and Precambrian iron formations, Chem. Geol., 75:1–42, 1989*)

Geothermal waters. Oxygen isotope fractionations between water and dissolved sulfate from deep boreholes have yielded isotopic temperatures in reasonable agreement with the maximum measured temperatures in the geothermal reservoirs at Wairakei, New Zealand; Otake, Japan; and Larderello, Italy. The rate of isotope exchange between water and dissolved sulfate is sufficiently rapid to assure isotopic equilibrium in all reservoirs of significant size having temperatures above 140°C (280°F).

The rate of isotopic exchange is sufficiently slow, however, that no significant reequilibration is expected when geothermal waters ascend and escape to the surface via natural hot springs of significant flow rate. Unfortunately, the isotopic compositions of spring waters can be altered by boiling or dilution with near-surface waters. Surface meteoric waters usually contain less ^{18}O because they have not undergone the isotopic exchange with hot rocks experienced by deep geothermal waters. Boiling tends to increase both ^{18}O and D in the residual waters. Furthermore, oxidation of hydrogen sulfide can contribute sulfate to spring waters and thereby alter their sulfate compositions. Despite these difficulties, subsurface reservoir temperatures can be estimated from spring water isotopic analyses if appropriate corrections are made for boiling, dilution, and oxidation as indicated by chemical analyses. In this manner, the temperature of the deep reservoir at Yellowstone National Park (Wyoming, Montana, Idaho) has been estimated to be 360°C (680°F).

An additional isotopic thermometer that has been applied to geothermal systems is the partitioning of ^{13}C between carbon dioxide and methane.

Hydrothermal ore deposits. Although the majority of stable isotope studies of ore deposits are motivated by a desire to elucidate the origins of hydrothermal fluids, isotopic temperatures have also been obtained for several deposits. Sulfur isotope fractionations between coprecipitated sphalerite and galena in Providencia, Mexico, have yielded temperatures that are usually within 20°C (68°F) of the filling temperatures of associated fluid inclusions. A similar investigation of the Darwin lead-zinc replacement ores in California yielded reasonable isotopic temperatures ranging from 380°C (720°F) to 270°C (520°F). However, the wide range of sulfur isotope fractionations observed in low-temperature Mississippi Valley–type

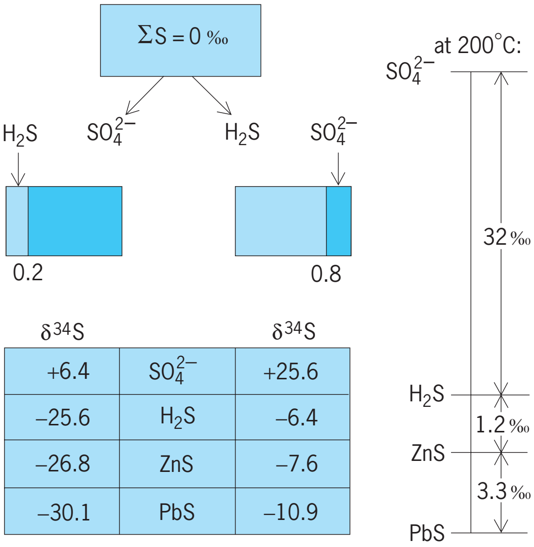


Fig. 6. Variation of $\delta^{34}S$ of sulfate, H_2S, and sulfide minerals with variation in H_2S/SO_4^{2-} of a hydrothermal solution at $T = 200°C$ (392°F), and $\delta^{34}S$ total = 0‰. (*After R. O. Rye and H. Ohmoto, Sulfur and carbon isotopes and ore genesis: A review, Econ. Geol., 69:826–842, 1974*)

deposits indicates that isotopic equilibrium may not occur in these deposits.

The effects of variations in H_2S/SO_4^{2-} ratio (or oxygen fugacity) on the isotopic compositions of hydrothermal minerals can be modeled (**Fig. 6**). It is clear that the oxidation of H_2S to SO_4^{2-} in a hydrothermal solution will have a profound influence on the sulfur isotope compositions of minerals forming from it. The consequent large isotopic variation observed in some deposits complicates the application of isotopic thermometry. *See* FUGACITY.

Sulfate-sulfide isotope fractionations in the metamorphosed Balmat-Edwards lead-zinc district, New York, indicate temperatures of about 620°C (1150°F) near the peak temperature of metamorphism. However, pyrite-sphalerite and sphalerite- galena fractionations indicate temperatures of 390°C (730°F) and 325°C (617°F), respectively. These are probably spurious temperatures caused by isotopic exchange during retrograde cooling.

Exceptionally large oxygen isotope differences have been observed between calcite and uraninite in hydrothermal veins. The δ values of calcite and uraninite in the Martin Lake mine, Saskatchewan, are +16.8 and −30‰, respectively. It is not clear if these values reflect a very large equilibrium fractionation, or deposition of the uraninite from isotopically light meteoric waters. Paleoclimate information has been gleaned from the isotopic composition of meteoric water trapped as fluid inclusions in epithermal fluorite from Idaho. *See* ORE AND MINERAL DEPOSITS.

Paleoclimates. The development of isotopic methods for determining the ancient temperatures at which fossil shells were grown has vastly increased understanding of the thermal history of the Earth's surface. The possibility of using that knowledge to predict the future course of climate is stimulating much research.

The most widely used isotopic thermometer is the fractionation of oxygen isotopes between the calcite shells of microscopic foraminifera and ocean water (**Fig. 7**). Isotopic variations exist for fossil planktonic foraminifera from Caribbean sediments during the past 730,000 years (**Fig. 8**). Seasonal variations are averaged out in such data because tens of individual shells are picked for each analysis from a section of sediment core representing hundreds of years of accumulation. If the isotopic composition of seawater were constant through time, the $^{18}O/^{16}O$ ratio of a fossil shell grown at equilibrium would depend only on its mean temperature of deposition. This is not the case because periodic accumulations of ^{18}O-depleted continental ice during the Pleistocene caused corresponding enrichments of ^{18}O within the oceans, thus enhancing the isotopic contrast between glacial and interglacial biogenic

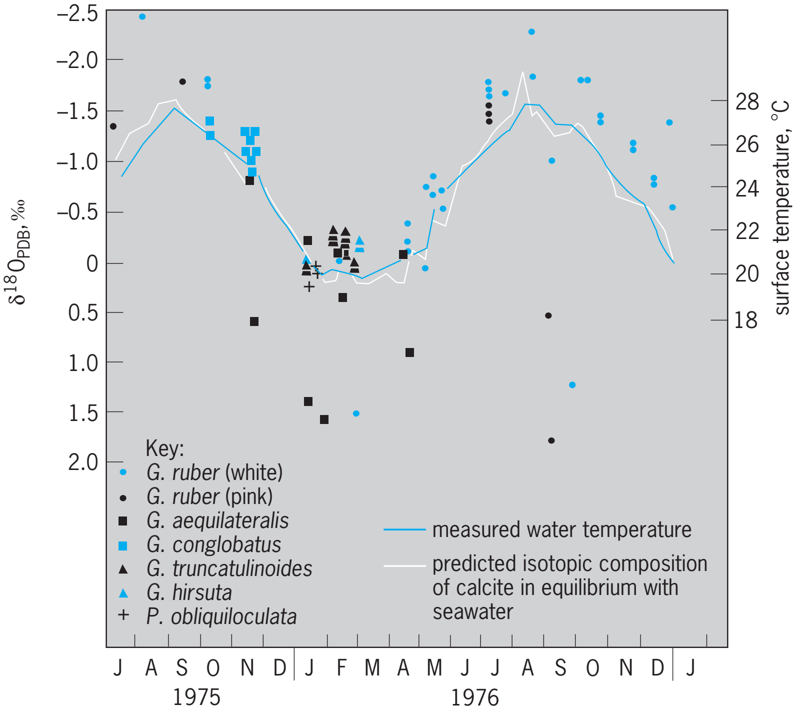


Fig. 7. Seasonal variation of $\delta^{18}O$ of living planktonic foraminifera (genera *Globigerina* and *Pachiderma*) off Bermuda. °F = (°C × 1.8) + 32. (*After D. F. Williams, A. W. H. Be, and R. G. Fairbanks, Seasonal oxygen isotopic variations in living planktonic foraminifera off Bermuda, Science, 206:447–449, 1979*)

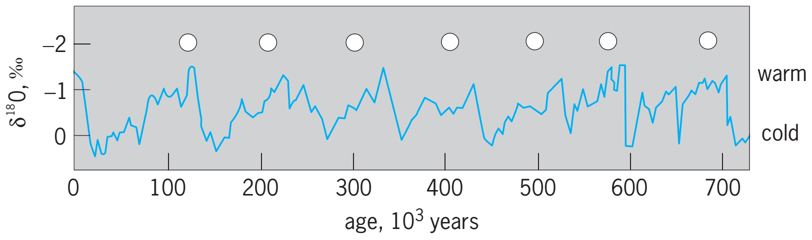


Fig. 8. Oxygen isotope variations in one species of planktonic foraminifera from Caribbean sediments during the past 730,000 years. The reference standard is PDB. The circles indicate times of maximum orbital eccentricity when northern summers occurred at perihelion and were relatively warm. (*After C. Emiliani, The cause of the ice ages, Earth Planet Sci. Lett., 37:349–352, 1978*)

calcium carbonate. This ice-volume influence during the past few glacial cycles can be estimated from analyses of benthic foraminifera that grew under consistently frigid temperatures. Seawater variation during the Pleistocene accounts for most of the 1.8‰ fluctuation in isotopic composition observed in planktonic foraminifera from tropical ocean sediments.

Isotopic analyses of benthic foraminifera from pre-Pleistocene sediments reveal variations in excess of possible ice-volume effects. Furthermore, the contrast between benthic and planktonic analyses is less in pre-Pleistocene sediments, presumably because deep water was warmer in preglacial times. Deep-sea temperatures were about 15°C (27°F) warmer 50 million years ago. Rapid, 1-part-per-million increases in the ^{18}O compositions of benthic foraminifera 35 and 15 million years ago probably indicate the initiation (temperature effect) and growth (ice-volume effect) of the Antarctic ice sheet, respectively. *See* CENOZOIC; PLEISTOCENE.

The recorded climatic cycles (Fig. 8) have led to several explanatory hypotheses. Statistical analysis has shown that a part of the variance in the climatic record is concentrated in three spectral peaks with periods of 100,000, 41,000, and 23,000 years. These closely match the periods of the Earth's orbital eccentricity, obliquity, and precession. This evidence supports the Milankovitch theory, which predicts extensive glaciation in the Northern Hemisphere a few thousand years in the future.

Orbital influences on solar radiation may not be the only cause of Pleistocene climate swings. For example, fluctuations in the carbon dioxide concentration in the atmosphere may also affect climate. Available data show an excellent correlation between the carbon dioxide trapped in Antarctic ice and the temperature variations derived from hydrogen isotope variations in the ice (**Fig. 9**). Heavy isotopes of hydrogen and oxygen are highly depleted in Antarctic snow because they preferentially "rain out" of air before it reaches the Antarctic interior. In addition, the abundances of these isotopes in precipitation correlate with air temperature. The mechanisms controlling atmospheric carbon dioxide concentration are complex and include ocean–atmosphere exchange, ocean mixing, biological productivity, volcanism, weathering, and fossil-fuel combustion.

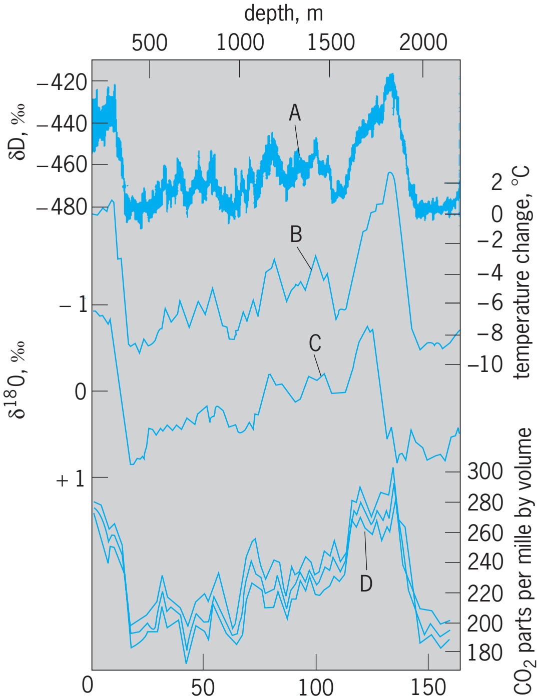


Fig. 9. Graphs showing correlation between carbon dioxide trapped in ice and inferred temperature variations. Curve A: hydrogen isotope profile in Vostok core, Antarctica [deuterium (D) per mille versus standard mean ocean water (SMOW)]. Curve B: temperature profile (relative to modern temperature) derived from curve A after correcting for isotopic changes in seawater derived from curve C. Curve C: oxygen (O) isotope changes in seawater derived from analyses for foraminifera. Curve D: carbon dioxide (CO_2) content of air trapped in ice (with uncertainty bands). (*After C. Lorius et al., Antarctic ice core: CO_2 and climatic change over the last climatic cycle, EOS, 69:681–684, 1988*)

In order to extend isotopic thermometry to Precambrian oceans, it is necessary to analyze materials that are less amenable than calcite to postdepositional isotopic exchange with circulating waters. Chert is considered the most suitable candidate for this role. The cherts that exhibit the highest δ^{18}O values within each age group are considered to be the most pristine; the other cherts have probably experienced isotopic exchange with adjacent rocks, with waters at elevated temperatures, or with isotopically light meteoric waters. The involvement of meteoric waters can be ascertained by means of D/H analyses of hydroxyl groups that are incorporated into the cherts.

There are three competing explanations of the trend toward higher δ^{18}O values with decreasing age. The first assumes fairly uniform temperatures and a progressive enrichment of ^{18}O in ocean water over geologic time. The second assumes that the ^{18}O content of ocean water is nearly constant over time because it is buffered by seawater–basalt interactions at 200–300°C (390–570°F) along axes of sea-floor spreading. The chert trend can then be interpreted in terms of progressive cooling of the Earth's surface temperature over time, with the climate being a steamy 70°C (160°F) about 3.4 billion years ago. The third possibility is that the ancient chert has not preserved its original isotopic composition. *See* CHERT; PALEOCLIMATOLOGY. G. Donald Garlick; Robert E. Criss

Bibliography. A. T. Anderson, Mineralogy of the Labrieville anorthosite, Quebec, *Amer. Mineral.*, 51:1671–1711, 1966; R. E. Criss, *Principles of Stable Isotope Distribution*, Oxford University Press, 1999; C. Emiliani, The cause of the ice ages, *Earth Planet. Sci. Lett.*, 37:349–352, 1978; R. W. Fairbridge (ed.), *Encyclopedia of Geochemical and Environmental Sciences*, Van Nostrand Reinhold, 1972; I. Friedman and J. R. O'Neil, *Compilation of Stable Isotope Fractionation Factors of Geochemical Interest*, U.S. Geol. Surv. Prof. Pap. 440-KK, 1977; R. T. Gregory, R. E. Criss, and H. P. Taylor H.P., Jr., Oxygen isotope exchange kinetics of mineral pairs in closed and open systems: Applications to problems of hydrothermal alteration of igneous rocks and Precambrian iron formations, *Chem. Geol.*, 75:1–42, 1989; C. Lorius et al., Antarctic ice core: CO_2 and climatic change over the last climatic cycle, *EOS*, 69:681–684, 1988; H. C. Urey, The thermodynamic properties of isotopic substances, *J. Chem. Soc.*, 1947:562–581, 1947; D. F. Williams, A. W. H. Be, and R. G. Fairbanks, Seasonal oxygen isotopic variations in living planktonic foraminifera off Bermuda, *Science*, 206:447–449, 1979.

Geologic time scale

An ordered, internally consistent, internationally recognized sequence of time intervals, each distinct in its own history and record of life on Earth, including the assignment of absolute time in years to each geologic interval. The geologic time scale (see **table**) has two essential components: a relative scale, consisting of named intervals of geologic history arranged in chronologic sequence from oldest (bottom) to youngest (top); and a numerical (or absolute) time scale, providing estimates of absolute ages for the boundaries of these intervals.

Relative time scale. The essential principles and concepts needed to develop a relative time scale for the Phanerozoic Eon (see table) had been recognized by the end of the eighteenth century. The principle of original horizontality implied that sediments, which eventually were lithified into sedimentary rocks, accumulated as essentially horizontal layers. The principle of superposition implied that in any undisturbed sequence of layered rocks the oldest layer was always on the bottom; hence, the bottom-to-top ordering of the geologic time scale. Uniformitarianism embodied the hypothesis that an understanding of modern geologic processes, such as weathering, erosion, and deposition of sediment, could be used to interpret the record of ancient processes, even though it is only partially complete. Igneous rocks, such as granites, were convincingly demonstrated to have been originally molten (magma) and were thus younger than any host rock into which the magma was injected. A conglomerate, very coarse detrital sedimentary rock, is younger than any of the pebbles or boulders contained within it. This is the principle of components. The application of these principles and concepts clearly demonstrated that the Earth had an unequivocally long history and facilitated interpretations about virtually all geologic materials formed as well as the determination of relative ages of different rocks, where continuously exposed. In addition, they were used to aid in correlating sequences of rocks by determining the comparable ages of rocks that were widely separated geographically. *See* CONGLOMERATE; DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; EROSION; IGNEOUS ROCKS; WEATHERING PROCESSES.

Geologic time scale

EON ERA Period [system] Epoch [series]	Age at beginning of interval, in million years	Interval length, in million years
PHANEROZOIC		
CENOZOIC		65
Quaternary (Q)*		1.8
Recent	0.01	0.01
Pleistocene	1.8	1.79
Tertiary (T)	65	63.2
Pliocene (Tpl)	5.3	3.5
Miocene (Tm)	23.0	17.7
Oligocene (To)	33.9	10.9
Eocene (Te)	55.8	21.9
Paleocene (Tp)	65.5	9.7
MESOZOIC	251	185.5
Cretaceous (K)	145.5	80
Jurassic (J)	199.6	54.1
Triassic (Tr)	251	51.4
PALEOZOIC	542	291
Permian (P)	299	48
Carboniferous (M,P)	359.2	60.2
Devonian (D)	416	56.8
Silurian (S)	443.7	27.7
Ordovician (O)	488.3	44.6
Cambrian (C)	542	53.7
PRECAMBRIAN		
PROTEROZOIC	2500	1958
LATE (Z)† (Neoproterozoic)	1000	458
Ediacaran	630	88
Cryogenian	850	220
Tonian	1000	150
MIDDLE (Y) (Mesoproterozoic)	1600	600
EARLY (X) (Paleoproterozoic)	~2500	900
ARCHEAN	3600	
LATE (W)	2800	300
MIDDLE (U)	3200	400
EARLY (V)	>3600	>400

*In parentheses are the symbols for the periods and epochs used on geologic maps and figures in North America, as well as other parts of the world.

†Letter designations of Precambrian age intervals are used by the U.S. Geological Survey.

Faunal succession. The problem of stratigraphic correlation was resolved through faunal succession, established by the British civil engineer William Smith in the early 1800s. Smith collected fossils across southern England and Wales. His observations showed that specific sequences of sedimentary rocks had a characteristic fossil content and that the fossils typical of a succession of layers were always found in a predictable order. If he found a fossiliferous layer in a new area, he could predict the fossils that would be found in younger or older layers. Using the principles of faunal succession and superposition, Smith was able to establish the sequence of sedimentary rocks over much of England and to produce the first regional geologic map, which showed the areal distribution of distinctive assemblages of rocks of different relative ages. The principle of faunal succession also permitted the correlation of British sedimentary rocks with sections in western Europe and ultimately with rocks worldwide. This was the foundation of historical geology. *See* FOSSIL; GEOLOGY; INDEX FOSSIL.

Correlation. The recognition that the fossil record was uniquely ordered through time, and that distinctive fossil assemblages within this orderly sequence could be used to determine approximate contemporaneity (the principle of correlation) among sedimentary rocks from different areas, allowed the major components of the relative geologic time scale to be compiled. Sequences of rocks with particular faunal and lithologic characteristics were given names reflecting either distinctive features or the geographic areas where they were best expressed. Examples include the coal-bearing sequences of England (Carboniferous), the prominent chalk cliffs of Dover (Cretaceous) [*creta* is Latin for chalk], and the rocks and fossils that lay below the Carboniferous and were particularly well exposed in the Devon area of southwest England (Devonian).

Distinctive sequences of sedimentary rocks and associated faunal assemblages soon became formalized as systems, and the time interval they represented became periods. Because sequences often contained considerable thicknesses of rocks and there were sequential differences in the faunal successions within them, they were divided into smaller intervals characterized solely by their faunas. The primary divisions, often simply designated as lower, middle, and upper, were later given formal names as series. Series were further divided into stages, which were subdivided into zones.

Because the fundamental principles of historical geology developed in England and rapidly spread to western Europe, almost all names from the major rock sequences that make up the relative geologic time scale had their origin in Europe. Ultimately, the principle of faunal succession facilitated correlation of rocks of Carboniferous, Cretaceous, and Devonian ages, and for many of the series, stages, and zones within them, in many parts of the world. In many cases, different local sequences of series, stages, and zones had been developed and, at least tentatively, named. A proliferation of named stratigraphic sequences led to the need for some international standardization of nomenclature. This process started at the first International Geological Congress in the late 1800s, and the sequence of systems, although lacking approximate absolute ages for their boundaries, was established by the end of the nineteenth century. Stabilization of international nomenclature for intervals within systems, and more accurate definition of the ages of interval boundaries, is an ongoing process that is the responsibility of working groups and subcommissions of the International Commission on Stratigraphy. *See* SEDIMENTARY ROCKS; STRATIGRAPHIC NOMENCLATURE.

Absolute time scale. Absolute age determinations involve estimating the absolute age of formation of geologic materials or the absolute age of a specific process that affected the materials. A better understanding of the regional history and relative ages of igneous and metamorphic rocks, and how igneous or metamorphic events compared in time with the deposition of sedimentary sequences, was made possible with the discovery of radioactivity in the late nineteenth century and the recognition that several important rock-forming minerals contained elements with radioactive isotopes. Demonstrated in the early twentieth century, isotopic age dating of geologic materials assumes that in a chemically closed system (such as a mineral grain in an igneous or metamorphic rock that has not experienced further metamorphism, melting, chemical weathering, or mechanical fracturing after grain growth) the decay products of radioactive isotopes of elements (such as potassium, uranium, thorium, or rubidium) that are found in sufficient abundance in particular minerals will be retained in the mineral grain below a specific, narrow range of isotopic blocking temperatures. The isotopic blocking temperature range is dependent on mineral structure, mineral grain size, and rate of thermal cooling. Once the rate of decay of a particular radioactive isotope (parent) has been determined, the ratio of the remaining parent to its daughter decay products retained in a closed system will provide an estimate of the numerical (absolute) age, as well as associated error of the decay system. Decay systems of this kind may be applied to common (for example, biotite, potassium feldspar, and amphibole) and less common but still ubiquitous (for example, zircon, sphene, and badellyite) minerals in a wide range of igneous and metamorphic rocks. *See* BASEMENT ROCKS; METAMORPHIC ROCKS; RADIOACTIVITY.

The decay rate for each radioactive isotope is different, and therefore some isotopic decay schemes are ideal for very old rocks and some are better suited for young rocks or surface deposits. A slightly different system involves carbon-14 (^{14}C), which is a radioactive isotope produced by cosmic-ray interaction with nitrogen-14 (^{14}N) in the upper levels of the atmosphere. The production of ^{14}C is influenced by the intensity of the Earth's geomagnetic field. ^{14}C decays to ^{14}N on a time scale of several thousands of years. Over time, the ratio of unstable ^{14}C to the far more common carbon isotope ^{12}C has become stabilized in the atmosphere, and this ratio enters into the carbon incorporated by every living organism, where it is maintained as long as the organism is alive. When an organism dies, the ^{14}C gradually decays to ^{14}N, and the time at which the organism died can be estimated by how much ^{14}C is retained in the remains (for example, bone, shell, or plant fragment). Because of the relatively rapid rate of decay of ^{14}C, this radioactive isotopic decay scheme can be used only to date geologically young materials (that is, less than about 70,000 years). Isotopic age determinations of specific minerals and, sometimes, whole rocks constitutes geochronology. *See* DATING METHODS; GEOCHRONOMETRY; RADIOCARBON DATING; ROCK AGE DETERMINATION.

Geochronology did not become of geologic importance until the 1940s, when development of mass spectrometers permitted accurate determination of minute quantities of individual isotopes. Continual refinement of the instruments used for isotopic analysis today permits high-precision age determinations of parts of individual mineral grains less than a few hundreds of micrometers in any dimension. The accuracy of refinements to the absolute geologic time scale continues to improve. *See* MASS SPECTROMETRY; MASS SPECTROSCOPE.

Parts of the relative time scale can be assigned absolute age estimates, and substantially refined, if a datable igneous rock such as a volcanic ash or lava flow is interlayered with fossiliferous strata. Few such fortuitous occurrences exist at or even near boundaries between specific intervals in the relative time scale. There are exceptions. In the 1990s, examination of strata deposited across the boundary between the Precambrian and the Cambrian periods (see table) in northern Siberia identified numerous horizons of volcanic origin, which contain grains of zircon that allow for high-precision uranium-lead (U-Pb) isotopic age determinations. Based on these data, the Precambrian-Phanerozoic (Cambrian) boundary was adjusted from about 570 million years ago to about 542 million years ago. Another example of a recent refinement is that of the age of the last full reversal of the Earth's magnetic field (from the Matuyama reverse polarity chron into the Brunhes normal polarity chron, in which we are now living). This is also taken as the base of the middle Pleistocene subseries. Early age determinations suggested that the Matuyama/Brunhes boundary was about 730,000 years ago. Recognition that young marine sedimentary sequences recorded astronomically tuned cycles led to the conclusion that the estimated age of the boundary could not be correct. Recent high-precision $^{40}Ar/^{39}Ar$ age determinations on lavas slightly older and younger than the boundary, as well as those that actually record the field polarity transition, have demonstrated that the actual age of the boundary is about 780,000 years. The boundary between the Cretaceous and the Tertiary (see table), the terminal extinction of the dinosaurs and many marine species, is now considered to be about 65 to 65.5 million years, depending on the time scale chosen, based on high-precision $^{40}Ar/^{39}Ar$ and U-Pb age determinations of meteorite impact–related materials from the Chixulub impact in the Yucatan peninsula, Mexico, and Deccan plateau basalts, which were erupted across the Cretaceous-Tertiary boundary, in India. In a general sense, geochronologists are making higher-precision estimates of the ages of specific geologic events to better define the geologic time scale. Intrusive igneous rocks that crosscut sedimentary rocks, igneous rock clasts in conglomerates, and very rare metamorphic rocks with preserved fossils can be used to provide some general definition of absolute ages of parts of the relative time scale. The geologic circumstances associated with each absolute age estimate dictate that many absolute age estimates for boundaries between intervals of the time scale must be interpolated between well-defined time points, and are thus constantly adjusted as better age data become available. The Phanerozoic geologic time scale (see table) is an evolving product of combining absolute and relative age components. All of the nomenclature has been long derived from the relative time scale, which is controlled by the fossil record, and absolute ages of specific boundaries are the best estimates available.

Based on isotopic age determinations, the age of events during the early, Precambrian part of Earth history, before there was a significant fossil record, has become far better understood. There is little international consensus for the nomenclature of the divisions of the Archean and Proterozoic eons. Geologists studying Precambrian rocks adhere to nomenclatures that are controlled solely by absolute ages, used to (loosely) define interval boundaries in the Precambrian (see table). *See* ARCHEAN; PRECAMBRIAN; PROTEROZOIC.

John W. Geissman; Allison R. Palmer

Bibliography. F. M. Gradstein, J. G. Ogg, and A. G. Smith (eds.), *A Geologic Time Scale*, 2004; W. B. Harland et al., *A Geologic Time Scale*, 1989; J. C. Reed et al. (eds.), *Precambrian: Conterminous United States*, vol. C-2 of *Geology of North America*, 1993; N. J. Snelling (ed.), *The Chronology of the Geologic Record*, Geol. Soc. (London) Mem. 10, 1985; S. Winchester, *The Map That Changed the World: William Smith and the Birth of Modern Geology*, 2001.

Geology

The science of the Earth. The study of the Earth's materials and of the processes that shape them is known as physical geology. Historical geology is the record of past events. *See* EARTH; EARTH SCIENCES.

Energy from two sources continually produces changes in the Earth. Radiant energy from the Sun causes ocean currents, winds, waves, rainfall, weathering, soil formation, and a myriad of other physical and chemical changes in the outermost rocky portion of the solid Earth (lithosphere), in the fluid envelopes of water (hydrosphere) and air (atmosphere), and in the totality of living matter (the biosphere). Heat energy inside the Earth causes slow convective movements deep in the Earth's interior. The internal motions break the rigid lithosphere into large fragments called tectonic plates, which move laterally at velocities up to around 5 in. (12 cm) a year. Collisions and other interactions between moving plates of lithosphere produce the Earth's gross topography—the ocean basins, mountain ranges, even the shapes of the continents themselves. *See* ATMOSPHERE; BIOSPHERE; HYDROSPHERE; INSOLATION.

Geologists examine rocks exposed at the Earth's surface and samples recovered from drilling. However, the radius of the Earth is 3982 mi (6371 km), and so the inner portions of the Earth must be studied remotely by means of the Earth's magnetic, electrical, gravitational, elastic, and other physical properties.

Following the spacecraft landings on the Moon, and space exploration of the other planets and their moons, the study methods of geology have been used in comparative planetology, in which the origin, development, and history of all solid bodies in the solar

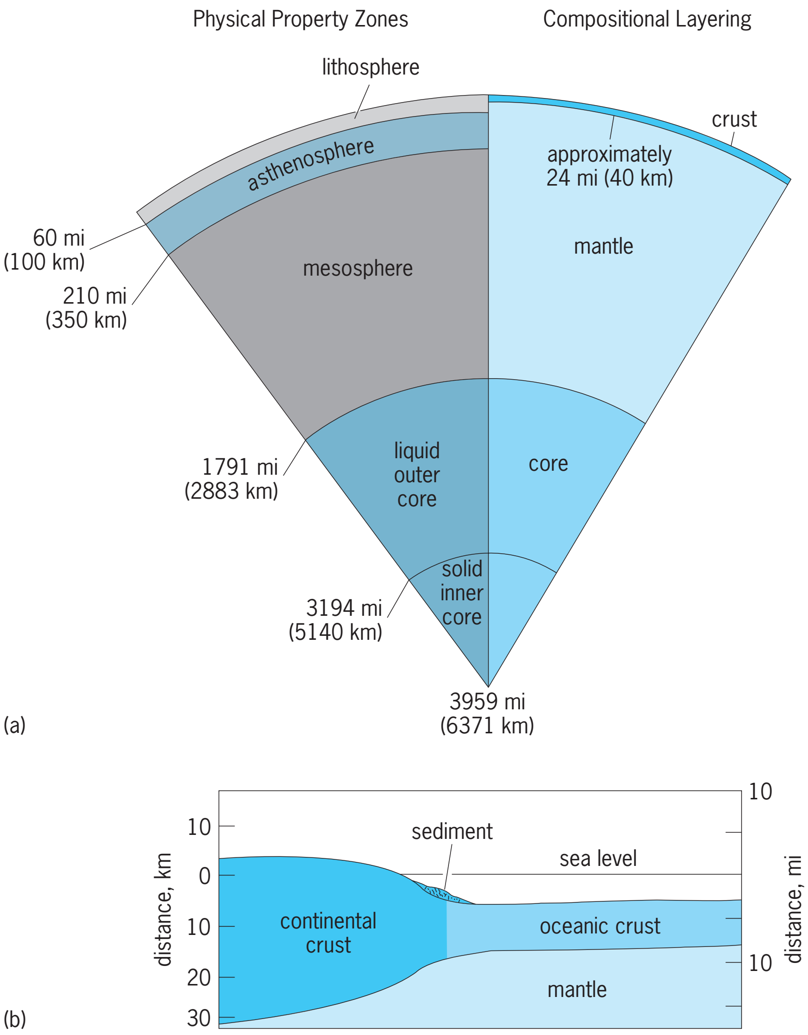


Fig. 1. Layers in the Earth. (*a*) Comparison of compositional layers and physical property changes in the Earth; temperature and pressure increase with depth. (*b*) Continental crust and oceanic crust.

system are compared. Geology became a universal science in the second half of the twentieth century, and an understanding of the geological evolution of the Moon, Mars, Venus, and other planetary bodies has provided a new perspective on the Earth's history. *See* PLANETARY PHYSICS.

Physical and historical geology. Geology is an interdisciplinary subject that overlaps and depends on other scientific disciplines. Physical geology is concerned primarily with the Earth's materials (minerals, rocks, soils, water, ice, and so forth) and the processes of their origin and alteration. Chemistry and physics are the two scientific disciplines most closely related—study of the chemistry of the Earth's materials is geochemistry, study of the physical properties of the Earth is geophysics. *See* GEOCHEMISTRY; GEOPHYSICS.

Historical geology is based on two complementary disciplines, stratigraphy and paleontology. Stratigraphy is the systematic study of stratified rocks through geologic time. The stratigraphic record reveals the sequence of events that have affected the Earth through eons of time. Absolute dates for the stratigraphic record are provided from geochemical studies of naturally occurring radioactive isotopes. Paleontology is the study of fossilized plants and animals with regard to their distribution in space and time. The fossil record is much more abundant in strata deposited during the present, or the Phanerozoic Eon, covering the past 545 million years. Paleontology is closely related to biology. The distinctions between physical and historical geology are more matters of convenience than substance, because it is increasingly clear, within the framework of plate tectonics, that all aspects of geology are interrelated. *See* PALEOBOTANY; PALEONTOLOGY; STRATIGRAPHY.

Internal structure and plate tectonics. The solid Earth is layered with respect to both composition and the physical properties (**Fig. 1**). The outermost compositional layer, the crust, is of two kinds: the oceanic crust, averaging about 6 mi (10) in thickness, everywhere underlies the ocean basins; the continental crust, averaging about 24 mi (40 km) in thickness, everywhere underlies the continents. Although the oceanic and continental crusts differ in composition, both are richer in silicon, aluminum, potassium, and sodium than the mantle which lies below. The mantle cannot be sampled directly, and its compositional heterogeneity is still uncertain. The innermost compositional layer is the core, a metallic mass composed largely of iron and nickel, with small amounts of silicon, oxygen, sulfur, and other chemical elements.

Layering of the Earth's physical properties arises from changes in pressure and temperature with depth. The outermost 60 mi (100 km) of the Earth, the lithosphere, is rigid and tough, and it can be fractured readily. From a depth of about 60–220 mi (100–350 km) is the asthenosphere, a region where temperatures are sufficiently high that rocks in the mantle are weak. Instead of fracturing, rocks in the asthenosphere flow and deform plastically. Beneath the asthenosphere is the mesosphere, a region where plastic properties decrease. Another distinct physical-property boundary occurs within the core—the outer core is molten, while the inner core, which has the same composition as the outer core, is solid. *See* ASTHENOSPHERE; EARTH INTERIOR; LITHOSPHERE.

Plate tectonics is one result of the Earth's layering of physical properties. Rigid plates of lithosphere float on the fluidlike asthenosphere. Continental crust and oceanic crust ride on top of lithospheric plates. Magma rising from deep in the mantle creates new oceanic crust where plates move apart. Where plates converge, the old oceanic crust is resorbed into the interior. Continental crust is characterized by a low density and cannot be dragged down into the mantle and resorbed in the same way as high-density oceanic crust. Instead, continental crust is continually moved around, colliding to form larger continents, fragmenting to form smaller units, in a fashion that endlessly changes the shapes of continents and the disposition of continents

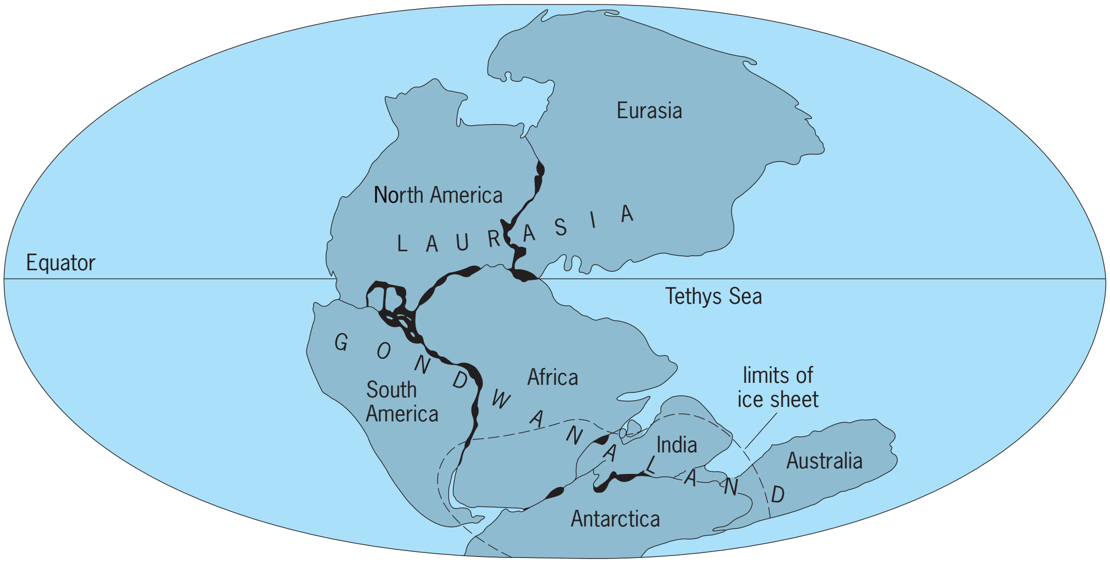


Fig. 2. Continents 250 million years ago, when they were assembled into a single large landmass called Pangaea. The northern portion of Pangaea is called Laurasia, the southern portion Gondwanaland. (*After B. J. Skinner and S. C. Porter, Physical Geology, Wiley, 1987*)

and ocean basins. Continental masses have collided and broken apart repeatedly throughout the Earth's history (**Fig. 2**). Geological activity is concentrated along plate boundaries and is most obviously manifested in earthquakes, volcanic eruptions, and the formation of mountain ranges such as the Himalaya and the Alps. *See* CONTINENT; CONTINENTAL DRIFT; CONTINENTS, EVOLUTION OF; EARTH, CONVECTION IN; EARTH CRUST; EARTHQUAKE; MAGMA; OROGENY; PLATE TECTONICS; VOLCANO.

Mineralogy. Minerals are the basic building blocks of rocks. About 3600 minerals have been identified, but fewer than 50 are common constituents in the types of rocks that are abundant in the Earth. The most common minerals in the crust are feldspars, quartz, micas, amphiboles, pyroxenes, olivine, and calcite. Each mineral is characterized by a distinct geometric packing of its constituent atoms, known as its crystal structure. Under ideal growth conditions the atomic packing is expressed as a crystal, which is a natural geometric solid bounded by plane faces. When a substance crystallizes in bulk, crowding of grains growing from neighboring centers prevents formation of recognizable crystals. An interlocking aggregate of irregularly shaped grains results. Most rocks have such an interlocking fabric of grains. Modern laboratories have effective devices for resolving the mineral content of rock materials; even the ultramicroscopic particles in clays are clearly defined under the electron microscope. *See* CRYSTAL STRUCTURE; ELECTRON MICROSCOPE; MINERAL; MINERALOGY.

Petrology. This is the study of rocks, their physical and chemical properties, and their modes of origin. Rocks are divisible into three primary families and three secondary ones. The primary families are igneous rocks, which have solidified from molten matter (magma); sedimentary rocks, made of fragments derived by weathering of preexisting rocks, of chemical precipitates from sea or lake water, and of organic remains; and metamorphic rocks derived from igneous or sedimentary rocks under conditions that brought about changes in mineral composition, texture, and internal structure (fabric). The secondary rock families are pyroclastic rocks, which are partly igneous and partly sedimentary rocks because they are composed largely or entirely of fragments of igneous matter erupted explosively from a volcano; diagenetic rocks are transitional between sedimentary and metamorphic rocks because their textures or compositions were affected by low-temperature, postsedimentation processes below conditions of metamorphism; migmatites are transitional between metamorphic and igneous rocks because they form when metamorphic rocks are raised to temperatures and pressures so that small localized fractions of the rock start to melt but the melting is insufficient for a large body of magma to develop. *See* PETROGRAPHY; PETROLOGY; ROCK.

Igneous rocks. Igneous rocks are formed as either extrusive or intrusive masses; that is, they are solidified at the Earth's surface or deep underground. Both kinds range widely in composition; silica, the most abundant ingredient, varies from about 40% to more than 75%. The most silica-rich varieties of igneous rock such as granite and granodiorite tend to be found in the continental crust. Basalt and gabbro, both containing about 50% silica, are the primary igneous rocks of the oceanic crust. *See* SILICA MINERALS.

Intrusive bodies (plutons) of granite and other silica-rich igneous rocks that formed at various depths are most numerous in mountain zones for two reasons: first, mountain belts have been much deformed, and abundant evidence indicates that crustal disturbance has favored igneous action;

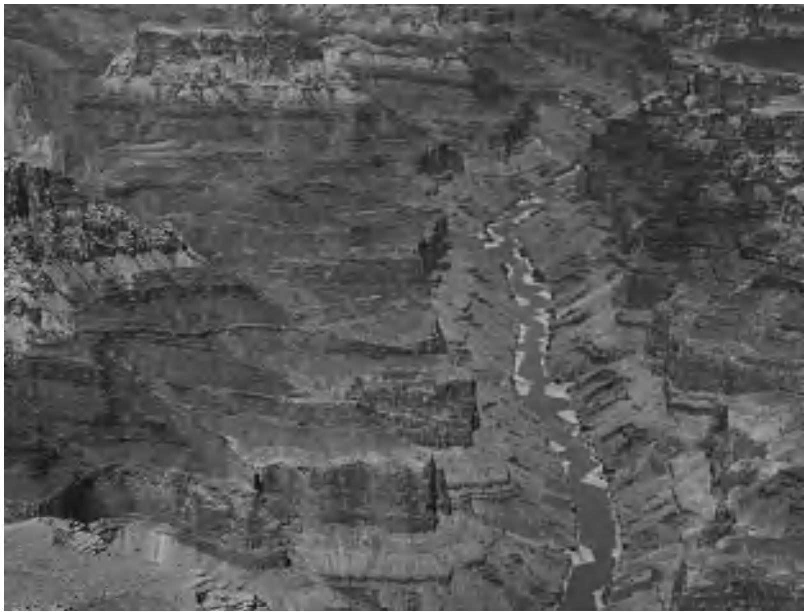


Fig. 3. Nearly horizontal sedimentary rocks approximately 4000 ft (1200 m) thick along the Colorado River Canyon, Arizona. Some tilting of layers is discernible in the left section. (*Spence Air Photos*)

second, uplifts in mountain lands have permitted erosion to great depths so that plutonic masses are exposed. Evidence suggests that many large plutons solidified in magma reservoirs that formerly supplied volcanoes in action. *See* PLUTON.

Volcanic materials are erupted through two kinds of openings—central vents and long fissures. Central eruptions build up conical mountains; the materials are in part products of explosion, pyroclastic rocks, and partly interspersed lava flows. Lavas issuing from fissures do not build up volcanic cones but instead form vast fields of volcanic rock, chiefly basalt. The most extensive and most active volcanic fissures lie beneath the sea along the centers of the mid-ocean ridges. There, basaltic magma formed in the mantle rises and creates new oceanic crust. *See* BASALT; IGNEOUS ROCKS; LAVA; MID-OCEANIC RIDGE.

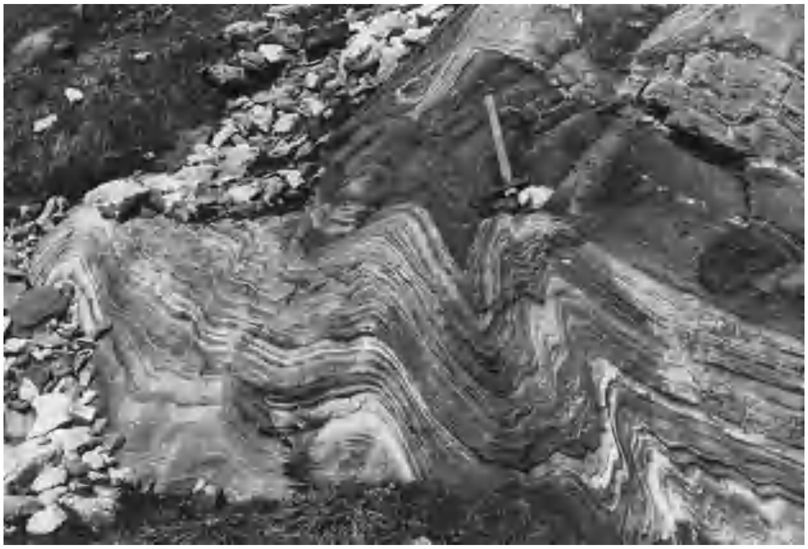


Fig. 4. Strata deformed by the compressive forces that created the Alps, near Val Malenco, northern Italy. (*Courtesy of Kurt Bucher*)

Sedimentary rocks. Bedrock exposed to the atmosphere and hydrosphere is broken into pieces, large and small, which are moved by running water and other agents to lower ground, and spread in sheets over river floodplains, lake bottoms, and sea floors. Dissolved matter is carried to seas and other bodies of water, and some of it is precipitated chemically and by action of organisms. The material deposited in various ways becomes compacted, and in time much of it is cemented into firm sedimentary rock. Generally, the deposition is not continuous but recurrent, and sheets of sediment representing separate events come to form distinct layers of sedimentary rock. The individual layers are beds or strata, and the rocks are said to be stratified (**Fig. 3**).

Large areas in every continent are underlain by sedimentary rocks that represent deposits during many periods of the Earth's history. In part, these bedded rocks are nearly horizontal, as they were originally; but in many places, particularly along the margins of present or past tectonic plates, elongate belts of bent and fractured strata can be observed in mountain belts (**Fig. 4**).

The principal kinds of sedimentary rock are conglomerate, sandstone, siltstone, shale, limestone, and dolostone. There are many other kinds, and some, though less important quantitatively, have large practical value; examples include common salt, gypsum, rock phosphate, iron ore, and coal. *See* SEDIMENTARY ROCKS.

Metamorphic rocks. These rocks have been developed from earlier igneous and sedimentary rocks by heat and pressure, especially in mountain zones found along plate margins, and near large masses of intrusive igneous rock. Thermal metamorphism adjacent to plutons results from rising temperatures, often with addition of new elements by circulating fluids, but without pronounced deformation of strata. Common effects are hardening and crystallizing of the affected rock, with changes in mineral composition.

Dynamic metamorphism involves elevated temperatures but also results from shearing stresses in rocks subjected to high pressure; effects are development of cleavage planes and growth of platy minerals in parallel arrangement. Dynamic metamorphism is most commonly observed in mountain belts where the crust is thickened and deformed due to plate collisions.

The common metamorphic rocks are in the two general classes, foliated (including slate, phyllite, schist, and gneiss) and nonfoliated (including marble and quartzite). *See* METAMORPHIC ROCKS; METAMORPHISM.

Regolith. Bedrock at and near the Earth's surface is subject to mechanical and chemical changes in a complex group of processes called weathering. Blocks and small chips that become detached due to mechanical fragmentation are especially vulnerable to chemical attack, which makes radical changes in the mineral content. Some soluble products of alteration are removed by percolating water; and in the less soluble residue the most common constituents are quartz and clay, which are the basic minerals of soil. The effectiveness of weathering and the nature of its products are controlled by climate, topography, kinds of bedrock, and other variables. Organic processes play a major role in chemical weathering, which is most effective under conditions that favor development of bacteria, plants, and ground-dwelling animals. The blanket of weathered material that covers most of the Earth's surface is called the regolith.

Weathering prepares the way for removal of rock materials and reshaping of land surfaces by several agents of erosion. The most obvious of these agents is running water, which during a single rainstorm may cut deep gullies into plowed fields and sweep vast quantities of soil, sand, and coarser debris into brooks and eventually into channels of major streams. Abrasion by such moving loads deepens and widens stream channels in hard bedrock. Study of drainage systems brings conviction that even the largest and deepest valleys have been fashioned by the action of running water. *See* WEATHERING PROCESSES.

The pull of gravity causes the regolith to move downslope, from high points to low. Soil on slopes, even those covered with grass and other vegetation, creeps slowly downward. Blocks dislodged from cliffs build steep masses of sliderock which slowly migrate downslope. Frequently, in mountain lands great masses, including loose material and bedrock, rush down as landslides. All downslope movements caused by gravity are termed mass wasting. *See* LANDSLIDE; MASS WASTING; REGOLITH.

Water moving through underground openings dissolves and carries away great quantities of material. Caverns, large and small, are a conspicuous result. In high latitudes and in some mountain regions, glaciers are powerful eroding agents. In arid regions, quantities of sand and dust are moved by wind. Large-scale erosion along the coasts of seas and lakes is performed by waves and currents. *See* CAVE; DESERT EROSION FEATURES; GLACIATED TERRAIN.

Each major agent of erosion fashions characteristic features in landscapes; the subdiscipline of geology devoted to studies of these features is geomorphology. The net tendency of erosion is to reduce the height of landmasses. Various stages in the history of reduction are indicated by forms of valleys and slopes and by relations of land surfaces to the underlying bedrock. Some wide regions have been uplifted, and the streams have been rejuvenated after an advanced stage was reached in a cycle of erosion. Rejuvenation of an eroded land surface is believed to result, at least in part, from plate tectonics. *See* EROSION; GEOMORPHOLOGY.

Sediments and sedimentation. Sediment now deposited provides essential clues to understanding processes of past sedimentation and hence of the origin of sedimentary rock. Sediment and sedimentary rock are all important in geology. On the basis of depositional environment, sediments are assigned to three categories: terrestrial, those laid down on land; marine, those deposited on sea floors; and mixed terrestrial-marine, those laid down in transitional zones such as deltas, marine estuaries, and areas between high and low tide. In each major group the sediments are further described as clastic (consisting of rock fragments) and chemical (formed either as inorganic precipitates or partly through organic agencies). Study of modern sediments in the several environments takes account of physical peculiarities and the included remains of organisms.

Terrestrial sediments are extremely variable. They are laid down chiefly through the agencies of mass wasting, running water, glacier ice, and wind. The deposits are in large part temporary, as the tendency is for them to shift downslope and seaward in continued erosion of the lands. Marine sediments, deposited beneath oceans, comprise material derived from the land, from shells and skeletons of marine animals and plants, from the ocean by chemical precipitation, and from space (particles of meteorites). *See* MARINE SEDIMENTS; NEARSHORE PROCESSES.

Tectonics and structural geology. Through careful study of rock masses, it is possible to distinguish primary structures (such as bedding in sedimentary rocks, and flow structures in igneous rocks) that were acquired as a result of the genesis of the rock, from secondary structures that result from later deformation. Significant features in sedimentary rocks make them especially valuable for registering later changes in form.

Deformation may proceed rapidly or slowly. Rapid deformation usually involves large fractures (faults) with instantaneous displacements of several meters (**Fig. 5**). Rapid deformation is generally attended by strong earthquakes. Slow deformation leads to broad-scale warping, bending, and folding. The principal kinds of structural features that record past deformation are folds, joints, faults, cleavage, and unconformities. Tectonics is closely related to structural geology but deals with regional features,

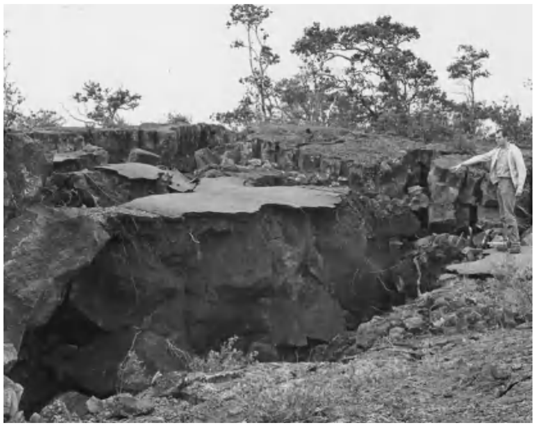


Fig. 5. Fault breaking through to the surface, near Hilo, Hawaii. (*R. S. Fiske, U.S. Geological Suvey*)

such as mountain ranges. *See* FAULT AND FAULT STRUCTURES; FOLD AND FOLD SYSTEMS; GEODYNAMICS; STRUCTURAL GEOLOGY.

The most pronounced crustal deformation tends to be found in mountain belts, where erosion has exposed exceptionally thick sections of sedimentary rocks. These rocks record long histories of slow subsidence and sedimentation, interrupted by large-scale deformation and uplift. The deformation is due to lateral compressive forces arising from plate tectonic movements. *See* CORDILLERAN BELT; MOUNTAIN SYSTEMS.

Major relief features of the Earth reflect differences in density of the underlying rocks. Continental rocks have appreciably lower average density than the basaltic rocks of the oceanic crust. Continents therefore stand high, while ocean basins are low. Great mountain blocks, such as the Alps and the Himalaya, represent thickened parts of continental crust. The condition of approximate balance among diverse parts of the crust is called isostasy (equal standing). *See* ISOSTASY.

Stratigraphy. Studies of sedimentary rocks in comparison with the many kinds of modern sediments provide a basis for recognizing conditions under which the rocks were formed. Generally, a close interpretation is possible; bodies of rock are confidently classified as deposits on floodplains, at margins of glaciers, in large lakes, in shallow seas near shore, or in deep-sea troughs. Each distinctive type of deposit represents a facies. A rock type that is essentially uniform over a considerable area constitutes a lithofacies. An assemblage of fossils that is nearly uniform in a large unit of sedimentary deposits, indicating an environment suited to certain forms of life, is a biofacies. Deposits formed at the same time may differ in both lithofacies and biofacies, reflecting differences in topography, climate, and other items of environment. *See* FACIES (GEOLOGY).

A fundamental principle in stratigraphic studies, known as the law of superposition, is that in a normal sequence of strata any layer is older than the layer next above it. This elementary law is of importance in studies of many belts of highly deformed rocks where thick sections of strata have been overturned, even completely inverted, and can be resolved only through criteria that indicate original tops of beds. Close matching of the many kinds of modern sediments with materials in sedimentary rocks formed over an immense span of time has established the uniformitarian principle, which holds that processes now operating on the Earth have operated in similar fashion through the ages.

The history of the Earth's crust is encoded chiefly in sedimentary rocks, which record a sequence of events, changing physical environments, developments in plant and animal life, and effects of crustal movements. Additional records are supplied by volcanic rocks, which in many areas are interlayered with, and grade into, sedimentary strata; by relations of intrusive igneous bodies to older and younger rocks; and by erosion surfaces, some displayed in present landscapes, others revealed in exposures of unconformities. The long history includes radical changes in physical geography, featuring a contest between land and sea; the birth, rise, and wasting away of successive mountain systems; and the evolution of living forms in seas and on lands. *See* UNCONFORMITY.

Each continent (other than ice-buried Antarctica) displays a wide lowland or platform that was occupied repeatedly by seas and is now mantled with little-deformed marine strata. The total deposit on each platform represents a long span of time and ranges from a few tens to a few thousand meters thick. The join between continental crust and oceanic crust marks the geological boundary between an ocean basin and a continent. The boundary is beneath the sea, at the foot of the continental slope, and it is here that sediment eroded from the continent is accumulated. The sites of such accumulations of sediment classically were called geosynclines. When continental collisions occur, the wedge of sedimentary strata draped along and over the continental margin becomes deformed through folding and faulting, and elevated into a mountain range (Fig. 5). Examples of this process can be found in the Appalachian Mountains and the Alpine-Himalayan chain. One of the triumphs of plate tectonics is the illumination it has thrown on this long-puzzling aspect of mountain range development. *See* CONTINENTAL MARGIN; GEOSYNCLINE; PHYSICAL GEOGRAPHY.

Geologic column and geologic time scale. At any one locality a sequence of sedimentary beds, from older to younger, can be determined through physical evidence. Persistence of some peculiar units may establish approximate correlations through moderate distances, occasionally hundreds of miles. But the key to relative dating of stratigraphic units and to confident worldwide correlations is provided by fossils of animals and plants, which record progressive evolution in living forms from ancient to recent times. Through correlation a worldwide composite diagram has been developed that combines, in chronological order, the succession of known strata on the basis of fossil contents or other evidence. This composite diagram is the geologic column.

Many of the oldest known sedimentary rocks are now highly metamorphosed, and any fossils they may have held must have been obliterated. Some thick sections of old strata that are not appreciably altered have yielded only sparse indications of life, such as patterns of marine algae and burrows made by worms or other lowly forms. Successively younger groups of strata hold abundant fossils of marine invertebrates, marine fishes, land plants that progressed from primitive to more modern kinds, reptiles, birds, and small mammals, followed by diverse and generally more advanced kinds of mammals, primitive humans, and finally modern humans. Some forms evolved rapidly, and short-lived species that were equipped to become widely dispersed are of greatest value for correlation.

Geologic column and scale of time

Eon	Era	Period	Epoch	Dates (10^6 years BP)
Phanerozoic	Cenozoic	Quaternary	Holocene	0.01
			Pleistocene	1.8
		Tertiary	Pliocene	5.3
			Miocene	23
			Oligocene	33.9
			Eocene	58.8
			Paleocene	65.5
	Mesozoic	Cretaceous		145
		Jurassic		200
		Triassic		251
	Paleozoic	Permian		299
		Pennsylvanian		318
		Mississippian		359
		Devonian		416
		Silurian		444
		Ordovician		488
		Cambrian		542
Proterozoic*	No subdivisions in wide use			2500
Archean*	No subdivisions in wide use			3600
Hadean	No subdivisions			4560

*Proterozoic plus Archean also called Precambrian.

Absolute dates can be given to points on the geologic column, and a geologic time scale developed thereby, through the use of certain naturally occurring radioactive isotopes. Such dating is known as geochronology. The names given to units of the geologic column (see **table**) are derived largely from the paleontologic record—or its absence. *See* GEOCHRONOMETRY; GEOLOGIC TIME SCALE.

Geologic mapping. A geologic map represents the lithology and, so far as possible, the geologic age of every important geologic unit in a given area. Each distinctive unit that can be shown effectively to the scale of the map is a geologic formation. A good topographic base map is essential for representing relations of bedrock to land surface forms. Cooperation of workers with specialized qualifications—for example, in petrology, paleontology, or structural geology—required for accurate mapping and description of complex areas. Large organizations, such as federal geological surveys and some commercial firms, possess diversified personnel, laboratories equipped for varied analyses, and special field equipment. Aerial photographic surveys serve as a guide in field work and help in plotting accurate locations. Satellite imagery has provided vast quantities of data for mapping. *See* AERIAL PHOTOGRAPH; REMOTE SENSING; TOPOGRAPHIC SURVEYING AND MAPPING.

A completed geologic map should indicate important structural details, such as inclinations of strata, locations of faults, and axial traces of folds. Usually the map is supplemented by vertical sections on which structural features seen at the surface are projected to limited depths. Maps of small scale may represent sedimentary rocks only according to the systems to which they belong. With larger scale a given system may be represented by several formations, each recording an important episode in the history of the region. In some European countries, geologic mapping has been completed to fairly large scales; but in all continents great areas have been mapped only in reconnaissance fashion.

Economic geology. A general knowledge of geology has many practical applications, and large numbers of geologists receive special training for service in solving problems met in the mining of metals and nonmetals, in discovering and producing petroleum and natural gas, and in engineering projects of many kinds. Human use of materials has become so great that waste materials are influencing natural geological processes. As a result, a new discipline, environmental geology, is starting to emerge. *See* ENGINEERING GEOLOGY; PETROLEUM GEOLOGY.

Brian J. Skinner

Bibliography. P. W. Birkeland, *Soils and Geomorphology*, 1999; H. Blatt, R. J. Tracy, and B. Owens, *Petrology: Igneous, Sedimentary, and Metamorphic*, 3d ed., 2005; R. J. Chorley, S. A. Schumm, and D. E. Sugden, *Geomorphology*, 1985; A. Cox and R. B. Hart, *Plate Tectonics: How It Works*, 1988; J. R. Craig, D. J. Vaughan, and B. J. Skinner, *Resources of the Earth*, 3d ed., 2001; G. H. Davis and S. J. Reynolds, *Structural Geology of Rocks and Regions*, 2d ed., 1996; G. M. Friedman, J. E. Sanders, and D. C. Kopaska, *Principles of Sedimentary Deposits: Stratigraphy and Sedimentation*, 1994; C. Klein, *Manual of Mineralogy*, revised 22nd ed., 2001; A. D. Miall, *The Geology of Stratigraphic Sequences*, 1996; A. D. Miall, *Principles of Sedimentary Basin Analysis*, 3d ed., 2000; B. J. Skinner and S. C. Porter, *The Blue Planet: An Introduction to Earth System Science*, 2d ed., 1999; B. J. Skinner and S. C. Porter, *The Dynamic Earth: An Introduction to Physical Geology*, 5th ed., 2003; S. M. Stanley, *Earth System History*, 2d ed., 2004; S. M. Stanley, *Exploring Earth and Life Through Time*, 1993.

Geomagnetic variations

Variations in the natural magnetic field measured at the Earth's surface and elsewhere in the Earth's magnetosphere (for example, at the geostationary orbit). These are field changes with periodicities from about 0.3 second to hundreds of years. (These boundaries are set to distinguish geomagnetic variations from the quasipermanent field and higher-frequency waves.) Many of these observed variations—from very short periods (seconds, minutes, hours) to daily, seasonal, semiannual, solar-cycle (11-year), and secular (~60–80 years) periods—arise from sources that either are external to the Earth (but superposed upon the larger, mainly dipolar field) or internal to the Earth (the magnetic-dipole and higher-harmonic trends and variations on the scales of hundreds and even thousands years). The daily and seasonal motions of the atmosphere at ionospheric altitudes cause field variations that are smooth in form and relatively predictable, given the time and location of the observation. During occasions of high solar–terrestrial disturbance activity that give rise to aurorae (northern and southern lights) at high latitudes, very large geomagnetic variations occur that even mask the quiet daily changes. These geomagnetic variations are so spectacular in size and global extent that they have been named geomagnetic storms and substorms, with the latter generally limited to the polar regions.

Solar quiet-time variations. The recurrent patterns of daily geomagnetic field changes (**Fig. 1**) arise in the upper atmosphere through dynamo current processes occurring at 100–120-km (60–75-mi) altitudes in the E ond lower F regions of the ionosphere. The charged particles of the daytime ionosphere are driven by thermotidal and wind forces through the main geomagnetic field to form a local current, extending primarily over the sunlit side of the Earth.

The ionization and the tidal and wind forces vary with time of day and season; the tidal and wind forces are dependent upon geographic location. However, the Earth's main field is offset from the geographic axis alignment and thus provides some geomagnetic coordinate organization to solar quiet-time (*Sq*) variations. Because the Earth's interior composition is electrically conducting, the solar quiet-time current in the ionosphere causes a secondary current to be induced within the Earth; the magnetic field

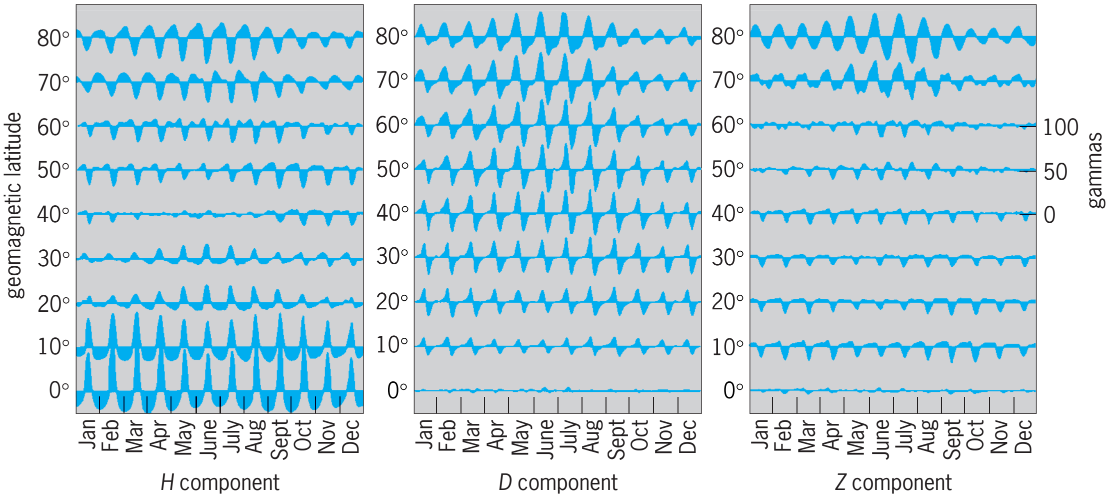


Fig. 1. Annual picture of the midmonth daily variations of quiet-time geomagnetic field, in local time, for the Northern Hemisphere from the Equator (0°) to 80° latitude displayed for *H*, *D*, and *Z* field components. The scale size between baselines is 50 nanotesla. (1 nT = 1 × 10^{-9} tesla = 1 gamma—a geomagnetic induction unit used in the past.)

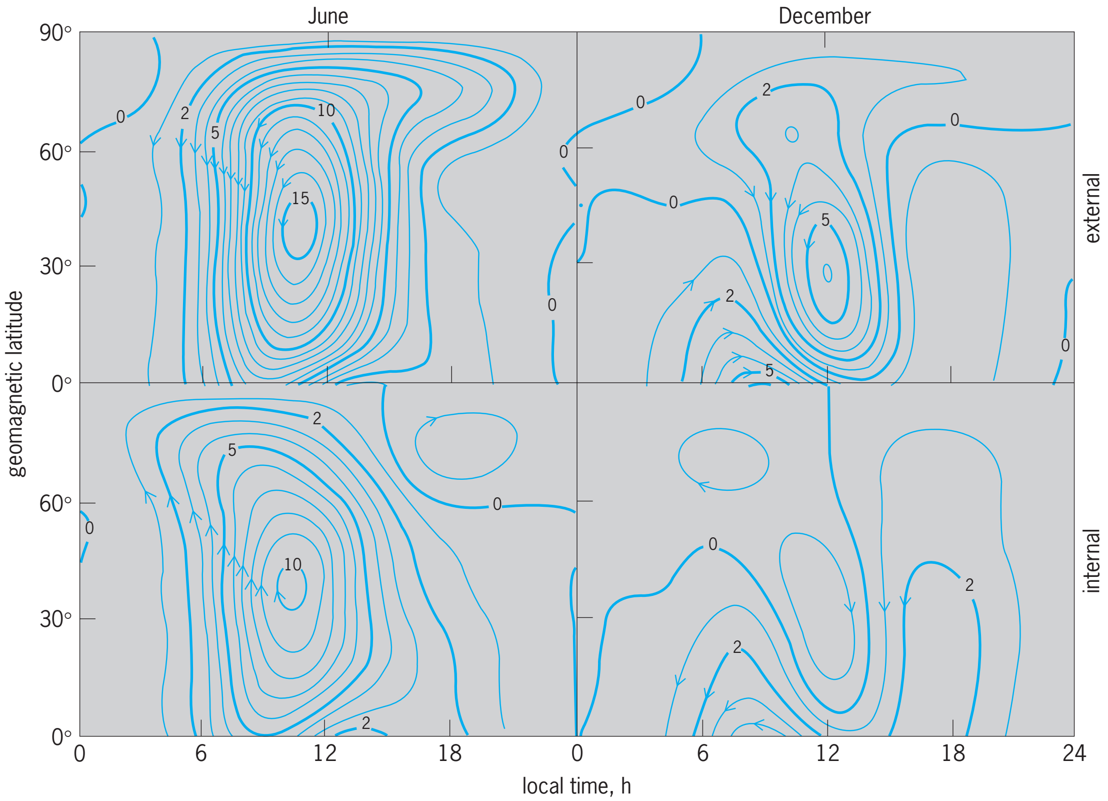


Fig. 2. Streamlines for the equivalent external ionospheric source currents (top) and internal induced currents (bottom) for solar quiet-time (*Sq*) variations of the geomagnetic field in the Northern Hemisphere for midmonth quiet conditions, in June and December. Each pattern, in local time versus geomagnetic (dipole) latitude coordinates, shows the equivalent current contours, with 10^{-4} ampere flowing in the region between contour lines. Arrows indicate the required flow direction. The midnight zero level that is assumed tor the display has no effect on the current pattern.

variations generated by these two currents are observed at the Earth's surface. In addition, a resolution of these external and internal parts of the observed fields can be used to calculate the remote deep-earth electrical conductivity profiles.

Figure 2 illustrates the computed "equivalent" external *Sq* current patterns in the ionosphere and those induced in the Earth (internal) for the summer and winter months of the Northern Hemisphere that could provide the observed fields. The denser the contours, the higher the surface field contribution.

The right-hand rule (with the right hand wrapped around the current vector and the thumb pointing in the current direction, the fingers point in the resulting magnetic field direction) can be applied showing the daily changes of field illustrated by Fig. 1. Near the Equator, the (almost) horizorital orientation of the main geomagnetic field lines causes an especially high ionospheric conductivity for the *Sq* system; as a result, there arises an intense eastward daytime current called the equatorial electrojet. In the polar cap region, there are quiet-time, Birkeland field-aligned currents flowing between the magnetosphere and ionosphere that add to the dynamo source of quiet-time geomagnetic field variations. Because magnetospheric behavior is sensitive to the direction of the interplanetary magnetic field (IMF) arriving with the solar wind particles blown out from the Sun, and because polar region field lines reach the outer parts of the magnetosphere, a signature of this interplanetary field is embedded in the polar cap *Sq* records. *See* ATMOSPHERIC TIDES; GEOMAGNETISM; INTERPLANETARY MAGNETIC FIELD; IONOSPHERE; MAGNETOSPHERE; SOLAR WIND.

Lunar variations. The seinidiurnal lunar tidal oscillations of the atmosphere drag the E-region ionization through the main field of the Earth and produce another dynamo current, *L*. The fields of this current are quite small, less than 10% of the *Sq* amplitude, and so special analytical methods are required to isolate their contribution to the observed geomagnetic records. The lunar tidal "day" is 50.5 minutes longer than the solar day, and the global *L*-amplitude patterns depend upon the twice-daily lunar tidal forces on the atmosphere, the E- and F-region ionization, and the direction and magnitude of the main magnetic field of the Earth.

Eclipse and solar flare effects. Temporary conductivity increases in the ionization due to direct x-ray radiation from solar flares, or decreases in ionization due to solar eclipses, can modify the *Sq* and *L* dynamo currents. Such variations appear as single half-cycle changes in the geomagnetic field that last from several minutes to about an hour, with maximum amplitudes rarely larger than 10 nanotesla. These small-amplitude variations are best observed on the extremely quiet *Sq*-condition days. *See* SUN.

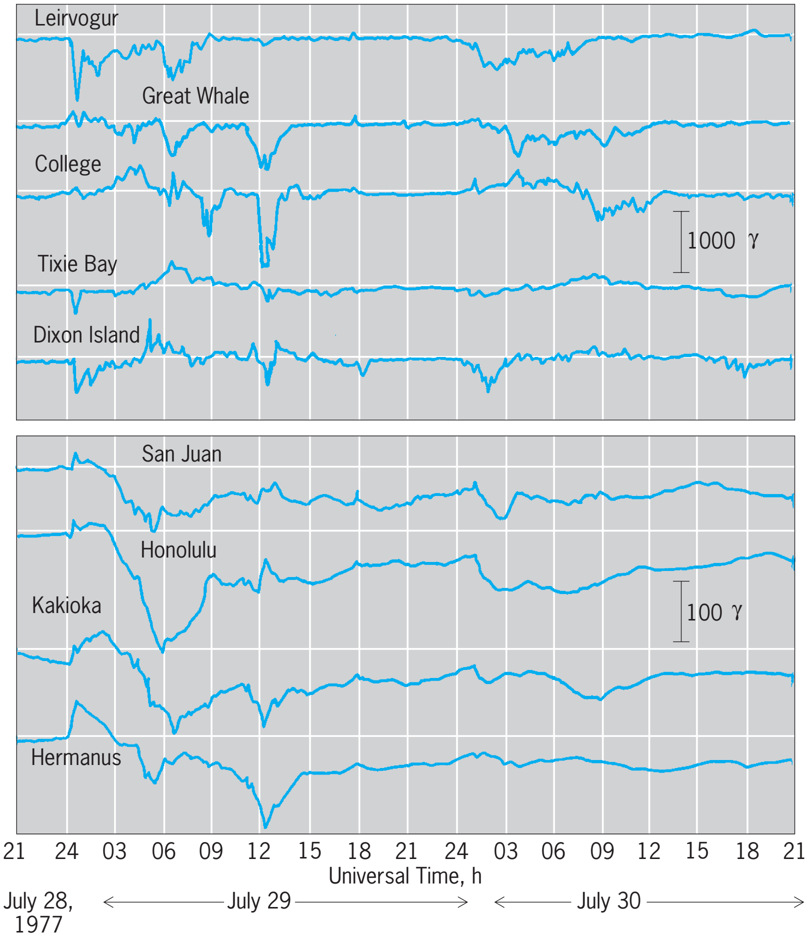


Fig. 3. Common scale magnetograms for the *H* component of a geomagnetic storm variation recorded at two groups of stations: the auroral zone (top) and the low-latitude regions (bottom). Gamma (γ) is the geomagnetic induction unit. (*Records prepared by National Geophysical Data Center, Boulder*)

Magnetic storms. Large geomagnetic disturbances (storms) are approximately one-hundredth of the Earth's main-field strength and are caused by shocks the magnetosphere experiences when significant solar wind disturbances (also known as coronal mass ejections, CME) arrive at the Earth's orbit. As a result, solar-wind-charged particles enter the magnetosphere and increase the Earth's ring current at 3.5 to 9 earth radii. The *H* (magnetic northward) component of the variation field typically shows the greatest arnplitudes (**Fig. 3**). *See* SOLAR CORONA.

Many storms have a similar appearance in the *H* component that is divisible into three parts: sudden commencement, or initial phase; main phase; and recovery phase. The sudden commencements are recorded simultaneously (within minutes) about the entire Earth. Usually there are a sudden onset and then the increase in the northward field strength that may continue for as long as several hours. During the main phase, the *H* component decreases, and this decrease often lasts longer than the initial phase and is several times larger in amplitude. The follow-on recovery to the quiet-time level takes longer than the other two phases and may even extend to several days. The more intense storms show an increase of both amplitude and duration. At some observatories, only the storm's main phase is observed.

Storm intensity and occurrence. The storms are most intense at the latitudes of the nightside auroral zone, where they can be about six to ten times larger than at middle latitudes. They show minimum amplitudes in the region of 20 to 30° geomagnetic latitude on the nightside of the Earth and have a secondary maximum near the dayside equatorial region. At middle latitudes, about 10 storms per year attain a magnitude of over 50 nT; about one or two storms per year attains over 250 nT. There is an increase in activity during the equinoxes, and the storm variation amplitudes are slightly larger in the winter hemisphere. The number and intensity of storms vary with changes in the sunspot number and lag behind the 11-year solar activity cycle by about a year or two. There are direct relationships of the storms with the solar outbursts and solar magnetic field orientation (which changes every 22 years), as well as the high-velocity solar wind and the interplanetary magnetic field direction.

The planetary-wide geomagnetic activity is measured at each magnetic observatory by the local *K* index, a quasilogarithmic scale indicating the range of most disturbed components of the geomagnetic field in a 3-hour interval. The *K* index is obtained from the range of the field changes about the estimated *Sq* variation during the same period; this value is normalized for the observatory location. An average of the indices from selected world observatories provides a "planetary" index, *Kp*. Storm magnitudes are arranged in size either by their highest *Kp* value or by the ring current index, *Dst*.

Polar substorms. A significant amount of energy can be delivered into the Earth's magnetosphere by a group of related physical processes called polar substorms. These phenomena occur when solar wind blows from the Sun at higher (500–800 km/s) than usual (350–400 km/s) velocities. If the incoming solar wind carries along a southward-directed magnetic field (that is, opposite to the Earth's field at its interface with the solar wind at about 10–12 earth radii distance from the Earth's center) for a prolonged time, a polar substorm can be triggered and then observed in the high-latitude magnetic and ionospheric parameter variations. There follows an explosive precipitation of particles into the midnight sector of the ionosphere, a spectacular increase in aurorae, a massive flow of field-aligned currents to and from the auroral region, and the dramatic development of geomagnetic westward and eastward electrojet currents in the auroral zone ionosphere (**Fig. 4**) and in the polar cap. These currents give rise to local heating of the high-altitude atmosphere and to the decrease and violent variations of the *H* magnetic field component seen on the magnetic storm records at auroral latitudes (top four records of Fig. 3). Very short period pulsations (from seconds to minutes) of the geomagnetic field, known as ultralow-frequency (ULF) micropulsations, that are identified with corresponding fluctuations of the auroral luminosity are observed at such times.

The substorm disturbance often proceeds through three stages: (1) the growth phase (of several tens of minutes), representing the time of injection of energy from the nightside magnetosphere; (2) the expansive or explosive phase (several minutes or more), when the disturbance rises to its maximum in amplitude and effective area; and (3) the decay phase (up to an hour or two) as the event subsides. Consecutive substorms can blend (or follow in a sequence) for several hours during the main phase of a magnetic storm. A peak of the substorm activity is usually restricted to a region of less than 5° in latitude and 100° in longitude at the Earth's nightside ionosphere. The ratio of the high- to low-frequency components of the geomagnetic variations decreases rapidly with distance from the disturbance center. Evidence of the substorm is carried to lower latitudes by electric and magnetic fields from the closure of strong auroral-region westward electrojet currents within the ionosphere and from the Birkeland currents of the magnetosphere. The auroral electrojet index, AE, compiled from magnetic records obtained at observatories located in the auroral latitudes, is used as a measure of the substorm intensity. *See* AURORA.

Ring current. During the main phase of the magnetic storms and often during the growth phase of the substorms, energetic ions and electrons are fed into a ringlike region at about 3–9 earth radii distance in the equatorial plane of the Earth. There the complicated field and particle interactions generate a westward-flowing ring current due to a charge separation as the energetic protons and electron move toward the Earth from the magnetospheric tail. By using the right-hand rule, it can be seen that this current causes a worldwide southward field roughly parallel to the Earth's dipole axis. The resulting depression of the measured fields about the entire Earth, particularly apparent at the lower and equatorial latitudes, contributes to the main phase of the geomagnetic storm. With the decay phase of the substorm, the source of maintenance protons disappears; this allows a slow decay of the ring current. This process is seen as a recovery phase of the magnetic storm, and it lasts a few hours to several days, depending upon the intensity of the substorms that might occur during the storm's main phase. A ring current index, *Dst*, is derived from the storm-time (*st*) disturbance (*D*) field levels of low-latitude stations, with the *Sq* variation removed from the data and only the axially symmetric contributions to the field being considered. Actually, at the onset of the storm the positive values of *Dst* represent the compression of the magnetosphere by the solar wind rather than a ring current (which causes negative *Dst* values). These compression and ring current characteristics can be seen as the initial arid recovery phases of the magnetic storms on the low-latitude observatory traces (bottom four) of Fig. 3.

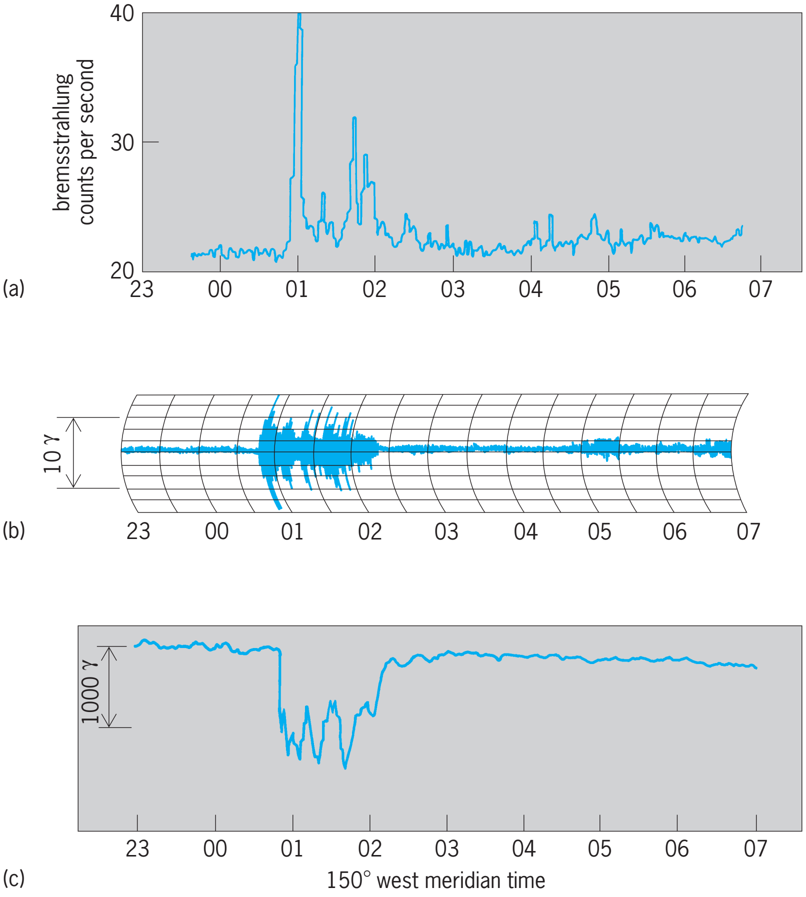


Fig. 4. Concurrent polar substorm phenomena observed at College, Alaska. (*a*) Low-energy electron precipitation into the ionosphere measured as the bremsstrahlung x-ray count rate. (*b*) Geomagnetic field micropulsations recorded by a north-south axis sensor. (c) *H* component of magnetic field. 1 gamma (γ) = 1 nT.

Rapid variations. The geomagnetic spectrum from several minutes to about a third of a second shows activity assolciated with solar–terrestrial disturbances. The irregularly shaped (on an amplitude–time trace) *Pi* pulsations are ideintified with the substorm onset (Fig. 4). The more smoothly varying (continuous) pulsations, *Pc*, also occur in association with the unsettled magnetospheric environment; they are recognized as having special period groups of several minutes (*Pc*4, *Pc*5), about 30 to 15 s (*Pc*3, *Pc*2), and about 0.5 to 5 s (*Pc*1) for which amplitudes near 10, 0.3, and 0.05 nT, respectively, are often reported.

These oscillations arise as hydromagnetic waves whose periods and amplitudes are governed by the charged-particle population and main-field configuration within the magnetosphere and by the transmission of energy from the magnetosphere into the ionosphere. All these variations except *Pc*1 are closely associated with auroral luminosity fluctuations. The *Pc*1 micropulsations (**Fig. 5**) have been shown to arrive at the high-latitude ionosphere as hydromagnetic waves that subsequently propagate in the F-region ionospheric duct

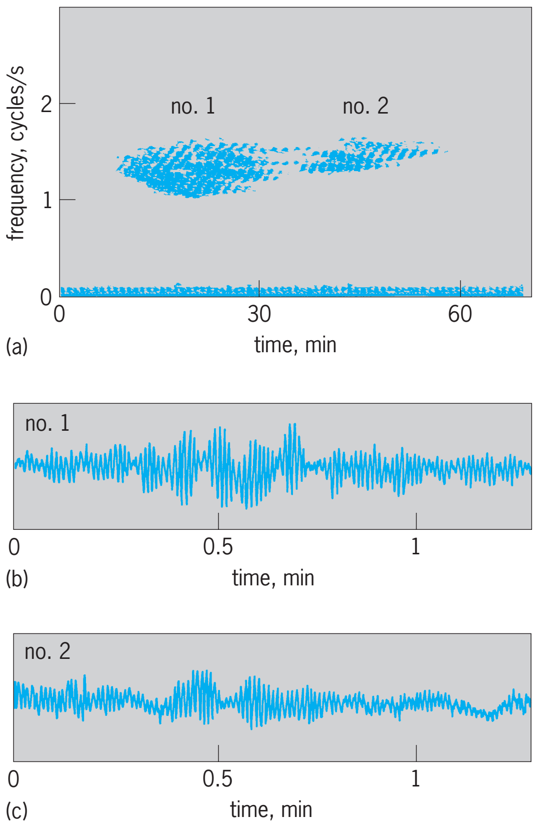


Fig. 5. Example of *Pc*1 geomaignetic pulsation event recorded at Boulder, Colorado. (*a*) Frequency-versus-time display of a two-part event. Note the unique rising frequency structure about a midfrequency near 1.5 cycles/s. (*b*) Amplitude-versus-time representation for no. 1 event and (*c*) no. 2 event. The beating appearance in the amplitudes is largely due to the overlapping elements of the rising frequency structure.

to the lower latitudes. The *Pc*1 pulsations occur most frequently during times of high magnetic activity in the week following major substorms.

Vladimir O. Papitashvili; Wallace H. Campbell

Bibliography. S.-I. Akasofu and S. Chapman, *Solar Terrestrial Physics*, Oxford Clarendon Press, 1972; W. H. Campbell, *Introduction to Geomagnetic Fields*, Cambridge University Press, 1997; S. Chapman and J. Bartels, *Geomagnetism*, vols. 1 and 2, Oxford University Press, 1940; J. Jacobs (ed.), *Geomagnetism*, vol. 3, Academic Press, New York, 1989; F. Lowes et al. (eds.), *Geomagnetism and Paleomagnetism*, 1988; A. Nishida, *Geomagnetic Diagnosis of the Magnetosphere*, Springer-Verlag, New York, 1978; T. A. Potemra (ed.), *Magnetospheric: Currents*, AGU Geophys, Monogr. 28, Washington, DC. 1984.

Geomagnetism

The magnetism of the Earth; also, the branch of science that deals with the Earth's magnetism. Formerly called terrestrial magnetism, geomagnetism involves any topic pertaining to the magnetic field observed near the Earth's surface, within the Earth, and extending upward to the magnetospheric boundary. Modern usage of the term is generally confined to historically recorded observations to distinguish it from the sciences of archeomagnetism and paleomagnetism, which deal with the ancient magnetic field frozen respectively in archeological artifacts and geologic structures. *See* PALEOMAGNETISM; ROCK MAGNETISM.

The primary component of the magnetic field observed at the Earth's surface is caused by electric currents flowing in its liquid core, and is called the main field. Vectorially added to this component are the crustal field of magnetized rocks, transient variations imposed from external sources, and the field from electric currents induced in the Earth from these variations.

Main geomagnetic field. The geomagnetic field is specified at any point by its vector **F**. Its direction is that of a magnetized needle perfectly balanced before it is magnetized, and freely pivoted about that point, when in equilibrium. The north pole of such a needle is the one that at most places on the Earth takes the more northerly position. Over most of the Northern Hemisphere, that pole dips downward (**Fig. 1**). The elements used to describe the vector **F** are *H*, the component of the vector projected onto a horizontal plane; its north and east components *X* and *Y*, respectively; *Z* the vertical component; *F* the magnitude of the vector **F**; the angles *I*, the dip of the field vector below the horizontal; and *D* the magnetic declination or deviation of the compass from geographic north. By convention, *Z* and *I* are positive downward, and *D* is positive eastward (or may be indicated as east or west of north). These elements can be related to each other by trigonometric equations. *See* MAGNETIC COMPASS.

However, when deriving and using accurate models of the field, a spherical coordinate system from

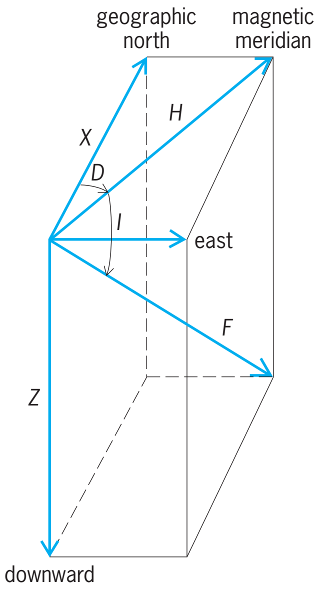


Fig. 1. Elements of the geomagnetic field. *D* = declination, *I* = inclination, *H* = horizontal intensity, *X* = north intensity, *Y* = east intensity, *Z* = vertical intensity, *F* = total intensity.

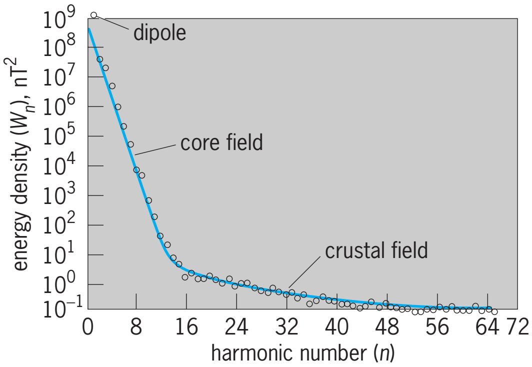


Fig. 2. Energy density spectrum of the geomagnetic field from an analysis of *Magsat* data at an average altitude of 260 mi (420 km).

the center of the Earth is used. The usual spherical coordinates are r, the distance from the Earth's center; θ, the geocentric colatitude; and ϕ, the east longitude. The components of field in this system are F_r, F_θ, and F_ϕ, which are related to those in the topocentric system by Eqs. (1)–(3), where δ is a positive angle

$$X = -F_\theta \cos\delta - F_r \sin\delta \tag{1}$$

$$Y = F_\phi \tag{2}$$

$$Z = -F_r \cos\delta + F_\theta \sin\delta \tag{3}$$

under 0.2°. It can be computed from trigonometric relations by using the accepted oblate spheroid for the Earth. *See* GEODESY.

It is possible to use a scalar potential function because very little current flows in the region between the Earth and the ionosphere, which begins at about 56 mi (90 km) altitude. However, currents above these heights do influence satellite observations, and their determination and removal from such data are the subject of current research. FORTRAN codes utilizing such coefficients to compute field are in general use. *See* MAGNETIC MONOPOLES; SPHERICAL HARMONICS.

A spectrum of the energy density of the field, as observed by the satellite *Magsat* at an average altitude of 260 mi (420 km) (**Fig. 2**), is computed by summing the squares of the coefficients for each n at a given value of r [Eq. (4)]. The spectrum reveals

$$W_n = (n+1)\left(\frac{a}{r}\right)^{2(n+2)} \sum_{m=0}^{n} \left[\left(g_n^m\right)^2 + \left(h_n^m\right)^2\right] \tag{4}$$

the presence of two different sources of the field: one in the core and one in the Earth's crust where the temperature is below the Curie point. This crustal magnetization can be completely due to the present field, or it can also have a remanent component of the ancient field frozen in the rocks. Near the Earth's surface the core field dominates up to about $n = 14$. Beyond $n = 14$ the crustal source is the more important. *See* CURIE TEMPERATURE.

Projection of the core component downward using the term

$$\left(\frac{a}{r}\right)^{2(n+2)}$$

shows a flat or so-called white spectrum just under the seismically determined core–mantle boundary, some 1800 mi (2900 km) below the surface. The crustal spectrum becomes flat no more than a few kilometers below the surface, as would be expected for sources at such a depth.

Magnetic poles. A magnetic pole is defined as a location where the field is vertically aligned, $H = 0$. Due to the presence of sometimes strong (for example, >1000 nT) magnetic anomalies at the Earth's surface, there are a number of locations where the field is locally vertical. However, those field components that extend to sufficient altitude to control charged particles can be accurately located by using the computations from a spherical harmonic expansion using degrees up to only about $n = 10$. Indeed, a pole can be defined by using only the main dipole ($n = 1$), or many terms. *See* AURORA.

The $n = 1$ poles are sometimes referred to as the geomagnetic poles, and those computed using higher terms as dip poles. The term geomagnetic could also refer to the eccentric geomagnetic pole, which can be computed from $n = 1$ and $n = 2$ harmonics so as to be the best representation of a dipole offset from the center of the Earth. The latter has been used as a simplified field model at distances of 3 or 4 earth radii. Due to the more rapid fall-off of the higher terms with distance from the Earth, the two principal poles approach those of the $n = 1$ term with increasing altitude, until the distortions due to external effects begin to predominate. *See* MAGNETOSPHERE.

Whereas for 1990–1995 the geomagnetic poles were located at 79°N and 71°W (and the corresponding southern antipodal points), the addition of higher-degree terms estimates the dip poles to be at 79°N, 105°W and 65°S, 138°E. Transient variations may change these positions up to several tens of kilometers over the course of a day, and more during times of magnetic disturbance.

A number of magnetic coordinate systems depend on the main magnetic field. The invariant latitude is based on the physics of trapped particles and used to organize data from the radiation belts. Corrected magnetic coordinates, based on the first few terms of the spherical harmonic expansion, are used for observations of conjugate geophysical phenomena such as aurora. Conjugate phenomena are those that occur at both ends of a magnetic field line extending from one hemisphere to another. An adjunct to such systems that becomes especially important in polar regions is a concept known as magnetic time, which is based on the $n = 1$ poles. Magnetic noon occurs when the point of observation lies between the $n = 1$ pole and the Sun, and magnetic midnight, when the pole lies between the observer and the Sun. More sophisticated transformations also use some of the lower degrees of n in defining magnetic time.

The distribution of the dip angle I over the Earth's surface can be indicated on a globe or map by contours called isoclines, along which I is constant. The isocline for which $I = 0$ (where a balanced magnetized needle rests horizontal) is called the dip

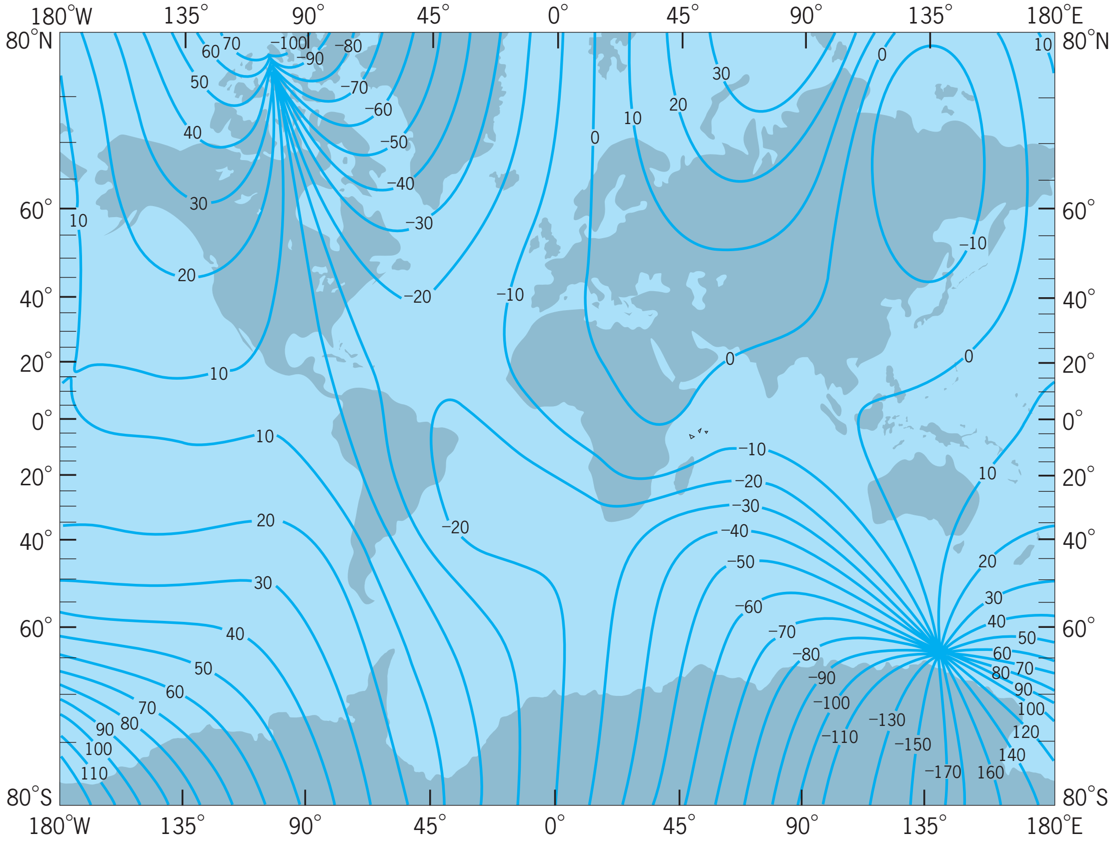


Fig. 3. Lines of equal geomagnetic declination in degrees for current epochs. (*U.S. Naval Oceanographic Office and British Geological Survey*)

equator. The dip equator is geophysically important because there is a region in the ionospheric E layer in which small electric fields can produce a large electric current called the equatorial electrojet. *See* GEOMAGNETIC VARIATIONS; IONOSPHERE.

Magnetic declination. A magnetized compass needle can be weighted to rest and move in a horizontal plane at the latitudes for which it is designed, thus measuring the declination D. The lines on the Earth's surface along which D is constant are called isogonic lines or isogones (**Fig. 3**). The compass points true geographic north on the agonic lines where $D = 0$. At nonpolar latitudes, D is a useful tool for marine and aircraft navigational reference. Indeed, isogones appear on navigation charts, electronic navigational aids are referenced to D, and airport runways are marked with D/10. A runway painted with the number 11 indicates that its direction has a compass heading of 110°. The compass needle becomes less reliable in polar regions because the horizontal component H becomes smaller as the magnetic poles are approached. *See* NAVIGATION.

Intensity patterns. The intensity of the field can also be represented by maps, and the lines of equal intensity are called isodynamic lines (**Fig. 4**). The dipole dominates the patterns of magnetic intensity on Earth in that the intensity is about double at the two poles compared to the value near the Equator. However, it can also be seen that the next terms of the spherical harmonic expansion also have a significant effect, in that there is a second maximum in Siberia, and an area near Brazil that is weaker than any other. This so-called Brazilian anomaly allows charged particles trapped in the magnetic field to reach a low altitude and be lost by collisions with atmospheric gases. The highest intensity of this smooth field is about 70 microteslas near the south magnetic pole in Antarctica, and the weakest is about 23 μT near the coast of Brazil.

Magnetic anomalies. The term anomaly has become clearer than it was previously because it is recognized that the geomagnetic field has a continuous spectrum (Fig. 2) but with two distinct contributors. Originally, the term meant a field pattern that was very local in extent; the modern definition is that portion of the field whose origin is the Earth's crust. The sizes of the strong and easily observable features are generally up to only a few tens of kilometers. Their intensity ranges typically from a few hundred nanoteslas up to several thousand, and they are highly variable depending on the geology of the region.

The former usage has continued in many practical applications where magnetic anomaly maps are constructed from a smooth contouring of aeromagnetic data, ignoring the larger scale but weaker anomalies. Such maps are derived from data taken on a number of flight lines after reduction, using an internationally agreed-upon reference model that is supposed to represent the noncrustal or core component. Such models have been termed IGRF for a

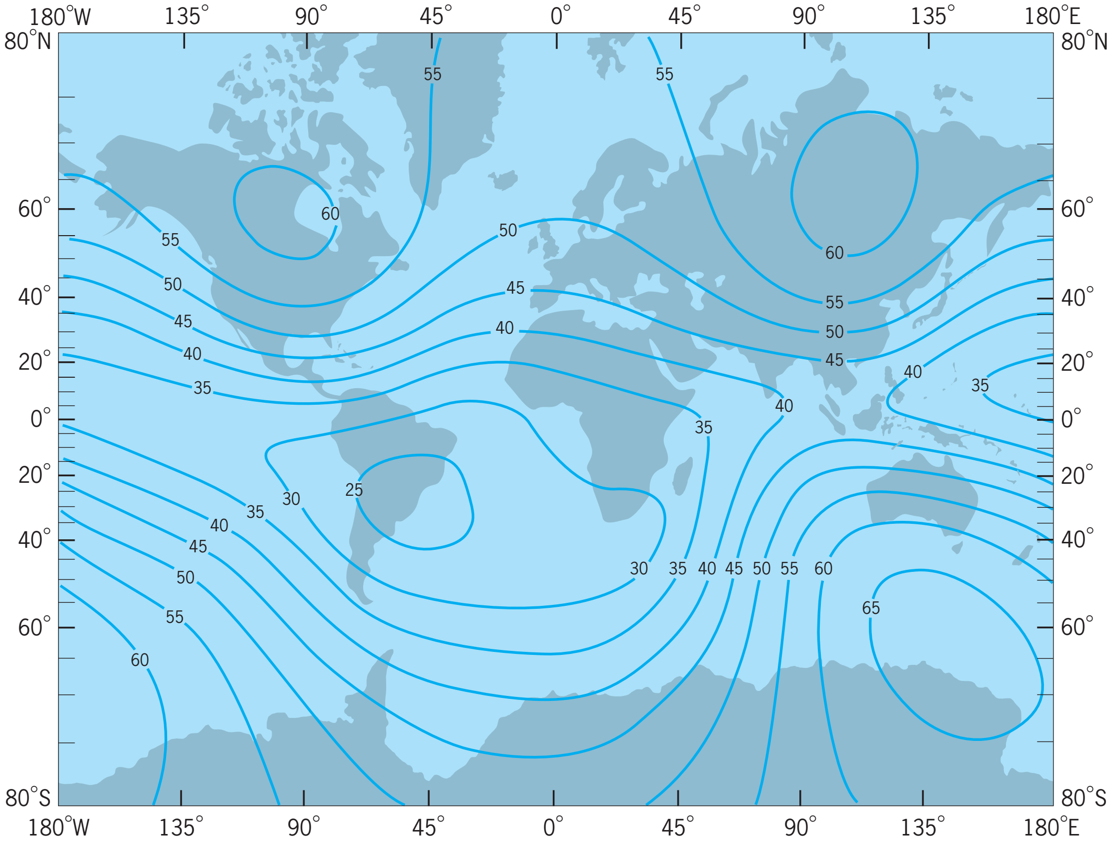


Fig. 4. Lines of equal magnetic intensity *F* in microteslas for recent epochs.

predicted International Geomagnetic Reference Field or DGRF where the attempt is made to correct past predictions and obtain a "definitive" set of spherical harmonic coefficients. Generally, the year or epoch of validity of such models accompanies the term. Thus, IGRF 1995 represents the best prediction of what the field will be at epoch 1995.0. To the present time, the IGRFs have been limited to $n = 8$ because of the difficulty of prediction of more detail with inadequate data, and the DGRFs to $n = 10$. The core field dominates that of the crust up to about $n = 14$ (Fig. 2), so that such models fall short of their goal. Neglect of the terms above $n = 10$ gives errors of the order of a few tens of nanoteslas in estimating the core contribution. Estimation of the core component is complicated by the presence in the lower-degree terms of contributions from crustal magnetization that are comparable in intensity to those seen above $n = 14$.

The best reduction of survey data that can be done is to remove the core component by subtracting a field computed by using the spherical harmonic models limited to $n = 14$ or 15. Formerly, it was thought that the regional or background field could be eliminated by treating data with a high-pass filter. Such filters can be devised to pass data with wavelengths shorter than some preset amount. However, such filtering also removes the long-wavelength components that exist even in the high-degree crustal field contribution. The background field removed with spherical harmonic of degree n has a scale size equal to $40{,}000/n$ (km).

With the advent of satellite surveys, the weaker, larger-size anomalies have become more visible and can be incorporated into spherical harmonic models. Those for Europe using harmonic degrees $n =$ 15 through 50 were derived from the 1979–1980 survey by *Magsat* (**Fig. 5**). The strongest anomaly was found near Kursk, Soviet Union. Surface surveys of this region reveal two narrow strips of anomaly 36 mi (60 km) apart running from the northeast to the southwest. The most disturbed part of the major (northerly) strip is only 1.2 mi (2 km) wide, although the strip is 150 mi (250 km) long; Z is everywhere above normal and ranges up to 190 μT.

Exploration geophysicists work on a much smaller scale and are more concerned with removing the spatial gradients of the observations than with reduction to an absolute level. The practice is to first reduce the contoured residuals of their data; that is, the observed field reduced by that computed from the IGRF, followed by a local two-dimensional fit to further eliminate an average slope across the maps. Such reductions depict the very local anomalies resulting from near-surface mineral deposits, but not the larger-scale features. Maps so constructed generally have discontinuities at their borders when an attempt is made to combine them with others from adjoining regions.

An isolated anomaly such as that at Kursk is

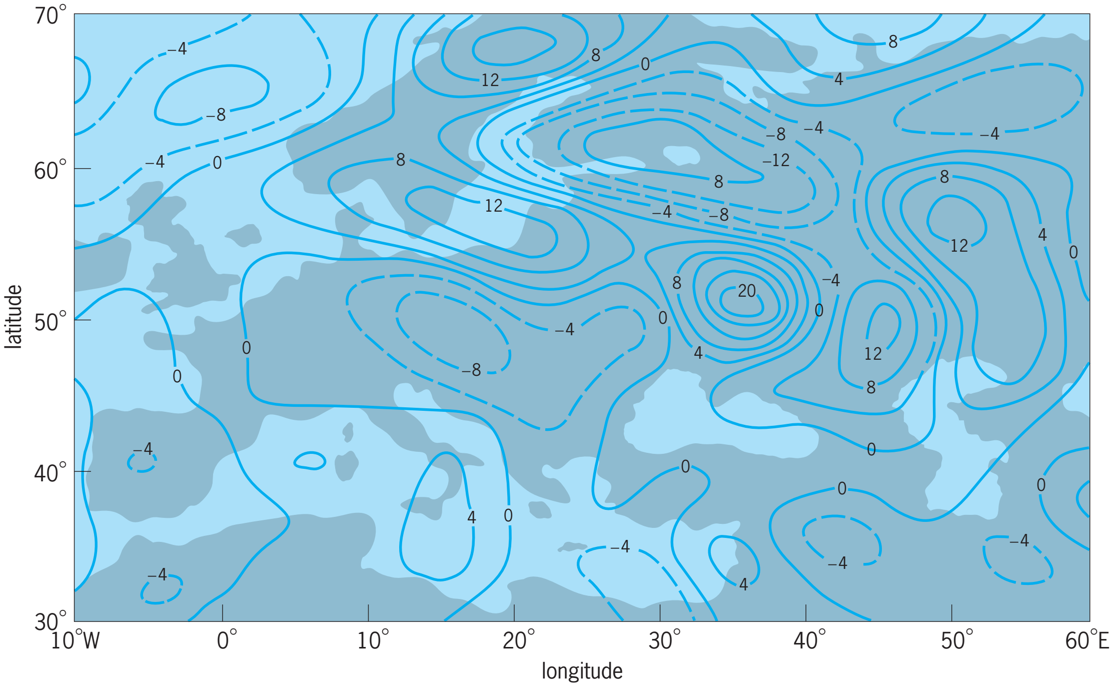


Fig. 5. Crustal anomaly field of the *Z* component as determined from a *Magsat* model (n = 15–50) over Europe for an altitude of 217 mi (350 km). Contours are in nanoteslas.

thought to be due to igneous intrusions. The other, broader patterns are assumed to be from differences of composition in the magnetic basement or crystalline rocks underlying sedimentary deposits, but still not so deep as to be hotter than the Curie point. The differences between continental and oceanic crusts are not clearly evident, nor do the patterns of anomalies about spreading centers show any obvious signatures at satellite altitudes. *See* DIPOLE; DIPOLE MOMENT; GEOPHYSICAL EXPLORATION; MAGNETITE; PLATE TECTONICS.

Secular magnetic variation. The main or core component of the geomagnetic field undergoes slow changes that necessitate continual adjustment of the model coefficients and redrawing of the isomagnetic maps. In any magnetic element at a particular place, the variation may be an increase or a decrease and is not constant in either magnitude or sign. This distribution of the rate for any element can be indicated on isoporic maps by lines (isopors) along which the rate is constant (**Fig. 6**). Typically, the pattern of isopors is more complex than that of the isomagnetic lines for the same element, partly because the spectrum of such change is not dominated by the dipole as is the case of the static field. There has been no accurate satellite monitoring since 1980. The greatest uncertainty in maps of secular magnetic variation are for the southern ocean areas where satellite data are inadequate to overcome the gaps in the distribution of surface magnetic observatories.

Studies indicate that the dipole component of the field 2000 years ago was about 50% stronger than the present. Its average decay rate has averaged about 0.05% per year (15 nanoteslas per year) since about 1840 when absolute measurements were first begun, but accelerated from 1970 to its 1994 value of 0.08% per year (24 nT/yr). However, there is also evidence that the decade of the 1940s showed a rate of only about 10 nT per year. A linear projection of the present rate would have the dipole decreasing to zero in less than 1500 years. Although archeomagnetic evidence indicates that the field has indeed decayed to near zero level within the last 50,000 years with a subsequent return to the present polarity, and paleomagnetic results show that the field has reversed its polarity many times since the Earth's formation (the last time, about a million years ago), there is no model that can predict the future course of field change.

One other smooth variation that has been detected in secular change is the apparent westward drift of some of the patterns of change. Isopors of the vertical field are seen to move preferentially to the west at about 0.2° per year. However, many changes are unpredictable. For example, what has been termed the magnetic jerk was first applied to a discontinuity noted in 1969 (**Fig. 7**). The term jerk is taken from the mechanical analog whereby the acceleration on a moving body is suddenly changed. Between the times of such changes, the secular variations appear to be smooth. Similar changes have probably gone unobserved because of the sparseness of the recorded data. The last irregularity occurred in 1983, though it has not yet been adequately analyzed.

Theory of core field and secular variation. Deriving a suitable model that explains the source of the Earth's magnetic field has been one of the most frustrating problems that theoreticians have faced. Starting with the physical laws that should govern the behavior of a highly conducting, rotating, spherical

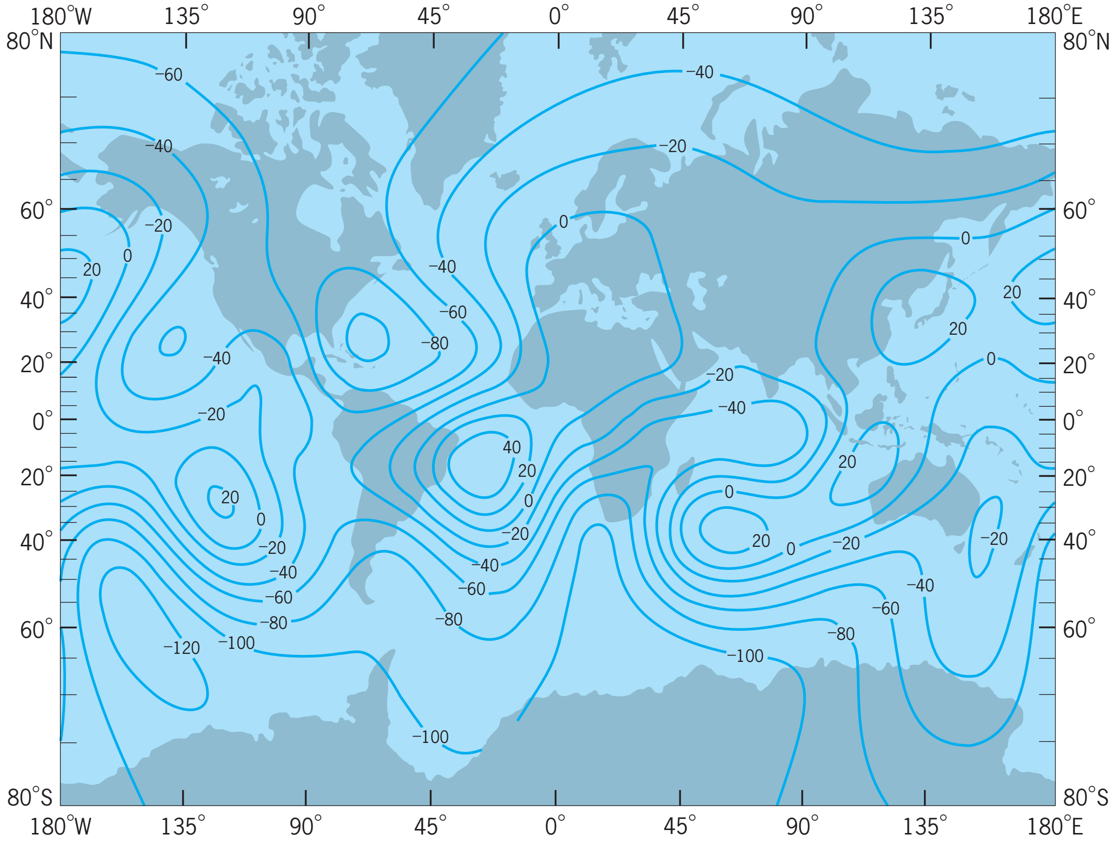


Fig. 6. Contours of equal annual rate of change (isopors) of the total intensity of the field in nanoteslas per year estimated for 1990. (*U.S. Naval Oceanographic Office and British Geological Survey*)

fluid and coming up with a model of the geomagnetic field is exceedingly difficult. The fact that the planets Mercury, Jupiter, Saturn, Uranus, and Neptune also have dipolelike fields indicates that it is a common process, most likely a hydromagnetic dynamo. Dynamo means that a current is generated as an electrical conductor is moved through a magnetic field, as in a dynamo supplying electrical power. *See* GEODYNAMO; JUPITER; MERCURY (PLANET); NEPTUNE; SATURN; URANUS.

Magnetic surveys and models. The main source of data for magnetic maps and models before the advent of satellites was fixed magnetic observatories. These stations, numbering about 140, provided the continuous record of changes in the magnetic field at their location. Their data are generally accurate and an excellent indicator of both secular change and the transient variations, but their global coverage is too sparse for a determination of the whole field. Spherical harmonic analyses based only on such data produce distorted results because of the large gaps in coverage, especially because of the sparseness of observing locations in southern oceanic regions.

An international effort known as the World Magnetic Survey (WMS) was initiated by the International Association of Geomagnetism and Aeronomy (IAGA) during the 1960s to intensify surface and aeromagnetic surveys. A panel was established under IAGA to provide standard magnetic models for public use, and to derive the first such model, IGRF 1965. This project also resulted in a bilateral agreement between the United States and the Soviet Union for the first sharing of satellite-derived data. The spacecraft that were part of this agreement were the POGO half of the OGO series (*Polar Orbiting Geophysical Observatory*; United States) and *Cosmos 49* (Soviet Union; see **table**).

Field models were initially constructed by using only the very accurate scalar data (errors less than 1 nT) from the OGO satellites series. However, it was found that such models were subject to a weakness known as the Backus ambiguity, whereby the vertical component (*Z*) of the field near the magnetic dip equator was poorly determined. It was thus

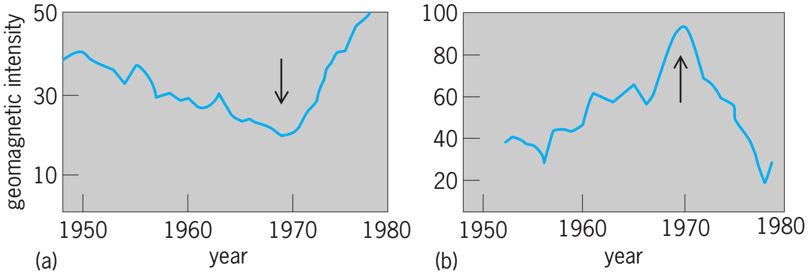


Fig. 7. Secular variation of the geomagnetic intensities (*a*) from the Hartland Magnetic Observatory in Great Britain and (*b*) from the Fredericksburg Observatory in the United States. The arrows point to the time of the jerk or abrupt chang of trend in late 1969 to early 1970. (*After V. Cortillot and J. L. LeMouël, Geomagnetic secular variations impulses, Nature, 311:709–716, 1984*)

Low-altitude satellite geomagnetic measurements, 1958–1993

	Orbit				
Spacecraft	Inclination, degrees	Altitude range, mi (km)	Interval	Instruments	Coverage
Sputnik 3	65	260–360 (440–600)	May–June 1958	Flux gates	Soviet Union
Vanguard 3	33	310–2250 (510–3750)	Sept.–Dec. 1959	Proton precession	Near ground stations
Kosmos 26	49	160–2421 (270–403)	Mar. 1964	Proton precession	Whole orbit
Kosmos 49	50	157–293 (261–488)	Oct.–Nov. 1964	Proton precession	Whole orbit
1964–83C	90	620–653 (1040–1089)	Dec. 1964–June 1965	Rubidium vapor	Near ground stations
OGO 2	87	248–906 (413–1510)	Oct. 1965–Sept. 1967	Rubidium vapor	Whole orbit
OGO 4	86	247–545 (412–908)	July 1967–Jan. 1969	Rubidium vapor	Whole orbit
OGO 6	82	238–659 (397–1098)	June 1969–June 1971	Rubidium vapor	Whole orbit
Kosmos 321	71	162–402 (270–403)	Jan.–Mar. 1970	Cesium vapor	Whole orbit
Magsat	97	211–347 (352–578)*	Nov. 1979–June 1993	Cesium/flux gates	Whole orbit
Dynamic Explorer 2	90	190–630 (309–1012)	Aug. 1981–Feb. 1983	Flux gates	
POGS	90	Circular 470 (750)	Jan. 1991–Sept. 1993	Flux gates	
UARS	57	Circular 360 (585)	Sept. 1991–1994	Flux gates	
Freja	60	370–960 (600–1550)	Oct. 1992–1994	Flux gates	

*The altitude decayed throughout the interval. Some data were taken as low as 110 mi (190 km).

necessary to combine surface vector data with those from satellites. The coefficients of the models were expanded into series, such as $g(t) = g_o + \dot{g}(t - t_o)$, where the secular change terms $\dot{g}$ were determined primarily by the observatory data.

The *Magsat* survey (see table) was the first spacecraft attempt to obtain accurate vector data. The problem in such projects is to determine the spacecraft orientation, or attitude, to sufficient accuracy. Whereas optical pumping magnetometers (such as used on the OGO spacecraft) are accurate to better than 1 nT, this corresponds in a 50,000-nT field to an angular accuracy of 1/50,000 radians, or about 4 seconds of arc. *Magsat* achieved a nominal 20 seconds accuracy during its short data collection interval. The field models for 1980 were determined accurately up to about $n = 50$, beyond which the coefficients reached the noise level (Fig. 2).

Since 1980, there has not been a spacecraft designed to accurately follow secular change or to improve the crustal field definition pioneered by the Kosmos, OGO, and *Magsat* surveys. The satellites after 1980 (see table) make field observations with scalar errors greater than 20 nT and have little knowledge of the attitude. Field models that can be derived with post-1980 surface and satellite data can only bound the errors and provide models useful for applications where high accuracy is not required. *See* MAGNETISM.

Joseph Cain

Bibliography. G. Backus, R. Parker, and C. Constable, *Foundations of Geomagnetism*, 2005; W. H. Campbell, *Introduction to Geomagnetic Fields*, 2d ed., 2003; J. A. Jacobs (ed.), *Geomagnetism*, 4 vols., 1987-1991; J. A. Jacobs, *Reversals of the Earth's Magnetic Field*, 2d ed., 1993; F. J. Lowes et al., *Geomagnetism and Paleomagnetism*, 1988.

Geometric phase

A unifying mathematical concept that describes the relation between the history of internal states of a system and the system's resulting orientation in space. Under various aspects, this concept occurs in geometry, astronomy, classical mechanics, and quantum theory. In geometry it is known as holonomy. In quantum theory it is known as Berry's phase, after M. Berry, who isolated the concept (which was already known in special cases) and explained its wide-ranging significance.

Examples. A few examples of phenomena in which geometric phases appear are discussed below. The first concerns a man sitting on a platform, free to rotate about a perpendicular axis, that is, a soda-fountain stool. The problem is how can he rotate himself without touching anything external or shifting his position on the stool. The solution does not seem so difficult. He might, for example, extend his arm and rotate it to the right. Then he and the platform will rotate together by a smaller amount to the left. This procedure does not yet provide a fully satisfactory solution, however, because he remains with his arm sticking out. He would like a solution such that he returns to his initial posture. Simply to reverse the motion will not do, of course, because during the reverse motion the rotation of his body and the platform will also be canceled. The solution involves pulling the arm back close to his body before undoing its rotation. At the end of this cycle, he and the platform will have rotated together through a nonzero angle, and he will have returned to his initial posture.

A more elaborate form of the same problem concerns an astronaut initially at rest, that is, not rotating in space. The astronaut has the problem of reorienting herself. By simple maneuvers of the type just discussed, extending, rotating, and retracting the arm in different directions, the astronaut can indeed point herself in any desired direction, while returning to her original posture.

The same problem, essentially, must be solved by a cat dropped from rest in an arbitrary orientation. It must perform such contortions to reorient itself and land feet first.

A person's ability to control his orientation contrasts profoundly with his inability to perform a displacement in space. It is a theorem of mechanics that the center of mass of an isolated body moves at a constant velocity, independently of whatever machinations the body might perform.

In the examples discussed so far, a surprising nonzero net rotation or reorientation is found after a cycle involving no net change in configuration, for a system basically at rest. A more complex situation, important for several of the more interesting applications of the geometric phase, occurs when a nonzero rate of rotation is expected, but this rate is manipulated by cyclic changes of configuration.

A classic realization of this situation concerns a spinning particle (henceforth termed simply a spin) interacting with a magnetic field whose magnitude is constant but whose direction can vary. If the direction of the magnetic field were fixed, then the spin would precess around the magnetic field direction, say at a rate ω. Thus after time t, the spin will have precessed through the angle $\theta_{\text{static}} = \omega t$. Now, if the direction of the field is slowly varied through a cycle before it is restored to its original direction, Berry showed that the total angle is instead $\theta_{\text{Berry}} = \omega t + \gamma_{\text{Berry}}$, where γ_{Berry} is half the solid angle swept out by the direction of the field. Thus, the value of γ_{Berry} is independent of almost all details of the cycle and has a purely geometric character. *See* PRECESSION; SPIN (QUANTUM MECHANICS).

Spins can be manipulated, and their directions monitored, by using the techniques of nuclear magnetic resonance. In this way, effects, such as the one just described and others more intricate, have been exhibited experimentally. *See* NUCLEAR MAGNETIC RESONANCE (NMR).

General concept. With these examples, it is possible to gain some understanding of the general concept. A system is envisioned—stool sitter, skater, astronaut, cat, spin in magnetic field—whose possible states can be visualized as points in a suitable abstract space. At the same time, the system has some position or orientation in another space, which in all these examples (but not always) is just ordinary physical space. A history of internal states can be represented by a curve in the first space; the effect of this history on the disposition of the system, by a curve in the second space. The mapping between these two curves is described by the geometric phase. Especially interesting is the case when a closed curve (cycle) in the first space maps onto an open curve in the second, for then there is no net change in internal state, yet the disposition of the system with respect to the outside world is altered.

The power of the geometric phase ideas is that they make it possible, in complex dynamical problems, to find some simple universal regularities without having to solve the complete equations. Significant uses of these ideas include demonstrations of the fractional electric charge and quantum statistics of the quasiparticles in the quantum Hall effect, and of the occurrence of anomalies in quantum field theory. *See* ANYONS; HALL EFFECT; QUANTUM FIELD THEORY.

Frank Wilczek

Bibliography. A. Shapere and F. Wilczek (eds.), *Geometric Phases in Physics*, 1989.

Geometrical optics

The geometry of light rays and their images, through optical systems. Geometrical optics is by far the oldest model proposed for accounting for the behavior of light, going back to classical Greece. It was not until around the beginning of the nineteenth century that the wave nature of light was seriously considered, and in the modern view of the nature of light, geometrical optics as a fundamentally correct model is simply wrong. In spite of this geometrical optics is remarkably robust, remaining as a most practical tool in the solution of optical problems. It has been applied to analyzing laser resonators, to solving problems in interference and diffraction, and even to analyzing the behavior of waveguides, where at first glance it would seem to be totally inappropriate. These developments have been made possible by the generation of "fictitious" rays (all rays are fictitious) or by attributing to rays properties which cannot be accounted for in a strictly geometrical optical model. Nevertheless, the principal application of geometrical optics remains in the field of optical design, where it has been employed since the first optical instruments were developed in the early seventeenth century. This article concentrates on this application. *See* DIFFRACTION; INTERFERENCE OF WAVES; LASER; WAVEGUIDE.

Basic concepts. Light is a form of energy which flows from a source to a receiver. It consists of particles (corpuscles) called photons. All photons of a single pure color have the same energy per photon. Different colors have different energies.

For green light there are 2.5×10^{18} photons per joule of energy, or 2.5×10^{18} photons per second per watt of power. This factor increases for red light and diminishes for blue. *See* LIGHT; PHOTON.

Refractive index and dispersion. The speed with which the particles travel depends on the medium. In a vacuum this speed is 3×10^8 m · s^{-1} (1.86×10^5 mi · s^{-1}) for all colors. In a material medium, whether gas, liquid, or solid, light travels more slowly. Moreover, different colors travel at different rates. This change in speed is a property of the medium, and is measured by the ratio of the speed in a vacuum to the speed in the medium. This ratio is called the refractive index of the medium. The variation in refractive index with color is called dispersion. Generally speaking, blue light travels more slowly than red light, so the medium has a higher refractive index for blue than for red light. In order to deal more precisely and quantitatively with color, and therefore with refractive index and dispersion, it is customary to identify the different colors by their wavelengths, even though waves have no place in geometrical theory. *See* COLOR; DISPERSION (RADIATION); REFRACTION OF WAVES.

Rays. The paths that particles take in going from the source to the receiver are called rays. The particles never interact with each other. The time that it takes a particle to travel from one point to another along a ray is determined by the speed at which the particle travels, which is determined by the refractive index of the medium. The product of the refractive index and the path length is called the optical path length along the ray. The optical path length is equal to the distance that the particle would have traveled in a vacuum in the same time interval.

Sources and wavefronts. A point source is an infinitesimal region of space which emits photons. An extended source is a dense array of point sources. Each point source emits photons along a family of rays associated with it. For each such family of rays there is also a family of surfaces each of which is a surface of constant transit time from the source for all the particles, or alternatively, a surface of constant optical path length from the source. These surfaces are called geometrical wavefronts (**Fig. 1**), although, again, waves have no place in geometrical theory. The reason for this name is that they are often good approximations to the wavefronts predicted by a wave theory. In isotropic media, rays are always normal to the geometrical wavefronts, and the optical path length from one geometrical wavefront to another is the same for all the rays in the family.

Ray paths. The ray path which any particle takes as it propagates is determined by Fermat's principle, which states that the ray path between any two points in space is that path along which the optical path length is stationary (usually a minimum) among all neighboring paths. In a homogeneous medium (one with a constant refractive index) the ray paths are straight lines. In a homogeneous medium containing a point source, the family of rays from the source all radiate outward and the associated geometrical wavefronts are concentric spherical shells.

In a system that consists of a sequence of separately homogeneous media with different refractive indices and with smooth boundaries between them (such as a lens system), the ray paths are straight lines in each medium, but the directions of the ray paths will change in passing through a boundary surface. This change in direction is called refraction, and is governed by Snell's law, which states that the product of the refractive index and the sine of the angle between the normal to the surface and the ray is the same on both sides of a surface separating two media. The normal in question is at the point where the ray intersects the surface (**Fig. 2**).

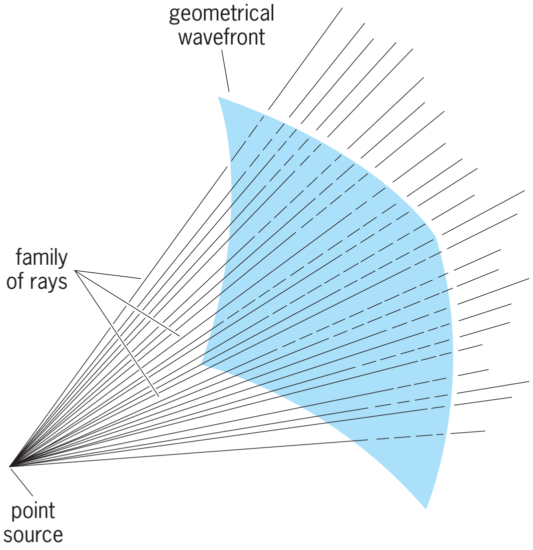


Fig. 1. Geometrical wavefront for a family of rays associated with a point source.

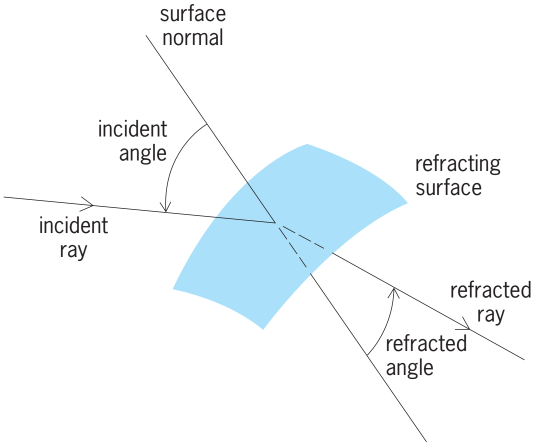


Fig. 2. Refraction of a ray at a smooth surface between two homogeneous media.

Because this product is the same on both sides of the interface, the angle of the ray in the medium with the lower refractive index is always larger than the corresponding angle on the other side of the surface. When the sine of the larger angle becomes one, the largest value it can have, the angle itself is 90°, but the other angle is smaller. The latter angle is called the critical angle for the two media. If the light is traveling from the denser (higher-index) medium to the less dense medium, there is nothing to prevent the incidence angle from being larger than the critical angle, in which case the light is totally reflected by the surface and does not pass through to the other side. This phenomenon is called total internal reflection. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

Ideal image formation. The primary area of application of geometrical optics is in the analysis and design of image-forming systems. An optical image-forming system consists of one or more optical elements (lenses or mirrors) which when directed at a luminous (light-emitting) object will produce a spatial distribution of the light emerging from it which more or less resembles the object. The latter is called an image. *See* OPTICAL IMAGE.

In order to judge the performance of the system, it is first necessary to have a clear idea of what constitutes ideal behavior. Departures from this ideal behavior are called aberrations, and the purpose of optical design is to produce a system in which the aberrations are small enough to be tolerable. *See* ABERRATION (OPTICS).

First consider the requirements for ideal behavior. In an ideal optical system the rays from every point in the object space pass through the system so that they converge to or diverge from a corresponding point in the image space. This corresponding point is the image of the object point, and the two are said to be conjugate to each other (object and image functions are interchangeable). Another way of expressing the same thing is that an ideal optical system converts every spherical geometrical wavefront in the object space into a spherical geometrical wavefront in the image space with the image point located at its center of curvature.

It is not enough, however, to have a system which forms a perfect point image for every point object. The image points must be in the proper geometrical relationship to constitute a good image.

Since there is a one-to-one correspondence between the conjugate object and image points, the geometry of the object and image spaces must be connected by some mapping transformation. The one generally used to represent ideal behavior is the collinear transformation. The intrinsic properties of the collinear transformation are as follows. If three object points lie on the same straight line, they are said to be collinear. If the corresponding three image points are also collinear, and if this relationship is true for all sets of three conjugate pairs of points, then the two spaces are connected by a collinear transformation. In this case, not only are points conjugate to points, but straight lines and planes are conjugate to corresponding straight lines and planes.

It is attractive to have an ideal behavior in which for every point, straight line, or plane in the object space there is one and only one corresponding point, straight line, or plane in the image space. This does not, however, guarantee that the three-dimensional mapping is distortion-free.

Another feature usually incorporated in the ideal behavior is the assumption that all refracting or reflecting surfaces in the system are figures of revolution about a common axis, and this axis of symmetry applies to the object-image mapping as well. With this axial symmetry, an object line which coincides with the axis has as its conjugate an image line also coinciding with the axis. Therefore every object plane containing the axis, called a meridional plane, has a conjugate which is a meridional plane coinciding with the object plane, and every line in the object-space meridional plane has its conjugate in the same plane. In addition, every object plane perpendicular to the axis must have a conjugate image plane which is also perpendicular to the axis, because of axial symmetry. In the discussion below, the terms object plane and image plane refer to planes perpendicular to the axis unless otherwise modified.

An object line parallel to the axis will have a conjugate line which either intersects the axis in image space or is parallel to it. The first case is called a focal system, and the second an afocal system.

Focal systems. The point of intersection of the image-space conjugate line of a focal system with the axis is called the rear focal point (**Fig. 3**). It is conjugate to an object point on axis at infinity. The image plane passing through the rear focal point is the rear focal plane, and it is conjugate to an object plane at infinity. Every other object plane has a conjugate located at a finite distance from the rear focal point, except for one which will have its image at infinity. This object plane is the front focal plane, and its intersection with the axis is the front focal point. Every object line passing through the front focal point will have its conjugate parallel to the axis in image space.

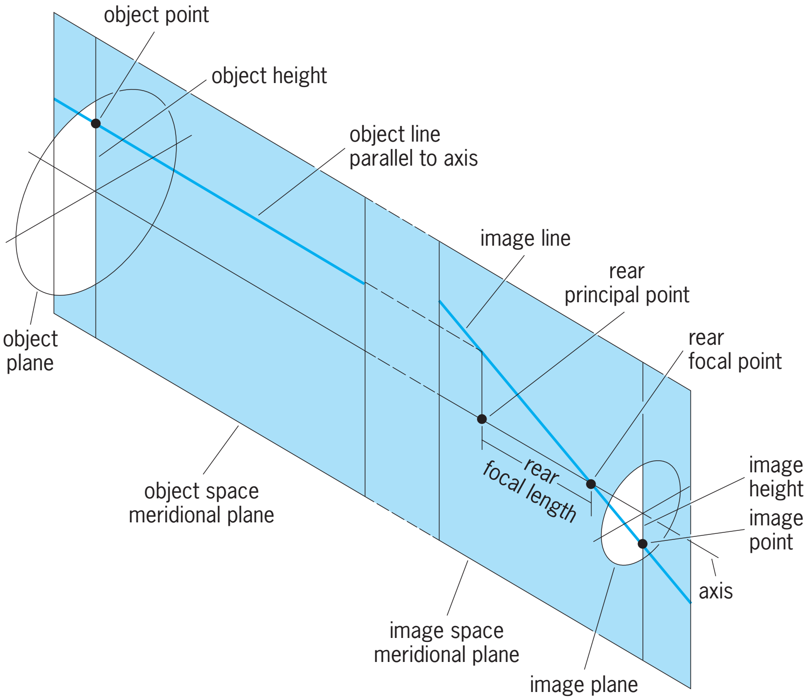


Fig. 3. Focal system.

Now take an arbitrary object plane and its conjugate image plane. Select a point off axis in the object plane and construct a line parallel to the axis passing through the off-axis point. The conjugate line in image space will intersect the axis at the rear focal point and the image plane at some off-axis point. The distance of the object point from the axis is called the object height, and the corresponding distance for the image point is the image height. The ratio of the image height to the object height is called the transverse magnification, and is positive or negative according to whether the image point is on the same or the opposite side of the axis relative to the object point.

Every pair of conjugate points has associated with it a unique transverse magnification, and a given transverse magnification specifies a unique pair of conjugate planes. The rear focal plane and its conjugate have a transverse magnification of zero, whereas the front focal plane and its conjugate have an infinite transverse magnification.

The conjugate pair which have a transverse magnification of $+1$ are called the (front and rear) principal planes. The intersections of the principal plane with the axis are called the principal points. The distance from the rear principal point to the rear focal point is called the rear focal length, and likewise for the front focal length. *See* FOCAL LENGTH.

The focal points and principal points are four of the six gaussian cardinal points. The remaining two are a conjugate pair, also on axis, called the nodal points. They are distinguished by the fact that any conjugate pair of lines passing through them make equal angles with the axis. The function of the cardinal points and their associated planes is to simplify the mapping of the object space into the image space.

In addition to the transverse magnification, the concept of longitudinal magnification is useful. If two planes are separated axially, their conjugate planes are also separated axially. The longitudinal magnification is defined as the ratio of the image plane separation to the object plane separation, and is proportional to the product of the transverse magnifications of the two conjugate pairs. In the limit of the separation between the pairs approaching zero, the longitudinal magnification becomes proportional to the square of the transverse magnification.

Only in planes perpendicular to the axis is there an undistorted mapping, because only in such planes is the magnification a constant over the field. An object plane not perpendicular to the axis has a conjugate plane also inclined to the axis, and the magnification varies over the field, resulting in keystone distortion.

Afocal systems. In the case of afocal systems, any line parallel to the axis in object space has a conjugate which is also parallel to the axis (**Fig. 4**). Conjugate planes perpendicular to the axis are still uniquely related, but the transverse magnification is constant for the system, and the same value is applied to every pair of conjugate planes. The longitudinal magnification, also constant for the system, is proportional to the square of the transverse magnification regardless of the separation of the pair of conjugate planes. Cardinal points do not exist for afocal systems.

The most common use of an afocal system is as a telescope, where both the object and the image are at infinity. The apparent sizes of the object and image are determined by the angular subtense, which is defined, for finite object or image distance, as the ratio of the height to the distance from the observer, and, for infinite distance, by holding this ratio constant as the distance approaches infinity. The concept of angular magnification is therefore useful; this is defined as the ratio of the transverse magnification to the longitudinal magnification, and is inversely proportional to the transverse magnification. The power of a telescope or a pair of binoculars is the magnitude of the angular magnification. *See* MAGNIFICATION; TELESCOPE.

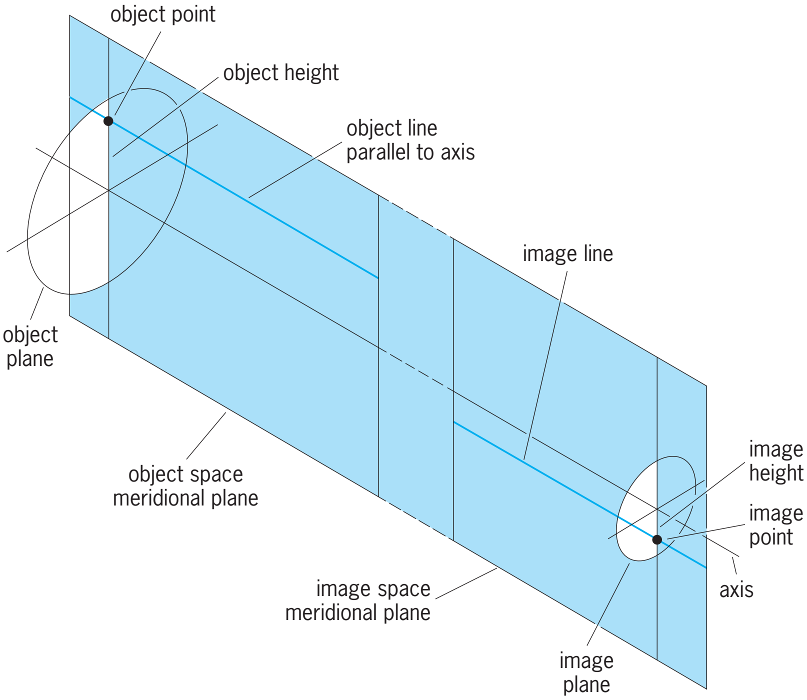


Fig. 4. Afocal system.

Paraxial optics. The above discussion of the properties of the ideal image-forming system did not consider the optical properties of the system where the ray paths are determined by Snell's law. Real optical systems do not, in fact cannot, obey the laws of the collinear transformation, and departures from this ideal behavior are identified as aberrations. However, if the system is examined in a region restricted to the neighborhood of the axis, the so-called paraxial region, where angles and their sines are indistinguishable from their tangents, a behavior is found which is exactly congruent with the collinear transformation. Paraxial ray tracing can therefore be used to determine the ideal collinear properties of the system. If a more extended region than the paraxial is considered, departures from the ideal collinear behavior are observed because additional terms in a series expansion for the trigonometric functions must be taken into account. Nevertheless, even in this extended region the paraxially determined quantities represent the first-order behavior of the system.

Ray tracing is usually done by using a cyclic algorithm, with the coordinate system being transferred from surface to surface through the system as the ray tracing progresses. Two operations are involved, a refraction and a transfer. In refraction, given the incident ray direction and position on the surface, Snell's law is used to determine the emergent direction of the ray. In transfer, the position of the ray on the next surface is determined.

In paraxial ray tracing, which uses the paraxial approximation to Snell's law, the operations of refraction and transfer are simple linear operations, and only two rays need be traced in order to locate the cardinal points of the system. Once the cardinal points have been determined, there are simple equations available to locate any pair of conjugate planes and their associated magnification.

Apertures. The above discussion does not take into account the fact that the sizes of the elements of the optical system are finite, and, in fact, for any optical system, the light that can get through the system to form the image is limited. To determine how much light can get through the system, consider the tracing of a sequence of rays from the axial object point meeting the first surface at increasingly greater heights. Eventually one of these rays and all others beyond it will be blocked by the edge of some aperture in the system. Assuming the aperture is circular and centered on the axis, all the object rays inside the cone determined by the ray which just touches the edge of the aperture will get through the system and participate in the formation of the image. The aperture which limits the cone of rays admitted is called the aperture stop of the system.

An observer who looks into the front of the system from the axial object point sees not the aperture stop itself (unless it is in front of the system), but the image of it formed by the elements preceding it. This image of the aperture stop is called the entrance pupil, and is situated in the object space on the system. The image of the aperture stop formed by the rear elements is the exit pupil, and is situated in the image space of the system. The beam of light which would go from the object point to the image point on axis is spindle-shaped, with a circular cross section everywhere (**Fig. 5*a***).

Light from the edge of the object field will also be limited by the same aperture stop and, to first order, will have the same circular cross section at each surface as the axial beam but will be displaced laterally except at the aperture stop and the pupils. The beam will consist of a sequence of sections of skew circular cones (Fig. 5*b*). The connection between the entrance pupil and the object-space segments of the axial and the off-axis beams is shown in Fig. 5*c*.

It is also possible that a portion of the beam will be further blocked by the edge of some element which, although large enough to admit the axial beam, will not be large enough to admit all of the off-axis beam. This latter effect is called vignetting.

Marginal ray and chief ray. The characteristics of the axial beam are completely specified by a ray traced from the axial object point toward the edge of the entrance pupil. This ray will graze the edge of the aperture stop and emerge from the edge of the exit pupil proceeding to the axial image point. Such a ray is often called a marginal ray, although other names are also common.

The characteristics of the beam from the edge of the field are the same as those of the axial beam except for the displacement of the center of the beam at each surface. Thus the characteristics of this off-axis beam are described by the marginal ray, determining the cross-sectional radius, and a ray traced from the edge of the object field toward the center of the entrance pupil. This ray will pass through the center of the aperture stop and emerge from the center of the exit pupil proceeding to the image point at the edge of the image field. This ray is commonly called the chief ray.

These two rays, the marginal ray and the chief ray, are all that need be traced to determine the first-order properties of the image-forming system.

Lagrange invariant. At any surface in the system the marginal ray and the chief ray are determined by their heights at the surface and angles they make with respect to the axis. Moreover, they are connected by an interesting invariant relationship, often called the Lagrange invariant. This is obtained by multiplying the height of each ray by the product of the refractive index and the angle of the opposite ray, and taking the difference between the resulting two products.

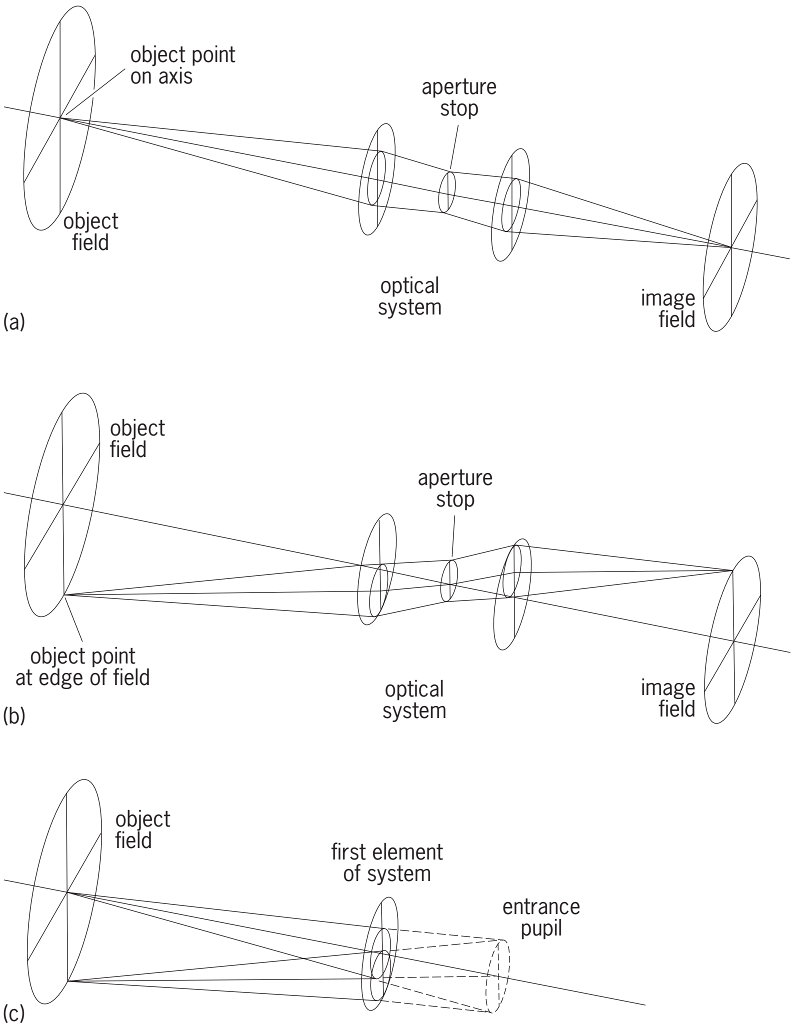


Fig. 5. Optical system with aperture stop. (*a*) Axial beam determined by the aperture stop. (*b*) Off-axis beam determined by the aperture stop. (*c*) Entrance pupil and its connection with the axial and off-axis beams in object space.

Because the Lagrange invariant holds at all surfaces throughout the optical system, in particular the object, the image, and the pupils, it is useful in solving many problems in the early stages of laying out an optical system, especially those involving a sequence of subsystems. It also plays a major role in the radiometry of image-forming systems; the total amount of light which can pass through a system is proportional to the square of the Lagrange invariant.

Although the determination of the first-order properties of an optical system is of major importance, this is only the beginning of the task of designing a good optical system. It is then necessary to trace real rays, determine the aberrations, and adjust the system parameters to improve the system's performance while maintaining its first-order behavior. *See* BINOCULARS; CAMERA; EYEPIECE; GUNSIGHTS; LENS (OPTICS); MIRROR OPTICS; OPTICAL MICROSCOPE; OPTICAL PRISM; OPTICAL PROJECTION SYSTEMS; OPTICAL SURFACES; OPTICAL TRACKING SYSTEMS; PERISCOPE; RANGEFINDER (OPTICS). Roland V. Shack

Bibliography. G. L. James, *Geometrical Theory of Diffraction*, rev. ed., 1986; M. Katz, *Introduction to Geometrical Optics*, 1994; R. Kingslake, *Optical*

System Design, 1983; M. V. Kline and T. E. Furtak, *Optics,* 2d ed., 1986; W. Smith, *Modern Optical Engineering*, 3d ed., 2000.

Geometry

A branch of mathematics concerned with the properties of space, including points, lines, curves, planes and surfaces in space, and figures bounded by them.

Euclidean geometry. Geometry as a high school subject is largely based on the *Elements* of Euclid of Alexandria, a 13-volume work written about 300 B.C., of which the first 6 volumes present plane geometry, the next 4 are concerned with numbers and length, and the last 3 develop solid geometry. Euclid's *Elements* present in a logical order a sequence of over 400 mathematical propositions or theorems, proved on the basis of a set of 10 axioms or postulates which are assumed to be "self-evident" or true without proof—for example: "two points determine a line." Many theorems involve properties of triangles, circles, and other geometric figures, concerning lengths, angles, and congruence.

The most famous of Euclid's assumptions is the fifth, called the parallel postulate, which asserts that: given a line l and a point P not on l, there is one and only one line through P in the plane containing l and P that does not intersect l. Many unsuccessful attempts were made over the centuries to deduce Euclid's parallel postulate from the other axioms. Equivalent to this postulate is the assumption that the sum of the interior angles in a triangle is 180°. In a triangle bounded by arcs of three great circles on a sphere, the sum of the angles always exceeds 180°. *See* EUCLIDEAN GEOMETRY.

Noneuclidean geometries. K. F. Gauss is credited with discovering an "elliptic" noneuclidean geometry in which there are no parallels, but the other euclidean axioms are satisfied. One way to model it is to call each diameter of a sphere a "point," and call each plane through the center (intersecting the sphere in a great circle) a "line." Then each two points determine a unique line and each two lines determine one point.

Another type of noneuclidean geometry, called hyperbolic geometry, was discovered about 1830 by J. Bolyai and N. I. Lobachevski. Through a given point P not on a line l, there are many lines that do not meet l. This geometry may be modeled by defining as "lines" those circular arcs that meet a large circle C at right angles and restricting the term "points" to those within the absolute circle C. Then two points determine a unique line, but there are many lines through a point P not on a line l that do not intersect l. Since consistent geometries exist in which the parallel postulate does not hold, this postulate is independent of the others. *See* NONEUCLIDEAN GEOMETRY.

Projective geometry. Projective geometry studies geometric properties invariant under projections (as in photography) which may distort angles and preserve neither lengths nor ratios of lengths, whereas euclidean geometry is confined to properties of figures such as angles and ratios of lengths that are invariant under rigid motions, reflections, and similarity transformations. If A, B, C, and D are four collinear points, the so-called cross ratio $(BA/BC) \div (DA/DC)$ is preserved under projection. To each set of parallel lines in euclidean geometry is added a "point at infinity" or "vanishing point" in projective geometry, so that each two lines meet in just one point. Lines joining a point P not on l to four points A, B, C, and D on l have the same cross ratio as the four points. A comprehensive theory is derived from the invariance of cross ratio, including a complete theory of conic sections. Hyperbolic, elliptic, and euclidean geometries may be studied in terms of projective transformations that fix respectively a real nondegenerate conic, or an imaginary nondegenerate or degenerate line conic, the latter consisting of two imaginary points at infinity that lie on all circles, but on no other conics. *See* PROJECTIVE GEOMETRY.

Algebraic geometry. Algebraic geometry is a study of solutions of systems of polynomial equations $f_j(x) = 0$ in several variables x_i thought of as coordinates of a point $x = (x_1, x_2, \ldots, x_n)$ in n-dimensional space. Classical studies emphasize properties of an irreducible algebraic plane curve $f(x_1, x_2) = 0$ (f an irreducible polynomial), such as singular and multiple points and genus. The curve is rational (of genus 0) if x_1 and x_2 can be expressed as rational functions of a parameter t. Two curves $f(x_1, x_2) = 0$ and $g(u_1, u_2) = 0$ are birationally equivalent if both the coordinates of each curve can be expressed as rational functions of the coordinates of the other. In modern studies, the coefficients in the polynomials f_j are chosen from an arbitrary field k, and solutions x are sought in an algebraically closed field K containing k. The set X of all points x at which all the f_j vanish is called an algebraic closed set, or variety. If X is not empty, then the set of all polynomials $f(x)$ that vanish on X form an ideal with a finite basis in the ring of polynomials. Modern algebraic geometry studies these ideals and corresponding varieties. *See* POLYNOMIAL SYSTEMS OF EQUATIONS; RING THEORY.

Differential geometry. Differential geometry uses the tools of the calculus to study properties of curves and surfaces, usually in a euclidean space where the squared distance between nearby points is given in rectangular coordinates by the pythagorean formula $ds^2 = dx^2 + dy^2 + dz^2$. Arc length s is defined along "smooth" curves. Geodesics on a surface, like great circles on a sphere, are curves of shortest length between points not too far apart. Curvature is defined at each point of a smooth space curve, and torsion (twisting) at points not belonging to a straight portion of the curve; both curvature and torsion are constant on a helix (similar to a coiled spring). Curvatures of surfaces play an important role. *See* DIFFERENTIAL GEOMETRY.

Riemannian geometry. Riemannian geometry, named for G. F. B. Riemann, generalizes the concepts of differential geometry to noneuclidean spaces of any number n of dimensions in which there may or may not be any "straight" lines of infinite extent. Points P are specified by coordinates x^i like longitude and latitude on the Earth's surface.

Squared distance is expressed by a quadratic form $ds^2 = \int g_{ij} dx^i dx^j$, summed over i and j from 1 to n, where the functions g_{ij} are components of a "metric tensor" that is a scalar function of n vectors, varying in value from point to point. Under a differentiable change of coordinates (somewhat more general than the change from rectangular to spherical coordinates in euclidean analytic geometry), the components g_{ij} are changed. But geodesic paths and a certain curvature obtained from second derivatives of the g_{ij} are independent of the choice of coordinates x^i and describe intrinsic properties of the space. *See* RIEMANNIAN GEOMETRY.

Concepts of riemannian geometry are employed in Albert Einstein's theory of relativity. In his special theory (1905), the invariant interval between events in space-time is $ds^2 = dt^2 - (dx^2 + dy^2 + dz^2)/c^2$, where c is the velocity of light, about 3×10^8 m/s (1.86×10^5 mi/s). Thus neither the apparent time interval dt between events nor the space interval but a combination of the two is the same for all observers. In his general theory (1916), Einstein introduces a more general riemannian metric with g_{ij}'s representing the local gravitational potential, and explains the acceleration ascribed by Newton to a gravitational force as acceleration due to motion along curved world lines in space-time that are bent by the gravitational field. *See* RELATIVITY; SPACE-TIME. J. Sutherland Frame

Bibliography. E. Ballico (ed.), *Projective Geometry with Applications*, 1994; I. Chavel, *Riemannian Geometry: A Modern Introduction*, 1995; H. S. M. Coxeter, *Introduction to Geometry*, 2d ed., 1989; A. Helfer, *Introduction to Modern Differential Geometry*, 1991; W. V. Hodge and D. Pedoe, *Methods of Algebraic Geometry*, 3 vols., 1994; I. R. Shafarevich (ed.), *Algebraic Geometry I: Algebraic Curves, Algebraic Manifolds and Schemes*, 1994.

Geomorphology

The study of landforms, including the description, classification, origin, development, and history of planetary surface features. Emphasis is placed on the genetic interpretation of the erosional and depositional features of the Earth's surface. However, geomorphologists also study primary relief elements formed by movements of the Earth's crust, topography on the sea floor and on other planets, and applications of geomorphic information to problems in environmental engineering.

Geomorphologists analyze the landscape, a factor of immense importance to humankind. Their purview includes the structural framework of landscape, weathering and soils, mass movement and hillslopes, fluvial features, eolian features, glacial and periglacial phenomena, coastlines, and karst landscapes. Processes and landforms are analyzed for their adjustment through time, especially the most recent portions of Earth history.

History. Geomorphology emerged as a science in the early nineteenth century with the writings of James Hutton, John Playfair, and Charles Lyell. These men demonstrated that prolonged fluvial erosion is responsible for most of the Earth's valleys. Impetus was given to geomorphology by the exploratory surveys of the nineteenth century, especially those in the western United States. By the end of the nineteenth century, geomorphology had achieved its most important theoretical synthesis through the work of William Morris Davis. He conceived a marvelous deductive scheme of landscape development through the action of geomorphic processes acting on the structure of the bedrock to induce a progressive evolution of landscape stages.

Perhaps the premier geomorphologist was Grove Karl Gilbert. In 1877 he published his report "Geology of the Henry Mountains." This paper introduced the concept of equilibrium to organize tectonic and erosional process studies. Fluvial erosion was magnificently described according to the concept of energy. Gilbert's monograph "Lake Bonneville" was published in 1890 and described the Pleistocene history of the predecessor to the Great Salt Lake (**Fig. 1**). The monograph is a masterpiece of dynamic analysis. Concepts of force and resistance, equilibrium, and adjustment—these dominated in Gilbert's study of geomorphology. He later presented a

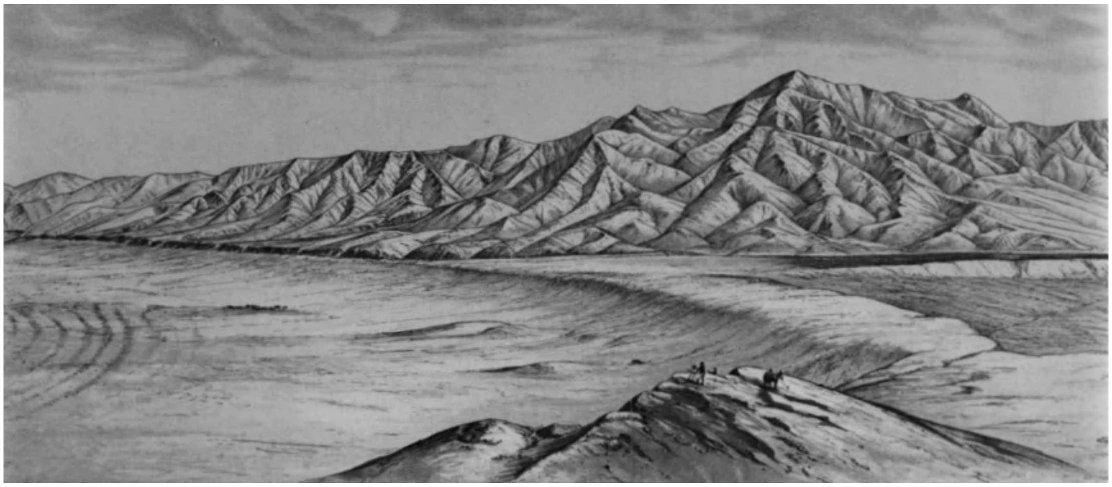


Fig. 1. The great bar of Pleistocene Lake Bonneville at Stockton, Utah. (*After G. K. Gilbert, U.S. Geol. Surv. Monogr. 1, 1890*)

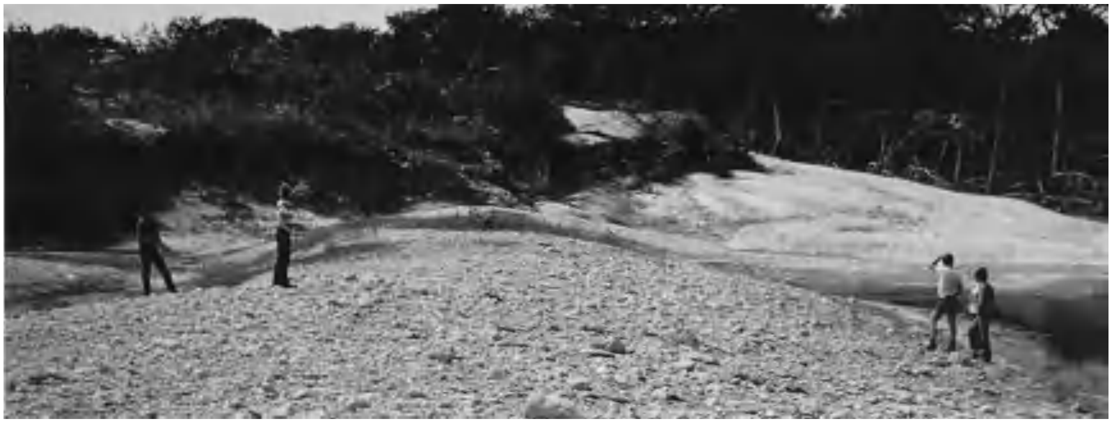


Fig. 2. Surveying large transverse gravel bars created by flooding of the Medina River, Texas, in August 1978.

thorough analysis of fluvial sediment transport and the environmental effects of altered fluvial systems. He even made a perceptive study of the surface morphology of the Moon.

Despite Gilbert's example, geomorphologists in the early twentieth century largely worked on landscape classification and description according to Davis's theoretical framework. Toward the middle of the twentieth century, alternative theoretical approaches appeared. Especially in France and Germany, climatic geomorphology arose on the premise that distinctive landforms and processes are associated with certain climatic regions. Geomorphology since 1945 has become highly diversified, with many groups specializing in relatively narrow subfields, such as karst geomorphology, coastal processes, glacial and Quaternary geology, and fluvial processes.

Process geomorphology. Modern geomorphologists emphasize basic studies of processes presently active on the landscape (**Fig. 2**). This work has benefited from new field, laboratory, and analytical techniques, many of which are borrowed from other disciplines. Geomorphologists consider processes from the perspectives of pedology, soil mechanics, sedimentology, geochemistry, hydrology, fluid mechanics, remote sensing, and other sciences. The complexity of geomorphic processes has required this interdisciplinary approach, but it has also led to a theoretical vacuum in the science. At present many geomorphologists are organizing their studies through a form of systems analysis. The landscape is conceived of as a series of elements linked by flows of mass and energy. Process studies measure the inputs, outputs, and transfers for these systems. Although systems analysis is not a true theory, it is compatible with the powerful new tools of computer analysis and remote sensing. Systems analysis provides an organizational framework within which geomorphologists are developing models to predict selected phenomena.

The future. Geomorphology is increasing in importance because of the increased activity of humans as a geomorphic agent. As society evolves to more complexity, it increasingly affects and is threatened by such geomorphic processes as soil erosion, flooding, landsliding, coastal erosion, and sinkhole collapse. Geomorphology plays an essential role in environmental management, providing a broader perspective of landscape dynamics than can be given by standard engineering practice.

The phenomenal achievements of nineteenth-century geomorphology were stimulated by the new frontier of unexplored lands. The new frontier for geomorphology in the late twentieth century lies in the study of other planetary surfaces (**Fig. 3**). Each new planetary exploration has revealed a diversity of processes that stimulates new hypotheses

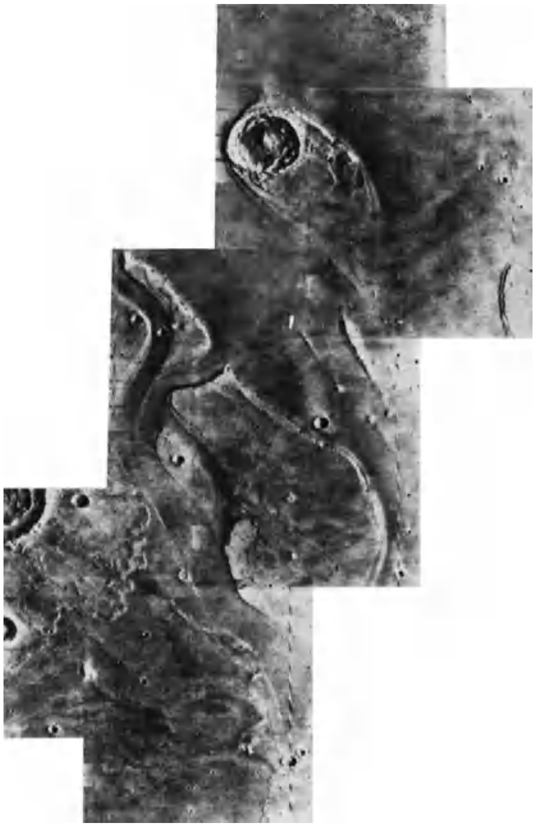


Fig. 3. Streamlined uplands and large sinuous channels in the Chryse Planitia region of Mars. (*National Aeronautics and Space Administration*)

for features on Earth. Geomorphology must now solve the mysteries of meteor craters on the Moon and Mercury, great landslides and flood channels on Mars, phenomenally active volcanism on Io, and ice tectonics on Ganymede. *See* COASTAL LANDFORMS; EROSION; GLACIATED TERRAIN; KARST TOPOGRAPHY.

Victor R. Baker

Bibliography. V. R. Baker and S. J. Pyne, G. K. Gilbert and modern geomorphology, *Amer. J. Sci.*, 278:97–123, 1978; A. L. Bloom, *Geomorphology*: *A Systematic Analysis of Late Cenozoic Landforms*, 3d ed., 1997; A. F. Pitty, *Geomorphology: Themes and Trends*, 1985; D. F. Ritter, *Process Geomorphology*, 3d ed., 1995.

Geophagia

Soil ingestion by animals. Grazing animals such as sheep and cattle ingest varying amounts of soil when they graze herbage contaminated with it.

Pastures become contaminated with soil when livestock walk across the herbage, particularly in wet conditions, as the treading action pushes the plants against the soil surface while soil also brushes off their hooves. An increase in the amount of soil ingested by the animals is associated with an increase in the ash content of the feces. Thus, the daily intake of soil can be determined from the daily fecal output and the fecal ash content.

The amount of soil ingested by sheep and cattle is influenced by soil type, stock density, earthworm activity, management practices, and various seasonal factors. Soils that are well drained and have a strong structure do not break up so readily and contaminate the herbage as is the case for poorly drained, weak-structured soils. When the density of stock grazing in a given area of herbage is increased, the amount of treading is increased, while the herbage is grazed more closely. The overall effect is that more soil is transferred to the herbage and ingested. Earthworm casts deposited at the soil surface are also ingested when the herbage is closely grazed.

Geophagia is subject to seasonal variations. The wetter and cooler conditions of autumn and winter result in muddier herbage and an increase in soil ingestion by grazing animals. During the spring and summer, the greater growth of the herbage and drier conditions result in cleaner herbage, and there is a marked decrease in intake of soil.

The ingestion of soil affects the teeth of sheep. For instance, under the conditions prevailing in intensive systems of sheep farming, the incisor teeth of the sheep slowly wear down, so that the older animals have only a small amount of tooth showing above the gum. The rate of teeth wear has been found to be associated with the amount of soil ingested, the causative agent being the abrasive action of the soil.

Soil can be a source of mineral nutrients. Since soils are higher than herbage in iron, manganese, zinc, copper, cobalt, selenium, and iodine, they may contribute to the mineral nutrition of the grazing animals. *See* AGRICULTURAL SCIENCE (ANIMAL); AGRICULTURE; SOIL CHEMISTRY.

Neville D. Grace

Geophysical exploration

Making, processing, and interpreting measurements of the physical properties of the Earth with the objective of practical application of the findings. Most exploration geophysics is conducted to find commercial accumulations of oil, gas, coal, or other minerals, but geophysical investigations are also employed with engineering objectives, in studies aimed at predicting the nature of the Earth for the foundations

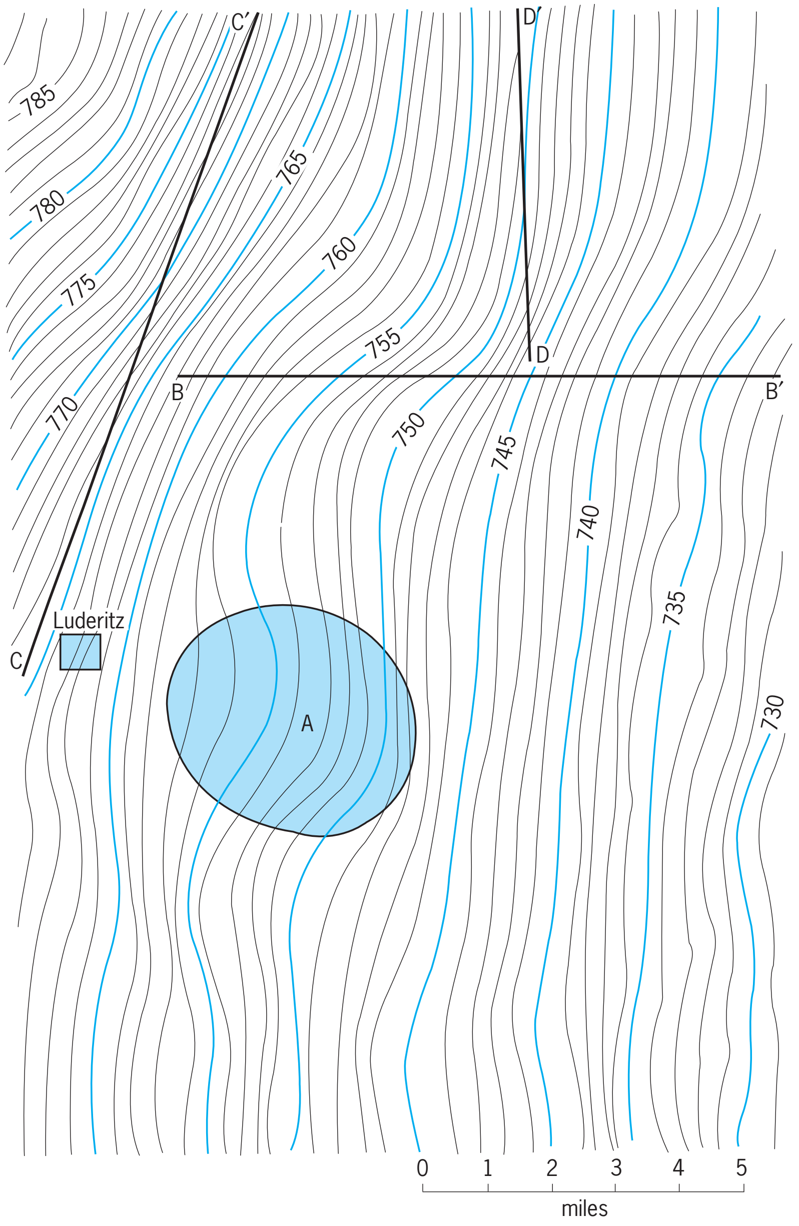


Fig. 1. Bouguer gravity of a portion of the Perth Basin, Western Australia. Contour interval is 1 mGal; datum is arbitrary. Departures from regularity are called anomalies. The bulge *A* results from an uplift area, the contour offset along *BB′* is an east-west trending fault, downthrown to the south. Other faults (*CC′* and *DD′*) are indicated by a closer spacing of the contours; both are downthrown to the east. Some of the variations in contour spacing result from measurement errors, some from interpolating between control points. 1 mi = 1.6 km. (*Western Australian Petroleum Pty. Ltd.*)

of roads, buildings, dams, tunnels, nuclear power plants, and other structures, and in the search for geothermal areas, water resources, pollution, archeological ruins, and so on.

Geophysical exploration, also known as applied geophysics or geophysical prospecting, is often divided into subsidiary fields according to the property being measured, such as magnetic, gravity, seismic, electrical, electromagnetic, thermal, or radioactive. *See* GEOPHYSICS.

General principles. A number of principles apply to most of the different types of geophysical exploration. Occasionally, prospective features can be mapped directly, such as iron deposits by their magnetic effects, but most features are studied indirectly by measuring the properties or the geometry of rocks that are commonly associated with certain mineral deposits.

Ordinarily, an anomaly is sought, that is, a departure from the uniform geologic characteristics of a portion of the Earth (**Fig. 1**). The primary objective of a survey is usually to determine the location of such departures. Sometimes, areas of anomalous data are obvious, but more often they are elusive because the anomaly magnitude is small compared to the background noise or because of the interference of the effects of different features. A variety of averaging and filtering techniques are used to accentuate the anomalous regions of change.

An anomaly usually seems smaller as the distance between the anomalous source and the location of a measurement increases (**Fig. 2**). Hence, a nearby source usually produces a sharp anomaly detectable only over a limited region, although possibly of large magnitude in this region. The detail of measurement required to locate anomalies must be compatible with the depth of the sources of interesting anomalies. If the source of an anomaly is deep in the Earth, then the anomaly is spread over a wide area, and its magnitude is small at any given location. As the depth of the anomaly increases, more sensitive instruments are needed because the effects become much smaller. Hence, the depth of the feature sought governs both the amount of detail and the precision required in measurements. Many of the differences in geophysical methods derive from the different depths of interest. Engineering, mineral, and ground-water objectives are usually shallow, often less than 100 ft (30 m), whereas petroleum and natural gas accumulations are usually quite deep: 0.6–5 mi (1–8 km).

Geophysical data usually are dominated by effects that are of no interest, and such effects must be either removed or ignored to detect and analyze the anomalous effects being sought. Noise caused by near-surface variations is especially apt to be large. The averaging of readings is the most common way of attenuating such noise.

The interpretation of geophysical data is almost always ambiguous. Since many different configurations of properties in the Earth can give rise to the same data, it is necessary to select from among many possible explanations those that are most probable, and (usually) to select from among the probable explanations the few that are most optimistic from the point of view of achieving set objectives. In engineering and pollution applications, the most pessimistic interpretation may be sought so that the worst situation can be studied.

Geologic features affect the various types of measurements differently; hence more can be learned from several types of measurements than from any one alone. Combinations of methods are particularly useful in mining exploration. In petroleum surveys,

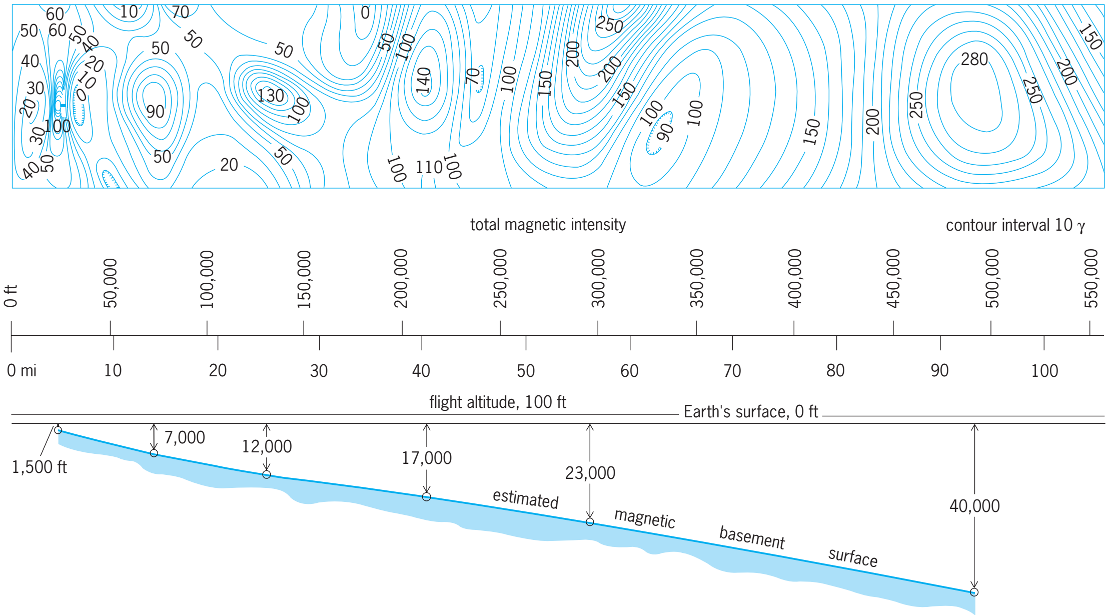


Fig. 2. Portion of a magnetic map (top) and interpretation (below). On the map, sharper features at the left indicate basement rocks are shallow, whereas broad anomalies to the right indicate deep basement. 1 ft = 0.3 m; 1 mi = 1.6 km. (*After S. L. Hammer*, *4th World Petroleum Congress Proceedings*, *Rome*, *Sec. 1*, *1955*)

magnetic, gravity, and seismic explorations are apt to be performed in that sequence, which is the order of relative cost. Less expensive methods are used first to narrow down the region to be explored by more expensive methods.

Usually the properties that geophysicists are able to measure are not directly related to objectives of interest; hence some association must be developed between the measured properties and features of interest. Geophysical measurements can determine something about the location, lateral extent, and depth of the source of an anomaly but not its specific cause. For example, an anomalous area located by ground-penetrating radar might be caused by a multitude of things, many of them of no interest. Interpretation utilizes other known information; for example, knowledge that potentially hazardous materials have been buried in the area might suggest this explanation for an anomaly. Reasoning is usually somewhat inferential; an ore often associated with a geological feature might be found by looking near geophysical indications of these features. The inference that the factors that produced a particular structural feature may have also affected sedimentation may lead to the discovery of a stratigraphic accumulation. *See* STRATIGRAPHY.

Magnetic surveying. Rocks and ores containing magnetic minerals become magnetized by induction in the Earth's magnetic field so that their induced field adds to the Earth field. Magnetic exploration involves mapping variations in the magnetic field to determine the location, size, and shape of such bodies. *See* GEOMAGNETISM.

The magnetic susceptibility of sedimentary rock is generally orders of magnitude less than that of igneous or metamorphic rock. Consequently, the major magnetic anomalies observed in surveys of sedimentary basins usually result from the underlying basement rocks. Determining the depths of the tops of magnetic bodies is thus a way of estimating the thickness of the sediments. Gradiometers are used to detect sedimentary structures such as faults. *See* ROCK MAGNETISM.

Except for magnetite and a very few other minerals, mineral ores are only slightly magnetic. However, they are often associated with bodies such as dikes that have magnetic expression so that magnetic anomalies may be associated with minerals empirically. For example, placer gold is often concentrated in stream channels where magnetite is also concentrated.

Instrumentation. Several types of instruments are used for measuring variations in the Earth's magnetic field. Because the magnetic field is a vector quantity, its magnitude and direction can be measured or, alternatively, components of the field in different directions. Usually, however, only the magnitude of the total field is measured.

Optically pumped, proton and fluxgate magnetometers are used extensively in magnetic exploration. Although it is fairly easy to achieve magnetometers with even greater sensitivity, those used in exploration typically are accurate to 0.1–10 nanoteslas (0.1–10 gammas), which is compatible with uncertainties in noise background. More sensitive magnetometers such as the cryogenic superconducting quantum interference device (SQUID) are sometimes used where more precise measurements are needed. *See* MAGNETOMETER; SQUID.

Field methods. Most magnetic surveys are made by aircraft, because a large area can be surveyed in a short time, and thus the cost per unit of area is kept very low. Aeromagnetic surveying is especially adapted to reconnaissance, for locating those portions of large, unknown areas that contain the best exploration prospects so that future efforts can be concentrated there. Other types of measurements (for example, electromagnetic, gamma-ray) are often made simultaneously.

The spacing of measurements must be finer than the size of the anomaly of interest. Petroleum exploration usually concentrates only on large anomalies, hence a survey for such objectives may involve flying a series of parallel lines spaced 0.6–2 mi (1–3 km) apart, with tie lines (perpendicular lines) every 6–9 mi (10–15 km) to assure that the data on adjacent lines can be related properly. The flight elevation is usually 900–3000 ft (300–1000 m). In mineral exploration, lines are usually located much closer—sometimes less than 300 ft (100 m) apart—and the flight elevation is as low as safety permits. Helicopters are sometimes used for mineral magnetic surveys.

Aircraft are usually equipped with Global Positioning System (GPS) and Doppler radar navigation and with both aneroid and radar altimeters so that the locations of measurements are known accurately. Aircraft also use radionavigation measurements and aerial photographs or other means to locate the aircraft. *See* AERIAL PHOTOGRAPHY; ALTIMETER; DOPPLER RADAR; SATELLITE NAVIGATION SYSTEMS.

The immediate product of aeromagnetic surveys is a graph of the magnetic field strength along lines of traverse. After adjustment, the data are usually compiled into maps on which magnetism is shown by contours (isogams) that connect points of equal magnetic field strength.

In ground mineral exploration, the magnetic field is measured at closely spaced stations. The effects of near-surface magnetic bodies is accentuated over measurements made in the air. Magnetic surveys are also carried out on the ground to delineate near-surface features such as buried drums and tanks or archeological artifacts. Magnetic gradient measurements [measurements made at nearby points (sometimes 10 ft or 3 m apart) so that differences give the magnetic gradients] are especially sensitive to near-surface features.

Magnetic surveying is often done in conjunction with other geophysical measurements, because it adds only a small increment to the cost and the added information often helps in resolving interpretational ambiguities.

Data reduction and interpretation. The reduction of magnetic data is usually simple. Often, measurement conditions vary so little that the data can be interpreted directly, or else require only network adjustments to minimize differences at line intersections. The magnetic field depends on the elevation at which it is measured, but data can be continued to a different elevation; that is, the magnetic field at one elevation can be determined from knowledge of the field at a different elevation. Where different parts of an area have been surveyed at different elevations, continuation can be used to reconcile them. In surveys of large areas, the variations in the Earth's overall magnetic field may be removed (magnetic latitude correction). In exceptional cases, such as in land surveys made over very irregular terrain, as in bottoms of canyons where some of the magnetic sources may be located above the instrument, reduction of the data can become difficult.

The sharpness of a magnetic anomaly depends on the distance to the magnetic body responsible for the anomaly. Inasmuch as the depth of the magnetic body is often the information being sought, the shape of an anomaly is the most important aspect. Modeling is used to determine the magnetic field that would result from bodies of certain shapes and depths. The model anomalies are examined for a parameter of shape that is proportional to the depth (**Fig. 3**). The shape parameter is measured on real anomalies and scaled to indicate how deep the body responsible for the anomaly lies. Such estimates are typically accurate to 10–20%, sometimes better.

Iterative modeling techniques are used in more detailed studies. The field indicated by a model is subtracted from the observed field to give an error field. Then the model is changed to obtain a new error field. This process is repeated until the error field is made sufficiently small. The model then represents one possible explanation of the anomaly.

One commonly employed interpretation technique is to assume that the measured field results from near-vertical dikes or the edges of thin horizontal sheets. Seven or so successive equally spaced measurements can then be solved for the depth and other parameters of anomalies. Computers solve for each successive set of measurements, and then sophisticated filtering techniques eliminate solutions that are not consistent.

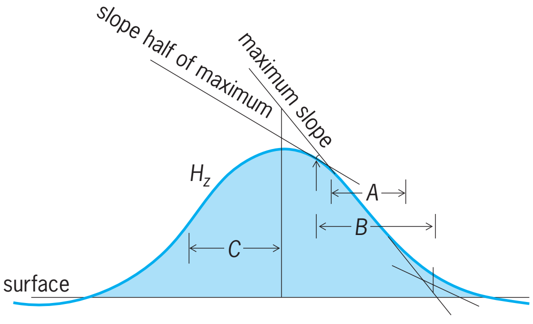


Fig. 3. Variation of shape measurements across an anomaly. Among the shape factors sometimes measured are *A*, the distance over which the slope is maximum; *B*, the distance between points where the slope is half of the maximum slope; *C*, half the width of the anomaly at half the peak magnitude. Such shape measurements multiplied by index factors give estimates of the depth of the body responsible for the anomaly.

Gravity exploration. Gravity exploration is based on the law of universal gravitation: the gravitational force between two bodies varies in direct proportion to the product of their masses and in inverse proportion to the square of the distance between them.

Because the Earth's density varies from one location to another, the force of gravity varies from place to place. Gravity exploration is concerned with measuring these variations to deduce something about rock masses in the immediate vicinity. *See* EARTH, GRAVITY FIELD OF.

Vertical density changes affect all stations equally and so do not produce easily measured effects. Gravity field variations are produced by lateral changes in density. Absolute density values are not involved, only horizontal changes in density. The product of the volume of a body and the difference between the density of the body and that of the horizontally adjacent rocks is called the anomalous mass.

Gravity surveys are used more extensively for petroleum exploration than for metallic mineral prospecting. The size of ore bodies is generally small; therefore, the gravity effects are quite small and local despite the fact that there may be large density differences between the ore and its surroundings. Hence, gravity surveys to detect ore bodies have to be very accurate and very detailed. In petroleum prospecting, on the other hand, the greater dimensions of the features more than offset the fact that density differences are usually smaller. *See* PETROLEUM; PROSPECTING.

Instrumentation. The most common gravity instrument in use is the gravity meter or gravimeter. The gravimeter basically consists of a mass suspended by springs comprising a balance scale. The gravimeter can be balanced at a given location, then moved to another location, and the minute changes in gravitational force required to rebalance the instrument can be measured. Hence the gravimeter measures differences in a gravity field from one location to another rather than the gravity field as a whole.

A gravimeter is essentially a very sensitive accelerometer, and extraneous accelerations affect the meter in the same way as the acceleration of gravity affects it. Typically, gravimeters read to an accuracy of 0.01 mGal, which amounts to 1/100,000,000 of the Earth's gravitational field. Anomalies of interest in petroleum exploration are often of the magnitude of 0.5 to 5 mGal. Extremely sensitive gravimeters measuring to microgal accuracy are used in boreholes to locate cavities such as caves and tunnels, and in archeological applications to search for burial chambers. *See* ACCELEROMETER; GRAVITY METER.

Field surveys. Almost all gravity measurements are relative measurements; differences between locations are measured although the absolute values remain unknown. Ordinarily, the distance between the stations should be smaller than one-half the depth of the structures being studied.

Gravity surveys on land usually involve measurements at discrete station locations. Such stations are spaced as close as a few meters apart in some mining or archeological surveys, about 0.3 mi (0.5 km) for petroleum exploration, and 6–10 mi (10–20 km) for some regional geology studies. While it is desirable to have gravity values on a uniform grid, often this is not convenient, and so stations are located on traverses around loops. For petroleum exploration, the gravimeter might be read every 0.3 mi (0.5 km) around loops of about 4 by 6 mi (6 by 10 km). Helicopters are used for transport between stations in areas of difficult terrain. Location and elevation are then determined by an inertial navigation system, also carried by the helicopter.

The gravity field is very sensitive to elevation. An elevation difference of 9 ft (3 m) represents a difference in gravity of about 1 mGal. Hence, elevation has to be known very accurately, and the most critical part of a gravity survey often is determining elevations to sufficient accuracy.

Gravity measurements can be made by ships at sea. Usually the instrument is located on a gyrostabilized platform which holds the meter as nearly level as possible. The limiting factor in shipboard gravity data is usually the uncertain velocity of the meter, especially east to west, since the ship is moving. The velocity of a ship traveling east adds to the velocity because of the rotation of the Earth. Consequently, centrifugal force on the meter increases and the observed gravity value decreases (Eötvos effect).

Gravity measurements are also sometimes made by lowering a gravimeter to the ocean floor and balancing and reading the meter remotely. Gravity measurements have been made by aircraft using techniques like those used at sea, but are not sufficiently accurate to be useful for most exploration.

Specialized gravimeters are used to make measurements in boreholes. The main difference between gravity readings at two depths in a borehole is produced by the mass of the slab of earth between the two depths; this mass pulls downward on the meter at the upper level and upward at the lower level. Thus the difference in readings depends on the density of this slab. In sedimentary rocks, the borehole gravimeter is used primarily for measuring porosity. The density of most minerals in sedimentary rocks is about the same, but very different from water-filled pore spaces.

Variations in the Earth's gravity field affect sea level. Orbiting satellites can measure their elevation with respect to sea level with sufficient accuracy to map variations in the Earth's field over the oceans. Satellites can measure gravity anomalies at sea that are larger than about 5 mGal and 15 mi (25 km) width.

Data reduction. Gravity measurements have to be corrected for factors other than the distribution of the Earth's mass. Meters drift or change their reading gradually because of various reasons. The Sun and Moon pull on the meter in different directions during the course of a day. The gravitational force varies with the elevation of the gravimeter both because at greater elevations the distance from the Earth's center increases (free-air correction) and because mass exists between the meter and the reference elevation, which is usually mean sea level (Bouguer correction). Gravity varies with latitude because the Earth's equatorial radius exceeds its polar radius and because centrifugal force resulting from the Earth's rotation varies with latitude. Nearby terrain affects a gravimeter; mountains exert an upward pull, valleys cause a deficit of downward pull. Thus the effects of nearby elevation differences add, whether the differences are positive or negative. This is the most critical correction to be made to gravity data in areas of rough terrain.

Gravity measurements that have been corrected for all of these effects are called Bouguer anomalies, or free-air anomalies if the Bouguer correction has not been made. They therefore represent the effects of local masses within the Earth, that is, effects for which corrections have not been made. Most gravity maps display contours (isogals) of free-air or Bouguer anomaly values.

Data interpretation. The most important part of gravity interpretation is locating anomalies that can be attributed to mass concentrations being sought, isolating these from other effects (Fig. 1). Separating the main part of the gravitational field, which is not of interest (the regional), from the parts attributed to local masses, the residuals, is called residualizing.

Many techniques for gravity data analysis are similar to those used in analyzing magnetic data. Shape parameters are used to determine the depth of the mass's center. Another widely used technique is model fitting: a model of an assumed feature is made, its gravity effects are calculated, and the model is compared with field measurements.

Continuation is a process by which calculations are made from measurements of the gravity field over one surface to determine what values the field would have over another surface. A field can be continued if there is no anomalous mass between the surfaces. Continuing the field to a lower surface produces sharper anomalies as the anomalous mass is approached. However, if the process is carried too far, instability occurs when the anomalous mass is reached. The technique, however, is very sensitive to measurement uncertainties and often is not practical with real data.

Seismic exploration. The seismic method is the predominant geophysical method. Seismic waves are generated by one of several types of energy sources and detected by arrays of sensitive devices called geophones or hydrophones. The most common measurement made is of the travel times of seismic waves and the amplitude of the waves, with less attention being given to changes in their frequency content or wave shape.

The seismic method is divided into two major classes, refraction and reflection, and two types based on the objectives, exploration and reservoir studies. Method classification depends on whether the predominant portion of wave travel is horizontal or vertical, respectively.

Principles of seismic waves. A change in mechanical stress produces a strain wave that radiates outward as a seismic wave, because of elastic relationships. The radiating seismic waves are like those that result from earthquakes, though much weaker. Most seismic work involves the analysis of P waves (compressional waves) in which particles move in the direction of wave travel, analogous to sound waves in air. S waves (shear waves) are occasionally studied, but most exploration sources do not generate very much shear energy. Surface waves, especially Rayleigh waves, are also generated, but these are mainly a nuisance because they do not penetrate far enough into the Earth to carry much useful information. Recording techniques are designed to discriminate against them. *See* SEISMOLOGY.

A seismic wave is a vector, involving both magnitude and direction. Historically, measurements have been made of only the vertical component of motion (with geophones), or only of the magnitude (with hydrophones). Attention is now being shifted to measuring all components of wave motion in order to study the conversion of P-waves to S-waves, and vice versa, and anisotropy.

The amplitude of a seismic wave reflected at an interface depends on the elastic properties, often expressed in terms of seismic velocity and density on either side of the interface. When the direction in which the wave is traveling is perpendicular to the interface, the ratio of the amplitudes of reflected and incident seismic waves is given by the normal reflection coefficient $R_\perp$ as shown in Eq. (1), where

$$R_\perp = \frac{\Delta(\rho V)}{2(\rho \bar{V})} \qquad (1)$$

$\Delta(\rho V)$ is the change in the product of velocity and density and $(\rho \bar{V})$ is the average of the product of velocity and density on opposite sides of the interface. The relationships are much more complicated where wave travel is not perpendicular to interfaces. The variation of the reflection coefficient with angle of incidence is now routinely used to indicate the kind of fluid in the pore space and the lithology.

Seismic waves are bent when they pass through interfaces, and Snell's law holds, shown in Eq. (2),

$$\frac{\sin \sigma_1}{V_1} = \frac{\sin \sigma_2}{V_2} \qquad (2)$$

where σ_i is the angle between a wavefront and the interface in the ith medium where the velocity is V_i (**Fig. 4**). Because velocity ordinarily increases with depth, seismic-ray paths become curved concave-upward (**Fig. 5**).

The resolving power (ability to separate features) with seismic waves depends inversely on their wavelength λ and is often thought of as of the order of λ/4. The wavelength is often expressed in terms of the wave's velocity and frequency f: $\lambda = V/f$. Most seismic work involves frequencies from 15 to 70 Hz, and most rocks have velocities from 4500 to 18,000 ft/s (1500 to 6000 m/s) so that wavelengths range from 90 to 900 ft (30 to 300 m). Usually, the frequency becomes lower and the velocity higher as depth in the Earth increases, so that wavelength increases and resolving power decreases. Very shallow, high-resolution work involves frequencies higher than those cited above, and long-distance refraction (and earthquakes) involve lower frequencies. *See* ECHO SOUNDER.

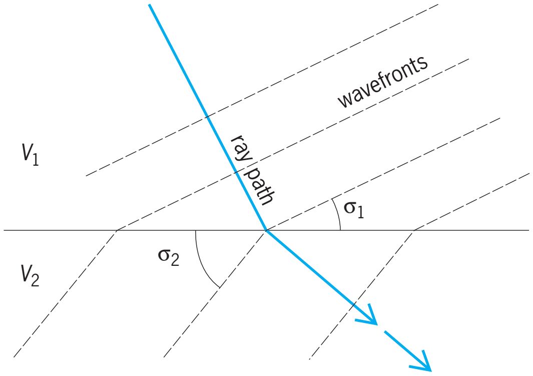


Fig. 4. Bending of seismic waves at interface; $V_2 > V_1$.

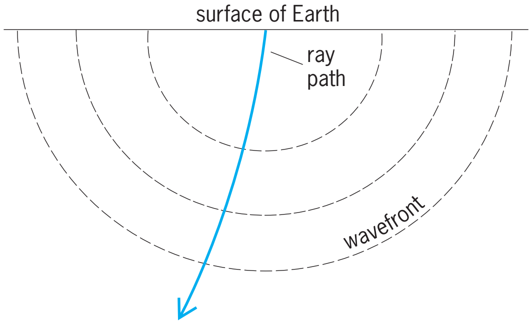


Fig. 5. Wavefronts become more widely spaced and ray paths become curved as velocity increases with depth.

Reflection exploration. Seismic-wave energy partially reflects from interfaces where velocity or density changes. The measurement of the arrival times of reflected waves (**Fig. 6**) thus permits mapping the interfaces that form the boundaries between different kinds of rock. This, the predominant geophysical exploration method, can be thought of as similar to echo sounding. When a seismic source S generates seismic energy, it is received at detectors located at intervals, say from A to B. The distance to the reflector can be obtained from the arrival time of the reflection if the velocity is known. If the reflector dips towards A, the reflection will arrive sooner at B

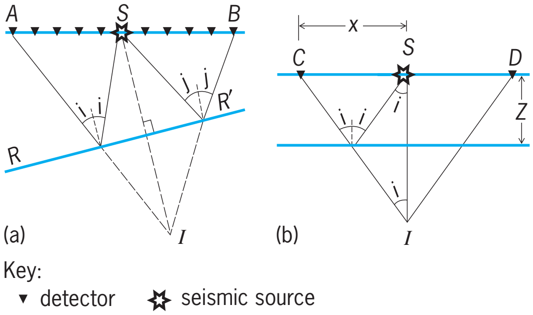


Fig. 6. Measuring reflected seismic wave energy. (*a*) Dipping reflector. (*b*) Flat reflector.

than at A; the difference in arrival times is a measure of the amount of dip.

For a flat reflector (Fig. 6*b*) and constant velocity V, the arrival time at detector C is t_c and the arrival time at a detector at the source S is $t_0 = 2Z/V$, where Z is the depth. From the pythagorean theorem (for the triangle CSI; Fig. 6*b*), Eqs. (3) and (4) are obtained.

$$(Vt_c)^2 = (Vt_0)^2 + x^2 = (2Z)^2 + x^2 \qquad (3)$$

$$V = x\left(t_c^2 - t_0^2\right)^{1/2} \qquad (4)$$

These give both the values of V and Z. Similar relationships can be used for nonflat reflectors or nonconstant velocity to yield velocity information.

Usually a number of detector groups are used, and the arrival of reflected waves is characterized by coherency. Thus, if all of the detectors in a line move in a systematic way, a seismic wave probably passed. Multiple detectors make it possible to detect coherent waves in the presence of a high noise level and also to measure distinguishing features of the waves. *See* SEISMIC EXPLORATION FOR OIL AND GAS.

Refraction exploration. Refraction seismic exploration involves rocks characterized by high seismic velocity. Wavefronts are bent at interfaces (**Fig. 7**) so that appreciable energy travels in high-velocity members and arrives earlier at detectors distant from the source than energy that has traveled in overlying lower-velocity members. Differences in arrival time at different distances from the energy source yield information on the velocity and attitude (dip) of the high-velocity member.

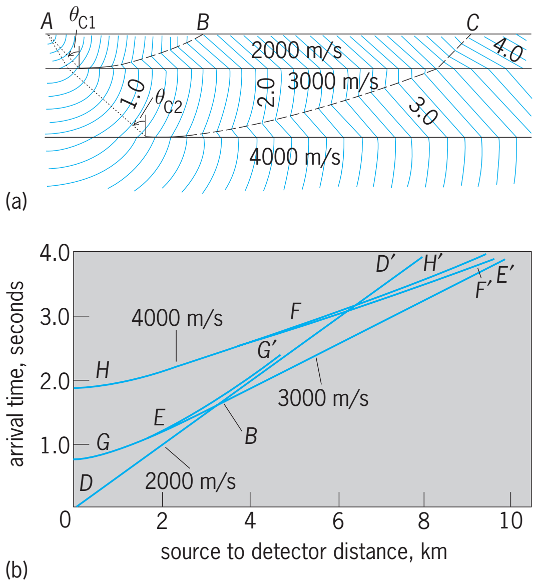


Fig. 7. Refractive seismic exploration data. (*a*) Section through model of layered earth. Curves (wavefronts) indicate location of seismic energy at successive times after a shot at *A*. Beyond *B*, energy traveling in the second layer arrives first; beyond *C*, that in the third layer arrives first. (*b*) Arrival time as a function of source-to-detector distance. *DD*′, direct wave; *EE*′, refracted wave (head wave) in the second layer; *FF*′, refracted wave in the third layer; *GG*′, reflection from interface between first and second layers; *HH*′, reflection from that between second and third layers. 1 km = 0.6 mi. 1 m/s = 3.3 ft/s.

A variant of refraction seismic exploration is the search for high-velocity masses in an otherwise low-velocity section, by looking for regions where seismic waves arrive earlier than expected. Such arrivals, called leads, were especially useful in locating salt domes in Louisiana, Texas, Mexico, and Germany in the late 1920s and 1930s.

Refraction seismic techniques are used in groundwater studies, engineering geophysics, and mining to map the water table and bedrock under unconsolidated overburden, with objectives such as foundation information or locating buried stream channels in which heavy minerals might be concentrated or where water might accumulate. Refraction techniques are also used in petroleum exploration and for crustal studies.

Channel-wave exploration. Seismic waves can become trapped once they are generated in low-velocity formations. The low-velocity formation constitutes a waveguide, and the waves are called channel, guided, or seam waves. Coal often satisfies waveguide requirements, and channel waves are used for determining the continuity of coal beds. The objective usually is to ascertain that the coal measures are not interrupted by faults or channels that would interfere with the operation of longwall mining machines. Sources and geophones are located in the coal bed in mining tunnels; and both reflection and transmission ray paths are used, the former where sources and geophones are in the same tunnel and the latter where they are in different tunnels. Channel waves are also sometimes studied in borehole-to-borehole measurements to ascertain the continuity of reservoirs.

Instrumentation. Detectors of seismic energy on land (geophones or seismometers) are predominantly electromechanical devices. A coil moving in a uniform magnetic field generates a voltage proportional to the velocity of the motion. Usually the coil has only one degree of freedom and is used so it will be sensitive to vertical motion only. Three mutually perpendicular elements in a three-component detector are coming into increased use to distinguish the type of waves (compressional, shear, Rayleigh, and such) or to determine the direction from which the waves come. *See* SEISMOGRAPHIC INSTRUMENTATION.

Detectors in water are usually piezoelectric. Pressure changes produced as a seismic wave passes distort a ceramic element and induce a voltage between its surfaces. Such detectors are not directionally sensitive. *See* HYDROPHONE.

Detectors are usually arranged in groups (arrays) spread over a distance and connected electrically so that, in effect, the entire group acts as a single large detector. Such an arrangement discriminates against seismic waves traveling in certain directions. Thus a wave traveling horizontally reaches different detectors in the group at different times, so that wave peaks and troughs tend to cancel; whereas a wave traveling vertically affects each detector at the same time, so that the effects add. The principles of seismic array design are similar to those in radio antenna design.

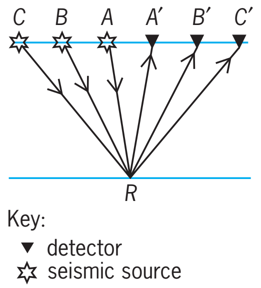


Fig. 8. Reflection of seismic energy obtained from reflection point *R* (common midpoint) by positioning a source at *A* and a detector at *A′*. The same point is involved with a source at *B* and a detector at *B′*, or *C* and *C′*. The redundancy in measuring reflections from the same point many times (often 48- to 96-fold) permits sorting out different kinds of waves.

The signal from the detectors is transmitted to recording equipment over a cable, streamer, or radio link, and then amplified and recorded. The output level from a geophone varies tremendously during a recording. Seismic recording systems are linear over ranges of 120 dB or more. Seismic amplifiers employ various schemes to compress the range of seismic signals without losing amplitude information. They also incorporate adjustable filters to permit discriminating on the basis of frequency.

Many sources are used to generate seismic energy. The classical land energy source is an explosion in a borehole drilled for the purpose, and solid explosives continue to be used extensively for work on land and in marshes. The explosion of a gas mixture in a closed chamber, a dropped weight, a hammer striking a steel plate, and other sources of impulsive energy are used in land work. An air gun, which introduces a pocket of high-pressure air into the water, is the most common energy source in marine work. Other marine energy sources involve the explosion of gases in a closed chamber, a pocket of high-pressure steam introduced into the water, the discharge of an electrical arc, and the sudden mechanical separation of two plates (imploder).

An oscillatory mechanical source (vibroseis) is the predominant source being used on land. Such a source introduces a long wave train so that individual reflection events cannot be resolved without subsequent processing (correlation with the input wave train), which, in effect, compresses the long wave train and produces essentially the same result as an impulsive source.

Field techniques. Most petroleum exploration seismic work has been carried out along lines of survey often run parallel to each other at right angles to the geological strike with occasional perpendicular tie lines, often run on a regular grid. Long lines many kilometers apart are sometimes run for regional information, but lines are often concentrated in regions in which anomalies have been detected by previous geophysical work. Most seismic work has the objective of mapping interfaces continuously along the seismic lines to map the geological structure.

Geophone groups are spaced 80–300 ft (25–100 m) apart with 48–120 adjacent groups of 6–24 geophones each being used for each recording. The source is sometimes located at the center of the active groups (split spread), sometimes at one end (end-on spread).

Following a recording, the layouts and sources are advanced down the line by some multiple of the group interval for redundant coverage (**Fig. 8**).

Small marine operations, often called profiling, may consist of an energy source and a short streamer containing a number of hydrophones and feeding a recorder. Larger marine operations (**Fig. 9**) involve ships 180 ft (60 m) or more in length towing a streamer 1 to 3 mi (2 to 8 km) long with 250 groups of hydrophones spaced along the streamer. An energy source is towed near the ship. Recordings are made as the ship is continuously under way at a speed of about 6 knots (3 m/s).

The foregoing methods acquire data along lines of traverse, but most seismic work today is designed to acquire data uniformly over an area. Such methods are known as three-dimensional, and they result in acquiring a volume rather than a cross section of data. A variety of geophone and source arrangements are used on land, most often with several parallel lines of geophones and perpendicular lines of sources. Often more than 1000 geophone groups will be recorded for each source location. Most marine three-dimensional data are acquired by ships towing two sources and up to 12 streamers pulled to the sides of the ship by paravanes, so that several closely spaced parallel lines of data are acquired on a single traverse. Except for requiring more data channels, instrumentation and methods are similar to those used for two-dimensional data acquisition. *See* OCEANOGRAPHY.

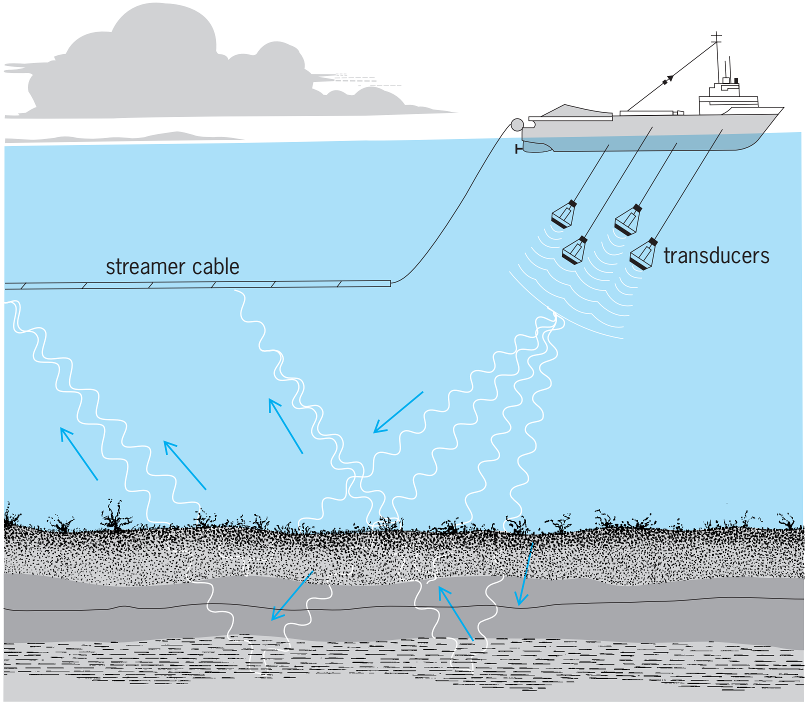


Fig. 9. Seismic surveying at sea is done by towing a long streamer containing detectors that sense seismic energy. Seismic energy is generated by several air guns that are also towed from the ship.

Three-dimensional methods generally require more precision in determining source and detector locations. The Global Positioning System and electronic distance instruments are used on land. Marine surveys use highly redundant location measurements of several different kinds. Global Positioning System instruments usually locate the ship and streamer tail buoys; acoustic transducers measure distances between the sources, streamers, and ship; and magnetic compasses within the streamers determine their orientations. A large volume of positioning data is acquired and reduced by computer in real time so that corrections can be made immediately.

Data reduction and processing. Seismic data are corrected for elevation and near-surface variations on the basis of survey data and observations of the travel time of the first energy from the source to reach the detectors, which usually involves travel either in a direct path or in shallow refractors.

Most seismic data processing is done either to reduce the noise so that structural and stratigraphic features can be seen more clearly, or to reposition features to display correctly their positions relative to each other so that they can be located by drilling a well. Seismic data processing is one of the larger users of giant computers.

Almost all data are processed by computers, with the first step often being editing, wherein data are merged with identifying data, rearranged, checked for being either dead or wild (with bad values sometimes replaced with interpolated values), time-shifted in accordance with elevation and near-surface corrections that have been determined in the field, scaled, and so on.

Following the editing, different processing sequences may follow, including (1) filtering (deconvolution) to remove undesired natural filter effects, trace-to-trace variations, variations in the strength or wave shape of the source, and so on; (2) grouping according to common midpoint (Fig. 8) or some other arrangement; (3) analyzing to see what velocity values will maximize coherency as a function of source-to-detector distance; (4) statistically analyzing to see what trace shifts will maximize coherency; (5) trace-shifting according to the results of steps 3 or 4; (6) stacking by adding together a number of individual traces; (7) migrating by rearranging and combining data elements in order to position reflection events more nearly under the surface locations where the appropriate reflecting surface is located; (8) another filtering; and (9) displaying of the data.

The techniques for processing two- and three-dimensional seismic data are nearly the same. When data are acquired only along lines of traverse, there is always ambiguity as to the directions from which the waves come, which produces ambiguity as to where features are located perpendicular to the line of acquisition. This ambiguity is removed with three-dimensional data sets, resulting in major improvements in the ability to resolve features. Three-dimensional data also permit many ways of displaying data that help in interpreting the significance of features.

Data interpretation. The travel times of seismic reflections are usually measured from record section displays (**Fig. 10**), which result from processing. Appropriate allowance (migration) must be made because reflections from dipping reflectors appear at locations downdip from the reflecting points. Allowance must also be made for variations in seismic velocity, both vertically and horizontally. Seismic events other than reflections must be identified and explained.

In petroleum exploration the objective is usually to find traps, places in which porous formations are high relative to their surroundings and in which the overlying formation is impermeable. If oil or gas,

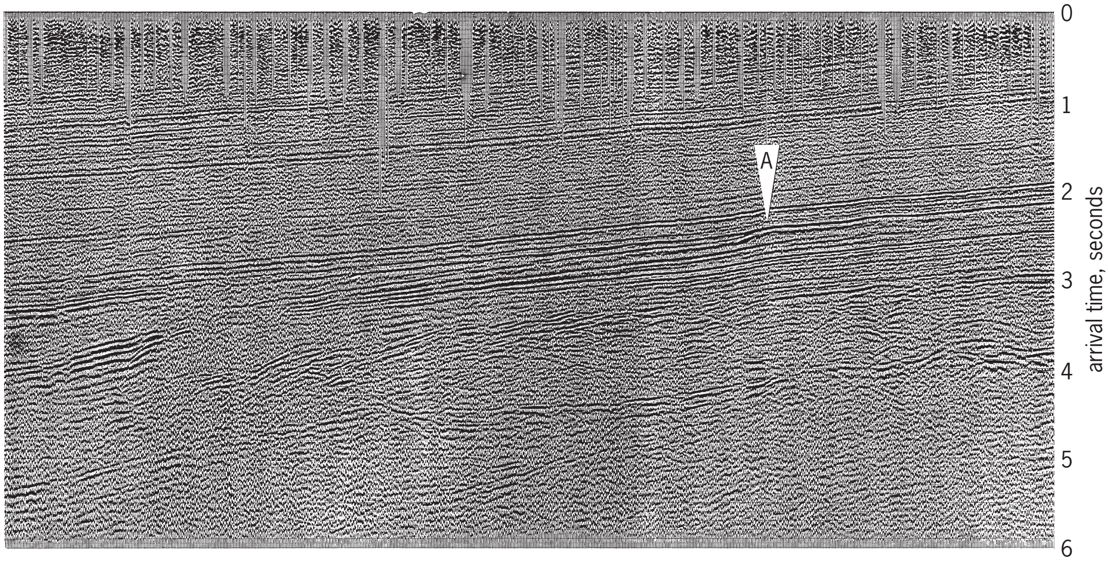


Fig. 10. Seismic record section in East Texas after processing. The rock bedding giving the various reflections was nearly horizontal when deposited, and the present dips result from tilting and other deformation subsequent to deposition. The reflection event at *A* is attributed to the Edwards Reef, which formed a barrier at the time the adjacent sediments were deposited. Downdip to the left of the reef formations the Woodbine sands can be seen thinning and pinching out in the updip direction. These are productive of oil and gas in this area. (*Grant Geophysical*)

which are lighter than water, are present, they float on top of the water and accumulate in the pores in the rock at the trap. Seismic exploration determines the geometry, hence where traps might be located. However, it usually is not possible to tell conclusively from seismic data alone whether oil or gas was ever generated, whether the rocks have porosity, whether overlying rocks are impermeable, or whether oil or gas might have escaped or been destroyed, even if they were present at one time.

In addition to mapping the structural patterns within the Earth, seismic data are analyzed for evidence that might identify the nature of the formations, the environment in which they were formed, and the nature of the fluid in the pore spaces. To reconstruct the geologic history, often several reflections at different depths are mapped, and attempts are made to reconstruct the position of the deeper reflectors at the time of deposition of shallower rocks.

After some experience has been developed in an area, patterns in the seismic data that distinguish certain reflectors or certain types of structure or stratigraphy often can be recognized. Seismic velocity measurements are helpful here. Seismic stratigraphy is considered an important aspect of sequence stratigraphy. *See* SEQUENCE STRATIGRAPHY.

In relatively unconsolidated sediments and in some other circumstances, gas and oil may lower the seismic velocity or rock density or both sufficiently to produce a distinctive reflection, usually evidenced by strong amplitude (a "bright spot") and other distinguishing features (**Fig. 11**). Coal and peat beds are also characterized by reflections of strong amplitude.

Much interpretation, including almost all three-dimensional interpretation, is done at computer workstations. The interpreter views data, base maps, and the progressing interpretation on computer screens. It is possible to call up any of the available data, display the data in various ways in various colors, pick events representing various horizons with the aid of an automatic picker, pick faults, carry out computations, and manipulate the data in various ways. In a three-dimensional interpretation, the interpreter can slice through the three-dimensional volume in many ways, including vertical zig-zag sections, slices at constant travel time, slices along picked horizons, slices parallel to faults, and combinations of horizontal and vertical displays. Workstations also make it easier to incorporate geologic, well-log, engineering, and production data into interpretation. The result is a much more complete and precise interpretation than conventional methods produce.

Seismic methods can also be applied to petroleum engineering. A number of techniques, including three-dimensional, delineate oil and gas fields in enough detail to permit increased recovery. Correlation of seismic amplitude with porosity, permeability, and other rock properties permits a more complete description of inhomogeneities, which affect fluid flow through reservoirs. Seismic methods, especially three-dimensional methods, had major impact on the discovery and production of oil and gas in the 1990s. This was especially important in revealing pools of hydrocarbons that had been missed or bypassed and thus increasing production from,

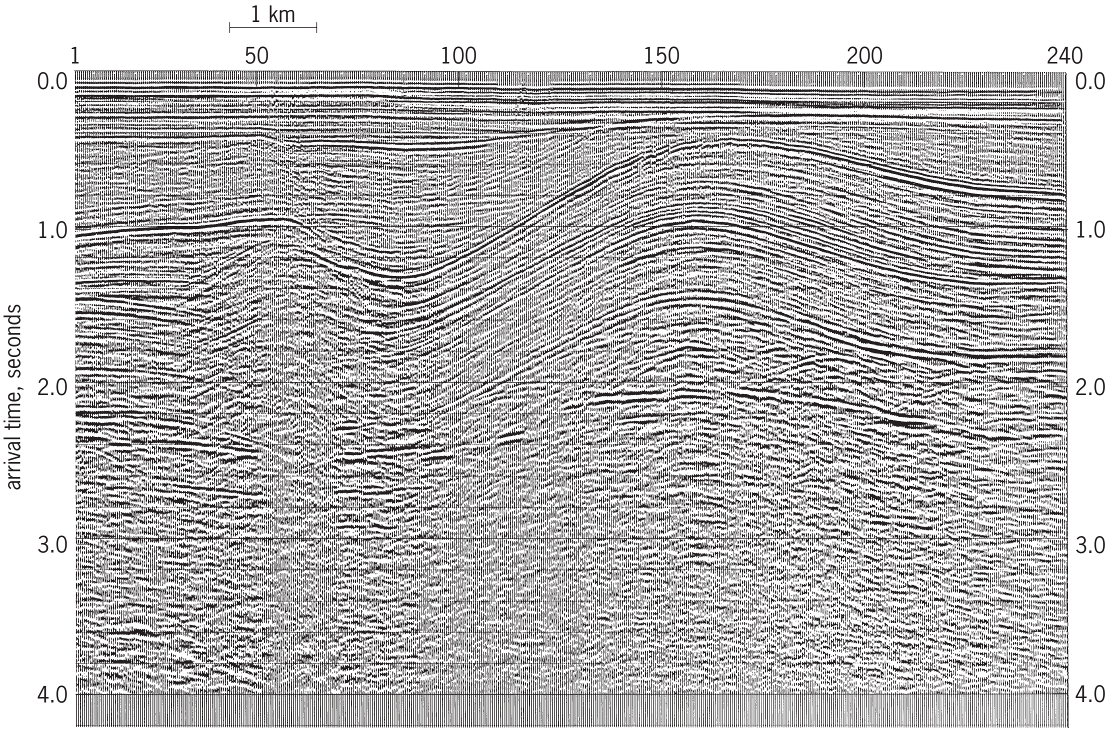


Fig. 11. Seismic record section from the North Sea. The layer from about 2 to 2.3 s at the right edge is composed primarily of salt that has undergone plastic flow. The folding of the rock formations is attributed to this salt movement. The upfolding (an anticline) has produced a trap containing gas at 1.1 s. The scale at the top refers to source points, based on a specified beginning point. 1 km = 0.6 mi. (*Grant Geophysical*)

and lengthening the life of, oil and gas fields. Seismic methods have the potential under some circumstances to monitor the flow of fluids as production progresses. Such information would permit improved hydrocarbon recovery.

Electrical and electromagnetic exploration. Variations in the conductivity or capacitance of rocks form the basis of a variety of electrical and electromagnetic exploration methods, which are used in metallic mineral prospecting and in engineering, ground-water, and other surveys with shallow objectives. Both natural and induced direct current and low-frequency alternating currents are measured in ground surveys, and ground and airborne electromagnetic surveys involving the lower radio frequencies are made.

Natural currents in the Earth, called telluric current, affect large areas. The current density of telluric currents varies with rock conductivity. Comparisons are made between readings observed simultaneously at various locations and at a reference location in order to determine resistivity differences between the locations.

Changes in electrical current flow give rise to associated magnetic fields, and the converse is also true, according to Maxwell's equations. Natural currents are somewhat periodic. Magnetotellurics involves the simultaneous measurement of natural electrical and magnetic variations from which the variation of conductivity with depth can be determined. *See* GEOELECTRICITY; ROCK, ELECTRICAL PROPERTIES OF.

Certain mineral ores store energy as a result of current flow and, after the current is stopped, transient electrical currents flow. This phenomenon is called induced polarization. Observations of the rate of decay of these transient currents are studied in time-domain methods.

Alternating currents tend to flow along the surface of conductors rather than in their interior. The thickness which contains most of the current is called the skin depth. The skin depth, in meters, is given by $(2/\sigma\mu\omega)^{1/2}$, where σ is the conductivity in mhos per meter, μ is permeability in henrys per meter, and ω is angular frequency in radians per second. Since the skin depth becomes greater as frequency becomes lower, measurements at different frequencies give information on the variation of conductivity with depth. Methods in which apparent resistivity is determined as a function of frequency are called frequency-domain methods.

Direct-current and low-frequency alternating-current ground surveys are carried out with a pair of current electrodes, by which electrical current is introduced into the ground, and a pair of potential electrodes across which the voltage is measured. The equipment is often simple, consisting essentially of a source of electrical power (battery or generator), electrodes and connecting wires, ammeter, and voltmeter. A key problem here is providing equipment that will generate enough electrical or electromagnetic energy in the ground but is reasonably portable. The depth of current penetration depends on the geometry of disposition of electrodes, on the frequency used, and on the conductivity distribution. There are two basic types of measurement: (1) electrical sounding, wherein apparent resistivity is measured as the electrode separation is increased—these measurements depend mainly on the variation of electrical properties with depth; (2) electrical profiling, in which variations are measured as the electrode array is moved from location to location.

Electromagnetic methods generally involve a transmitting coil, which is excited at a suitable frequency, and a receiving coil, which measures one or more elements of the electromagnetic field. The receiving coil is usually oriented in a way that minimizes its direct coupling to the transmitter, and the residual effects are then caused by the currents that have been induced in the ground. A multitude of configurations of transmitting and receiving antennas are used in electromagnetic methods, both in ground surface and airborne surveys. In airborne surveys, the transmitting and receiving coils and all associated gear are carried in an aircraft, which normally flies as close to the ground as is safe. Airborne surveys often include multisensors, which may record simultaneously electromagnetic, magnetic, and radioactivity data along with altitude and photographic data. Sometimes several types of electromagnetic configurations or frequencies are used.

The effective penetration of most of the electromagnetic methods into the Earth is not exceptionally great, but they are used extensively in searching for mineral ores within about 300 ft (100 m) of the surface. Electrical methods are effective in exploring for ground water and in mapping bedrock, as at dam sites. They are used also for detecting the position of buried pipelines and in land-mine detection and other military operations.

Ground-penetrating radar can provide images of shallow features in a manner analogous to seismic methods, and it is increasingly used in environmental studies to locate waste dumps and other features that may be pollution threats, in studying archeological sites, and for other applications. Ground-penetrating radar reflects because of changes in electrical resistivity, especially from metallic objects such as drums of waste. The principal feature limiting such radar penetration into the earth is the presence of ground water, so this radar is used more in arid regions than where the water table is near the surface.

Radioactivity exploration. Natural radiation from the Earth, especially of gamma rays, is measured both in land surveys and airborne surveys. Natural types of radiation are usually absorbed by a few feet of soil cover, so that the observation is often of diffuse equilibrium radiation. The principal radioactive elements are uranium, thorium, and the potassium isotope ^{40}K. Radioactive exploration has been used primarily in the search for uranium and other ores, such as niobium (columbium), which are often associated with them, and for potash deposits. The scintillation counter is generally used to detect and measure the radiation. *See* SCINTILLATION COUNTER.

Remote sensing. Measurements of natural and induced electromagnetic radiation made from high-flying aircraft and Earth satellites are referred to collectively as remote sensing. This comprises both the observation of natural radiation in various spectral bands, including both visible and infrared radiation, and measurements of the reflectivity of infrared and radar radiation. *See* REMOTE SENSING.

Well logging. A variety of types of geophysical measurements are made in boreholes, including self-potential, electrical conductivity, velocity of seismic waves, natural and induced radioactivity, and temperature variations. Borehole logging is used extensively in petroleum exploration to determine the characteristics of the rocks that the borehole has penetrated, and to a lesser extent in mineral exploration.

Measurements in boreholes are sometimes used in combination with surface methods, as by putting some electrodes in the borehole and some on the surface in electrical exploration, or by putting a seismic detector in the borehole and the energy source on the surface. *See* WELL LOGGING. R. E. Sheriff

Bibliography. A. R. Brown, *Interpretation of Three-Dimensional Seismic Data*, 5th ed., 1999; M. B. Dobrin and C. Savit, *Introduction to Geophysical Prospecting*, 1988; R. E. Sheriff, *Encyclopedic Dictionary of Exploration Geophysics*, 3d ed., 1991; R. E. Sheriff, *Geophysical Methods*, 1989; R. E. Sheriff and L. P. Geldart, *Exploration Seismology*, 2d ed., 1995; W. M. Telford et al., *Applied Geophysics*, 1990; K. H. Waters, *Reflection Seismology: A Tool for Energy Resource Exploration*, 3d ed., 1992.

Geophysical fluid dynamics

The branch of physics that studies the dynamics of naturally occurring large-scale flows in both the atmosphere and oceans. Examples of such flows are weather patterns, atmospheric fronts, the Gulf Stream, coastal upwelling, and El Niño. The fluids are either air or water in a moderate range of temperatures and pressures. *See* EL NIÑO; GULF STREAM.

Because of their large scale (from tens of kilometers up to the size of the planet), geophysical flows are strongly influenced by the diurnal rotation of the Earth, which is manifested in the equations of motion as the Coriolis force. Another fundamental characteristic is stratification, that is, density heterogeneity within the fluid in the presence of the Earth's gravitational field, which is responsible for buoyancy forces. Thus, geophysical fluid dynamics may be considered to be the study of rotating and stratified fluids. It is the common denominator of dynamical meteorology and physical oceanography. *See* CORIOLIS ACCELERATION; EARTH; METEOROLOGY; OCEANOGRAPHY.

Historical development. Although geophysical fluid dynamics first became recognized as a scientific discipline in the late 1950s, its foundations can be traced to developments in fluid mechanics over a much longer period of time. During the nineteenth and early twentieth centuries, a number of mathematicians (for example, P. S. de Laplace), fluid dynamicists (Lord Kelvin, H. von Helmholtz, G. I. Taylor), meteorologists (L. F. Richardson, V. Bjerknes, C. G. Rossby), and oceanographers (A. Defant, V. W. Ekman) made significant contributions to the study of large-scale flows that later became incorporated in the discipline of geophysical fluid dynamics. *See* FLUID MECHANICS.

During World War II the demand for improved forecasts of the weather and sea state provided a major impetus and resulted in a new generation of dynamical meteorologists and physical oceanographers who continued to make contributions long after the war ended. The main catalyst, however, was the launching of the geophysical fluid dynamics Summer Program at the Woods Hole Oceanographic Institution, in Massachusetts, which began in 1959.

Dynamics of rotating flows. The first of the two distinguishing attributes of geophysical fluid dynamics is the effect of the Earth's rotation. Because geophysical flows are relatively slow and spread over long distances, the time taken by a fluid particle (be it a parcel of air in the atmosphere or water in the ocean) to traverse the region occupied by a certain flow structure is comparable to, and often longer than, a day. Thus, the Earth rotates significantly during the travel time of the fluid, and rotational effects enter the dynamics. Fluid flows viewed in a rotating framework of reference are subject to two additional types of forces, namely the centrifugal force and the Coriolis force. (Properly speaking, these originate not as actual forces but as acceleration terms to correct for the fact that viewing the flow from a rotating frame—the rotating Earth in the case of geophysical fluid dynamics—demands a special transformation of coordinates.) Contrary to intuition, the centrifugal force plays no role on fluid motion because it is statically compensated by the tilting of the gravitational force caused by the departure of the Earth's shape from sphericity. Thus, of the two, only the Coriolis force acts on fluid parcels. History reveals that Laplace wrote the correct equations of fluid flow in a rotating frame for a study of ocean tides some decades before G. G. de Coriolis, but somehow the name of the latter has remained attached to the terms representing the rotational effect in the equations of motion. *See* EARTH ROTATION AND ORBITAL MOTION.

The importance of the Coriolis force in the dynamics is measured by the Rossby number, Ro [Eq. (1)], where U is a typical velocity value within

$$\mathrm{Ro} = \frac{U}{fL} \tag{1}$$

the flow, $f = 2\Omega \sin$ (latitude) the Coriolis parameter [the Earth's angular rotation rate $\Omega = 7.29\ 10^{-5}$/s], and L a representative distance along the flow. The smaller this number, the stronger the effect of rotation on the flow.

At low speeds or long length scales, not atypical of geophysical flows, the Rossby number may fall far below unity. In such case, the rotating effects become dominant, and the balance of horizontal forces is primarily an equilibrium between the pressure-gradient force and the Coriolis force [Eqs. (2) and (3)]. Here ρ is the density, f the Coriolis

$$-\rho f v = -\frac{\partial p}{\partial x} \qquad (2)$$

$$+\rho f u = -\frac{\partial p}{\partial y} \qquad (3)$$

parameter, u and v the eastward and northward velocity components respectively, p the pressure, and x and y the eastward and northward coordinates respectively. This balance is called geostrophy. *See* GEOSTROPHIC WIND.

One remarkable property of geostrophic flows is their vertical rigidity: All fluid parcels on the same vertical share the same velocity and therefore move as one rigid column. This is known as the Taylor-Proudman theorem. A corollary is another property: If there is any vertical velocity, fluid columns must go up and down without shrinking or stretching, occasionally forcing isolation of entire fluid columns above bottom bumps or dips, called Taylor columns. In the atmosphere and oceans, this property is manifested by the tendency of slow flows to follow topographic contours.

When the Rossby number is less than unity but not extremely small, the dynamics are less degenerate, and flow fields can exhibit a number of time-dependent behaviors, including inertia-gravity waves and vorticity waves. Inertia-gravity waves are the classical surface gravity waves with a modification due to the Coriolis effect. Their frequency-wavenumber relationship is shown in Eq. (4). Here g is the Earth's

$$\omega = \sqrt{ghk^2 + f^2} \qquad (4)$$

gravitational acceleration, h the fluid's resting depth, f the Coriolis parameter, and k the wavenumber (2π divided by the wavelength). One property of these waves is that, unlike their purely gravitational counterparts, their frequency cannot fall below a minimum (f), lest they become evanescent (exponential instead of periodic behavior in one of the spatial directions). The Kelvin wave is the special case of a wave that is propagating along a boundary (shoreline) and evanescent away from it. This wave also has the property of having no velocity component transverse to the direction of propagation.

The other kind of waves is specific to rotating fluids. Instead of gravity, their restoring force is a vorticity gradient, which is created either by the Earth's curvature or by a bottom slope. There are thus two kinds of vorticity waves: planetary and topographic waves. Synoptic weather formations at midlatitudes share characteristics with planetary waves, whereas topographic waves tend to be associated with airflow over mountain ranges and ocean currents along the continental slope. A third source of vorticity gradient is a gradient in the horizontal shear of a horizontal flow. In this case, however, a portion of the kinetic energy stored in the flow is easily transferred to the wave, and the result is a growing wave that greatly distorts the original flow field. This is called barotropic instability. *See* WAVE MOTION; SURFACE WAVES.

In the vicinity of a top or bottom boundary, friction can become important, and the combination of rotation with friction leads to the development of Ekman layers. These layers differ from their cousins in classical fluid mechanics by having a well-defined thickness and a helicoidal flow field. An unexpected result is the existence of a significant angle between the frictional stress causing the boundary layer and the direction of the fluid transport within it. *See* FRICTION.

Dynamics of stratified flows. Variations of moisture in the atmosphere, of salinity in the ocean, and of temperature in either can modify the density of the fluid to such an extent that buoyancy forces become comparable to other existing forces. The fluid then has a strong tendency to arrange itself vertically so that the denser fluid sinks under the lighter fluid. The resulting arrangement is called stratification, the second distinguishing attribute of geophysical fluid dynamics. The greater the stratification in the fluid, the greater the resistance to vertical motions, and the more potential energy can affect the amount of kinetic energy available to the horizontal flow. A practical measure of stratification is the Brunt-Vaisala frequency, N, defined from its square according to Eq. (5), where $\rho(z)$ is the density function of the

$$N^2 = -\frac{g}{\rho}\frac{d\rho}{dz} \qquad (5)$$

vertical z and decreasing upward ($d\rho/dz < 0$). In many ways, the Brunt-Vaisala frequency N is to stratified fluids what the Coriolis parameter f is to rotating fluids, and a close analogy can be developed between stratified and rotating fluids—namely, vertical properties of the latter are similar to horizontal properties of the former. The dimensionless number of stratified flows corresponding to the Rossby number of rotating flows is the internal Froude number, Fr [Eq. (6)], where H is the height of the fluid region

$$\mathrm{Fr} = \frac{U}{NH} \qquad (6)$$

under consideration. *See* DENSITY.

Just as the water-air density discontinuity can support gravity waves, the gradual density variation of a stratified fluid can support gravitational waves, called internal waves. Like their surface counterparts, internal waves can be affected by the Earth's rotation, and unlike them, they can propagate not only horizontally but also vertically. Internal waves are ubiquitous in the ocean, where their frequency spectrum very often displays a standard structure suggesting saturation. In the atmosphere, mountain waves and cloud rolls are manifestations of internal waves.

On occasion, a source of energy is sufficiently intense to create vertical motions that overcome the stratification, and the result is vertical mixing over a partial extent of the fluid system. A common situation is one where turbulence induced by a shear stress in the proximity of a boundary generates a mixed boundary layer. In the ocean, a mixed layer is created almost daily in the top few meters under the action of surface winds, and it also exists at the bottom of the ocean wherever near-bottom currents can generate enough turbulence to overcome the stratification. *See* TURBULENT FLOW.

Another mechanism capable of erasing an existing stratification is thermal convection. Whenever a fluid is heated from below or cooled from above, it becomes top heavy, and negative buoyancy forces generate velocities that rearrange the fluid toward a more stable state. During daytime, the ground is warmed by the Sun and effectively heats the atmosphere from below. The result is a state of convection that gradually erases the previous nocturnal stratification and creates a well-ventilated layer called the atmospheric boundary layer. The turbulence accompanying this state of convection greatly facilitates the dispersal of pollution. In lakes and oceans, heat loss at the surface during winter generates convection that can mix the water over several tens or hundreds of meters, depending on the intensity of the cooling and on the strength of the stratification created during the previous summer. *See* CONVECTION (HEAT).

Dynamics of rotating and stratified flows. A quantity central to the understanding of geophysical flows, which are simultaneously rotating and stratified, is the potential vorticity, q. This quantity incorporates both rotation and stratification [Eq. (7)] and has the

$$q = \left(f + \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\frac{\partial \rho}{\partial z} + \left(\frac{\partial u}{\partial z}\frac{\partial \rho}{\partial y} - \frac{\partial v}{\partial z}\frac{\partial \rho}{\partial x}\right) \quad (7)$$

property of being nearly conserved by fluid parcels as they travel with the flow. One situation in which potential-vorticity conservation rules the fate of the fluid is geostrophic adjustment. This occurs whenever a heterogeneous fluid initially away from equilibrium seeks a lower level of energy; potential-vorticity conservation then limits the amount of vertical stretching or squeezing that every parcel can undergo, and the result is a final state with a residual motion in geostrophic balance. Many atmospheric and oceanic density fronts exhibit attributes of geostrophic adjustment.

The interplay between rotation and stratification effects creates a preferential length scale at which a large portion of the energy available in geophysical flows tends to be concentrated. This privileged length, L, called the radius of deformation, arises when the Rossby and Froude numbers are equal [Eq. (8)]. Numerous structures such as jets, vortices,

$$L = \frac{NH}{f} \quad (8)$$

and fronts have horizontal dimensions that are comparable to the radius of deformation. In the atmosphere, the radius of deformation sets the size of synoptic weather patterns (500–1000 km; 310–620 mi), whereas in the ocean it is usually smaller (10–50 km; 6–30 mi) and related to the width of currents such as the Gulf Stream.

Because they do not correspond to a state of absolute minimal energy, geostrophically adjusted states retain a certain amount of energy that can still be made available for subsequent motions. The process by which a geostrophic state relaxes toward a yet lower energy level is called baroclinic instability. The instability begins with a horizontal meandering of the flow that has a vertical phase shift creating a mutual reinforcement of the meanders at the different vertical levels of the fluid. Growing meanders eventually detach as vortices, and the ultimate situation is a complex pattern of a wavering flow embedded in an assortment of interacting vortices. The preferential length scale of the meanders and vortices is again the radius of deformation.

Geophysical flows are replete with vortices, resulting from baroclinic instability. Their interactions generate highly complex flows not unlike those commonly associated with turbulence. Unlike classical fluid turbulence, however, geophysical flows are wide and thin (with, furthermore, a high degree of vertical rigidity as a result of rotational effects), and their turbulence is nearly two-dimensional.

Applications. Geophysical fluid dynamics finds much of its use in dynamical meteorology and physical oceanography. In meteorology, geophysical fluid dynamics has been the key to understanding the essential properties of midlatitude weather systems, including the formation of cyclones and fronts. Geophysical fluid dynamics also explains the dynamical features of hurricanes and tornadoes, sea and land breezes, the seasonal formation and break-up of the polar vortex that is associated with high-latitude stratospheric ozone holes, and a host of other wind-related phenomena in the lower atmosphere. *See* CYCLONE; HURRICANE; TORNADO.

In oceanography, successes of geophysical fluid dynamics include the explanation of major oceanic currents, such as the Gulf Stream. Coastal river plumes, coastal upwelling, shelf-break fronts, and open-ocean variability on scales ranging from tens of kilometers to the size of the basin are among the many other marine applications. The El Niño phenomenon in the tropical Pacific is rooted in processes that fall under the scope of geophysical fluid dynamics. *See* EL NIÑO; GULF STREAM.

Highly successful applications of geophysical fluid dynamics have also been made to other planetary-scale fluids, such as the molten rock in the Earth's outer core, Jupiter's Great Red Spot, and convective gases in stars, including the Sun. *See* JUPITER; SUN.

Benoit Cushman-Roisin

Bibliography. B. Cushman-Roisin, *Introduction to Geophysical Fluid Dynamics*, Prentice Hall, 1994; A. E. Gill, *Atmosphere-Ocean Dynamics*, Academic Press, 1982; J. Pedlosky, *Geophysical Fluid Dynamics*, 2d ed., Springer-Verlag, 1987; H. M. Stommel and D. W. Moore, *An Introduction to the Coriolis Force*, Columbia University Press, 1989.

Geophysics

The study of the Earth and its relations to the rest of the solar system using the principles and practices of physics. Geophysics is considered by some to be a branch of geology, by others a branch of physics. It is distinguished from geology by its use of instruments to make direct and indirect measurements of the phenomena being studied in contrast to the more direct observations of geology, and by its concern with other members of the solar system. *See* EARTH; GEOLOGY; SOLAR SYSTEM.

Subdivisions of Geophysics

Geophysics is divided into a variety of subdisciplines. These can be grouped according to the portion of the Earth with which each is concerned, although there is much overlap. Solid-earth geophysics is concerned with the Earth's interior; oceanography and hydrology with the aqueous parts; meteorology with the lower atmosphere; and aeronomy with the upper atmosphere. Because students of the rest of the solar system use methods similar to those employed by geophysicists to study the Earth, geophysics has grown to encompass studies of the other planets, the Sun, and the space in which these bodies and the Earth move.

Solid-earth geophysics. This discipline is subdivided, according to the methods used to study the Earth, into the sciences of geodesy, geologic thermometry, seismology, and tectonophysics.

Geodesy. Geodesy is the science of the shape and size of the Earth. Estimates of its diameter were made at least as early as the second century B.C. A basic gridwork of carefully located points on the Earth's surface was started early in the seventeenth century, and today forms the basis of all surveys and maps of the Earth's surface. *See* GEODESY.

From the dimensions of the Earth and its gravity, its mass can be calculated. The distribution of mass in the interior is determined from the variation of gravity from place to place on and above the surface and from calculations involving the moment of inertia and the velocities of seismic waves in the Earth's interior. Precise measurements of the orbits of satellites map the gravitational field in the space surrounding the Earth. *See* EARTH, GRAVITY FIELD OF.

Geothermal studies. Heat is escaping continually from the Earth's interior by conduction and by volcanic processes. Measurements of the variations in the rate of upward heat flow provide evidence of the processes active in the Earth's interior. Studies of the thermal conductivity of rocks and of the rate of heat generation by disintegration of radioactive trace elements in the rocks make it possible to estimate the temperature distribution in the Earth's interior. The history of temperature variations in the interior can be modeled by using different assumptions of starting conditions and comparing the results with the limited knowledge of present conditions to test ideas on the evolution of the Earth. *See* EARTH, HEAT FLOW IN; GEOLOGIC THERMOMETRY.

Volcanic processes supplement the heat loss and are an important area of study because of their danger to human activities and because they provide information on how ore deposits form. Studies of volcanic processes on the ocean floor have revealed an environment for living creatures whose energy sources are not derived from sunlight, and may yield clues to the early evolution of life on Earth. *See* DEEP-SEA FAUNA; PREBIOTIC ORGANIC SYNTHESIS; VOLCANOLOGY.

Seismology. This is the science of earthquakes and other ground vibrations. Seismological studies may lead to the development of techniques for reliable earthquake prediction. In addition, earthquake studies may provide better understanding of the nature of the resultant motions, so that buildings and other structures can be designed to withstand their vibrations. By using earthquake vibrations, the structure of the Earth's interior and the patterns of deformation of the Earth can be mapped. *See* EARTHQUAKE; SEISMOLOGY.

Study of the times of passage of seismic waves through the Earth's interior is the principal source of information on the distribution of different types of rocks. Studies of the waves' rate of absorption and coefficients of reflection give information on the physical properties of different layers of the Earth. From the motions recorded at places on the surface, it is possible to locate where within the Earth each earthquake occurs, details of the mechanism of generation of the waves, and the amount of energy released. This provides clues as to the processes of change occurring in the Earth's interior which lead to short- and long-range deformation of the surface.

Because large explosions produce seismic waves which are recorded at great distances from the source, much effort has been given to seismic research as a means of monitoring compliance with a nuclear test ban treaty.

Tectonophysics. Also called geodynamics, this is the science of the deformation of rocks. It consists of tectonics, the study of the broader structural features of the Earth and their origins, as in mountain building; and rock mechanics, the measurement of the strength and related physical properties of rocks. *See* CONTINENTS, EVOLUTION OF; GEODYNAMICS; OROGENY; PLATE TECTONICS; ROCK MECHANICS.

Because the increase of temperature and pressure with depth in the Earth profoundly changes the physical behavior of rocks, a proper understanding of processes in the Earth's interior is possible only when based on the results of laboratory experiments carried out under extreme conditions. Pressures at depths greater than a few tens of miles can be duplicated only under very transient conditions, and usually not at the high temperatures which are prevalent. The atomic structure of minerals goes through phase changes at high pressures which result in the occurrence at great depths of mineral varieties different from those encountered at the surface, with resultant uncertainty as to composition of the Earth's deepest layers. The Earth's interior is likely to remain incompletely explored for a long time in spite of the best efforts of geophysicists. *See* EARTH INTERIOR; HIGH-PRESSURE MINERAL SYNTHESIS.

Hydrospheric geophysics. Water is the compound that makes the Earth unique among large bodies in the universe. It occurs as a gas, a liquid, a solid, and as a component (often a trace element) in rocks. That part of the Earth's water which occurs as a liquid and as ice, largely free to move relatively easily from place to place, is called the hydrosphere. Specialized branches of geophysics have evolved to study water in this region. *See* HYDROSPHERE.

Oceanography. Scientific study of the oceans is concerned with the shape and structure of the ocean basins, the physical and chemical properties of seawater, ocean currents, waves and tides, thermodynamics of the ocean, and the relations of these to the organisms which live in the sea. Knowledge of the ocean is important because it is a source of food, much of the world's commerce travels across its surface, and its heat balance is a major factor affecting weather worldwide. *See* HEAT BALANCE, TERRESTRIAL ATMOSPHERIC; MARINE GEOLOGY; OCEANOGRAPHY.

Hydrology. Fresh water in lakes, in streams, and in the pores of near-surface rocks is the concern of hydrology. Every organism requires water to live. The distribution and purity of the water supply are consequently important. Surface water also carries away much of society's wastes, both sanitary and industrial. Hydrology is concerned with the chemical, physical, and biological processes by which these wastes are changed and removed as water moves through and over the ground toward the oceans. The constant exchange of water between the Earth's interior, the hydrosphere, and the atmosphere is an important aspect of hydrology. The availability of pure water for personal, agricultural, and industrial uses is so essential that pollution control is one of the most rapidly growing scientific and technological problems of modern society. *See* GROUND-WATER HYDROLOGY; HYDROLOGY; WATER POLLUTION.

Glaciology. An appreciable fraction of Earth's water occurs in the form of snow and ice. If all of the water frozen in glaciers were to melt, the ocean surface would rise several hundred feet, changing coastlines substantially. Glaciology, which is often considered a branch of hydrology, is the scientific study of the distribution and movement of this frozen water, how it accumulates and melts, and what it does to the underlying rocks as it flows over them. *See* GLACIOLOGY; TERRESTRIAL WATER.

Meteorology. Meteorology is the study of the composition and movements of the mass of air known as the atmosphere; of the interaction of the atmosphere with living organisms, the hydrosphere, and the solid earth; and of the flow of energy within the atmosphere and to and from the space beyond. An important goal of meteorology is to predict changes in atmospheric conditions, the weather and climate, from observations of current and past conditions and theoretical calculations based on models of how the atmosphere behaves. For this purpose, temperature, pressure, and the moisture conditions in the atmosphere are measured continuously or periodically at a large number of places at and above the Earth's surface, reported to data-processing centers, and used to make regular weather predictions. These are widely distributed by government agencies, private companies, and the mass media, and are used to guide individual and organizational activities. *See* AGRICULTURAL METEOROLOGY; CLIMATE MODIFICATION; INDUSTRIAL METEOROLOGY.

Short-term weather predictions attempt to describe conditions to be expected within a few hours or days; but attempts are made at predicting long-term trends for weeks, months, or even years ahead as well. Data gathered by satellites orbiting the Earth play an important role in making predictions and understanding atmospheric processes. It has become the responsibility of government to warn the public of upcoming extreme weather. *See* ATMOSPHERE; MESOMETEOROLOGY; METEOROLOGY; SATELLITE METEOROLOGY; WEATHER FORECASTING AND PREDICTION; WEATHER MODIFICATION.

Aeronomy. There is no distinct upper boundary to the atmosphere, which becomes progressively less dense with elevation, eventually merging with the Sun's extended atmosphere through which Earth moves. However, the properties of the atmosphere undergo a radical change at an elevation of about 60 mi (100 km), where the atmosphere becomes so highly ionized that its electrical properties become important in controlling its behavior. Aeronomy is the science of this upper part of the atmosphere. Solar radiations are mainly responsible for this ionization, although cosmic rays play a role. Molecules are dissociated into their component atoms, and selective diffusion results in an upward concentration of hydrogen, which may be lost continuously from the Earth. Free electrons as well as ions are present. The electrical particles move, forming currents which induce fluctuations of the Earth's magnetic field. *See* COSMIC RAYS; IONOSPHERE; UPPER-ATMOSPHERE DYNAMICS.

The impact of charged particles from the Sun can be so violent as to produce magnetic storms which disrupt radio communication. It also produces the aurora in high latitudes. Farther out, moving charged particles carried by the Earth in its motion around the Sun form a layer called the magnetosphere, whose shape is distorted by the flow of material radiated from the Sun. The charged layers of the atmosphere protect the Earth from the strongest of the Sun's radiations. Knowledge of the nature and variability of these radiations and their effects on the Earth and its inhabitants has been accumulating rapidly only in the last few decades, and much remains to be discovered. Exploration of the upper atmosphere and the outlying space became possible only with development of rocket probes and placing of artificial satellites in orbits around the Earth. As these researches into the outer fringes of the Earth's atmosphere have progressed, they have overlapped astronomical studies of the Sun to such a degree that solar science has become largely accepted as a part of geophysics. *See* AERONOMY; AURORA; MAGNETOSPHERE; SOLAR WIND; SUN.

Planetology. Until the exploration of space using rockets, the only information available about planets and natural satellites was obtained by use of astronomical telescopes. With the advent of rocket probes, the same diversity of measurements made on the Earth could be carried out on each of the other bodies. Because the methods of measurement used are similar to those used by geophysicists to examine the Earth, studies of the other planets and their moons have become a branch of geophysics known as planetology. *See* MOON; PLANET; PLANETARY PHYSICS.

Overlapping Fields

In addition to the regional subdivisions of geophysics defined above, there are other overlapping distinct specialties.

Geomagnetism. Geomagnetism is concerned with a detailed description of the Earth's magnetic field and its changes and with the magnetic properties of rocks. Because the Earth's magnetic field is used as a guide for navigation and to locate north in surveying, knowledge of its current patterns and changes both at the Earth's surface and in the surrounding space is important. The magnetism of rocks is also useful in delineating the history of changes in the Earth, providing the principal line of evidence showing how the present patterns of continents and oceans have evolved from very different arrangements in the past. Variations in magnetic field strength are also useful in mapping variations in the distribution of rocks of different compositions.

Changes in the magnetic field induce electric currents in the Earth's interior and in the atmosphere. Because magnetism and electricity are so closely linked, geoelectricity is generally considered a part of geomagnetism. Patterns of electric currents can be used to map the variations in the electrical conductivity of buried rocks, providing evidence of the Earth's internal structure. *See* GEOELECTRICITY; GEOMAGNETISM; ROCK MAGNETISM.

Geochronology. This field deals with the dating events in the Earth's history. The principal technique is based on radioactive disintegrations: the proportion of daughter to parent elements in a mineral or rock is a measure of the age of the material. Other methods depend on the redshift of the spectra of distant stars, the rate of recession of the Moon, annual variations in the growth rates of plants and animals including fossils, and the rates of erosion and sedimentation. *See* DATING METHODS; GEOCHRONOMETRY.

Geocosmogony. This is the study of the origin of the Earth. The many hypotheses fall into two groups: those which postulate that the Earth is primarily an aggregate of once smaller particles, and those which claim that it is essentially a fragment of a larger body. Current speculation favors the former theory. Geocosmogony is initimately linked with the origin of the solar system and the Milky Way Galaxy. Many lines of evidence suggest that the formation of the Earth was a typical minor event in the evolution of the Milky Way or of the universe as a whole, occurring 5–10 × 10^9 years ago. *See* COSMOCHEMISTRY; MILKY WAY GALAXY.

Exploration and prospecting. Geophysical techniques are widely used not only to study the general structure of the Earth but also to prospect for petroleum, mineral deposits, and ground water and to map the sites of highways, dams, and other structures. Seismic methods are the most widely used, but electrical, electromagnetic, gravity, magnetic, and radioactivity surveying methods are also well developed. Many types of geophysical surveys can be made by lowering measuring apparatus into boreholes. *See* GEOPHYSICAL EXPLORATION; PROSPECTING; WELL LOGGING.

B. F. Howell, Jr.

Bibliography. J. De Bremaecker, *Geophysics: The Earth's Interior,* 1985; C. M. R. Fowler, *The Solid Earth: An Introduction to Global Geophysics*, 2d ed., 2004; M. N. Hill et al., *The Sea*, vols. 1–8, 1962–1983; P. Kearey, M. Brooks, and I. Hill, *An Introduction to Geophysical Exploration*, 3d ed., 2002; F. K. Lutgens and E. J. Tarbuck, *The Atmosphere: An Introduction to Meteorology*, 10th ed., 2006; W. M. Kaula, *An Introduction to Planetary Physics*, 1976; P. V. Sharma, *Geophysical Methods in Geology*, 2d ed., 1986.

Geostrophic wind

A hypothetical wind based upon the assumption that a perfect balance exists between the horizontal components of the Coriolis force and the horizontal pressure gradient force per unit mass, with the implication that viscous forces and accelerations are negligible. Application of the geostrophic wind facilitates an approximation of the wind field from the pressure data over vast regions in which few wind observations are available.

Bases of the approximation. The geostrophic wind blows parallel to the isobars (lines of equal pressure) with lower pressure to the left of the direction of the wind in the Northern Hemisphere and to the right in the Southern Hemisphere. Its speed is given by the equation below, where ρ is the density of air, Ω is

$$V_{\mathrm{geo}} = \frac{1}{2\rho\Omega\,\sin\phi}\,\frac{\partial p}{\partial n}$$

the angular speed of rotation of the Earth about its axis, p is the atmospheric pressure, ϕ is the latitude, and n is a coordinate normal to the isobars and directed toward higher pressure. The approximation now known as the geostrophic wind was first derived empirically by C. H. D. Buys-Ballot in 1857 and has been known as Buys-Ballot's law.

The geostrophic wind represents a good approximation to the actual wind at elevations greater than about 3000 ft (1 km), except in instances of strongly curved flow, large variations in wind speed, and in the vicinity of the Equator.

Thermal wind. This is a term denoting the net change in the geostrophic wind over some specific vertical distance. This change arises because the rate of change of pressure in the vertical is different in two air columns of different air density, so that the horizontal component of the pressure gradient force per unit mass varies in the vertical. The thermal wind is directed approximately parallel to the isotherms of air temperature with cold air to the left and warm air to the right in the Northern Hemisphere, and vice versa in the Southern Hemisphere. Thus, for example, the increasing predominance of westerly winds aloft may be viewed as a consequence of the warmth of tropical latitudes and the coldness of polar regions. *See* CORIOLIS ACCELERATION; GRADIENT WIND; WIND; WIND STRESS.

Frederick Sanders; Howard B. Bluestein

Geosynthetic

Any synthetic material used in geotechnical engineering. *See* MANUFACTURED FIBER.

Geotextiles

Geotextiles are used with foundations, soils, rock, earth, or other geotechnical material as an integral part of a manufactured project, structure, or system. These textile products are made of synthetic fibers or yarns, constructed into woven or nonwoven fabrics that weigh from 3 to 30 oz/yd^2 (100 to 1000 g/m^2). Geotextiles are more commonly known by other names, for example, filter fabrics, civil engineering fabrics, support membranes, and erosion control cloth.

Permeable geotextiles perform three basic functions in earth structures: separation (the fabric provides a boundary that segregates materials); reinforcement (the fabric imparts tensile strength to the system, thereby increasing its structural stability); and filtration (the fabric retains soil particles while allowing water to pass through). Such geotextiles can thus be adapted to numerous applications in earthwork construction. The major end-use categories are stabilization (for roads, parking lots, embankments, and other structures built over soft ground); drainage (of subgrades, foundations, embankments, dams, or any earth structure requiring seepage control); erosion control (for shoreline, riverbanks, steep embankments, or other earth slopes to protect against the erosive force of moving water); and sedimentation control (for containment of sediment runoff from unvegetated earth slopes). No one geotextile is suited for all these applications. Each use dictates a specific fabric requirement to resist installation stresses and to perform its function once installed.

Stabilization applications. Soft or low-strength soils on a project site present costly problems in the construction and maintenance of haul roads, storage yards, railroads, and other areas which must support vehicular traffic. Poor soil conditions also cause rapid deterioration of paved and unpaved roads, city streets, and parking lots. Problems are caused by subgrade deformation and intermixing between subgrade soil and aggregate

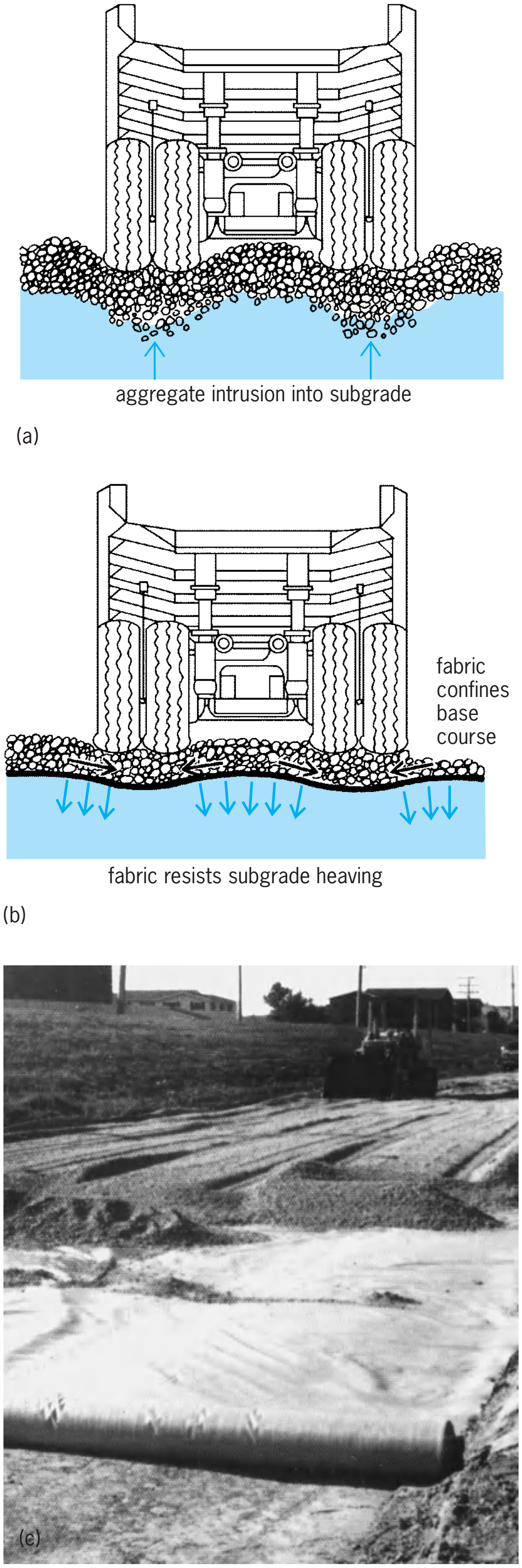

aggregate intrusion into subgrade
(a)
fabric confines base course
fabric resists subgrade heaving
(b)
(c)

Fig. 1. Geotextiles are used to stabilize soils. (*a*) A rutted and unstable surface results from subgrade deformation and intermixing between subgrade soil and aggregate base. (*b*) Geotextiles resist rutting through separation and reinforcement (*from Mirafi Fabrics for the Mining Industry, Celanese Fibers Marketing Co.*). (*c*) A highway base course constructed with geotextile between subgrade and aggregate base (*from Mirafi Family of Construction Fabrics/MPB8, Celanese Fibers Marketing Co.*).

base, resulting in a rutted and unstable surface that impedes or even prohibits the traffic flow (**Fig. 1***a*).

Geotextiles can eliminate or reduce the effect of these soft-soil problems through separation and reinforcement. When placed over a soft soil, the geotextile provides a support membrane for placement and compaction of aggregate base. The fabric barrier prevents aggregate particles from intruding into the soft soil and prevents soil particles from pumping up into the aggregate layer. As a continuous membrane between soil and aggregate base, the geotextile helps confine the aggregate against lateral and vertical movement. This confinement maintains the density and hence the load-distributing characteristics of the aggregate. The fabric also resists the upward heaving of subgrade between wheel paths. If the subgrade is extremely soft and cannot support vehicle loads, the fabric will act as a reinforcing membrane to assist the subgrade in supporting loads (Fig. 1*b* and *c*).

Drainage applications. Soil moisture control through drainage is essential to maintain stability in pavements, foundations, cut slopes, and earth dams. Drainage is accomplished by providing a trench or blanket of porous rock for soil moisture to seep into. A perforated pipe is often installed within the porous rock to collect the moisture and transport it to an outlet. To ensure effective performance and long life, drain structures need a filter to retain soil particles that would clog the drain.

Graded aggregate filters are conventionally used. A properly designed graded aggregate will confine or retain the soil, thus preventing significant particle movement. If, however, the drainage aggregate is too coarse, the voids between rocks at the soil-aggregate interface will be too large for soil particles to bridge across. The resulting lack of soil-particle confinement will result in erosion when water seeps out of the soil, that is, soil piping (**Fig. 2***a*). *See* GROUND-WATER HYDROLOGY.

The pore structure and permeability of some geotextiles are similar to those of graded aggregate filters. These fabrics can provide the same particle retention at the soil-drain interface while permitting unrestricted flow of water from the adjacent soil. Geotextiles can thus eliminate the need for graded aggregate filters in drains. When drainage fabric is used, no special aggregate gradation is required, because the fabric prevents soil from washing into the drain (Fig. 2*b* and *c*). *See* DAM; FOUNDATIONS; PAVEMENT.

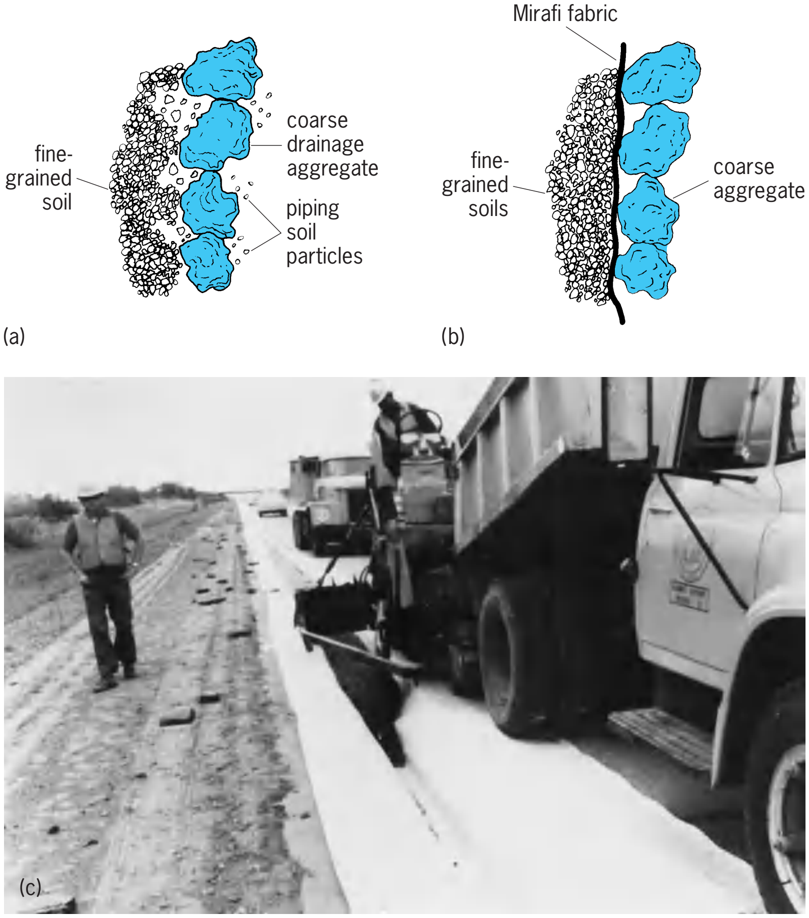


Fig. 2. Geotextiles find wide use for drainage. (*a*) Without a filter, soil particles may wash into and clog the drain. (*b*) Drainage fabric provides retention at the soil-drain interface, while allowing water to pass through (*from Mirafi Fabrics for the Mining Industry, Celanese Fibers Marketing Co.*). (*c*) Drain trench lined with a geotextile and backfilled with coarse aggregate.

Erosion control applications. Embankments along coastal shorelines and inland waterways are subjected to wave and current action that can cause severe erosion, instability, and even destruction of the earth slope. To protect against erosion, earth slopes have often been covered with armor (riprap, concrete blocks, concrete slabs). Despite the armor covering, erosion can still occur, undermining the armor's foundation. To assure long-term stability and performance, the erosion-control structure must include a barrier that shields the soil surface from scouring. This barrier should be permeable so that any moisture seeping from the soil slope can escape without buildup of hydrostatic pressure. Traditionally, granular filters of specially graded sand, gravel, or stone have been used to prevent slope erosion beneath the armor. But granular filters are expensive, particularly when not locally available, and even when properly installed are subject to erosive forces that can wash them away.

Some geotextiles are ideal erosion control barriers. Erosion control fabrics will shield a soil slope from the erosive force of moving water, and they are permeable so that seepage from the earth slope can pass through freely. These fabrics will remain intact, covering the soil slope as long as the armor stone remains above it. *See* EROSION.

Sedimentation control. Severe erosion can occur during earthwork construction or mining operations when protective vegetation is removed and soil slopes are left temporarily unprotected. Resultant sediment runoff can create serious downstream damage, for example, contaminated waterways, clogged storm drains, or sediment-covered forests or

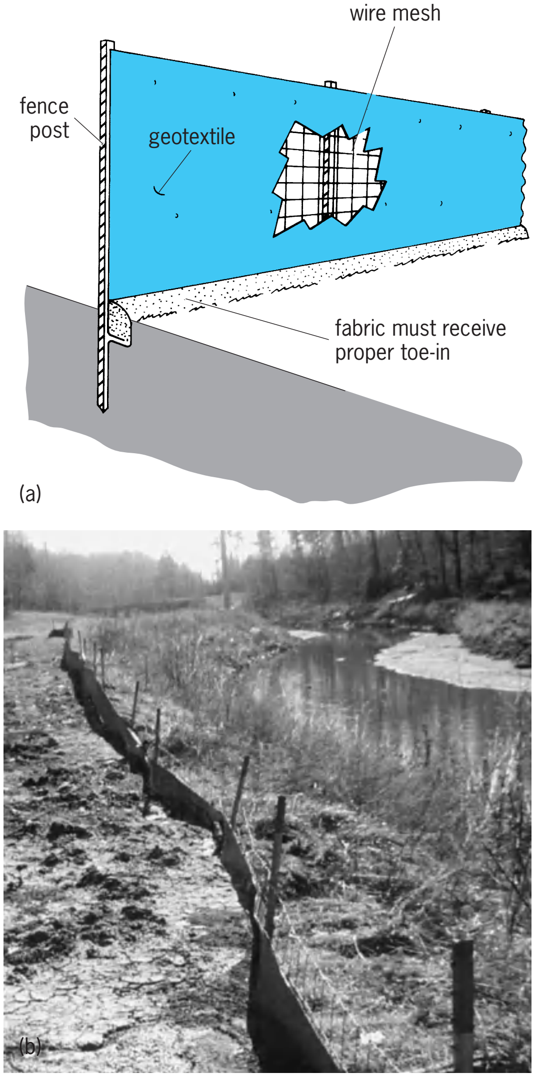


Fig. 3. A silt fence is a fabric-lined fence structure. When lined with a geotextile, it retains sediment runoff, preventing contamination of adjacent waterways. (*Mirafi Fabrics for the Mining Industry, Celanese Fibers Marketing Co.*)

pastures. Government agencies at the federal and state level have recognized the problems associated with sedimentation, and legislation has been passed requiring the control of sediment runoff from any disturbed land area. As a result, earthwork contractors and mine operators are faced with the responsibility and cost of controlling sediment runoff from work sites.

Silt fences constructed with geotextiles offer a cost-effective solution to sedimentation control. A silt fence is a fabric-lined structure installed on, or at the base of, an unvegetated slope. It acts as a water-permeable barrier that retains sediment runoff from the slope. The silt fence can be thought of as a temporary impoundment structure that forms a sediment pond aboveground (**Fig. 3**). When installed along the perimeter of a construction site, a silt fence can prevent sediment from leaving the disturbed area. By installing silt fences along stream banks, sediment can be kept from reaching the waterway. Robert G. Carroll, Jr.

Geomembranes

A geomembrane is any impermeable membrane used with soils, rock, earth, or other geotechnical material in order to block the migration of fluids. These membranes are usually made of synthetic polymers in sheets ranging from 0.01 to 0.14 in. (0.25 to 3.5 mm) thick. Geomembranes are also known as flexible membrane liners, synthetic liners, liners, or polymeric membranes.

Early liners included clay, bentonite, cement-stabilized sand, and asphalt. Such liners, however, have measurable permeability or tend to crack and fissure when exposed to certain environments and chemicals. Traditional soil liners such as clay are regarded as good barriers if their permeability is 10^{-7} cm/s. Geomembranes, by contrast, have no true permeability if there are no leaks or holes in the membrane, and they possess inherent flexibility to accommodate geotechnical settlement and shifting. Migration through a geomembrane takes place at extremely low levels via chemical diffusion. Effective permeabilities for geomembranes can be calculated from measured diffusion rates, and vary from 10^{-11} to 10^{-13} cm/s, depending on the type of polymeric membrane. Modern geomembranes are commonly made of medium-density polyethylenes that are very nearly high-density polyethylenes (HDPE), several types of polyvinyl chloride (PVC), chlorosulfonated polyethylene (a synthetic rubber), ethylene propylene diene monomer (EPDM), and several other materials. Some geomembranes require reinforcement with an internal fabric scrim for added strength, or plasticization with low-molecular-weight additives for greater flexibility. *See* POLYMER.

Applications. Geomembranes are able to contain fluids, thus preventing migration of contaminants or valuable fluid constituents. Since they prevent the dispersal of materials into surrounding regions, geomembranes are often used in conjunction with soil liners, permeable geotextiles, fluid drainage media, and other geotechnical support materials. The major application of geomembranes has been containment of hazardous wastes and prevention of pollution in landfill and surface impoundment construction (**Fig. 4**). They are also used to a large extent in mining to contain chemical leaching solutions and the precious metals leached out of ore, in aquaculture ponds for improved health of aquatic life and improved harvesting procedures, in decorative pond construction, in water and chemical storage-tank repair and spill containment, in agriculture operations, in canal construction and repair, and in construction of floating covers for odor control, evaporation control, or wastewater treatment through anaerobic digestion. *See* HAZARDOUS WASTE.

Basin liners. When geomembranes are used as basin liners, they are often applied in direct contact and on top of soil liners. Such a composite lining system offers extra security for containment because the soil liner backs up the geomembrane. Rolls of geomembrane are deployed downslope, with seams parallel to the slope of the basin. Texture-surface geomembranes have been developed to enhance the stability of steep slopes.

Final cover. Final caps form covers over waste impoundments or other containments. In final-cover applications, geomembranes prevent intrusion of precipitation into isolated and contained areas and, in addition, prevent the escape of gases from the containment. Often the surface over which the final cover is applied is unstable or subject to settlement. Seaming can take place on floating barges, if necessary, or floated into position in the construction of floating covers.

Final covers over landfills are often subject to deformation from differential settling of the subsoils. Geomembranes capable of extreme elongation and flexibility should be selected to accommodate the differential settlement, so that they can be stretched a great distance before tearing and can be bent (flexed) with ease. Textured-surface geomembranes are desirable for final covers because they impart long-term slope stabilities.

Vertical cut-off walls. Geomembranes are hung vertically in repair of tank linings and in construction of vertical barrier walls in soils. They can be used either alone or in conjunction with bentonite slurry trenches to cut off and isolated sections of ground water in cases of pollution remediation, salt-water intrusion, or dike construction. Geomembrane panels are inserted directly into these cut-off walls to depths of 100 ft (30 m) by using vibratory hammers and sheet pile drivers. The adjacent panels are not joined by heat seaming but are interlocked through prefabricated complementary joint sections; the locking mechanism is engaged as the panels are slid alongside one another into the cut-off wall.

Design and installation. A geomembrane is not intended to provide tensile support or load-carrying capacity. Many applications, in waste containment particularly, require maximum durability and longevity, with a lifetime of as much as hundreds of years. Therefore, in addition to selecting materials that are suitably resistant to chemicals and to wear, it is desirable to limit fatigue-inducing tension stresses through good design and installation practice.

Soil subgrades and cover soils must be free of sharp objects such as sticks and angular stones that would introduce puncture stresses on a linear used as a containment barrier. Heavy equipment used to place covering soils over the synthetic liner should be separated from the geomembrane by a suitable thickness of the soil. If necessary, geotextiles are placed above or below a geomembrane in order to increase puncture resistance.

Seaming. Joining adjacent panels of geomembrane is usually accomplished through the application of heat. Some heat seaming (welding) is done in factories during construction of prefabricated panels. However, for most panel installations, some if not all seams must be constructed in the field. Types of welding include extrusion, hot-wedge, and hot-air.

In extrusion welding, a strand of molten polymer is deposited at the edge of the overlapped geomembrane panels, bonding the sheets together.

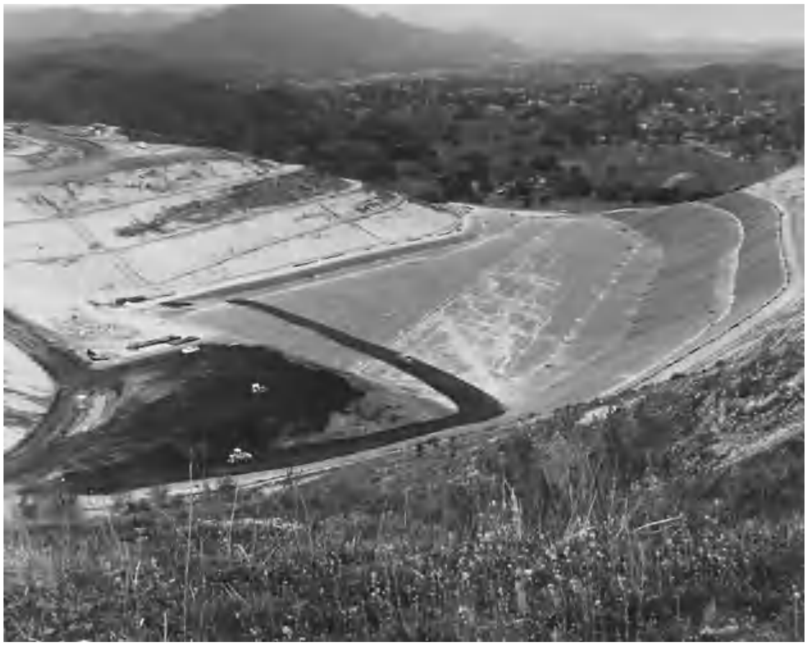


Fig. 4. Calabasas solid-waste landfill in Los Angeles County, California, being lined with geomembrane.

In hot-wedge seaming, a wedge of hot steel is passed between the overlapped panel edges, followed directly by pressure rollers effecting the seam. The most advanced variation creates two welded tracks separated by an air gap. This allows testing by air pressure in the gap to detect potential leaks in the seam (**Fig. 5***a*).

In the hot-air welding method the sheets are melted by hot air blown between them, and then the molten sheets are pressed together. This welding system is not considered sufficiently consistent for final seaming operations; it is generally used only in conjunction with extrusion welding to keep the panels stationary before application of the extrusion weld.

Alternatively, solvent-adhesive seams can be effected for certain geomembranes. In solvent-adhesive welding operations the solvent-adhesive mixture is brushed onto panel edges, and various methods are used to press the panel edges together to form the seam. *See* ADHESIVE.

Testing. Geomembrane seams are tested very carefully to ensure that they are free of leaks and adequately joined. Nondestructive testing is used to check the entire seamed distance for leaks, and if any are found they are repaired.

Destructive testing is used to assess the quality of the bonding between panels. It requires that samples of the finished seam be removed and pulled apart in order to observe bonding. True welding is required, as opposed to surface attachments analogous to those in soldered metals. It is important to determine if welds are fully integrated because a mere surface attachment can be disrupted by absorbed chemicals. Samples are pulled apart by bending back the top panel from the overlapped section of the bottom panel in an attempt to peel the surface between

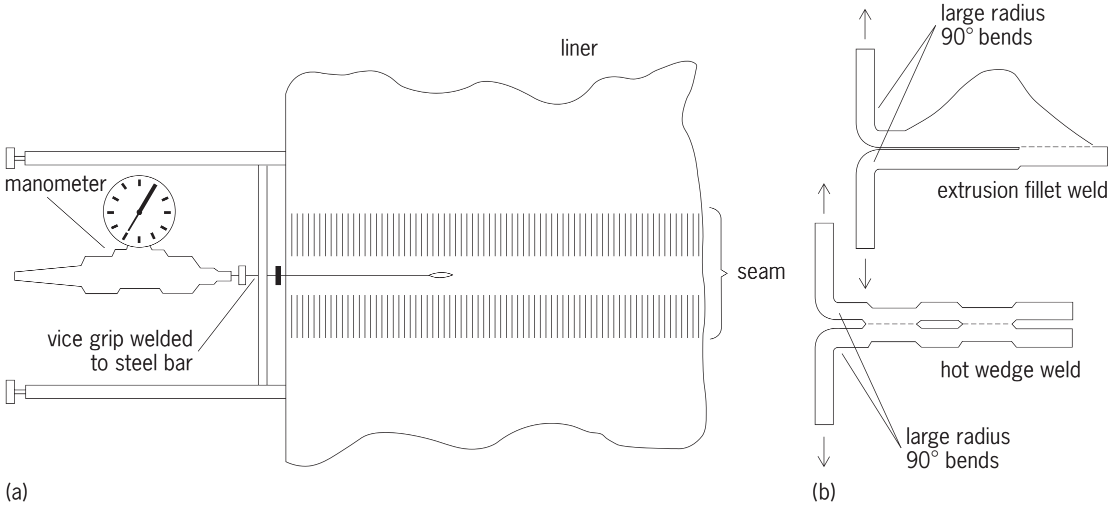


Fig. 5. Schematic diagrams of testing. (*a*) Nondestructive seam air-pressure test for leaks. (*b*) Destructive peel testing of both extrusion and dual-track hot-wedge seams to assess quality of welding, showing specimen configuration for peeling apart seams.

the seamed panels (Fig. 5*b*). If true welding has occurred, the resulting tear is through one of the panels as opposed to delamination (peeling). In contrast to nondestructive testing, destructive testing is limited to an interval basis. *See* ENGINEERING GEOLOGY; SOIL MECHANICS. Mark Cadwallader; Hal Pastner

Bibliography. J. E. Fluet, Jr. (ed.), *Geotextile Testing and the Design Engineer*, 1987; R. M. Koerner and R. F. Wilson-Fahmy (eds.), *Geosynthetic Liner Systems: Innovations, Concerns, and Designs*, 1996.

Geothermal power

Thermal or electrical power produced from the thermal energy contained in the Earth (geothermal energy). Use of geothermal energy is based thermodynamically on the temperature difference between a mass of subsurface rock and water and a mass of water or air at the Earth's surface. This temperature difference allows production of thermal energy that can be either used directly or converted to mechanical or electrical energy.

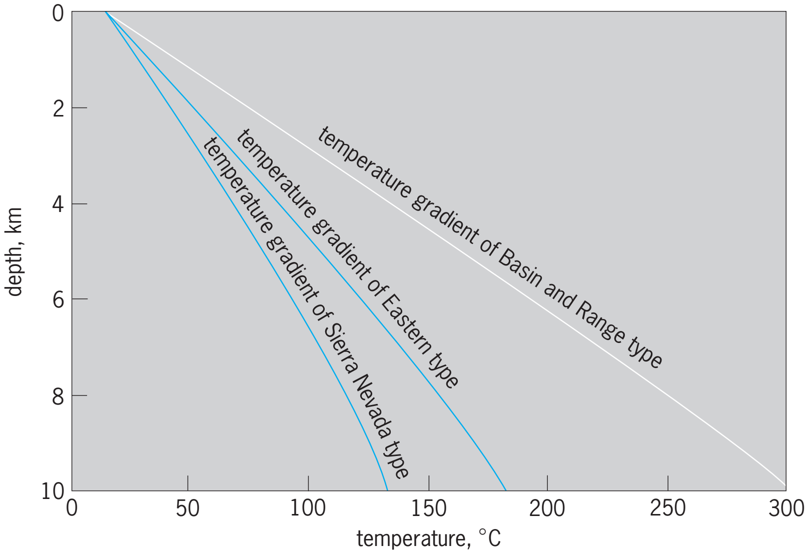


Fig. 1. Calculated average conductive temperature gradients in representative heat-flow provinces of the United States. 1 km = 0.6 mi. °F = (°C × 1.8) + 32. (*After D. E. White and D. L. Williams, eds., Assessment of Geothermal Resources of the United States–1975, USGS Circ. 726, 1975*)

Characteristics

Temperatures in the Earth in general increase with increasing depth, to 400–1800°F (200-1000°C) at the base of the Earth's crust and to perhaps 6300–8100°F (3500–4500°C) at the center of the Earth. Average conductive geothermal gradients to 6 mi (10 km; the depth of the deepest wells drilled to date) are shown in **Fig. 1** for representative heat-flow provinces of the United States. The heat that produces these gradients comes from two sources: flow of heat from the deep crust and mantle; and thermal energy generated in the upper crust by radioactive decay of isotopes of uranium, thorium, and potassium. The gradients of Fig. 1 represent regions of different conductive heat flow from the mantle or deep crust. Some granitic rocks in the upper crust, however, have abnormally high contents of uranium and thorium and thus produce anomalously great amounts of thermal energy and enhanced flow of heat toward the Earth's surface. Consequently thermal gradients at shallow levels above these granitic plutons can be somewhat greater than shown on Fig. 1. *See* EARTH, HEAT FLOW IN.

The thermal gradients of Fig. 1 are calculated under the assumption that heat moves toward the Earth's surface only by thermal conduction through solid rock. However, thermal energy is also transmitted toward the Earth's surface by movement of molten rock (magma) and by circulation of water through interconnected pores and fractures. These processes are superimposed on the regional conduction-dominated gradients of Fig. 1 and give rise to very high temperatures near the Earth's surface. Areas characterized by such high temperatures are the primary targets for geothermal exploration and development.

Commercial exploration and development of geothermal energy to date have focused on natural geothermal reservoirs—volumes of rock at high temperatures (up to 662°F or 350°C) and with both high porosity (pore space, usually filled with water) and high permeability (ability to transmit fluid). The thermal energy is tapped by drilling wells into the reservoirs. The thermal energy in the rock is transferred by conduction to the fluid, which subsequently flows to the well and then to the Earth's surface.

Natural geothermal reservoirs, however, make up only a small fraction of the upper 6 mi (10 km) of the Earth's crust. The remainder is rock of relatively low permeability whose thermal energy cannot be produced without fracturing the rock artificially by means of explosives or hydrofracturing. Experiments involving artificial fracturing of hot rock have been performed, and extraction of energy by circulation of water through a network of these artificial fractures may someday prove economically feasible.

There are several types of natural geothermal reservoirs. All the reservoirs developed to date for electrical energy are termed hydrothermal convection systems and are characterized by circulation of meteoric (surface) water to depth. The driving force of the convection systems is gravity, effective because of the density difference between cold, downward-moving, recharge water and heated, upward-moving, thermal water. A hydrothermal convection system can be driven either by an underlying young igneous intrusion or by merely deep circulation of water along faults and fractures. Depending on the physical state of the pore fluid, there are two kinds of hydrothermal convection systems: liquid-dominated, in which all the pores and fractures are filled with liquid water that exists at temperatures well above boiling at atmospheric pressure, owing to the pressure of overlying water; and vapor-dominated, in which the larger pores and fractures are filled with steam. Liquid-dominated reservoirs produce either water or a mixture of water and steam, whereas vapor-dominated reservoirs produce only steam, in most cases superheated.

Natural geothermal reservoirs also occur as regional aquifers, such as the Dogger Limestone of the Paris Basin in France and the sandstones of the Pannonian series of central Hungary. In some rapidly subsiding young sedimentary basins such as the northern Gulf of Mexico Basin, porous reservoir sandstones are compartmentalized by growth faults into individual reservoirs that can have fluid pressures exceeding that of a column of water and approaching that of the overlying rock. The pore water is prevented from escaping by the impermeable shale that surrounds the compartmented sandstone. The energy in these geopressured reservoirs consists not only of thermal energy, but also of an equal amount of energy from methane dissolved in the waters plus a small amount of mechanical energy due to the high fluid pressures. *See* AQUIFER; GROUNDWATER HYDROLOGY.

Use of Geothermal Energy

Although geothermal energy is present everywhere beneath the Earth's surface, its use is possible only when certain conditions are met: (1) The energy must be accessible to drilling, usually at depths of less than 2 mi (3 km) but possibly at depths of 4 mi (6–7 km) in particularly favorable environments (such as in the northern Gulf of Mexico Basin of the United States). (2) Pending demonstration of the technology and economics for fracturing and producing energy from rock of low permeability, the reservoir porosity and permeability must be sufficiently high to allow production of large quantities of thermal water. (3) Since a major cost in geothermal development is drilling and since costs per meter increase with increasing depth, the shallower the concentration of geothermal energy the better. (4) Geothermal fluids can be transported economically by pipeline on the Earth's surface only a few tens of kilometers, and thus any generating or direct-use facility must be located at or near the geothermal anomaly.

Direct use. Equally important worldwide is the direct use of geothermal energy, often at reservoir temperatures less than 212°F (100°C). Geothermal energy is used directly in a number of ways: to heat buildings (individual houses, apartment complexes, and even whole communities); to cool buildings (using lithium bromide absorption units); to heat greenhouses and soil; and to provide hot or warm water for domestic use, for product processing (for example, the production of paper), for the culture of shellfish and fish, for swimming pools, and for therapeutic (healing) purposes. J. Patrick Muffler

Major localities where geothermal energy is directly used include Iceland (30% of net energy consumption, primarily as domestic heating), the Paris Basin of France (where 140–160°F or 60–70°C water is used in district heating systems for the communities of Melun, Creil, and Villeneuve la Garenne), and the Pannonian Basin of Hungary.

Electric power generation. The use of geothermal energy for electric power generation has become widespread because of several factors. Countries where geothermal resources are prevalent have desired to develop their own resources in contrast to importing fuel for power generation. In countries where many resource alternatives are available for power generation, including geothermal, geothermal has been a preferred resource because it cannot be transported for sale, and the use of geothermal energy enables fossil fuels to be used for higher and better purposes than power generation. Also, geothermal steam has become an attractive power generation alternative because of environmental benefits and because the unit sizes are small (normally less than 100 MW). Moreover, geothermal plants can be built much more rapidly than plants using fossil fuel and nuclear resources, which, for economic purposes, have to be very large in size. Electrical utility systems are also more reliable if their power sources are not concentrated in a small number of large units.

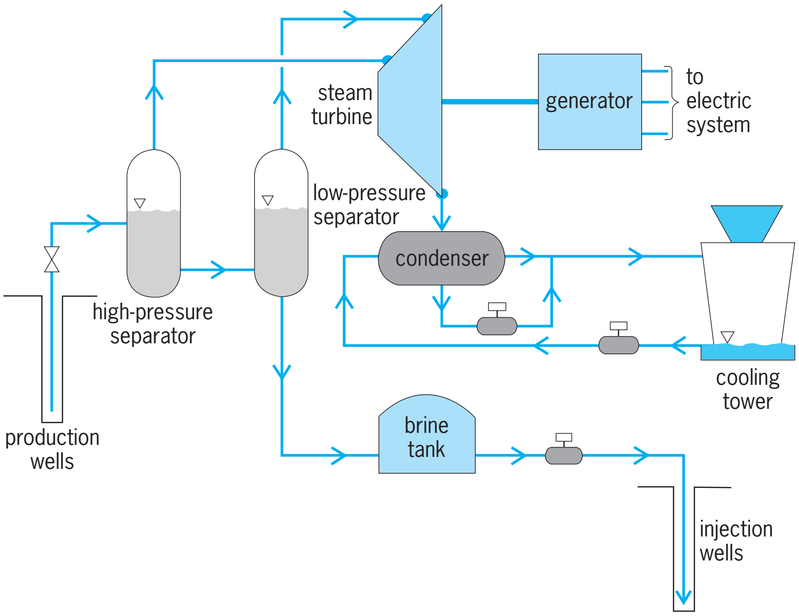


Fig. 2. Schematic diagram of the steam flash process.

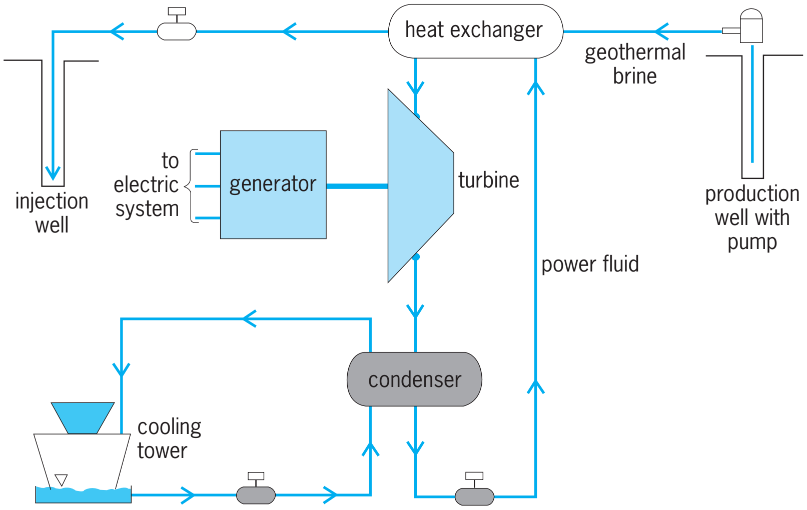


Fig. 3. Schematic diagram of the binary process.

In the United States a law was passed in 1978 that required the output from geothermal power generation projects (and others not based on fossil fuel resources, and cogeneration projects) to be purchased by electrical utilities at the cost that was avoided by the utility as a result of obtaining the power from a geothermal power plant. The legislation is called the Federal Public Utility Regulatory Policies Act (PURPA) and has created an incentive for the development of geothermal power projects.

The process used for generating power varies in accordance with the characteristics of the geothermal resource. The characteristics that affect the process are the temperature, the suspended and dissolved solids in the resource, and the level of noncondensable gases (primarily carbon dioxide) entrained in the geothermal brine, or steam. Almost all resources discovered to date are of the hydrothermal type (pressurized hot water) which can be produced from a well by two methods. If the temperature of a hydrothermal resource is below 400°F (204°C), a geothermal well can be produced with a pump, which maintains sufficient pressure on the geothermal brine to keep it as pressurized hot water. For hydrothermal resources over 400°F, the more suitable method of production is to flow the wells naturally, yielding a flashing mixture of brine and steam from the wells.

Steam flash process. The most common process is the steam flash process (**Fig. 2**), which incorporates steam separators to take the steam from a flashing geothermal well and passes the steam through a turbine that drives an electric generator. For the greatest efficiency in this process, a double-entry turbine is utilized which enables the most amount of steam available in the production from the geothermal well to be converted to electric power. If the resource has a high level of suspended and dissolved solids, it may be necessary to incorporate scaling control equipment in the steam flash vessel at the front of the plant and solids-settling equipment at the tail end of the plant. This will keep the process equipment from becoming plugged and allows a clean residual brine to be maintained for reinjection into the reservoir. If there are significant amounts of noncondensable gases, it may be necessary to install equipment to eject these gases out of the condenser to keep the back pressure on the system from rising and thereby cutting down on the efficiency of the process.

There are, at present, two resources in operation that have “dry steam,” which is produced from the wells directly. These are very easy to convert to electric power and use the above described process without the necessity of the separation and brine injection equipment.

Binary process. A more efficient utilization of the resource can be obtained by using the binary process (**Fig. 3**) on resources with a temperature less than 360°F (180°C). This process is normally used when wells are pumped. The pressurized geothermal brine yields its heat energy to a second fluid in heat exchangers and is reinjected into the reservoir. The second fluid (commonly referred to as the power fluid) has a lower boiling temperature than the geothermal brine and therefore becomes a vapor on the exit of the heat exchangers. It is separately pumped as a liquid before going through the heat exchangers. The vaporized, high-pressure gas then passes through a turbine that drives an electric generator. The vapor exhaust from the turbine is then condensed in conventional condensers and is pumped back through the heat exchangers. There is a distinct environmental advantage to this process since both the geothermal brine and power fluid systems are closed from the atmosphere. Hydrocarbons, such as isobutane and propane, are common power fluids used in this process. *See* ELECTRIC POWER GENERATION.

Thomas C. Hinrichs

Production and Pollution Problems

The chief problems in producing geothermal power involve mineral deposition, changes in hydrological conditions, and corrosion of equipment. Pollution problems arise in handling geothermal effluents, both water and steam.

Mineral deposition. In some water-dominated fields there may be mineral deposition from boiling geothermal fluid. Silica deposition in wells caused problems in the Salton Sea, California, field; more commonly, calcium carbonate scale formation in wells or in the country rock may limit field developments, for example, in Turkey and the Philippines. Fields with hot waters high in total carbonate are now regarded with suspicion for simple development. In the disposal of hot wastewaters at the surface, silica deposition in flumes and waterways can be troublesome.

Hydrological changes. Extensive production from wells changes the local hydrological conditions. Decreasing aquifer pressures may cause boiling water in the rocks (leading to changes in well fluid characteristics), encroachment of cool water from the outskirts of the field, or changes in water chemistry through lowered temperatures and gas concentrations. After an extensive withdrawal of hot water from rocks of low strength, localized ground subsidence may occur (up to several meters) and the original natural thermal activity may diminish in intensity. Some changes occur in all fields, and a good understanding of the geology and hydrology of a system is needed so that the well withdrawal rate can be matched to the well's long-term capacity to supply fluid.

Corrosion. Geothermal waters cause an accelerated corrosion of most metal alloys, but this is not a serious utilization problem except, very rarely, in areas where wells tap high-temperature acidic waters (for example, in active volcanic zones.) The usual deep geothermal water is of near-neutral pH. The principal metal corrosion effects to be avoided are sulfide and chloride stress corrosion of certain stainless and high-strength steels and the rapid corrosion of copper-based alloys. Hydrogen sulfide, or its oxidation products, also causes a more rapid degradation than normal of building materials, such as concrete, plastics, and paints. *See* CORROSION.

Pollution. A high noise level can arise from unsilenced discharging wells (up to 120 decibels adjusted), and well discharges may spray saline and silica-containing fluids on vegetation and buildings. Good engineering practice can reduce these effects to acceptable levels.

Because of the lower efficiency of geothermal power stations, they emit more water vapor per unit capacity than fossil-fuel stations. Steam from wellhead silencers and power station cooling towers may cause an increasing tendency for local fog and winter ice formation. Geothermal effluent waters liberated into waterways may cause a thermal pollution problem unless diluted by at least 100:1.

Geothermal power stations may have four major effluent streams. Large volumes of hot saline effluent water are produced in liquid-dominated fields. Impure water vapor rises from the station cooling towers, which also produce a condensate stream containing varying concentrations of ammonia, sulfide, carbonate, and boron. Waste gases flow from the gas extraction pump vent.

Pollutants in geothermal steam. Geothermal steam supplies differ widely in gas content (often 0.1–5%). The gas is predominantly carbon dioxide, hydrogen sulfide, methane, and ammonia. Venting of hydrogen sulfide gas may cause local objections if it is not adequately dispersed, and a major geothermal station near communities with a low tolerance to odor may require a sulfur recovery unit (such as the Stretford process unit). Sulfide dispersal effects on trees and plants appear to be small. The low radon concentrations in steam (3–200 nanocuries/kg or 0.1–7.4 kilobecquerels/kg), when dispersed, are unlikely to be of health significance. The mercury in geothermal stream (often 1–10 microgram/kg) is finally released into the atmosphere, but the concentrations created are unlikely to be hazardous. *See* AIR POLLUTION.

Geothermal waters. The compositions of geothermal waters vary widely. Those in recent volcanic areas are commonly dilute (<0.5%) saline solutions, but waters in sedimentary basins or active volcanic areas range upward to concentrated brines. In comparison with surface waters, most geothermal waters contain exceptional concentrations of boron, fluoride, ammonia, silica, hydrogen sulfide, and arsenic. In the common dilute geothermal waters, the concentrations of heavy metals such as iron, manganese, lead, zinc, cadmium, and thallium seldom exceed the levels permissible in drinking waters. However, the concentrated brines may contain appreciable levels of heavy metals (parts per million or greater).

Because of their composition, effluent geothermal waters or condensates may adversely affect potable or irrigation water supplies and aquatic life. Ammonia can increase weed growth in waterways and promote eutrophication, while the entry of boron to irrigation waters may affect sensitive plants such as citrus. Small quantities of metal sulfide precipitates from waters, containing arsenic, antimony, and mercury, can accumulate in stream sediments and cause fish to derive undesirably high (over 0.5 ppm) mercury concentrations. *See* WATER POLLUTION.

Reinjection. The problem of surface disposal may be avoided by reinjection of wastewaters or condensates back into the countryside through disposal wells. Steam condensate reinjection has few problems and is practiced in Italy and the United States. The much larger volumes of separated waste hot water (about 55 tons or 50 metric tons per megawatt-electric) from water-dominated fields present a more difficult reinjection situation. Silica and carbonate deposition may cause blockages in rock fissures if appropriate temperature, chemical, and hydrological regimes are not met at the disposal depth. In some cases, chemical processing of brines may be necessary before reinjection. Selective reinjection of water into the thermal system may help to retain aquifer pressures and to extract further heat from the rock. A successful water reinjection system has operated for several years at Ahuachapan, El Salvador.

A. J. Ellis

Bibliography. H. C. H. Armstead, *Geothermal Energy*, 2d ed., 1983; R. Bowen, *Geothermal Resources*, 2d ed., 1989; M. Economides and P. Ungemach (eds.), *Applied Geothermics*, 1987; L. Edwards et al. (eds.), *Handbook of Geothermal Energy*, 1982; J. Elder, *Geothermal Systems*, 1981; M. A. Grant et al., *Geothermal Reservoir Engineering*, 1983; K. Wohletz and G. Heiken,*Volcanology and Geothermal Energy*, 1992.

Geraniales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the superorder Rosidae of Eudicotyledon. The order consists of 6 families (Francoaceae, Geraniaceae, Greyiaceae, Ledocarpaceae, Melianthaceae, Vivianiaceae), 15 genera, and approximately 700 species. The Geraniaceae constitute the vast majority of the order and are temperate herbs or soft shrubs with deeply cleft or compound leaves (see **illus.**). The other families are mainly woody and are found in South America or Africa. Flowers typically have 5 sepals and petals, 10 stamens, and 5 fused carpels that separate from the central axis of the pistil when in fruit. Many of the species possess volatile compounds, as in *Pelargonium* (Geraniaceae) and *Melianthus* (Melianthaceae). *See* MAGNOLIOPHYTA; PLANT KINGDOM; ROSIDAE.

K. J. Sytsma

A common eastern United States species of geranium (*Geranium maculatum*), which is characteristic of the order Geraniales. (*A. W. Ambler, National Audubon Society*)

Gerbil

The common name for 88 species of small rodents comprising the subfamily Gerbillinae in the family Muridae (see **table**). They inhabit the desert regions, steppes, and sandy wastes of Asia, Africa, and southern Russia, often living many miles from water.

Genera and common names of gerbils

Genus	Numbers of species	Common name
Gerbillus	38	Northern pygmy gerbils
Microdillus	1	Somalian gerbil
Gerbillurus	4	Southern pygmy gerbils
Tatera	12	Large naked-soled gerbils
Taterillus	8	Small naked-soled gerbils
Desmodillus	1	Cape short-eared gerbil
Desmodilliscus	1	Brauer's gerbil
Pachyuromys	1	Fat-tailed gerbil
Ammodillus	1	Walo
Sekeetamys	1	Bushy-tailed jird
Meriones	16	Jirds
Brachiones	1	Przewalski's gerbil
Psammomys	2	Fat sand rats
Rhombomys	1	Great gerbil

Morphology. Most gerbils are about the size of a rat and are protectively colored to resemble the desert sands in which they live. They have a head and body length of 50–200 mm (2–8 in.), a tail length of 56–245 mm (2.2–9.8 in.), and a weight of 10–227 g (0.4–8 oz). They have large ears, large hindfeet, well-developed claws, and a long tail. Some have hair on the soles of the feet and bristles on the toes, presumably to prevent them from sinking into the loose soil. The smallest is the Brauer's gerbil with a head and body length of 41–74 mm (1.6–2.9 in.), a tail length of 33–49 mm (1.3–1.9 in.), and a weight of 6–14 g (0.2–0.5 oz). The fat-tailed gerbil (*Pachyuromys*) has a tail only half as long as its body. The tail is thickened and sausage-shaped, and serves as a fat-storage organ for this species, which does not store food in its burrows.

Many small mammals (gerbils, kangaroo rats, jerboas, golden hamsters, elephant shrews, etc.) living in hot arid conditions have relatively large auditory bullae (the bulbous bones enclosing their hearing apparatus). The arid air of the desert has poor sound-carrying qualities in comparison to cooler humid air. Thus, the external ears are large and very sensitive and are said to be suitable for receiving long-range low-frequency sounds. The inflated bullae increase the volume of the air-filled chambers surrounding the middle ear, which in turn reduces the resistance to the inward movement of the tympanic membrane. The malleus, part of which is greatly lengthened, is allowed to rotate more freely and has increased leverage. This transforms relatively weak vibrations of the greatly enlarged tympanic membrane into more powerful movements which are transmitted to the incus and the small footplate of the stapes, which rests against the oval window of the inner ear. *See* EAR (VERTEBRATE).

Gerbils display varying degrees of saltatory (jumping) locomotion (see **illustration**). They are often called antelope rats because of the way they move about. They hop rather than scamper or scurry in typical rat fashion. In this respect they remind one of jerboas, or jumping rodents; indeed, "gerbil" is just another form of the name "jerboa." *See* JERBOA.

Habitat and behavior. Gerbils are primarily nocturnal mammals. They are active year-round, although

The gerbil has a long tail that is used for balance when it hops.

their activity may be reduced during the winter in some areas. They feed on seeds, grasses, and roots, which for the most part contain a little moisture. Food may be stored for winter use. Some species may also include insects, birds, and bird eggs in their diet.

Fat sand rats (*Psammomys*) have the most efficient kidney known, but the necessity for efficiency is not a lack of water but an abundance of salt. Sand rats are desert rodents, but their habitat is restricted to the edges of saline and brackish waters. Sand rats feed on plants containing 80–90% water with about twice the salinity of seawater. The plentiful salt content makes it necessary for the kidney of the sand rat to be able to greatly concentrate the urine.

Gerbils live in burrows excavated in the sand. Burrows are variable in structure and may contain several levels with nest chambers, lengthy tunnels, and food storage chambers. Gerbils may be sociable and share community tunnels, or they may lead a solitary life. Males and females usually do not nest together.

The largest of the gerbils are the great gerbil and the large naked-soled gerbils. The great gerbil (*Rhombomys*) lives in central Asia from the Caspian Sea to southern Mongolia and north-central China, Iran, Afghanistan, and western Pakistan. Adults have a head and body length of 150–200 mm ($5\ ^{3}/_{4}$–$7\ ^{3}/_{4}$ in.) and a tail length of 130–160 mm (5–6 in.). The fur is thick and soft, and the tail is hairy, almost bushy. The soles of the feet are hairy, and the toes end in large claws. Females are polyestrous and produce two or three litters annually. Litter size may range from one to fourteen, and the young are born after a gestation of 23–32 days. Females may reach sexual maturity at 3–4 months. Longevity ranges 2–4 years. This species is considered a pest in Central Asia, where it damages crops, railway embankments, and irrigation channels. This species is also a reservoir for plague.

The large naked-sole gerbils (*Tatera*) have a head and body length of 90–200 mm ($3\ ^{1}/_{2}$–$7\ ^{1}/_{2}$ in.), a tail length of 115–245 mm ($4^{1}/_{2}$–$9^{1}/_{2}$ in.), and a weight of 30–227 g (1–8 oz). The body is heavy and ratlike. The dorsal coloration ranges from grayish to brownish; the underparts are white or whitish. The soles of the feet are naked. These gerbils usually walk on all four limbs, and when alarmed they escape by bounding as high as 1.5 m (4.9 ft). One species is said to be able to cover 3.5 m (11.5 ft) in one leap.

The Mongolian gerbil (*Meriones unguiculatus*) has become the most widely known species of gerbil since the 1950s because it has become a laboratory and pet animal and is now commercially traded almost worldwide. This species is gregarious and makes a good pet in captivity. *See* RODENTIA.

Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, vol. 3, McGraw-Hill, 1990; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Germ layers

The first cell layers that appear in the development of all animals, and from which the embryo body and extraembryonic membranes (when present) are constructed. The cells of the early embryo (the blastula) are organized into supracellular units, the germ layers. These are more or less distinct anatomically but do not necessarily have sharp boundaries of demarcation, in part because one layer gives rise to another. Germ layers are universal among animal embryos; they are an efficient method for establishing discontinuities of architectural importance and for setting aside what will become lineages of cells with different fates. *See* DEVELOPMENTAL BIOLOGY.

Types. Traditionally, three germ layers are recognizable: ectoderm, endoderm, and mesoderm as the outer, inner, and middle layers, respectively. Coelenterates and sponges possess only two of these layers—ectoderm and endoderm—and so are diploblastic. All other metazoans (multicellular animals) have the three germ layers and so are triploblastic. In recent years, a fourth germ layer—the neural crest—has been recognized as a distinctive layer restricted to vertebrates, which are therefore quadroblastic. The evolutionary innovations that followed the evolution of mesoderm (and that characterize all invertebrates other than sponges and coelenterates) are paralleled by evolutionary innovations that followed the evolution of the neural crest (which characterizes all vertebrates). *See* METAZOA; VERTEBRATA.

Endoderm and ectoderm are present in unfertilized eggs; they are set aside in the egg during oogenesis as a result of gene products deposited into the egg by the mother. Consequently, endoderm and ectoderm are regarded as primary germ layers—they arise first in individual development and were the first to appear phylogenetically. Mesoderm and neural crest arise later in development through inductive interactions: mesoderm from interactions between endoderm and ectoderm; neural crest following interaction between mesoderm and neurectoderm (neural ectoderm) and from interactions within the neurectoderm. Consequently, mesoderm and neural crest are regarded as secondary germ layers—they arise secondarily in individual development and evolved subsequent to primary germ layers. The posterior region of many, if not all, vertebrate embryos does not form from germ layers but from a germinal zone of cells that does not separate into germ layers. As this occurs at neurulation (differentiation of nerve tissue and formation of the neural tube), we speak of primary neurulation (from germ layers) and secondary neurulation (from a germinal zone).

Three of the germ layers are named in accordance with their definitive positions in the spherical type of gastrula seen in sea urchins and amphibians, in which the ectoderm surrounds the embryo, the endoderm has invaginated as the precursor of the embryonic gut, and the mesoderm has delaminated to lie between ectoderm and endoderm. The neural crest is named from its topographical position in the crests of the neural folds at the onset of neurulation. In blastoderm types of blastulae, as in birds, a more or less distinct and complete lower endodermal layer (hypoblast) and an upper layer (epiblast) containing the prospective ectodermal and mesodermal layer-forming cells are present. In gastrulae of reptiles, birds, mammals, and a few invertebrate types, where the spherical form of the blastula has been modified into a two-layered blastodisc, the definitive positions of the layers remain essentially the same. The terms epiblast, mesoblast, and hypoblast are sometimes used as synonyms for ectoderm, mesoderm, and endoderm, respectively. However, these are geographical (topographical) terms; early in development the epiblast can contain future ectoderm and mesoderm or even all three future germ layers. Therefore, these terms should not be used as synonyms for germ layers. *See* BLASTULATION; GASTRULATION.

Also, the terms ectoderm and mesoderm should not be confused with the terms epithelia and mesenchyme; the former are germ layers, the latter subsequent types of cellular organization. Epithelia are sheets of polarized laterally connected cells (in simplest form, one cell layer thick) situated on a basement membrane secreted and deposited by the epithelial cells. Mesenchyme is a meshwork of unconnected cells that reside within extracellular matrix (ECM) which they synthesize and deposit. The ECM may be solid as in bone or fluid as in blood. Epithelia and mesenchyme can arise from all four germ layers, so the names for germ layers and for cellular organization should not be conflated. Sometimes, we use the terms mesenchyme for mesenchyme of mesodermal origin, and ectomesenchyme or endomesenchyme for mesenchyme of ectodermal or endodermal origin. *See* EMBRYONIC DIFFERENTIATION; EMBRYONIC INDUCTION.

Origin. The germ-layer structure of embryos has been known for almost 200 years. In 1817 Heinrich Christian Pander described the three-layered structure of the chick blastoderm, and within a decade (1828–1837) Karl Ernst von Baer recognized that the layer concept held true for many types of embryos, both vertebrate and invertebrate. During the latter half of the nineteenth century, the following concept of the origin of the germ layers gradually developed. The blastula was considered to be a single-layered hollow sphere which became converted into a two-layered gastrula by a process of invagination or delamination of cells from one wall of the blastula. The outer layer thus became the ectoderm; the inner layer, the endoderm. A third layer, the mesoderm, then arose from part of either the inner or outer layer, depending upon the group of animals; sometimes the terms endomesoderm and ectomesoderm are used to reflect the endodermal or ectodermal origins of mesoderm.

This concept of the origin of one germ layer from another during the process of gastrulation came to be accepted as the general and basic method of development of most metazoans. Today, we recognize that specific germ-layer-forming regions are present prior to the actual arrangement of the cells into distinct layers. We speak of these as future or presumptive germ layers, a state normally found in zygotes (for the two primary germ layers), in blastulae for mesoderm and in neurulae for neural crest. The fate of presumptive germ layers is not fixed; transplanting future ectoderm from a frog blastula into the future endoderm of another blastula results in the “ectodermal” cells forming endoderm. If the same transplantation is done using gastrula-stage ectoderm, we find that the cells no longer can change their fate—they form a piece of skin (an ectodermal derivative) within the gut (an endodermal derivative).

Fate of the layers. Once the germ layers have become established in their definitive anatomical positions, their further development consists, among other changes, in the segregation of subdivisions within them and in their morphological rearrangement. In vertebrates the ectodermal layer becomes subdivided into a skin-forming area and a neural-forming area. The latter area folds up to form the neural tube, from the crest of which will arise neural crest cells. The mesodermal layer in many invertebrates and vertebrates separates into two sublayers, one of which becomes applied to the overlying ectoderm, the other to the underlying endoderm. The space between the sublayers develops into the coelomic cavity. In vertebrates the portion of the mesoderm closest to the midline boxy axis (paraxial mesoderm) undergoes segregation into dorsal mesoderm to form a medial and dorsal, axial notochord, a rodlike structure around which the vertebral column is later formed. The notochord does not extend into the head; in the head, prechordal mesoderm and neural crest both form mesenchyme, with head muscles arising from mesoderm and the craniofacial skeleton arising from mesoderm and from neural crest. Lateral to the notochord, the mesoderm differentiates into blocks of mesoderm known as somites (one member of each pair on either side of the notochord and developing spinal cord), which later contribute to formation of skeletal muscles, vertebrae, and other structures. Initially epithelial, somites transform from an epithelial to a mesenchymal organization, providing a nice example of the distinction between mesoderm and mesenchyme and of the origin of an epithelium from mesoderm. Still more laterally, the ventral mesoderm (the lateral plate) is further subdivided into an outer, or somatic, and inner, or splanchnic, layer, with the coelomic space between. An area of unsegmented mesoderm between the somites and the lateral plate, known as the intermediate mesoderm, contributes to the kidneys. This positioning and subdivision of primary and secondary germ layers constitute the basic or primitive body plan characteristic of all vertebrates. Further development of the germ layers consists of the formation of all the tissues and organs of the adult, except in some species of the germ cells, which arise from a separate germ plasm. The following list indicates the germ layer origin of some common organs and tissues of adult vertebrates.

Ectoderm
- Epidermis (hair, feathers, lens of eye)
- Central nervous system (brain, spinal cord, cranial and spinal nerves)
- Epithelia of sense organs, mouth, salivary glands
- Pituitary gland
- Enamel of teeth

Mesoderm
- Muscle
- Cartilage
- Bone
- Blood vessels
- Blood
- Kidney
- Gonads
- Liver (in part)
- Thyroid (in part)

Endoderm
- Epithelia of pharynx, thyroid, lungs
- Inner lining of digestive tract
- Bladder (in part)
- Urethra (in part)
- Liver (in part)

Neural crest
- Cartilage
- Bone
- Dentine of teeth
- Adrenal gland (part)
- Septa and valves of the heart
- Peripheral nervous system
- Enteric ganglia

Brian K. Hall; Nelson T. Spratt

Bibliography. S. F. Gilbert and S. R. Singer, *Developmental Biology*, 8th ed., 2006; B. K. Hall, *The Neural Crest in Development and Evolution*, 1999; J. M. W. Slack, *Essential Developmental Biology*, 2d ed., 2006; L. Wolpert et al., *Principles of Development*, 3d ed., 2006.

Germanium

A brittle, silvery-gray, metallic chemical element, Ge, atomic number 32, atomic weight 72.59, melting point 937.4°C (1719°F), and boiling point 2830°C (5130°F), with properties between silicon and tin. Germanium is distributed widely in the Earth's crust in an abundance of 6.7 parts per million (ppm). Germanium is found as the sulfide or is associated with sulfide ores of other elements, particularly those of copper, zinc, lead, tin, and antimony. *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

Germanium has a metallic appearance but exhibits the physical and chemical properties of a metal only under special conditions since it is located in the periodic table where the transition from nonmetal to metal occurs. At room temperature there is little indication of plastic flow and consequently it behaves like a brittle material.

As it exists in compounds, germanium is either divalent or tetravalent. The divalent compounds (oxide, sulfide, and all four halides) are easily reduced or oxidized. The tetravalent compounds are more stable. Organogermanium compounds are many in number and, in this respect, germanium resembles silicon. Interest in organogermanium compounds has centered around their biological action. Germanium in its derivatives appears to have a lower mammalian toxicity than tin or lead compounds.

The properties of germanium are such that there are several important applications for this element, especially in the semiconductor industry. The first solid-state device, the transistor, was made of germanium. Single-crystal germanium is used as a substrate for vapor-phase growth of GaAs and GaAsP thin films in some light-emitting diodes. Germanium lenses and filters are used in instruments operating in the infrared region of the spectrum. Mercury-doped and copper-doped germanium are used as infrared detectors; synthetic garnets with magnetic properties may have applications for high-power microwave devices and magnetic bubble memories; and germanium additives increase usable ampere-hours in storage batteries.

Paul S. Gleim

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; R. Hull and J. C. Bean (eds.), *Germanium Silicon: Physics & Materials*, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; Z. Rappoport (ed.), *The Chemistry of Organic Germanium, Tin and Lead Compounds*, 2003.

GERT

A procedure for the formulation and evaluation of systems using a network approach. Problem solving with the GERT (graphical evaluation and review technique) procedure utilizes the following steps:

1. Convert a qualitative description of a system or problem to a generalized network similar to the critical path method—PERT type of network.

2. Collect the data necessary to describe the functions ascribed to the branches of a network.

3. Combine the branch functions (the network components) into an equivalent function or functions which describe the network.

4. Convert the equivalent function or functions into performance measures for studying the system or solving the problem for which the network was created. These might include either the average or variance of the time or cost to complete the network.

5. Make inferences based on the performance measures developed in step 4.

Both analytic and simulation approaches have been used to perform step 4 of the procedure. GERTE was developed to analytically evaluate network models of linear systems through an adaptation of signal flow-graph theory. For nonlinear systems, involving complex logic and queuing situations, *Q*-GERT was developed. In *Q*-GERT, a simulation of the network is performed in order to obtain statistical estimates of the performance measures of interest.

GERT networks have been designed, developed, and used to analyze production lines, quality control in manufacturing systems, capacity of air terminal cargo facilities, equipment allocation in construction planning, planning and control of marketing research, and system reliability. In recent time, GERT has become an obscure method which has seen little use. *See* DECISION THEORY; PERT; SIMULATION.

A. Alan B. Pritsker

Bibliography. U. Brandes and T. Erlebach (eds.), *Network Analysis: Methodological Foundations*, 2005; A. A. B. Pritsker, *GERT: Graphical Evaluation and Review Technique*, RM-4973-NASA, 1966; A. A. B. Pritsker, *Modeling and Analysis Using Q-GERT Networks*, 2d ed., 1980.

Gestation period

In mammals, the interval between fertilization and birth. It covers the total period of development of the offspring, which consists of a preimplantation phase (from fertilization to implantation in the mother's womb), an embryonic phase (from implantation to the formation of recognizable organs), and a fetal phase (from organ formation to birth).

In some mammals, the gestation period is extended by a special process known as delayed implantation or embryonic diapause: the fertilized egg develops in the normal way to form a hollow ball of cells (the blastocyst) and development is then arrested. Attachment of the blastocyst to the inner wall of the womb (implantation) is usually the next step in other mammals, but in cases of delayed implantation the blastocyst remains free in the womb and does not become attached until a later stage. In some cases, the period of delay before implantation is fixed. One example is the European roe deer, in which delayed implantation seems to ensure an optimal interval between mating in late summer and early fall, and birth in the spring. In other cases, the period of delayed implantation is variable. For instance, in many marsupials (such as the kangaroo) it is typical for a mother to have offspring in the pouch and one or more blastocysts free in the uterus. If the pouch offspring fail to survive, a free blastocyst can implant rapidly, thus permitting the mother to give birth again rapidly. For this reason, in these marsupials and in other mammals with an optional delay the gestation period can vary (see **table**).

Gestation periods of selected mammals

Mammal	Days	Mammal	Days
Baboon	173	Horse	337
Camel	402	Human	267
Cat	63	Lion	109
Chimpanzee	235	Macaque	165
Cheetah	93	Marmoset	148
Cow	284	Mole	35
Dog	63	Mouse	19
Elephant		Opossum	13
(Indian)	626	Orangutan	250
Giraffe	456	Rat	21
Gorilla	259	Sheep	148
Guinea pig	67	Wolf	63
Hamster	20	Zebra	360
Hedgehog	33		

There is widespread confusion over the duration of the gestation period in humans because of the way in which it is defined medically. The time of ovulation, and hence the time of fertilization, is difficult to determine in humans, so for purely practical reasons doctors measure the duration of pregnancy as the interval between the last menstrual period and birth, which is typically about 40 weeks or 280 days. For comparison with other mammals, however, the true gestation period between fertilization and birth in humans is about 267 days.

The length of the gestation period in mammals depends primarily on body size and the state of development of the offspring at birth. Large-bodied mothers have big offspring that take longer to develop, and development is also prolonged for offspring that are born at an advanced stage of development. There is, in fact, a major distinction between different groups of mammals according to the state of development of the offspring at birth. In some mammal groups (for example, marsupials, most insectivores, carnivores, and most rodents) the mother gives birth to a litter of poorly developed offspring (altricial offspring) that are typically kept in some kind of nest for some time after birth until they have grown enough to move around independently. Such altricial offspring are usually hairless at birth, and the eyes and ears are sealed off by membranes.

Marsupials are superaltricial, as their offspring are extremely poorly developed at birth and usually crawl into a pouch of some kind to continue their development before emerging into the outside world. In other mammal groups (for example, hoofed mammals, primates, whales, and dolphins), mothers typically give birth to a single, well-developed offspring, and a nest is not usually required. Such precocial offspring usually have a thick coat of fur at birth, and the eyes and ears are typically open. This distinction between mammal groups probably has a long evolutionary history. Fossil specimens from the Eocene Period (some 50 million years ago) show that relatives of the modern horse already had a single, well-developed offspring.

Because gestation periods generally increase in a regular fashion with the body size of the mother, this trend must also be taken into account when making comparisons between mammal species. For instance, the gestation period of the elephant (660 days) is about $2^1/_2$ times as long as the true human gestation period (267 days), but this difference can be attributed almost entirely to the enormous difference in body size. Once body size effects and special cases have been excluded, it emerges that marsupials have by far the shortest gestation periods. Among placental mammals, altricial mammals have markedly shorter periods than precocial mammals. Indeed, for any given body size, a precocial mammal has a gestation period three to four times longer than an altricial mammal. The human gestation period of 267 days is, for example, almost three times longer than the gestation period of 93 days for the cheetah, which has a similar body weight but gives birth to altricial offspring. Compared to all other mammals, human beings are found to have one of the longest gestation periods relative to body size.

One remarkable feature of mammalian gestation periods is that they show very little variability within a species. After excluding exceptional cases, departures from the average usually lie in a range of no more than ±4%. This is one of the smallest degrees of variability found in any biological dimension. *See* FERTILIZATION (ANIMAL); PREGNANCY; REPRODUCTIVE SYSTEM. R. D. Martin

Geyser

A natural spring or fountain which discharges a column of water or steam into the air at more or less regular intervals. It may be regarded as a special type of spring. Perhaps the best-known area of geysers is in Yellowstone Park, Wyoming, where there are more than 100 active geysers and more than 3000 noneruptive hot springs. Other outstanding geysers are found in New Zealand and Iceland. The most famous geyser is probably Old Faithful (see **illus.**) in Yellowstone Park, which erupts about once an hour. Then for about 5 min the water spouts to a height of 100–150 ft (30–45 m). Other geysers are less regular, but some intermittently discharge water and steam to heights of 250 ft (75 m) or more.

Old Faithful, Yellowstone Park, Wyoming. (*National Park Service, U.S. Department of the Interior*)

The eruptive action of geysers is believed to result from the existence of very hot rock, the relic of a body of magma, not far below the surface. The neck of the geyser is usually an irregularly shaped tube partly filled with water which has seeped in from the surrounding rock. Far down the pipe the water is at a temperature much above the boiling point at the surface, because of the pressure of the column of water above it. Its temperature is constantly increasing, because of the volcanic heat source below. Eventually the superheated water changes into steam, lifting the column of water out of the hole. The water may overflow gently at first but, as the column of water becomes lighter, a large quantity of hot water may flash into steam, suddenly blowing the rest of the column out of the hole in a violent eruption. *See* SPRING (HYDROLOGY). Albert N. Sayre; Ray K. Linsley

Ghost image (optics)

A feature or shape at the focal plane of a camera or other optical instrument that is not present in an actual scene, or an unfocused duplicate image that is overlaid upon a desired image. Ghost images, or ghosts, are caused by reflections from the surfaces of lenses or windows. Each glass surface divides incoming light into two parts: a refracted part that passes through the surface, and a reflected part that is turned back. If the reflected light is turned back again by reflection from another glass surface or a mirror, it may travel to the focal plane to form a ghost image. Ghost images may appear as an out-of-focus blur or smudge, a sharp circle or polygon with the shape of the camera iris or other aperture, or a false image of an object within a scene. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

Antireflection coatings. The brightness of ghost images is greatly reduced by applying antireflection coatings to lenses and windows. Antireflection coatings are thin films of transparent material that reduce the reflectance of a glass surface. A common antireflection coating is a single layer of magnesium fluoride (MgF_2). Most glasses reflect a little more than 4% of the light from each surface. A single layer of magnesium fluoride with a thickness equal to one-quarter of the wavelength of light reduces the reflectance of glass to a little more than 1%. An even lower reflectance is obtained by applying two or more layers with properly chosen materials and thicknesses. Design of antireflection coatings is a specialized discipline in which an optimum mix of coating materials and layer thicknesses is developed with the help of computers and coating-design software. *See* INTERFERENCE FILTER.

Optical instrument design. Ghost images may be reduced or eliminated by properly designing the components of an optical instrument. The simplest way to eliminate ghosts is to use mirrors instead of lenses. When this is not possible, ghosts are controlled by dispersing them or moving them. A ghost is dispersed by adjusting the shape, location, or number of optical components so as to spread the ghost image over a broad area of the focal plane. As the size of the ghost increases, its brightness falls until it is no longer visible. A ghost image is moved by tilting the surfaces that cause the ghost. Tilting a surface changes the direction of the reflected light, which moves the ghost off the image plane. Tilting a single surface on a lens or window causes one edge of the part to be thicker than the other edge; this configuration is called a wedge. Ghosts caused by two reflections within a single lens element or window are not affected by tilting the entire component. Redirection of light from one tilted surface is undone by the tilt in the other surface.

Designing optical components that disperse and move ghosts is done with the help of computers and optical design software. Actions that remove ghost images often reduce image quality or increase the cost and complexity of an instrument. A computer calculates how changes to an instrument affect both the desired image of a scene and the brightness and shape of ghosts. A designer can then choose component shapes and materials that give the best performance.

Optical design software works by representing the propagation of light as lines, or rays, which lie along the path that light takes through an optical instrument; this technique is called ray tracing. At each lens surface, rays are split in two—one ray tracing the path of the transmitted light and the other ray tracing the path of reflected light. Numbers are assigned to each ray that represent the transmitted or reflected power. Rays that finally intersect the focal plane, along with their assigned power, are collected or binned together to calculate the size and brightness of each ghost. If the ghosts are too bright, changes that reduce them are made to the optical design. The effect of these changes on both image quality and ghost brightness is assessed, and adjustments are made in the design until an acceptable balance is achieved between the brightness of the ghost and other measures of the performance and cost of the instrument. Modern computers and software are capable of tracing many millions of rays, so accurate prediction of size, brightness, and cause of ghosts is a common part of instrument design. *See* GEOMETRICAL OPTICS.

Other sources. Reflections from lenses and windows are the most common causes of ghost images, but there are other sources. Optical components are often placed within a mechanical housing and then held in place by rings of metal. Bright sources, often outside the field of view of the instrument, illuminate the interior of the lens housing and retaining rings. Reflections from metal mounting surfaces may travel to the focal plane and form streaks or spots known as flares. Flares are also caused by reflection from the edges of lenses. Flares of this type are suppressed by applying black paint to the edge of an offending lens. Ideally, the index of refraction for the paint matches the index of refraction of the glass, so light that illuminates the edge of the lens passes easily into the paint and is absorbed. Diffuse scatter from housing walls often produces broad patches of illumination referred to as veiling glare. This kind of ghost is reduced by placing a thin metal annulus, called a vane, within the housing to block reflected and scattered light, or by machining fine grooves or threads in the walls of the housing to scatter light away from the focal plane.

Narcissus. Thermal infrared cameras work at wavelengths from 8 to 13 micrometers. At these long wavelengths, which cannot be seen by the eye, all warm objects produce infrared radiation. Infrared cameras that use cold detectors may exhibit a dark ghost around the center of their focal plane; this ghost is called narcissus. It is present if the detector sees a ghost image of itself reflected in one or more lenses of the camera. Because the detector is cold, it produces less infrared radiation than warm objects within the camera, so a ghost image of the detector is dark instead of bright. *See* INFRARED IMAGING DEVICES; INFRARED RADIATION. Gary L. Peterson

Bibliography. E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*, Wiley, New York, 1996; H. A. Macleod, *Thin-Film Optical Filters*, 3d ed., Institute of Physics Publishing, Bristol, U.K., and Philadelphia, 2001; W. J. Smith, *Modern Optical Engineering*, 3d ed., McGraw Hill, New York, 2000.

Giant nuclear resonances

Elementary modes of oscillation of the whole nucleus, closely related to the normal modes of oscillation of coupled mechanical systems. Giant nuclear resonances occur systematically in most, if not all, nuclei, with oscillation energies typically in the range of 10–30 MeV. Among the best-known examples is the giant electric dipole (E1) resonance, in which all the protons and all the neutrons oscillate with opposite phase, producing a large time-varying electric dipole moment which acts as an effective antenna for radiating gamma rays. *See* GAMMA RAYS.

Giant resonances are usually classified in terms of three characteristic quantum numbers: L, S, and T, where L is the orbital angular momentum, S is the (intrinsic) spin, and T is the isospin carried by the resonance oscillation. The number L is also the multipole order, with possible values $L = 0$ (monopole), $L = 1$ (dipole), $L = 2$ (quadrupole), $L = 3$ (octupole), and so on. The spin quantum number S is either 0 or 1. The $S = 0$ resonances are often called electric, and the $S = 1$ ones magnetic (EL or ML, where L is the multipole order), stemming from the fact that these giant resonances have strong decay modes involving the emission of either electric (for EL resonances) or magnetic (for ML resonances) multipole photons of the same multipole order as the resonance. A giant resonance with $S = 0$ corresponds to a purely spatial oscillation of the nuclear mass (or charge density), while one with $S = 1$ corresponds to a spin oscillation. The isospin quantum number T, which is also either 0 or 1, determines the relative behavior of neutrons versus protons; in a $T = 0$ or isoscalar giant resonance, the neutrons and protons oscillate in phase, whereas in a $T = 1$ or isovector resonance the neutrons and protons oscillate with opposite phase. Examples of well-established giant resonances are the E0, E1, and M1 isovector modes, and the E0 and E2 isoscalar modes. Other types of giant resonances also exist, notably the giant Gamow-Teller resonance. *See* MULTIPOLE RADIATION; NUCLEAR MOMENTS.

These resonances are called giant because of their great strength, 50–100% of the theoretical limit, concentrated in a compact energy region. The oscillation energy is characteristic of the type of giant resonance and usually varies smoothly with mass. It is determined by the restoring force and the nuclear mass; the force is due to the nuclear attraction between nucleons, the most important part being the component of the same multipole order as the giant resonance.

Width. The width of a giant resonance generally contains contributions from three different sources: decay, damping, and fragmentation. If the resonance has sufficient energy, as most do, then the decay width stems primarily from the emission of nucleons and, in some cases, composite particles. Damping width is associated with mixing (damping) of the giant resonance into the sea of more complicated, underlying states, a process analogous to the frictional damping of a classical oscillator.

In the absence of decay and damping width, a giant resonance ideally would occur at a single sharp energy in a spherical nucleus. The degree to which the strength is fragmented among several different energies is accounted for by the fragmentation width.

Giant electric dipole (E1) resonance. This, the oldest and best known of the nuclear giant resonances, is the dominant feature in reactions initiated by gamma rays (**Fig. 1**). The absorption of a gamma ray induces the giant E1 oscillation, which breaks up, in this case, by emitting neutrons. This resonance is also the dominant feature in the reverse process, in which gamma rays are produced by proton and neutron bombardments of nuclei. The resonance, which is isovector ($L = 1$, $S = 0$, $T = 1$), occurs at an energy $E_x \cong 32A^{-1/3} + 20.6A^{-1/6}$ MeV (where A is the mass number of the nucleus) in medium and heavy nuclei (for example, 16.5 MeV at $A = 100$), and somewhat lower in light nuclei. Its strength is essentially equal to the theoretical limit. In deformed nuclei the resonance splits into several overlapping components (two if the shape is axially symmetric). In the most strongly deformed nuclei the resonance is partially resolved into two separate components.

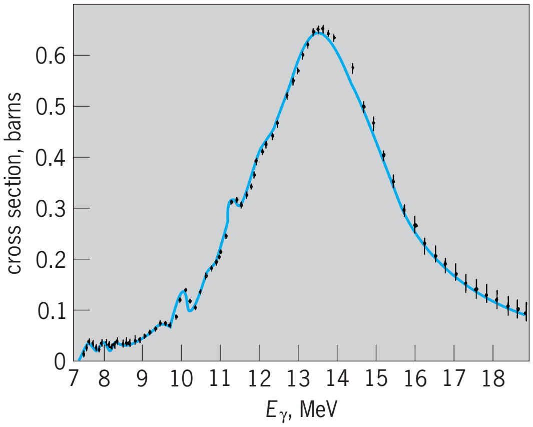


Fig. 1. **Cross section of neutrons produced by gamma-ray bombardment as a function of the gamma-ray energy E_γ. The large peak is due to the giant electric dipole resonance. 1 barn = 10^{-28} m^2. (*After A. Veyssiere et al., Photoneutron cross sections of ^{208}Pb and ^{197}Au, Nucl. Phys., A159:561–576, 1970*)**

Reaction selectivity. Because the different types of giant resonances often overlap in energy, particular resonances must be selectively excited in order to clearly delineate their properties. This is analogous to using the right driving force to excite particular normal modes of a coupled mechanical system. Inelastic alpha-particle scattering is very effective for studying isoscalar giant electric resonances. Both giant magnetic and electric resonances are studied by inelastic electron scattering, whereas charge changing reactions like the (*p, n*) reaction (neutron production by proton bombardment) are very useful for probing spin and isospin ($S = 1$; $T = 1$) modes.

Giant E0 and E2 resonances. The isoscalar giant E0 (electric monopole; $L = 0$, $S = 0$, $T = 0$) resonance lies at about $80A^{-1/3}$ MeV (17 MeV at $A = 100$), very close in energy to the giant E1 resonance, whereas the isoscalar giant E2 (electric quadrupole; $L = 2$, $S = 0$, $T = 0$) resonance lies somewhat lower, at about $65A^{-1/3}$ (14 MeV at $A = 100$). Both are strongly excited in forward-angle inelastic scattering of energetic alpha particles (**Fig. 2**). The angular dependence of the giant resonance excitation probability (cross section) is very different for the E0 and E2 resonances (Fig. 2), a feature which was essential in identifying the E0 mode.

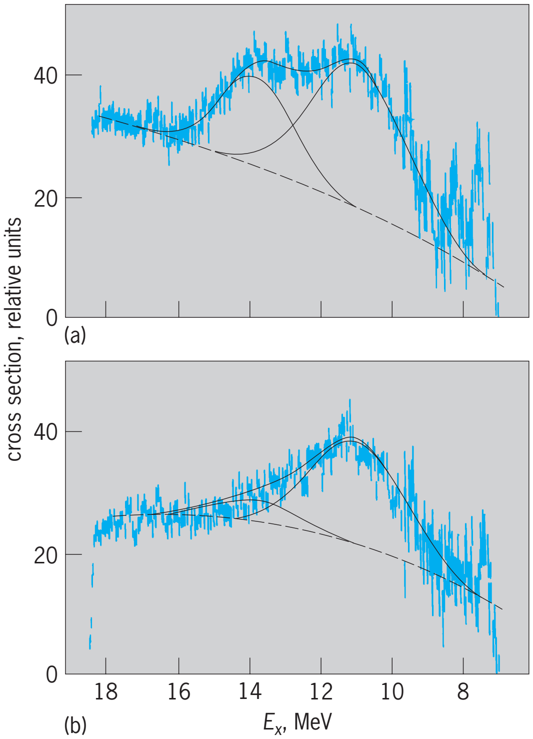


Fig. 2. Partial spectra of 120-MeV alpha particles scattered inelastically from lead-208 at angles of (*a*) 0° to 1.5° and (*b*) 1.5° to 3°. The broad peaks at excitation energies E_x near 11 and 14 MeV are the isoscalar giant quadrupole resonance and the isoscalar giant monopole resonance, respectively. (*After S. Brandenberg et al., Decay of the isoscalar giant monopole resonance in ^{208}Pb, Nucl. Phys., A466:29–69, 1987*)

The E0 resonances are spherically symmetric. The isoscalar E0 resonance is called the breathing mode, as the whole nucleus undergoes a purely radial oscillation, alternately expanding and contracting. The isoscalar E0 resonance energy is important in determining the nuclear compressibility.

The isovector E0 giant resonance ($L = 0$, $S = 0$, $T = 1$) is also a purely radial oscillation but one in which the protons oscillate against the neutrons. As a result, it lies at a higher energy than the isoscalar E0 resonance, at about $60A^{-1/6}$ MeV (28 MeV at $A = 100$). The isovector E2 giant resonance ($L = 2$, $S = 0$, $T = 1$) lies at around $130A^{-1/3}$ MeV (also 28 MeV at $A = 100$).

Giant Gamow-Teller resonance. In ordinary nuclear beta decay, a neutron inside a nucleus is transformed into a proton, and an electron and an antineutrino are produced. In one of the simplest types of beta decay, called Gamow-Teller decay, the transformed neutron is otherwise undisturbed, except that its spin may be reversed. As a result, the nucleus usually gains a small amount of energy. If beta decay involved a higher energy transfer to the nucleus, it would drive the giant Gamow-Teller resonance, which is a pure spin oscillation where the neutron spin and the proton spin oscillate out of phase ($L = 0$, $S = 1$, $T = 1$). A giant Gamow-Teller resonance is a strong feature in the (p, n) reaction in which neutrons emerge at 0° from nuclei struck by energetic protons (**Fig. 3**). This reaction substitutes a proton for a neutron in the nucleus via a spin-dependent interaction, in a manner analogous to beta decay but with a much larger energy transfer.

The properties of the giant Gamow-Teller resonance are important in certain problems in nuclear astrophysics. For example, the low-energy tail of the giant Gamow-Teller resonance in ^{37}Ar is important in determining the expected rate of the inverse beta decay reaction $\nu_e + {}^{37}\mathrm{Cl} \rightarrow {}^{37}\mathrm{Ar} + e^-$, which is used to detect neutrinos produced by nuclear reactions in the Sun. *See* RADIOACTIVITY; SOLAR NEUTRINOS.

Giant resonances in highly excited nuclei. Studies of the giant electric dipole resonance have been extended to highly excited hot nuclei. These studies provide unique information about the properties of such nuclei, in particular their shape. The shape sensitivity arises from the resonance splitting in a deformed nucleus, as described above. The size of the splitting gives the magnitude of the deformation, whereas the relative strength of the components determines the sense of the deformation: prolate (football-shaped) or oblate (doorknob-shaped).

The highly excited nuclei are formed by heavy-ion fusion, in which an energetic nucleus from an accelerator collides and fuses with a target nucleus. The large energy transferred to the excited nucleus is distributed between giant resonances and other, mostly random forms of energy analogous to heat, which is characterized by a nuclear temperature T.

These studies demonstrate the persistence of ground-state deformation in highly excited nuclei, with temperatures such that $kT \cong 1$ MeV (k = Boltzmann's constant). On the other hand, nuclei that are spherical at low temperature and spin undergo large thermal-shape fluctuations at temperatures $kT \geq 1$ MeV. Furthermore, hot nuclei that are spherical at low spin become deformed (most likely oblate-shaped) at higher spin. The same tendency is shown by a spherical liquid drop (or spherical water balloon) to become oblate when rotated.

Multiple giant resonances. The double giant E1 resonance is made up of two giant E1 resonances coexisting in the same nucleus. It is a so-called two-phonon giant resonance, or, in other words, a giant

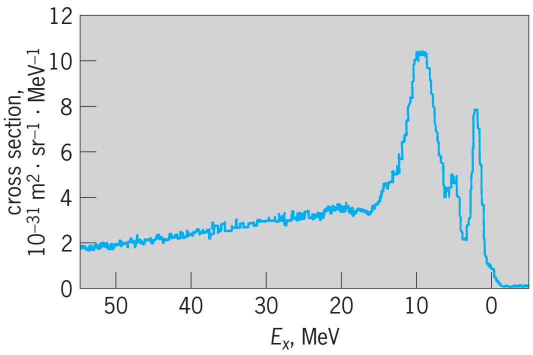


Fig. 3. Spectrum of neutrons emitted at 0° by bombarding zirconium-90 with 200-MeV protons. The broad peak at excitation energy E_x near 9 MeV is the giant Gamow-Teller resonance. (*After T. N. Taddeucci, Polarization transfer in (p, n) reactions, Can. Phys., 65:557–565, 1987*)

E1 resonance built upon a giant resonance. Double giant E1 resonances were first produced in (π^+, π^-) reactions, in which two neutrons inside the nucleus are transformed into protons and the resonance is excited. In subsequent studies, the double giant E1 resonance has been produced in distant collisions of heavy target nuclei with heavy projectiles moving at relativistic speeds. In such collisions, the repulsive electric force between the two positively charged nuclei has sufficiently high frequency Fourier components to induce a two-step process in which the normal (single) giant E1 resonance is first excited, and then, before the excited nucleus has a chance to decay, it is struck by a second so-called virtual photon, producing the double resonance. *See* FOURIER SERIES AND TRANSFORMS.

Giant resonances in supernovae explosions. Giant resonances play an important role in energetic nuclear reactions occurring in nature. Among the best examples are supernovae explosions. In a supernova, a massive star, having burned most of its nuclear fuel, undergoes gravitational collapse. During the collapse, nuclei in the core region produce neutrinos by a process called electron capture, which is a kind of inverse beta decay. The rate of these reactions, which cool the core and accelerate the collapse, depends on the properties of the giant Gamow-Teller resonance.

The stellar collapse continues until the stiffness of nuclear matter resists further compression, at which point a shock wave is created which rebounds out-ward and blows off the mantle of the star. The strength of the shock wave is directly related to the nuclear compressibility discussed above in the context of the giant isoscalar E0 resonance.

Left behind in the central region is a hot neutron star, which cools by emitting neutrinos. These higher-energy neutrinos travel outward and heat the nuclei in the mantle via inelastic scattering reactions which excite various giant resonances. This heating of the mantle, which can occur before the shock wave arrives, may contribute to the explosion. Certain elements found in nature may have been produced primarily as giant resonance decay products in these reactions. *See* GRAVITATIONAL COLLAPSE; NEUTRINO; NEUTRON STAR; NUCLEAR REACTION; NUCLEAR SPECTRA; NUCLEAR STRUCTURE; RESONANCE (QUANTUM MECHANICS); SUPERNOVA.

Kurt A. Snover

Bibliography. B. L. Berman and S. C. Fultz, Measurements of the giant dipole resonance with monoenergetic photons, *Rev. Mod. Phys.*, 47:713–761, 1975; Proceedings of the Gull Lake Conference of Giant Resonances, *Nucl. Phys.*, A569:1c–420c, 1994; Proceedings of the Workshop on Isovector Excitations in Nuclei, *Can. J. Phys.*, 65:535–698, 1987; K. A. Snover, Giant resonances in excited nuclei, *Annu. Rev. Nucl. Part. Sci.*, 36:545–603, 1986; J. Speth, *Electric and Magnetic Giant Resonances in Nuclei*, 1991; A. Van der Woude, Giant resonances, *Prog. Part. Nucl. Phys.*, 18:217–293, 1987; S. E. Woolsey et al., The neutrino process, *Astrophys. J.*, 356:272–301, 1990.

Giant star

An intermediate state in the evolution of a star in which it swells to enormous proportions before its death. During the longest and most stable phase of a star's life, the star, like the Sun, derives its energy from the thermonuclear fusion of hydrogen into helium deep in its dense, hot (10^7 K and up) core. It is then said to be on the main sequence. When the hydrogen fuel is gone and the central energy source is thereby exhausted, the core contracts and heats under the action of gravity, fresh hydrogen is ignited in a shell that surrounds the spent core, and the star becomes much more luminous, larger, and cooler at its surface. The lower surface temperature produces a redder color, hence the common term red giant. Stars like the Sun brighten by a factor of 1000 and grow in radius by a factor of 100 to about half the size of Earth's orbit (4.7×10^7 mi or 7.5×10^7 km).

There are actually three separate giant states. The first, described above, is terminated when the core temperature climbs so high (over 10^8 K) that the helium ignites and fuses into carbon. This event stabilizes the star; though the star is still a giant, it then contracts and dims as it quietly fuses its helium. When this helium is exhausted, the earlier behavior is repeated. The core contracts and is finally stabilized by electron degeneracy, becoming essentially a white dwarf. Helium then fuses to carbon in a shell around the core, and farther out hydrogen fuses to helium. The star then enters the asymptotic giant branch of the Hertzsprung-Russell diagram and swells to enormous proportions, perhaps two astronomical units (1.9×10^8 mi or 3×10^8 km), becoming even redder than before. It may pulsate and be seen as a long-period variable star, and loses much or most of its mass through a strong wind. *See* HERTZSPRUNG-RUSSELL DIAGRAM; STELLAR EVOLUTION.

James B. Kaler

Bibliography. I. Iben, Jr., Stellar evolution within and off the main sequence, *Annu. Rev. Astron. Astrophys.*, 5:571–626, 1967; I. Iben, Jr., and A. Renzinin, Asymptotic giant branch evolution and beyond, *Annu. Rev. Astron. Astrophys.*, 21:271–342, 1983; J. B. Kaler, *The Cambridge Encyclopedia of Stars*, 2006; J. B. Kaler, *Stars and Their Spectra: An Introduction to the Spectral Sequence*, 1997; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millennium*, 3d ed., 2007.

Giardiasis

A disease caused by the protozoan parasite *Giardia lamblia*, characterized by chronic diarrhea that usually lasts 1 or more weeks. The diarrhea may be accompanied by one or more of the following: abdominal cramps, bloating, flatulence, fatigue, or weight loss. The stools are malodorous and have a pale greasy appearance. Infection without symptoms is also common. As with most other protozoa inhabiting the intestinal tract, the life cycle of *Giardia* involves two stages: trophozoite and cyst.

Trophozoites stay in the upper small-intestinal tract, where they actively feed and reproduce. When the trophozoites pass down the bowel, they change into the inactive cyst stage by rounding up and developing a thick exterior wall, which protects the parasite after it is passed in the feces. People become infected either directly by hand-to-mouth transfer of cysts from feces of an infected individual or indirectly by drinking feces-contaminated water. After the cyst is swallowed, the trophozoite is liberated through the action of digestive enzymes and stomach acids, and becomes established in the small intestine.

Epidemiology. Giardiasis occurs worldwide. Surveys conducted in the United States have demonstrated *Giardia* infection rates ranging from 1 to 20%, depending on the geographic location and age of persons studied. In community epidemics caused by contaminated drinking water, as many as 50 to 70% of the residents have become infected. Outbreaks also occur among backpackers and campers who drink untreated stream water. Both human and animal (beaver) fecal contamination of stream water has been implicated as the source of *Giardia* cysts in waterborne outbreaks. *Giardia* species in dogs and possibly other animals are also considered infectious for humans.

Epidemics resulting from person-to-person transmission occur in day-care centers for preschool-age children and institutions for the mentally retarded. Infants and toddlers in day-care centers are more commonly infected than older children who have been toilet-trained.

Why some people become ill when infected with *G. lamblia* and others do not has not been fully explained. Host immunity undoubtedly plays a role, but the exact immune mechanisms involved have not been identified. A number of other factors, such as the number of *Giardia* cysts swallowed (dose), varying virulence between *Giardia* strains, and origin of the parasite (human or animal), have been postulated, but not proved, as having an influence on the clinical course of infection. *See* EPIDEMIC; EPIDEMIOLOGY.

Diagnosis. The diagnosis of *Giardia* infection is most commonly made by identifying the causative agent, *G. lamblia*, in the feces. It is also possible to identify the parasite in digestive juices or biopsy material taken from the small intestine. In individuals with watery diarrhea, trophozoites are most commonly found in stools, but a few cysts may also be present. After the acute stage has passed, stools are more often semiformed or formed, and contain the more hardy cyst form of the parasite. Because *Giardia* cysts are passed in the feces on an intermittent basis, a minimum of three stool specimens (one every other day) should be obtained and examined to minimize the chance of missing an infection. The parasites may be stained in iodine or by more permanent staining methods for purposes of differentiating them from other bowel-inhabiting protozoa.

Treatment. Three drugs are available in the United States for the treatment of giardiasis: quinacrine, metronidazole, and furazolidone. Quinacrine is considered the drug of choice for adults and older children. Although quinacrine is effective in young children, the drug frequently causes vomiting in this age group. Metronidazole gives cure rates similar to quinacrine, and is generally well tolerated by both adults and children. Furazolidone is also an effective drug; it is the only anti-*Giardia* preparation that is supplied in pediatric suspension.

Prevention. Epidemic giardiasis most commonly results from ingestion of contaminated drinking water. The long-term solution to municipal waterborne outbreaks requires improvement in, and widespread use of, water filtration equipment in the water treatment process. Many cities in the United States rely solely on chlorination to disinfect drinking water; however, the amount of chlorine used does not kill *Giardia* cysts. Backpackers and campers should not drink water directly from streams or lakes, no matter how clean the water looks. If stream water must be used for drinking, it should be boiled for 1 min to kill *Giardia* as well as other infectious organisms that might be present. Chemical disinfectants such as laundry bleach or tincture of iodine may also be used to disinfect water of uncertain purity. These products work well against most bacterial and viral organisms, but are not considered as reliable as heat in killing *Giardia*. If water is cloudy, it should be strained through a clean cloth into a container to remove any sediment or floating matter. Then the water should be treated with chemicals.

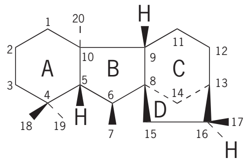
Person-to-person transmission of *Giardia* can be prevented by practicing good personal hygiene and maintaining a sanitary environment. Hand washing should be stressed, especially after using the toilet or handling soiled diapers of infants. Quick and thorough cleanup of fecal accidents at home or in institutions also reduces the risk of spreading *Giardia* to others. *See* MEDICAL PARASITOLOGY.

Dennis D. Juranek

Bibliography. *Health Information for International Travel*, HEW Publ. (CDC) 1999; *Waterborne Transmission of Giardiasis*, U.S. Environmental Protection Agency, EPA-600/9-79-001, June 1979.

Gibberellin

Any of the members of a family of higher-plant hormones characterized by the *ent*-gibberellane skeleton shown below. Some of these compounds have profound effects on many aspects of plant growth and development, which indicates an important regulatory role.



The discovery of gibberellins dates back to Japan in the 1920s, when plant pathologists were trying to understand, and ultimately control, the bakanae or

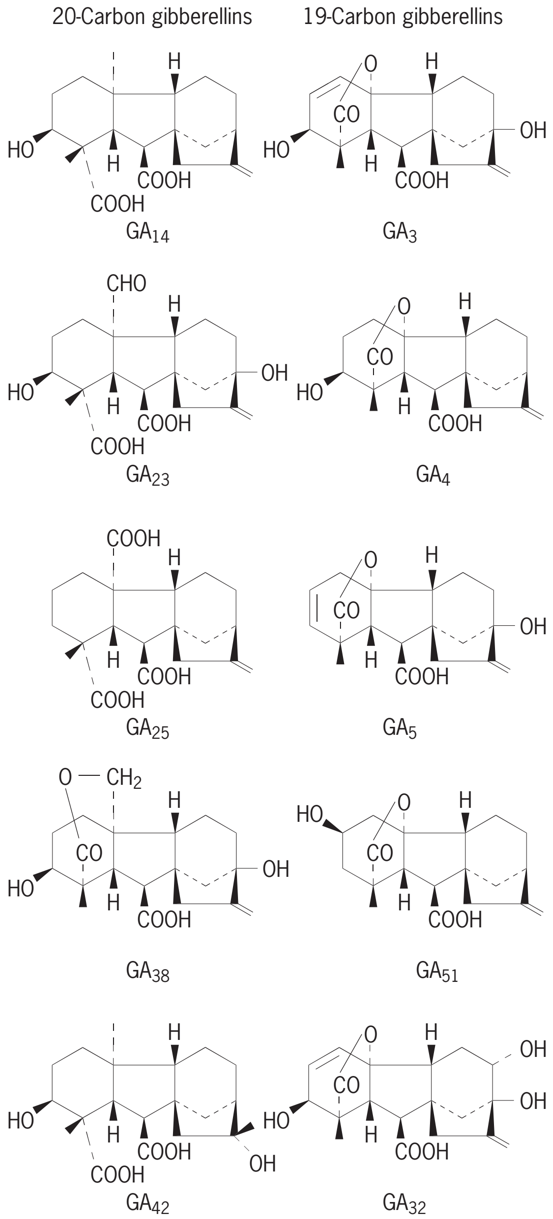


Fig. 1. Structural formulas of representative 20- and 19-carbon gibberellins.

"foolish seedling," disease of rice, which is caused by the fungus *Gibberella fujikuroi*. However, the chemical structure of the compound was not elucidated until 1954, when scientists reported the structure of gibberellic acid (GA_3) [**Fig. 1**]. *Gibberella fujikuroi* was subsequently shown to produce numerous other gibberellins, although GA_3 is by far the most abundant.

The discovery that gibberellic acid could restore normal growth in certain dwarf mutants of corn and peas led to the conclusion that gibberellic acid, or compounds structurally similar to it, is an endogenous growth regulator in higher plants. In 1958, the first higher-plant gibberellin, GA_1, was isolated from immature seeds of runner bean, and more than 70 gibberellins have subsequently been identified in all families of angiosperms and gymnosperms as well as several species of other fungi and ferns.

Nomenclature. All gibberellins are assigned a number and referred to as gibberellin $A_1 \ldots A_x$, or simply as $GA_1 \ldots GA_x$. In order for a compound to be given the GA designation, three criteria must be met: the structure and stereochemistry must be known; the compound must have the *ent*-gibberellane ring structure with a carboxyl group at carbon 7; and the compounds must occur naturally in plants or fungi. Compounds with unknown or incompletely known structures, such as those that have been detected in mosses, liverworts, algae, and bacteria, are referred to as gibberellin-like substances.

Chemistry. Gibberellins are tetracyclic diterpene acids with the *ent*-gibberellane skeleton. The two classes are based on the presence of the 19-carbon gibberellins and the 20-carbon gibberellins (Fig. 1). The 19-carbon gibberellins, which have lost carbon 20, contain a five-member lactone bridge that links carbons 4 and 10. The 19-carbon gibberellins, formed from 20-carbon gibberellins, are the biologically active forms. Gibberellins also vary according to the position and number of hydroxyl groups linked to the carbon atoms of the *ent*-gibberellane skeleton. Hydroxylation has a profound influence on biological activity. For example, compounds with a hydroxyl group on carbon 2 (such as GA_8, GA_{29}, and GA_{51}) are almost totally inactive and are generally thought to be the metabolic deactivation products of more active gibberellins. In contrast, gibberellins with a hydroxyl group on carbon 13 (GA_5 and GA_{20}) or carbon 3 (GA_4 and GA_7) possess much higher biological activity. The dihydroxylated gibberellins GA_1 and GA_3, which contain hydroxyl groups on carbons 3 and 13, are generally the most active.

Analytical methods. With the exceptions of *G. fujikuroi* and immature seeds of certain higher plant species, tissue levels of gibberellins are normally quite low, usually 1 to 100 nanograms per kilogram fresh weight. This complicates analysis of gibberellins, because extensive purification is often required before the actual analysis can be performed.

Analysis of gibberellin begins with extraction from the plant tissue using an organic solvent such as methanol or acetone. The organic extract is usually subjected to one or more procedures, such as solvent-solvent partitioning, that separate gibberellins from major plant constituents, including chlorophyll and other nonpolar pigments. The partially purified gibberellin extract is then subjected to chromatography, which affords additional purification and separates individual gibberellins from one another. In the past, this was achieved by paper or thin-layer chromatography; however, the use of high-performance liquid chromatography has largely replaced these two techniques. *See* CHROMATOGRAPHY; LIQUID CHROMATOGRAPHY.

The final step in gibberellin analysis is detection. Early work relied on gibberellin-specific bioassays that exploited the growth response of seedlings to applied gibberellins. Dwarf genotypes of rice, corn, and peas that have reduced levels of endogenous gibberellins have been most useful because of their sensitivity. In all these systems, the amount of growth elicited is proportional to the amount of gibberellin applied. Another widely used bioassay is based on the promotion of starch hydrolysis in de-embryonated cereal seeds: the amount of starch hydrolysis is proportional to the concentrations of applied gibberellin. In practice, the response of the bioassay to several known concentrations of a gibberellin—which is usually GA_3, because of commercial availability—is determined and plotted graphically to construct a standard curve. The amount of gibberellin in a plant tissue extract can then be estimated from the standard curve by interpolation.

Bioassays provide a highly specific, simple, and inexpensive method for detection of gibberellins, but they have several inherent disadvantages. First, the degree of the response is a function of the logarithm of the concentration, so that a tenfold increase in the concentration of applied gibberellin only doubles the response, making it difficult to accurately measure small changes in GA levels. Second, bioassays give little structural information about the gibberellin, so that although the amount of GA-like substances present in an extract can be estimated, little can be deduced about exactly which gibberellin is being measured. Finally, and most important, not all gibberellins elicit the same response in a given bioassay. Some, such as GA_1 and GA_3, are highly potent in all bioassays, whereas gibberellins with hydroxyl groups on carbon 2 (for example, GA_8, GA_{29}, and GA_{51}) are almost totally inactive. *See* BIOASSAY.

A combination of gas chromatography and mass spectrometry overcomes many of the disadvantages of bioassays because such a technique provides definitive structural as well as quantitative information about the compounds being analyzed. A sample is injected into a gas chromatograph, which separates the components in the mixture. The gaseous component is fed into the mass spectrometer, where it is bombarded with a stream of electrons. As a result, the molecules break apart and form a series of charged fragments. Each compound has a unique pattern of fragment size and quantity, and so an unknown compound can be identified when its mass spectrum matches that of a reference compound. The mass spectrometer can also measure gibberellin levels, because the number of ions produced during fragmentation is proportional to the amount of the compound injected. *See* GAS CHROMATOGRAPHY; MASS SPECTROMETRY.

Biosynthesis. Gibberellins are a small subset of the terpene family of compounds. As with other terpenoids, gibberellin biosynthesis begins with mevalonic acid (**Fig. 2**). The *ent*-gibberellane skeleton is built by successive combinations of 5-carbon isoprene units. The 20-carbon diterpene geranylgeranyl pyrophosphate, with four isoprene units, is cyclized to form kaurene, which is a tetracyclic compound that resembles the *ent*-gibberellane skeleton except that ring B contains six carbons instead of five, as in gibberellins (Fig. 2). The conversion of kaurene to a gibberellin starts with the sequential oxidation of carbon 19, followed by hydroxylation at carbon 7 to form *ent*-7-α-hydroxykaurenoic acid (Fig. 2). Contraction of ring B by the extrusion of carbon 7 leads to the formation of GA_{12}-aldehyde. All gibberellins are subsequently derived from this compound by three major metabolic activities: (1) the oxidation of the carbon-7 aldehyde to a carboxyl group; (2) the formation of 19-carbon gibberellins by the sequential oxidation of carbon 20 to an aldehyde, followed by its removal and subsequent formation of a lactone bridge from carbon 4 to carbon 10; and (3) hydroxylation of one or more carbons contained in the *ent*-gibberellane skeleton. This process can occur in either 19- or 20-carbon gibberellins. In corn and peas, the formation of GA_{12}-aldehyde is followed by early hydroxylation at carbon 13 (GA_{53}); converson to GA_{20}, a 19-carbon gibberellin; and hydroxylation at carbon 3 (GA_1), and then at carbon 2 (GA_8). Alternatively, GA_{20} can be hydroxylated once at carbon 2 to form GA_{29} (**Fig. 3**). Hydroxylation at carbon 2 serves as a deactivation process. *See* TERPENE.

Glucose conjugates of gibberellins have been detected in many plant tissues, especially in seeds. Linkage to glucose takes place through either a carboxyl group (glucose ester) or an ether at a hydroxyl group (glucoside). Conjugation of gibberellins to glucose serves as another deactivation process in addition to hydroxylation at carbon 2. The conjugates may also serve as storage forms of gibberellin. *See* CONJUGATION AND HYPERCONJUGATION.

Gibberellin biosynthesis and metabolism takes place in three stages, based on characteristics of the component enzymes and properties of their substrates. First, soluble cytoplasmic enzymes convert mevalonic acid to kaurene. Most of the enzymes are common to the biosynthesis of other terpenes, and their substrates are highly soluble in water, because many contain a pyrophosphate group. The second stage begins with kaurene and ends with the formation of GA_{12}. Stage II oxidative enzymes are microsomal, and they require molecular oxygen (O_2) and the reduced form of nicotinamide adenine dinucleotide phosphate (NADPH); they are mixed-function oxidases involving the participation of cytochrome P-450. In addition, the substrates are highly nonpolar and not readily soluble in water. Stage III reactions include the conversion of 20-carbon gibberellins to 19-carbon gibberellins and all hydroxylations. These enzymes are soluble oxidases that contain iron and require NADPH and α-ketoglutarate as cofactors. *See* CYTOCHROME.

Of all gibberellins native to a given species, only one is believed responsible for biological activity, the others being either precursors to or deactivation products of that one gibberellin. It has been shown that GA_1 controls stem growth in both corn and peas and may also be important for biological action in most other species.

Determination of the gibberellin biosynthesis pathway has practical implications, such as the control of the stature of plants. Synthetic plant-growth regulators called growth retardants are used commercially to regulate plant growth and many of these compounds reduce endogenous gibberellin levels by

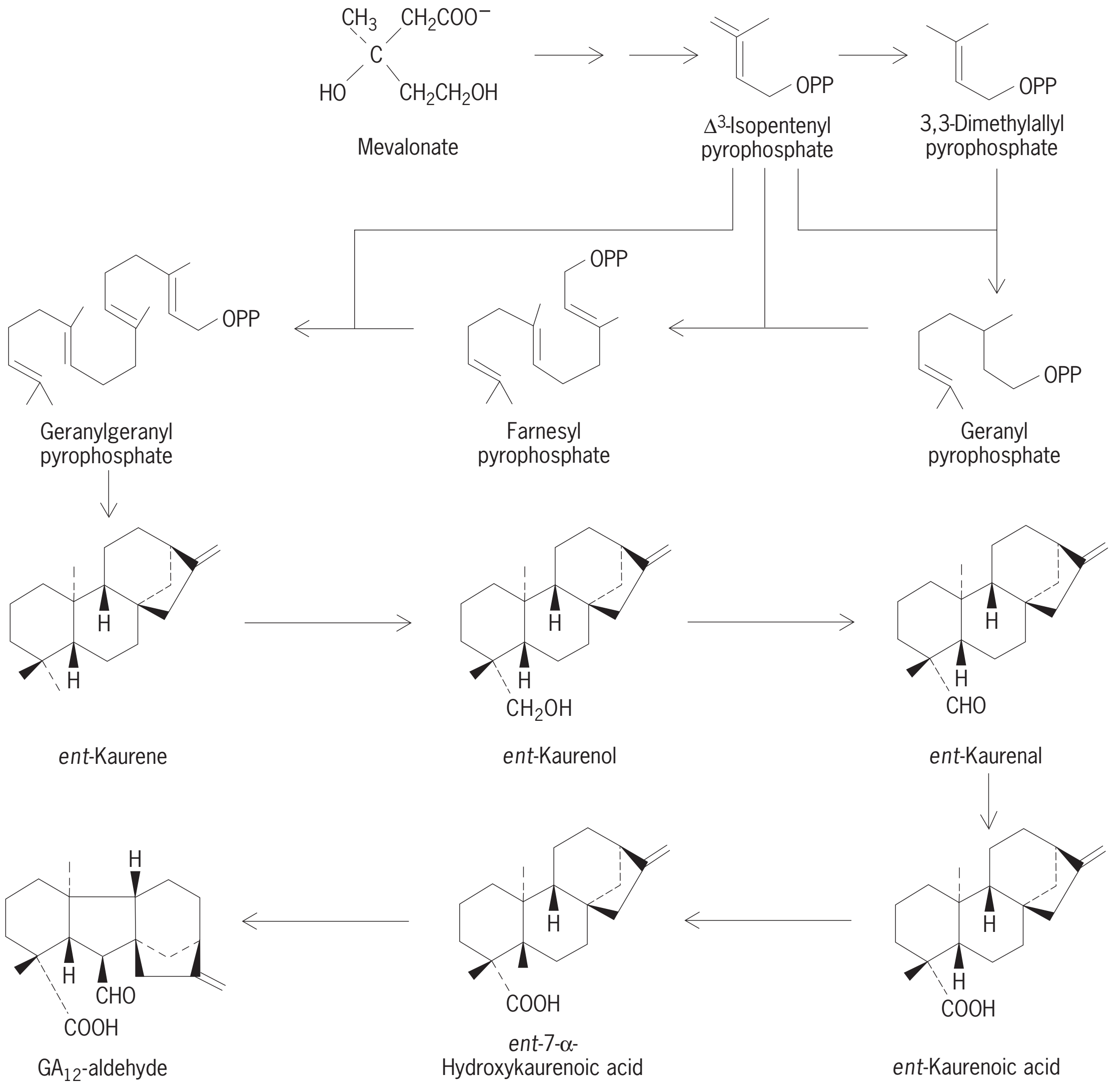


Fig. 2. Biosynthetic pathway of gibberellins from mevalonic acid in higher plants and the fungus *Gibberella fujikuroi*.

inhibiting gibberellin biosynthesis. Some growth retardants inhibit the cyclization of geranylgeranyl pyrophosphate to kaurene; others block the oxidation of kaurene to kaurenoic acid.

Control of growth and development. The involvement of gibberellins in specific aspects of plant growth and development can be inferred from several lines of experimentation.

Stem elongation. Probably the best-defined role for gibberellins in regulating the developmental processes in higher plants is stem growth. The absolute need for gibberellins in this process has been demonstrated by restoration of normal growth with gibberellin application to gibberellin-deficient dwarf mutants of corn, rice, peas, and other species. Rosette plants, which represent a special case of dwarfism, grow as dwarfs until they receive some inductive environmental stimulus such as a change in day length or temperature; gibberellins can often substitute for the stimulus. In the case of spinach, which requires long days to initiate stem growth, the inductive stimulus dramatically increases the level of certain gibberellins through higher activity of specific enzymes in the gibberellin biosynthetic pathway. The cellular basis for gibberellin-induced stem growth can be either an increase in the length of pith cells in the stem, as appears to be the case in lettuce hypocotyls, or primarily the production of a greater number of cells, as in many rosette plants following an inductive stimulus. *See* PHOTOPERIODISM.

Seed dormancy and germination. Freshly shed seeds from many species are often unable to germinate under ideal conditions; these seeds are said to be dormant. Depending on the species, overcoming dormancy requires a period of dry storage (afterripening) or an environmental stimulus such as light or low temperature. Applied gibberellins can often promote germination of dormant seeds, a capability suggesting that gibberellins are involved in the process of breaking dormancy.

Gibberellins are intimately involved in other aspects of seed germination as well. In the early stages of germination, the stored reserves that nourish the young seedling are mobilized until its photosynthetic

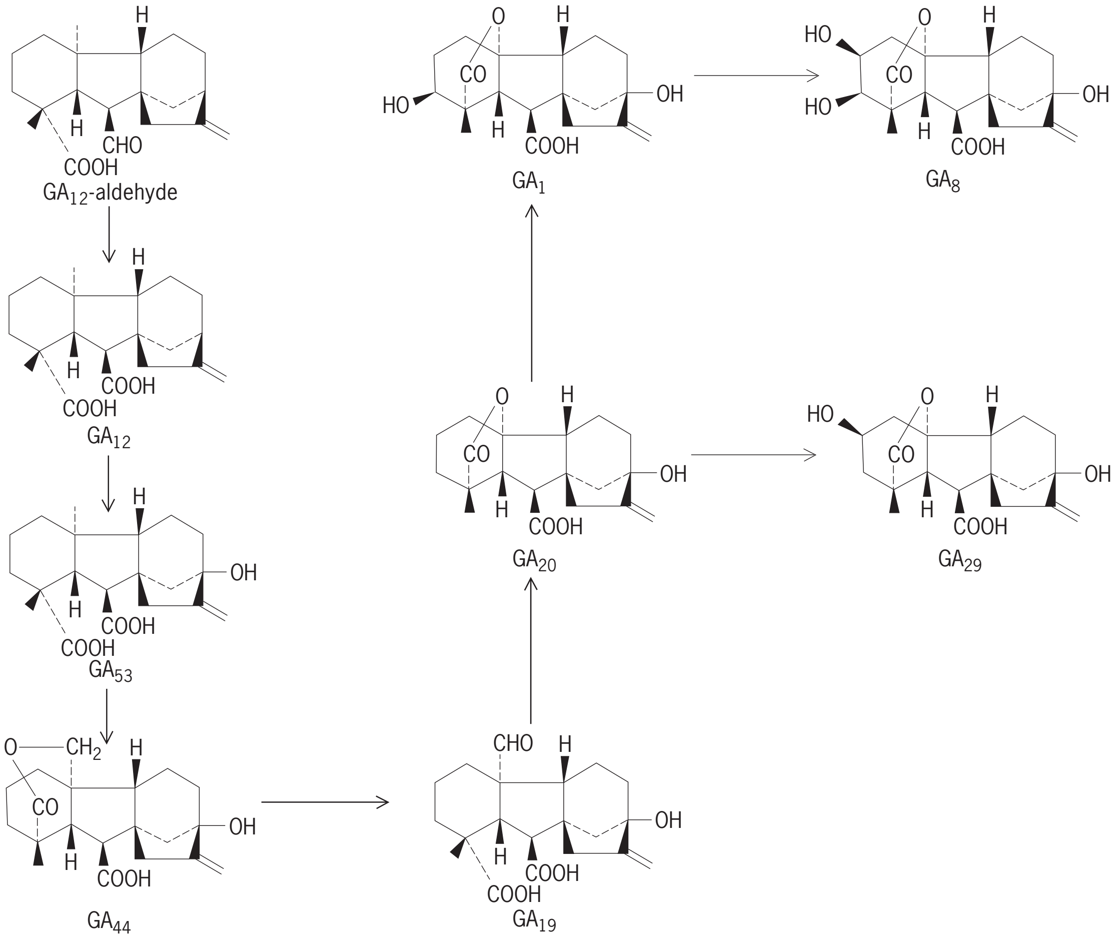


Fig. 3. Predominant pathway of gibberellin metabolism in corn and peas.

apparatus develops sufficiently. In the case of cereal grains, the breakdown of starch to glucose in the endosperm begins a few hours after the imbibition of water begins. Gibberellin in the seed embryo is believed to signal starch hydrolysis following action of the enzyme α-amylase, which is synthesized and released by aleurone cells that envelop the endosperm. Starch hydrolysis does not occur in de-embryonated seeds. Studies of molecular aspects of this process indicate that gibberellin causes increased transcription of the gene coding for the α-amylase enzyme, which results in higher levels of its messenger ribonucleic acid. *See* DORMANCY; RIBONUCLEIC ACID (RNA); SEED.

Flowering. Applied gibberellins promote or induce flowering in plants that require either cold or long days for flower induction. Gibberellin is probably not the flowering hormone or floral stimulus, because the floral stimulus appears to be identical or similar in all response types.

The sexuality of imperfect flowers (flowers with only male or female parts) is genetically determined, although environmental factors such as day length or temperature can be overriding factors. The application of gibberellins often modifies sex expression, usually causing an increase in the number of male flowers; in corn, however, feminization occurs. Gibberellin-deficient dwarf varieties of corn have male flowers on the ear (female inflorescence) indicating a natural role for gibberellins in sex expression. *See* FLOWER; PLANT GROWTH.

Commercial uses. Although gibberellins have limited use in agriculture compared with other agricultural chemicals such as herbicides, several important applications have been developed, including the production of seedless grapes. Application of gibberellin at bloom results in increased berry size and reduced berry rotting. Gibberellins are also used to increase barley malt yields for brewing and to reduce the time necessary for the malting process to reach completion. *See* MALT BEVERAGE.

Gibberellins have found significant applications in plant breeding. Many conifers do not flower until they are at least 10 years old, but the application of gibberellins can bring on cone production in juvenile plants. Thus, practical genetic improvement programs can be hastened by shortening the juvenile periods. In biennial vegetables such as carrots or brussels sprouts, seed can be produced in one season instead of the normal two following applications of gibberellin.

Other uses for gibberellin in agriculture include reduction of rind discoloration in citrus fruits, increased yield in sugarcane, stimulation of fruit set in fruit trees, and increased petiole growth in celery. *See* PLANT HORMONES.

James D. Metzger

Bibliography. A. Crozier (ed.), *The Biochemistry and Physiology of Gibberellins*, 1983; P. J. Davies (ed.), *Plant Hormones and Their Role in Plant Growth and Development*, 1987; J. E. Graebe, Gibberellin biosynthesis and control, *Annu. Rev. Plant Physiol.*, 38:419–465, 1987; D. S. Letham, P. B. Goodwin, and T. J. V. Higgins (eds.), *Phytohormones and Related Compounds: A Comprehensive Treatise*, 1978.

Ginger

An important spice or condiment; also the plant from which it is obtained, *Zingiber officinale*, of the ginger family (Zingiberaceae). The plant is a native of southeastern Asia. It is an erect perennial herb (see **illus.**) having thick, scaly, branched rhizomes which contain starch, gums, an oleoresin (gingerin) responsible for the pungent taste, and an essential oil which imparts the aroma. The rhizomes, dug up after the aerial parts have withered, are treated in different ways to produce green ginger or dried ginger. Ginger is used in medicine, in culinary preparations (soups, curries, puddings, pickles, gingerbread, and cookies), and for flavoring beverages such as ginger ale and ginger beer. The plant is grown in China, Japan, Sierra Leone, Jamaica, Queensland, and Indonesia. *See* SPICE AND FLAVORING; ZINGIBERALES.

Ginger (*Zingiber officinale*). (*USDA*)

Perry D. Strausbaugh

Ginkgoales

An order of gymnosperms in the class Ginkgoopsida (Pinophyta) with only one extant species, *Ginkgo biloba* (the maidenhair tree). Leaves identified as *Ginkgo* appeared first in the Upper Triassic, but *Ginkgo*-like leaves have been discovered as early as the Permian. During the Jurassic and Cretaceous periods, *Ginkgo* was a large taxon containing many species and having a circumpolar distribution in the Northern Hemisphere. The group diminished in size during the Early Cretaceous and into the Tertiary periods. It is probable that *Ginkgo*, considered sacred by the Chinese, would have become extinct had it not been cultivated in Chinese temple gardens, where the ginkgos became magnificent, old specimens several hundred feet tall.

Although *Ginkgo* is superficially different from other gymnosperms, details of seed morphology and reproduction clarify its nature as a gymnosperm. Young trees are excurrent; that is, they have a straight trunk from which lateral branches diverge (**Fig. 1**), but with age their excurrent habit often becomes obscured by the development of several major branches. The plant produces both long and short (spur) shoots. Long shoots, which result from extensive internodal growth, have alternately arranged leaves. Short shoots, characterized by almost no internodal growth, develop in the axils of leaves on long shoots and bear terminal clusters of leaves. The fan-shaped leaves, with dichotomous venation, are highly distinctive and may have a deep, apical notch (**Fig. 2***a*). Those borne on short shoots often lack the apical notch.

Ginkgo is dioecious (male cones and ovules are borne on different trees). Male cones, produced in the axils of leaves on short shoots, consist of an axis bearing helically arranged stamenlike structures, each of which terminates, usually, in a pair of microsporangia in which pollen is produced (Fig. 2*b*). Ovules commonly occur in terminal pairs on slender

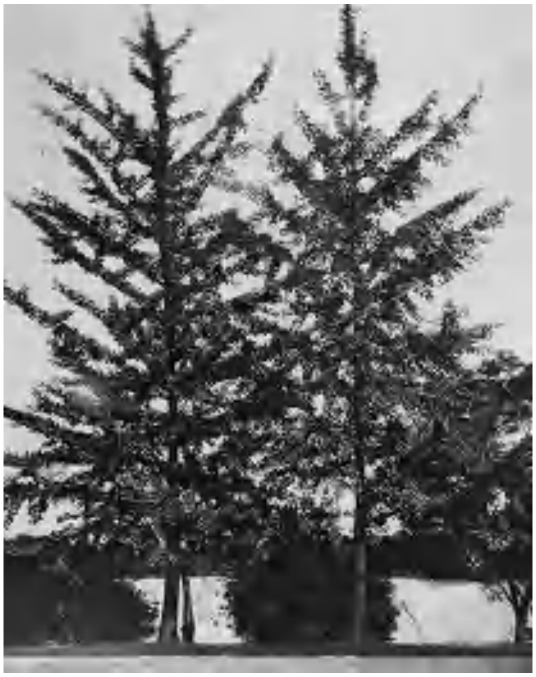


Fig. 1. Two maidenhair trees (*Ginkgo biloba*).

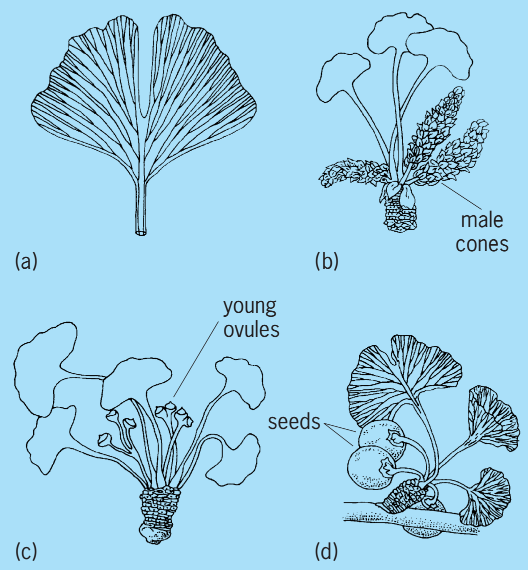


Fig. 2. Essential features of *Ginkgo biloba*. (*a*) Leaf showing deep apical notch and open, dichotomous venation of blade. (*b*) Short (spur) shoot with mature male cones and leaves lacking an apical notch. (*c*) Short shoot bearing young ovules with conspicuous basal collars. (*d*) Short shoot with mature seeds. (*After G. M. Smith et al., A Textbook of General Botany, 5th ed., Macmillan, 1953*)

stalks borne among the leaves of short shoots. Each ovule is enclosed basally by a collar (Fig. 2*c*). As the ovule matures, the outer integument becomes thick and fleshy, giving the seed the appearance of a fruit (Fig. 2*d*). Upon disintegration, the fleshy integument releases an unpleasant odor.

Ginkgo is used commonly as a shade tree in cities around the world because of its beauty and its resistance to disease and the effects of automotive pollution.

Several extinct genera are thought to be closely related to *Ginkgo*. Among these are *Trichopitys* and *Sphenobaiera*, which first appear in the early Permian, and *Baeira*, *Arctobaiera*, and *Eretmophyllum* from the Mesozoic. The reproductive structures of these ginkgophytes are unknown or not well understood. While their leaf form and stomatal characteristics resemble those of *Ginkgo*, it is not clear whether they should be included in Ginkgoales or assigned to one or more separate orders. *See* EMBRYOBIONTA; GINKGOOPSIDA; PINOPHYTA. Charles B. Beck

Bibliography. H. C. Bold, C. J. Alexopolous, and T. Delevoryas, *Morphology of Plants and Fungi*, 5th ed., 1990; R. F. Scagel, *Plants: An Evolutionary Survey*, 1984; W. N. Stewart and G. W. Rothinell, *Paleobotany and the Evolution of Plants*, 2d ed., 1993; T. N. Taylor and E. L. Taylor, *Biology and Evolution of Fossil Plants*, 1993.

Ginkgoopsida

A class of largely extinct gymnosperms (Pinophyta). Included orders are Calamopityales, Callistophytales, Arberiales, Peltaspermales, Ginkgoales, Leptostrobales, Caytoniales, Pentoxylales, and Ephedrales. The most ancient taxa, Calamopityales and Callistophytales, lived during the Carboniferous; the Arberiales, from late Carboniferous into the Triassic; the Peltaspermales, from Permian through Jurassic; the Leptostrobales and Caytoniales, from Triassic into the Cretaceous; and Ginkgoales, predominantly, from Triassic into the Cretaceous periods, with one species, *Ginkgo biloba*, persisting to the present. Ephedrales is the only largely extant group, but has a pollen fossil record beginning in the Upper Triassic. *See* CAYTONIALES; EPHEDRALES; GINKGOALES.

These taxa, divergent in many characteristics, are unified by the presence in all of platyspermic (bilaterally symmetrical) seeds lacking cupules. Presumed seeds of the most primitive order (Calamopityales) are of the *Lyrasperma* type, where the integument is fused with the nucellus over most of its length, and the nucellus, uncovered at its apex, terminates in a flasklike, pollen-catching organ called a salpinx. In all other orders, the seed is of the *Callospermarion* type, in which the integument, largely free from the nucellus, covers it except for a small opening at the apex, the micropyle. In some taxa, the seed may be secondarily platyspermic (derived from a radially symmetrical ancestor). In the most primitive, and some other, taxa of ginkgoopsids, the seeds and microsporangia are thought to have been borne on pinnately (featherlike) branched fertile structures. During the evolution of more advanced taxa, the individual microsporangia aggregated, fused to form synangia, and shifted onto leaves, causing additional changes to the microsporophylls and phyllosperms (seed-bearing leaves). In particular, the phyllosperms became greatly modified, often into peltate structures (as in advanced Peltaspermales) and stemlike structures (as in *Ginkgo*). *See* PINOPHYTA; PLANT KINGDOM. Charles B. Beck

Bibliography. C. B. Beck (ed.), *Origin and Evolution of Gymnosperms*, 1988; K. U. Kramer and P. S. Green (eds.), *The Families and Genera of Vascular Plants: Pteridophytes and Gymnosperms*, vol. 1, 1990.

Ginseng

The common name of the genus *Panax*, a group of perennial herbs of the aralia family (Araliaceae), native to the woodlands of the North Temperate Zone. *Panax schinseng* of Manchuria, extensively cultivated, was in such demand among the Chinese that the supply became insufficient. Then *P. quinquefolius* of eastern North America was discovered, and soon it was being exported to China in large quantities (see **illus.**). The price paid for the dried roots was so high that in a relatively short time the collectors nearly exterminated the plants. The Chinese used ginseng as a general panacea for many

***Panax quinquefolius*, ginseng, showing shoot and base.**

ills, but there is no evidence that the drug has therapeutic value. *See* APIALES.

Perry D. Strausbaugh; Earl L. Core

Giraffe

Member of the family Giraffidae represented by a single species, *Giraffa camelopardalis*. The only other living species in this family is the okapi (*Okapia johnstoni*); many fossil species are known. The giraffe occurs in the savanna regions of tropical Africa and the okapi ranges through the forested areas of the Congo. Both species are ruminants and belong to the mammalian order Artiodactyla (even-toed ungulates).

Giraffe (*Giraffa camelopardalis*). (*Photo by Arthur J. Emmrich; © 1999 California Academy of Sciences*)

The giraffe is the tallest of all mammals and may reach a height of 18 ft or 5.5 m (see **illustration**). The neck is long because of the extreme elongation of the neck vertebrae rather than an increase in their number. Giraffes are browsing animals, feeding mainly on acacia tree leaves. They are well adapted to this mode of feeding, having long lips and a prehensile tongue that can extend up to 20 in. (50 cm). These features allow them to pluck leaves from the trees and to avoid the thorns. There are two prominent horns on the forehead which are bony outgrowths covered by skin, and there is a short mane along the back of the neck. While these animals may weigh 1 ton (0.9 metric ton), they are agile and can travel at a good rate of speed. The senses of sight, hearing, and smell are well developed, and danger can be sensed at considerable distances. The giraffe lives in small herds with many females and usually one mature and several immature males. Old males are excluded and lead a life of isolation. Giraffes are social animals and may be seen with zebra, ostrich, and gnu. Gestation for the giraffe lasts about 15 months, and a single young is born. The young is about 6 ft (2 m) tall at the time of birth.

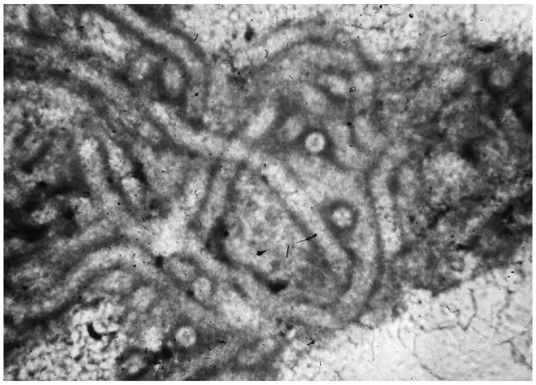
The okapi was not discovered until about 1900. It is not a common animal and lives in inaccessible areas of the eastern Congo. It is cryptically colored, having a hazel coat and striped hindquarters, and blends into its environment. The head shape, the lips, the tongue, and the horns of the male are the same as the giraffe's, but the neck is not elongate. The okapi closely resembles the extinct *Palaeotragus*, a ruminant that preceded the giraffe and occurred in Greece during the Miocene. The okapi is a nocturnal animal that lives singly or in pairs. It is a browser and quite dependent upon water. A single young is born after a gestation period of about 14 months. These animals breed readily in captivity, are rather shy, and become quite docile. *See* ARTIODACTYLA.

Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Girvanella

A genus of fossil algae. *Girvanella* is characterized by flexuous, tubular filaments of uniform diameter, composed of thick, calcareous walls (see **illustration**). External diameters average between 10 and 30 micrometers, although specimens less than 10 μm and up to about 100 μm have been identified as *Girvanella.* Filaments may occur free (unattached), but usually occur in groups, twisted together to form nodules and encrusting masses on



***Girvanella* in a thin section of Cambrian limestone. Tube diameter is about 20 m.**

various objects. The genus is intergrown with encrusting foraminifers in some Paleozoic limestones.

Girvanella is now generally placed in the blue-green algae (Cyanophyta), although in the past it has also been described variously as a foraminifer, sponge, and green algae. This genus is interpreted to be the calcified sheath of a variety of filamentous blue-green algae, similar to several living types.

Girvanella is a very common fossil, with a worldwide distribution. Occurring mainly in marine rocks, it has been reported from the Cambrian to Cretaceous. The apparent absence of *Girvanella* in rocks younger than Cretaceous age has not been satisfactorily explained. *See* ALGAE. John L. Wray

Bibliography. R. Riding, *Girvanella* and other algae as depth indicators, *Lethaia*, 8:173–179, 1975; J. L. Wray, *Calcareous Algae*, 1977.

Glacial geology and landforms

The scientific study of the processes and impacts of ice sheets, valley glaciers, and other ice masses on the Earth's surface, both on land and in ocean basins. The processes include understanding how ice masses move, erode, transport, and deposit sediment. The impacts on glaciated landscapes are enormous in terms of topographic change and floral and faunal modification. In those areas peripheral to glaciated areas, drainage patterns are altered, and climatic, vegetation, and soil conditions are severely changed. In addition, glacial geology involves studying the causes of glaciation, the chronology of glaciation in geologic time (the retreat and advance of ice masses at all scales), glacial stratigraphy, sea-level change, and how glaciations affect oceans, climate, flora, fauna, and human society globally. Closely allied to glacial geology are studies into the physics of ice masses (glaciology), global climatology (paleoclimatology), and paleoenvironmental reconstructions (paleoecology). *See* GLACIOLOGY; PALEOCLIMATOLOGY; PALEOECOLOGY.

Glacial sediments, both lithified and unlithified, are found on every continent, and the present oceanic basins of the Earth are covered in great thicknesses with glacial sediments. At the maximum of the past Pleistocene glaciation, sea level was at least 120 m (390 ft) below the present level. As a result, temporary land bridges appeared in the immediate postglacial period between Siberia and Alaska and between South Asia and the Indonesian Archipelago, and large parts of the North Sea in Europe were dry land. Such land bridges acted as corridors through which flora, fauna, and humans colonized North America, northwestern Europe, and parts of Australasia.

The impact of glacial geology on society is enormous, especially in the midlatitudes and poleward in both the Northern and Southern hemispheres. The soils, ground water, and construction are affected by glacial sediments and glaciated terrains. Cities such as Boston, Massachusetts, have been built on glacial landscapes such that urban planning and transportation systems must adapt to the topography. The siting of dams, the intricacies of ground-water pathways, and the location and discovery of minerals are tied to glaciations and problems in glaciated terrains. Of the total freshwater on Earth, 98% is held in ice sheets and other ice masses. (Freshwater is <2.5% of total water resources on Earth.) The impact of glaciation in the United States on its population is shown in **Fig. 1**, where considering land area in proportion to population, it can be seen that at least 60% of the population lives and works in areas once covered by the last glaciation (the Wisconsinan). In Canada, the comparative figure would be closer to 99% of the population, and in western Europe at least 50–60% of the population inhabits glaciated terrain.

Glacial History

Today, approximately 10% of the Earth's surface is covered by ice masses. The largest ice masses are the Greenland Ice Sheet and the West and East Antarctic Ice Sheets. The remaining ice masses are various mountain glaciers, small ice caps and fields, and ice shelves attached to the fringes of the Antarctic and the Canadian Arctic Archipelago (**Fig. 2***a*). Present ice-sheet thickness varies from around 2500 m (8200 ft) in central Greenland to over 4600 m (15,000 ft) in Antarctica. Valley glaciers and smaller ice masses are considerably thinner. During the maximum extent of the last glaciation in North America (Late Wisconsinan, Quaternary Period), the thickness of the Laurentide Ice Sheet in the region of Hudson Bay would have been similar to the Antarctic today. *See* ANTARCTICA; ARCTIC AND SUBARCTIC ISLANDS.

The extent of glaciation during the Quaternary appears to have covered at least 30% of the continents, but if the oceanic basins in the Arctic, north Pacific, north Atlantic, and Southern oceans are included, over 60% of the Earth's surface would have been icebound (Fig. 2*b*). It is likely that similar percentages of ice cover would have applied in earlier geological times.

Global glaciations can now be detected in all geological periods except the Jurassic. The Earth has alternated between greenhouse (warm, nonglacial periods, also know as interglacials) and icehouse

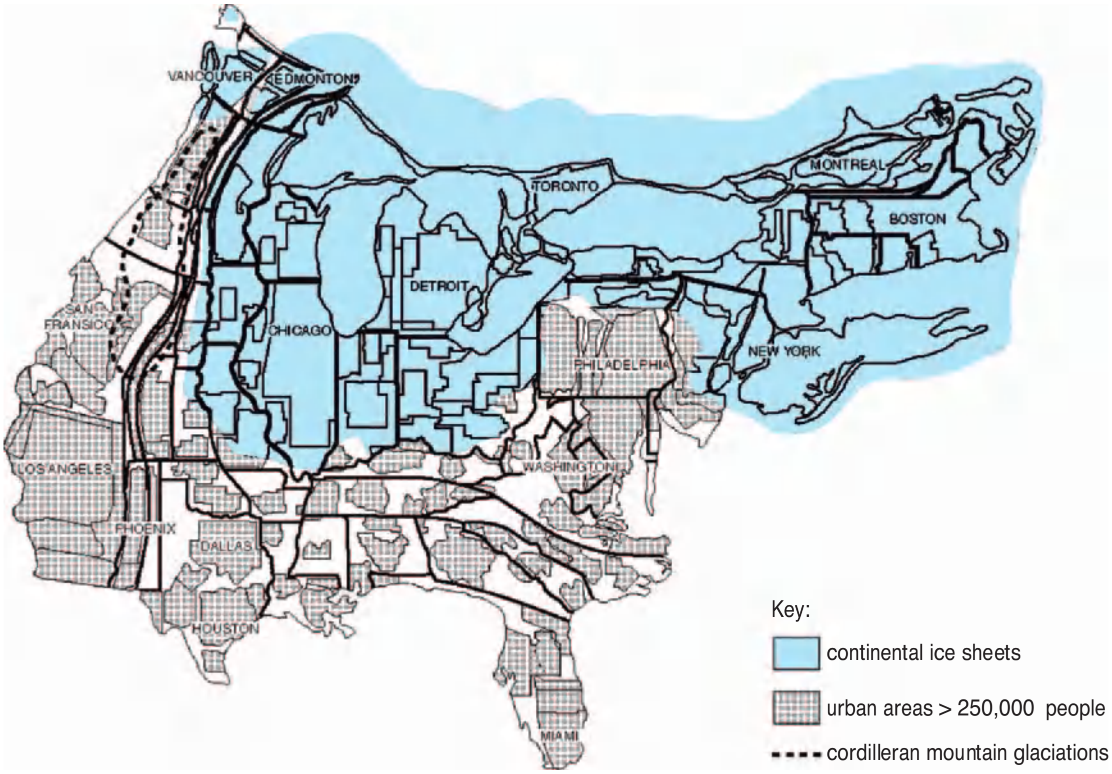


Fig. 1. Isodemographic map (1975/1976) overlain by the areal extent of the Laurentide and Cordilleran ice sheets at the last glacial maximum limits, 18,000 years ago. (*Modified from the Cart Department of Queens University, Kingston, Ontario, and R. C. Berg et al.*)

(cold, glacial periods) conditions in cycles ranging from 100,000 to 150,000 years. Recognition of the repetition of glaciation is a relatively new concept. In the nineteenth century, glacial lithified sediment (the Talchir Boulder Beds) was recognized in India, as well as the Permian glacial sediments of Shropshire, England, but a full recognition of repeated global glaciations became established only in the midtwentieth century. *See* GEOLOGIC TIME SCALE.

For a long time in Europe and North America, it was thought that a single glacial period [the Pleistocene (the ice age)] was to have occurred, giving it a unique position within the geological record. The Pleistocene is of immense importance since not only are the flora and fauna we see today largely a by-product of that glacial period, but also the emergence of human ancestors appears to stem from that period. In addition, the Pleistocene was a period of considerable upheaval in terms of plant speciation and the demise of several larger animals such as the wooly rhinoceros, giant elk, and mammoth.

When ice ages began to be recognized in Europe, a subdivision of four major glaciations was established. This fourfold division also was used in North America. Only with oceanic deep-drilling programs in the north Atlantic and Pacific oceans did it become apparent that, instead of four, 17 to 20 major glaciations had occurred during what has become a much longer and colder Pleistocene Period. Where in the past it was thought that glaciations worldwide were synchronous in their advances and retreats, it has become apparent that ice-mass margins varied considerably between continents and along different sections of the margin of the same ice sheet. For example, along the margins of the Laurentide Ice Sheet in North America, retreat and advance varied considerably between the east coast of New England and the Canadian Maritimes, as compared with the margins along the prairie regions of the Dakotas and the Canadian Prairie Provinces. This lack of synchroneity, certainly over periods less than was 2000 years, suggested a much more complex glacial geology than was previously considered. As dating methods advanced and the use of oxygen isotope ratios established warm and cold periods from data supplied from ocean drill cores, a more exact method of determining glacial and nonglacial periods has emerged using isotope stages. *See* PLEISTOCENE.

It appears we now live in an interglacial period that may be prolonged due to global warming and greenhouse gases. In the longer time span, it is possible that early accelerated global warming will lead to increased precipitation in the midlatitudes, which may have a negative feedback effect, thereby increasing the chance for a return to global glaciation. *See* GLACIAL HISTORY.

Causes of Glaciation

The causes of glaciation can be summarized by six possible mechanisms: cosmic, volcanic, tectonic, climatological, the "snowball earth," and astronomical. The repetitive nature of glaciation over geologic time must be of a paramount consideration. It has long

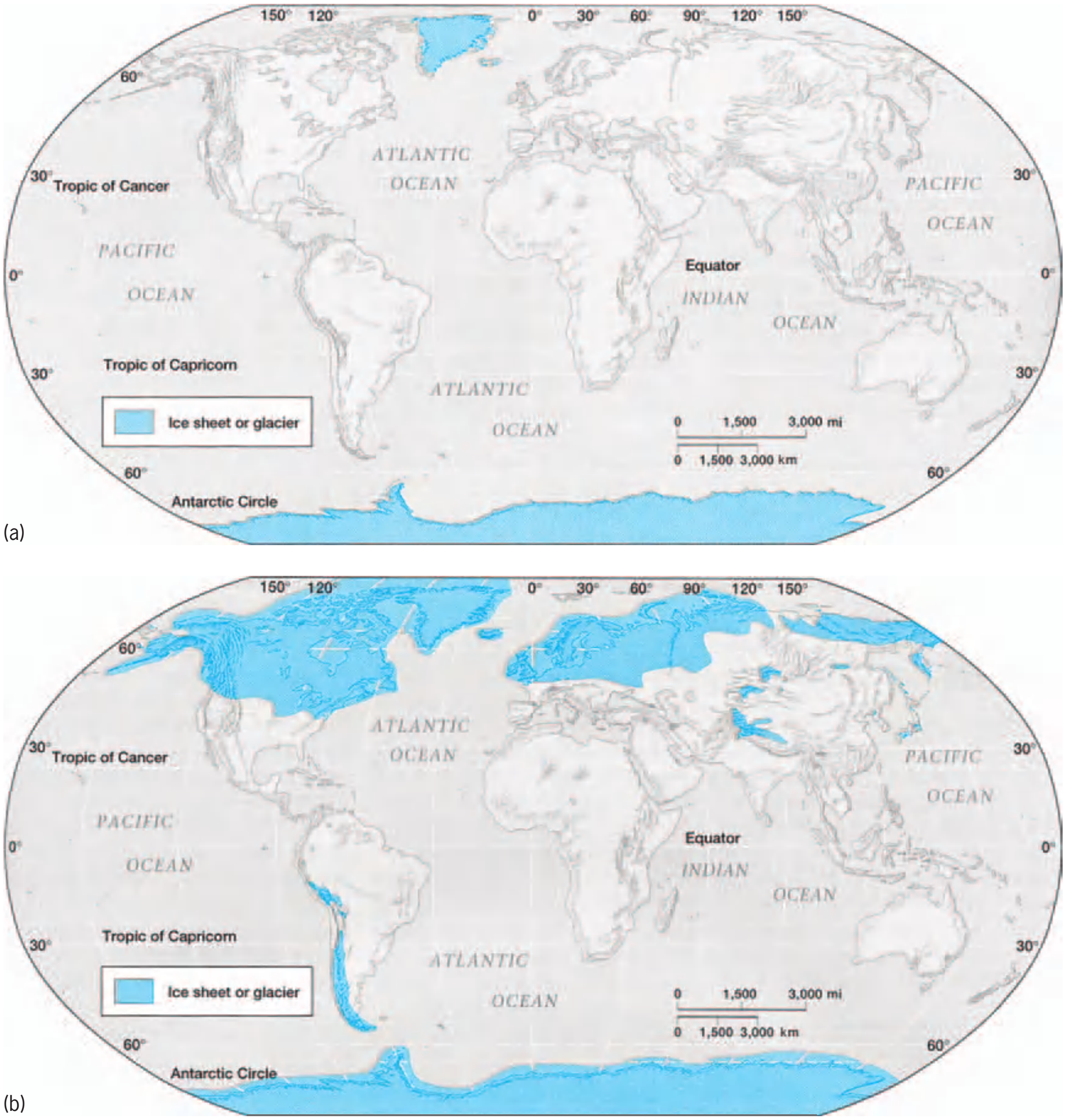


Fig. 2. Worldwide glaciation: (*a*) present distribution and (*b*) 18,000 years ago at the maximum of the Late Wisconsinan. (*Modified from T. L. McKnight and D. Hess, Physical Geography, 6th ed., 2000*)

been recognized that a reduction in solar energy striking the Earth's surface has the effect of lowering the mean annual surface temperatures by a few degrees (1–2°C), which is sufficient for snow banks and snowfields in high mountains to survive summer heat and gradually begin accumulating snow, leading to a sufficient thickness of snow that will transform to glacial ice. *See* INSOLATION; SNOWFIELD AND NÉVÉ.

Cosmic dust storms have been suggested as a mechanism of solar radiation reduction. They do occur and over the short term reduce overall solar radiant energy, but since they are of a nonrepetitive nature they cannot be a significant cause of glaciation.

Similar reductions in solar radiation occur with major volcanic eruptions when vast plumes of volcanic dust are shot high into the atmosphere and circle the Earth in the jet stream for considerable time. In 1991, the explosive volcanic eruptions from Mount Pinatubo in the Philippines caused a general global reduction in annual surface temperature of 0.3–1°C. However, this was a nonrepetitive event and not a major causative mechanism for global glaciation.

Over geologic time, continental landmasses have changed isostatically with respect to sea level, and landmasses such as the Tibetan Plateau have risen due to plate collision and orogenic forces. It has been suggested that as landmasses rise in elevation they may have been the sites of earlier glacial initiations. Again, even though much as this fact has been established, it is not repetitive and is therefore not a prime mechanism for glaciation. *See* PLATE TECTONICS.

That a climatological mechanism may trigger glaciation by changing patterns of seasons with much colder winters and cooler summers as was witnessed in medieval to midnineteenth-century Europe cannot be denied, but such a sequence of events is unlikely to cause repeated global glaciation.

In recent years, the snowball theory, or runaway albedo effect, has been proposed in which the Earth became completely encased in a vast ice sheet, with even the oceans covered by thick sea ice. Evidence from pre-Quaternary sediments points to glaciations in the Neoproterozoic (1000–543 million years ago) that seem to have begun at low latitudes close to sea level. Such a glaciation would have required the ocean to cool to at least 30°N and S latitude, at which point a runaway albedo effect would have rapidly caused the Earth to lose energy by reflecting longwave radiation and becoming entirely ice covered. Alternatively, the Earth's tilt (at least 54°) then was greater than today (23.5°), and that may have caused great surface temperature anomalies. It is possible that the composition of the atmosphere was markedly different from today in terms of greenhouse gas effects which, when added to plate tectonic motion of continental landmasses, may have led to glaciation. Evidence from the Neoproterozoic points to the viability of the concept, but since that time glaciation has not occurred in the same globally cataclysmic manner. The snowball earth remains an interesting and controversial theory but does not seem to explain any of the global glaciations since the Neoproterozoic. *See* ALBEDO.

The most likely cause of glaciation on a global scale, but allied to global variations in climate and ocean currents, appears to be solar forcing. This astronomical theory has its beginnings in the midnineteenth century when James Croll and, later, Milan Milankovitch suggested that the movement of the Earth around the Sun (eccentricity) in association with the tilt of the Earth in relation to the Sun (tilt) and the Earth's wobble on its axis (precession), when in synchroneity, might be sufficient to reduce the Earth's surface temperature enough to initiate glaciation. This mechanism, coupled to other oceanic and climatological changes, appears to be the most likely mechanism to trigger repeated global glaciation. *See* EARTH ROTATION AND ORBITAL MOTION; PRECESSION OF EQUINOXES.

It is apparent that no single factor can be said to cause global glaciation. A combination of astronomical and climatological factors in relation to ocean currents and surface ocean temperatures, coupled with the plate tectonic movement of landmasses and the overall atmospheric quantities of greenhouse gases, appear to be necessary to trigger a global glaciation. Much remains to be understood about global climatic conditions and changes in relation to the many of factors discussed above. *See* CLIMATE HISTORY.

Landscape Development

In the Pleistocene, it seems that the vast ice sheets covering the Earth's surface began a relatively slow buildup to repeated maximum extensions into the midlatitudes (Fig. 2*b*). These ice sheets had an enormous impact in moving across terrain by eroding, transporting, and depositing vast quantities of sediments both on land and in ocean basins. The surface topography of the glaciated continents was totally altered with distinctive landscapes. Landforms were repeatedly overprinted until the terrains now seen in the northern United States and throughout Canada and northern Europe were finally exposed from beneath the ice some 10,000 years ago. Although the ice-sheet buildup was relatively slow, the melting of the ice sheets was remarkably fast, with sea level rising very quickly. Since the ice sheets isostatically had depressed the continental landmasses, in many instances sea level rose above present levels, drowning large coastal areas. Later, the continental landmasses adjusted to the loss of the ice-sheet load, and slowly raised shorelines and cliffs began to reappear above sea level. In Europe, these raised beaches and strandlines became the focus of the colonization pathways of early humans as they moved into northern Europe.

Glaciated landscapes are dominated by the effects of erosion, especially as seen on bedrock surfaces in the form of glacial scratches (striae) and chattermarks and the effects of high-pressure meltwater scour (**Fig. 3**). The evidence of glacial transport is in the boulder trains and isolated erratic boulders left strewn across many glaciated landscapes (**Fig. 4**). Evidence of glacial deposition occurs as thick glacial sediments, landforms, glacial lake sediments, and the immense thicknesses of glacial sediments within marine environments.

Glacial erosion. Glacial erosion is fundamental to the production of sediments for transport and deposition. Erosion processes occur in all glacial environments. Although limited, erosion in the form of abrasion and meltwater action occurs on the surfaces of ice masses, especially during the summer months when melted ice and snow move across glacier surfaces. In mountainous areas, considerable windblown debris and avalanched and mass-movement debris often end up on top of the ice mass (that is, the supraglacial environment), where debris may

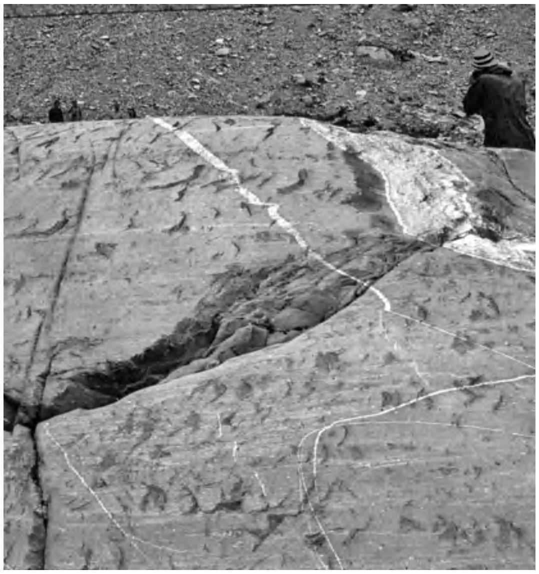


Fig. 3. Striations and chattermarks on bedrock in front of Omsbreen, Norway, ice moving from right to left. People are shown for scale.

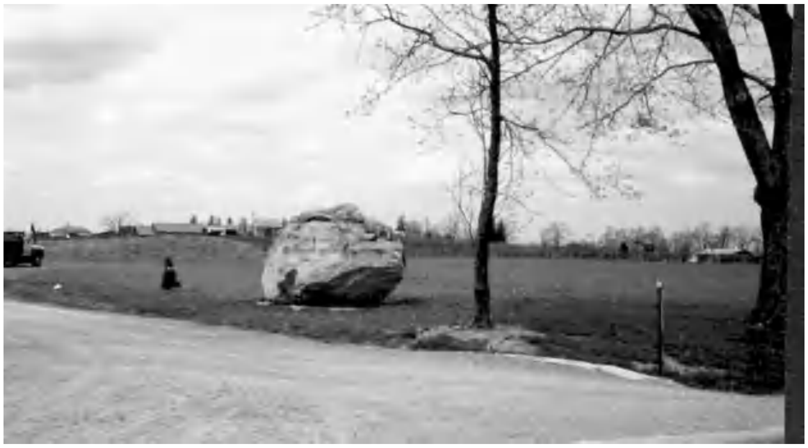


Fig. 4. Erratic boulder, approximately 3 m (10 ft) high, sitting on the side of a field in southern Ontario, Canada.

suffer abrasion from percussion and from meltwater action. Even more limited is the amount of erosion that can occur within ice masses (that is, the englacial environment) in tunnels and galleries. Glacial debris is transported spasmodically, depending on meltwater activity within an ice mass, the concentration of debris within the ice mass, and the amount of debris that enters from the supraglacial environment via crevasses and meltwater moulins (cylindrical shafts), the connectivity of englacial tunnels, and meltwater discharge in the summer melt season. Significant volumes of debris are released from ice masses along lateral and frontal margins (that is, the proglacial environment) in valley glaciers or along the front margins of large ice sheets or into oceanic basins where ice margins are floating or have ice shelves attached. Debris exiting into the proglacial environment is subject to substantial erosion largely due to meltwater transport and mass-movement activity. The dominant location of glacial erosion occurs below the glacier (that is, the subglacial environment), where there is a high level of stress, considerable meltwater activity often under hydrostatic pressures, and vast sources of erodable material.

Processes. Glacial erosional processes can be subdivided as abrasion, plucking/quarrying, meltwater action, chemical action, and freeze–thaw processes. In general, all of these processes may operate on the same rock surfaces, such that examples of glacial erosion typically possess all their characteristic marks.

Abrasion is the wear or attrition of bedrock surfaces and rock fragment surfaces by the scouring processes of debris-laden ice and meltwater. The debris within the base of the ice or debris moving within high-pressure meltwater passes over the surface of bedrock or other rock surfaces, scratching and wearing down the surface. The evidence of abrasion is seen as minute scratches, or striae, on rock surfaces, or extremely smooth rock faces with tortuous geometries illustrative of rapid wear. These latter forms of erosion are called P-forms (**Fig. 5**). To be effective, such abrasion processes demand high basal-ice debris concentrations or high debris content in high-pressure meltwater streams, debris that is sharp, angular, and harder than the rock surfaces to be cut, sufficiently high basal-ice pressures, and an effective means of evacuation of the abraded debris. This last requirement is essential so that the abrasion process can continue, that is, not become clogged. The production of such immense volumes of abraded debris is apparent in the downstream "milky" nature of glacial streams, which exhibit a blue or greenish-blue color due to the high content of fine debris known as glacial milk or rock flour.

Plucking or quarrying is a set of processes that remove fractured, jointed, or disaggregated rock

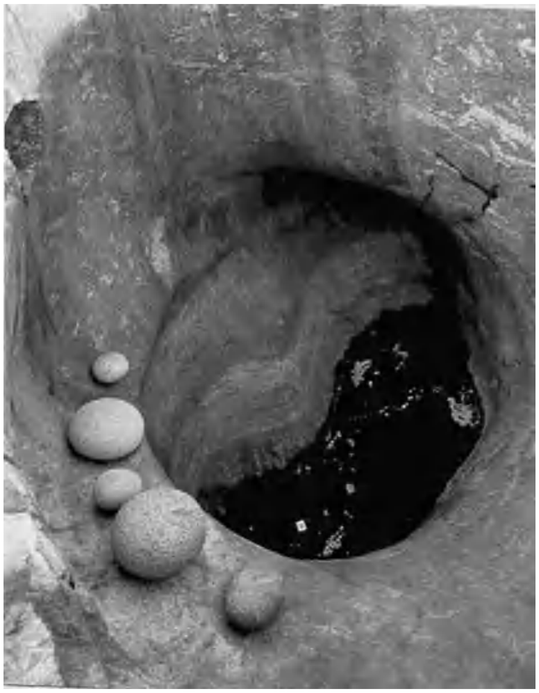


Fig. 5. Large pothole (glacial P-form), Finland. (*Geological Survey of Finland*).

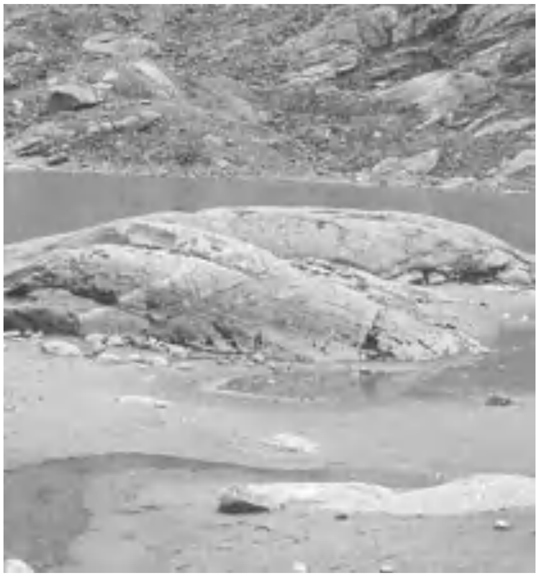


Fig. 6. Roche moutonée, approximately 25 m (80 ft) in length in front of the Nigardsbreen Glacier, Norway, with ice moving from right to left.

fragments from bedrock or the surface of rocks. Typically, rocks fracture under tensile or compressive stresses produced by the overlying moving ice mass or by percussive processes, whereby rock fragments crash against rock surfaces and produce flakes and shards of rock. Rock plucking can produce enormous boulders as well as minuscule rock flakes. Where large bedrock knobs have been overridden, roche moutonée may be produced as distinctive erosional landforms with smooth up-ice (stoss) sides and steep craggy down-ice (lee) sides (**Fig. 6**). Where percussive plucking occurs, tiny chattermarks and lunate fractures are developed on rock surfaces.

Meltwater, in association with high debris content and high-pressure discharge, creates other forms of abrasive wear on rock surfaces. A little understood process is chemical weathering beneath ice masses, where hydrostatically pressured meltwater, high stress levels, and fluctuating temperatures lead to carbonate, silicate, and iron solution and reprecipitation. Characteristically, such chemical processes can usually be observed as distinctive stains or "trails" on rock surfaces, often on the lee side of bedrock protrusions. In a glacial environment, freeze-thaw processes (seasonal and diurnal) are active, typically producing large volumes of frost-shattered rock. *See* WEATHERING PROCESSES.

Landforms. Glacial erosional landforms can be subdivided into areal and linear forms. The range of scale of these landforms can be immense, from centimeter-sized "rat-tails" and "flutes" on rock surfaces (**Fig. 7**) to roche moutonnée tens of meters in height (Fig. 6).

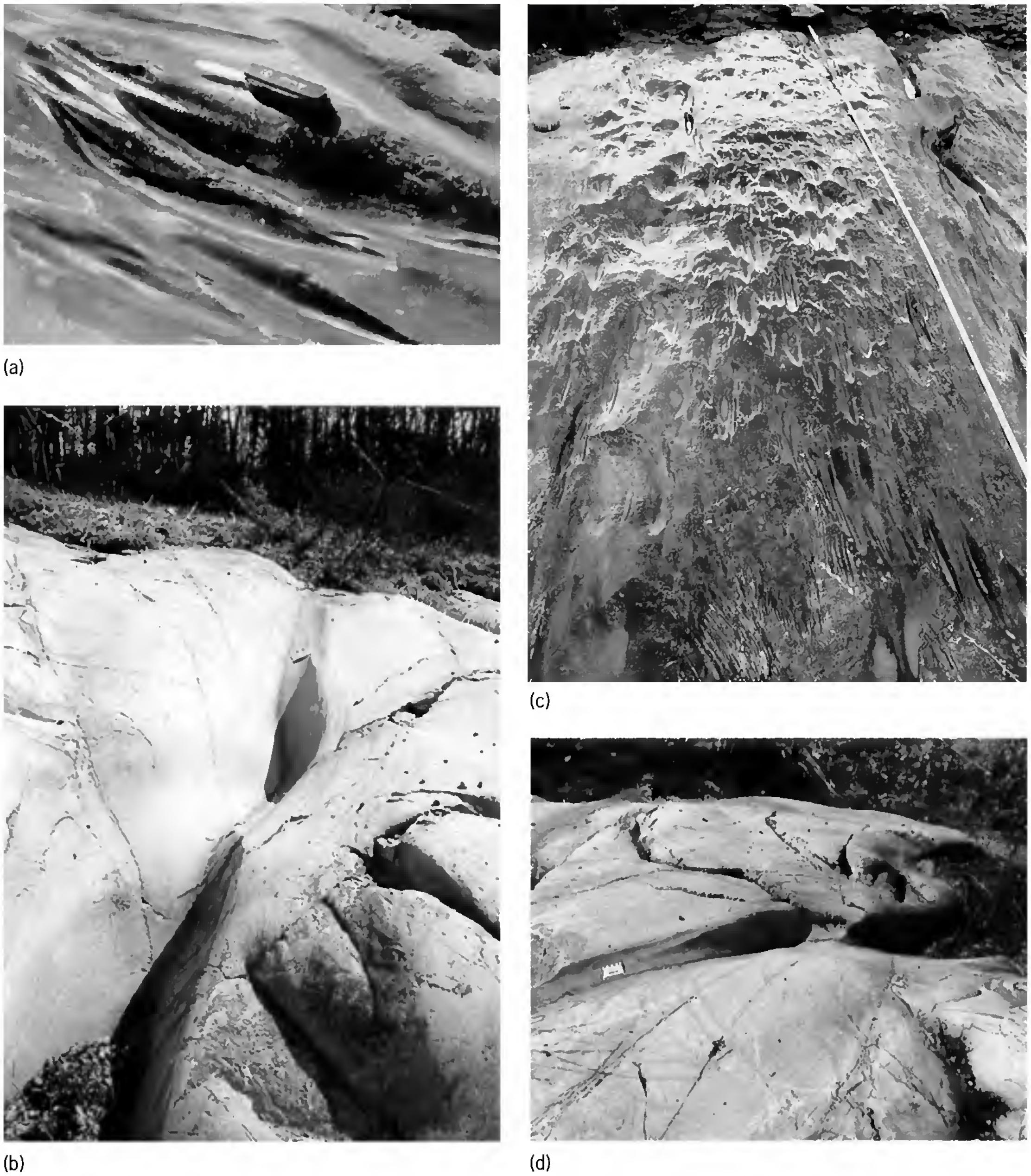


Fig. 7. Fluted and streamlined bedrock forms: (*a*) small-scale flutes, Wilton, Ontario; (*b*) P-forms in bedrock, Espanola, Ontario; (c) subglacial meltwater erosion marks; and (*d*) P-forms, Espanola, Ontario (scale card is 8.5 cm long).

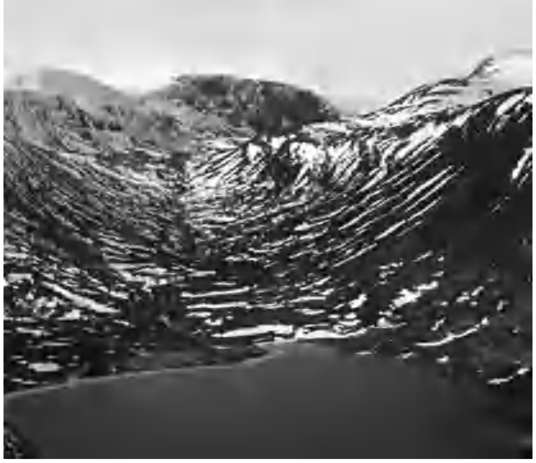


Fig. 8. U-shaped valley, Muick, Scotland.

Vast areas of the Canadian and Fenno-Scandian shields best portray widespread areal glacial erosion. Linear forms occur in all glaciated regions, such as fiords, troughs (U-shaped valleys) [**Fig. 8**], finger lakes, tunnel valleys, and, at the much smaller scale, fluted bedrock knobs (P-forms), and roche moutonnée. All these landforms reflect the erosive power of glaciers, ice sheets, and associated meltwater action.

Landscape systems. Perhaps the best example of a glaciated landscape system is high mountain glaciation, where pyramidal peaks (horns), arêtes, cirques, tarn lakes, rock steps (riegels), and hanging valleys are found in close association (**Fig. 9**). No similar landscape system exists for ice-sheet erosive landscapes except for the areal scour, bedrock trough, and fluted-bedrock-knob landscapes of the Canadian Shield.

Glacial transport. In altering a landscape through erosion and subsequent deposition, glaciers and ice sheets transport vast quantities of sediment from fine-grained clays and silts to huge boulders. Sediment is transported on the surface of the ice as supraglacial debris, within the ice as englacial debris, beneath the ice as subglacial debris, and beyond the ice margins as proglacial sediment. Each environment affixes the transported sediment with a potentially distinctive signature that, although complex due to overprinting and retransportation, can be used to differentiate glacial sediment types.

Glacial deposition. The deposition of glacial sediment is largely a function of the means by which that sediment has been transported. Since ice masses scavenge sediments from their glaciated basins (in the case of ice sheets, this may be continental landmasses), the provenance of glacial sediment is typically immense, containing far-traveled sediment and boulders as well as local rock materials. The idea that glacial sediments are principally composed of far-traveled rocks is erroneous since most glacial sediments reflect a provenance only a few tens of kilometers up-ice from any point of deposition.

Glacial deposits can be subdivided into unsorted sediments with a wide range of particle sizes, and sorted sediments deposited by meltwater. Unsorted sediments are typically referred to as till, or glacial diamicton, and are exceedingly complex in their sedimentology and stratigraphy. They occur in all subenvironments of the glacier system and range from subglacial high-clay consolidated lodgement tills, to coarse supraglacial flow tills, to subglacial and submarginal areas of basal ice melting, leading to melt-out tills. In recent years, this classification with reference to subglacial tills is probably inaccurate. It would be more rigorous to call these tills tectomicts—products of the complex subglacial environment. *See* TILL.

Sorted, or fluvioglacial, sediments range from fine-grained clays deposited in glacial lakes (glaciolacustrine), to coarser sands and gravels deposited in front of ice masses as outwash fans (sandur), to rainout sediments deposited in marine settings (glaciomarine).

Processes. The processes of glacial deposition include direct smearing-on of subglacial debris as lodgement tills, mass movement of sediments from the frontal and lateral margins of glaciers as flow tills, and direct ablation of ice and melt-out tills formed in

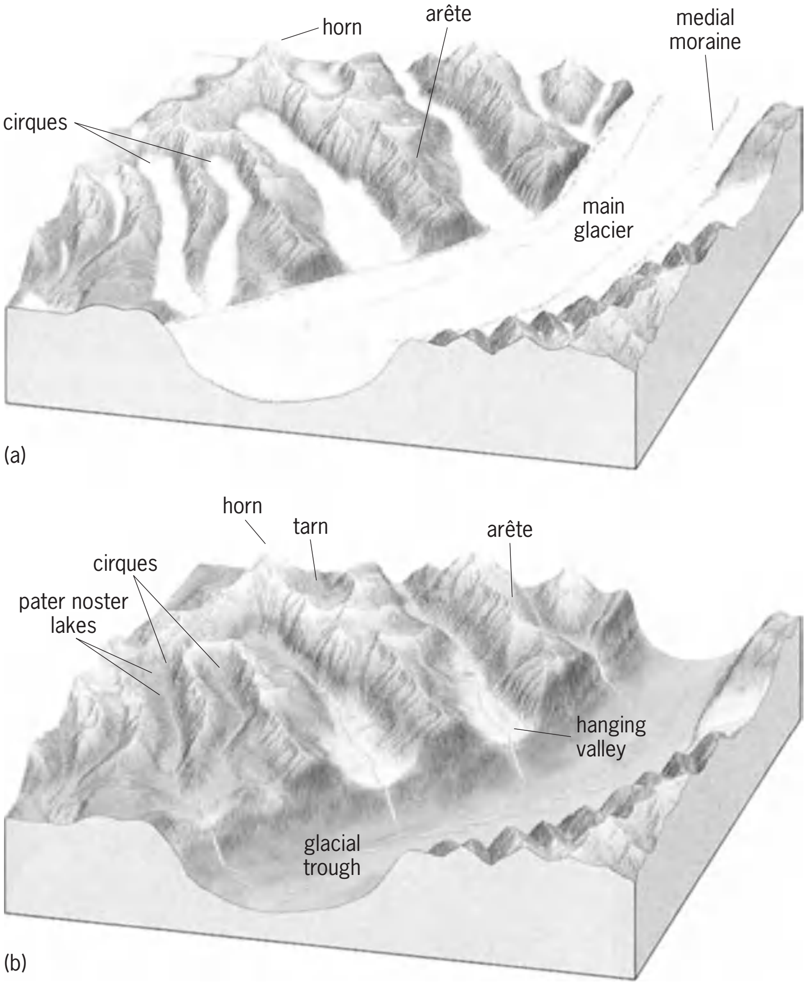


Fig. 9. Erosional landscape system for alpine glaciation. (*a*) Region during period of maximum glaciation. (*b*) Glaciated topography. (*Modified from E. J. Tarbuck and F. K. Lutgens, 2003*)

front of ice masses in proglacial areas and in melt-out within subglacial cavities and caverns.

Landforms. Glacial landforms, or bedforms, range from those formed transverse to ice motion and parallel to ice motion, to unoriented nonlinear forms, ice-marginal forms, and fluvioglacial landforms. Although a complex number of glacial depositionary forms occur in most glaciated areas, the dominant landforms include Rogen moraines, drumlins and fluted moraines, hummocky moraines, end moraines, and eskers and kames.

Rogen moraines, or ribbed moraines, are a series of conspicuous ridges transverse to ice movement. These ridges typically rise 10–20 m (33–66 ft) in height, 50–100 m (164–328 ft) in width, and are spaced 100–300 m (328–984 ft) apart. They tend to occur in large numbers as fields in Quebec and Finland. They are composed of a range of subglacial sediments and are thought to be formed by the basal deformation of the underlying sediment, possibly at times in association with a floating ice margin (**Fig. 10**).

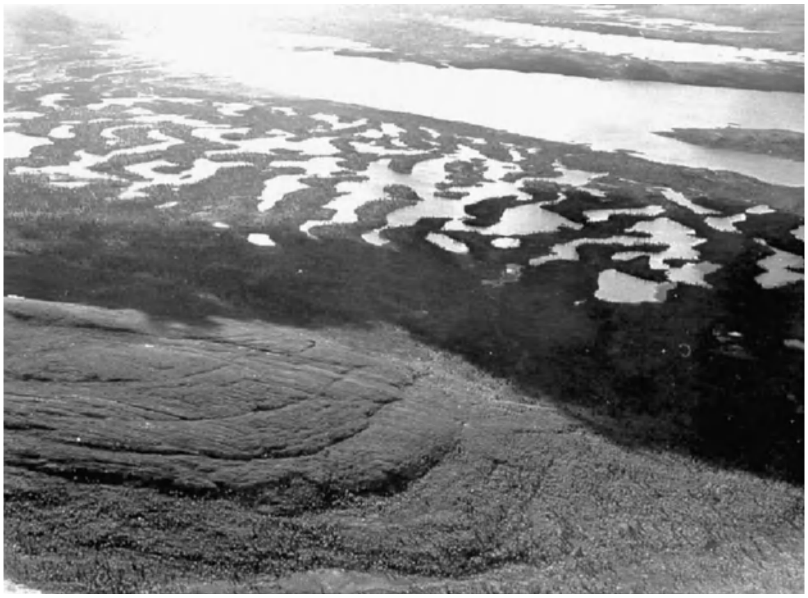


Fig. 10. Rogen moraine near Uthusslön, Sweden. (*Courtesy of Jan Lunqvist*)

Drumlins are one of the most known glacial landforms. A considerable literature exists as to their formation, and debate continues on their mechanism of origin. Drumlins are streamlined, roughly elliptical, or ovoid-shaped hills with a steep stoss side facing up-ice and a gentler lee side facing down-ice (**Fig. 11**). Drumlins range in height from a few meters to over 250 m (820 ft) and may be from 100 m (330 ft) to several kilometers in length. They tend to be found in vast swarms or fields of many thousands. In central New York State, over 70,000 occur. Likewise, in Finland, Poland, Scotland, and Canada, vast fields exist. Drumlins are composed of a wide range of dominantly subglacial sediments but may contain boulder cores, bedrock cores, sand dykes, and “rafted” nonglacial sediments. It seems likely that these landforms were developed below relatively fast-moving ice, beneath which a deforming layer of sediment developed inequalities and the preferentially stiffer units became nuclei around which sediment plasters and the characteristic shape evolved. Other hypotheses of formation are the streamlining by erosion of preexisting glacial sediments, changes in the dilatancy of glacial sediments leading to stiffer nuclei at the ice-bed interface, fluctuations in pore-water content and pressure within subglacial sediments again at the ice–bed interface, and the infilling of subglacial cavities by massive subglacial floods.

Fluted moraines are regarded as subglacial, streamlined bedforms akin to drumlins (**Fig. 12**). These landforms are typically linked in formation with drumlins and Rogen moraines. They tend to be much smaller in height and width as compared to drumlins but may stretch for tens of kilometers in length. They are composed of subglacial sediments. Fluted moraines have been found to develop in the lee of large boulders but may also be attenuated drumlin forms, the result of high basal-ice shear stress and high ice velocities. In central New York State, fluted moraines occur beside and among the large drumlin fields.

Hummocky moraine is a term used to denote an area of terrain in which a somewhat chaotic deposition pattern occurs, similar to a series of small sediment dumps. These moraines rarely rise above a few meters but may exist over considerable areas. Hummocky moraine often marks locations where massive downwasting of an ice mass may have occurred.

End moraines or terminal, retreat, or frontal moraines occur at the front margins of ice masses. Typically, these landforms, transverse to ice motion, may be a few meters to tens of meters in height and, if built up over several years, may be even higher and of considerable width (1–5 km or 0.6–3 mi). These moraines contain subglacial and supraglacial debris, and often have an arcuate shape closely mirroring the shape of an ice front. Since these moraines mark the edge of an ice mass at any given time, the longer the ice remains at that location, the higher and larger the moraines become. As an ice mass retreats with periodic stationary periods (stillstands), a series or sequence of moraines develop that mark the retreat stages (**Fig. 13**). Vast sequences of moraines can be observed in the midwestern states of Illinois, Ohio, and Michigan.

Eskers are products of subglacial meltwater streams in which the meltwater channel has become blocked by fluvioglacial sediments. Eskers are long ridges of sand and gravel that run across the landscape (**Fig. 14**). They range in height from a few meters to >50 m (160 ft) and may run for tens of kilometers. In Canada, some eskers cross the Canadian Shield for over 100 kilometers (60 mi). These landforms often have branching ridges and a dendritic morphology. Eskers are dominantly composed of fluvioglacial sand and gravel, with distinctive faulted strata along the edges where ice-tunnel walls melted

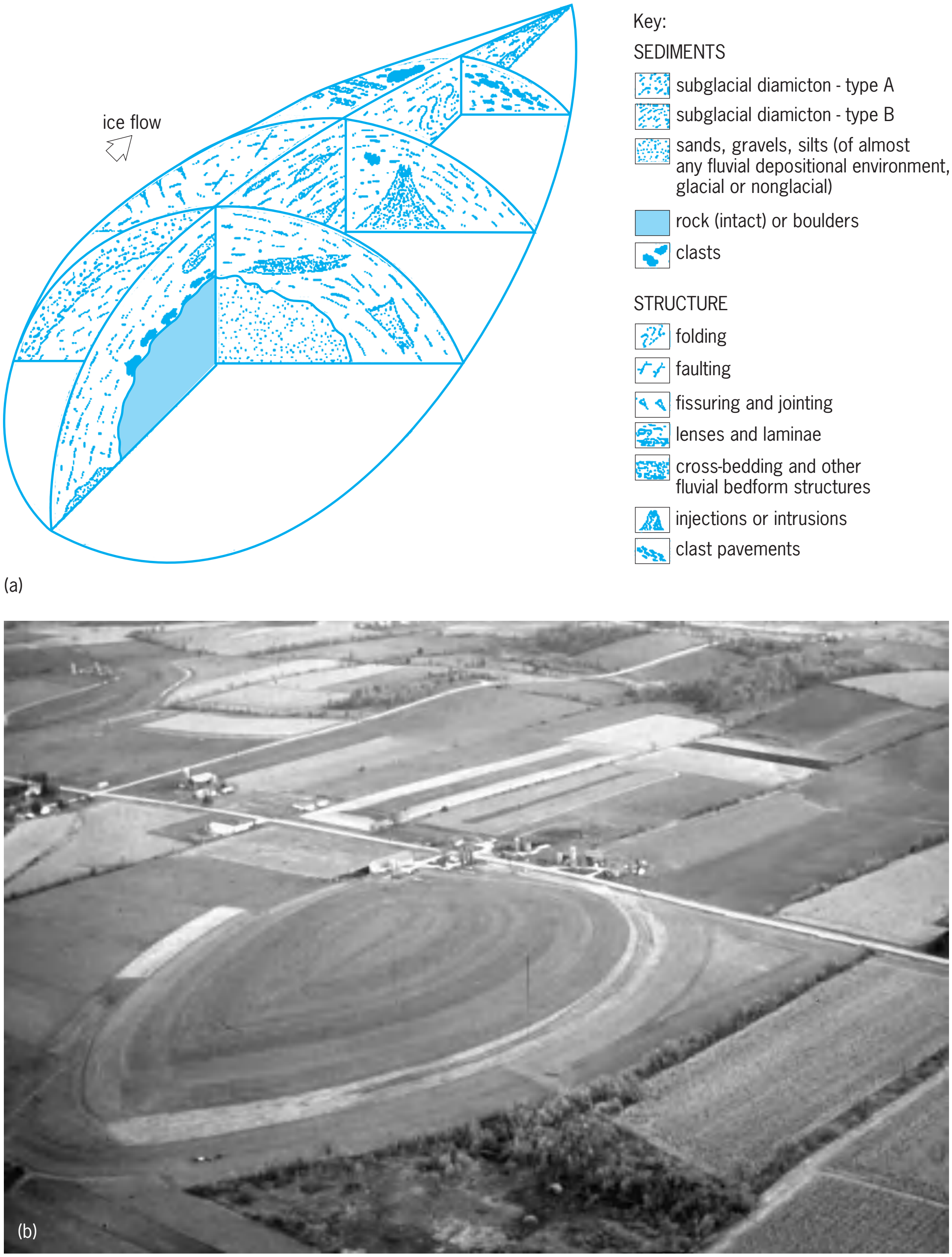

ice flow
Key:
SEDIMENTS
subglacial diamicton - type A
subglacial diamicton - type B
sands, gravels, silts (of almost any fluvial depositional environment, glacial or nonglacial)
rock (intact) or boulders
clasts
STRUCTURE
folding
faulting
fissuring and jointing
lenses and laminae
cross-bedding and other fluvial bedform structures
injections or intrusions
clast pavements
(a)
(b)

Fig. 11. Drumlin S. (*a*) General model, showing the variability of the internal composition. (*b*) Green Bay Lobe, Wisconsin. Ice direction from the upper right to lower left. (*Courtesy of D. Setz*)

and collapsed. In many instances, eskers are observed "running" uphill or obliquely crossing over drumlin—evidence of their formation within high-pressure subglacial meltwater tunnels.

In many glaciated areas during ice retreat, large, roughly circular dumps of fluvioglacial sand and gravel occur. These forms, called kames, appear to form when infilled crevasses or buried ice melts, leaving behind the sediment in a complex but chaotic series of mounds. The term kame or kame delta has been applied to fluvioglacial sediments that formed temporary deltas on entering now-gone glacial lakes. In Finland, a long line of such kame deltas formed into a moraine-like series of linear ridges (Salpausselkä) transverse to the ice-front retreat.

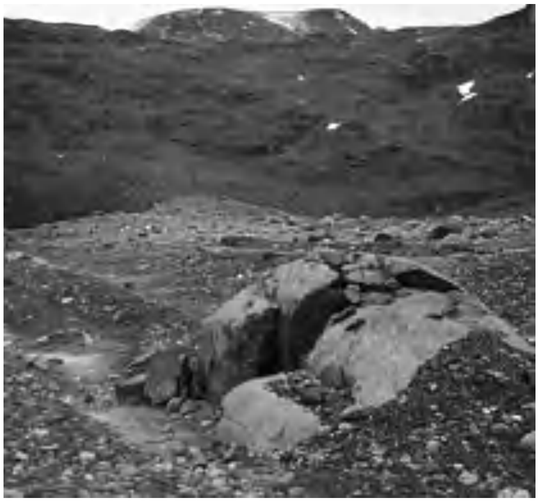


Fig. 12. Fluted moraine, Storbreen Glacier, Norway. Note boulder at he head of the flute; ice direction bottom right to middle left.

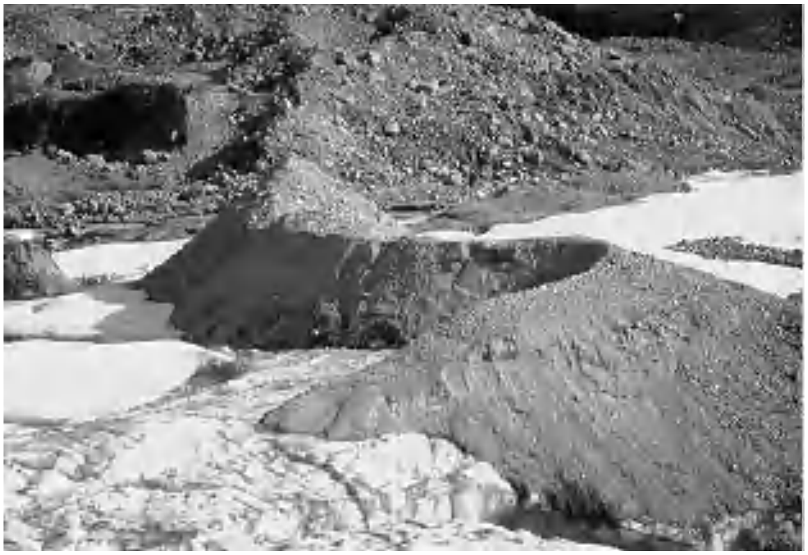


Fig. 14. Esker ridge, Bylot Island, Canada. (*Courtesy of C. Zdanowicz*)

Distinctive suites of glacial depositional landforms can be found in valley glacier and ice-sheet settings (**Fig. 15**). Typically, these sequences of landforms are often overprinted due to subsequent glaciations, but in the midwestern United States such suites of depositional landforms from the last glaciation (Late Wisconsinan) can be clearly discerned (**Fig. 16**). Distinctive landscape systems attributable to the marginal areas of ice sheets also carry a unique supraglacial suite of landforms, as can be observed in parts of the midwestern United States.

Periglacial effects. Areas beyond the ice limits are strongly influenced by glaciation due to deteriorating climatic conditions, the deposition of windblown dust (loess), the divergence of river systems where headwaters or partial drainage basins may be intersected by advancing ice, the impact of outburst floods from ice fronts (jökulhlaups), and major faunal and floral changes due to encroaching ice. In terms of the impact of ice masses on human life and

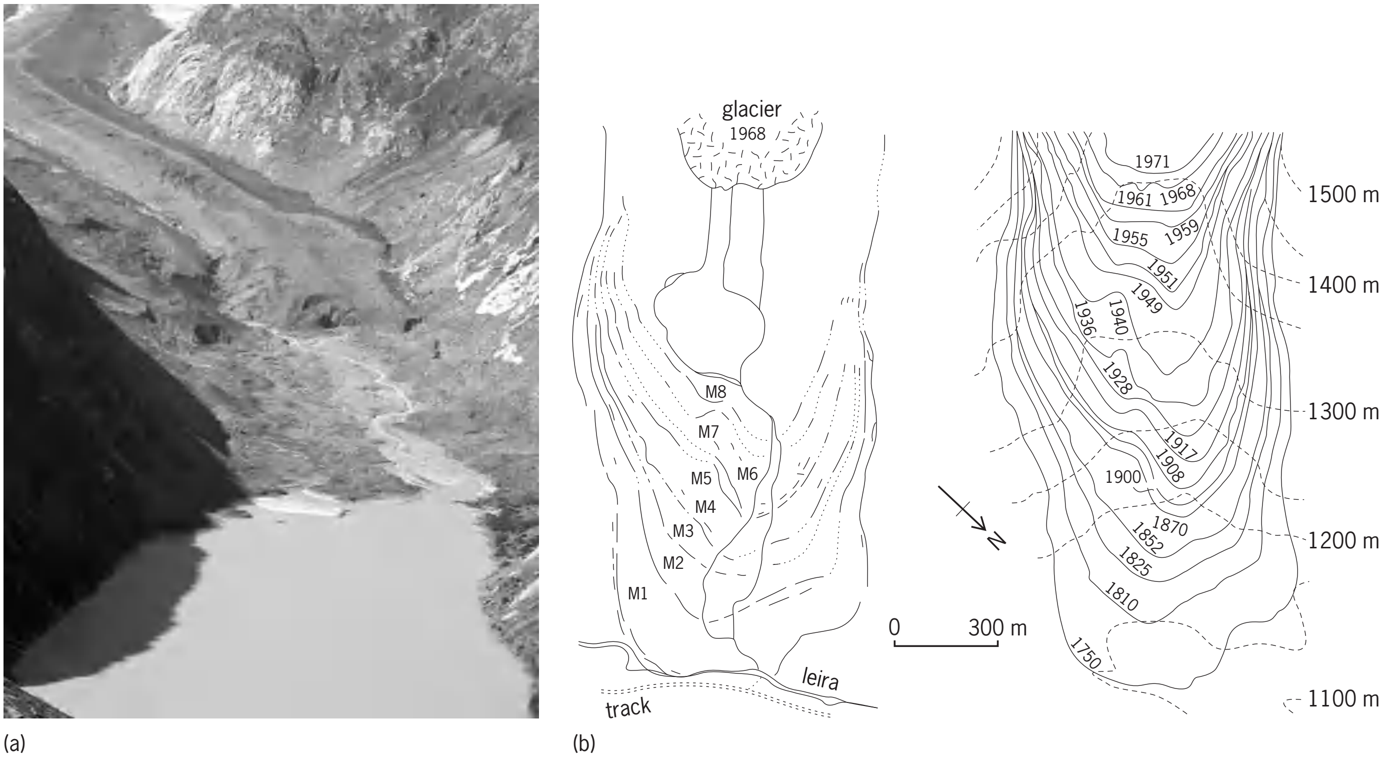


Fig. 13. End moraine. (*a*) Findelen Glacier, Switzerland (*Courtesy of J. Matthews*). (*b*) Sequence of retread of end moraines from Storbreen Glacier, Norway.

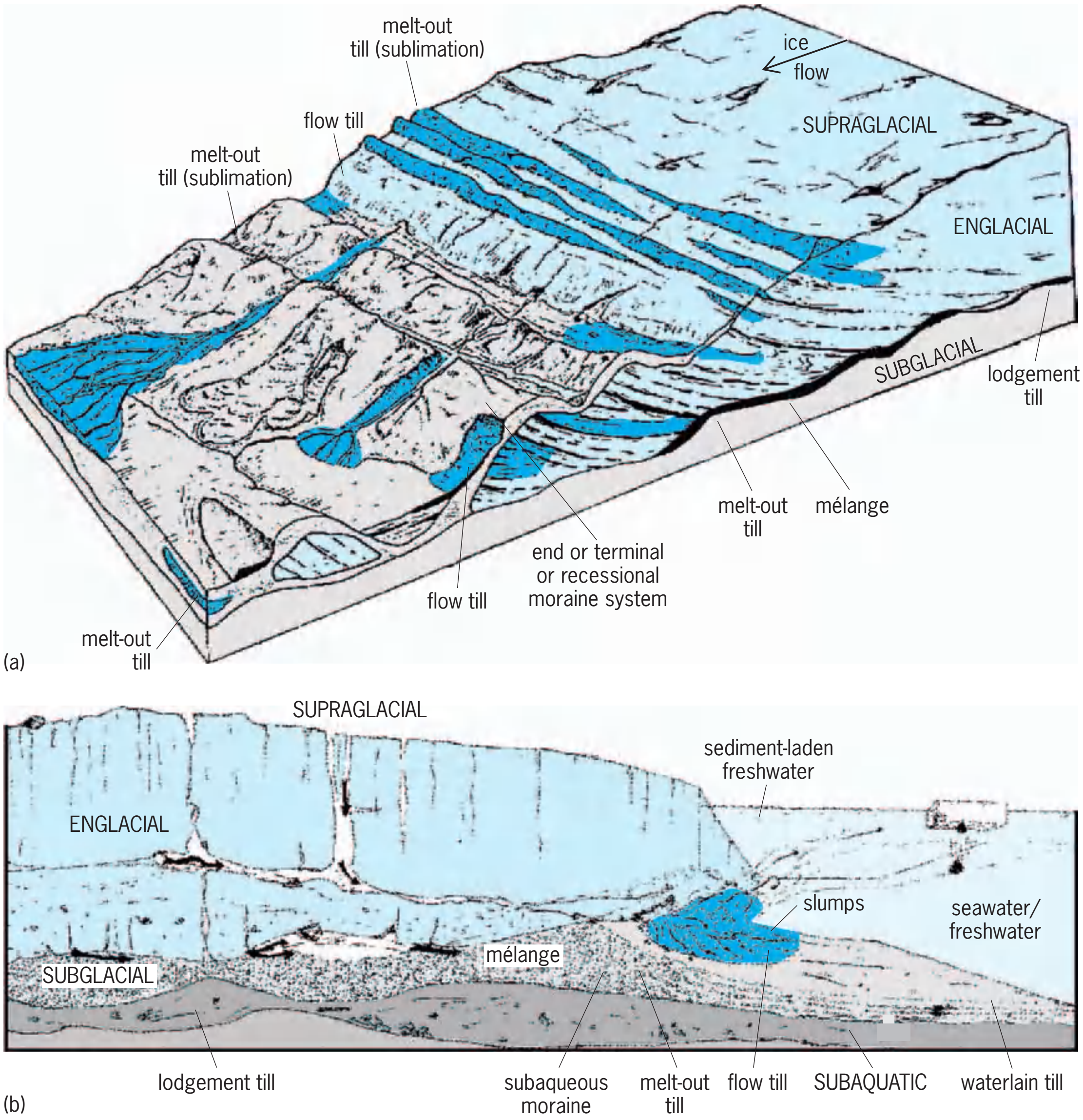


Fig. 15. Models of (*a*) land-based glacial depositional system and (*b*) depositional system of an ice mass with a marine moraine.

society, the only evidence is somewhat anecdotal, and at the close of the Late Wisconsinan may have acted to spur human migration, for example, through the short-lived ice-free corridor between the Cordilleran and Laurentian ice sheets in central Alberta, Canada, into the prairies to the south.

Within a few hundreds of kilometers of the vast ice sheets that covered North America and Europe, climatic conditions must have been severe, with strong katabatic cold winds descending from the ice. Associated with the poor climates, periglacial conditions must have prevailed in which the ground became permanently frozen to a considerable depth. The effects of periglacial activity produced frozen ground phenomena such as the movement of sediments down slopes, ice and sand wedges, localized ice lenses, and large ice-cored mound (pingo) formation. Dramatic changes in vegetation types and animal life occurred in the Northern Hemisphere, with the southern migration of many species. Unlike today, for example, central-southern Texas would have supported hardwood forests of the type now found in Ohio and Pennsylvania.

Due to the katabatic winds, fine sediments were picked up from the proglacial areas along the margins of the ice sheets, and the dust was transported away from the ice and deposited as thick, massive loess sediments (glacioaeolian sediments). Considerable thicknesses of these sediments occur in the midwestern United States, especially in Iowa and Kansas, as well as in central Europe, Hungry, and China. *See* LOESS.

Where ice advance cuts across drainage divides, rivers reduced and occasionally diverted around end moraines. In some instances, vast outpourings of meltwater (jökulhlaups) flooded from the ice fronts leading to distinctive heavily dissected terrains being formed (badlands). Such an occurrence is recorded in the Columbia River Badlands of Washington and Oregon states, where a huge lake

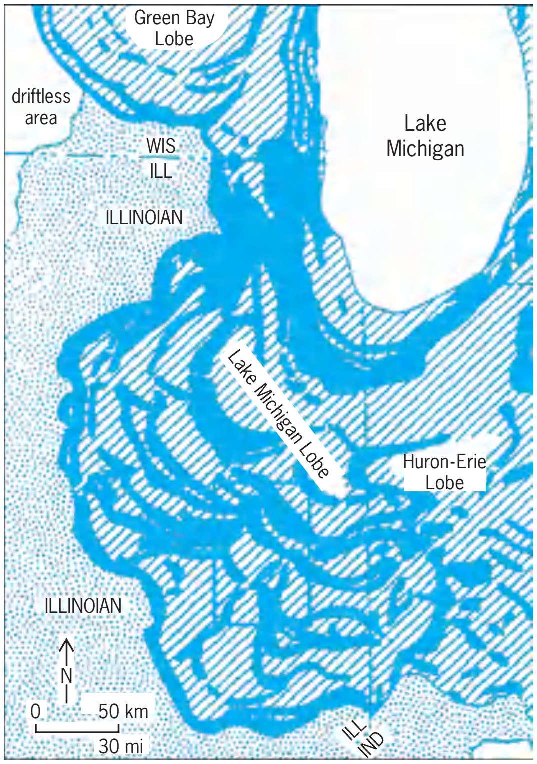


Fig. 16. Late Wisconsinan end moraines of the Green Bay Lobe, Lake Michigan Lobe, and Huron-Erie Lobe. (*Modified from J. C. Frye and H. B. Willman, 1973*)

formed and then dramatically drained (Lake Missoula floods).

John Menzies

Bibliography. D. I. Benn and D. J. A. Evans, *Glaciers and Glaciation*, 1998; M. R. Bennett and N. F. Glasser, *Glacial Geology*, 1996; J. Menzies (ed.), *Modern and Past Glacial Environments*, rev. student ed., 2002; D. Mickelson and J. Attig (eds.), Glacial Processes Past and Present, Spec. Pap. 337, Geological Society of America, 1999.

Glacial history

The glacial history of the Earth is complex and extends back in geological time to the Proterozoic and possibly the Archean. Global glaciations have occurred during every geological period except the Jurassic. This persistence of global glaciation, as well as repeated and continuing glaciation in high mountainous areas, wherever altitudes exceed the local snowline, is so consistent that one might argue that the Earth is essentially a glacial planet. Evidence suggests that glacial epochs have repeatedly occurred almost every 100,000 to 150,000 years. *See* CLIMATE HISTORY; GEOLOGIC TIME SCALE; GLACIOLOGY.

Context. Since medieval times, there have been curious explanations of features that are now recognized as the result of glaciation, such as erratic boulders once regarded as the putting stones of giants, or meltwater potholes as the devil's punchbowls. The Norsemen in Iceland and Greenland, as reflected in the tenth- and eleventh-century sagas, recognized the power and impact of glacier advance on the surrounding terrain in terms of glacial erosion scouring bedrock surfaces, the advance and retreat of glaciers across pastureland, and the subsequent loss or reappearance of the land. In Switzerland, France, Germany, and Norway, the effects of glaciation were recognized by the 1700s. During the twelfth to eighteenth centuries, a cold climate period, known as the Little Ice Age, affected much of northern Europe, resulting in famines because of the drastic reductions in crop yields. During this period, glaciers expanded in the Swiss and French Alps and overran pastureland, and permanent snow beds in the highlands of Europe and in snowbound mountain passes were reported by travelers. This Little Ice Age also occurred in North America in what is referred to as the Neoglacial.

It was not realized until the 1840s that significant glaciations had affected Europe, when Louis Agassiz, professor at the University of Neuchâtel, Switzerland, and later at Harvard University, published *Études sur les Glaciares*. By the early part of the twentieth century, it was largely accepted that major glaciations had affected the Earth.

Although Agassiz's work is often hailed as the precursor to the realization that the Earth had been repeatedly glaciated during the Ice Age, it was apparent to several early geologists that ancient global glaciations could be detected from the examination of much older rocks. As early as 1855, it was suggested that the Permian rocks of Shropshire, England, were glacial in origin. Similarly, the Permo-Carboniferous Talchir boulder beds of India and others at Hallet's Cove in Australia testified to massive glaciations. In northern Ontario, the Northwest Territories of Canada, Wyoming, Virginia, Greenland, Spitsbergen, and Siberia, lithified sediments (tillites/diamictites) have been reported as evidence of ancient glaciations. Many of the locations where ancient glacial sediments have been found occur today in tropical areas such as the Amazon Basin, Namibia, and Mauritania. Long before the understanding of plate tectonics and continental drift, paradoxical questions arose as to how such glaciations could have occurred.

By the early 1900s, a fourfold sequence of glacial advance, followed by interglacial periods of warmer temperatures (akin to today's climate), had been derived for Europe based largely upon examination of the north alpine German river terraces by A. Penck and E. Brückner (see **table**). By the mid-1900s, the idea that these glaciations had spread to the other parts of Europe and North America was generally accepted, such that a fourfold set of glaciations, called the Pleistocene, became the accepted norm worldwide. *See* PLEISTOCENE.

With increasing marine exploration and the development of various dating techniques, especially oxygen isotope dating, it was increasingly clear that instead of four major global glaciations in the Pleistocene, there were more than 17 and possibly even more than 20 (**Fig. 1**). Since ocean basins are superb repositories of terrestrial sediment with little or no erosion, complete sequences of sediment

Major Pleistocene glacials and interglacials of North America and Europe following the classical system

European Alps	Northwest Europe	Britain	North America
Würm	**Weichsel**	**Devensian**	**Wisconsinan**
R/W	*Eem*	*Ipswichian*	*Sangamon*
Riss	**Warthe/ Saale/ Drenthe**	**Wolstonian**	**Illinoian**
M/R	*Holstein*	*Hoxnian*	*Yarmouth*
Mindel	**Elster**	**Anglian**	**Kansan**
G/M	*Cromerian*	*Cromerian*	*Aftonian*
Gunz			**Nebraskan**

*From bottom to top, the time sequence is from earlier to later Pleistocene. Interglacials in italics; glacials in bold.

layers, which reflect terrestrial processes and climatic conditions, are found in ocean deeps. Thus, repeated glaciation on land is well preserved and represented in the ocean sediments. *See* DATING METHODS; MARINE SEDIMENTS; PALEOCLIMATOLOGY.

Impact of glaciation. Apart from the obvious impact of glacier ice in eroding, transporting, and depositing vast volumes of sediments, glaciation has led to many indirect effects on the landscape and life. With the advance and retreat of vast ice sheets and the more localized fluctuations of mountain glaciers, considerable impact can be seen on local and regional fauna and flora. In some cases, especially at the close of the ice ages, extinction of species seems to have occurred, such as the woolly mammoth, the saber-toothed tiger, and the giant Irish elk. The migration of plants and animals occurred across the rapidly cooling and tundralike central plains of North America in advance of ice expansion from the Laurentide ice sheet and associated ice caps and fields. With the retreat of the ice, a northern migration ensued. Similar migrations can be seen in Europe, Asia, New Zealand, and South America. *See* BIOGEOGRAPHY; POSTGLACIAL VEGETATION AND CLIMATE.

In direct impacts upon the biogeography can be seen in increasing periods of aridity and later in periods of higher-than-normal precipitation (pluvials), the diversion of surface streams and rivers, and the huge fluctuations in some major river discharge regimes (for example, the Mississippi), causing significant changes thousands of kilometers downstream and into neighboring oceans. In the North Atlantic, vast outpourings of sediment and associated meltwater can be observed in the detection of Heinrich events that tell of vast changes in the subglacial glaciodynamics of the Laurentide ice sheet. As ice sheets grow and advance, crustal depression (isostatic depression) occurs. Following ice retreat, the reverse, or crustal rebound, occurs. Even today in northern United States and Canada, this rebound is ongoing. *See* ISOSTASY.

In close association with crustal depression and rebound is the fact that ice sheets act as enormous reservoirs of water evaporated from the ocean and deposited as snow on distant ice sheets. The ice sheets hold this water for thousands of years, resulting in a net decrease in ocean level (eustatic change). It is likely that sea level fell by as much as 120 m (390 ft) at the maximum extension of the last Pleistocene ice sheets. Unlike crustal rebound, which is

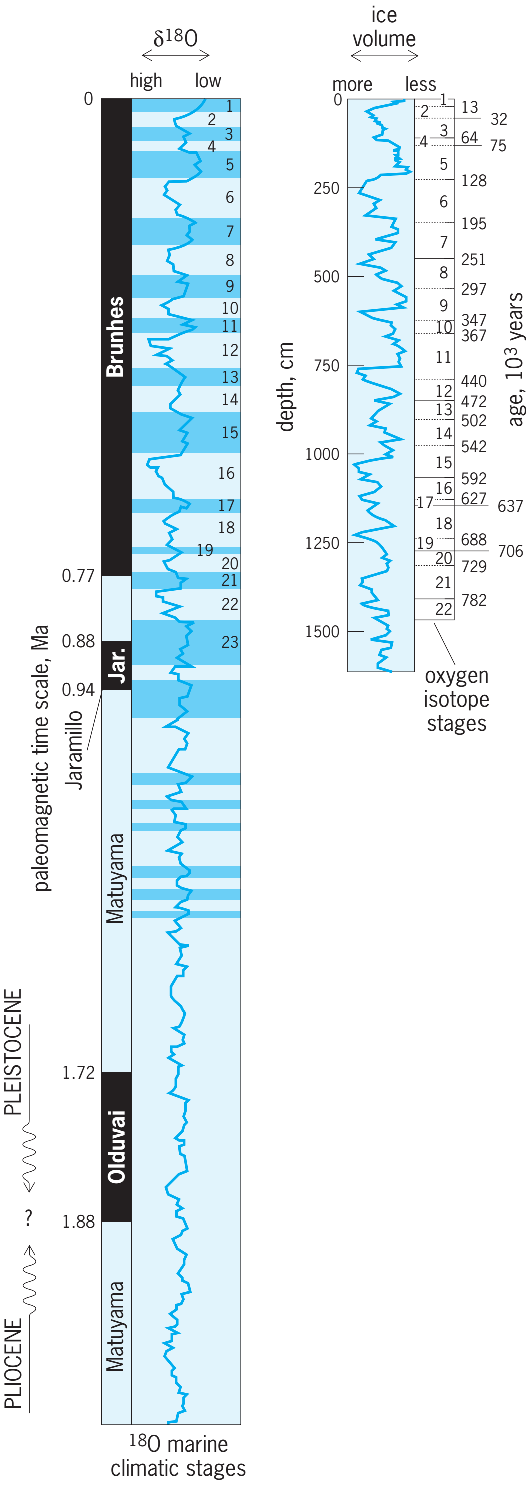


Fig. 1. Oxygen isotope records for past 1.88 million years, indicating global ice volume. In isotope stages 1 to 23, that indicates glaciations.

initially rapid but quickly slows down, sea level rises very rapidly following ice retreat and wastage. The impact of such sea-level rise is to drown coastlines, which reemerge from the sea once crustal rebound catches up as evidenced by raised beaches and abrasion platforms and exhumed spits and bars.

During and following the advance and retreat of global ice masses in the Pleistocene as well as dramatic shifts in biogeographic zones, early human evolution and later human migrations appear to have taken place. At the height of the late Pleistocene, migration of the first people into North America occurred across Beringia (dry land where the Bering Strait is now) and then southward, presumably following migrating caribou along the corridor formed between the relatively nearby margins of the western edge of the Laurentide ice sheet and eastern margin of the coalescent edges of the Cordilleran ice sheet in central Alberta and Montana. In Europe, human migration along the edges of the emerging coastlines of western Europe seems to have taken place close to the end of the Pleistocene.

Cause of global glaciation. The cause of global glaciation has long been debated. From the midnineteenth century to about 1970, a unique, nonrepetitive causation was looked for, such that ice ages were viewed as singular aberrations of unusual climatic conditions. It slowly became apparent that global climatic conditions were cyclic. Although other critical factors might intervene, the dominant causation of global glaciation must be climatic- and oceanic-based, tied directly to solar fluctuations in association with the Earth's solar orbit and tilt (**Fig. 2**). *See* EARTH ROTATION AND ORBITAL MOTION; INSOLATION; PRECESSION OF EQUINOXES.

As early as 1864, J. Croll had surmised that global cold periods leading to global glaciations were connected to the relationship between the Earth and the Sun. This idea, known as solar forcing, was expanded in the 1920s by M. Milankovitch. In the past several decades, these ideas have been refined and perfected to include the influence and impact of oceanic currents and global weather patterns in relation to Rossby waves and ocean surface temperatures (for example, El Niño), such that today a much sounder basis for understanding and potentially predicting global climate changes exists. However, the impact or onset of human-forced climatic global warming remains a central issue still debated, and its repercussions could be globally catastrophic.

Distribution of glaciers and ice sheets. The term global glaciation is perhaps imprecise. Although roughly synchronous major glaciation occurred in both the Northern and Southern hemispheres, ice never extended much beyond 40°N or 40°S latitude. The debate that surrounds the snowball Earth hypothesis suggests that ice, once north and south of 30°, developed a runaway feedback related to global albedo, causing a catastrophic glaciation that covered the whole Earth. Based upon existing geological evidence, it is likely that this hypothesis is inaccurate. *See* ALBEDO.

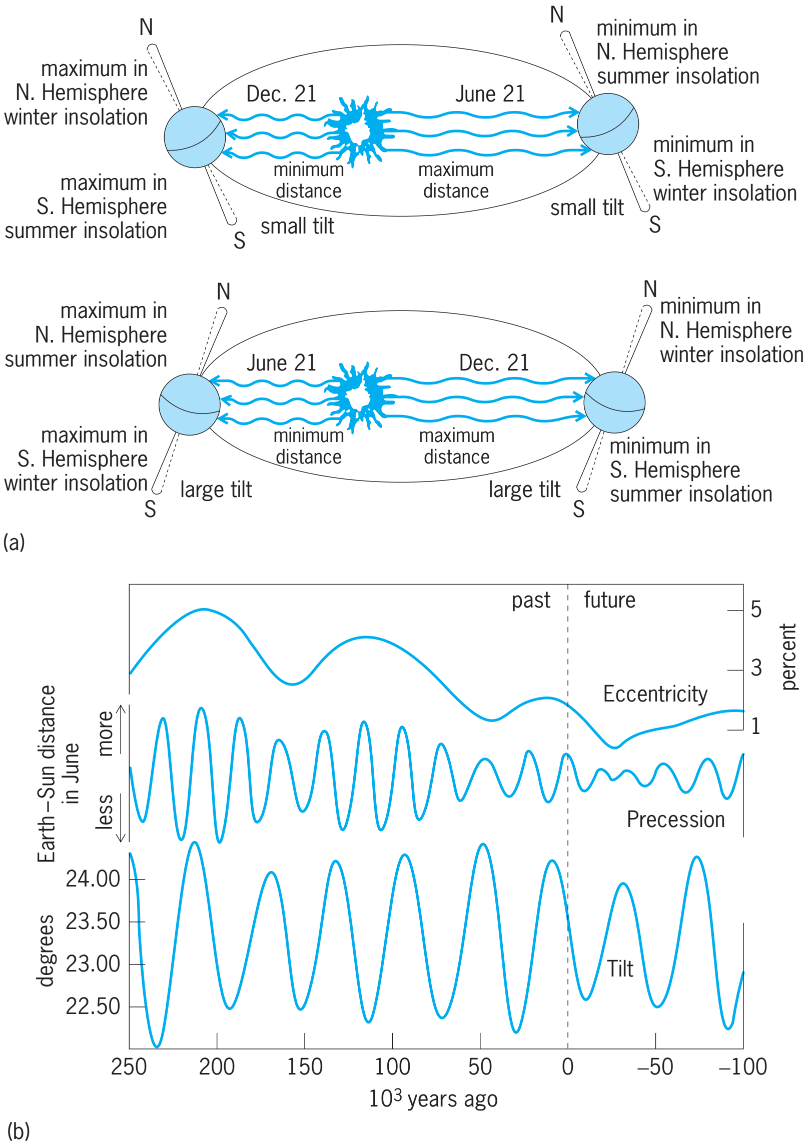


Fig. 2. **Models of the Earth's orbital elements: eccentricity, tilt, and precession. (*a*) Relationship of orbital elements in a calendar year (*adapted from W. F. Ruddiman and H. E. Wright, eds., North America and Adjacent Oceans During the Last Deglaciation, 1987*). (*b*) Changes in orbital elements over the past 205,000 years and 100,000 years into the future (*adapted from J. Imbrie and K. Imbrie, Ice Ages: Solving the Mystery, 1979*)**

The distribution of ice cover during the Pleistocene is shown in **Fig. 3**, where approximately 30% of the continents are covered. If the ocean basin areas that were covered by ice are also included, the portion would rise to around 60%. Glacier buildup, leading to the development of vast ice sheets, requires high snow precipitation, low summer melt, and the survival of snow cover from winter through the following summer to the next winter. As snow accumulates, the intense reflection off the snow (its albedo) acts to maintain a low-temperature surface, reducing snowmelt and maintaining snow cover. Therefore, snow accumulates in those areas of the Earth's surface where low temperatures prevail, such as high mountain areas, polar areas, and mid- to upper-latitude continental areas, such as central Canada and central Russia. In the Suthern Hemisphere, the landmass that lies beneath the Antarctic acted as a location for gradual snow accumulation. In the

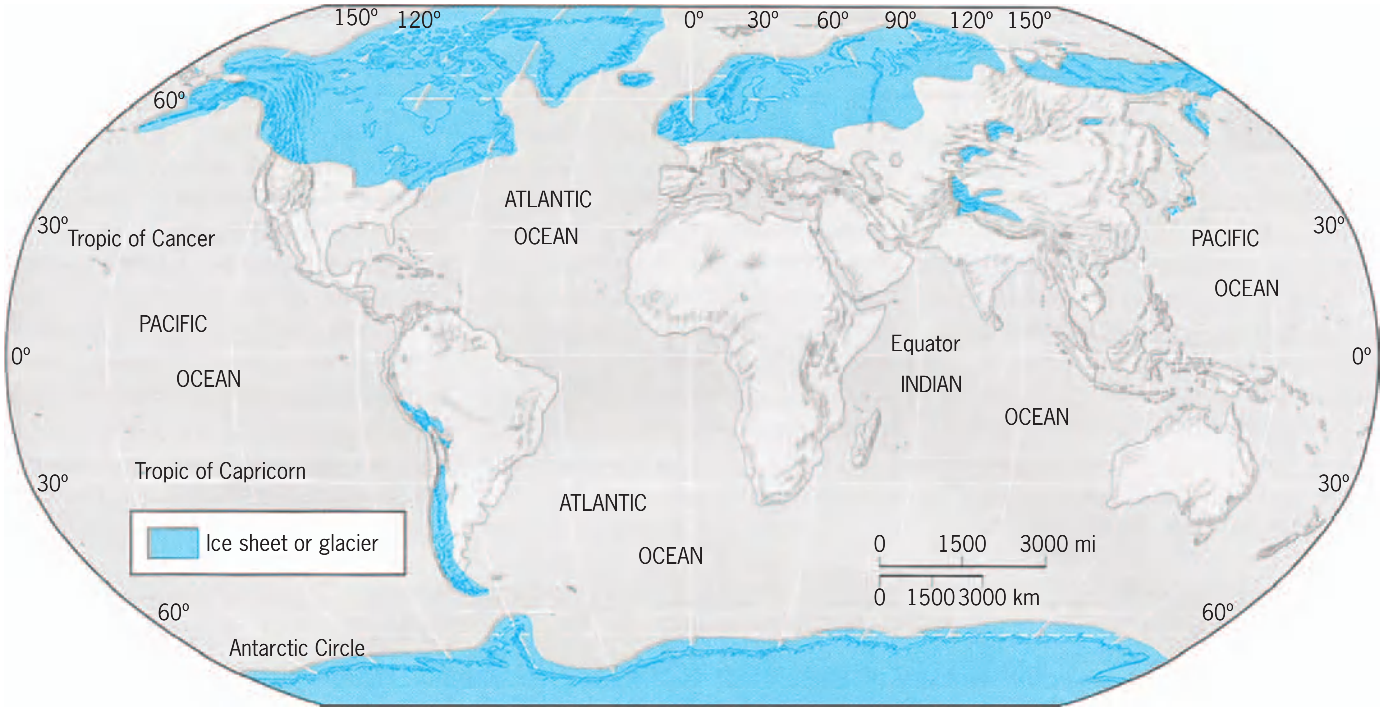


Fig. 3. Extent of worldwide glaciation approximately 18,000 years ago at the maximum of the late Wisconsinan. (*Modified from T. L. McKnight and D. Hess, Physical Geography, 6th ed., 2000*)

Northern Hemisphere with only the Arctic Ocean and no substantive landmass, snow accumulated on Greenland, the Canadian Arctic islands, central Canada, central Russia, and the Ural Mountains, as well as on high mountainous areas on all continents such as the Himalayas, the Pamirs and Tien Shan, the European Alps, the Carpathians, the Norwegian and Scottish Highlands, the Pyrenees, and the Rocky Mountains. In the Southern Hemisphere other than in Antarctica, glaciers developed in the Andes, the Southern Alps of New Zealand, the highlands of New Guinea, and the mountains of East Africa.

Evidence of glaciation. Terrestrial and marine evidence supports the glacier footprint in most parts of the world. Terrestrial clues are found in the following: specific sediment types indicative of glacial processes; the presence of erratic boulders far-traveled from a known source; distinctive landforms and landform suites (such as end moraine sequences, eskers, and drumlins); as well as the presence of buried organics, such as pollen grains, which can be traced to distinctive vegetation communities typical of arctic conditions. The presence of glacier ice on land or on fringing continental shelves is reflected in marine environments by specific sediment types with dropstones (carried by icebergs) and evidence of ice-rafted debris, as well as by the presence in stratigraphic sequences of suites of sediments containing distinctive organic remains indicative of varying terrestrial and sea temperature changes. *See* GLACIAL GEOLOGY AND LANDFORMS; MORAINE; PALYNOLOGY; POLLEN.

More recently, it has been possible to use sophisticated techniques, such as optical luminescence and oxygen isotope ratios, to date and climatically characterize specific sediment strata or horizons within sediment packages. There are instances where the specific origin of possible glacial sediment remains in doubt, even given the geological or stratigraphic context in which it is found. Considerable research remains to be done in differentiating glacial sediments both on land and in ocean basins. This problem is accentuated when attempting to characterize older lithified sediments that have suffered considerable geochemical and diagenetic changes. Glacial sediments, especially on land, often are geologically ephemeral, easily eroded or reworked, relatively thin and not widespread, and retain signatures that can easily be mistaken for sediments from other geological environments. *See* DIAGENESIS.

Pleistocene glaciations. It is now widely accepted that the Pleistocene began approximately 1.8 million years ago. It seems likely that the climate following the end of the Pliocene rapidly shifted to glacial conditions in both hemispheres. The onset of cold climatic conditions led to the buildup and the formation of the Laurentide and Cordilleran ice sheets covering most of the northern half of North America, and the European ice sheet expanded from the Alps into Germany and France. Meanwhile, the vast Fenno-Scandian ice sheet covered most of northern Europe, expanding into Poland, the Netherlands, and northern Germany. A vast ice sheet covered northern Russia, with a considerable extension across the Arctic Islands into the Arctic Ocean. In the Southern Hemisphere, the Antarctic ice sheet, which seems to have been in existence since the Oligocene, expanded into the Southern Ocean, while glaciers expanded and traveled to the coasts in South Island, New Zealand, and to the Antarctic and Pacific oceans

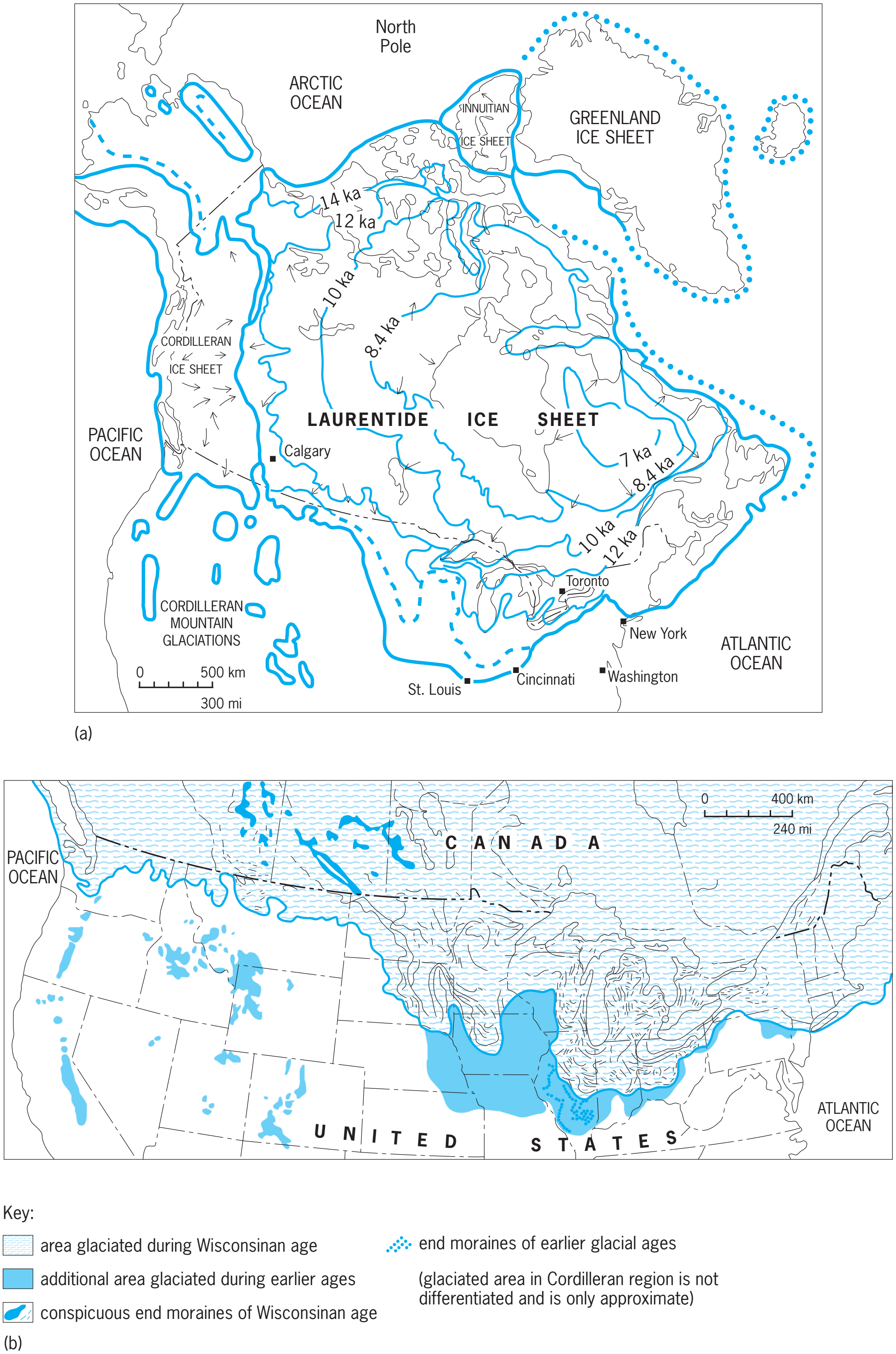
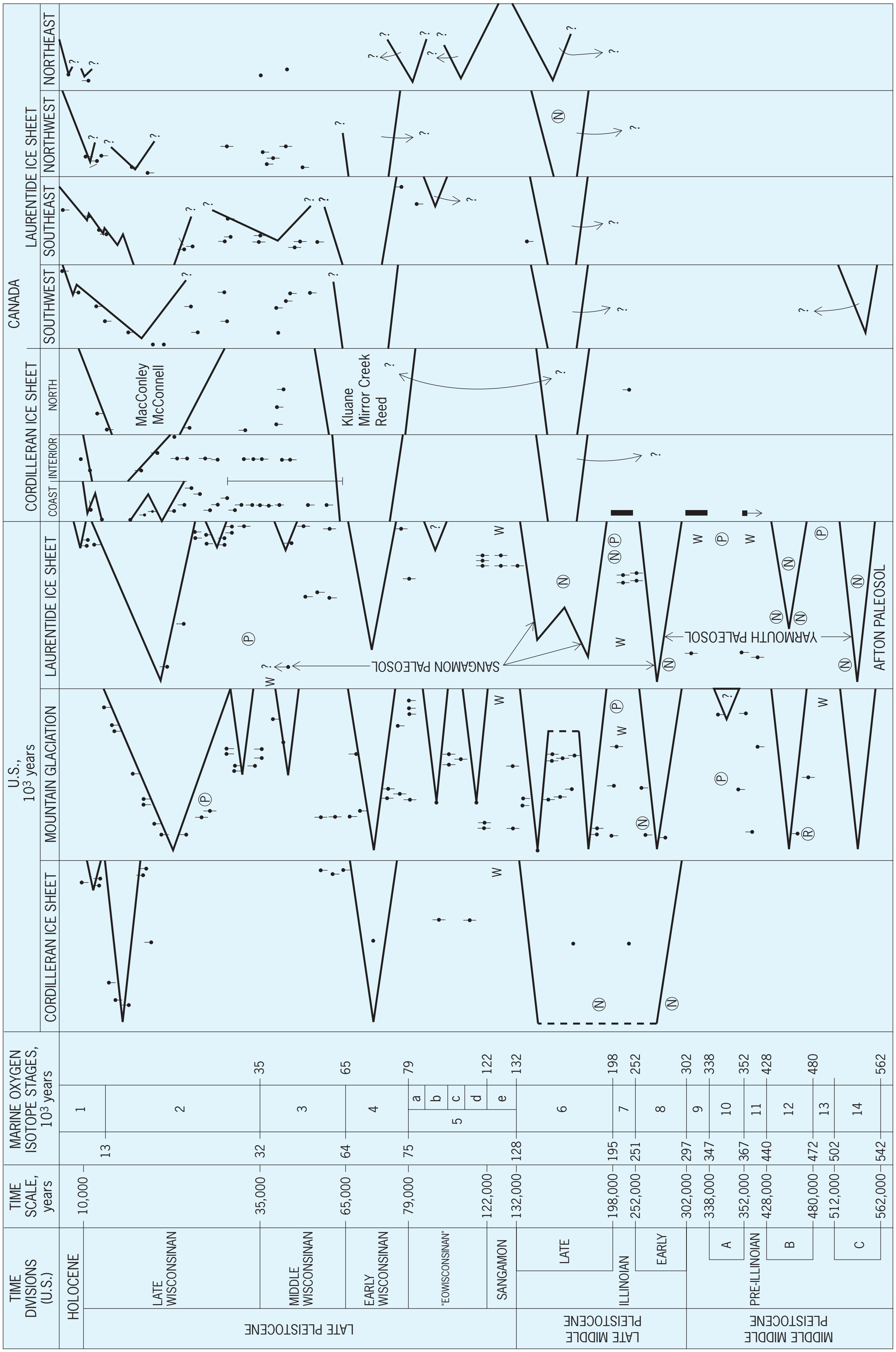


Fig. 4. Pleistocene glaciation. (*a*) Approximate maximum extent of glaciation and main ice sheets in North America in the late Wisconsinan 18,000 years ago and subsequent retreat margins to approximate position 7000 years ago; ka = 1000 years (*adapted from Fulton, 1989, and Clark, 1997*). (*b*) Extent of late Wisconsinan glaciation along the southern margin of the Laurentide and Cordilleran ice sheets 18,000 years ago (*modified from R. F. Flint, Glacial and Quaternary Geology, 1971*).



(continued on next page)

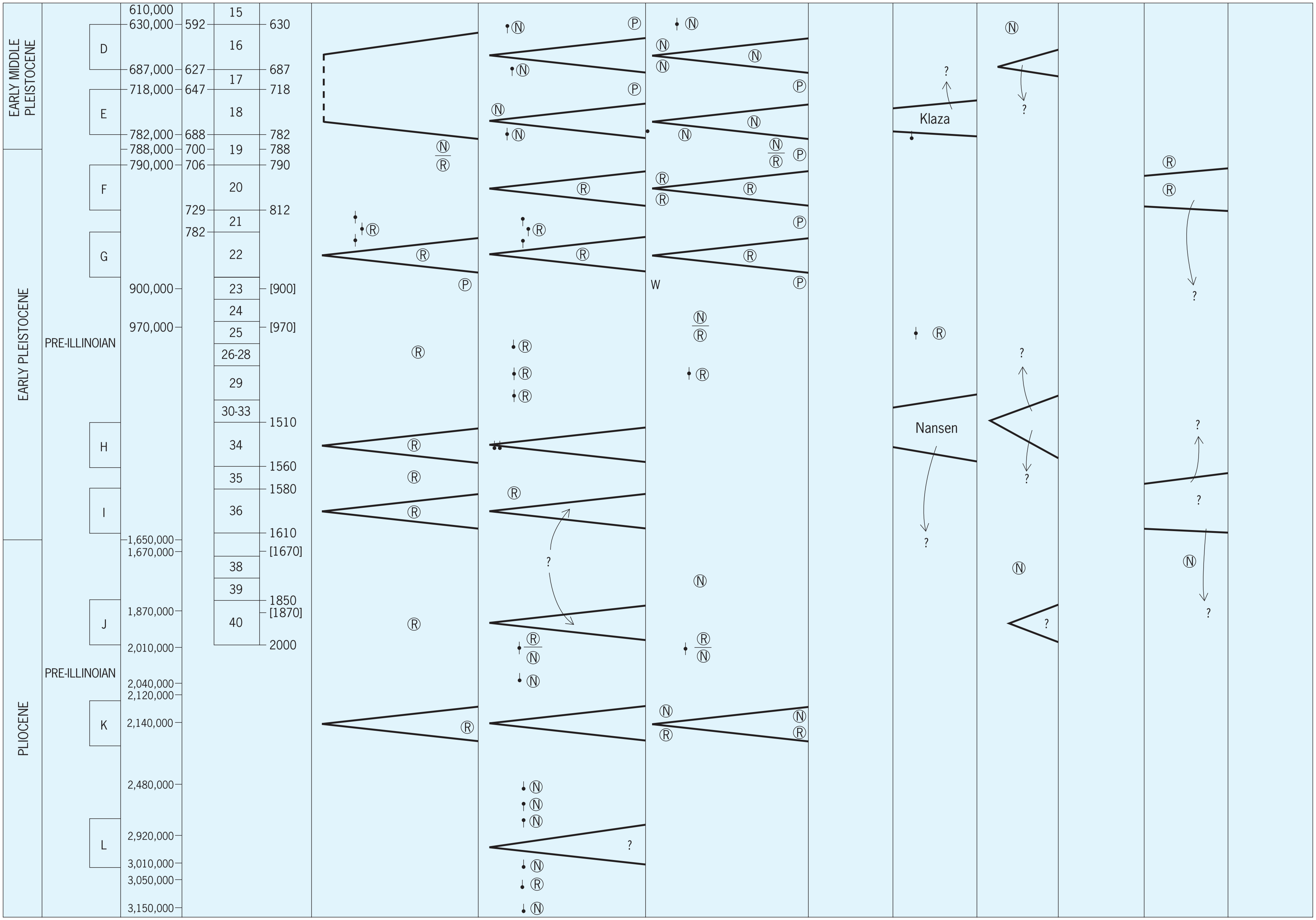


Fig. 5. Correlation chart of late Cenozoic (Pliocene and Pleistocene periods) glaciations in North America. (*Modified from Bowen et al., 1986*)

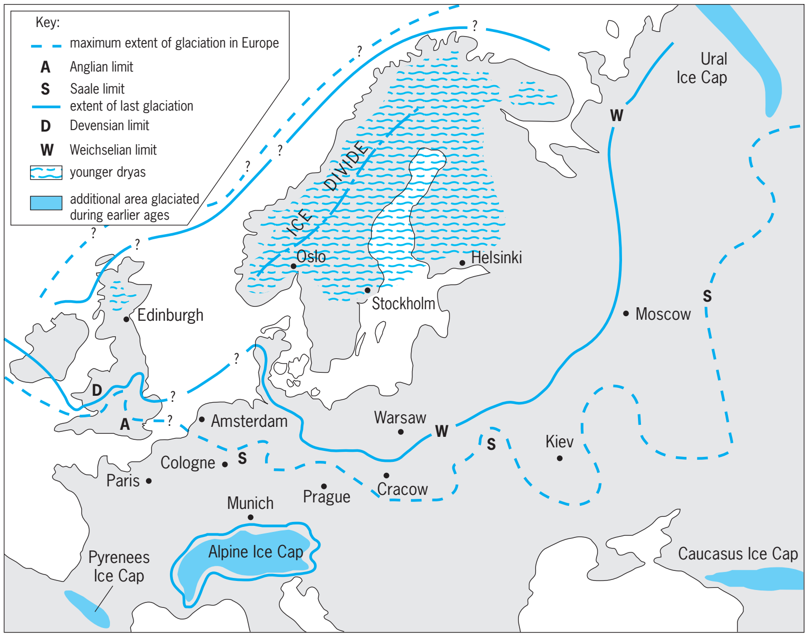


Fig. 6. Extent of late Pleistocene glaciations in Europe approximately 18,000 years ago. (*Modified from T. Nilsson, The Pleistocene: Geology and Life in the Quaternary Ice Age, 1983*)

in southern Patagonia in South America (Fig. 3). Elsewhere, smaller, independent ice sheets and ice fields covered much of the United Kingdom and Ireland, the highlands of New Guinea, the Sierra Nevadas in the western United States, other highlands in Central America, and the Andes from the tropics of Ecuador and Venezuela to the dry Andean Plateau of Bolivia and Peru. *See* OLIGOCENE; PLIOCENE.

The Pleistocene must be understood in the context of repeated phases of global glacier buildup, followed by major ice-sheet retreat and the resumption of warm interglacial times (100,000 years in length) in which climatic conditions returned to temperatures similar to today. At other times, ice-sheet retreat was much less dramatic, and interstadial conditions of cool, wet climates prevailed for short periods (10,000 years), followed quickly by ice-sheet advances (**Figs. 4** and **6**). In all the continents affected by the Pleistocene glaciations, similar advances/retreats and interglacial and interstadial phases generally occurred, but the specific details and timing of events differed not only among ice sheets but also along the various edges of the same ice sheet. At the time scale of thousands of years, there was a synchroneity. But at shorter periods, the idiosyncrasies of topographic variations, the local and regional fluctuations in ice discharge and mass balance, and local and regional basal glaciodynamic and regional weather patterns were such that synchroneity of glacial and nonglacial events, and of glacial pulses, retreats, surges, and rapid downwasting permutations caused huge variations in ice-front behavior.

The end of the Pleistocene occurred about 12,000 years ago, when the ice sheets began to melt, climate returned to conditions similar to today, sea level rapidly rose, and crustal rebound started. However, this approximate date needs considerable qualification since the ice sheet retreated from most of the northern United States somewhat earlier, following the maximum extension of the ice 18,000 years ago. In contrast, the resumption of warmer, nonglacial conditions did not begin in the northerly parts of Canada until 6000 years ago. Similar variations occurred in Europe and in the Southern Hemisphere.

With the end of the Pleistocene, the present interglacial period began. Considerable climatic fluctuations have continued to occur as witnessed by the Little Ice Age of the twelfth to eighteenth centuries, by the much warmer period of the Hypsithermal (climatic optimum) 9200–7300 years ago, and the possibly human-aided, global warming of the next few centuries. **Figure 7** shows the fluctuations in temperature and consequent ice-sheet advances for the past 150,000 years and the projections for the next 150,000 years, in which it is conjectured that following a period of global warming a sequence of glacial and interglacial periods will resume.

Pleistocene of North America. The major glaciations of North America can be divided into early, middle, and late Pleistocene. Within these time divisions, the major glacial periods of the Pre-Illinioan, Illinionan, and Wisconsinan can be subdivided according to those in the United States related to the Cordilleran ice sheet, mountain glaciation, and Laurentide ice sheet (Fig. 5). In Canada, a twofold subdivision of

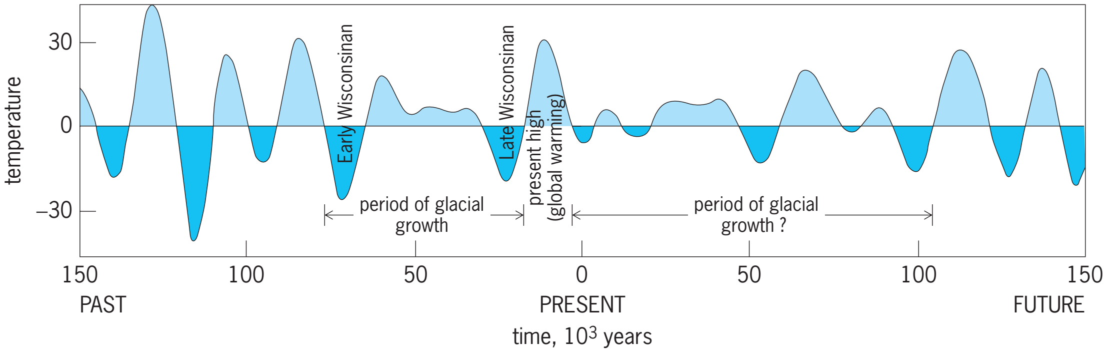


Fig. 7. Insolation curves for 45°N covering the past 150,000 years and the future 150,000 years. Past and future glaciations are shown below the zero degrees line. (*Adapted from W. S. Broecker and J. van Donk, J. Rev. Geophys. Space Phys., 8:167–198, 1970*)

Cordilleran and Laurentide ice-sheet glaciations can be distinguished.

Evidence based upon tills indicates that the Laurentide ice sheet had expanded into Iowa by the late Pliocene (glaciation K in Fig. 5). Glaciofluvial sediments in the western prairies of Canada are considered late Pliocene or early Pleistocene in age. From the early Pleistocene onward, a series of glacial advances and retreats can be traced throughout North America. The maximum extent of the last (late Wisconsinan) Laurentide ice sheet occurred along the southern margin between 24,000 and 14,000 years ago, while along the northern and eastern margins in Canada it occurred between 12,000 and 8000 years ago. The maximum extension of mountain glaciation seems to have occurred around 22,000 years ago, while the maximum extent of the Cordilleran ice sheet occurred 15,000 to 14,000 years ago. At the regional level in the Great Lakes basin, a complex sequence of events followed the retreat of the maximum ice-sheet cover approximately 18,000 years ago. Multiple lobes and ice streams formed and subdivided as the general northward retreat of the ice continued, with ice retreating from the Lake Ontario basin by around 11,000 years ago and from the north shore of Lake Superior by around 8000 years ago. *See* TILL.

The future. Predictions of future glacial periods can be made based on our past understanding of global climatic change and the tendency for the Earth's climate to swing between warmer and colder phases (Fig. 7). To date, the best climate models only suggest possible scenarios, typically based upon imperfect knowledge. All indications point to accelerated global warming, concomitant sea-level rise, and in the short time dramatic changes of climate and weather patterns, which some source; suggest are already occurring. Such changes over the next 50 to 100 years could lead to local and regional famine, coastal flooding, and socioeconomic and political upheaval at a scale never before witnessed. Glacial history is so wrapped up in the evolution and development of human civilization that it is crucial we understand past glacial history to predict potential global environmental change. John Menzies

Bibliography. R. M. Alley, *The Two-Mile Time Machine: Ice Cores, Abrupt Climate Change, and Our Future*, 2000; B. M. Fagan, *The Little Ice Age: How Climate Made History 1300-1850*, 2001; J. Grove, *The Little Ice Age*, 1988; J. Imbrie and K. Imbrie, *Ice Ages: Solving the Mystery*, 1979, reprint 2005; E. LeRoy LaDurie, *Times of Famine, Times of Feast: The History of Climate since the Year 1000*, 1971; P. A. Mayewski and F. White, *The Ice Chronicles: The Quest To Understand Global Climate Change*, 2003; J. Menzies (ed.), *Modern and Past Glacial Environments*, rev. student ed., 2002.

Glaciology

A broad field encompassing all aspects of the study of ice. While many glaciologists focus their attention on glaciers, the largest ice masses on Earth, glaciology also includes the study of ice that forms on rivers, lakes, and the sea; ice in the ground, including both permafrost and seasonal ice such as that which disrupts roads in the spring; and ice that crystallizes directly from the air on structures such as airplanes and antennas. All forms of snow research, including snow hydrology and avalanche forecasting, fall under the broad rubric of glaciology. Even planetary geologists are involved, as two of the moons of Jupiter, Ganymede and Callisto, are believed to be composed largely of ice. This article, however, will be restricted to discussion of glaciers.

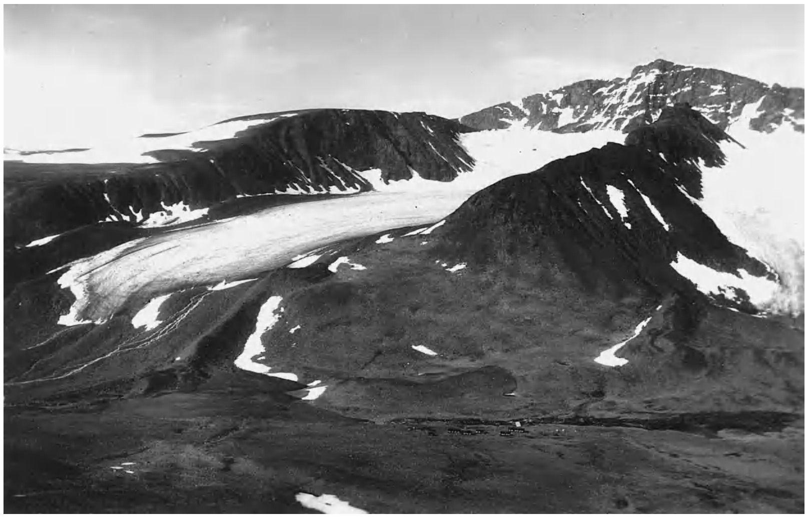


Fig. 1. Storglaciären, a small valley glacier in northern Sweden.

Classification of glaciers. Glaciers are classified principally on the basis of size, shape, and temperature. Cirque glaciers occupy spectacular steep-walled, overdeepened basins a few square kilometers (1 km^2 = 0.36 mi^2) in area, called cirques. Most cirques are in high mountain areas that have been repeatedly inundated by ice. The cirques and the deep valleys leading away from them were, in fact, eroded by larger glaciers over the past 3 million years. *See* CIRQUE.

As a cirque glacier expands, it is usually constrained, at least initially, to move down such a valley. It then becomes a valley glacier (**Fig. 1**). Where such a valley ends in a deep fiord in the sea, the glacier is called a tidewater glacier. *See* FIORD.

In contrast, some glaciers are situated on relatively flat topography. Such glaciers can spread out in all directions from a central dome. When small, on the order of a few tens of kilometers across, these are called ice caps. Large ones, like those in Antarctica and Greenland, are ice sheets. *See* ANTARCTICA; ICE FIELD.

Thermally, glaciers are usually classified as either temperate or polar. In the simplest terms, a temperate glacier is one that is at the melting point throughout. The term melting point is used in this context rather than 0°C (32°F), because the temperature at which ice melts decreases as the pressure increases. Thus, the temperature at the base of a temperate glacier that is 500 m (1700 ft) thick will be about −0.4°C (31.2°F), but if heat energy is added to the

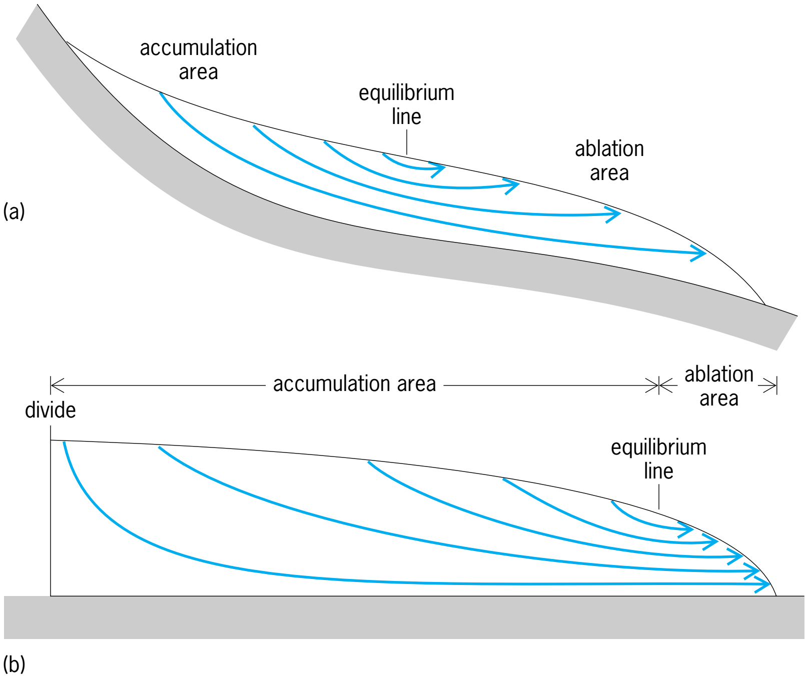


Fig. 2. Longitudinal profiles of (*a*) a valley glacier and (*b*) a continental ice sheet. The equilibrium line on the valley glacier is near the middle of the glacier, while that on the ice sheet is much nearer the margin.

ice, it can melt without an increase in temperature. Most valley glaciers are temperate.

In polar glaciers the temperature is below the melting point nearly everywhere. The temperature of a polar glacier increases with depth, however, because the deeper ice is warmed by heat escaping from within the Earth and by frictional heat generated by deformation of the ice. Thus, at its base, a polar glacier may be frozen to the substrate or may be at the melting point. Ice caps and ice sheets are normally polar, as are some valley glaciers in high latitudes.

As was the case with the classification based on size and shape, there is a continuum of thermal regimes in glaciers. The most common intermediate type has a surficial layer of cold ice, a few tens of meters (1 m = 3.28 ft) thick in its lower reaches, but is temperate elsewhere. Such glaciers are sometimes called subpolar or polythermal.

Glacier growth and mass balance. Glaciers exist because there are places where the climate is so cold that some or all of the winter snow does not melt during the following summer. The next winter's snow then buries that remaining from the previous winter, and over a period of years a thick snow pack or snowfield develops. Deep in such a snow pack the snow is compacted by the weight of the overlying snow. In addition, evaporation of water molecules from the tips of snowflakes and condensation of this water in intervening hollows results in rounding of grains. These processes of compaction and metamorphism gradually transform the deeper snow, normally known as firn, into ice. Melt water percolating downward into this firn may refreeze, accelerating the transformation. *See* SNOWFIELD AND NÉVÉ.

Ice is a mineral, like quartz or diamond. It has a well-defined crystal structure consisting of sheets of hexagonal rings stacked one upon the other so that individual hexagons lie one above another. These stacks form a tube. In the science of crystallography, the direction parallel to the axis of this tube is called the *c* axis of the mineral. *See* CRYSTALLOGRAPHY.

Crystal structures, natural or otherwise, are never perfect. Crystal lattices are distorted by imperfections called dislocations. When stressed, these dislocations move through a crystal by breaking and reforming one atomic bond at a time, resulting in deformation. This is why glaciers can flow. Were crystals perfect, many such bonds would have to break and reform simultaneously in order for deformation to occur, and this would require vastly higher stresses.

Ice does not deform significantly if the stresses are purely hydrostatic, as in a lake. However, the surfaces of snowfields are sloping, and this results in nonhydrostatic forces. These forces push the ice toward areas where the surface elevation is lower. Because the ice is weakened by dislocations, it responds to this force by flowing. When flow rates exceed a few meters per year, the ice mass is properly called a glacier.

Glaciers obviously flow downhill. Not so obvious, however, is the fact that here the term "hill" is defined by the slope of the glacier surface, not by the slope of the bed. Glaciers can in fact, flow up along a bed that is sloping in a direction opposite to that of the glacier surface.

Places where winter snowfall exceeds summer melt, which typically occurs either at high altitudes or at high latitudes, serve as accumulation areas for glaciers (**Fig. 2**). At high altitudes, the ice flows down valleys to lower elevations where temperatures are warmer (Fig. 2*a*). The part of a glacier lying in an area where summer temperatures are high enough to melt some of the ice flowing in from higher elevations in addition to all of the previous winter's snow is called the ablation area. In the case of ice caps or ice sheets at high latitudes, such as in Greenland, the ice either flows to the ocean where it breaks off readily and floats away (a process called calving), or it flows to either lower elevations or more southerly latitudes where temperatures are again high enough to melt both the winter snow and some of the inflowing ice (Fig. 2*b*). *See* ICEBERG.

The boundary between the ablation area and the accumulation area at the end of a melt season is called the equilibrium line (Fig. 2). The position of the equilibrium line for a given year can be determined by locating the upper limit of extensive bare ice just before the first major autumn snowfall. The altitude of the equilibrium line will be higher during warm years or years with little winter snowfall, and lower during cold years or years with a lot of snow fall. *See* SNOW LINE.

When, during a given year, the mass of snow added in the accumulation area of a glacier exceeds the mass of ice lost from the ablation area, the glacier is said to have had a positive mass balance. If such a situation persists for several years, the glacier will advance to lower elevations or more temperate latitudes, thus increasing the size of its ablation area

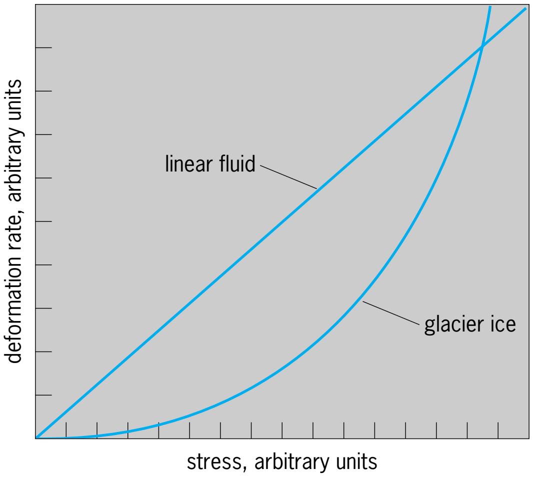


Fig. 3. Plots of deformation rate versus stress for a nonlinear substance such as glacier ice and for a linear fluid such as water.

and the mass loss. Conversely, persistent negative mass balances lead to retreat. Contrary to one implication of the word retreat, a retreating glacier does not flow backward. Rather, a glacier retreats when the ice flow toward the terminus is less than the melt rate at the terminus.

Because of time lags in a glacier flow system, valley glaciers may take decades, and continental ice sheets millennia, to respond to a change in mass balance. For example, studies of cores from the Greenland ice sheet suggest that the increase in temperature that precipitated the end of the last ice age occurred 11,600 years ago over a period of only a few years. However, the Laurentide ice sheet that covered the northern half of North America persisted until approximately 8000 years ago.

Flow of ice. Ice is a nonlinear material. Its deformation rate, $\dot{\epsilon}$, is proportional to the cube of the stress, τ, thus $\dot{\epsilon} = (\tau/B)^3$, where B is a measure of the viscosity of the ice. The cubic dependence means that at higher stresses the deformation rate increases very rapidly with any increase in stress. This behavior is distinct from that of linear fluids such as water (**Fig. 3**). *See* VISCOSITY.

Although this relation appears simple, it is not. In order to calculate τ, for example, one has to know all of the stresses acting at any given point in a glacier, and there are nine of them—three normal and six shear. (Fortunately, three of the shear stresses equal the other three, which simplifies the problem slightly.) The same applies to $\dot{\epsilon}$. In addition, B is not constant; it decreases as the temperature increases, so warm ice deforms more readily than cold ice. In ice near the melting point, viscosity also decreases with increasing interstitial water content. The water forms films along crystal boundaries, and this enables crystals to slip past one another. Finally, the viscosity is sensitive to the orientation of crystals in the ice. The reason for this is that ice deforms about 10 times more easily when the *c* axes are perpendicular to the plane of shear instead of parallel to it. Once the ice has undergone a small amount of deformation, crystals that are favorably oriented for deformation tend to grow at the expense of others. Thus, in ice that has been deformed appreciably, the viscosity tends to be lower than in undeformed ice.

Owing to these complications, sophisticated computer models are necessary for the most accurate calculations of glacier flow. Under suitable circumstances, however, approximations can be made that lead to simple but useful relations. One such result predicts that the flow rate of ice should be proportional to the fourth power of its thickness. Thus, a 10% decrease in thickness results in about a 35% decrease in flow rate. Such a decrease in flow rate would reduce the input of ice to the lower reaches of a glacier appreciably. Unless the ablation rate were to decrease by a similar amount, the glacier terminus would retreat significantly. This is the reason for the frequently observed evidence indicating that a relatively modest decrease in thickness of a glacier was associated with a substantial retreat of the snout. For example, Storglácieren (Fig. 1) has retreated several hundred meters from its 1915 maximum position, but has only thinned a few tens of meters.

Sliding and till deformation. Both temperate glaciers and polar glaciers that are at the melting point at the base are able to move over the bed by sliding. In addition, a glacier resting on a bed of unconsolidated material often drags this material along. This material is commonly a mixture of clay, silt, sand, and gravel formed by glacier erosion and known as till. The thickness of the deforming till layer may range from centimeters to a few meters. *See* TILL.

Sliding can occur over a deforming substrate or over bedrock. Sliding speeds are normally lower over bedrock because the roughness of the surface is greater.

Two basic processes are involved in sliding: regelation and plastic flow. In regelation, ice melts on the upglacier sides of bumps, where the pressure is high and the melting point is depressed. The meltwater flows past the bump and refreezes in its lee, where the pressure is lower and the temperature is higher. The latent heat of freezing is conducted from the warmer lee side to the colder stoss or upglacier side where it melts more ice.

Plastic flow is simply the process whereby ice deforms around a bump. During plastic flow, ice may separate from the bed in the lee of a bump, leaving a cavity that is generally filled with water.

Regelation is most important on small bumps, because the path along which heat is conducted from the lee to the stoss side of the bump is shorter and the temperature gradient is higher. Plastic flow is most important on large bumps, because the stress scales with the size of the bump and the flow rate is proportional to the cube of the stress. The two processes are of roughly equal importance for bumps that are about 2–8 cm (1–3 in.) high and a meter or so in length.

Foliation. Glaciers commonly appear to be banded (Fig. 1). This appearance, a result of intercalation of layers of ice with differing bubble and dirt contents, is known as foliation. Variations in bubble content commonly arise from the vagaries of snow accumulation and meltwater penetration, and are thus subparallel to the original layering of the snow in the accumulation area. In addition, in areas where stresses in the ice near the surface are tensile, fractures called crevasses may form. Crevasses often become filled with snow in such a way that the filling has either a higher or lower bubble content than the surrounding ice. Variations in dirt content are most common in ice near the base of a glacier, and result from episodic entrainment of the debris.

By the time ice reaches the snout of a glacier, these layers have been stretched to such an extent that they are commonly less than 1% of their original thickness, and even layers that were once nearly vertical, such as crevasse fillings, have been rotated so that they are subparallel to the bed or valley sides. This is particularly true of ice that has moved near the bed or sides. It is then usually impossible to determine the origins of variations in bubble content, but the variations themselves are still readily visible.

The planes separating layers of ice with differing bubble or dirt content have been referred to as shear planes, but this is not accurate. They may be subparallel to the plane of maximum shear strain rate, particularly near the bed, but they normally are not surfaces of localized intensive shear.

Water movement. In temperate ice there are veins along the intersections of each three neighboring ice crystals. Water can move through these veins and in so doing dissipate viscous or frictional heat. This heat melts ice, thus enlarging the veins.

Larger veins carry more water per unit of wall area, and they are thereby enlarged at the expense of smaller ones. At depths of several meters, tubes a few millimeters (25 mm = 1 in.) in diameter may be found which join downward to make successively larger conduits, forming an arborescent (treelike) network. Trunk conduits in the network eventually reach the bed. Thence, some of the water may percolate into the substrate, but most flows as streams along the bed, either in conduits melted upward into the ice or, more rarely, in channels cut downward into the bedrock.

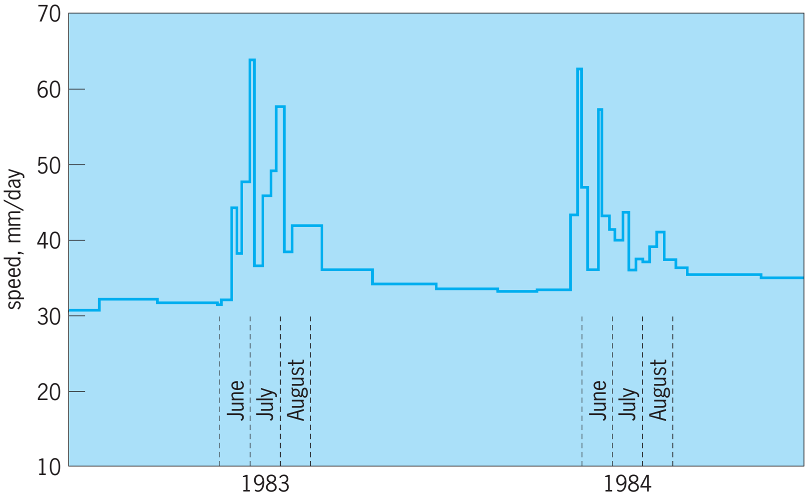


Fig. 4. Speed of a stake planted in Storglaciären (Fig. 1), over a 2-year time span.

Despite the vein system, ice near the surface of a glacier is not very permeable. However, crevasses can reach depths of a few tens of meters, and they thus give surface water access to the millimeter-scale and larger conduits at these depths. Although, initially, these conduits may not be large enough to carry all of the water tumbling into the crevasse, the viscous heat dissipated in the conduits rapidly enlarges them.

Crevasses move with the ice, and may eventually reach a place where stresses at the surface are compressive. The crevasse then closes, but the water easily maintains an opening, called a moulin, in the glacier surface.

The pressure at depth in a glacier is largely dependent on the ice thickness. However, the water pressure in conduits, although variable, is almost always lower than that in the ice. Conduits are thus constantly being squeezed closed by plastic flow of the ice. This closure is offset by melting of conduit walls by viscous heat. The equilibrium conduit size for any given water discharge is that at which closure and melting are balanced. If closure is too fast, the conduit becomes constricted and the pressure builds up in it, thus decreasing the rate of closure.

Because of the pressure distribution at the bed, subglacial streams do not, in general, follow topographic valleys in the bed. Rather, they flow along valley sides, flow uphill, and even cross ridges. Where melting of debris-laden ice forming the sides of subglacial conduits delivers to the streams more sediment than they can carry, sinuous ridges of gravel, called eskers, are deposited. Major esker systems in places such as Maine and Minnesota reveal the courses of vanished subglacial streams. *See* ESKER.

Water flow into a glacier varies on time scales of hours to months. It is highest in the summer, when snow and ice are melting and rainstorms provide additional flow. This is when the conduit system is largest. In the autumn and winter, when water inputs to the glacier decline, often becoming negligible, conduits close. Thus in the spring, when water fluxes increase again, conduits are small and the drainage system is deranged. Water pressures may then build up to high levels for a few days.

Water at a glacier bed normally has access to hundreds of cavities. When water pressures are high, each of these cavities becomes a hydraulic jack. The water pushes upglacier against the rock and down-glacier against the roof of the cavity. This additional downglacier force makes the glacier slide faster (**Fig. 4**). Such accelerations are likely to be most dramatic early in the melt season before the conduit system adjusts to renewed water input. However, they can also occur in the middle of the melt season when warm weather or rain increases the water input. Smaller accelerations may also occur on a diurnal time scale, particularly on sunny days, in response to the peak in melt that occurs in the early afternoon.

Glacier surges. Under certain rather rare circumstances, the high subglacial water pressures that develop in the spring do not dissipate quickly but persist for weeks. This occurs under glaciers that have been thickening for several years or decades but have not advanced appreciably as a result of the thickening. On these occasions, the increase in sliding speed resulting from the increased water pressure inhibits development of an integrated subglacial conduit system, so water pressures remain high. The glacier then may advance at speeds of meters to tens of meters per day, in what is known as a surge.

During a surge, ice that has been accumulating on higher parts of the glacier is transferred to lower areas. The upper part of the glacier becomes thinner and the surface slope decreases. Eventually, a conduit system becomes established, the water pressure drops, and the surge ends. The cycle then starts over again, with years or decades of gradual thickening as the glacier builds toward another surge.

Ice streams. Flow as rapid as that in surging glaciers is also found in some parts of the Antarctic Ice Sheet. These high flow rates occur in streams of ice tens of kilometers wide and hundreds of kilometers long, and are sustained for centuries. These streams are bounded not by valley sides but by ice that is moving much more slowly. Ice Stream B, for example, which drains to the Ross Ice Shelf in West Antarctica, has a maximum speed of 825 m/yr, while ice on either side of it is moving at only 10–20 m/yr. Roughly parallel to Ice Stream B are four other ice streams, some of which are presently inactive. One of these, Ice Stream C, shut down abruptly about a century ago. The high speeds of these ice streams are attributed to slippery conditions at the bed, where high water pressures reduce friction between the ice and the bed. Changes in paths of water flow at the bed are believed to be responsible for the changes in ice stream activity. If rising sea level eventually destabilizes the Ross Ice Shelf, discharge of ice through these ice streams could lead to rapid drawdown of the West Antarctic Ice Sheet, thus accelerating the rise in sea level.

Jökulhlaups. Among the hazards associated with glaciers are jökulhlaups, or sudden releases of water from lakes dammed by glaciers. Jökulhlaup is an Icelandic word; it has entered the vocabulary of geology because such floods are common in Iceland where localized volcanic heat is responsible for the presence of deep lakes surrounded by ice. In other regions, the lakes are more commonly formed where a glacier in a trunk valley extends across the open mouth of a tributary valley.

The basic cause of jökulhlaups is easy to visualize. Ice floats in water, so when such a lake becomes sufficiently deep the water forces its way out beneath the ice dam. However, some jökulhlaups occur before the water actually becomes deep enough to float the ice.

Once water begins to escape under such an ice dam, the conduit may be enlarged rapidly, resulting in a catastrophic release and total draining of the lake. This is most likely where the surface area of the lake is large, as the lake level then remains high while the conduit is being enlarged. High lake levels result in high water pressures in the conduits, and this inhibits closure of the conduit by plastic flow. Conduits thus can become quite large. When the surface area of the lake is small, water pressures in the developing conduit may drop quickly as the water level falls, and the conduit may squeeze shut before the lake drains completely.

Among the largest floods in the geologic record are those resulting from repeated catastrophic draining of an ice-dammed lake in Montana, Lake Missoula, during the ice ages. About a third of eastern Washington was covered by these floods, which had discharges estimated to have exceeded 17×10^6 m^3/s (4.5×10^9 gal/s). Smaller jökulhlaups have occurred during recorded history in places such as Alaska, Greenland, Norway, the Alps, and the Himalayas, and in many cases they have been the cause of property damage and loss of life.

Glacier erosion. Despite the fact that ice is relatively soft, glaciers are able to erode rock as hard as granite. The two principal erosional processes are abrasion and quarrying. Abrasion occurs when rocks in the sole of a glacier are forced into contact with the bedrock over which the ice is moving, much as sandpaper may be forced into contact with a board. The contact force depends not on the thickness of the ice but on the rate of melting by geothermal and frictional heat at the bed. In the absence of such melting, a rock in the sole of the glacier would simply be forced upward into the ice. However, when melting occurs, ice flows around the rock toward the bed to replenish that lost. The drag thus exerted on the rock forces it into intense contact with the bed.

Bedrock surfaces that have been abraded are smooth on a scale of decimeters. The surface generally appears polished, but deep scratches, formed by larger stones, mar the polish. These scratches are called striations.

Quarrying is a more complicated process. It involves fracturing of the bedrock and entrainment of blocks of rock isolated by these fractures. At first it was questioned whether a glacier could break solid rock. However, it is now believed that repeated water-pressure fluctuations in subglacial cavities cause propagation of microcracks that are present in any apparently solid rock, and that these microcracks eventually join to isolate blocks. Propagation of the microcracks is a result of fatiguing, much as a wire bent repeatedly in the same place will break.

Entrainment is also facilitated by water-pressure fluctuations. Increases in water pressure in cavities transfer some of the weight of the glacier from bedrock bumps to the water. Pressure-release freezing may then occur on top of the bump. This increases the traction that the glacier exerts on the rock.

Water-pressure fluctuations are likely to be highest and most frequent in areas of crevassing, where water has ready access to the glacial drainage system. One such area of water input is a crevasse called the bergschrund that forms at the head of the glacier (**Fig. 5**). Another site is where convexities in the

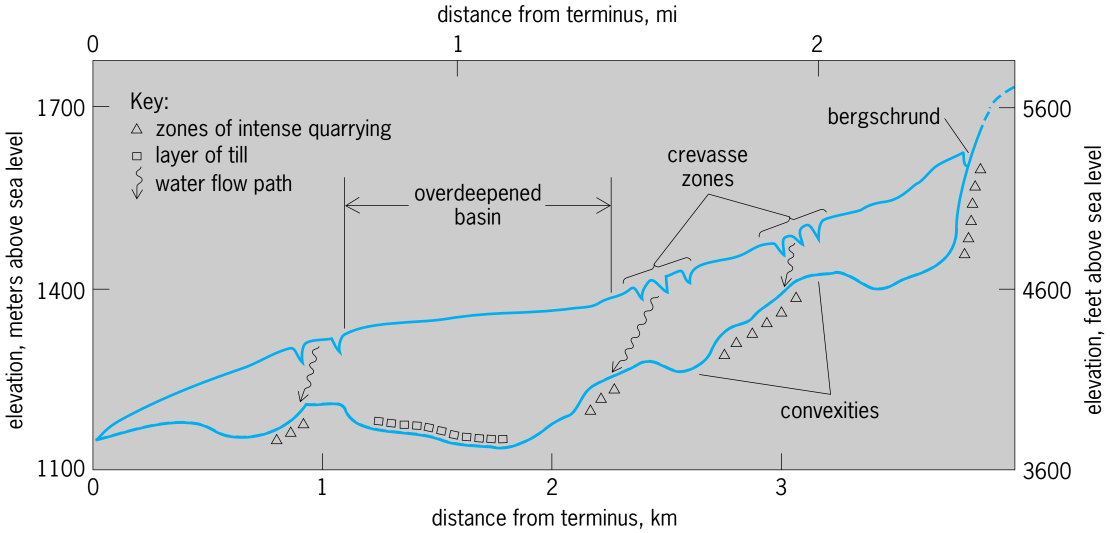

distance from terminus, mi
0
1
2
1700
1400
1100
elevation, meters above sea level
5600
4600
3600
elevation, feet above sea level
Key:
△ zones of intense quarrying
□ layer of till
water flow path
overdeepened basin
crevasse zones
bergschrund
convexities
0
1
2
3
distance from terminus, km

Fig. 5. Longitudinal profile of Storglaciären (Fig. 1) showing relation of crevasses to convexities in the glacier bed and to zones of intense quarrying. (*After R. LeB. Hooke, Positive feedbacks associated with the erosion of glacial cirques and overdeepenings, Geol. Soc. Amer. Bull., 103:1104–1108, 1991*)

glacier bed lead to tensile stresses and hence crevassing at the surface. Thus, the areas of most intense quarrying are the cirque headwall and the steep lee faces of these convexities. In the latter case, there is a positive feedback loop: the convexity causes crevassing that admits water to the lee face, where further quarrying accentuates the convexity.

Because it is localized on and thus accentuates steep faces, quarrying is responsible for some of the most dramatic attributes of glaciated valleys, and hence for the spectacular scenery that is associated with glaciated landscapes in mountainous areas (**Fig. 6**).

Water flowing along a glacier bed from areas of thick ice to areas of thinner ice must warm up in order to remain at the melting point. Some of the viscous energy dissipated by this water is consumed in warming it. When water flows up an adverse slope leading out of an overdeepened basin, the energy available may not be adequate to maintain the water at the melting point. Freezing then occurs, constricting the subglacial conduits and forcing the water to follow englacial (inside the glacier) paths. In the absence of subglacial water flow, the products of glacier erosion are not flushed out, but accumulate to form a protective layer that inhibits further erosion of the bed. For this reason, convexities are not scoured away, but become amplified by the erosional processes. *See* GLACIATED TERRAIN.

Fig. 6. Glaciated mountain landscape.

Glacier deposition. The load of rock debris that glaciers acquire by quarrying and abrasion is eventually deposited when the ice melts. Among the most important types of deposits are moraines. Moraines are ridges of debris deposited directly from the ice at the glacier margin (Fig. 1). They may be only a few meters in width and height or reach several kilometers in width and tens of meters in height. Because water has played little role in their deposition, they are typically heterogeneous, with abundant silt- and clay-sized material.

The time required for the formation of large moraines ranges from decades to centuries. They mark positions where the ice margin was stable for a long period of time. Where they are associated with organic material that can be dated by radiocarbon techniques, they are important in reconstructing the glacial history of an area; thus moraines are valuable in paleoclimatic studies. *See* DATING METHODS; MORAINE; PALEOCLIMATOLOGY.

Climatic record in glaciers. In inland areas of the Antarctic and Greenland ice sheets, it is so cold that there is no melting, even in summer. The snow that is deposited there, together with any impurities from the atmosphere, is compacted into ice with little or no chemical alteration. Accordingly, it preserves a record of conditions at the time of deposition. For this reason, much effort has been focused on obtaining continuous cores through these ice sheets.

One of the most important paleoclimatic signals in glacier ice is provided by the ratio of the heavy isotope of oxygen (^{18}O) to the normal isotope, (^{16}O). Snow deposited during warmer periods contains more ^{18}O. By making detailed measurements of the ratio of these two isotopes, the difference between snow deposited in winter and that deposited in summer can be detected. Thus, much as trees can be dated by counting rings, scientists can date glacier ice that is up to several thousand years old. Using isotopes, it is easy to detect the onset and termination of ice ages (**Fig. 7**).

Recently, traces of the variations in bubble content and ice texture in annual layers have been found preserved in ice up to several thousand years old. This has enabled scientists to study the rate of climate change at the end of the last ice age. As the climate warmed, the accumulation rate on the Greenland Ice Sheet increased by a factor of nearly two, because the warm air held more moisture. Melting at the ice sheet margins, however, more than balanced the increase in accumulation and resulting increase in flow, so the ice sheet retreated.

This change in climate occurred over a period of about 3 years. An abrupt change of this magnitude today would wreak havoc with agriculture throughout the world.

Air bubbles that are trapped when snow is compacted into ice contain a record of the composition of the atmosphere. Among the constituents of such air bubbles is carbon dioxide (CO_2). By knowing the age of the ice from the $^{18}O/^{16}O$ ratio, a record of variations in atmospheric CO_2 can be reconstructed (Fig. 7). From this, it is clear that high atmospheric CO_2 concentrations are associated with warm periods in the past, and conversely. The postindustrial increase in CO_2 concentration in the atmosphere from about 280 ppm to about 360 ppm is as large as the change at the end of the last ice age, about 12,000 years ago. If this recent increase in CO_2 concentration leads to an increase in worldwide temperature comparable to that at the end of the ice age, glaciers will retreat, leading to a rise in sea level of 1–2 m (3–6 ft) in the twenty-first century. This, in conjunction with shifts in agricultural belts and other changes, could lead to major dislocations of human activities.

Another trace chemical present in ice is lead (Fig. 7). Variations in the concentration of lead in cores from Greenland indicate that lead is about 10 times more abundant in the atmosphere today than it was before humans learned how to use it. Such findings, in conjunction with medical research into health hazards posed by lead, have resulted in laws limiting its use in products such as paint and gasoline. *See* CLIMATE HISTORY.

Glaciers of the ice ages. Glacier ice contains more ^{16}O than does seawater. When moisture from the oceans begins to accumulate as snow on the land, leading to growth of ice sheets, the $^{18}O/^{16}O$ ratio in the ocean increases. The $^{18}O/^{16}O$ ratios in shells of microscopic animals that live in the oceans reflect this change. Thus, through study of cores of sediment from the ocean floor, a record of the advance and retreat of continental ice sheets can be reconstructed.

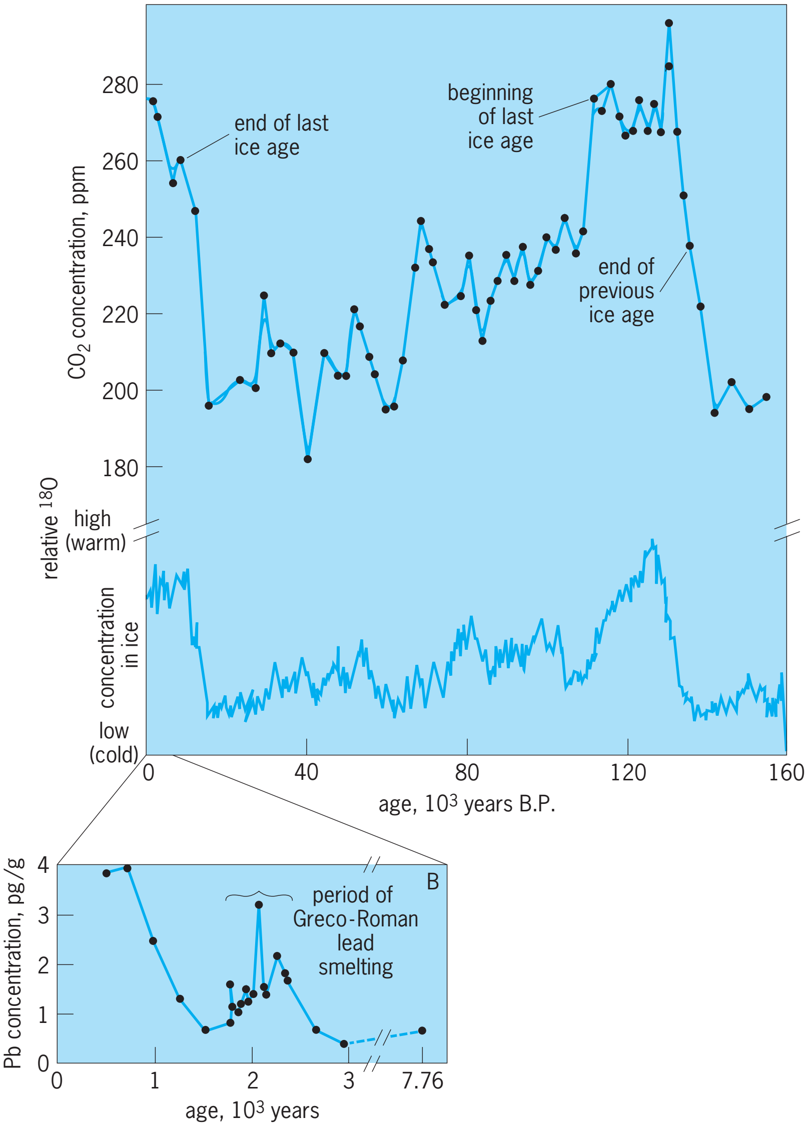


Fig. 7. Plot of $^{18}O/^{16}O$ ratio and of CO_2 and Pb variations with time in ice cores. The broad peak in lead concentration about 2000 years ago coincides with the period of maximum production of lead by the Roman Empire.

This record reveals that over the past 3 million years situations like that at present, with relatively little ice outside the polar areas, are exceptional. Such comparatively ice-free periods tend to last about 10,000 years, the length of time since the end of the last ice age. Then, in response to well-known and predictable changes in the Earth's orbit that lead to cooler summers in the Northern Hemisphere, ice sheets begin to advance. The advance, punctuated by partial retreats, culminates after about 90,000 years. When roughly half of North America and Europe is again covered by ice, it will be difficult for the world of the year 90,000 to support a human population like the present one. *See* GLACIAL EPOCH; GLACIAL GEOLOGY.

Roger LeB. Hooke

Bibliography. D. I. Benn and D. J. A. Evans, *Glaciers and Glaciation*, 1997; M. R. Bennett and N. F. Glasser, *Glacial Geology: Ice Sheets and Landforms*, 1996; D. Drewry, *Glacial Geologic Processes*, Edward Arnold, London, 1986; M. Hambrey and J. Alean, *Glaciers*, 2d ed., 2004; W. S. B. Paterson, *The Physics of Glaciers*, 3d ed., 1999.

Gland

A structure which produces a substance or substances essential and vital to the existence of the organism and species. Glands are classified according to (1) the nature of the product; (2) the structure; (3) the manner by which the secretion is delivered to the area of use; and (4) the manner of cell activity in forming secretion. A commonly used scheme for the classification of glands follows.

I. Morphological criteria

A. Unicellular (mucous goblet cells)

B. Multicellular

1. Sheets of gland cells (choroid plexus)

2. Restricted nests of gland cells (urethral glands)

3. Invaginations of varying degrees of complexity

a. Simple or branched tubular (intestinal and gastric glands)—no duct interposed between surface and glandular portion

b. Simple coiled (sweat gland)—duct interposed between glandular portion and surface

c. Simple, branched, acinous (sebaceous gland)—andular portion spherical or ovoid, connected to surface by duct

d. Compound, tubular glands (gastric cardia, renal tubules)—branched ducts between surface and glandular portion

e. Compound tubular-acinous glands (pancreas, parotid gland)—branched ducts, terminating in secretory portion which may be tubular or acinar

II. Mode of secretion

A. Exocrine—the secretion is passed directly or by ducts to the exterior surface (sweat glands) or to another surface which is continuous with the external surface (intestinal glands, liver, pancreas, submaxillary gland)

B. Endocrine—the secretion is passed into adjacent tissue or area and then into the bloodstream directly or by way of the lymphatics; these organs are usually circumscribed, highly vascularized, and usually have no connection to an external surface (adrenal, thyroid, parathyroid, islets of Langerhans, parts of the ovary and testis, anterior lobe of the hypophysis, intermediate lobe of the hypophysis, groups of nerve cells of the hypothalamus, and the neural portion of the hypophysis)

C. Mixed exocrine and endocrine glands (liver, testis, pancreas)

D. Cytocrine—passage of a secretion from one cell directly to another (melanin granules from melanocytes in the connective tissue of the skin to epithelial cells of the skin)

III. Nature of secretion

A. Cytogenous (testis, perhaps spleen, lymph node, and bone marrow)—gland "secretes" cells

B. Acellular (intestinal glands, pancreas, parotid gland)—gland secretes noncellular product

IV. Cytological changes of glandular portion during secretion

A. Merocrine (sweat glands, choroid plexus)—no loss of cytoplasm

B. Holocrine (sebaceous glands)—cells undergo dissolution and are entirely extruded, together with the secretory product

C. Apocrine (mammary gland, axillary sweat gland)—only part of the cytoplasm is extruded with the secretory product

V. Chemical nature of the product

A. Mucous goblet cells (submaxillary glands, urethral glands)—the secretion contains mucin

B. Serous (parotid gland, pancreas)—secretion does not contain mucin

Development. Glands arise from the epithelium either by modification of cells in situ or as outgrowths. In situ glands are either unicellular and the result of cellular modifications, epithelial sheets of contiguous single cells, or intraepithelial groups or nests of cells (**Fig. 1**). Outgrowth glands are multicellular and arise as evaginations which grow into the surrounding tissue and may incorporate this tissue as part of the gland's structure. Outgrowths may be hollow or solid and of three main types: (1) those glands which detach themselves entirely from parental epithelium, such as the thyroid; (2) solid, club-shaped masses which become tubelike, and may be simple tubular, simple branched tubular, or compound tubular structures, of which the secreting portion of the gland is tubular, straight, or coiled (**Figs. 2** and **3**); and (3) bulbous outgrowths which form sac-shaped secreting units (**Fig. 4**). Each unit is called an acinus or alveolus. Such glands may be simple acinus, simple branched acinous, and compound acinous. Compound glands, either tubular (Fig. 3) or acinous (Fig. 4), are distinguished from simple glands by the presence of many ducts. The various ducts eventually join one or more common ducts which lead to the surface outlet. Compound glands incorporate surrounding tissues, including blood, lymph, and nerve tissues, into their substance as development proceeds. Such glands become large and have a tendency to subdivide into smaller divisions or lobes. Some compound glands are composed of tubular

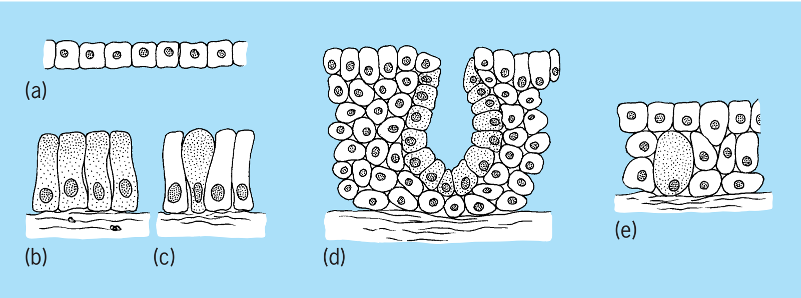


Fig. 1. Intraepithelial gland types. (*a*) Embryonic epithelium. (*b*) Epithelial sheet of glandular cells. (*c*) Unicellular goblet gland. (*d*) Multicellular gland. (e) Leydig's cell.

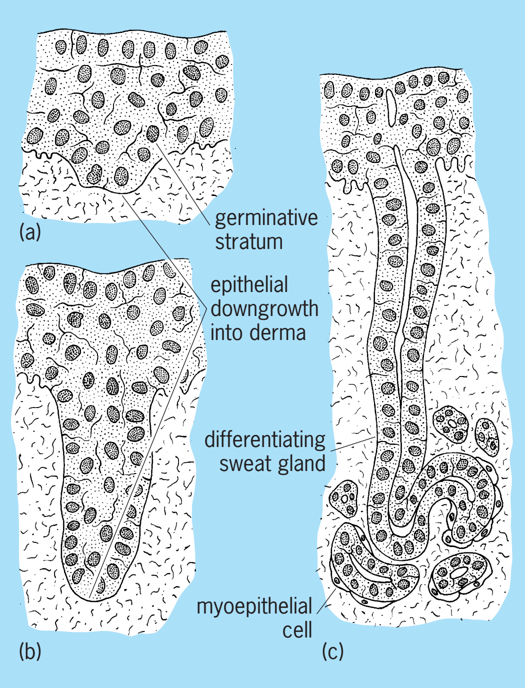


Fig. 2. Development of sweat gland as outgrowth from epithelial layer. (*a*, *b*) Epithelial downgrowth from germinative stratum. (*c*) Differentiating gland.

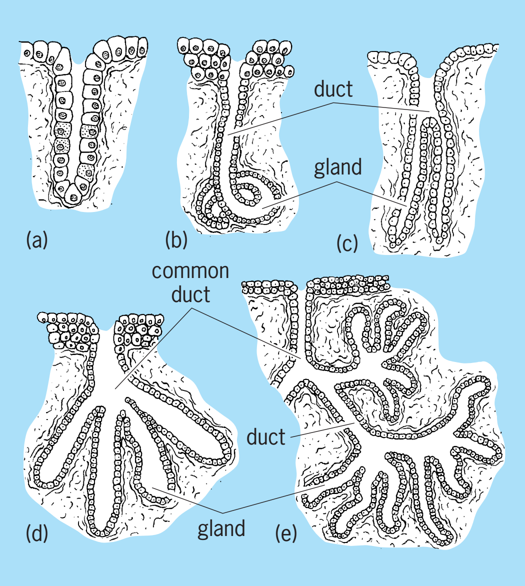


Fig. 3. Examples of tubular glands. (*a*) Simple (gland of Lieberkühn). (*b*) Coiled (sweat gland). (*c*) Simple branched (gastric gland). (*d*) Simple branched (Brunner's gland). (*e*) Compound (liver).

and acinous structures. They consist of a series of branched tubules which possess saclike, acinous outgrowths from their walls or distal ends.

Secretions. Glands may be classified on the basis of the kinds of secretions they produce. Sebaceous glands secrete oil or oily materials; serous glands, albuminous, watery material; mucous glands, a gelatinous substance; cytogenous glands liberate living cells; and granular glands produce and secrete granules.

Skin glands. All skin glands which include the in situ and epithelial-outgrowth glands are exocrine.

In situ glands are unicellular and hence represent modified single epithelial cells. They remain at the surface of, or within, the epithelial layer of the skin. Small unicellular glands together with larger single-cell glands known as club cells are abundant in the epidermis of fishes. They are mucus-secreting and odoriferous. The glands of Leydig are unicellular structures present in the epidermis of larval urodele amphibia such as salamanders and newts, and in adult urodeles, such as *Necturus*, which retain a quasi-larval condition in the adult. They are found also in gymnophionan amphibia. Leydig cells resemble the club cells of fishes. They secrete mucus and some may produce a poisonous substance. Unicellular hatching glands are present in the epidermis of the snout region of frog and toad larvae previous to hatching, and probably also in the larvae of most other amphibia. Their secretion digests the egg capsule and permits the larva to hatch.

Epithelial-outgrowth glands in their development present bulbous, epithelial outgrowths into the subepithelial tissues, and they may be classified into two groups. One group contains simple unbranched, or slightly branched acinous glands of fishes and amphibians. In most cases, they are mucus-secreting, although in some fishes and amphibians they are poison-secreting. The second group are simple unbranched, branched acinous, or tubuloacinous glands of reptiles, birds, and mammals. The secretion is thick and sebaceous. These glands are abundant in mammals and generally associated with the hair follicles. However, in some areas of the mammalian skin such as the nose, lips, nipples, upper eyelids, around the anus, and external genitals, these glands arise independently as invaginations of the germinative stratum of the epidermis. Compound

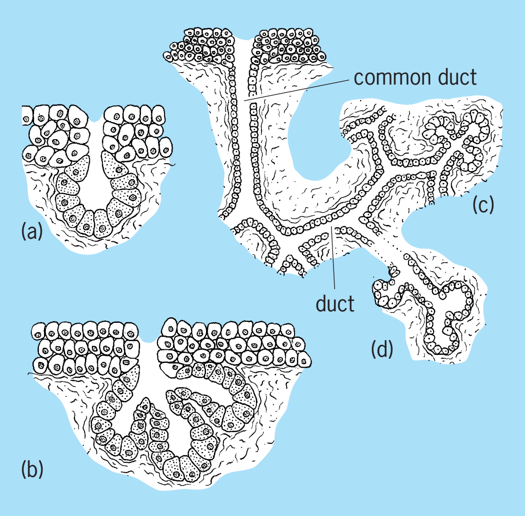


Fig. 4. Examples of typical acinous (alveolar) glands. (*a*) Simple. (*b*) Branched. (*c*) Compound. (*d*) Tubuloacinous, in which the distal secreting units are simple tubes with acinous side chambers.

tubular, nasal glands (salt glands) are situated near the eyes in marine birds and reptiles. These salt glands eliminate excess salt taken in by drinking seawater and with food. The fluid produced by salt glands is many times as salty as the bird's blood or body fluids; the kidneys produce more fluid, but with a much lower sodium chloride content. *See* ENDOCRINE SYSTEM (VERTEBRATE); LACRIMAL GLAND; POISON GLAND; SALT GLAND; SCENT GLAND; SEBACEOUS GLAND; SWEAT GLAND; UROPYGIAL GLAND.

Olin E. Nelsen

Glanders

A contagious zoonosis affecting primarily horses, mules, and donkeys and caused by the bacterium *Burkholderia* (*Pseudomonas*) *mallei*. Glanders primarily involves the respiratory systems, skin, and lymphatics. It is marked by a purulent inflammation of mucous membranes and an eruption of nodules on the skin, forming deep ulcers. Glanders was once common throughout the world but is now found only in the Mideast and parts of Africa, Russia, Asia, and South America. *Burkholderia mallei* is a gram-negative, non-acid-fast, nonsporulating, nonmotile, unencapsulated bacillus occasionally showing bipolar staining; it is obligately aerobic and oxidase-positive. It is closely related to *B. pseudomallei*, the cause of melioidosis, a disease of humans and animals in Southeast Asia and northern Australia. *Burkholderia mallei* is highly infectious for humans, who may acquire it by handling or treating glanderous animals or during laboratory investigations. Untreated acute disease in humans has a 95% mortality rate within 3 weeks. There is no vaccine against infection. *Burkholderia mallei* is considered a category B biological agent because it is moderately easy to disseminate, causes moderate morbidity with high mortality in untreated cases, and requires special enhanced laboratory diagnostic capability. *See* PSEUDOMONAS.

Epidemiology and transmission. Glanders is usually contracted by ingestion of contaminated food or water, by contact, or by inhalation of infectious droplets. The organism survives for months under most warm conditions. All equids are highly susceptible, particularly donkeys and mules. The disease is most frequent in endemic areas in the rainy season. Historically, glanders has been a scourage of military horses, in which the disease rapidly become endemic in wartime. During World War I it caused great loss on the eastern front in Russia. In World War II it was the subject of intense study as a biological weapon in Japan, Russia, and the United States, and caused heavy losses in horses and mules in the Far East.

Clinical findings and pathology. The disease is usually acute and often fatal in donkeys and mules, and chronic in horses, some of which may ultimately recover but continue to carry *B. mallei*. It is characterized by formation of nodules and ulcerations of the skin and respiratory membranes and by granulomatous nodules in the lungs, lymphatic channels, and lymph nodes. Acute disease is often rapidly fatal because of multiplication and dissemination of the organism in the bloodstream.

Humans develop a chronic or acute form with painful cutaneous ulcerating nodules, lymphangitis, and abscessation as in equids. Nodules progress to pyemia, pneumonia, and death within 3 weeks unless antibiotic treatment is provided.

Diagnosis and treatment. *Burkholderia mallei* may easily be cultured on agar medium supplemented with glycerol. The addition of bacitracin, polymixin B, and actidione is useful to suppress contaminants. The colonies are small, round, translucent, and yellowish and give off a slight fruity odor. The polymerase chain reaction (PCR) may be used for rapid, safe identification.

Although *B. mallei* is sensitive to sulfonamides and tetracyclines, affected horses are not usually treated since the destruction of cases is extremely effective for control and eradication. Essential components of diagnosis include clinical examinations at frequent intervals to detect the cutaneous and nasal forms, immunological tests to detect serum antibody, and skin and intradermopalpebral (within the skin of the eyelid) injection of mallein, a glycoprotein of *B. mallei*, to detect hypersensitivity. *See* AGGLUTINATION REACTION; COMPLEMENT-FIXATION TEST; HYPERSENSITIVITY.

Control and prevention. Early detection and elimination of infected animals is the key control measure. Detection and destruction of infected animals and regional quarantine measures have been very effective in control and eradication. Stables used by affected horses should be depopulated for several months before introducing susceptible stock. John F. Timoney

Bibliography. A. Bauernfeind et al., Molecular procedure for the rapid detection of *Burkholderia mallei* and *Burkholderia pseudomallei*, *J. Clin. Microbiol.*, 36:2737–2741, 1998; Glanders, in *Manual of Standards for Diagnostic Tests and Vaccines*, pp. 576–581, World Organization for Animal Health, Paris, 2000; J. Lopez et al., Characterization of experimental equine glanders, *Microb. Infect.*, 5(12):1125–1131, 2003.

Glass

Materials made by cooling certain molten materials in such a manner that they do not crystallize but remain in an amorphous state, their viscosity increasing to such high values that, for all practical purposes, they are solid. Materials having this ability to cool without crystallizing are relatively rare, silica (SiO_2) being the most common example. Although glasses can be made without silica, most commercially important glasses are based on it. *See* AMORPHOUS SOLID.

Glass products are enormously varied, including windows, doors, bottles and vials, tableware, optical instruments, fiber optics, mirrors, colored filters, chemical apparatus, pipe, crucibles, thermal and electrical insulation, cloth, and automobile, boat, and aircraft bodies. It is even possible to make a glass in which a photographic image can be developed.

Chemical properties. Chemically, most glasses are silicates. Silica by itself makes a good glass (fused silica), but its high melting point (3133°F or 1723°C) and its high viscosity in the liquid state make it difficult to melt and work. Fused silica products are expensive and are used only when their special properties (low thermal expansion, high softening point, light transmission characteristics, and corrosion resistance) are essential. *See* SILICA MINERALS.

To lower the melting temperature of silica to a more convenient level, soda (Na_2O) is added in the form of sodium carbonate (soda ash). This has the desired effect, but unfortunately the resulting glass has no chemical durability and is soluble even in water (water glass).

To increase durability, lime (CaO) in the form of calcium carbonate (limestone) is added to the glass to form the basic soda-lime-silica glass composition that is used for the bulk of common glass articles, such as bottles and window glass. Although these are the main ingredients, commercial glass contains other oxides (aluminum and magnesium oxides) and ingredients to help in melting, oxidizing, fining, or decolorizing the glass batch.

Special kinds of glass have other oxides as major ingredients. For example, boron oxide (B_2O_3) is added to silicate glass to make a low-thermal-expansion glass for chemical glassware that must withstand rapid temperature changes, for example, Pyrex glass. This type of glass is known as a borosilicate. Also, lead oxide (PbO) is used in optical glass and lead crystal because it gives a high index of refraction. Many other special types of glass have been developed for particular uses; some of these are discussed in this article.

Structure. Physically, glass is an arrangement of atoms quite like that in the liquid state; that is, it has no long-range order. In a crystal, the atoms are arranged in a regular, repeating pattern, but in a glass, although the arrangement in the neighborhood of a single type of atom may be the same throughout the glass (four oxygens around each silicon), the overall structure lacks periodicity of atomic arrangement (**Fig. 1**).

Melting. Production of glass articles begins with batch mixing of raw materials (sand, soda ash, and limestone) and their melting. For small production and special glass, melting may be done in pots or crucibles containing up to 1 ton of glass. Several pots may be heated by one central furnace into which the pots are inserted and later removed through large doors.

Larger batches are melted in large covered furnaces or tanks to which heat is supplied by a flame playing over the glass surface. Usually, these glass tanks are fired regeneratively; that is, the hot exhaust gases pass through an open brick lattice (checker) on one side of the furnace and heat it. After about 30 min, the flow is reversed, combustion air is brought in through the hot checkers and preheated, and the exhaust gases go out through checkers on the opposite side of the tank. Regeneration conserves energy and increases the flame temperature. Most glass tanks are fired by gas or oil; however, auxiliary heating with electricity (boosting) is common in the United States and all electric melting is widely used in Europe.

A glass tank may be either a batch (day tank) or a continuous type. The latter is divided into two sections joined by a narrow passage or throat. The raw materials and the scrap glass (cullet) to be remelted are charged continuously into one end of the first or melting section, where they are melted, mixed thoroughly, and refined (freed from bubbles) at high temperature (2910°F or 1600°C). The molten glass is then cooled as it passes through the throat to the second or conditioning section and exits the tank. It is cooled to the temperature (about 2012°F or 1100°C) for forming (shaping), and is taken by forehearths to the forming operation, which is different for each type of product. Continuous tanks range in capacity from 50 to 500 tons (45 to 450 metric tons) of glass per day.

Heat treatment. After forming, glass must be slowly cooled or annealed, usually in a long oven called a lehr. The purpose of annealing is to reduce the internal stresses, which can degrade the strength or optical properties of glass; for example, large internal stresses can cause the glass to crack during cooling and make the glass birefringent. These stresses

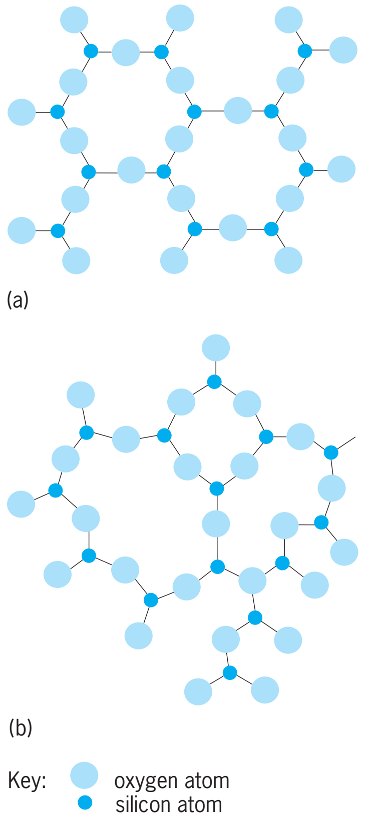


Fig. 1. Comparison of arrangements of atoms in crystals and glass. (*a*) Regular arrangement of atoms in crystal. (*b*) Disordered arrangement of atoms in glass.

are created when, because of temperature variations throughout the piece, different parts of the ware become rigid at different times. The opposite of annealing is tempering, in which the glass is rapidly chilled by a blast of air. This results in large compressive stresses in the surface of the glass. Because glass invariably fails by tension, this compressive stress must be overcome before the glass breaks. Thus tempered, glass is very strong and is used for doors and windows. Glass that is annealed usually breaks irregularly into fragments called shards. If tempered glass breaks, it shatters into many small pieces of roughly rectangular shape because of high internal stresses.

Products. The most common glass products are container ware and flat glass. The latter is principally used in transportation and architecture. Flat glass is produced by the float process, developed in the 1950s and considered one of the important technological achievements of the twentieth century (**Fig. 2**). Here a stream of viscous glass from a long melting furnace is poured continuously (floated) onto a shallow pool of molten tin at about 1920°F (1050°C). The glass spreads to a width of 130 in. (3.3 m), forms a sheet or ribbon with parallel surfaces, and attains a natural equilibrium thickness of about 0.25 in. (6.8 mm). As it continues down the tin bath, it cools and is sufficiently solid to be lifted off the bath exit at 1200°F (650°C). It then travels through an annealing lehr before being cut to size. Thinner glass may be made by stretching the ribbon with top rollers, and thicker glass may be made by partially damming the glass on the tin surface. The tin bath must be sealed to maintain an inert atmosphere and prevent oxidation of the tin.

Float glass has completely displaced two older flat-glass processes. Sheet glass with a wavy but brilliant fire-polished surface is made by drawing molten glass upward and solidifying it; and plate glass with parallel surfaces is made by drawing glass through rollers, cooling and annealing it, and then grinding and polishing both surfaces. Advantages of float glass are high optical quality, no waviness, fire-polished surface, and large thickness range [less than 0.04 in. (1 mm) to over 1 in. (25 mm)]; the process is continuous, and well suited to computer control and on-line coating application.

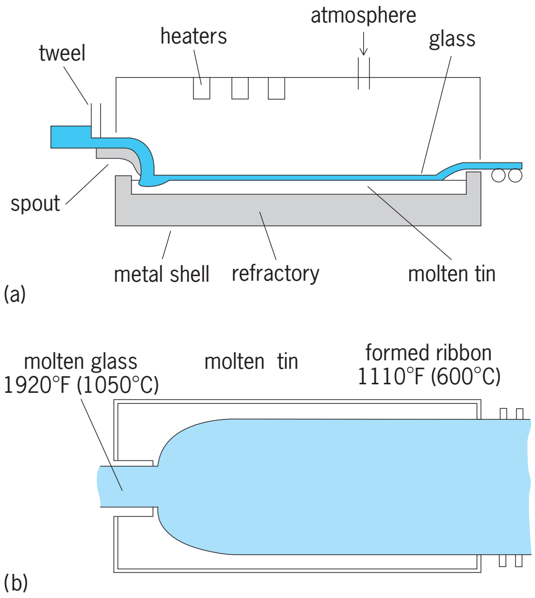


Fig. 2. Float process for manufacture of flat glass. (*a*) Float bath. (*b*) Glass sheet with equilibrium thickness. (*After W. C. Hynd, Flat glass manufacturing processes, in D. R. Uhlmann and N. J. Kreidl, eds., Glass: Science and Technology, vol. 2: Processing 1, Academic Press, 1984*)

Flat glass has come to be considered a commodity, and value-added products are made by bending, tempering, laminating, encapsulating, double-glazing, and coating the glass. Glass is bent for greater freedom of form, and it is tempered to make it stronger and safer. Laminated safety glass used in windshields is made by bonding two sheets of thin glass (0.08–0.12 in. or 2–3 mm) with a transparent organic material under heat and pressure. Thin film coatings on flat glass modify light and heat transmission; they may be electrically conductive, reflective, semireflective, antireflective, colored, low-emissivity, and so forth. Electrochromic coatings darken the glass when voltage is applied. *See* DIELECTRIC MATERIALS; ELECTROCHROMIC DEVICES; OPTICAL MATERIALS.

Container ware is largely made on automatic machines; some special pieces may be made by hand blowing into molds, and some very special ware may be made by the free or offhand blowing of a skilled glass blower. In the automatic process (**Fig. 3**), a stream of glass flows from the forehearth and is cut by shears into individual gobs, which are fed to a blank mold. Here the gob is formed into a rough blank, or parison, by either a plunger or compressed air; at this stage the opening of the bottle gets its final shape. The blank mold opens and the parison, supported by the bottle opening, is transferred to the final or blow mold, where it is blown to shape. The bottle is then transferred to a traveling belt, which carries it through the annealing lehr and on to automated inspection and packing.

Electric light bulb envelopes are molded on a special machine which converts a fast-moving ribbon of glass into over 10,000 bulbs per minute.

Glass fibers are either continuous for reinforcement or discontinuous for insulation. Continuous fibers are usually made by extruding molten glass through multiple (about 2000) nozzles in a platinum bushing built into the bottom of a forehearth; they may also be made by remelting and extruding glass marbles. After extrusion, fibers are drawn rapidly [up to 240 ft/s (80 m/s)] to attenuate their diameter, coated with bonding and protective chemicals (collectively called a size), and wound together on a spool or collet. Most continuous fiber is an E-glass composition, which contains the oxides of silicon, boron, calcium, and aluminum but no soda or other alkali. Continuous glass fiber is used to weave glass cloth or is chopped and used to reinforce plastics.

Glass wool is most often made by spinning, or forcing molten glass centrifugally out of small holes (around 20,000) in the periphery of a rapidly rotating steel spinner, and attenuating the fibers by

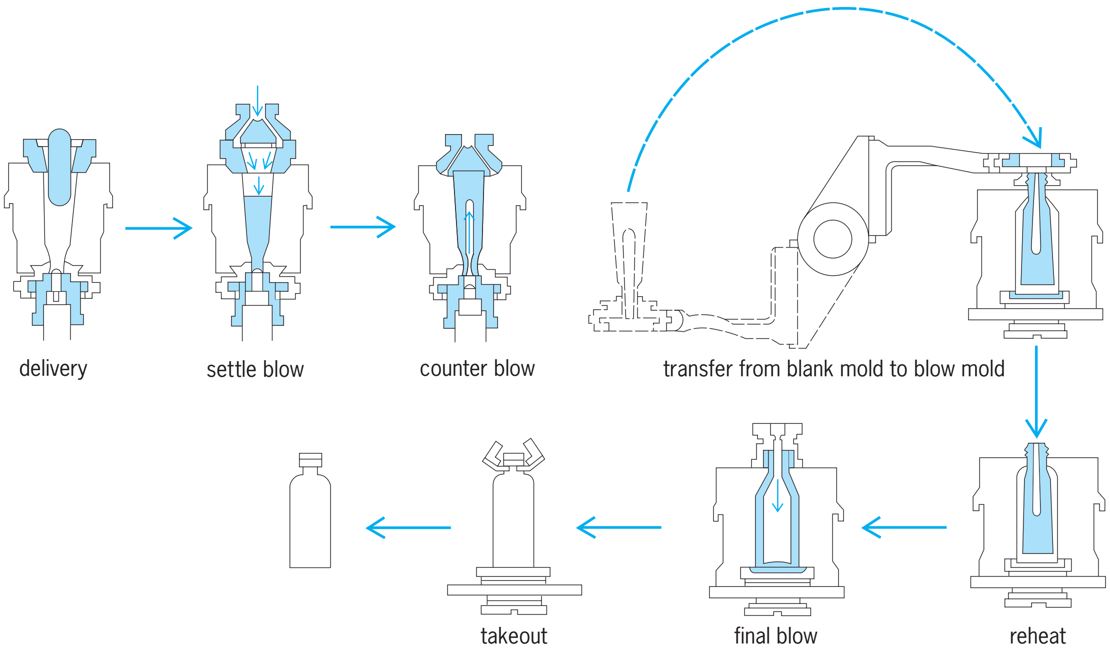


Fig. 3. Steps in automatic bottle manufacture. (*Glass Packer*)

entrainment with gas burners. As the fibers fall to a conveyor, they are sprayed with a binder that preserves their open structure. The resulting glass wool mat is mainly composed of still air, which accounts for its excellent insulation properties. Since these fibers are of extremely small diameter (about 5 micrometers) and thus have a large surface area, they are susceptible to chemical corrosion. Glass wool is similar in composition to window glass, but contains 5–10% boric oxide to increase the chemical durability. Many other processes exist for making discontinuous glass fiber. When made from impure natural materials or from slags, it is often called rock or mineral wool.

Optical fibers are hair-thin transparent glass fibers that transmit light around curves and over long distances by total internal reflection. Many fibers are combined into one fiber bundle or cable. A fiberscope allows image transmission for remote inspection or in medical procedures. Fiber optics for telecommunication carry digital pulses of light, which may be amplified by using a fiber laser amplifier. Power fibers carry high-power laser light used in welding or surgery. Fiber-optic sensors measure ambient conditions wherever a fiber optic can be inserted, such as pH and chemical composition in a reaction vessel, air quality, or blood gases and pressure. *See* FIBER-OPTIC SENSOR; FIBER-OPTICS IMAGING; OPTICAL FIBERS.

Foam glass, used where a self-supporting insulation material is needed, is formed by creating a large volume of small bubbles in a matrix of glass.

Optical glass is generally melted in platinum-lined tanks or pots, from which it is drawn or cast into blanks which can be ground to final shape. An alternate method is to cool the glass in the pot, break it apart, and select flaw-free pieces. Because an important requirement in optical glass is clarity, the purest raw materials are used.

Special glass. Besides the fairly common types of glass mentioned above, there are hundreds of special types; for example, the milk-white opal glass used for dinnerware and bottles, which owes its appearance to tiny crystals grown in the glass.

The process of crystallization in a glass, known as devitrification, is generally avoided; however, in some cases it is desired. An example is a type of glass ceramic that is shaped in the same manner as an ordinary glass, and then, by appropriate heat treatment, converted to a completely crystalline material. It is transparent because of the extremely small crystallite size (60 nanometers) and has outstanding thermal shock resistance since its thermal expansion is nearly zero (comparable to fused silica).

Another special glass is sensitive to light, which creates a latent image within the glass, the image being developed by heat treatment. This glass is used to make signs, and radio and clock faces. A variant of this glass is one in which the portion exposed to light is more soluble than the unexposed portions; it can be used to form, photographically, a fine intricate mesh or screen.

So-called solder glass has a relatively low softening point (below 930°F or 500°C). It is used to join two pieces of higher-melting glass, without softening and deforming them. The solder glass is applied to the joint as a powder and melted to produce a seal, as in a vacuum tube.

Vycor glass is a nearly pure silica glass formed without the production problems of fused silica. The starting glass, a soda borosilicate, is formed and, under heat treatment, separated into two phases, one of which can be leached out. The remaining porous material is nearly pure SiO_2 and can be converted to a dense, clear glass by heating. During the latter treatment, it shrinks considerably.

Chemical vapor deposition produces very pure silica glass for fiber optics by hydrolysis of silicon tetrachloride in a flame and densification of the silica soot so produced. *See* VAPOR DEPOSITION.

Properties. The most important properties are viscosity; thermal expansion; strength; index of refraction; dispersion; light transmission (both total and as a function of wavelength); corrosion resistance; and electrical properties. These properties depend on glass composition, so glasses with specific values of these properties can often be tailor-made for specific applications.

The viscosity of glass is important mainly in its manufacture. The viscosity increases continuously as the temperature decreases, the logarithm of the viscosity being nearly proportional to the reciprocal of the temperature. Important points on the viscosity-temperature curve are the melting range (viscosity about 100 poises), the working range (10^3–10^6 poises), the softening point ($10^{7.6}$ poises), the annealing point (10^{13} poises), and the strain point ($10^{14.5}$ poises); the annealing range lies between the last two points. Traditionally the unit for viscosity is the poise (a relic cgs unit) but the pascal-second (Pa-s) the SI unit, is used increasingly in glass science; 10 poises = 1 Pa-s.

The coefficient of thermal expansion of various glass families ranges from close to zero to over 200 (units of $10^{-7}/°C$). Silica, titania-silica ultralow-expansion glass, and specially designed glass ceramics have a coefficient of thermal expansion of nearly zero (from 0 to 5.5) and are used for telescope mirror supports, where good dimensional stability on temperature change is necessary. Glass ceramics are used for cook tops and ovenware, where outstanding thermal shock resistance is necessary. Common soda-lime glass has a coefficient of thermal expansion of 92, and it is subject to breakage when cooled rapidly. Alkali borosilicates have a coefficient of thermal expansion of 32 and are intermediate in thermal shock resistance. In making glass-to-metal seals and in enameling glass onto metal, it is important to match the coefficient of thermal expansion of the metal with that of the glass; this is generally possible because the thermal expansion may be fixed at any desired value by adjusting glass composition suitably. *See* THERMAL EXPANSION.

The theoretical strength of glass is extremely high, about 40 gigapascals (GPa; 6,000,000 lb/in.^{2}), but this is reduced markedly in practice because of the presence of many submicroscopic flaws on the glass surface. These act as stress concentrators and effectively control practical glass strength. The flaws may be mechanical in origin (handling the glass surface inevitably introduces flaws) or particulate (surface devitrification, dust, metal particles from forming equipment, and so forth). High-strength glass fibers may have strengths as high as 14 GPa (2,100,000 lb/in.^{2}); commercial glassware is 50–150 megapascals (MPa; 7000–22,000 lb/in.^{2}); and abraded glass may be as low as 15 MPa (2200 lb/in.^{2}). Hard oxide and lubricating coatings are applied to glass bottles to reduce friction, thereby reducing surface damage and maintaining bottle strength.

Glass articles may be strengthened by prestressing the glass to put the surface in compression. This may be done by several means: thermal tempering or toughening by quenching (automotive glass except windshields; doors; tableware), chemical strengthening (eyewear, basketball backboards), enameling, lamination of core and clad glasses with different thermal expansions, or controlled surface devitrification. Tempered glass is most common; it is five times as strong as annealed glass and is more resistant to thermal shock.

Glass breaks with no warning by brittle fracture. The low practical strength of glass and its brittle fracture are perhaps the most serious limitations on its use as an engineering material.

Index of refraction and its variation with wavelength (dispersion) depend on glass type and composition. Crown glasses (such as soda-lime glass) have low dispersion, while flint glasses (such as lead glass) are higher. Because of the dispersion of colors by glass, the image produced by white light through a simple lens is blurred, a phenomenon called chromatic aberration. This is reduced considerably by combining a crown glass with a flint glass in a compound lens. *See* REFRACTOMETRIC ANALYSIS.

Light transmission is important for materials such as colored glass and filters. However, it also determines the tint of ordinary window and bottle glass. The main coloring impurity, iron oxide, enters the glass via impure raw materials and imparts a greenish tinge, which may be seen by looking edgewise through "clear" float glass. In nominally clear float glass the iron concentration is 0.1% iron(III) oxide (Fe_2O_3). For optical glass, in which maximum transmission is needed, the iron oxide must be kept below 0.01% through beneficiation of raw materials, and its oxidation state must be controlled in order to minimize absorption. In fiber optics, with even lower absorption requirements (about 0.1 dB/km), impurities must be held to parts per million. If a clear glass is desired and total transmission is not important, it can be achieved by adding a so-called decolorizer such as selenium or cobalt, which colors the glass blue and balances the green iron absorption; the resulting glass has a light gray coloration that cannot be easily detected by the eye.

The electrical conductivity of glass, as for most insulators, increases with temperature, the logarithm of the conductivity varying linearly with the reciprocal of the absolute temperature. The surface of glass can be made conducting by a transparent tin oxide coating.

Dimensional stability is of importance in precision instruments, such as clinical thermometers, in which a delayed dimensional change in the glass can destroy the calibration of the instrument. If the glass is not carefully annealed and aged, it undergoes a compaction with time because the atoms slowly draw nearer to each other, a condition which is more stable at room temperature. This leads to changes in properties such as density, index of refraction, and strength. J. F. McMahon; J. Wenzel

Colored and stained glass. Stained glass is colored by any of several means and assembled to produce a varicolored mosaic or representation. Stained glass windows, with the pieces arranged either to produce a pattern or a picture, have long been a decorative feature of buildings and especially old churches. Stained glass is also assembled into chandeliers and ornamental objects. The glass is colored by one of three means: fusing with metallic oxides, enameling, or painting. The addition of small percentages of metal compounds, usually oxides, to molten glass or pot metal produces throughout the glass a color characteristic of the compound.

The glass can also be coated with transparent metallic oxides and fired to bond the enameling to the glass base. Transparent pigments can be painted onto the glass in solid colors, or in several colors to render a picture; the pigments are then baked or burned onto the glass. With a hot iron, the glazier cuts the glass to shape, joins it into the design with channeled lead strips, and solders the joints. Either the tracery of the window or iron frames support the finished glaziery.

From the ancient art of making stained glass has evolved the modern technology of controlling the transmittance and color of glass. The process of adding metallic oxides to glass is widely used to produce colored glass for other than decorative purposes. Among these technological uses are the screening out of ultraviolet portions of solar radiation by display-window glass to decrease fading of objects on exhibition, the filtering out of infrared radiation especially in motion-picture and slide projectors by means of heat filters, and the reducing of visual contrast by means of a tinted auto windshield. Enameled glass is used for washable wall coverings and for luminous signs, and colored lenses in traffic lights, railroad and airport signals, and brake lights.

Melting highly colored glasses with flames poses problems because radiation from the heat of the flame cannot penetrate the glass melt, and the furnace bottom remains cold. Such glasses are often melted electrically. Forming is also affected to a lesser degree. J. Wenzel

Bibliography. N. P. Bansal and R. H. Doremus, *Handbook of Glass Properties*, 1996; C. A. Harper, *Handbook of Ceramics Glasses, and Diamonds*, 2001; G. McLellan and E. Shand (eds.), *Glass Engineering Handbook*, 1984; H. Rawson, *Glasses and Their Applications*, 1991; J. E. Shelby, *Introduction to Glass Science and Technology*, 2d ed., 2005; F. Tooley (ed.), *The Handbook of Glass Manufacture*, 2 vols., 1984; D. Uhlmann and N. Kreidl (eds.), *Optical Properties of Glass*, 1991; A. Varshneya, *Fundamentals of Inorganic Glasses*, 1994.

Glass switch

A glassy, solid-state device used to control the flow of electric current. Useful solid-state devices can be made from glassy as well as crystalline semiconductors. Crystals possess long-range order; that is, given the position of any particular atoms and the orientation of the neighboring atoms, the location of any other atom is known, no matter how far away from the atom under consideration. A glass is a special case of a noncrystalline class of materials, namely, amorphous solids. These do not exhibit long-range order, although they tend to have the same local structure (that is, short-range order) as the corresponding crystal. A glass is an amorphous solid that is formed by cooling rapidly from the liquid phase.

The first applications of glassy semiconductors were switches made from chalcogenide (that is, alloys containing tellurium, selenium, or sulfur) glasses. The two basic structures are known as the Ovonic Threshold Switch (OTS) and the Ovonic Memory Switch (OMS). They are active devices consisting simply of a thin film (about 10–100 nanometers thick) of glass between two metallic contacts. The device characteristics depend on the bulk properties of the semiconductor material rather than on the contacts. Consequently, the switches are symmetrical in that they respond identically to voltages and currents of both polarities. The OMS and OTS differ in terms of the composition of their amorphous semiconductor thin-film materials and their functional performance.

Amorphous semiconductors can be formulated so they can be doped by adding small amounts of impurities to change their electrical properties in the same way as crystalline semiconductors, or they can be designed to be insensitive to the effects of impurities. The glass switches typically use impurity-independent material compositions, and they are also highly resistant to the effects of radiation.

Device characteristics. Both the OTS and OMS show a rapid and reversible transition between a highly resistive and a conductive state effected by applied electric fields. The main difference between the two devices is that, after being brought from the highly resistive state to the conducting state, the OTS returns to its highly resistive state when the current falls below a holding current value. On the other hand, the OMS remains in the conducting state until a current pulse returns it to its highly resistive state. The OMS thereby remembers the last applied switching command, and it is from this property that the device receives its name.

Intermediate resistance states are also possible for OMS devices, which can be used in applications requiring a "gray scale." In all of these devices, the transitions between states are completely reversible.

The composition of the active material determines whether the device functions as an OTS or OMS, and also affects the values of certain device parameters. The device geometry, such as thickness and cross-sectional area of the active film, also affects the numerical values of the device parameters.

In the OTS (**Fig. 1**), conduction in the highly resistive state follows Ohm's law at fields below about 10^4 V/cm. At higher fields the dynamic resistance R_{dyn} decreases monotonically with increasing voltage. Typical values are $R_{\text{dyn}} = 2$–10 megohms at 1 V and $R_{\text{dyn}} = 0.1$–0.5 megohm just prior to breakdown.

When the applied voltage exceeds a threshold voltage V_T, the OTS switches along the load line to the conducting state. The transition time τ_t of this switching process has been shown to be less than 150 picoseconds. V_T is a function of both film thickness and active material composition and can be obtained in the range 2–300 V.

Current in the conducting state can be increased or decreased without significantly affecting the voltage drop across the device; the dynamic resistance is of the order of 1–10 ohms. Most of the voltage falls near the two contacts, due to barriers induced prior to switching; this accounts for about a 0.4-V drop. The field across the bulk is only about 1 kV/cm.

If the current is decreased below a critical value I_H, the OTS switches back to the original highly resistive state. I_H depends on circuit parameters and also can be varied; typical values are 0.1–1 mA.

The foregoing description of the static characteristics of the OTS holds for a slowly varying applied voltage. Upon application of a fast-rising pulse somewhat in excess of threshold voltage, the OTS ordinarily does not immediately switch to the conducting state but remains in the high-resistance state for a period of time τ_d, called the delay time. The magnitude of this delay is strongly dependent on the extent to which the threshold voltage is exceeded. For an applied voltage pulse slightly greater than V_T, τ_d can be several microseconds. However, it rapidly decreases with increasing voltage in excess of threshold, and essentially vanishes above a critical voltage that is proportional to film thickness. Above this point, the speed of switching is only circuit-limited, and total switching times less than 150 picoseconds have been observed.

Switching is an electronic effect, induced by the appearance of a critical electric field across a part or all of the film. The field induces a sharp transition which increases the free-carrier concentration by a factor of about 10^8. This results in a constant current density of approximately 10 kA/cm^2. At such current densities Joule heating effects are negligible.

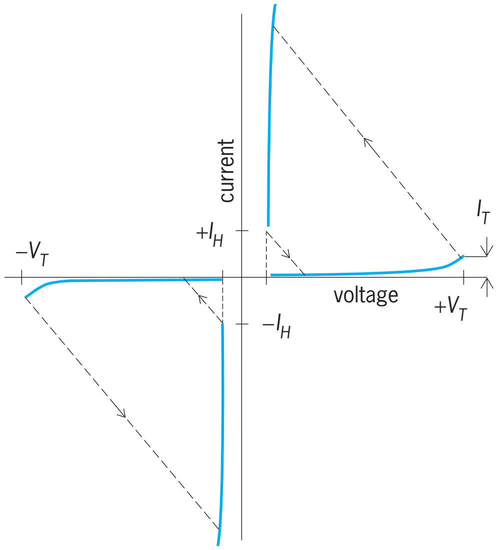


Fig. 1. OTS current-voltage characteristics.

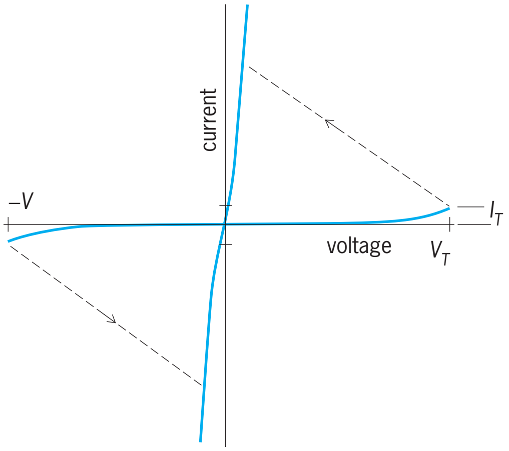


Fig. 2. OMS current-voltage characteristics.

The high conduction occurs primarily through a central filament whose area varies proportionally with the current.

In the OMS, the properties of the highly resistive state are essentially the same as for the OTS (**Fig. 2**). The OMS is switched to the conductive state when the threshold voltage is exceeded. After switching, however, the OMS is designed so that the amorphous semiconductor material within the conducting filament region can undergo a change in atomic structure. This change results in a change in electrical conductivity which remains after power is removed. The structural change is reversible by providing another pulse with a different current profile to the OMS so the device can be repeatedly programmed into a selected conductivity state.

Amorphous switch materials. Materials used in an OTS are specifically chosen to meet several important requirements. First, the film should have high electrical resistivity to avoid Joule heating effects. Second, the amorphous phase should be sufficiently stable to prevent crystallization or other structural changes. Third, the material should be chosen so that irreversible breakdown does not occur. All of these conditions can be fulfilled by the use of chalcogenide glasses. Chalcogenides are alloys one of whose major components comes from one of the elements in group 16 of the periodic table, ordinarily selenium or tellurium. These atoms ordinarily form polymeric-type chains in the solid state. Their electronic structure is such that they necessarily contain large but equal densities of positively and negatively charged traps in the amorphous phase. These traps keep the film resistivity high and retard irreversible breakdown. Structural stability is maintained by using alloys with relatively large densities or cross-linking atoms, ordinarily from groups 14 and 15 of the periodic table. *See* ELECTRICAL BREAKDOWN; TRAPS IN SOLIDS.

Memory material is chosen, in general, to contain relatively weaker bonds and a smaller density of cross-linking atoms, so that structural changes are more easily attained. A class of very rapidly crystallizing memory material alloys have been used to fabricate OMS devices that can be programmed at voltages less than 3 V, with high speed (less than 1 nanosecond), extremely long cycle life (greater than 10^{13}), and the capability to be directly reprogrammed into a number of selected intermediate conductivity states. These materials can exist in the amorphous state, the fully crystalline state, and states of partial crystallinity, without compositional phase separation. Consequently, changes in the atomic structure of these materials can occur without atomic diffusion, allowing for rapid programming and very long reprogramming life.

Applications. Although Ovonic switches can be fabricated from bulk amorphous material, they are most conveniently produced as thin-film structures in which the active material and electrodes are vacuum-deposited or sputtered films, photolithographically defined. This economical process is compatible with transistor technology and with the methods used to produce passive components. It also lends itself to the fabrication of densely packed arrays.

Nonvolatile memories. Integrated arrays of OMSs can be used as electrically erasable programmable read-only memories (EEPROMs). Readout of these devices is extremely rapid, limited only by the readout circuit's characteristics; and, because of the capability of the OMS to be programmed into a number of selected conductivity states, more than one bit of data can be stored in each memory cell. This gives the OMS-based EEPROM an advantage over conventional EEPROM devices in terms of data-storage density, and consequently, cost per bit of stored information. Additionally, because of the high programming speed and very long cycle life of the device, an opportunity exists to develop unique semiconductor memory devices based on the OMS that can play the role in a computer of both nonvolatile archival memory and high-speed system memory. Availability of these devices can greatly simplify computer architecture and improve processing speed. *See* COMPUTER STORAGE TECHNOLOGY; SEMICONDUCTOR MEMORIES.

Neural networks. The OMS can be used as an electrically reconfigurable electrical interconnection in neural networks. In this application it provides a simple, high-density means to accomplish the many thousands or millions of programmable electrical interconnections required in practical neural-network circuits that can exhibit artificial intelligence. The thin-film nature of the OMS also allows for the design of multiple-layer, three-dimensional memory or neural-network circuits with data-storage or circuit-interconnect densities significantly greater than can be accomplished by using crystalline semiconductor technology. *See* NEURAL NETWORK.

Photographic films. Films have been developed in which the quality and amount of structural changes can be controlled by the amount of energy incident upon the film. These films are the only non-silver-based film with high resolution and amplification. They can also be used in updatable imaging products.

Transistors. A transistor, using an OTS as the emitter, has been developed. This can be used as a threshold amplifier, as a threshold latching amplifier, or as the basis for a computer using ternary logic. In the conducting state, OTSs can be used to provide a constant current density and to inject hot electrons into crystalline-semiconductor devices. Their threshold characteristics provide the opportunity for another important control function in these applications.

Other promising application areas include ac control where use is made of the inherent symmetry of these components, and microwave generation made possible by the inherently fast switching transition. It is anticipated that the exploration of new applications for amorphous switches will accelerate as knowledge of them becomes more widespread. *See* ALLOY STRUCTURES; AMORPHOUS SOLID; GLASS; SEMICONDUCTOR. Stanford R. Ovshinsky; David Adler

Bibliography. D. Adler, *Sci. Amer.*, 236(5):36–48, May 1977; D. Adler et al., Threshold switching in chalcogenide-glass thin films, *J. Appl. Phys.*, 51:3289–3309, 1980; S. R. Ovshinsky, *Disordered Materials: Science and Technology*, 1991; S. R. Ovshinsky, Optically induced phase changes in amorphous materials, *J. Cryst. Solids*, 141:200–203, 1992.

Glass transition

The transition which occurs when a liquid is cooled to an amorphous or glassy solid. This can occur only if the cooling rate is fast enough to prevent crystallization which would otherwise occur if time had been sufficient for the sample to reach true equilibrium at each temperature. Since the crystal is invariably the thermodynamically stable low-temperature phase, the glass transition corresponds to a transition from a high-temperature liquid into a nonequilibrium metastable low-temperature solid. Experimentally, this transition occurs at a temperature T_g where the shear viscosity becomes immeasurably large, greater than 1×10^{15} poises (1×10^{14} pascal-seconds). Changes in the heat capacity and coefficient of thermal expansion are also generally observed near T_g. *See* AMORPHOUS SOLID; CRYSTAL STRUCTURE; VISCOSITY.

Basic phenomena. The liquid state of matter is characterized by zero shear modulus and a nonvanishing fluidity. (Fluidity is the reciprocal of viscosity.) When a liquid is slowly cooled, the system usually crystallizes through a discontinuous or first-order phase transition (that is, one with nonvanishing latent heat) at a melting temperature T_m. At T_m, the fluidity vanishes abruptly, and the system is in a new, stable-ordered, solid state, that is, a crystal. However, in many systems, it is possible to supercool the liquid to temperatures below T_m by rapid cooling. In those systems in which crystallization can be avoided, several interesting phenomena occur. The fluidity is observed to decrease continuously, reaching a limiting value so small that it is unobservable and the time required to relax a shear becomes very large, of the order of days or longer. When this has occurred, large-scale flow processes cease and the material appears solid on human time scales. This new state,

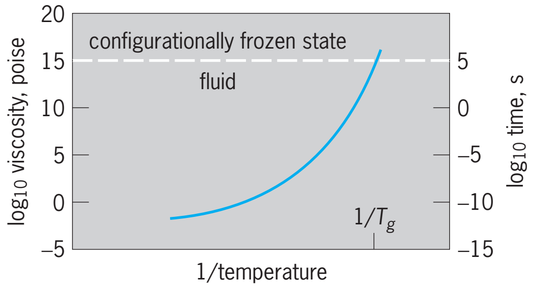


Fig. 1. Plot of the logarithms of viscosity and of the time constant for flow versus reciprocal temperature for a material cooled from the liquid to the glassy state. [$\log_{10}$ viscosity, Pa · s] = [$\log_{10}$ viscosity, poise] − 1.

which is referred to as a glass, differs from the crystal in that it is disordered and only metastable thermodynamically. Put simply, the glassy state is an extension of the liquid state, in which the viscosity (**Fig. 1**) increases above 1 × 10^{15} poises (1 × 10^{14} Pa · s). *See* PHASE TRANSITIONS.

Nonequilibrium character. All known systems are less stable in their glassy state than in some crystalline state. This is consistent with the observation that T_g is always lower (usually by a significant amount) than the melting temperature T_m. However, it is probably also true that the glass transition can occur in all liquids, provided crystallization can be bypassed. For many organic and polymeric systems, the difficulty of molecular packing and the steric hindrances are sufficient to prevent crystallization, and glass formation in these systems is relatively easy. In other systems, for example, metallic systems, rapid quench rates on the order of 1 × 10^6 K/s (2 × 10^{6}°F/s) may be necessary to avoid crystallization. For still other systems, particularly monatomic liquids with simple spherically symmetric interactions such as liquid argon, glass formation has never been observed in the laboratory. However, in computer simulations, where much faster cooling rates are achievable (1 × 10^{12} to 1 × 10^{14} K/s or 2 × 10^{12} to 2 × 10^{14}°F/s), such materials have formed glasses. This suggests that any system can be quenched from the liquid state to an amorphous glassy state assuming that the system can be cooled rapidly enough.

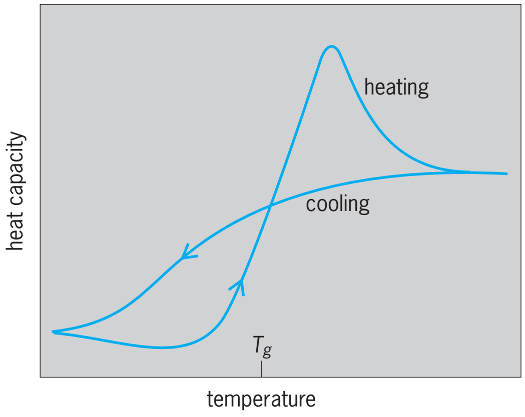


Fig. 2. Plot of heat capacity versus temperature for a system cooled and then reheated through the glass-transition region.

Variation of heat capacity. During the freezing of the supercooled liquid into a glass, the heat capacity and coefficient of thermal expansion change markedly within a narrow temperature range near T_g. This decrease in the heat capacity to a level near that of the crystalline system (**Fig. 2**) reflects a loss of configurational freedom which the sample undergoes at the glass transition. The magnitude of this drop in heat capacity in going from the equilibrium liquid to the glass is usually very large, approximately 50% for most glass formers, though it is barely detectable in the tetrahedrally coordinated network glasses silicon dioxide (SiO_2) and germanium dioxide (GeO_2). The exact value of T_g and the temperature dependence of the heat capacity depend on the heating and cooling rate of the measurement and on the thermal history of the sample. This behavior reflects that the observed changes in the heat capacity are not the result of a true equilibrium-phase transition but are a consequence of the system falling out of equilibrium when the time scale of the measurement becomes comparable to the relaxation time of the system. Below T_g, the relaxation rates are so slow that the system acts as a rigid solid, even though it is still flowing, although at a rate too slow to be measurable. *See* HEAT CAPACITY.

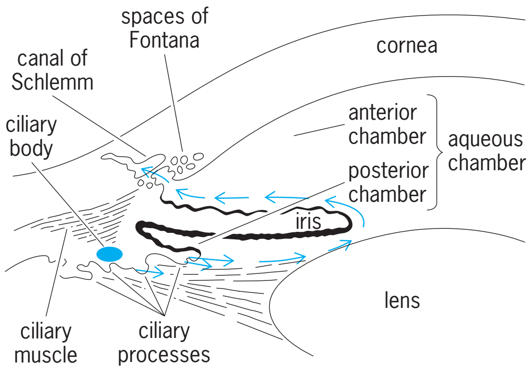
Related materials. The rapid increase in response times that characterize a dense supercooled liquid just above the glass transition has also been observed for a wide variety of disordered systems. These systems include a class of magnetic systems referred to as spin glasses, in which the interaction between the magnetic moments are "in conflict" with each other (for example, AuFe and CuMn); certain disordered insulators, in which highly localized electrons interact via the long-range Coulomb potential (the Coulomb glass); and orientational dipolar glasses (KBr:KCN), in which randomly located elastic-dipole defects interact via the strain field. These systems differ from ordinary liquids in that they do not have an ordered low-temperature phase, such as a crystal. Instead, the most stable state is disordered. However, as the temperature is lowered, these systems all reach a characteristic temperature in which their response becomes so slow that they essentially freeze, unable to reach their lowest free-energy state, very similar to the glass transition in ordinary liquids. *See* GLASS; METALLIC GLASSES; SPIN GLASS.

Gary S. Grest

Bibliography. S. R. Elliott (ed.), *Physics of Amorphous Materials*, 2d ed., 1990; J. P. Hansen, D. Levesque, and J. Zinn-Justin (eds.), *Liquids, Freezing and the Glass Transition*, 1990.

Glaucoma

A disease of the eye in which damage is caused by elevated pressure within the eye. The incidence in persons over the age of 40 is about 0.5%, making glaucoma one of the most common and serious eye disorders, surpassed in the United States only by cataracts



Pathway of circulation of aqueous humor in the eye.

as a cause of blindness. *See* CATARACT; VISUAL IMPAIRMENT.

The normal pressure within the eye measures 10–20 mmHg (1.3–2.7 kilopascals) and is maintained by a delicate balance between the inflow and outflow of fluid called aqueous humor. The aqueous humor is produced by the ciliary body and passes from the posterior chamber, that is, the area behind the iris, through the pupillary space into the anterior chamber, that is, the space in front of the iris (see **illus.**). It then drains through the trabecular meshwork into the canal of Schlemm, which leads to the exterior of the eye. *See* EYE (VERTEBRATE).

Classification. Four major types of glaucoma have been identified: (1) congenital or infantile glaucoma, which is evident at birth or shortly thereafter; (2) primary open-angle glaucoma, the most common kind, which is usually painless and is marked by a blockage in the outflow channels; (3) primary (acute) angle-closure glaucoma, in which the root of the iris interrupts drainage and causes a sudden painful blockage and acute rise in internal pressure; and (4) subacute or chronic angle-closure glaucoma, in which the root or base of the iris falls across the drain temporarily but repeatedly, resulting in transient increases in pressure but with scarring after each episode until the drain can no longer become unblocked. Secondary glaucoma results from inflammation, injury, surgery, or eye diseases such as swollen cataract.

Symptoms and effects. Infantile glaucoma often enlarges the eye because of a lack of rigidity in its white coat, known as the sclera; the condition is called buphthalmos. Angle-closure glaucoma is marked by pain, sudden visual loss, and a steamy or cloudy cornea; the semidilated and fixed pupil does not respond to changes in light intensity. In chronic angle-closure glaucoma, the periodic rises in pressure are accompanied by the above symptoms and also by halo vision, that is, the person sees haloes around a light source. In this type of glaucoma the pressure rises and falls as the angle reopens, and adhesions of scar tissue, called synechiae, form. The attacks continue until the drainage channel is sealed.

In the most common type of glaucoma, the asymptomatic open-angle form, the chief threat is a gradual, imperceptible loss of vision. The disease is bilateral, with progressive field loss mostly in the periphery where it escapes notice. For people over the age of 30 or 40, intraocular pressure should be measured every 2 or 3 years. Open-angle glaucoma is a heritable condition, but no racial or sexual patterns have been found.

Elevated intraocular pressure without apparent damage is a condition known as ocular hypertension. It should be considered a type of glaucoma. Frequent observation of the optic disk is recommended, with treatment as indicated.

Tests. A number of diagnostic tests are known. Eye pressure is measured by a tonometer, an instrument with weights and a scale. Applanation tonometry, which uses a slit lamp or special microscope, is more accurate. Gonioscopy, performed in conjunction with the slit lamp, is an examination of the drain or angle for susceptibility to sudden closure or gradual obstruction. Tonography provides a tracing, called a tonogram, which can be interpreted much like a cardiogram. It is usually obtained electronically by means of a special device attached to the eye to record the eye pressure over a 4-min period. These measurements make it possible to calculate the rate of elimination of aqueous humor from the eye. A low rate may be first sign of open-angle glaucoma.

So-called provocative tests are performed in open-angle or closed-angle glaucoma to elicit a pressure rise; they can help to provide a diagnosis when eye pressures are normal under ordinary conditions. The visual field is very important in diagnosing and following chronic glaucoma. It can be plotted and measured in various ways, from the simple tangent screen to the more sophisticated Goldmann perimeter.

Treatment. Blindness can usually be prevented by early treatment to maintain normal eye pressure. Eye drops or oral medication is usually effective, reducing the flow of aqueous humor into the eye and increasing its outflow. If medication fails, as in angle-closure glaucoma, or if sudden complete blockage occurs, surgery is indicated. Eye pressure is then lowered by creating an artificial outlet, usually with the argon or yttrium aluminum garnet (YAG) laser. The surgical procedure called peripheral iridectomy has largely been superseded by laser treatment. Congenital glaucoma, which occurs as a defect in development of the angle, must be treated surgically, however.

Open-angle glaucoma, or chronic painless glaucoma, can also be treated by trabeculoplasty with the argon laser, which is often successful. If it fails, a new outlet can be opened by a surgical procedure known as trabeculectomy. Lasers are preferred, if only because they carry less risk of infection than conventional surgery. *See* LASER. Jack Hartstein

Bibliography. G. K. Krieglstein, *Glaucoma*, 1993; M. B. Shields, *Textbook of Glaucoma*, 4th ed., 1997.

Glauconite

The term glauconite as currently used has a twofold meaning. It is used as both a mineralogic and morphologic term. The mineral glauconite is defined as an illite type of clay mineral. It is dioctahedral with considerable replacement of aluminum by iron and magnesium. Structural substitutions result in a charge deficiency in both the tetrahedral and octahedral sheets, and interlayer cations seem to balance both of these charges. Calcium and sodium as well as potassium are the interlayer cations. A fundamental characteristic of glauconite is that the unit cell is composed of a single silicate layer rather than the double layer of most other dioctahedral micas. *See* CLAY MINERALS; ILLITE.

Glauconite is known to occur in flakes and as pigmentary materials. When used in the morphological sense, the term glauconite often refers to small, green, spherical, earthy pellets. Some of these pelletal varieties are composed solely of the mineral described above, others are a mixed-layer association of this mineral and other three-layer structures.

The mineral glauconite is formed during marine diagenesis. Because of the frequent association of glauconite with organic residues, it has been generally concluded that the presence of organic material is necessary for the formation of this mineral. Glauconite forms in a marine environment from a variety of raw materials. It is known to form as an alteration product of biotite mica, when the alteration takes place under reducing conditions. The deposition of sediment must be slow during this mineralogic transformation to allow complete alteration before burial. If burial is accomplished before alteration, the biotite mica persists. *See* AUTHIGENIC MINERALS; DIAGENESIS.

Relatively shallow water is another requisite for the formation of glauconite. It has been shown that glauconite forms in the shallow sea in agitated waters which are not highly oxygenated. A reducing environment is also probably necessary for the formation of glauconite.

The magnesium content of glauconite is very uniform, and the ratio of ferric to ferrous iron is rather constant. This suggests that a critical content of magnesium and a particular oxidation-reduction potential might be required for this mineral to form. It has been suggested that certain concentrations of sodium and potassium ions are also necessary for glauconite formation, since the ratio of these cations in the interlayer positions is rather distinctive.

Glauconite readily occurs in pellet form and has been identified in both recent and ancient sediments. It is a major component in some "greensand" deposits and has been used commercially for the extraction of potassium from such sources. *See* MARINE SEDIMENTS.

Floyd M. Wahl; Ralph E. Grim

Glaucophane

A monoclinic sodic amphibole with composition close to $Na_2(Mg_3Al_2)Si_8O_{22}(OH)_2$. This mineral exhibits a characteristic blue color with distinct pleochroism from colorless to lavender blue when viewed in thin section by plane-polarized light. Outcrops of glaucophane-rich metamorphic rocks are commonly blue and tend to have good foliation; these rocks are termed blueschists. *See* AMPHIBOLE; BLUESCHIST; PLEOCHROISM.

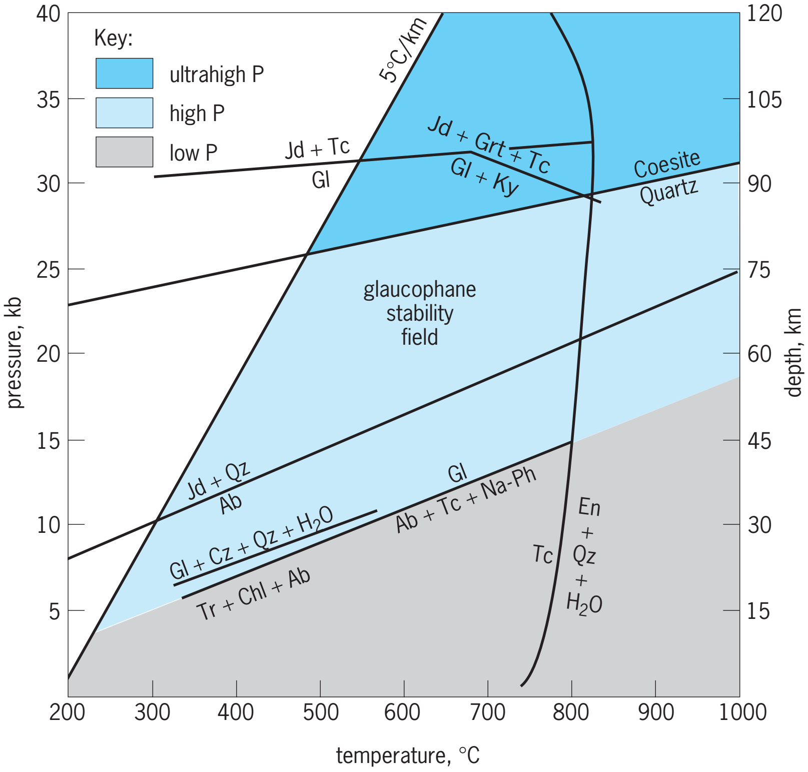
Composition. Classification of amphiboles by the International Mineralogical Association is based on the standard amphibole formula $AB_2{}^{VI}C_5{}^{IV}T_8O_{22}(OH)_2$. Sodic amphiboles are defined to include monoclinic amphiboles in which $Na_B > 1.50$. Analyses of naturally occurring glaucophanes show complex chemical substitution of octahedrally coordinated magnesium (Mg) by ferrous iron (Fe^{2+}) and of aluminum (Al) by ferric iron (Fe^{3+}). Glaucophanes have $Fe^{2+}/(Fe^{2+} + Mg)$ ratios less than 0.5 and ${}^{VI}Al > {}^{VI}Fe^{3+}$, whereas magnesioriebeckites have ${}^{VI}Al < {}^{VI}Fe^{3+}$. Sodic amphiboles with higher $Fe^{2+}/(Fe^{2+} + Mg)$ ratios include ferroglaucophanes (${}^{VI}Al > {}^{VI}Fe^{3+}$) and riebeckites (${}^{VI}Al < {}^{VI}Fe^{3+}$). Glaucophanes may also contain minor calcium (Ca) substitution of sodium (Na), with the $Ca/(Na + Ca)$ ratio less than 0.25; this ratio is between 0.25 and 0.75 for winchite-ferrowinchite, and greater than 0.75 for tremolite-actinolite. These amphiboles have a vacant A site. Substitution of $Na_A + {}^{IV}Al = {}^{IV}Si$ results in nyböite and its ferric and ferro equivalents. Nyböites, similar to other sodic amphiboles, have been reported in subduction-zone metamorphic rocks such as the Nyböeclogite from Norway and the Donghai eclogite from eastern China. The increasing use of electron microprobe analyses of natural glaucophanes during the last three decades shows that sodic amphiboles exhibit complex compositional variations even in a single crystal. The difference in various substitutions of sodic amphibole depends on the pressure-temperature conditions and rock compositions. *See* ECLOGITE.

Stability. Stability relations of glaucophane and other sodic amphiboles have been extensively investigated. Experimental studies of synthetic iron-free glaucophane suggest a very high pressure stability field (>12 kilobars or 1200 megapascals), but natural glaucophane containing 9 wt % total iron as ferrous oxide (FeO) has been experimentally shown to decompose only at pressures below 4 kbar (400 MPa) and temperatures above 550°C. Many workers studying various blueschist terranes have noted that the ferric oxide (Fe_2O_3) content of glaucophane significantly extends its stability toward lower pressures.

End-member glaucophane has not been conclusively synthesized in the laboratory, and the pressure-temperature stability fields of glaucophane-magnesioriebeckite assemblages are highly dependent on bulk rock composition (hence, the mineral assemblage). Despite problems in laboratory studies and differences in experimental results, the available data suggest that formation of glaucophane requires pressures higher than 4 kbar (400 MPa; greater than about 7.5 mi or 12 km depth) and temperatures lower than 800°C (see **illus.**). This conclusion is consistent with investigations of natural occurrences in many blueschist and eclogite terranes within the Eurasian continent, Central America, and the circum-Pacific and Himalayan-Alpine orogenic belts. Within these regions, glaucophane occurs together with lawsonite, jadeitic pyroxene, and, with increasing temperature, epidote and garnet in addition to chlorite, white mica, and stilpnomelane. These assemblages define the glaucophane-schist or blueschist facies of metamorphism. The aluminum oxide (Al_2O_3) content of the glaucophane-magnesioriebeckite assemblage coexisting with the epidote + actinolite + chlorite + albite + quartz assemblage decreases systematically with decreasing pressure, and has been used to estimate pressure (hence, depth) for blueschist formation in California, Japan, New Caledonia, and New Zealand.

Both experimental data and thermodynamic calculations show that glaucophane of the end-member composition is stable over a wide range of pressure-temperature conditions from 7 to 31 kbar at 400°C and 11 to 32 kbar at 600°C (see illus.). Reaction of glaucophane to form jadeite + talc defines the upper pressure limit of glaucophane; the equilibrium line was calculated at 30.5 kbar and 600°C with a gentle positive pressure-temperature slope. Similarly, the calculated maximum pressure of the reaction glaucophane + kyanite = jadeite + garnet + talc lies at 33 kbar at about 640°C and 30 kbar at about 750°C. Nyböite, like glaucophane, has not been synthesized from its own or other composition, and amphiboles with about 70 mol % nyböite are stable to pressures greater than 32 kbar at 600–900°C, but break down to albite + nepheline + sodian phlogopite at 15 kbar and 600–900°C. Apparently, both glaucophane and nyböite are stable within the stability field of coesite, a high-pressure polymorph of quartz. These sodic amphiboles have been identified in coesite-bearing ultrahigh-pressure metamorphic rocks from the Dora-Maira massif of the Italian Western Alps and the Dabie Mountains of central China, with the estimated minimum pressures of 27–29 kbar, equivalent to mantle depths of nearly 80–90 km. *See* MASSIF.

Tectonic significance. Glaucophane is an index mineral of blueschist, which is generated under unusually high pressures at low temperatures in a tectonic environment exclusively associated with a subducted lithospheric slab or related tectonic loading. The glaucophane-bearing assemblages occur in recrystallized graywackes and pelitic rocks and in metabasites and metacherts of oceanic affinity; they are typically found in subduction zone complexes at plate boundaries, a setting first recognized in the Jurassic and Cretaceous Franciscan Complex of northern California. Blueschists are most common and best developed in Mesozoic and Cenozoic terranes: some Paleozoic and even latest Precambrian blueschists have been described in Russia and China. Blueschists formed earlier in geologic time may have been eroded or been recrystallized under normal geothermal conditions. The preservation of glaucophane in blueschists of continental or island arc margins indicates either rapid uplift or maintenance of low geothermal gradients by steady-state subduction for tens of million years. *See* GRAYWACKE; SUBDUCTION ZONES.



Pressure-temperature diagram showing the stabilities of glaucophane and glaucophane-bearing assemblages. Tc, talc; Jd, jadeite; Gl, glaucophane; Cz, clinozoisite; Qz, quartz; Tr, tremolite; Chl, chlorite; Grt, garnet; En, enstatite; Na-ph, sodium phologopite; Ab, albite; H_2O, water. 1 kilobar = 10^2 MPa, 1 km = 0.6 mi. The geothermal gradient with increasing temperature at 5°C per kilometer depth is indicated. (*Modified after J. H. Carman and M. C. Gilbert, 1983; S. Maruyama and others, 1986; R. Y. Zhang and J. G. Liou, 1994*)

Porphyroblastic sodic amphibole is also common as a primary phase in eclogites of many continent collisional orogens, including those in the western Alps, central China, and southern Urals. However, the development of glaucophane in some eclogites, including those tectonic blocks of the Franciscan mélange in California, has been explained as the beginning of a retrograde evolution from the stable conditions of the primary eclogite paragenesis. This reaction is accompanied by a variation in the distribution of iron and magnesium between the amphibole, garnet, and omphacite. In some pre-Alpine high-grade metamorphic rocks, amphiboles of the glaucophane-magnesioriebeckite series developed in the pre-eclogite and eclogite stages of the early Alpine orogeny; they persisted into the post-eclogite stage, before being replaced by blue-green Na-Ca amphibole during the post-eclogitic and late Alpine stages. Compositions of different-stage glaucophane crystals vary even in a single rock and reflect different prograde and retrograde pressure-temperature conditions. *See* OROGENY; PORPHYROBLAST.

J. G. Liou; R. Y. Zhang; S. Maruyama

Bibliography. J. H. Carman and M. C. Gilbert, Experimental studies on glaucophane stability, *Amer. J. Sci.*, 283:A414–A437, 1983; W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 28: *Double-Chain Silicates*, 2d ed., 1997; B. E. Leake et al., Nomenclature of amphibole: Report of the

Subcommittee on Amphiboles of the International Mineralogical Association, Commission on New Minerals and Mineral Names, *Amer. Mineral.*, 82:1019–1037, 1997; J. G. Liou and S. Maruyama, Parageneses and compositions of amphiboles from Franciscan jadeite-glaucophane type facies series metabasites at Cazadero, California, *J. Metamor. Geol.*, 5:371–395, 1987; S. Maruyama, M. Cho, and J. G. Liou, Experimental investigations of blueschist-greenschist transition equilibria: Pressure dependence of Al_2O_3 contents in sodic amphiboles—a new geobarometer, *Geol. Soc. Amer. Mem.*, no. 164, pp.1–16, 1986; S. Maruyama, J. G. Liou, and M. Terabayashi, Blueschists/eclogites of the world and their exhumation, *Int. Geol. Rev.*, 38:490–596, 1996; R. Y. Zhang and J. G. Liou, Coesite-bearing eclogite in Hanah Province, central China: Detailed petrology, glaucophane stability and PT path, *Eur. J. Mineral.*, 6:217–233, 1994.

Glazing

The application of finely ground glass, or glass-forming materials, or a mixture of both, to a ceramic body and heating (firing) to a temperature where the material or materials melt, forming a coating of glass on the surface of the ware. Glazes are used to decorate the ware, to protect against moisture absorption, to give an easily cleaned sanitary surface, and to hide a poor body color.

Glazes are classified and described by the following characteristics: surface—glossy or matte; optical properties—transparent or opaque; method of preparation—fritted or raw; composition—such as lead, tin, or boron; maturing temperatures; and color. Opaque glazes contain small crystals embedded in the glass, but special glazes in which a few crystals grow to recognizable size are called crystalline glazes. A glaze may be applied during the firing; such a glaze is called salt glaze. Common salt, NaCl, or borax, $Na_2B_4O_7 \cdot 10H_2O$, or a mixture of both is introduced into the kiln at the finishing temperature. The salt evaporates and reacts with the hot ware to form the glaze. This type of glaze has been applied to sewer pipe and some fine stoneware.

The most important factor in compounding a glaze, after a suitable maturing temperature has been obtained, is the matching of the coefficient of thermal expansion of the glaze and the body on which it is applied. A slightly lower coefficient for the glaze will place it in compression (the desired condition) when the ware cools. The reverse state (with the glaze in tension because it has a higher coefficient) leads to the formation of fine hairline cracks, a condition known as crazing. *See* CERAMICS; GLASS; METAL COATINGS.

J. F. McMahon

Bibliography. E. Cooper, *The Potter's Book of Glaze Recipes*, rev. ed., 2004; G. Daly, *Glazes and Glazing Techniques*, 1995; R. Hopper, *The Ceramic Spectrum: A Simplified Approach to Glaze and Color Development*, 2d ed., 2001.

Glide-path indicator

An aircraft landing instrument that provides the pilot with a set of vertical and horizontal cross pointers that indicate deviation from a radio-transmitted course to the threshold of the runway. A dual-frequency transmitter sends out one frequency to the right of the runway centerline and a second frequency to the left of the runway centerline. The reception of these signals in the aircraft biases the vertical needle to the left or right depending on the position of the aircraft relative to the transmitted 5° localizer path. Simultaneously, another dual-frequency transmitter causes a horizontal needle to indicate high or low as the aircraft descent path is compared to the transmitted 3° glide path. This indicator, when properly used with the other navigation instruments, allows approaches to be made to within 200 ft (60 m) of altitude and 0.5 mi (0.8 km) of the runway threshold, at which time the approach transfers to a visual type, or is aborted if no visual contact is made. *See* INSTRUMENT LANDING SYSTEM (ILS).

James W. Angus

Bibliography. D. A. Lombardo, *Advanced Aircraft Systems*, 1993; B. L. Stevens and F. L. Lewis, *Aircraft Control and Simulation*, 1992.

Glider

An unpowered flying device that attempts to copy the flight of soaring birds as accurately as possible.

Early development. Otto Lilienthal made hundreds of flights with several designs of hang gliders before his fatal crash in 1896. His gliders were of rigid construction, generally without movable control surfaces, and were controlled by shifting the pilot's weight. These gliders had no landing gear other than the pilot's legs.

The next major advance in glider design was made by Wilbur and Orville Wright, who were inspired by Lilienthal. The Wrights' 1902 glider in its final version made hundreds of perfectly controlled glides and set a distance record of 622 ft (189 m) and a duration record of 26 s on the dunes of Kitty Hawk, North Carolina. This glider did not depend on weight shift for control, but had aerodynamic controls consisting of movable elevator, rudder, and wing tips. This system, in principle, has been used to the present time, even on very large and fast gliders and powered aircraft. The success with this glider led the Wrights to construct a slightly larger aircraft, with engine and propellers, that enabled them to take off and fly from level ground for the first time on December 17, 1903. The first successful powered vehicle, now called an airplane, was not a practical or useful aircraft, but with engineering development did reach that goal in a few years. *See* AIRPLANE; FLIGHT CONTROLS.

Methods of flight. In October 1911, Orville Wright made a gliding flight of nearly 10 min duration, and demonstrated that gliders could stay up for long periods in the rising air caused by the wind blowing against a sand dune or hill. This condition of flight,

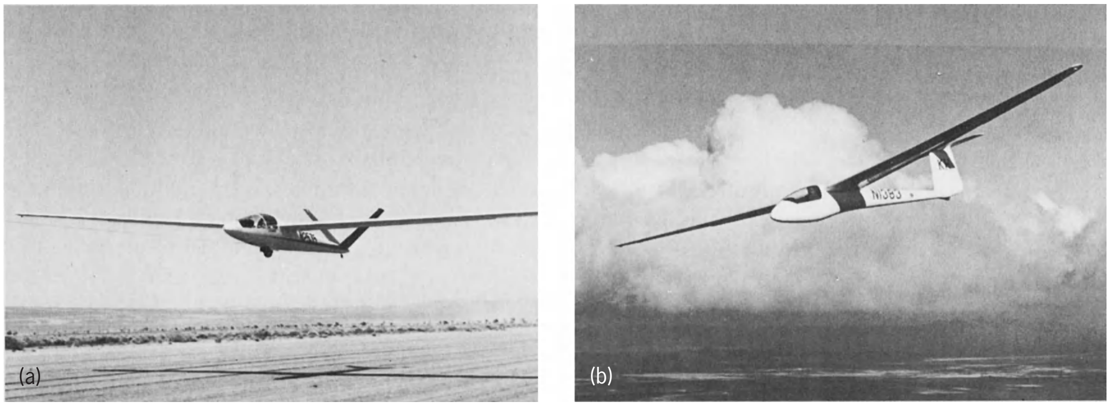


Fig. 1. **Sailplanes. (*a*) In tow during launch. (*b*) In soaring flight. (*Photographs by George Uveges*)**

called slope soaring, was the basic method of soaring flight until about 1930. Thermal soaring, the next step, was accomplished by flying in areas of rising convection currents, which are almost always present to some degree in the atmosphere. The development of thermal soaring and its practice is now principally due to the use of the variometer, calibrated as a sensitive rate-of-climb instrument, which enables the pilot to find the thermal and make the best use of it. By the use of thermal flight, the modern glider can fly almost anywhere in the world for extended time and distances over 500 mi (800 km) in one flight. Other methods of soaring make use of clouds and standing-wave phenomena in the atmosphere. These techniques have enabled gliders to achieve altitudes of 46,000 ft (14,000 m) under the proper conditions. High-performance gliders (sailplanes; **Fig. 1**) may be launched by towing behind powered aircraft to a release height of about 2000 ft (700 m), by winch-launching to about 800 ft (250 m), or by car towing, which is used to a lesser extent. Some gliders have been fitted with a small motor and propeller, which enables them to take off and climb to an altitude where rising air permits them to soar unpowered. Francis M. Rogallo

Sailplanes. Sailplane construction traditionally has been of wood and plywood, although the use of aluminum alloy has become common. The greater strength and stiffness of modern aluminum alloy permits higher aspect ratios and improved performance. The use of fiber glass as primary structure has also come into prominence, since it is possible to produce the external shapes in accurate molds with greater precision, resulting in improved performance.

Sailplanes are equipped with dive brakes on the wings for emergency descent and for landing in small areas. Properly designed brakes hold the aircraft to its "never-exceed" dive speed in a vertical dive. The laminar-flow airfoil achieves its lowest drag in a certain range of lift coefficient and corresponding speed range. By the use of properly designed flaps, this range can be shifted to higher or lower speeds at the pilot's control, which provides improved performance at a wider range of speeds. *See* AIRFOIL.

As sailplane performance increases, each new drag item becomes important, and there has been a tendency to use reclining, and almost full reclining, positions for the pilots. This permits fuselages with overall heights as low as 30 in. (75 cm) but introduces control system and visibility problems. *See* FUSELAGE.

Ernest Schweizer

Flexible-wing gliders. In the late 1950s the National Aeronautics and Space Administration (NASA) investigated various methods of returning crewed spacecraft to Earth. Two kinds of glider were investigated: aircraft with rigid delta or lifting body shapes and very high landing speeds, which later evolved into the space shuttle; and flexible-wing craft which could be packed like parachutes and deployed for slow, controlled landings in almost any open field, a concept that had been proposed 10 years earlier. After NASA demonstrated flexible-wing capability, many segments of the Department of Defense and industry also became interested in flexible-wing gliders for a variety of applications. Crewed and radio-controlled flights were made with flexible-wing gliders with or without power or towed by cars or aircraft. Some gliders were completely flexible, and some were stiffened with springy battens, aluminum tubes, or fabric tubes either pressurized or ram-inflated. Although extensive military and space applications are still undeveloped, flexible wings have had an effect on sport flying devices such as kites, hang gliders, and deployable gliders or gliding canopies used by sky divers. The completely flexible, deployable gliders have maximum glide ratios of only 3 to 4, which are very low compared to those of sailplanes, but a substantial change from the zero glide of the traditional parachute or the glide ratio of less than unity obtained from modified parachutes. Flexible, deployable gliders are used by sky divers who have demonstrated great ability to maneuver, penetrate the wind, and land on a chosen spot.

Modern foot-launched hang gliders with aluminum tube frames are the result of many improvements by private individuals and small manufacturers. Hang gliders have flown over 100 mi (160 km) in straight-line distance, have reached about 20,000 ft (6000 m) altitude, and have remained aloft more than 15 h. But it is not so much their performance that makes hang gliders popular, as their low cost, their

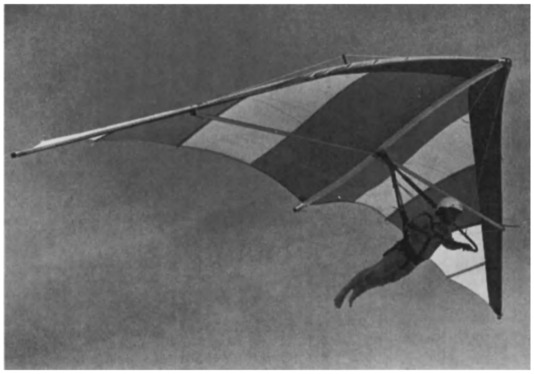


Fig. 2. Hang glider in flight. (*Photography by J. Foster Scott*)

convenience of folding into a small package for transport or storage, and the fact that no license is required for glider or pilot. Hundreds of thousands of people have learned to fly hang gliders (**Fig. 2**).

Propellers and motors of about 10 hp (7.5 kW) have been attached to some hang gliders, enabling them to take off from level ground and climb in still air if necessary, while still able to soar in updrafts with the engine off. Some hang gliders have wheels that can be used optionally. In the summer of 1979 five powered hang gliders took off from California and, after many stops for rest and refueling, landed on the east coast. *See* FLIGHT. Francis M. Rogallo

Bibliography. D. Mondey (ed.), *International Encyclopedia of Aviation*, 1977.

Global climate change

The periodic fluctuations in global temperatures and precipitation, such as the glacial (cold) and interglacial (warm) cycles of the Pleistocene (a geological period from 1.8 million to 10,000 years ago). Presently, the increase in global temperatures since 1900 is of great interest. Many atmospheric scientists and meteorologists believe it is linked to human-produced carbon dioxide (CO_2) in the atmosphere.

Greenhouse effect. The greenhouse effect is a process by which certain gases (water vapor, carbon dioxide, methane, nitrous oxide) trap heat within the Earth's atmosphere and thereby produce warmer air temperatures. These gases act like the glass of a greenhouse: they allow short (ultraviolet; UV) energy waves from the Sun to penetrate into the atmosphere, but prevent the escape of long (infrared) energy waves that are emitted from the Earth's surface. *See* ATMOSPHERE; GREENHOUSE EFFECT.

Global warming. Human-induced changes in global climate caused by release of greenhouse gases into the atmosphere, largely from the burning of fossil fuels, have been correlated with global warming. Since 1900, the amount of two main greenhouse gases (carbon dioxide and methane) in the Earth's atmosphere has increased by 25%. Over the same period, mean global temperatures have increased by about 0.5°C (0.9°F).

Of all the greenhouse gases produced by humans, the most concern centers on carbon dioxide. Not only is carbon dioxide produced in much greater quantities than any other pollutant, but it remains stable in the atmosphere for over 100 years. Methane, produced in the low-oxygen conditions of rice fields and as a by-product of coal mining and natural gas use, is 100 times stronger than carbon dioxide in its greenhouse effects. Methane, however, is broken down within 10 years.

Chlorofluorocarbon (CFC) pollution, from aerosol propellants and coolant systems, affects the Earth's climate because CFCs act as greenhouse gases and they break down the protective ozone (O_3) layer. The ozone layer normally prevents most UV radiation originating in outer space from entering the Earth's atmosphere. A thinner ozone layer allows more UV radiation to penetrate the Earth's atmosphere. CFC production has declined since the late 1970s, with further reductions imminent, but the benefit of this decrease will be realized only slowly because CFCs remain in the atmosphere up to 100 years.

Other pollutants released into the atmosphere are also likely to influence global climate. Sulfur dioxide (SO_2) from car exhaust and industrial processes, such as electrical generation from coal, cool the Earth's surface air temperatures and counteract the effect of greenhouse gases. Nevertheless, there have been attempts in industrialized nations to reduce sulfur dioxide pollution because it also causes acid rain. Since sulfur dioxide remains in the atmosphere for only a week, reduction of sulfur dioxide emissions will immediately lessen its impact on global climate. *See* AIR POLLUTION; OZONE.

Predictions. A rise in mean global temperatures is expected to cause changes in global air and ocean circulation patterns, which in turn will alter climates in different regions. While many regions have already warmed [the United States and western Europe had temperature elevations of about 0.4°C (0.7°F) during the twentieth century], some areas may experience cooling trends. Changes in precipitation have already been detected. In the United States, total precipitation has increased, but it is being delivered in fewer, more extreme events, making floods (and possibly droughts) more likely. *See* OCEAN CIRCULATION.

Range shifts. One impact of global warming on wildlife has been changes in the distribution of a species throughout the world. By analyzing preserved remains of plants, insects, mammals, and other organisms which were deposited during the most recent glacial and interglacial cycles, scientists have been able to track where different species lived at times when global temperatures were either much warmer or much cooler than today's climate. The range of most species was several hundred kilometers closer to the Poles or several hundred meters higher in elevation during times when the Earth was 4–5°C (7–9°F) warmer than it is today. Likewise, during glacial periods, species lived closer to the Equator and at lower elevations than they do now. Several studies have documented poleward and upward shifts of many plant and insect species during the current warming trend. In the western United States, the Edith's checkerspot butterfly lives, on average, 92 km (57 mi) farther north and 124 m (407 ft) farther up the mountains than it did in the early part of the twentieth century. During the same time period, many species of mountain plants have shifted to higher elevations in the Swiss Alps at rates up to 4 m (13 ft) per decade. The magnitude of these shifts in species' ranges northward and upward parallels the magnitude of warming that those regions have experienced. Climate is presumed to have been the driving force for the changes.

Phenological shifts. Changes in the timing of growth and breeding events in the life of an individual organism, called phenological shifts, have resulted from global warming. The beginning of spring is determined by length of the day and by local climatic conditions. People have long been interested in the events that mark the beginning of spring, such as blooming of the first spring flower, leafing out of trees, and nest building by birds. There have been changes in the timing of these events over the last few decades. In the Northern Hemisphere, trees such as oak, birch, and maple are leafing up to 20 days earlier. Almost one-third of British birds are nesting earlier (by 9 days) than they did 25 years ago; the other two-thirds have not changed. Five out of six species of British frog are laying eggs 2–3 weeks earlier.

Community reassembly. Community reassembly refers to changes in the species composition of communities. Communities are assemblages of interacting species living in the same area. Not all species have the same response time to environmental change For example, the Edith's checkerspot butterfly has moved almost exactly as much as predicted by the climatic change (see **illus.**). Alpine plants have lagged behind, moving at a rate of only 4 m (13 ft) per decade, when an immediate response to the warming trend should have resulted in shifts of 8–10 m (26–33 ft) per decade.

Further, not all species in a community will be equally limited by climate. Some will be more sensitive to small changes than others. Some will be primarily restricted by nonclimatic factors, such as soil type, or competitive displacement (the inability to successfully live in an area because a second species dominates local resources).

Euphydryas editha **(female), high elevation (10,000 ft or 3050 m) in the Sierra Nevada.**

These differences in response to large climatic changes can be seen in the fossil record. Because not all species were moving at the same rate, communities were broken apart, and sometimes new communities with no modern-day counterpart formed (nonanalog communities).

Extinctions. Extinction is the end of the existence of a species, but the term is also applied to the loss of a distinct subspecies or species within a given geographic area. To data, there have been no extinctions of species directly attributable to climate change. However, there is mounting evidence for drastic regional declines. The abundance of zooplankton (microscopic animals and immature stages of many species) has declined by 80% off the California coast. This decline has been related to gradual warming of sea surface temperatures. Zooplankton are a major food source for oceanic wildlife, and the decline of this food supply has been harmful to many birds, fish, and mammals. The sooty shearwater has declined by 90% since 1987. Populations of Cassin's auklet and rockfish have also decreased. A puzzling observation is that these species are not simply moving northward to colder waters where the zooplankton supply is still healthy, but seem unable to alter their behavior to respond to the changed environment. This very rigid, sedentary lifestyle does not bode well for their long-term survival, and may even be driving such species toward extinction. *See* CLIMATE HISTORY; CLIMATE MODIFICATION; CLIMATIC PREDICTION; EXTINCTION (BIOLOGY). Camille Parmesan

Bibliography. Intergovernmental Panel on Climate Change, *Climate Change 1995: 2d Assessment Report, Working Groups I and II*, Cambridge University Press, 1996; Intergovernmental Panel on Climate Change, *The Regional Impacts of Climate Change*, Cambridge University Press, 1998; P. M. Kareiva, J. G. Kingsolver, and R. B. Huey (eds.), *Biotic Interaction and Global Change*, Sinauer Associates, 1993; R. L. Peters and T. E. Lovejoy (eds.), *Global Warming and Biological Diversity*, Yale University Press, 1992.

Globe (Earth)

A sphere with a map of the world on its surface, the traditional model of the Earth. The use of a round object whose surface is equidistant at every point from the center ignores the Earth's true shape. Satellite measurements from space show that the Earth is slightly flattened at the Poles and bulges in the Southern Hemisphere. However, when scaled down to the size of a globe, these irregularities disappear. Compared with flat maps, all of which distort the Earth's surface patterns, the graphical representation produced by a spherical terrestrial globe eliminates serious inaccuracies and distortions of the relative size, direction, and shape of areas on Earth.

The world map may be drawn, engraved, or painted directly on the globe's surface, but it is more commonly a series of gores, or map segments in

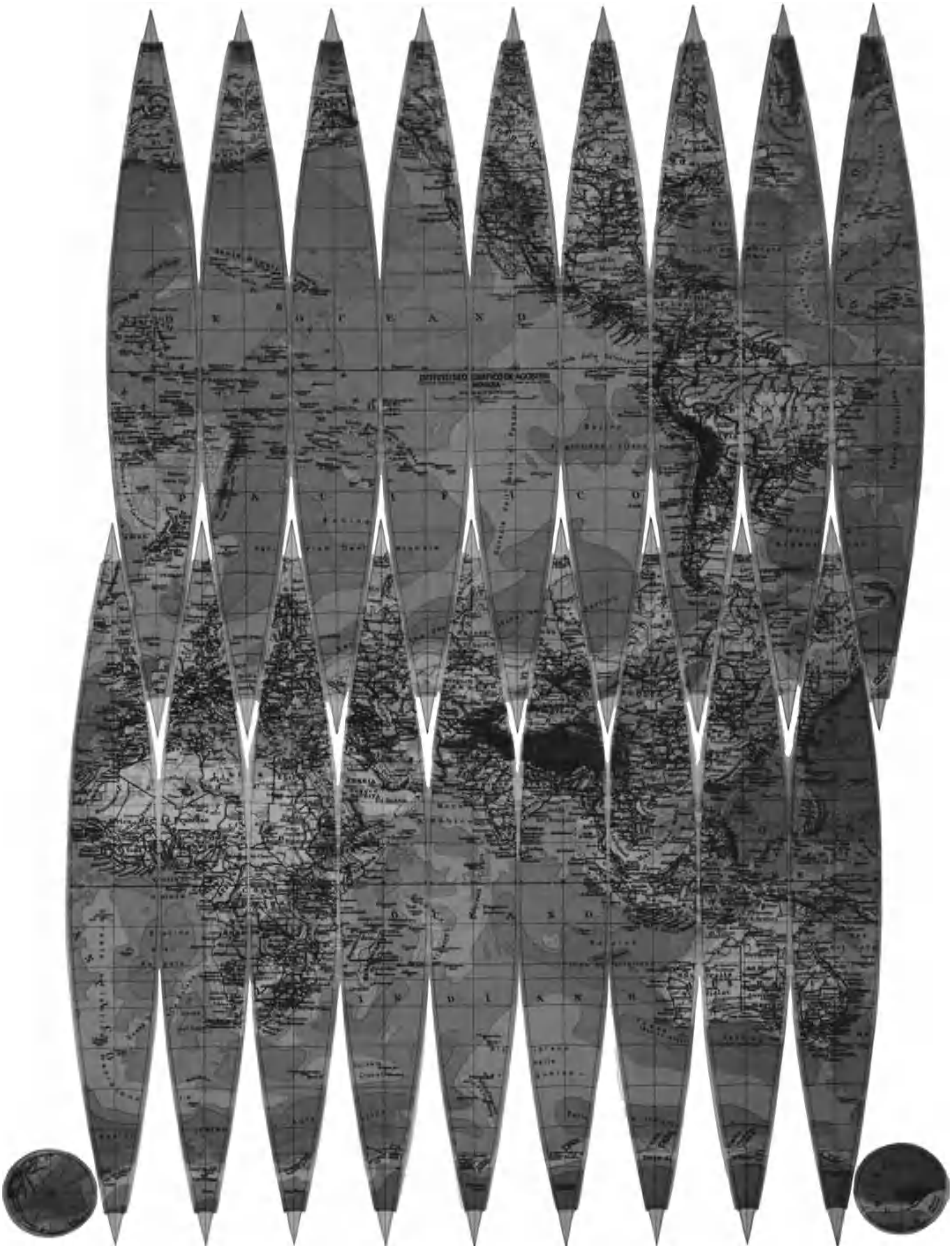


Fig. 1. Globe gores from collections of the Library of Congress. (*Istituto Geografico di Agostini, Novara, Italy*)

other designs, affixed to a cardboard, papier-maché, plastic, wooden, or metal ball (**Fig. 1**). Transparent globes with the map on the inner surface prevent soiling or wear. The surface of a raised relief globe has protrusions representing the Earth's landforms. To suggest mountains and basins, the vertical scale of these surface irregularities is exaggerated. With the development of sensors capable of measuring the ocean depths, recent globes depict relief on the ocean floor as well as on land. Globes constructed from a compilation of satellite images of Earth below depict continent shapes through the Earth's atmosphere.

Uses. Globes are artistically interesting and scientifically useful. Their value is in stimulating sound concepts of worldwide patterns, in rectifying errors induced by the limitations of flat maps, and in measuring distance and time. Long employed as aids in navigation, in teaching of earth sciences, and as library or parlor ornaments, globes are also used as toys, games, advertisements, exhibits, references for travel, and a means to illustrate missile and satellite paths. *See* CARTOGRAPHY.

Modern globes often have special attachments to enhance their utility. A meridian ring, extending from pole to pole, is calibrated in degrees to measure latitude. The longitude of points directly beneath that ring is indicated at the intersection of the ring with the equatorial scale. A horizon ring at right angles to the meridian ring may be calibrated in miles or meters, degrees, and hours to expedite distance and time measurement.

A hinged horizon ring may be lifted to serve as a meridian ring or placed in an oblique position to show great circle routes and distances. Some globes have a time disk loosely capped over the North Pole (**Fig. 2**). When set to the time of a point directly under the meridian ring, it shows the simultaneous time at other longitudes around Earth. The International Date Line is usually plotted in mid-Pacific with an analemma, shaped like a number 8, near it, which shows the latitude at which the Sun is directly overhead on each day of the year. Additional attachments convert globes into sundials, navigation instruments, and models for satellite orbits. As educational devices, globes are often lighted from the inside and may have transparent hemispheric covers on which to write. For interactive learning, a computerized globe was developed which speaks answers to questions about the location of any point on its surface that is touched. Handy inflatable plastic globes are available for the classroom. *See* INTERNATIONAL DATE LINE.

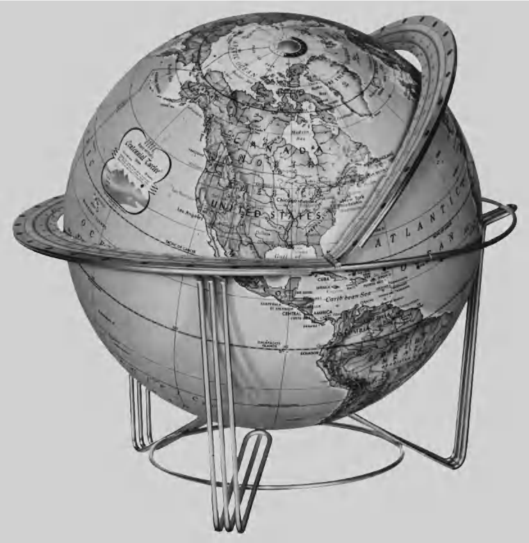


Fig. 2. Globe with hinged horizon ring. At the North Pole is a small disk which rotates independently to show simultaneous times over the Earth. (*Rand McNally and Co.*)

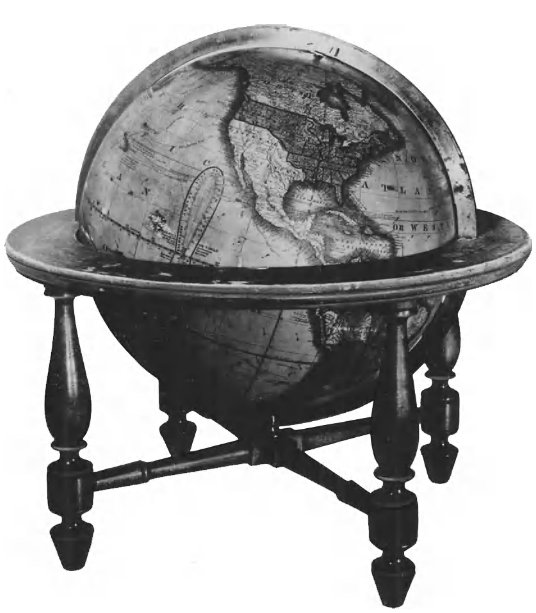


Fig. 3. Wilson globe, from collections of the Library of Congress. This globe is dated 1822. Note that the upper part of the analemma extends north of the horizon ring.

Sizes and mountings. Few globes are more than 18 in. (45 cm) in diameter, and most are 12 in. (30 cm) or smaller. A 52-ft (16-m) globe was exhibited at the Pan Pacific Exposition in 1915. At the Christian Science Publishing House in Boston, people enter a 30-ft (9-m) translucent ball to view the world map from the inside. On an outdoor motor-driven spindle, the 25-ft (8.5-m) Babson Globe was completed in 1955 in Wellesley, MA. A 75-in. (2-m) relief globe was built in 1957 by the Geo-Physical Map Co., now part of Rand McNally and Co. A 50-in. (1.3-m) globe, designed especially for White House use during World War II, had a free-rolling base. Globes in stands become decorative furniture. Others sit on tilted spindles or ride free in cradles or rings. Modern plastics and air compressors allow portable globes large enough for people to stand in to be set up for temporary display.

Famous globes. In the 2d century B.C., Greek geographer Crates of Mallus constructed the earliest known globe. The Germanic Museum in Nürnberg holds the oldest terrestrial globe in existence. This 20-in. (51-cm) original built in that city by Martin Behaim in 1492 shows more than 1100 place names, is made of pasteboard and gypsum supported by a wooden frame, and is covered with parchment. Of almost equal age is an engraved and gilded copper ball, 7 in. (17 cm) in diameter, discovered in 1850 at Laon, France. During the late Renaissance, the Church of Rome foreswore its 100-year-old ban on the idea of a round Earth, and the forging of ornate metal globes became important work in the coppersmith's craft. Other sixteenth- and seventeenth-century globes were made by European cartographers such as Mercator, Hondius, Heyden, Blaeu, Delisle, and Senex. The most publicized were made by Coronelli for Louis XIV of France. His largest, made in 1680, was 15 ft (4.5 m) in diameter, covered with engraved gores, and equipped with a door to permit approximately 30 people to stand inside. Scores of famous globes reside in the headquarters of the Coronelli-Weltbund der Globusfreunde in Vienna. The New York Public Library holds a globe made in 1506, one of the earliest constructed after the discovery of America. James Wilson built the first globes made in the Americas at Bradford, VT, in 1810 (**Fig. 3**). Near-globes, such as the icosahedron of 20 spherical triangles or the dymaxion of six squares and eight triangles, are made today.

Arch C. Gerlach; Ute J. Dymon; Nancy L. Winter

Bibliography. J. Campbell, *Map Use and Analysis*, 3d ed., 1998; J. Goss, *The Mapmaker's Art: An Illustrated History of Cartography*, 1993; A. H. Robinson, *Elements of Cartography*, 6th ed., 1994.

Globule

A small, opaque nebula seen in silhouette against a rich star field or a bright nebula. Globules were first cataloged in the 1920s. In 1947, B. J. Bok called attention to their potential significance for star formation, and since then they have been commonly known as Bok globules. A globule is a region of the interstellar medium containing a high density of interstellar grains that obscure the more distant background stars and cause the region to appear as a dark nebula in optical photographs. Only relatively nearby globules can be identified, because if there are many stars in front of the nebula the contrast with the background is too weak. The distances to such nebulae have been estimated by counting the number of stars photographed in the line of sight and comparing this count with an analogous star count in a clearer comparison field. Comparative star counts have also been used to evaluate the degree of extinction produced by the grains in the nebula; if the optical properties of the grains are known, the total number of grains in a cut through the nebula can be estimated. *See* INTERSTELLAR EXTINCTION; NEBULA.

The material contained in interstellar grains represents only a small fraction (about 1%) of the total mass of a globule; most of its mass is in gaseous form. The grains play a key role in shielding the nebular gas from the surrounding starlight, thus creating an environment in which molecules can survive and interact. Radio astronomers have been able to detect carbon monoxide (CO) emission lines in dark nebulae, and in the densest cores of nebulae even heavier molecules have been identified. Dark nebulae are now commonly called molecular clouds: the most abundant molecule is molecular hydrogen, but it is difficult to detect directly. Molecular clouds are of great interest as the regions in which stars are formed. *See* COSMOCHEMISTRY; INFRARED ASTRONOMY.

Nearby globules have been surveyed from the radio emission lines of carbon monoxide, ammonia (NH_3), and other molecules. These data, when combined with optical estimates of distances, have led to fairly accurate determinations of the dimensions of globules. Careful studies of ammonia emissions indicate that a typical (molecular core) globule has a diameter of 0.1 parsec (1 parsec = 1.9×10^{13} mi or 3.1×10^{13} km) and mass four times that of the Sun; larger, more massive globules are also known. The kinetic temperature in a globule is low; it is estimated to be about 10 K (−442°F). *See* RADIO ASTRONOMY.

Calculations predict that in the absence of internal support a typical globule undergoes gravitational collapse in less than a million years to produce one or more protostars. These young objects radiate in the infrared, a region of the spectrum to which the nebular grains are transparent. The *Infrared Astronomical Satellite (IRAS)* has mapped the sky at several infrared wavelengths, and astronomers have identified many protostars in the dense cores of molecular clouds. Some but not all globules have given birth to protostars, and these probably represent the youngest stars of the Milky Way. *See* INFRARED ASTRONOMY; STELLAR EVOLUTION. Beverly T. Lynds

Bibliography. E. Levy, J. I. Lunine, and M. S. Matthews (eds.), *Protostars and Planets*, vol. 3, 1993; B. Lynds (ed.), *Dark Nebulae, Globules, and Protostars*, 1971.

Globulin

A general name for any member of a heterogeneous group of serum proteins precipitated by 50% saturated ammonium sulfate, and thus differing from albumin, the protein present in greatest concentration in normal serum. Originally, the globulins were further subdivided on the basis of solubility (pseudoglobulin) or insolubility ("true globulins") in salt-free pure water. However, the sharp demarcations implied by these definitions do not exist since globulins show varying degrees of solubility in water and in solutions of various ionic strengths. *See* PROTEIN; SERUM.

The introduction of electrophoresis during the 1930s permitted separation of albumin from globulin on the basis of differing electric charge, and of subdivision of the globulins into alpha, beta, and gamma globulins on the basis of relative mobility at alkaline pH (8.6). However, each of these subgroups, though electrophoretically homogeneous, consists of a great variety of proteins with different biological properties and markedly different sizes and chemical properties other than net charge. Thus the α_2-globulins, for example, as defined by moving boundary or paper electrophoresis, contain proteins ranging in molecular weight from approximately 50,000 to approximately 1,000,000 (α_2-macroglobulin), each with differing functions; for example, haptoglobin binds free serum hemoglobin, ceruloplasmin is involved in copper transport and metabolism, and the α_2-macroglobulin binds or stabilizes certain enzymes, such as trypsin and plasmin. *See* BLOOD; ELECTROPHORESIS.

The beta globulins are similarly heterogeneous, containing transferrin (the iron-binding protein), various complement components, beta lipoproteins, and other biologically active proteins. The globulins of slowest electrophoretic mobility, the gamma globulins, contain a family of at least five distinct proteins, each associated with antibody activity. *See* ANTIBODY; COMPLEMENT. H. Hugh Fudenberg

Bibliography. F. W. Putnam (ed.), *The Plasma Proteins*, vols. 1–2, 1975; N. J. Russell and G. M. Powell, *Blood Biochemistry*, 1983.

Glow discharge

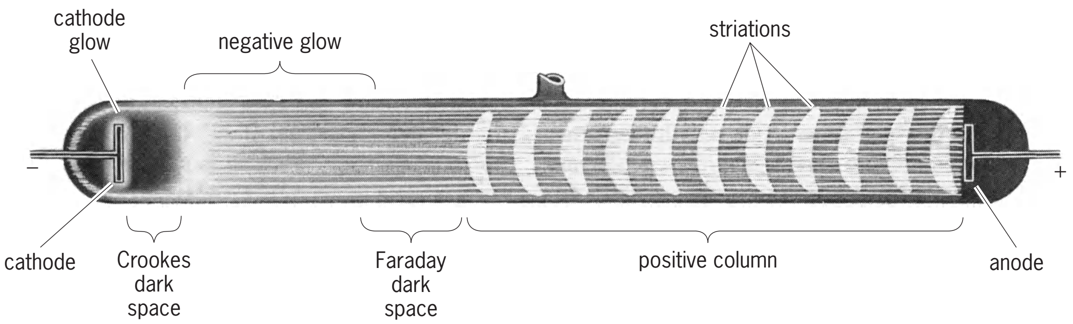
A mode of electrical conduction in gases. Glow discharge commonly occurs under conditions of relatively low pressure and generally in the pressure range of 1–10 mm of mercury (100–1000 pascals). The discharge typically gives off light, so that the region of the discharge appears to glow with considerable intensity. This glow is quite diffuse as contrasted to a higher-pressure discharge, such as a high-pressure arc. Typical currents may be of the order of tens or hundreds of milliamperes, whereas the potential drop may be of the order of 100 V.

The most important application of the glow discharge is in the so-called voltage regulator or voltage reference tube. This device maintains a relatively constant difference of potential across itself as the current is varied over an appreciable range, and consequently is very useful in cases where a constant reference potential is required.

In terms of the potential-current characteristic, the glow discharge occurs after the potential has been increased so that the Townsend region has been passed. Thus the discharge is field-sustained. On the other hand, a continued increase in current leads first to the region referred to as the abnormal glow and beyond this to the arc discharge. The transition from the abnormal glow to the arc generally is almost discontinuous and is accompanied by a spark. For a discussion of this relationship *see* ARC DISCHARGE; ELECTRIC SPARK; ELECTRICAL CONDUCTION IN GASES; TOWNSEND DISCHARGE.

Regions of discharge. There are three main regions of interest in the glow discharge, similar to those in the arc. These are the cathode fall, the positive column, and the anode region. These will be discussed separately, but it is appropriate first to examine some of the general features of the mode (see **illus.**). The appearance is that of successive more or less well-defined luminous and dark regions. Starting from the cathode, there is a dark space which generally extends for a few millimeters, the Aston dark space. This is followed by a luminous region, also of limited extent, known as the cathode glow. This is succeeded by a somewhat longer dark space, designated the Crookes or Hittorf dark space. After this comes the negative glow region, the boundaries of which are rather poorly defined. Following this is the Faraday dark space, which is also more extensive and poorly defined. This changes gradually into the positive column which is luminous and of length determined by the pressure and distance between electrodes. This region may or may not contain striations, and if present they may be either stationary or moving. At the end of the positive column is a thin layer of greater luminosity, designated the anode glow. Between this and the anode is the anode dark space.

Cathode fall. The events occurring at the cathode are vital to the discharge. The current in the cathode circuit is primarily due to positive ions. However, it is necessary to produce enough electrons at the cathode to maintain the discharge. These electrons gain energy as they move in the electric field toward the anode, and produce excitation and ionization. It appears that these electrons are secondary electrons resulting from positive ion bombardment of the cathode surface. The drop in potential which occurs at the cathode depends on the kind of gas and the cathode material. Generally, this potential drop is a large fraction of the total potential drop across the



Glow discharge at approximately 0.1 mm (13 Pa) pressure, showing successive more or less well-defined luminous and dark regions. (*After J. B. Hoag and S. A. Korff*, *Electron and Nuclear Physics*, *3d ed.*, *Van Nostrand*, *1948*)

discharge. The production of secondary electrons by this means is rather inefficient, which explains why the drop must be large.

Electrons starting at rest from the cathode must gain energy before they can produce excitation. This can be accomplished only by motion in the electric field, and hence there is a minimum distance which the electrons must move before they can produce excitation and consequent light emission. This explains the existence of the Aston dark space. It might be thought that the cathode glow could be explained by this also, but it is not likely that much of the light from this region is brought about by the secondary electrons. It appears that most of this light results from the positive ions that have struck the cathode and are returning to the ground state as neutral atoms. There are two facts of importance in this connection. First, the electron density is still rather low at such a short distance from the cathode. Second, the wavelengths present in the radiation indicate transitions involving states of a rather high degree of excitation. These high-energy states probably could not be produced by the electrons from the cathode at this point.

The Crookes dark space is actually a region of nearly uniform electric field. Most of the cathode drop occurs in this region, and here the positive ions gain most of their energy before striking the cathode. The electrons from the cathode gain enough energy here to produce cumulative ionization near the end of the space. In the negative glow, which follows, the potential is relatively constant. Here electrons, both from the cathode and from cumulative ionization, lose energy by inelastic collisions and produce a large amount of excitation. The boundary at the anode end of this space is poorly defined because of the broad distribution in electron energy. The slowing down of the electrons at the end of this region results in a negative space charge. Thus the electrons that move into the Faraday dark space gain energy.

Positive column. The beginning of the positive column is the result of excitation by these electrons. The situation in the positive column is the result of a balance between several processes. There is a nearly uniform potential drop which results in ionization throughout this entire region. On the other hand, there must be a loss of ions to make up for this production. This takes place primarily by diffusion to the walls, although there is also recombination. The electrons with their greater mobility diffuse to the walls, producing a slight negative potential relative to the center of the discharge. This negative potential both limits further electron diffusion and produces positive ion diffusion outward. This process is known as ambipolar diffusion. The positive column is not essential in maintaining the discharge. If the distance between electrodes is changed, with the pressure and current held constant, the extent of the positive column and the potential across the discharge change accordingly. The features of the anode and cathode regions remain unaltered under such a change up to the point where the positive column no longer exists.

A feature of this region is a succession of alternately luminous and dark regions, called striations, which usually occur when the discharge is operated at relatively high pressure; they may be stationary or moving. Their presence is related to the fact that in general the atomic species in the discharge are deexcited in times short compared to those required for them to diffuse through the positive column. Pure, inert gases do not show the effect, probably because they are excited into metastable, long-lived states.

Anode region. At the anode end of the positive column, the positive ions are repelled. This produces an increase in electric field, which causes the electrons to gain energy and excite more effectively. Thus the positive column ends in a region of increased luminosity, the anode glow.

Other aspects. There are many other aspects of the glow discharge that are interesting and important. One such phenomenon is cathode sputtering. Here the positive ions that are accelerated into the cathode knock out atoms or groups of atoms from the surface. Another aspect is that of abnormal glow. The voltage across the discharge remains nearly constant while the current is increased in the normal glow mode. This current increase is accompanied by an increase in the area of the cathode glow. When the glow has completely covered the cathode, a further current increase results in an increase in the cathode potential drop, and hence the potential drop across the discharge. This is the abnormal glow. It is characterized by more intense light emission and increased sputtering. *See* SPUTTERING.

It should be stated that many of the details of the discharge are uncertain. The processes are generally quite complicated. Reliable and accurate measurements are difficult at best, and most of the information is of a qualitative nature. Glenn H. Miller

Bibliography. B. Chapman, *Glow Discharge Processes: Sputtering and Plasma Etching*, 1980; M. N. Hirsh and H. J. Oskam (eds.), *Gaseous Electronics*, vol. 1: *Electrical Discharges*, 1978; R. K. Marcus (ed.), *Glow Discharge Spectroscopies*, 1993; Yu. P. Raizer, *Gas Discharge Physics*, 1991, reprint 1997.

Glucagon

The protein hormone secreted by the alpha cells of the pancreas which is known to influence a wide variety of metabolic reactions. Glucagon, along with insulin and other hormones, plays a role in the complex and dynamic process of maintaining adequate supplies of blood sugar (glucose). Glucagon has often been called the hyperglycemic-glycogenolytic factor because it causes the breakdown of liver glycogen to sugar (a process known as glycogenolysis) and thereby increases the concentration of sugar in the bloodstream (a condition known as hyperglycemia). Glucagon may also be involved in the regulation of protein and fat metabolism, gastric acid secretion and gut motility, excretion of electrolytes (such as sodium, potassium, and chloride) by the kidney, contractility of heart muscle, and release of insulin from the pancreas. On the basis of present evidence, glucagon may be considered to act in a close and subtle cooperation with other hormones and with enzymes to help control the dynamic processes of life. Glucagon is used in human medicine chiefly in certain diabetic conditions when a dangerously low blood sugar must be rapidly raised. *See* CARBOHYDRATE METABOLISM; DIABETES; ENDOCRINE MECHANISMS; ENZYME; GLYCOGEN; INSULIN; LIVER; PANCREAS.

Discovery and isolation. The history of glucagon is closely interwoven with that of insulin. During 1922–1927 several investigators were surprised to observe that injections of crude preparations of insulin caused an unexpected, rapid hyperglycemia followed by the expected drop in blood sugar. Interest in the dramatic, lifesaving effects of insulin greatly overshadowed this hyperglycemic activity found in extracts of pancreas. However, J. R. Murlin and his associates investigated the hyperglycemic effect, became convinced that it was caused by another substance in the pancreas, and termed that substance glucagon, meaning "mobilizer of sugar." Interest in this "contaminant" of insulin lay partially dormant for about two decades until 1953, when A. Staub, L. Sinn, and O. K. Behrens succeeded in purifying and crystallizing glucagon. The availability of the pure hormone led to an intensive study of its chemical and biological properties.

Assay. All measurements of glucagon potency are related arbitrarily to the crystalline hormone; no international standard has been designated as yet. The specific biological potency is usually determined by injecting glucagon intravenously or subcutaneously into cats or rabbits and measuring the degree of hyperglycemia. As little as 0.2–0.3 microgram injected intravenously into a fasted adult cat causes an increase (within 15 min) of 30–50 milligrams of sugar per 100 milliliters of blood. Glucagon may also be measured by a means of a "radioimmune" assay, in which radioactively labeled glucagon and antibodies to glucagon are employed. *See* BIOASSAY.

Chemistry. Glucagon is a single-chain polypeptide composed of 29 amino acid residues (**Fig. 1**), and has a molecular weight of 3485. The complete chemical synthesis of glucagon was achieved late in 1967.

Despite the fact that glucagon and insulin are difficult to separate, showing a similarity in some chemical and physical properties, the actual chemical composition and amino acid sequences of the two hormones are quite different. Glucagon is only slightly soluble in water between pH 4 and pH 8. It forms threadlike polymers, called fibrils, when heated in dilute acids. The hormone is very rapidly cleaved by enzymes such as trypsin and chymotrypsin, leading to loss of hyperglycemic activity. This is the chief reason that glucagon must be injected rather than taken by mouth. William W. Bromer

Biological activity. The mechanism by which glucagon brings about an increase in blood sugar level by inducing the breakdown of liver glycogen has been the subject of extensive research over the past few decades. Glucagon molecules bind to glucagon-specific protein receptors in the membrane of target cells. When these receptors bind glucagon, they interact with other proteins in the cell membrane called G proteins and "activate" these protein switches. The activated G proteins then activate an enzyme on the inner surface of the cell membrane called adenylate cylase. Activated adenylate cyclase then produces a small molecule called 3′,5′-cyclic adenosine monophosphate (cAMP), commonly referred to as a second messenger molecule (glucagon being the first messenger in this case). It is cAMP that activates the glycogen breakdown to sugar in the liver cells. *See* SIGNAL TRANSDUCTION.

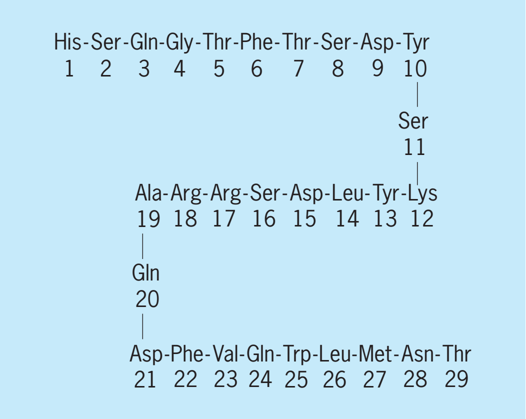

His-Ser-Gln-Gly-Thr-Phe-Thr-Ser-Asp-Tyr
1 2 3 4 5 6 7 8 9 10
Ser
11
Ala-Arg-Arg-Ser-Asp-Leu-Tyr-Lys
19 18 17 16 15 14 13 12
Gln
20
Asp-Phe-Val-Gln-Trp-Leu-Met-Asn-Thr
21 22 23 24 25 26 27 28 29

Fig. 1. Amino acid sequence in a molecule of glucagon. The abbreviations indicate: histidine (His), serine (Ser), glutamine (Gln), glycine (Gly), threonine (Thr), phenylalanine (Phe), aspartic acid (Asp), tyrosine (Tyr), lysine (Lys), leucine (Leu), arginine (Arg), alanine (Ala), valine (Val), tryptophan (Trp), methionine (Met), and asparagine (Asn).

Cloning and analysis of the gene encoding glucagon has led to the discovery of two glucagon-like peptides called GLP-1 (glucagon-like peptide 1) and GLP-2 (glucagon-like peptide 2). In recent years, research has focused on production and biological activity of GLP-1 and GLP-2. Glucagon, GLP-1, and GLP-2 are all encoded by the same gene. This gene produces a protein called proglucagon in cells in the pancreas, gastrointestinal tract, and brain. In the pancreas cells, proglucagon is processed and converted into glucagon. In the gastrointestinal tract and brain cells, proglucagon is processed and converted into GLP-1 and GLP-2 (**Fig. 2**). *See* GENE ACTION.

GLP-1. GLP-1 has the opposite effect of glucagon. It lowers blood sugar levels by a variety of means: it stimulates the production of insulin, which enhances

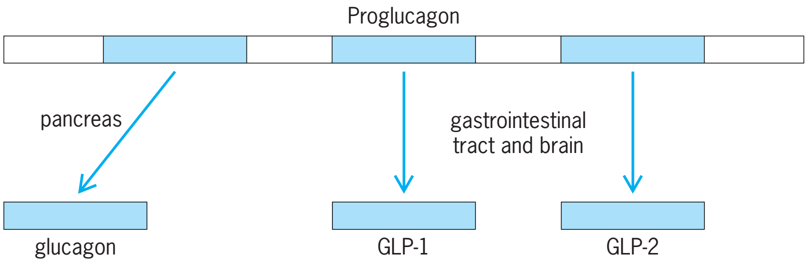


Fig. 2. Proglucagon protein is processed in the pancreas to yield the glucagon peptide and is processed in the gastrointestinal tract and brain to yield the GLP-1 and GLP-2 peptides.

the consumption of sugar by tissues; it inhibits gastric secretions, which decreases the absorption of sugars following meals; and it works within the central nervous system to reduce appetite, which also leads to low sugar levels. The ability of GLP-1 to stimulate insulin secretion has generated a great deal of interest in the use of GLP-1 in therapies for type II diabetes, in which insulin secretion is impaired. Studies have shown that injections of GLP-1 indeed stimulate insulin secretion, but the effect is short lived. Currently, extensive work is under way to find modifications of GLP-1 that will extend its half-life and make GLP-1 an effective therapy for type II diabetes.

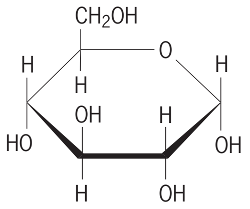
GLP-2. Less is known regarding the role of GLP-2. It is an intestinal growth factor that stimulates the division of the cells of the gastrointestinal lining. Thus GLP-2 enhances nutrient absorption during digestion. Current research is aimed at using GLP-2 in clinical treatments for intestinal damage and other related problems. M. Todd Washington

Bibliography. T. M. Devlin (ed.), *Textbook of Biochemistry with Clinical Correlation*, 5th ed., Wiley, New York, 2001; D. Voet and J. G. Voet, *Biochemistry*, Wiley, New York, 1995.

Glucose

A simple sugar, or monosaccharide, with the chemical formula $C_6H_{12}O_6$; also called dextrose, glucopyranose, grape sugar, and corn sugar. *See* MONOSACCHARIDE.

Chemistry. Glucose is a carbohydrate with six carbon atoms and has the structure



Glucose is synthesized in nature by photosynthetic organisms from carbon dioxide and water according to the general equation:

$$\underset{\text{Water}}{6H_2O} + \underset{\text{Carbon dioxide}}{6CO_2} + \underset{\text{Energy}}{\text{sunlight}} \longrightarrow \underset{\text{Glucose}}{C_6H_{12}O_6} + \underset{\text{Oxygen}}{6O_2}$$

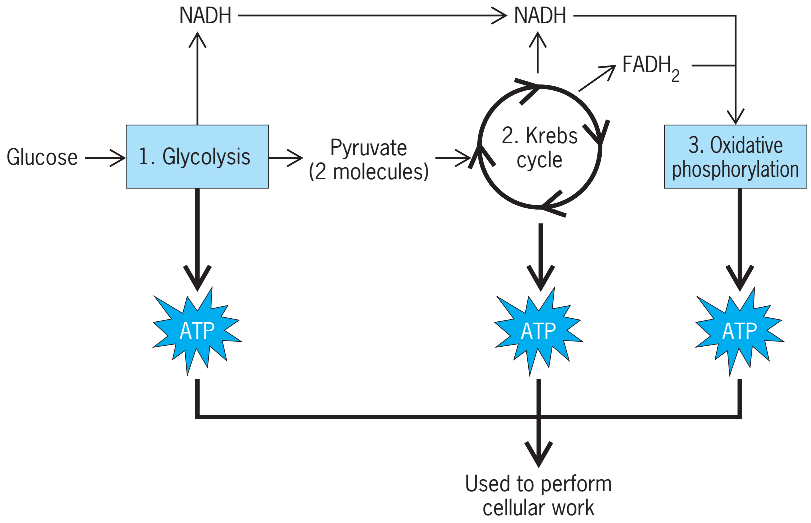
It is stored as starch in plants and as glycogen in animals. Starch and glycogen are polymers of glucose, in which individual glucose molecules have been strung together in long chains that efficiently store the chemical energy contained in the individual glucose molecules. In addition, glucose molecules serve a variety of functions in the natural world. Many of these uses are structural. For example, cellulose, which forms the cell walls of plants, is a polymer of glucose. In addition, the exoskeleton of insects and the cell walls of fungi are formed from a chemical derivative of glucose that has been assembled into chitin, a lengthy polymer. In its various polymeric forms, glucose is likely the most abundant organic molecule in the biosphere. *See* CARBOHYDRATE; CELLULOSE; GLYCOGEN; STARCH.

Metabolism. Glucose is a major source of energy for humans and other animals. However, our bodies do not maintain a large reservoir of individual glucose molecules. Rather, animals store glucose in their bodies in the form of glycogen, a large polymer of glucose found primarily in liver and muscle tissue. When our bodies need energy, for example between meals or during periods of extended fasting, individual glucose molecules can be extracted from glycogen. In the form of glycogen, glucose is securely stored as a large and stable polymer, yet it is readily accessible when needed by the body. After being released from glycogen stores, glucose molecules travel in the bloodstream to the various tissues, where they are absorbed into individual cells. Once inside a cell, glucose undergoes a series of chemical reactions that allows the cell to tap the energy contained within its chemical bonds. This process is referred to as the catabolism, or breakdown, of glucose. The overall equation for the catabolism of glucose is

$$\underset{\text{Glucose}}{C_6H_{12}O_6} + \underset{\text{Oxygen}}{6O_2} \longrightarrow \underset{\text{Carbon dioxide}}{6CO_2} + \underset{\text{Water}}{6H_2O} + \underset{\text{Heat and ATP}}{\text{energy}}$$

Since the catabolism of glucose involves the consumption of oxygen, this process is also called cellular respiration. The energy in the above equation represents heat energy, which is dissipated and lost, and energy from the chemical bonds of glucose that has been captured in the chemical bonds of the organic molecule adenosine triphosphate (ATP). ATP is a convenient energy source that can be used to drive a variety of chemical reactions. For example, cells harness the energy contained in the chemical bonds of ATP to perform many kinds of cellular work, such as the contraction of muscles and the biosynthesis of organic molecules. *See* ADENOSINE TRIPHOSPHATE (ATP).

The overall equation for glucose catabolism shown above takes place in cells as a series of chemical reactions that are grouped into three distinct steps. In the first step, glucose undergoes glycolysis, that is, the splitting of glucose. In this process, one glucose molecule (which contains six carbon atoms) is split into two molecules of pyruvate (which contains three carbon atoms). Glycolysis allows the cell to transfer some of the energy contained within glucose to ATP and another organic molecule called NADH. In the second step, the pyruvate molecules enter another series of chemical reactions, called the Krebs cycle, which allows the cell to capture more of the energy of glucose in the form of NADH and yet another organic molecule called $FADH_2$. The NADH and $FADH_2$ molecules produced during glycolysis and the Krebs cycle are coupled to the synthesis of ATP by a process called oxidative phosphorylation, which is the third and final step of glucose catabolism. The complete catabolism of glucose generally yields about 36 molecules of ATP per molecule of glucose. The catabolism of glucose is summarized schematically in the **illustration**. *See* CARBOHYDRATE METABOLISM; ENERGY METABOLISM; METABOLISM.



Breakdown (catabolism) of glucose. Glucose is catabolized in three steps: The final product is ATP (adenosine triphosphate), which is an organic molecule that cells use as a source of energy to perform cellular work.

Disease. The brain and other organs depend on glucose as a source of energy. Humans and other animals do not generally consume large amounts of free glucose molecules from their diets. Rather, glucose is usually ingested as starch, which is a polymer of glucose, or as sucrose (table sugar), lactose (milk sugar), and maltose (malt sugar). The latter three sugars are disaccharides, in which a glucose molecule has been chemically linked to another carbohydrate molecule. Our bodies have the ability to convert the starch and disaccharide molecules into individual glucose molecules. Since glucose is shuttled between the tissues of our body by the cardiovascular system, the concentration of glucose in the blood is carefully regulated in order to meet the energy demands of the brain and other tissues. Indeed, humans and other mammals maintain remarkably stable blood glucose concentrations despite sporadic food intake. Between meals, blood glucose levels decline and the liver produces a compensatory amount of glucose from glycogen and other sources; after eating a meal, glucose is absorbed from the gut into the bloodstream. The hormone insulin causes this circulating glucose to be taken up by skeletal muscle and adipose (fat) tissue. This complicated three-way balance involving glucose absorption from the small intestine, output from the liver, and uptake by muscle and fat is normally carefully regulated such that the concentration of glucose in the blood is maintained in the range of 70–120 milligrams glucose per deciliter of blood (4–7 millimolar) in humans. Deviations from this narrow concentration range can be very detrimental to health. For example, low blood glucose levels can lead to seizures, coma, and even death due to in adequate supply of glucose to the brain. In contrast, a prolonged elevated blood glucose level, as occurs in the diabetic state, can result in many disease complications. These include kidney failure, blindness, nerve damage, and cardiovascular disease.

Of the several types of diabetes, Type 1 or insulin-dependent diabetes mellitus (IDDM) and Type 2 or non-insulin-dependent diabetes mellitus (NIDDM) are the most prevalent. Type 1 occurs when the body's own immune system mistakenly attacks and destroys a specific type of cell located in the pancreas called the beta cell. The beta cells produce insulin, so the loss of these cells means that the body can no longer make its own insulin. Type 1 diabetics must carefully monitor their own blood glucose levels and take insulin, usually by injection, prior to eating. Type 2 diabetes mellitus is a complicated disease that usually develops over time. Early in the disease progression, muscle and fat tissue no longer respond appropriately to insulin and fail to efficiently absorb glucose. To compensate, the beta cells increase their output of insulin in an effort to encourage muscle and fat cells to take up glucose. All too often, however, the extra demands placed upon the beta cells cause them to wear out. Indeed, over time the beta cells may become unable to produce enough insulin to coax muscle and fat cells to absorb glucose from the blood. This condition results in the buildup of glucose in the blood associated with diabetes and may lead to disease complications. *See* DIABETES; INSULIN.

Glucose transporter. Facilitated glucose transport is the movement of glucose across cell membranes that is driven by the glucose concentration gradient, but assisted (facilitated) by carrier proteins. It is energy-independent, and it is stereospecific in that only the D-glucose isomer is transported; the L-glucose isomer is excluded. This process occurs in all mammalian cells and is essential for the maintenance of whole-body glucose metabolism and energy balance. Currently, there are five established functional facilitative glucose transporters in mammalian cells, termed GLUT1, GLUT2, GLUT3, GLUT4, and GLUTx. Each of these transporters has distinct but overlapping tissue distributions, which underscore their specific physiologic function.

GLUT1 is generally expressed and is thought to be responsible for the basal (minimum) uptake of glucose. GLUT2 is predominantly expressed in the liver and pancreatic beta cells, where it functions as part of a sensor that mediates hepatic glucose output during states of fasting and insulin secretion in the postprandial (after a meal) absorption state. In contrast, neurons primarily express the relatively high-affinity GLUT3 isoform necessary to maintain high rates of glucose metabolism for energy production. GLUTx appears to provide important function during early embryogenesis, whereas GLUT4 is exclusively expressed in insulin-responsive tissues, adipose tissue, and striated (skeletal) muscle. These latter tissues provide the key functional elements responsible for the insulin stimulation of glucose uptake, and are key targets for disregulation in states of insulin resistance and diabetes.

Jeffrey Pessin; Robert Watson

Bibliography. G. I. Bell and K. S. Polonsky, Diabetes mellitus and genetically programmed defects in beta-cell function, *Nature*, 414:788–791, 2001; M. A. Lazar, How obesity causes diabetes: Not a tall tale, *Science*, 307:373–375, 2005; S. O'Rahilly, I. Barroso, and N. J. Wareham, Genetic factors in Type 2 diabetes: The end of the beginning?, *Science*, 307:370–373, 2005; C. J. Rhodes, Type 2 diabetes—a matter of β-cell life and death?, *Science*, 307:380–384, 2005; R. T. Watson, M. Kanzaki, and J. E. Pessin, Regulated membrane trafficking of the insulin-responsive glucose transporter 4 in adipocytes, *Endocr. Rev.*, 25:177–204, 2004; P. Zimmet, K. G. M. M. Alberti, and J. Shaw, Global and societal implications of the diabetes epidemic, *Nature*, 414:782–787, 2001.

Gluons

The hypothetical force particles believed to bind quarks into strongly interacting particles. Theoretical models in which the strong interactions of quarks are mediated by gluons have been successful in predicting, interpreting, and explaining many phenomena in particle physics, but free gluons remain undetected in experiments (as do free quarks). According to prevailing opinion, an individual gluon cannot be isolated.

Color. In 1961 M. Gell-Mann and Y. Ne'eman independently suggested that the strong (nuclear) interaction respected the unitary symmetry SU(3) and that the strongly interacting particles called hadrons could be classified according to the patterns prescribed by SU(3). The family groups, or supermultiplets, that emerged were confined to a few of the simplest possibilities permitted under SU(3) symmetry. Mesons, the hadrons with integral spin in units of $\hbar$ (Planck's constant h divided by 2π), occur only in families with 1 or 8 members. The baryons, which possess half-internal spin, fit into groups with 1, 8, or 10 members. Gell-Mann and G. Zweig separately showed in 1963 that this circumstance could be explained by the hypothesis that hadrons were composites of fundamental constituents that have come to be called quarks. In this quark model of hadrons, a meson is composed of one quark and one antiquark, and a baryon is composed of three quarks. All the hadrons then known could be built out of three different varieties (or flavors) of quarks, denoted up, down, and strange. Subsequent experiments have revealed the existence of three more flavors of heavy quarks: charm, bottom, and top. To account for the observed pattern of mesons and baryons, quarks must be spin-$^1/_2$ particles. *See* QUARKS; UNITARY SYMMETRY.

Although these rules reproduce the properties of the observed hadron states, they lead to a theoretical inconsistency. The characteristics of the unstable hadron resonance known as Δ^{++} (1232 MeV/c^2), which decays into a proton and a positively charged pi meson, require that it be composed of three quarks in a configuration that is symmetric under the interchange of any pair of quarks. However, according to the Pauli exclusion principle (which first emerged in the description of atomic structure), identical spin-$^1/_2$ particles cannot occupy the same quantum state. The quark model could be brought into agreement with the Pauli principle, without compromising any of its successes, if a new attribute were ascribed to the quarks that would make the three up quarks distinguishable. For fanciful reasons, this new attribute is now known as color, though it has no connection with the color of visible light. Quarks are said to come in three colors, most frequently given the arbitrary labels red, blue, and green. A Δ^{++} resonance composed of one red up quark, one blue up quark, and one green up quark will then have the observed properties and be consistent with the laws of quantum mechanics. In this picture the antiparticle of a red up quark is an anti-red anti-up quark, so that the mesons are described as colorless quark-antiquark pairs.

Support for the idea that each quark flavor comes in three distinguishable colors has come from many quarters including the rate at which strongly interacting particles are produced in electron-positron annihilations. Theoretical predictions for such observables are sensitive to the number of distinct quark species and thus to the number of colors.

The fundamental particles that do not experience strong interactions are the leptons, which like the quarks are spin-$^1/_2$ particles that are structureless at the current limits of resolution. The most familiar examples of leptons are the electron, the muon, and the neutrinos. Each lepton flavor comes in but a single species, which is to say that leptons are colorless. It is therefore appealing to regard color as the strong-interaction analog of the electric charge. Like electric charge, color cannot be created or destroyed in any of the known interactions; it is said to be conserved. *See* COLOR (QUANTUM MECHANICS); LEPTON.

Gauge symmetry. The existence of a conserved quantity is quite generally a consequence of a continuous group of symmetry transformations that leave the laws of physics unchanged in form. For example, the conservation of energy follows from the fact that physical laws depend upon the time interval between occurrences, and not upon an absolute time measured on some master clock. Translation in time (that is, the resetting of clocks) is a symmetry of the equations of physics. Symmetries relating to internal properties of particles, like electric charge, are known as internal symmetries. In the case of conservation of the color charge, a natural choice is the unitary group SU(3), now applied to color rather than flavor. A conservation law follows, by Noether's theorem, from invariance under a global (position-independent) continuous symmetry. If the equations of physics are required to be invariant in form under local symmetry transformations that may be different at every point in space and time, the interactions related to the symmetry are completely fixed. The manner in which this could be accomplished for any continuous symmetry was indicated by C. N. Yang and R. L. Mills in 1954. *See* SYMMETRY LAWS (PHYSICS).

Yang-Mills theory. It frequently happens that the symmetries respected by a phenomenon are recognized before a complete theory has been developed. The question thus arises as to whether a complete theory of nuclear forces could be deduced from a knowledge of the symmetry. Nuclear forces had long been known not to distinguish between the proton and neutron. From the point of view of nuclear forces, the designations proton and neutron are purely conventional. This symmetry among protons and neutrons is called isospin invariance. Yang and Mills investigated the consequences of the hypothesis that the nuclear force among protons and neutrons could be derived by imposing local isospin invariance (which is to say that the convention could be chosen independently at every point of space and time). In general, the requirement of local gauge invariance implies that the interaction must occur through the exchange of massless spin-1 bosons. One species of force particle corresponds to each conserved quantity. This made the Yang-Mills theory unacceptable as a description of nuclear forces. It predicted that nuclear forces were mediated by three massless "gauge bosons," whereas the short range (on the order of 10^{-15} m) over which nuclear forces are observed to act demands that the force particles be massive, as proposed by H. Yukawa. *See* GAUGE THEORY; I-SPIN; QUANTUM FIELD THEORY.

Quantum chromodynamics. Applying similar reasoning to the idea that a local color gauge symmetry should prescribe the strong interaction among quarks leads to the gauge theory of strong interactions that has been called quantum chromodynamics (QCD). The mediators of the strong interaction are eight massless vector bosons, which are named gluons because they make up the "glue" that binds quarks together. It is hoped that the infinite range of the forces mediated by the gluons may help to explain why free quarks have not been isolated. The gluons themselves carry color. Hence, strong interactions among gluons will also occur through the exchange of gluons. It is therefore believed that gluons, as well as quarks, may be permanently confined. According to this view, only colorless objects may exist in isolation. The notion of color confinement is supported by many calculations, notably those made in the discrete (lattice) version of QCD. A complete analytic proof has not yet been constructed. *See* QUANTUM CHROMODYNAMICS.

Experimental evidence. No evidence has been reported for isolated or free gluons. As indicated above, the current interpretation of quantum chromodynamics is that free gluons cannot exist. Therefore it is necessary to devise indirect means to test the idea that gluons exist with all the desired attributes.

Inelastic electron-proton scattering. Early support for the existence of an electrically neutral glue within the proton came from 1968 experiments on inelastic electron-proton scattering carried out at the Stanford Linear Accelerator Center (SLAC). These experiments indicated that the electrons were not scattered electromagnetically from the proton as a whole, but from individual pointlike charged objects subsequently identified with the quarks. They also showed that only about half of the energy of a rapidly moving proton is carried by its charged constituents. The remainder must then be borne by neutral constituents that do not interact electromagnetically. This role would naturally be played by the gluons.

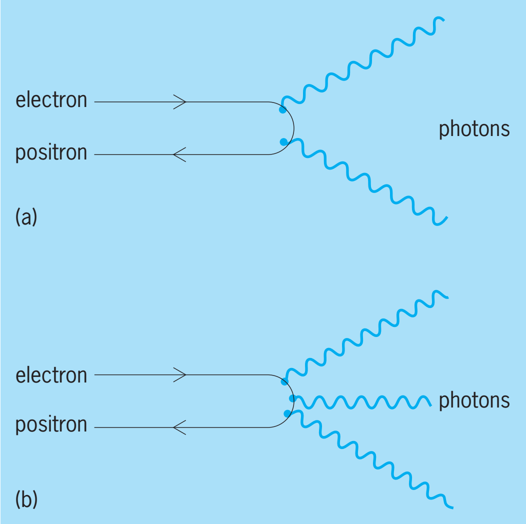


Fig. 1. Positronium decay. (*a*) Decay of parapositronium into two photons. (*b*) Decay of orthopositronium into three photons.

Charmonium lifetimes. Further evidence for the utility of the gluon concept was provided by the unusually long lifetime of the strongly decaying charmonium state J/ψ. In quantum electrodynamics the atom composed of an electron and an antielectron (positron) is known as positronium. Positronium occurs in two forms: orthopositronium, in which the electron and positron spins are aligned, and parapositronium, in which the spins are opposite. The electron and positron may annihilate into photons. The spinless parapositronium state may decay into two photons (**Fig. 1***a*), but the spin-1 orthopositronium state must decay into three photons (Fig. 1*b*). The difficulty of radiating an additional photon is reflected in the fact that orthopositronium lives

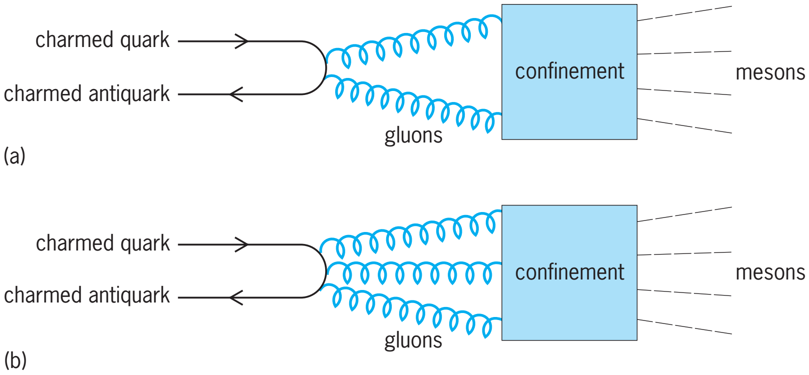


Fig. 2. Charmonium decay. (*a*) Decay of paracharmonium through a two-gluon semifinal state. (*b*) Decay of orthocharmonium through a three-gluon semifinal state.

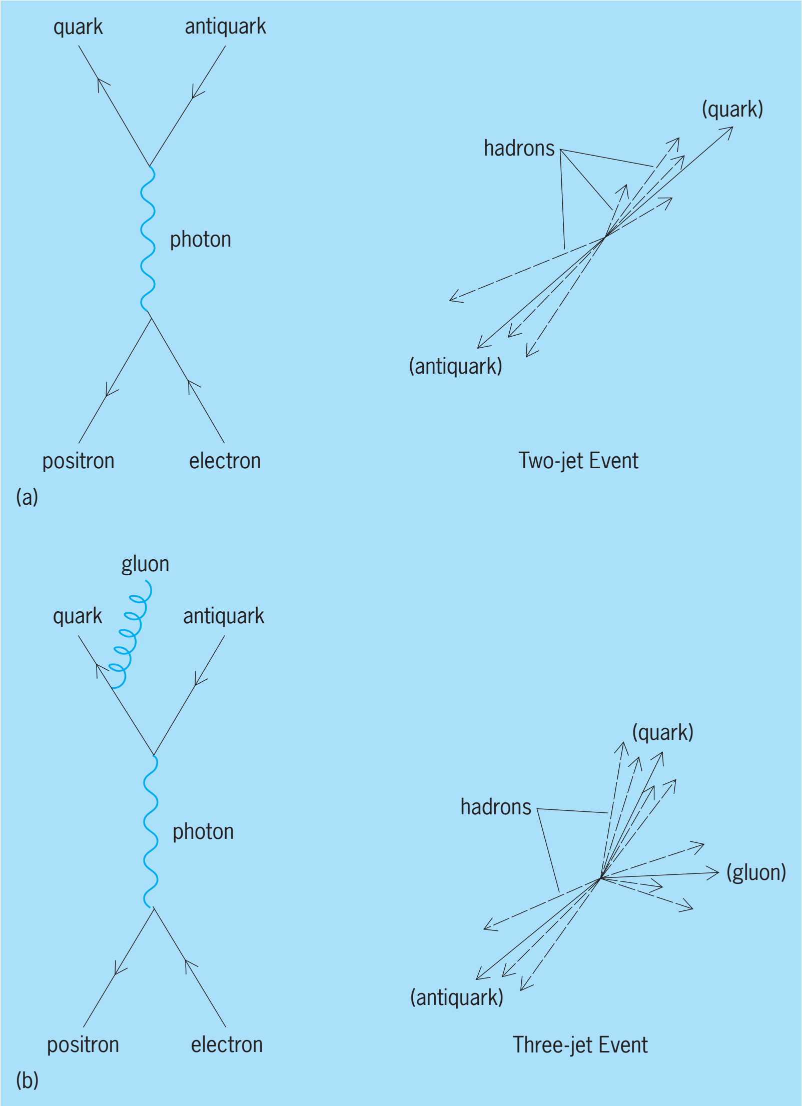


Fig. 3. Mechanisms for hadron production in electron-positron annihilation. (*a*) Two-jet event produced by the mechanism electron positron → quark + antiquark. (*b*) Three-jet event produced by the mechanism electron + positron → quark + antiquark + gluon. The observed hadrons are represented by broken lines.

1120 times longer than parapositronium. In similar fashion, charmonium, the strong-interaction "atom" composed of a charmed quark and a charmed antiquark, decays by the annihilation of the quark and antiquark into gluons. The gluons materialize into the observed hadrons through the action of the confinement mechanism, with unit probability. For the pseudoscalar paracharmonium level, designated η_c, the semifinal state is composed of two gluons (**Fig. 2***a*). The vector particle J/ψ, which corresponds to orthocharmonium, must decay into three gluons (Fig. 2*b*). The remarkably long lifetime of J/ψ and the large ratio of J/ψ to η_c lifetimes (approximately 500) argue for the aptness of the analogy. Decays of the still heavier quarkonium state upsilon also support this. *See* CHARM; J/PSI PARTICLE; POSITRONIUM.

Three-jet pattern. Compelling evidence for the existence of gluons was reported in 1979 by a number of experimental groups working at the high-energy electron-positron storage ring PETRA at the Deutsches Elektronen-Synchrotron (DESY) in Hamburg. It had earlier been established in experiments at SLAC and DESY that the dominant mechanism for hadron production in electron-positron annihilations is electron + positron → quark + antiquark, with the quarks materializing into hadrons. This interpretation explains the rate of particle production and the characteristic angular distribution of the sprays or jets of hadrons that emerge from the collisions (**Fig. 3***a*). If quantum chromodynamics is correct, one of the outgoing quarks may occasionally radiate an energetic gluon, just as a fast electron may radiate an energetic photon. When this happens, the hadrons may be expected to emerge in a three-jet pattern (Fig. 3*b*). Three-jet events are commonplace in electron-positron collisions at center-of-mass energies exceeding about 24 GeV. The characteristics of these events are consistent in every particular with the interpretation that the gluon radiation predicted by quantum chromodynamics is being observed indirectly.

High transverse-momentum jets. Because protons are composed of quarks and gluons, hard collisions of high-energy protons and antiprotons occur through scattering of gluons from gluons, or gluons from quarks or antiquarks, or quarks or antiquarks from quarks or antiquarks. Two-jet events have been measured in great detail in 630-GeV proton-antiproton collisions at the Superproton Synchrotron (SPS) Collider of the European Organization for Nuclear Research (CERN) near Geneva, Switzerland, and in 1.96-TeV proton-antiproton collisions at the Tevatron Collider at Fermilab in Batavia, Illinois. They are in complete accord with the predictions of quantum chromodynamics and cannot be understood without gluons.

Other implications. The existence of gluons with the properties implied by quantum chromodynamics has additional consequences. The interactions of quarks and gluons specified by quantum chromodynamics suggest the existence of a number of new species of hadrons. The most important of these from the gluon perspective are quarkless mesons, composed entirely of gluons. The simplest of these glueballs, as they are sometimes called, may be searched for in the radiative decay $J/\psi \rightarrow$ photon + 2 gluons, or in the decays of heavier quarkonium states. Several glueball candidates have been observed in this way, most notably the iota with mass 1440 MeV/c^2. Further experimental work will be required to determine whether the iota or other candidates must be identified as quarkless states. *See* MESON.

Further experimental tests of the properties that have been imputed to gluons now abound. The angular distribution of hadron jets in *W*-boson + 1-jet events probes the gluon spin. Multijet events produced in high-energy proton-antiproton collisions provide incisive tests of quantum chromodynamics and the nature of the gluon. Tests of the gluon spin are to be had from analysis of the pattern of scaling violations in inelastic lepton scattering, and from comparisons of the hadronic decay rates of various quarkonium states.

The existence of gluons with the canonical properties vindicates the idea of color gauge symmetry and provides strong encouragement for quantum chromodynamics and, by derivation, for grand unification theories of the strong, weak, and electromagnetic interactions. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; GRAND UNIFICATION THEORIES; STANDARD MODEL.

C. Quigg

Bibliography. F. E. Close, Hadron spectroscopy (theory), *Int. J. Mod. Phys.*, A20:5156–5163, 2005; R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and Collider Physics*, Cambridge University Press, 1996; S. L. Glashow, Quarks with color and flavor, *Sci. Amer.*, 233(4):38–50, October 1975; G. Kane, *Modern Elementary Particle Physics,* rev. ed., 1993; T. B. W. Kirk and H. D. I. Abarbanel (eds.), *Proceedings of the 1979 International Symposium on Lepton and Photon Interactions at High Energies*, Fermilab, Batavia, Illinois, 1980; Y. Nambu, The confinement of quarks, *Sci. Amer.*, 235(5):48–60, November 1976; A. Watson, *The Quantum Quark*, Cambridge University Press, 2004.

Glutathione

A biological molecule that is a tripeptide comprising the three amino acids glutamate, cysteine, and glycine, and is involved in multiple metabolic processes. While the cysteine residue is linked to the glycine residue through a normal α-amide bond, the glutamate residue is linked to the cysteine residue through an unusual γ-amide bond. Glutathione is found in reduced and oxidized forms (**Fig. 1**). The reduced form (GSH) contains only a single glutathione unit with the side chain of cysteine in the sulfhydryl form. The oxidized form (glutathione disulfide, GSSG) contains two glutathione units covalently attached through a disulfide bridge between the cysteine residues. *See* AMINO ACIDS.

Synthesis. Glutathione is synthesized from free amino acids in two steps, each catalyzed by a different enzyme (**Fig. 2**): [1] The enzyme γ-glutamylcysteine synthase joins one molecule of glutamate and one molecule of cysteine to form γ-glutamyl cysteine. To fuel the formation of γ-glutamyl cysteine, this enzyme hydrolyzes one molecule of adenosine triphosphate (ATP) to adenosine diphosphate (ADP) and inorganic phosphate (P_i). [2] The enzyme glutathione synthase joins one molecule of γ-glutamyl cysteine and one molecule of glycine to form one molecule of glutathione. This enzyme also fuels this reaction by hydrolyzing one molecule of ATP to ADP and P_i. These two steps of glutathione synthesis are also part of the γ-glutamyl cycle. *See* ADENOSINE TRIPHOSPHATE (ATP); ENZYME.

Role in oxidative metabolism. Glutathione plays an important role in regulating the oxidative state inside the cell. Critical to this function is the enzyme glutathione reductase, which catalyzes the conversion of one molecule of GSSG to two molecules of GSH. This enzyme contains a flavin adenine dinucleotide (FAD) electron-transferring cofactor and catalyzes this reaction in two steps: [1] The oxidized form of the enzyme binds one molecule of NADH. Two electrons are transferred from the NADH molecule to the FAD cofactor of the enzyme. This results in the reduced form of the enzyme and NAD^+, which is released. [2] The reduced form of the enzyme binds one molecule of GSSG. The two electrons are transferred from the FAD cofactor of the enzyme to the glutathione disulfide, resulting in the formation of two molecules of GSH, which are released.

Glutathione reductase maintains the cellular levels of GSH in 100-fold excess over the cellular levels of GSSG. This allows GSH to act as an intracellular reducing agent that maintains the sulfhydryl groups on the side chains of cysteine residues in proteins in their normal states. Two molecules of GSH can react with a protein containing an undesired disulfide between two cysteine residues to form a protein with two cysteine residues with correct sulfhydryl side chains and one molecule of GSSG.

Reactions utilizing glutathione. Several other enzymes utilize glutathione for a variety of metabolic processes. One of these enzymes is glutathione peroxidase, which catalyzes the conversion of two molecules of GSH and one organic hydroperoxide to one molecule of GSSG, one molecule of the corresponding alcohol, and one water molecule. This reaction is important in red blood cells, where hydrogen peroxide reacts with lipids in the cell membrane to form lipid peroxides that can rupture the cell. The action of glutathione peroxidase prevents

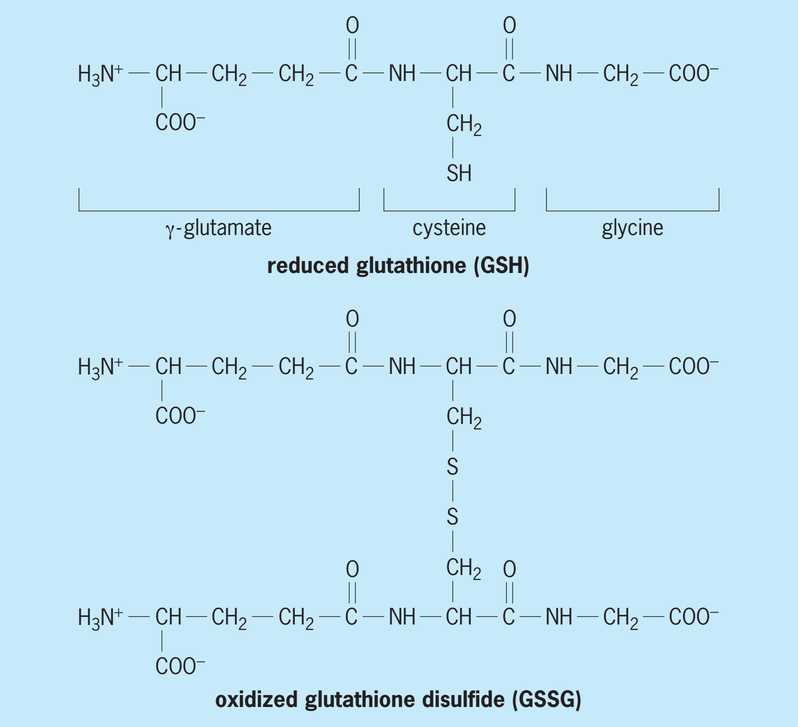

γ-glutamate
cysteine
glycine
reduced glutathione (GSH)
oxidized glutathione disulfide (GSSG)

Fig. 1. Reduced and oxidized forms of glutathione.

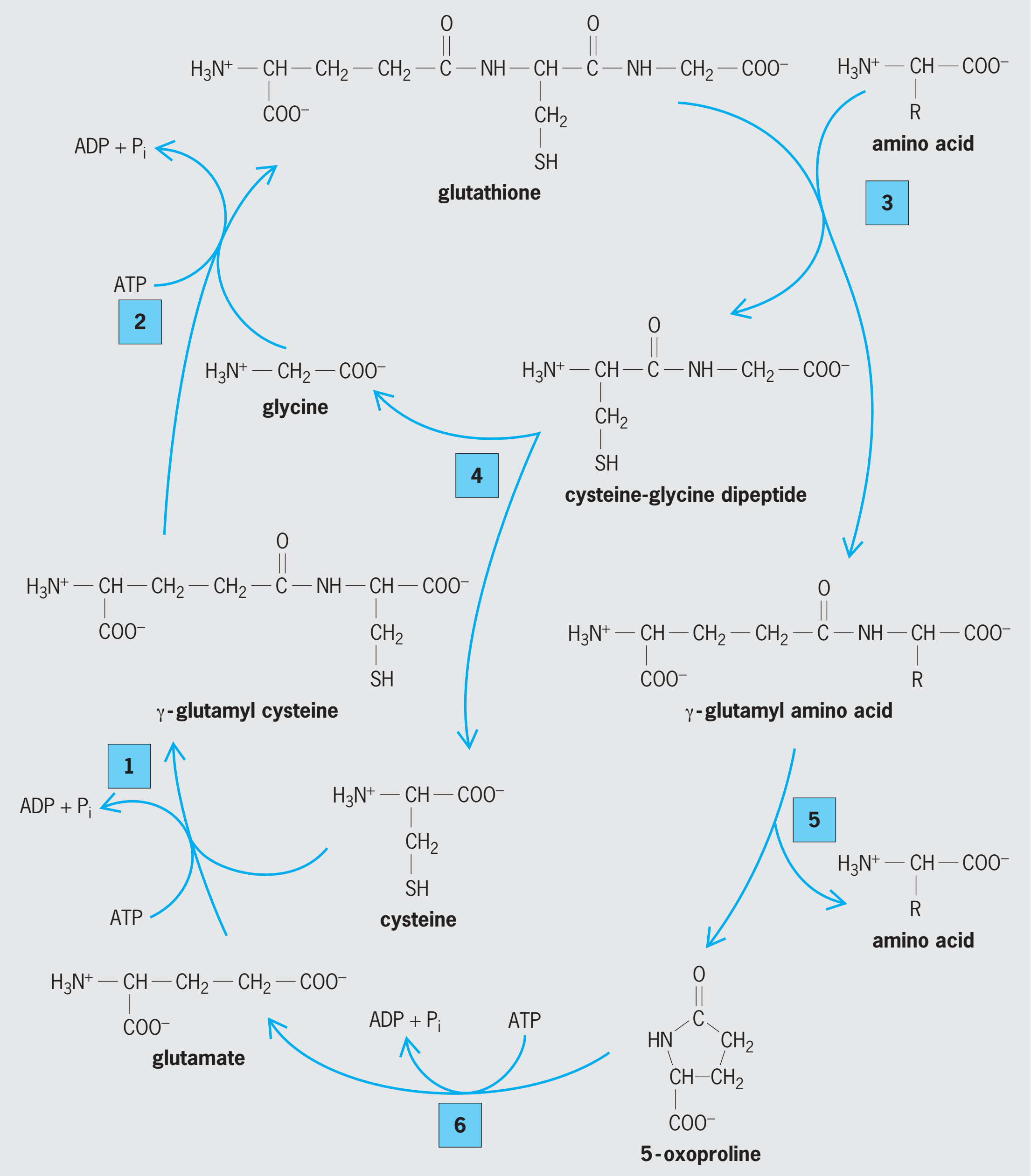


Fig. 2. γ-Glutamyl cycle. See text for description of steps.

cell rupturing by removing lipid peroxides from the cell membrane. *See* LIPID.

Another of these enzymes is glutathione-S-transferase, which catalyzes the covalent attachment of glutathione to a variety of normal metabolites or xenobiotics (chemicals not normally found within cells, such as toxins or drugs) through the sulfhydryl group of its cysteine residue. For example, this enzyme attaches glutathione to a compound called leukotriene A_4, a normal metabolite that is formed from arachidonic acid, to form leukotriene C4, a peptidoleukotriene that is a precursor of molecules involved in anaphylaxis, a severe and sometimes fatal allergic response. *See* ANAPHYLAXIS.

Glutathione also can attach itself nonenzymatically to proteins by forming a disulfide bond between the side chain of the cysteine in the glutathione and the side chain of a cysteine of the protein. Recent research shows that this "glutathionylation" is an important means of regulating the function of these target proteins. For example, glutathionylation has been observed with signal transduction components, such as *ras*, and transcription factors, such as *jun*, involved in promoting cell proliferation.

γ-Glutamyl cycle. Glutathione is also a key component of the γ-glutamyl cycle (Fig. 2), which is important for transporting amino acids from the outside of the cell to the interior of the cell. This cycle comprises six steps: [1, 2] As described above, glutathione is synthesized from glutamate, cysteine, and glycine. [3] The enzyme γ-glutamyl transpeptidase, which is found on the cell membrane, binds an amino acid from the exterior of the cell and one molecule of glutathione from the interior of the cell. This enzyme transfers the γ-glutamyl group from the glutathione to the amino acid and then releases both the γ-glutamyl amino acid and the cysteine-glycine dipeptide to the cell's interior. [4] The cysteine-glycine dipeptide is broken down into cysteine and glycine by an intracellular protease. [5] The enzyme γ-glutamyl cyclotransferase converts the γ-glutamyl amino acid to the free amino acid and 5-oxoproline. [6] The enzyme 5-oxoprolinase converts the 5-oxoproline to glutamate and hydrolyzes one molecule of ATP to ADP and P_i in the process.

M. Todd Washington

Bibliography. D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 4th ed., 2004; D. Voet and J. G. Viet, *Biochemistry*, 3d ed., 2004.

Glycerol

The simplest trihydric alcohol, with the formula $CH_2OHCHOHCH_2OH$. The name glycerol is preferred for the pure chemical, but the commercial product is usually called glycerin. It is widely distributed in nature in the form of its esters, called glycerides. The glycerides are the principal constituents of the class of natural products known as fats and oils.

Properties. When pure, glycerin is a colorless, odorless, viscous liquid with a sweet taste. It is completely soluble in water and alcohol but is only slightly soluble in many common solvents, such as ether, ethyl acetate, and dioxane. Glycerin is insoluble in hydrocarbons. It boils at 290°C (554°F) at atmospheric pressure and melts at 17.9°C (64.2°F). Its specific gravity is 1.262 at 25°C (77°F) referred to water at 25°C (77°F), and its molecular weight is 92.09. It has a very low mammalian toxicity.

Production. Glycerin was first discovered in 1779 by Carl W. Scheele, who made it by heating olive oil with litharge. Until after World War II, nearly all the glycerin of commerce was produced as a by-product in the manufacture of soap or from the hydrolysis (splitting) of fats and oils. However, a substantial portion of the material made in the United States is now prepared synthetically from propylene.

In the process of soapmaking, called saponification, fat reacts with aqueous sodium hydroxide. The crude product is coagulated by the addition of salt. The acid portion of the fat combines with the sodium hydroxide to form solid soap, and the glycerin liberated in the reaction remains in the salt solution, which is called spent lye.

In the process involving the hydrolysis of fats, they react with water to give the component acids and glycerin. An aqueous solution of the latter is produced, called glycerin sweet water. From this liquid and from spent lye from soapmaking, glycerin is obtained in much the same way. After a preliminary treatment to remove impurities, the water of solution is evaporated under reduced pressure. The residual glycerin is filtered while hot to remove precipitated salts. For most applications, it is necessary to refine it further by fractional distillation under reduced pressure.

Three synthetic routes from propylene to glycerin are used on a large scale. In the first of these, propylene is converted to allyl chloride, which is then treated with aqueous chlorine to make glycerin dichlorohydrins which may be directly hydrolyzed to glycerin. Alternately, the dichlorohydrins are hydrolyzed to epichlorohydrin with a further hydrolysis to glycerin. In a second process, propylene is oxidized in the vapor phase to acrolein. The acrolein is converted to glycerin by successive reactions with hydrogen peroxide, water, and hydrogen. In the third route, propylene is reacted with aqueous chlorine to make propylene oxide. This is isomerized to allyl alcohol, which in turn is reacted with peracetic acid and then with water to make glycerin.

Several grades of glycerin are marketed, including high gravity, dynamite, yellow distilled, USP (U.S. Pharmacopoeia), and CP (chemically pure). USP is water-white and suitable for use in foods, pharmaceuticals, and cosmetics, or for any purpose where the product is designed for human consumption.

Uses. Glycerin is used in nearly every industry. With dibasic acids, such as phthalic acid, it reacts to make the important class of products known as alkyd resins, which are used as coatings and in paints. Because of its valuable emollient and demulcent properties, it is used in innumerable pharmaceutical and cosmetic preparations. It is an ingredient of many tinctures, elixirs, cough medicines, and anesthetics. It is a basic medium for toothpaste.

In foods, it is an important moistening agent for baked goods and is added to candies and icings to prevent crystallization. It is used as a solvent and carrier for extracts and flavoring agents and as a solvent for food colors.

Because of its humectant properties, it is sprayed on tobacco before it is processed to prevent crumbling and is added to adhesives and glues to keep them from drying too fast. Many specialized lubrication problems have been solved by using glycerin or glycerin mixtures.

Many millions of pounds of glycerin are used each year to plasticize various materials. As much as 15% is added to cellophane to render it pliable. It is included in meat casings and special types of paper for the same purpose. Sheets and gaskets made from ground cork are plasticized with glycerin.

Of its chemical derivatives, the esters of glycerin are the most important. Nitroglycerin (glyceryl trinitrate) is used in the manufacture of dynamites and propellants. The mono- and diesters of higher fatty acids can be formed by direct reaction of glycerin with the acid or by a transesterification reaction of a glyceride with glycerin. These esters are used as emulsifiers in foods and preparation of baked goods and for modification of alkyd resins. *See* ALCOHOL; FAT AND OIL; FAT AND OIL (FOOD); POLYOL.

Philip H. Cook

Bibliography. J. A. Kent (ed.), *Riegel's Handbook of Industrial Chemistry*, 9th ed., 1992; *Kirk-Othmer Encyclopedia of Chemical Technology*, 3d ed., vol. 11, 1998; C. S. Miner and N. N. Dalton (eds.), *Glycerol*, ACS Monogr. 117, 1953.

Glycogen

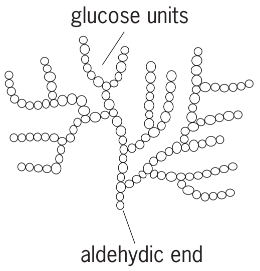
The primary reserve polysaccharide of the animal world. It is found in the muscles and livers of all higher animals, as well as in the cells of lower animals. Because of its close relationship to starch, it is often called animal starch, although glycogen is found in some lower plants, fungi, yeast, and bacteria. A polysaccharide similar to glycogen was isolated in one case from a higher plant, Golden Bantam sweet corn (*Zea mays*). *See* LIVER; STARCH.

Properties. Glycogen is a nonreducing, white, amorphous polysaccharide which dissolves readily in cold water, forming an opalescent, colloidal solution. It gives a reddish brown color with iodine, is precipitated by alcohol, and has a specific rotation $[\alpha]_D^{20}$ of approximately + 200°. It is very resistant to the action of alkalies and may be prepared by boiling liver or muscle tissue in 30% potassium hydroxide to destroy the proteins, and precipitating the glycogen with ethyl alcohol. *See* OPTICAL ACTIVITY.

The molecular weight of glycogen is usually very high, and it varies with the source and the method of preparation; molecular weights of the order of $1\text{–}20 \times 10^6$ have been reported.

In its biochemical reactions, glycogen is similar to starch. It is attacked by the same plant amylases that attack starch, and like starch, it is degraded to maltose and dextrins. Both glycogen and starch are broken down by animal or plant phosphorylase enzyme in the presence of inorganic phosphate with the production of α-D-glucose-1-phosphate. *See* CARBOHYDRATE METABOLISM.

Molecular structure. Chemical studies, based on methylation and periodate oxidation procedures, show glycogen to possess a branched structure similar to the amylopectin starch fraction. The molecules of both these polysaccharides consist of chains of D-glucose residues joined by α-1,4 linkages, having similar chains attached through α-1,6 linkages at the branch points (see **illus**.). Depending on the source, the average chain length of a branch (which is the average length in glucose units of the outer and inner branches) in a glycogen molecule is 11–18 D-glucopyranose units. In amylopectin the average chain length of a branch is 22–27.



Branched structure of glycogen.

Synthesis. About 1940 C. F. Cori and G. T. Cori first showed that glycogen could be synthesized in the test tube and that two enzymes, muscle phosphorylase and a branching enzyme which is present in animal tissues, are required for the formation of this important polysaccharide. The muscle phosphorylase catalyzes a stepwise transfer of α-D-glucosyl units from α-D-glucose-1-phosphate to a nonreducing end of a "primer" (acceptor substrate). Synthesis of polysaccharide from α-D-glucose-1-phosphate does not occur unless a small amount of starch, glycogen, or dextrin is present as a priming agent. In the presence of the primer, the enzyme adds D-glucose units to a preexisting polysaccharide chain, forming long linear chains joined through α-1,4 linkages. The reaction can be written as reaction (1).

$$x\ \alpha\text{-D-Glucose-1-phosphate} + \underbrace{(\text{D-glucose})_n}_{\text{Acceptor}} \rightleftharpoons \underbrace{(\alpha\text{-1,4-D-glucose})_{n+x}}_{\text{Glycogen chain}} + x \text{ phosphate} \tag{1}$$

The second enzyme (branching factor) has the ability to unite such chains through α-1,6 linkages, resulting in the formation of a highly branched glycogen structure. It was believed that the combined action of the two enzymes was responsible for the synthesis of glycogen in the animal body.

In 1957 L. F. Leloir and his coworkers obtained an enzymic preparation from liver and other animal tissues that catalyzed the transfer of D-glucose from uridine diphosphate D-glucose to an acceptor (glycogen) according to reaction (2).

$$x \text{ Uridine diphosphate-D-glucose} + \underbrace{(\alpha\text{-1,4-D-glucose})_n}_{\text{Acceptor}} \longrightarrow \underbrace{(\alpha\text{-1,4-D-glucose})_{n+x}}_{\text{Glycogen chain}} + x \text{ uridine diphosphate} \tag{2}$$

This reaction resembles that of the phosphorylase enzyme, but its equilibrium starting with uridine diphosphate D-glucose is more favorable for synthesis of glycogen than that starting with α-D-glucose-1-phosphate. Since the product obtained is branched, the enzyme preparations must contain another enzyme (6-glucosyltransferase), which transfers α-1,4-glucosyl chains produced by the first enzyme (glycogen glucosyltransferase) to form a branched molecule. The existence of two mechanisms of glycogen formation raises the question of which of the two mechanisms operates in the animal body. Evidence indicates that synthesis of glycogen in the body occurs through a transfer of D-glucose units from uridine diphosphate D-glucose by an enzyme, glycogen glycosyltransferase, whose function is entirely synthetic. When the chains become about 10 units long, portions are transferred to other chains forming α-1,6-glycosyl linkages by a branching enzyme. The function of phosphorylase is considered to be concerned chiefly with degradation of glycogen.

Metabolic pathways. The metabolic formation of glycogen from glucose in the liver is frequently termed glycogenesis. In fasted animals, glycogen formation can be induced by the feeding, not only of materials that can be hydrolyzed to glucose and other monosaccharides, such as fructose, but also of various other materials. A number of L-amino acids, such as alanine, serine, and glutamic acid, upon deamination in the liver give rise to substances, such as pyruvic acid and α-ketoglutaric acid, that can be converted in the liver to glucose units which are subsequently converted to glycogen. Furthermore, substances such as glycerol derived from fats, dihydroxyacetone, or lactic acid can all be utilized for glycogen synthesis in the liver. Such noncarbohydrate precursors are termed glycogenic compounds. The process of glycogen formation from these precursors is known as glyconeogenesis. The term glycogenolysis is used to connote glycogen breakdown. *See* POLYSACCHARIDE. William Z. Hassid

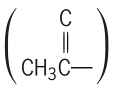
Bibliography. E. D. Atkins, *Polysaccharides*, vol. 8, 1986; R. W. Stoddart (ed.), *The Biosynthesis of Polysaccharides*, 1984.

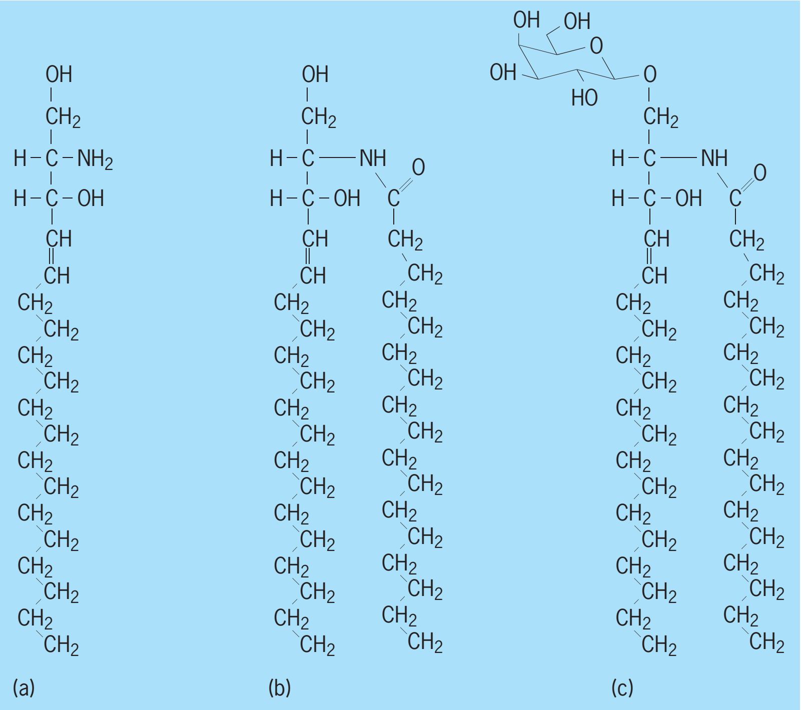
Glycolipid

One of a class of compounds having solubility properties of a lipid and containing one or more molecules of a covalently attached sugar.

Glycosphingolipids, the most abundant and structurally diverse type of glycolipids in animals, are glycosides of ceramide, a fatty acid amide of the amino alcohol sphingosine (see **illus.**). Cerebrosides are monosaccharide glycosides of ceramide, which generally contain either galactose or glucose. Galactosyl ceramide is enriched in brain tissue and is a major component of the myelin sheaths around nerves. Psychosine, *O*-sphingosyl galactoside, is the deacylated product of galactosyl ceramide. Glucosyl ceramide is present in the cell membranes of many cell types and is abundant in serum. Lipidoses (inherited lysosomal lipid storage diseases that lead to abnormal development and mental retardation) result from genetic defects in enzymes that degrade glycosphingolipids. For example, in Gaucher's disease, glucocerebrosidase, an enzyme that removes the glucose from glycosyl ceramide, is defective or missing.

Larger, neutral glycosphingolipids containing more than one sugar include lactosyl ceramide, abundant in leukocyte membranes; globosides; and other oligosaccharyl ceramides, some of which are important antigens defining blood groups. Gangliosides are oligosaccharyl ceramides, abundant in brain, spleen, erythrocytes, liver, and kidney, that contain glucose, galactose, *N*-acetylglucosamine, and sialic acids. Sialic acids are *N*-acyl derivatives of neuraminic acid, a nine-carbon sugar with a carboxylic acid group:





Components of glycosphingolipids. (*a*) Sphingosine. (*b*) Ceramide. (*c*) Galactocerebroside.

They occur at the terminus of many types of glycoconjugates, and display enormous structural diversity, with over 30 different structures known. Sulfatides are glycosphinglipids of various types that have sulfate esters, generally attached to galactose.

Glycosphingolipids carry blood group antigens and define tumor-specific or developmental antigens. In addition, they serve as receptors for many microorganisms and toxins, as modulators of cell surface receptors that mediate cell growth, and as mediators of cell adhesion. *See* ANTIGEN; CELL ADHESION.

Glycosyl phosphatidylinositols are a class of glycolipids that serve as membrane anchors for a multitude of proteins in organisms ranging from yeast to protozoa to humans. Glycosyl phosphatidylinositols consist of a molecule of phosphatidlyl inositol attached to a nonacetylated glucosamine and a trimannosyl core linked via an ethanolamine phosphodiester bridge to the C-termini of proteins. Furthermore, glycosyl phosphatidylinositol–core structures can have many different modifications, depending upon the protein and cell type. Lipophosphoglycans are glycosyl phosphatidylinositols attached to large polysaccharide structures that coat the surfaces of many parasitic protozoa, such as *Leishmania donovani*, the causative agent of visceral leishmaniasis (kala azar), afflicting millions of individuals each year. Lipophosphoglycans appear to protect these organisms from host defenses.

Mannosylphosphoryl dolichol, glucosylphophoryl dolichol, and oligosaccharyl phosphoryl dolichols are glycolipids with sugars attached to large polyisoprenoids (for example, C_{95} hydrocarbons) by phosphate esters. Dolichols are structurally related to cholesterol. Saccharylphosphoryl dolichols serve as important biosynthetic intermediates in the assembly of both asparagine-linked glycoproteins and glycosyl phosphatidylinositols. *See* GLYCOPROTEIN.

Glycosyl glycerides are glycolipids that have a structure analogous to phospholipids. They are the major glycolipids of plants and microorganisms but are rare in animals.

Bacteria produce a wide variety of glycolipids not easily categorized. Examples include fatty acid esters of carbohydrates, such as cord factor, an ester of the disaccharide trehalose with two molecules of the complex fatty acid, mycolic acid. Cord factor is a toxic component of the waxy capsular material of virulent strains of *Mycobacterium tuberculosis*, the causative agent of tuberculosis. Mycosides, glycolipids that are also found in tubercle bacilli, comprise long-chain, highly branched, hydroxylated hydrocarbon terminated by a phenol group, with the sugar glycosidically attached to the phenolic hydroxyl. *See* LIPID; SPHINGOLIPID; TUBERCULOSIS. Gerald W. Hart

Glycoprotein

A compound in which carbohydrate (sugar) is covalently linked to protein. The carbohydrate may be in the form of monosaccharides, disaccharides, oligosaccharides, or polysaccharides, and is sometimes referred to as glycan. The sugar may be linked to sulfate or phosphate groups. One to two hundred glycan units may be present in different glycoproteins. Therefore, the carbohydrate content of these compounds varies markedly, from 1% (as in the collagens), to 60% (in certain mucins), to >99% (in glycogen). *See* COLLAGEN; GLYCOGEN.

Glycoproteins are ubiquitous in nature, although they are relatively rare in bacteria. They occur in cells, in both soluble and membrane-bound forms, as well as in the intercellular matrix and in extracellular fluids, and include numerous biologically active macromolecules. A particularly rich source of glycoproteins is human serum, in which, of the more than 60 proteins that have been identified, only two—albumin and prealbumin—do not contain sugar. Another rich source is hen egg white, where probably all proteins, apart from lysozyme, have carbohydrate attached to their molecules. A number of glycoproteins are produced industrially by genetic engineering techniques for use as drugs; among them are erythropoietin, interferons, colony stimulating factors, and blood-clotting factors. *See* GENETIC ENGINEERING.

Composition and structure. Of the numerous monosaccharides found in nature, only a small number are common constituents of glycoproteins. In most glycoproteins, the carbohydrate is linked to the polypeptide backbone by either N- or O-glycosidic bonds. A different kind of bond is found in glycoproteins that are anchored in cell membranes by a special carbohydrate-containing compound, glycosylphosphatidylinositol, which is attached to the C-terminal amino acid of the protein. A single glycoprotein may contain more than one type of carbohydrate-peptide linkage. N-linked units are typically found in plasma glycoproteins, in ovalbumin, in many enzymes (for example, the ribonucleases), and in immunoglobulins. O-linked units are found in mucins; collagens; and proteoglycans (typical constituents of connective tissues), including chondroitin sulfates, dermatan sulfate, and heparin. *See* ALBUMIN; ENZYME; IMMUNOGLOBULIN.

Microheterogeneity. Within any organism, all molecules of a particular protein are identical. In contrast, a variety of structurally distinct carbohydrate units are found not only at different attachment sites of a glycoprotein but even at each single attachment site—a phenomenon known as microheterogeneity. For instance, ovalbumin contains one glycosylated amino acid, but over a dozen different oligosaccharides have been identified at that site, even in a preparation isolated from a single egg of a purebred hen.

Biosynthesis. The protein moieties of all glycoproteins are synthesized, on polyribosomes, in the same manner as nonglycosylated proteins. Formation of the carbohydrate units is catalyzed by glycosyltransferases, enzymes that transfer sugars from suitable donors to particular acceptors, and is not directed by a genetically specified template. Over 100 distinct glycosyltransferases, located in the rough endoplasmic reticulum of the cell and in the Golgi apparatus, participate in the biosynthesis of glycoproteins, each forming an individual glycosidic bond. In addition, many other enzymes are required for completion of the carbohydrate units, such as the attachment of sulfate or phosphate. O-linked units are formed by sequential incorporation of one monosaccharide at a time from a sugar nucleotide, first to a hydroxyl of a particular amino acid residue of a polypeptide, and then to a hydroxyl of the protein-bound sugars. *See* ENDOPLASMIC RETICULUM; GOLGI APPARATUS.

N-linked oligosaccharides are synthesized in a different, more complicated sequence of reactions. First, a single precursor oligosaccharide is formed, linked by pyrophosphate to a long-chain lipid. The oligosaccharide is then transferred from the lipid to an amide of an asparagine residue of a growing polypeptide chain. This occurs as the polypeptide emerges from the ribosome and enters the lumen of the rough endoplasmic reticulum. Subsequently, the polypeptide-linked oligosaccharide undergoes a series of processing reactions which begin in the rough endoplasmic reticulum and continue and end in the Golgi apparatus, eventually yielding the N-linked oligomannose units.

Catabolism. One pathway of glycoprotein degradation in the body starts with the stepwise hydrolysis of the major part of the glycans by the sequential action of lysosomal glycosidases, followed by the enzymatic disassembly of the protein and cleavage of the carbohydrate-peptide linkages. An alternative pathway starts with proteolysis of the polypeptide backbone and requires the participation of endoglycosidases, which release intact glycans from their linkages with amino acids. Individuals that lack one of the glycosidases or other enzymes (for example, sulfatases) needed for glycoprotein degradation suffer from serious, often fatal diseases. Examples are the mucopolysaccharidoses, such as the Hunter-Hurler syndrome. In these diseases, undegraded glycosaminoglycans accumulate in the tissues of the afflicted individuals, resulting in developmental abnormalities, skeletal deformations, and mental retardation. Another example is the I-cell disease, caused by lack of the first enzyme required for the formation of the mannose-6-phosphate on lysosomal enzymes, and thus the deficiency in the lysosomes of many of the hydrolases required for degradation of carbohydrate-containing compounds.

Functions. The carbohydrate may modify, in a variety of ways, the properties of the protein to which it is attached. Commonly, it changes the physicochemical properties of the protein, such as electrical charge, solubility, or viscosity, and increases its thermostability. Frequently, it protects the polypeptide backbone against proteolytic digestion. It is also required for the correct folding of proteins during their biosynthesis and for the stabilization of the three-dimensional structure of the mature glycoprotein.

Effects on biological activity. In a small number of glycoproteins the biological activity of the protein part is known to be affected by, or to depend on, the presence of the carbohydrate. A well-documented case is human chorionic gonadotropin, which loses its hormonal activity upon removal of its N-glycans. Another example is tissue plasminogen activator, a protease that induces clot lysis and is used clinically for this purpose. Two molecular species of the enzyme are found in tissues that are structurally identical except for the number of N-linked glycans (two or three); they differ markedly in their enzymatic activity. In general, however, the carbohydrate has no effect on the catalytic activity of glycoprotein enzymes, nor for example on the carbohydrate-binding properties or hemagglutinating activity of glycoprotein lectins.

Often, the carbohydrate affects the immunological properties of the protein to which it is attached, mainly by introducing novel antigenic determinants. This is illustrated by the A, B, and O blood group glycoproteins that are present on human erythrocytes, as well in other tissues, such as gastric mucosa and salivary glands. *See* BLOOD GROUPS.

Recognition determinants. Carbohydrates of glycoproteins can act as recognition determinants. Thus, infection by influenza virus is initiated by the attachment of the virus to sialic acid residues of the cell-surface glycoproteins. Many other systems in which carbohydrates play a similar role are known. For example, bacteria such as *Escherichia coli* adhere to the respiratory, intestinal, or urinary tract by binding to mannose residues on glycoproteins present on the epithelial cells lining these organs, a step prerequisite for the initiation of infection. Also, the binding of leukocytes to endothelial cells lining the blood vessels, crucial for the control of the movement of the leukocytes in the body, and for the initiation of inflammation, is mediated by complex oligosaccharides present on both the leukocytes and the endothelial cells. *See* CARBOHYDRATE; CELLULAR IMMUNOLOGY; GLYCOSIDE; MONOSACCHARIDE; OLIGOSACCHARIDE; POLYSACCHARIDE; PROTEIN.

Nathan Sharon

Bibliography. R. C. Hughes, *Glycoproteins*, 1983; N. Sharon and H. Lis, Carbohydrates in cell recognition, *Sci. Amer.*, 268(1):74–82, 1993.

Glycoside

A large important class of sugar derivatives in which the sugar is combined with a nonsugar. In their cyclic forms, monosaccharides (simple sugars) possess one carbon (C) atom (the anomeric carbon) that is bonded to two oxygen (O) atoms; one oxygen atom forms a part of the ring, whereas the other is outside the ring (exocyclic) and is part of a hydroxyl (OH) group. If the oxygen atom of the anomeric hydroxyl group becomes bonded to a carbon atom, other than that of a carbonyl (C═O) group, the resulting compound is a glycoside. A glycoside thus consists of two parts (**Fig. 1***a*): the sugar (glycosyl) unit, which provides the anomeric carbon, and the moiety (the aglycon), which is the source of the exocyclic oxygen and carbon atoms of the glycosidic linkage. Such compounds frequently are referred to as O-glycosides to distinguish them from analogs having a sulfur (thio- or *S*-glycosides), nitrogen (amino- or N-glycosides), or carbon (anomalously called C-glycosides) as the exocyclic atom on the anomeric carbon. *See* HYDROXYL; MONOSACCHARIDE.

The formation of glycosides is the principal manner in which monosaccharides are incorporated into more complex molecules. For example, lactose (Fig. 1*b*), the most abundant disaccharide in mammalian milk, has a glycosidic bond involving

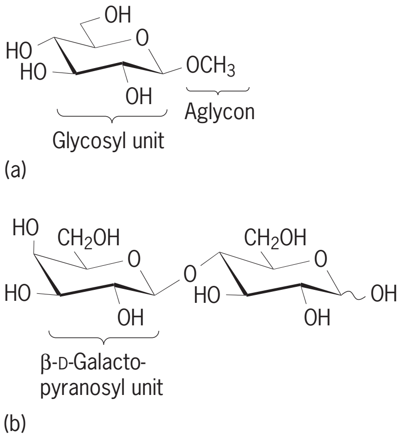


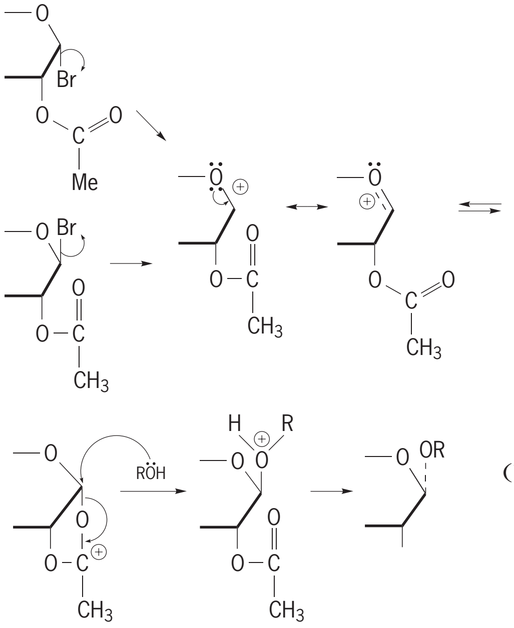
Fig. 1. Structural formulas of two glycosides. (*a*) Methyl β-D-glucopyranoside. (*b*) Lactose, 4-*O*-β-D-galactopyranosyl-D-glucopyranose; the wavy bond indicates that the group may have various orientations in space.

the anomeric carbon of D-galactose and the C-4 hydroxyl of D-glucose. The anomeric carbon atom can exist in either of two stereoisomeric configurations. In the D-series of carbohydrates a glycoside is designated as having the α configuration if the exocyclic oxygen is attached to the anomeric carbon by a bond that projects downward, or the β configuration if that bond projects upward, when the ring is oriented with the ring oxygen away from the viewer (Fig. 1*b*). In the L-series the orientation is reversed. The anomeric configuration is of immense importance to the chemistry and biochemistry of glycosides. For example, the principal structural difference between cellulose and amylose is that cellulose is β-glycosidically linked whereas amylose is α-linked. Humans are able to digest amylose but are unable to utilize cellulose for food. *See* CELLULOSE; LACTOSE; STEREOCHEMISTRY.

Chemical synthesis. The transformation of a monosaccharide into a glycoside involves the conversion of a hemiacetal (or hemiketal) into an acetal (or ketal). Such reactions normally are catalyzed by an acid and may involve a temporary loss of chirality (dissymmetry) at the anomeric carbon. Glycosides are susceptible to hydrolysis in aqueous acid, but they are stable under alkaline conditions. The classical (Fischer) method for synthesizing glycosides from unprotected (that is, underivatized) monosaccharides and low-molecular-weight alcohols involves heating the sugar and alcohol in the presence of an acid catalyst under anhydrous conditions. *See* ACETAL.

The simplicity of the method belies the complexity of the reactions; these give an equilibrium mixture of α- and β-pyranosides (having six-membered rings) and α- and β-furanosides (having five-membered rings) in a ratio that is dependent upon the monosaccharide, the alcohol, and the precise conditions of reaction. The furanosides are formed more rapidly, but pyranosides are more stable; thus, in general, the principal products at equilibrium are the pyranosides. The restrictions inherent in this method limit its synthetic usefulness.

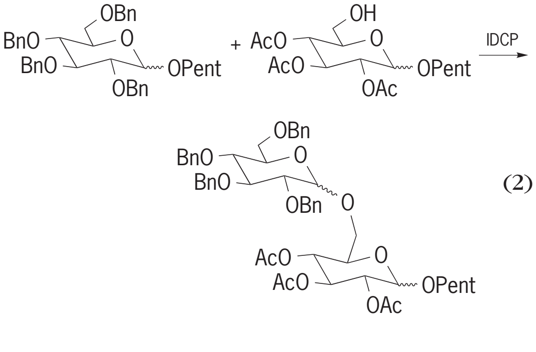
A much more common method is the Königs-Knorr reaction, which utilizes a carbohydrate derivative in which the anomeric hydroxyl group has been replaced by a halogen atom—usually bromine (Br)—and all of the other hydroxyl groups of the sugar have been converted to esters or ethers, thereby making them unavailable for glycoside formation. This fully protected glycosyl halide is capable of forming a glycosidic bond with virtually any compound having a free hydroxyl group. A large number of variations of this method have been devised; most use a Lewis acid, such as a salt of silver or mercury, as catalyst. This method affords a relatively high yield, and predictability as to the anomeric configuration of the glycoside. When the glycopyranosyl halide is protected with (participating) ester groups, the phenomenon of anchimeric assistance (neighboring-group effect) becomes operative, and the major product is the so-called 1,2-*trans*-glycopyranoside in which the functional group on C-2 is trans to the aglycon [reaction (1), where R is an alkyl or aryl group].



(1)

When the glycopyranosyl halide is protected with (nonparticipating) ether groups, the preferred product is the 1,2-*cis*-glycopyranoside. Thus, regardless of the glycosyl halide used, either an α- or a β-glycopyranoside can be formed as the major product by the appropriate selection of the protecting groups. A very powerful extension of this method employs *n*-pentenyl glycosides (specifically, 4-pentenyl), which are activated in the presence of halogen cations (halonium ions), usually supplied by iodonium collidine perchlorate (IDCP). A key feature is that if the *n*-pentenyl glycoside is protected by ester substituents it reacts much more slowly than one protected by ether substituents. The esterified glycoside is said to be disarmed and the etherified one armed with respect to their reactivity with halonium ions. Thus, when both are present in a reaction mixture the ether-protected (armed) 4-pentenyl glycosides react preferentially.

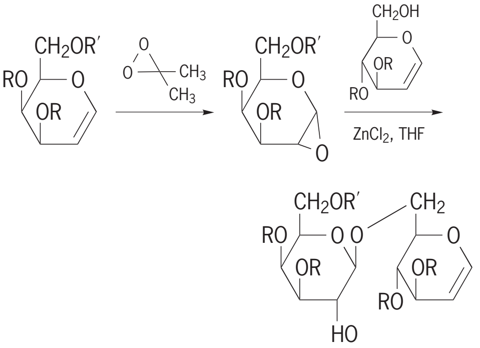
The sugars can be joined together by a glycosidic bond [reaction (2), where Ac is acetyl



(2)

[$H_3C—C(=O)—$], Bn is benzyl ($—CH_2—C_6H_5$), and Pent is 4-pentenyl]. A 4-pentenyl glycoside, fully substituted with ether groups, is reacted in the presence of IDCP with another 4-pentenyl glycoside having one free hydroxyl group and all the others esterified. The disarmed esterified glycoside reacts slowly with the free hydroxyl group, whereas the armed etherified glycoside reacts rapidly. Therefore, almost all of the reaction results in the replacement of the 4-pentenyloxy group of the armed glycoside by the free hydroxyl group of the disarmed glycoside.

Another ingenious method for glycosidic bond formation between sugar units uses an insoluble polymer as the site of reaction, and so is described as a solid-phase synthesis. Solid-phase syntheses of oligopeptides and oligonucleotides have been used for many years, but this method is the most practical application for oligosaccharides. These are formed by a sequence of couplings. The method begins by attaching a glycal (a monosaccharide having a double bond between carbons 1 and 2, and no hydroxyl group on either of those carbons) to the solid polymer; all of the free hydroxyl groups of the glycal are protected, and the double-bonded carbon atoms are incorporated into a three-membered cyclic ether (an epoxide). The use of 3,3-dimethyldioxirane effects this with high trans-stereoselectivity and yield [reaction (3), where R is benzyl (—CH_2—C_6H_5), R′ is a polymer, and THF is tetrahydrofuran]. The epoxide is reacted in tetrahydrofuran, in the presence of zinc chloride ($ZnCl_2$), with another partially protected glycal having one free hydroxyl group. The epoxide acts as the glycosyl donor and is opened by a back-side attack by the free hydroxyl of the glycosyl acceptor, thereby forming a glycoside and concurrently generating a free hydroxyl group at C-2. This valuable outcome can afford an acceptor for branched-oligosaccharide synthesis, or the hydroxyl group can be protected by esterification or etherification and the glycal epoxidized and used for further glycosidation. *See* ESTER; ETHER; OLIGOSACCHARIDE; ORGANIC SYNTHESIS; REACTIVE INTERMEDIATES.



(3)

Naturally occurring glycosides. A very large number of glycosides exist in nature, many of which possess important biological functions. In many of these biologically important compounds the carbohydrate portion is essential for cell recognition, the terminal sugar units being able to interact with specific receptor sites on the cell surface.

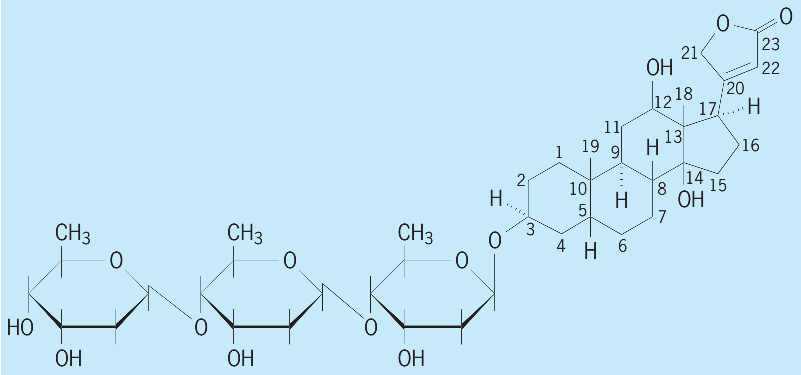


Fig. 2. Structural formula of digoxin, a cardiac glycoside.

One class of naturally occurring glycosides is called the cardiac glycosides because they exhibit the ability to strengthen the contraction of heart muscles. These cardiotonic agents are found in both plants and animals and contain complex aglycons, which are responsible for most of the drug action; however, the glycoside may modify the biological activity. The best-known cardiac glycosides come from digitalis and include the drug digoxin (**Fig. 2**). *See* DIGITALIS.

Glycosidic units frequently are found in antibiotics. For example, daunomycin and adriamycin, antibiotics used as antitumor agents, incorporate an aminodeoxy sugar unit glycosidically linked to a polycyclic aglycon (**Fig. 3***a*); and the important drug erythromycin A possesses two glycosidically linked sugar units (Fig. 3*b*). *See* ANTIBIOTIC.

Perhaps the most ubiquitous group of glycosides in nature is the glycoproteins; in many of them carbohydrates are linked to a protein by *O*-glycosidic bonds to L-serine, L-threonine, 5-hydroxy-L-lysine, or 4-hydroxy-L-proline. These glycoproteins include many enzymes (for example, the human intestinal disaccharidases), hormones, such antiviral compounds as interleukin-2, and the so-called antifreeze glycoproteins found in the sera of fish from very cold marine environments. *See* AMINO ACIDS; ANTIFREEZE (BIOLOGY); CARBOHYDRATE; ENZYME; GLYCOPROTEIN; HORMONE.

Glycolipids are a very large class of natural glycosides having a lipid aglycon. These complex glycosides are present in the cell membranes of microbes, plants, and animals. Animal glycolipids usually have ceramide derivatives as the aglycon. A ceramide is a general term for a component that consists of sphingosine (or one of its analogs) to which a long-chain fatty acid is connected by an amide linkage. If the glycosyl unit incorporates one or more *N*-acylneuraminic acid residues, the glycoside is called a ganglioside (Fig. 3*c*). Plant and microbial glycolipids are more diverse and include glycosides having *myo*-inositol derivatives, hydroxy fatty acids, or glycerol derivatives as the aglycon. Although their biological function is poorly understood, some are useful diagnostic antigens for certain types of lung

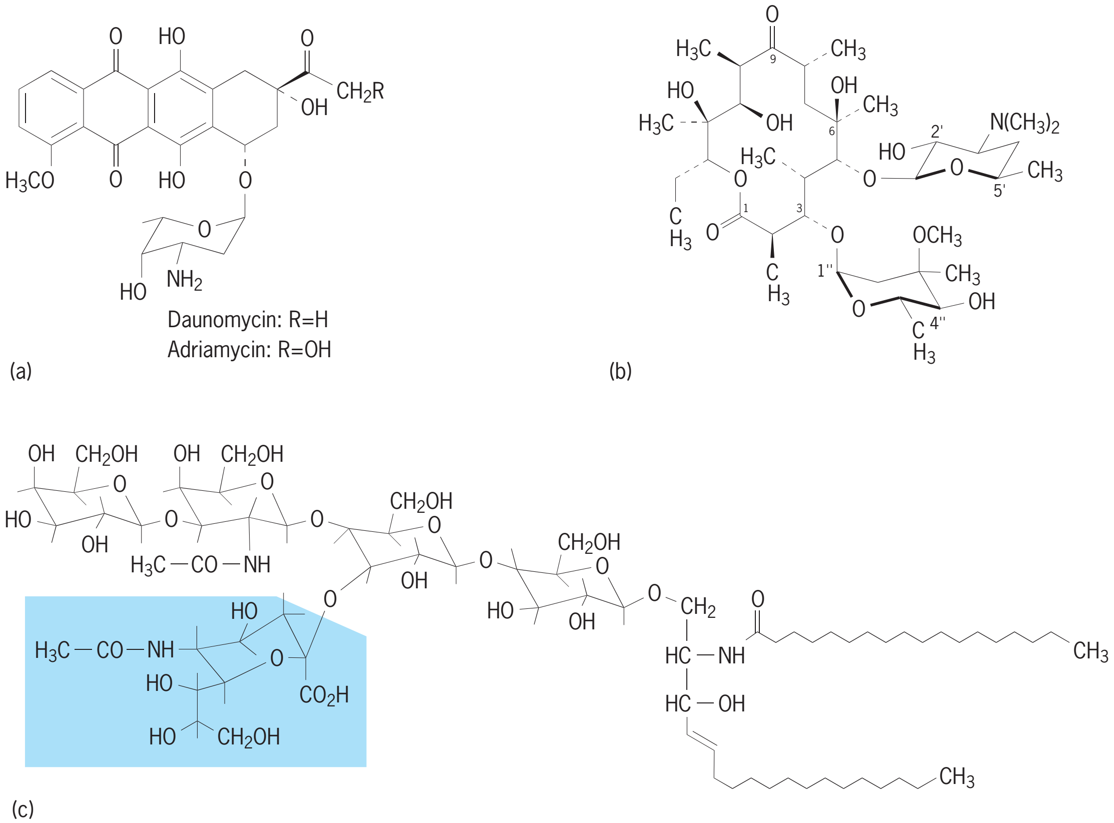


Fig. 3. Structural formulas of some naturally occurring glycosides. (*a*) Anthraquinonyl glucosaminosides (anthracycline antibiotics). (*b*) Erythromycin A, a macrolide (large-ring) glycoside antibiotic. [Me = methyl (—CH_3)]. (*c*) *N*-acetylneuraminic acid, a ganglioside; shaded area is the ceramide.

and colon carcinomas. *See* GLYCOLIPID; IMMUNOASSAY; LIPID. G. W. Hay

Bibliography. R. W. Binkley, *Modern Carbohydrate Chemistry*, 1988; A. E. Bochkov and G. E. Zaikov, *Chemistry of the O-Glycosidic Bond, Formation and Cleavage*, 1979; W. Pigman and D. Horton, *The Carbohydrates, Chemistry and Biochemistry*, 1972; L. G. Wade, Jr., *Organic Chemistry*, 4th ed., 1998.

Gnathostomata

A superclass of the subphylum Vertebrata (Craniata). Gnathostomes are animals with jaws, involving a vertical biting that developed from modified gill arches. They are further characterized by a notochord (an elongated dorsal cord of cells that is the primitive axial skeleton in all chordates) that is present in the ontogeny of all lineages but replaced by vertebral centra (the main bodies of vertebrae) in most taxa; early fossil forms with a bony exoskeleton that was lost in the higher lineages; limb girdles (secondarily lost in some forms) supporting paired appendages in all but the most primitive taxa; myelinated neurons; an adaptive immune system; intrinsic eye muscles; a sperm duct linked to the urinary system; and a distinct cerebellum. Some of the characters were carried over from the primitive superclass Agnatha, but jaws and paired appendages are unique to Gnathostomata. *See* ANIMAL EVOLUTION; JAWLESS VERTEBRATES; VERTEBRATA.

The gnathostomes are divided into three major grades, the hierarchical classification of which is shown below (the taxa are featured elsewhere in the Encyclopedia):

Superclass Gnathostomata
  Grade Placodermiomorphi
    Class Placodermi
  Grade Chondrichthiomorphi
    Class Chondrichthyes
  Grade Teleostomi
    Class: Acanthodii
      Actinopterygii
      Sarcopterygii

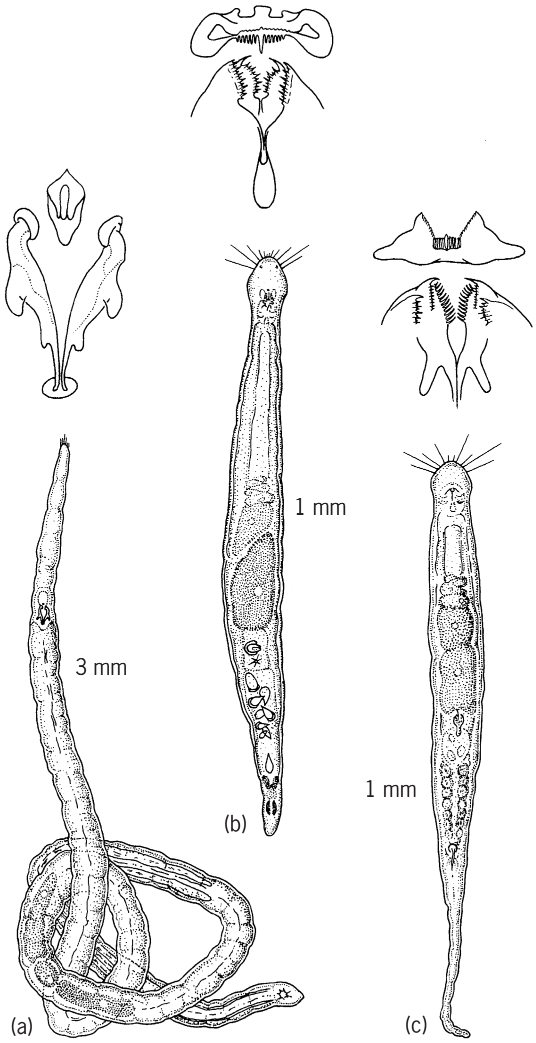
In the past, gnathostomes were recognized as two superclasses, Pisces and Tetrapoda, which constitute roughly the aquatic and the terrestrial vertebrates, respectively. The term Pisces is no longer recognized in classification, but "jawed fishes" is a convenient term for all taxa from placoderms (extinct armored fishes) to coelacanths (lobefin fishes) and lungfishes (sarcoptergyian fishes), with about 27,900 species. The remainder of the sarcopterygians are tetrapods, with about 26,750 species. The earliest gnathostomes, the placoderms, appeared in the fossil record no later than the Middle Silurian, about 425 million years before the present. *See* PISCES (ZOOLOGY); SARCOPTERYGII; TETRAPODA. Herbert Boschung

Bibliography. J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Gnathostomulida

A phylum of microscopic marine worms related to Rotifera and Micrognathozoa, mainly characterized by complex cuticular structures in the pharynx and a monociliated skin epithelium. Discovered in the 1920s on the German coast but not described until 1956, Gnathostomulida have since been reported from many sheltered sandy shores around the world. With fewer than 100 known species, this is one of the smallest animal phyla.

Morphology. Gnathostomulids range from 0.01 to 0.14 in. (0.3 to 3.5 mm) in length. They are worm-shaped, cylindrical, and semitransparent (or bright red), and sometimes have the external divisions of head and tail (see **illustration**). The skin is a one-layered epithelium that is completely monociliated; each of the polygonal epidermal cells bears only one cilium which can be up to 20 micrometers long. Parenchyma is poorly developed, and excretory organs may be present as three paired groups of protonephridia. The nervous system is largely basiepithelial, consisting of an unpaired buccal and unpaired frontal ganglion (brain) from which one to three pairs of longitudinal nerves originate. The sensory system usually consists of one or two pairs of simple bristles and three or four pairs of compound bristles (frontally and laterally) and a row of stiff cilia (dorsally on the head). In some genera, ciliary pits and spiral ciliary organs of unknown function are present in the anterior head region.



Representative Gnathostomulida showing details of basal plate and jaws. (*a*) *Haplognathia simplex*. (*b*) *Austrognathia riedli*. (*c*) *Gnathostomula mediterranea*.

Great diversity exists in the foregut area. The mouth, located ventrally in the head (in *Haplognathia* far from the anterior end), is equipped with complex cuticularized structures: sometimes with a jugum, a cartilaginous structure, in the upper lip, mostly with paired jaws and an unpaired basal plate in the lower lip area. The basal plate occurs in many different forms throughout the species but always bears lamellae or teeth in its center. The large mouth cavity is surrounded by a complex muscle apparatus (consisting of a dozen pairs of individual muscles) which becomes progressively more compact from *Haplognathia* to *Gnathostomula* types. The same is true for the jaws which vary from simple forceps types (*Haplognathia*) to complicated lamellar snap jaws with three rows of up to 60 teeth (see illustration). The midgut is simple; large cells surround a gut cavity; a permanent anus is lacking.

All gnathostomulids are hermaphrodites. The reproductive system consists of an unpaired dorsal ovarium, and paired (in most genera) or unpaired (in *Austrognathia*) caudolateral testes in the same specimen. In the majority of the genera, the male apparatus includes a simple tubular copulatory stylet, composed of a concentrically arranged bunch of cuticular rods and accompanied by glands. This type of male organ is always associated with a female organ consisting of a dorsal bursa (for sperm storage) with a lamellar wall on whose anterior extension a cuticularized opening ("mouthpiece") is found. *Haplognathia*, on the other hand, has no bursa at all, whereas *Austrognathia* often has a vagina and a bursa, the latter lacking cuticular structures. Sperm are very diverse, ranging from tiny aflagellar droplets (in the majority of genera) or large cones (in *Austrognathia*) to motile threads propelled by a single flagellum (as in *Haplognathia*).

Physiological activities. Fertilization is internal; after injection through the skin or the vagina and subsequent storage in the bursa or in the gut wall, the sperms one by one reach the mature egg. Oviposition (egg deposition or laying) is effected by rupture of the dorsal body epithelium, whereupon the egg becomes spherical and sticks to the substratum. Development is direct, with spiral cleavage and gastrulation as is typical for mesolecithal eggs (that is, eggs containing an intermediate amount of yolk). The hatching juvenile measures some 100 μm in length; of its pharynx apparatus, only the central part of the basal plate is developed. Despite its impressive armature, the sophisticated pharyngeal apparatus simply serves to scrape the molecular film of organic material [together with cyanobacteria (blue-green algae), bacteria, and fungal hyphae] off the sand grains. Body fragmentation is a phenomenon that cannot be explained as yet, whereas the formation of a mucous cyst, as observed in some *Austrognathia*, might help to overcome deterioration of the environment.

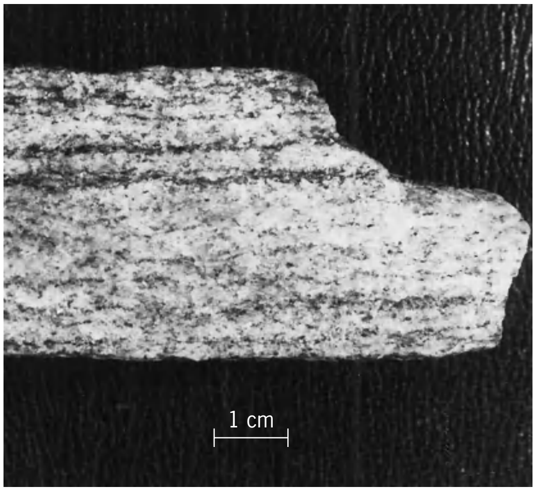
Ecology and distribution. The late discovery of the phylum can be attributed to at least three facts: (1) Sand rich in organic detritus [which is the typical biotope (an area of uniform environmental conditions and biota)] did not receive much attention in classical sand-fauna research. (2) Many gnathostomulids prefer a deeper (and mostly anaerobic) sand layer, especially the boundary zone between the oxidized surface and the reduced "black" layer, where steep surface chemical gradients occur. (3) Methods of extraction, mostly based on the fact that in a sediment sample the animals migrate to the surface, were such that H_2S-producing samples were discarded prior to the rather late appearance of gnathostomulids. The aspect of the biotope suggests that at least part of the metabolism (and obviously also development) takes place under anaerobic conditions. Investigations have shown that whenever the typical gnathostomulid biotope is encountered, it regularly produces several genera (up to six) and species (up to 15 so far), often in hundreds of specimens per liter of sand. Biotopes range in depth from the upper tidal to 1300 ft (400 m). The distribution of gnathostomulids is worldwide, the majority of localities being known from European coasts and the east coast of North America, with a growing number of reports from the Pacific Ocean.

Relationships. The gnathostomulids belong to the general grouping of "lower worms." First considered to be aberrant Turbellaria, they are now thought to be a separate phylum related to Rotifera and Micrognathozoa. A relationship with the enigmatic fossil conodonts could not be substantiated. *See* CONODONT; ROTIFERA. W. E. Sterrer; Rupert J. Riedl

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; M. V. Sørensen and W. Sterrer, New characters in the gnathostomulid mouth parts revealed by scanning electron microscopy, *J. Morphol.*, 253: 310–334, 2002; M. V. Sørensen, W. Sterrer, and G. Giribet, Gnathostomulid phylogeny inferred from a combined approach of four molecular loci and phylogeny, *Cladistics*, 22:32–58, 2006; W. Sterrer and S. Tyler (comp.), Taxonomic database of the Gnathostomulida, Version 1.0 [http://devbio.umesci.maine.edu/styler/gnathostomulida], 2003.

Gneiss

An ancient geological term for coarse-grained, banded crystalline rock. The word, first applied by German miners to rocks widely exposed in the Harz Mountains, was adapted to scientific usage in the late eighteenth century.



Gneiss formed by metamorphism of preexisting granite. Dark minerals are mica; light-colored minerals are quartz and feldspar. The streaky nature of banding is typical of gneisses. The sample is from the Great Smoky Mountains of North Carolina.

Gneiss is composed of mineral grains large enough to be seen with the naked eye (see **illus.**). Banding arises from segregation of the various minerals present, typically into dark- and light-colored layers. Individual bands are commonly 0.04 to 0.4 in. (1 mm to 1 cm) thick. Although individual mineral grains are often flattened parallel to banding, such shape orientation is not present in many gneisses. Sheetlike minerals such as micas may be present but form only a subordinate amount of the rock. Banded rock of coarse grain containing substantial amounts of such minerals is named schist. Crystalline rock which has flattened grains but lacks obvious banding is generally called leptite. *See* SCHIST.

Gneiss is defined by its texture, or arrangement of mineral grains, rather than by its mineral composition. However, the term gneiss is often taken to imply a mineral composition of granitic type, dominated by quartz and feldspar. Gneisses of other compositions are identified by qualifying terms such as compositional rock names, as in diorite gneiss and amphibolite gneiss, or a partial list of minerals present, as in biotite-plagioclase gneiss and hornblende-plagioclase gneiss. *See* FELDSPAR; QUARTZ.

Most gneisses are formed by recrystallization of preexisting rock during intense regional metamorphism. Shear stress present during such metamorphism causes formation of gneissic banding, although the exact mechanisms of this process are not well understood. Thus gneissic banding should never be mistaken for preexisting structures in the rock, such as sedimentary bedding. In most cases, recrystallization of minerals is dominant over shear, and individual minerals are not deformed. However, bent and fractured minerals may occur in mylonitic gneisses, in which shear was dominant over recrystallization. A variety of preexisting rock types (igneous, sedimentary, or metamorphic) may be converted into gneiss. Some granitic gneisses may be of direct igneous origin, having formed by the shear arising from differential flow of viscous granitic magma. *See* MYLONITE.

Gneisses typically occupy large areas within the high-grade cores of regional metamorphic belts. Such terranes are often difficult to understand, because the processes which cause formation of gneissic texture are also sufficient to obscure preexisting rock structures. High temperature and shear are sufficient to cause plastic flow of gneissic rock on a gigantic scale. Major, map-size structures which have been identified in such terranes include mantled gneiss domes and nappes.

Such conditions of metamorphism are probably brought about by deep tectonic burial and major regional compression. Thus gneissic terranes may be expected to form in areas of convergent plate tectonics. *See* METAMORPHIC ROCKS; METAMORPHISM; METASOMATISM; PLATE TECTONICS. David W. Mohr

Bibliography. M. J. Hibbard, *Petrography to Petrogenesis*, 1994; A. Miyashiro, *Metamorphic Petrology*, 1994; A. Spry, *Metamorphic Textures*, 1969.

Gnetales

The only order of the class Gnetopsida in the division Gnetophyta. There are three living families, each with a single genus: Ephedraceae (*Ephedra*; 65 species in arid regions), Gnetaceae (*Gnetum*; 29 species in the tropics), and Welwitschiaceae (*Welwitschia*; 1 species in Namibia). Among the gymnosperms, Gnetales are usually considered the closest living relatives of the angiosperms. This relationship is supported by such morphological features as wood vessels, netted venation, and double fertilization, but contemporary molecular evidence is ambiguous, sometimes favoring this relationship and sometimes implying a closer affinity with the conifers. The fossil record is sparse but increasing, perhaps extending back to the Triassic Period. Each modern genus is quite distinctive. Species of *Ephedra* (Mormon tea) somewhat resemble the spore-bearing horsetails (*Equisetum*), with scale leaves on photosynthetic, jointed stems. Most species of *Gnetum* are lianas, woody vines closely resembling the Australian flowering plant family Austrobaileyaceae. The only species of *Welwitschia* is unlike any other living plant. From its stumplike trunk, just two leaves grow ever wider and longer [up to 7 m (23 ft)], tattering at the tips. All three genera are dioecious, with separate male and female individuals. The pollen and seed cones are compound, with flowerlike buds in the axils of bracts. *See* CYCADOPSIDA; PINOPHYTA; PLANT KINGDOM. James E. Eckenwalder

Bibliography. W. E. Friedman (ed.), Biology and evolution of the Gnetales, *Int. J. Plant Sci.*, 157(6,suppl.):S1–S125, 1996.

Gnotobiotics

The science involved with maintaining a microbiologically controlled environment, and with the knowledge necessary to obtain and use biological specimens in this environment. The roots of the word are *gnotos*, meaning well known, and *biota*, the combined flora and fauna of a region.

All exposed surfaces of an animal are teeming with microbes. The contents of the large intestine may contain 3 trillion microbes per ounce (100 billion per gram), belonging to several hundred species. This microbiota (bacteria, viruses, molds, yeasts, protozoa, and small parasites) is so complex, so incompletely characterized (some cannot yet be grown in pure culture), and so subject to changes in relative proportions that knowledge of what that microbiota is and what it does cannot be directly gained. Pathogens are more amenable to study under such ill-defined conditions because of their dramatic, overriding effects on the host.

Even if the animal itself is the primary interest of a researcher, there is no direct way to determine how many of an animal's normal characteristics are truly its own, and how many involve interaction with or reaction against resident microbiota. The only way to determine this is comparison with animals that have no microbiota. If differences are found, then the role of individual microbial species can be studied by inoculating pure cultures of these species into the animals without a microbiota.

Thus, gnotobiotics evolved initially to answer questions about what difference the resident microbiota makes, and which members of the microbiota make the difference. Answers become more and more essential in going beyond the effects of pathogenic microbes to the harmful or helpful long-term effects of environmental chemicals, compounds produced in the host's own metabolism, and therapeutic drugs being tested for efficacy, toxicity, or carcinogenicity. The activity of the microbiota is proportionately much greater in laboratory animals than it is in humans, and could have a decisive effect on such chemicals, especially since the chemicals are often received in small doses. For example, intestinal microbes turn a minor component of the cycad bean, a South Pacific foodstuff, into a carcinogen. One of the best drugs against parasitic schistosomes in humans is turned into a carcinogen by a single species of intestinal streptococcus. Gnotobiotic studies are designed to detect such possibilities.

Gnotobiotes. Gnotobiote is the term applied to an animal (or plant) with a defined microbiota. The most simple of gnotobiotes, and the ultimate source of all other gnotobiotes, is the animal with no microbiota. Invertebrate animals of this type are most frequently called axenic. Vertebrate animals may also be called axenic, but are more frequently called germfree. These terms suffice until defined microbial species are added to the germfree or axenic animal. Then the axenic animals becomes gnotoxenic. The germfree animal, which was already a germfree gnotobiote, becomes an associated gnotobiote. As defined operationally by the Committee on Standards of the Institute of Laboratory Animal Resources, “A gnotobiote is one of an animal stock or strain derived by caesarean section or sterile hatching of eggs that is reared and continuously maintained with germfree technics under isolator conditions and in which the composition of any associated fauna and flora, if present, is fully defined, by accepted current microbiology.”

Gnotobiology. This is a term sometimes used to designate studies involving gnotobiotes, although it tends to suggest that there is a unified body of knowledge which results from studying gnotobiotes. In fact, the gnotobiote is a more precisely defined laboratory animal which helps elucidate biological phenomena in immunology, nutrition, physiology, oncology, gastroenterology, microbial ecology, gerontology, pathogenic microbiology, parasitology, and so on. Nevertheless, some of these findings serve a dual purpose; besides their contribution to other disciplines, they prove to be necessary knowledge for obtaining, maintaining, or intelligently using gnotobiotic animals, and thus, by definition, belong in the science of gnotobiotics.

History and significance. Louis Pasteur, the originator of the concept (not the term) gnotobiotics, suspected a profound symbiosis and even mutualism between a host animal and its microbiota. He said that if he had the time he would like to grow animals free of all germs to test his assumption that they would not be able to continue life. In 1886, M. Nencki, the discoverer of ptomaines, countered with chemical reasons to support the opposite hypothesis of E. Metchnikoff, namely, that much of an animal’s microbiota has a toxic, life-shortening effect and that the germfree animal would be better off by not having any. Early attempts at rearing germfree chickens and guinea pigs by both factions showed that animals could survive germfree, at least briefly, but poor health resulted; thus the experiments could be used to argue for or against either hypothesis.

It is now realized that both sides were right. Herbivorous animals consuming their natural diet are completely dependent on their microbiota to supply vitamin B_{12} and possibly other substances. G. Nuttall and H. Thierfelder, who in 1985, carried out the first experiment with germfree guinea pigs, discovered that they could rear them better with animal-derived feeds than with vegetable feeds; this was 55 years before identification of animal protein factor, which is actually a microbial product, vitamin B_{12}. However, the researchers were dissatisfied with the results of both diets and did not continue. Other one-shot experiments to test either hypothesis were carried out over the next three decades. They yielded only short-term results because of inadequate isolator design and materials, and because nutritional science was not yet equal to the task of providing diets which were adequate after sterilization and made up for any microbial contribution to digestion and to production of nutrients.

Furthermore, the very first germfree experiments with guinea pigs revealed an unexpected effect of microbes on the host. Germfree rodents and rabbits develop a conspicuous distention of the cecal pouch at the junction of the small and large intestines. The cecum is already relatively large in these species under open animal room conditions. The presence of this large amount of germfree cecal contents was subsequently found to decrease basal metabolism and cardiac output. The fact that surgically removing the cecum brought the germfree parameters toward those in germ-bearing animals indicated that the microbes’ effect on the cecum and on metabolism was the result of their removing some materials which were generated by the host during digestion and accumulated in the cecum. It has taken years of experimentation with different kinds of diet to reduce cecal distention to minor proportions.

As a result of knowing more of what microbes do to benefit their host and learning how the experimenter can substitute for these microbial effects by dietary manipulation, Metchnikoff’s hypothesis that the microbiota is life-diminishing can now be tested. In fact, germfree rats lived one-third longer than their open animal room counterparts (often referred to as conventional or classic animals). This is so not merely because they escape overt infections but because they show a much delayed development of nephrosis and of tumors and cancers.

In thus answering the question posed by Pasteur and Metchnikoff a century ago, the basic importance of gnotobiotic research is seen. The whole animal, integrating all the subtle, long-term effects of a microbiota, is the basis for a whole-life experiment. In an era when molecular biology, cell and tissue culture, and genetic engineering are used to find molecular explanations for biological phenomena, gnotobiotics is a reminder that it is the animal, at a higher level of organization, which points out the phenomena to be explained. *See* GENETIC ENGINEERING; MOLECULAR BIOLOGY; TISSUE CULTURE.

Building on the work of researchers during the 1930s and 1940s with animals in absolute isolation (**Fig. 1***a*), clinical surgeons and immunologists have learned that the use of antibiotics to eliminate major elements of the human microbiota, and the use of partial or complete isolation to prevent their reentry (**Fig. 2**), can greatly reduce infections in human patients with spontaneous immune deficiency or with iatrogenic immune deficiency resulting from procedures to cure cancer or prevent transplant rejection. Children with severe combined immune deficiency of genetic origin have lived as well as can be expected inside germfree isolators for periods up to 12 years. Knowledge and techniques developed with germfree animals are potentially relevant to such clinical applications. *See* IMMUNOLOGICAL DEFICIENCY.

Laboratory methods. Obtaining, maintaining, testing, and using gnotobiotes has become routine for the most common laboratory and domestic animals (rats, mice, pigs, guinea pigs, rabbits, chickens, quail, dogs, cats, primates, goats, cattle, and horses); the basic procedures are the same for all (Fig. 1*b*).

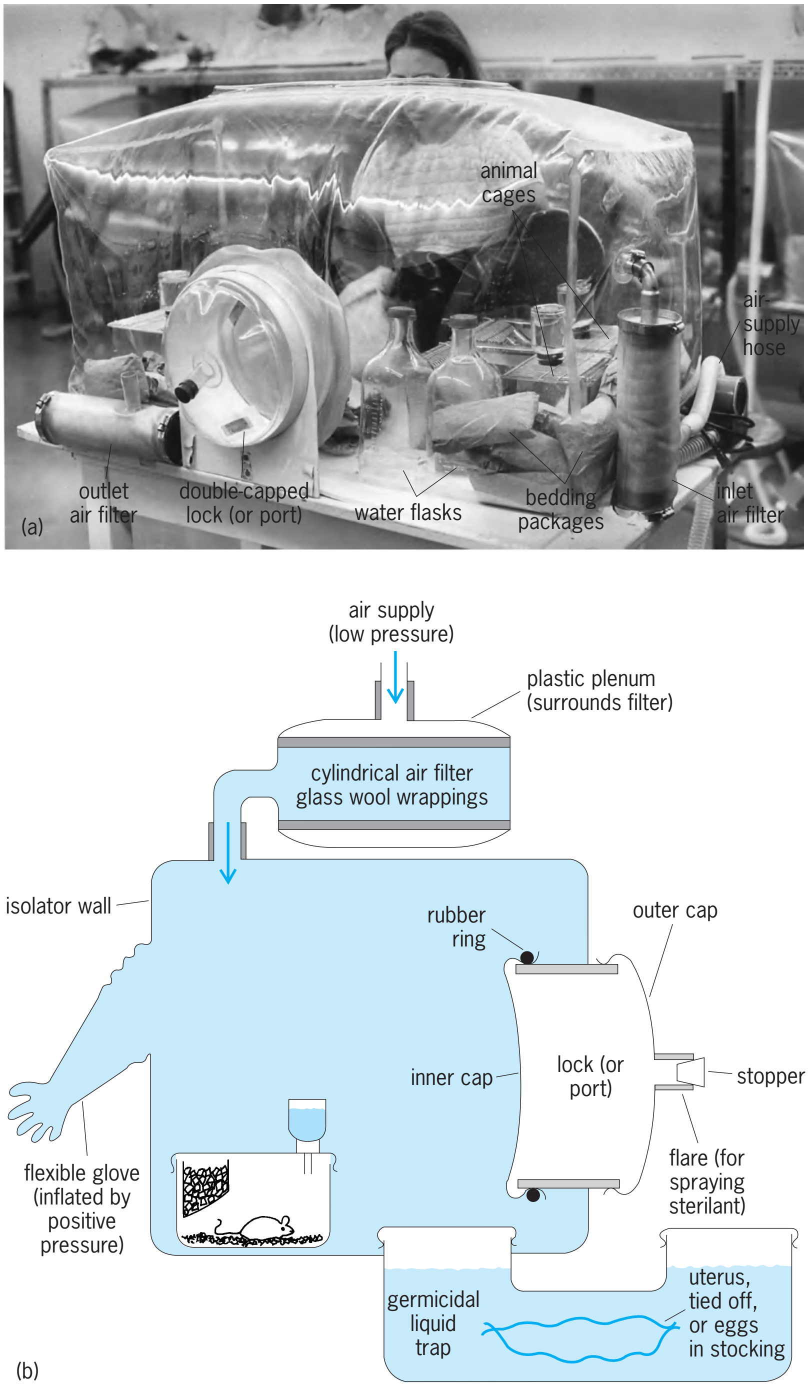

animal cages
air-supply hose
outlet air filter
double-capped lock (or port)
water flasks
bedding packages
inlet air filter
(a)
air supply (low pressure)
plastic plenum (surrounds filter)
cylindrical air filter glass wool wrappings
isolator wall
rubber ring
outer cap
inner cap
lock (or port)
stopper
flare (for spraying sterilant)
flexible glove (inflated by positive pressure)
germicidal liquid trap
uterus, tied off, or eggs in stocking
(b)

Fig. 1. Typical Trexler-type flexible film isolator. (*a*) Operating unit; operator is handling a bag of diet. (*b*) Diagrammatic cross section, including germicidal liquid trap to show one way of introducing new germfree animal lines.

Obtaining gnotobiotes. Obtaining an animal germfree or with so few associates that they can all be specified is the initial step. Nature protects the immunodeficient fetus in the womb or the chicken in the egg by keeping it germfree; to keep it so throughout life it must be removed aseptically into a germfree environment. This is done by cesarean section in a surgical isolator, with the germ-laden mother underneath the plastic floor of the isolator. The young are transferred to a rearing isolator for hand feeding. Alternatively the whole uterus may be removed, clamped shut, and passed through a germicidal trap in the same way that eggs are passed through (Fig. 1*b*).

Experience has shown that nature is not completely reliable in prenatal isolation of the young from all microbial forms. While there are extensive rat colonies which have never shown any evidence of bacterial, viral, fungal, or parasitic associates, gnotobiotic mice of all strains carry a leukemogenic virus which is present in cells of the unborn fetus, and generally remains latent unless activated by repeated small doses of radiation. Exhaustive tests are therefore essential before a cesarean-derived line can be declared germfree or gnotobiotic.

Chickens are easy to obtain germfree. The egg surfaces are chemically sterilized in a lock or trap (Fig. 1*b*), and the chicks hatch inside the isolator where they are able to feed without assistance. Rearing of the first germfree small mammals, however, forced gnotobiotic researchers to pioneer development of sterilizable substitutes for the milk of each species, and of miniaturized feeding devices. Standard laboratory-animal solid diets proved inadequate for germfree animals until modified, primarily in vitamin and mineral content, to compensate for losses during sterilization and the absence of microbial vitamin synthesis. Once reproducing germfree colonies were established, they not only maintained themselves but provided foster mothers for cesarean-derived young of new strains.

As an alternative to cesarean derivation, treatment of older animals with antibiotic mixtures and germicidal baths has proved quite effective in eliminating bacteria, fungi, and protozoa, but not any viruses that may be present. Several months of treatment, with frequent transfers to new isolators, are required.

Maintenance of gnotobiotes. Maintenance for long periods demands absolute barriers (Fig. 1). The barrier must be complete and must consist of materials proven impervious to microbes. Gloves that are an integral part of the barrier permit manipulation of the isolator contents. A sterilizable lock, with inner and outer doors, provides for entry of supplies (**Fig. 3**) and exit of wastes, and for transfers between isolators through a sterilizable corridor connecting two locks, similar to the connecting sleeve shown in Fig. 3. The lock may be replaced by a trap of liquid germicide in which the liquid forms a barrier to airflow and sterilizes the surfaces of objects passed through it (Fig. 1*b*). Air is sterilized by passage through fine glass-wool filters. Diets and bedding are not suitable for surface sterilants, and must be sterilized by steam, radiation, or penetrating gas in containers that can then be surface-sterilized or connected to a lock.

Isolating systems have included heavy-metal isolators capable of holding steam under pressure for maximum sterilizing efficiency; lighter-weight metal isolators that could be put inside a large steam autoclave; sterile rooms entered by operators in diving suits; rigid plastic boxes; and flexible, transparent plastic isolators (Trexler-type isolators) which now represent the most common, simple, and cheap type

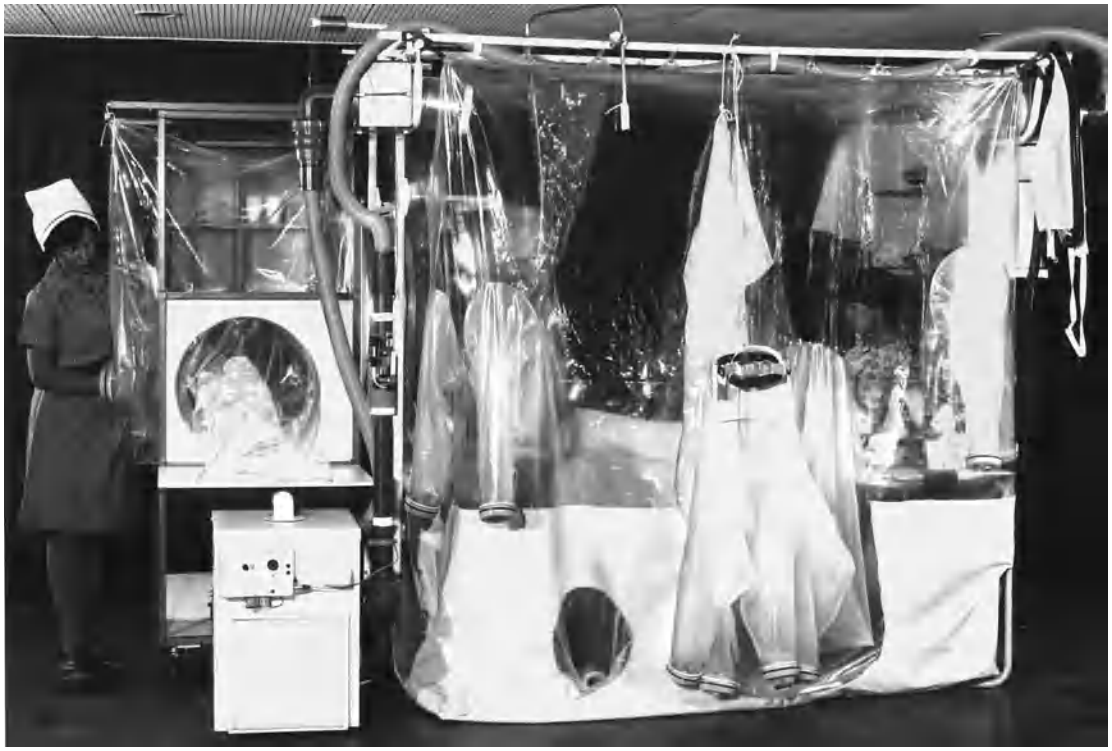


Fig. 2. Patient isolator used to protect immune-compromised patients from outside germs, and to protect hospital personnel from infection by patients with highly lethal tropical diseases. (*Vickers Americal Medical Corp.*)

of secure isolation for gnotobiotes (Fig. 1). This last type was made possible by the development of cold sterilants such as peracetic acid which can sterilize plastic and metal surfaces inside the isolator, the surfaces of objects passing through the lock, and the air initially within the isolator or lock.

Testing. Testing is required to verify both the germfree state of the initial animals and the successful maintenance of absolute barriers thereafter. The best methodologies must be applied: a variety of culture media and incubation conditions, light and electron microscopy, tissue culture, animal symptoms, and the animals' possession of antiviral antibodies. The maintenance of a reproducing colony gnotobiotically provides time for exhaustive testing of that line of animals since it was cesarean-derived. Thereafter, simpler testing of fecal matter for microscopically visible and cultivable bacteria usually suffices to detect breaks in the barrier, since the most ubiquitous microbes in the environment are relatively easy to detect, and a rare one is not likely to pass through a break alone.

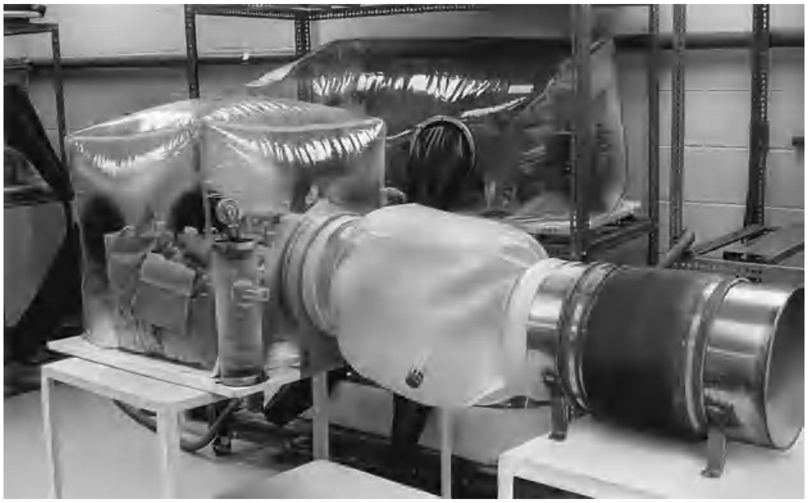


Fig. 3. Transfer sleeve connecting an isolator to a cylinder containing food and bedding which has been steam-sterilized. A thin plastic film covering the cylinder end inside the sleeve will be punctured from inside the isolator after the sleeve has been spray-sterilized. The same sleeve is used for sterile transfers between isolators.

Characteristics of germfree vertebrates. Germfree vertebrates obtained by the above operations reveal characteristics which are valuable data for potential users, demonstrate the far-reaching influence of the microbiota on the host, and provide new insights in basic and applied disciplines. In the following capsule summaries, characteristics have been selected in which the germfree animal differs notably from the usual or conventional animal.

Anatomy. Hearts, lungs, and livers of germfree animals are as much as 20% smaller than those of conventional animals. Germfree rodents and rabbits (not other species) have ceca about five times the conventional size, yet their small intestines are generally thinner-walled and have less surface area and white cells.

Function. Germfree intestines show slower peristalsis, reduced sloughing of intestinal surface cells, and higher enzyme content in those cells. Germfree intestinal capillaries are 20 times less responsive to Adrenalin than those of conventional animals.

Effects of gnotobiotic association

Host	Associated microbe	Effect
Chicken	*Bacillus subtilis*	Heart lesions and death
	Streptococcus faecium	Growth depression
	Clostridium perfringens	Growth depression; local immune response
	Streptococcus faecalis	Humoral immune response
	Lactobacillus species	Improved starch digestion; elimination of *Escherichia coli*
	Eimeria tenella	No clinical symptoms in associated gnotobiote; death of conventional host
Turkey	*Histomanas meleagridis*	Much milder disease in associated gnotobiote than in conventional host unless *Escherichia coli* added
Guinea pig	*Escherichia coli*	Enteritis and death
	Streptococcus faecalis	Prevention of *Escherichia coli* pathogenicity
	Entamoeba histolytica	No growth parasite in associated gnotobiote; death of conventional host
	Bacillus subtilis	Host death in presence of *Entamoeba histolytica*
	Nippostrongylus muris	Growth of rat parasite only in associated gnotobiote guinea pig
	Nematospiroides dubius	Growth of mouse parasite only in associated gnotobiote guinea pig
Mouse	*Nematospiroides dubius*	Slower worm growth in associated gnotobite than in conventional host; incomplete life cycle in associated gnotobiote
	Trichinella spiralis	Milder disease in associated gnotobiote than in conventional host
	Escherichia coli	Elimination of *Candida albicans*
	Escherichia coli plus *Proteus mirabilis* and *Streptococcus faecalis*	Elimination of *Vibrio cholerae*
	Mycoplasma pneumonia	No infection in associated gnotobiote; pneumonia in conventional host
Mouse irradiated	*Escherichia coli*	Shorter survival in associated gnotobiote than in conventional host
	Lactobacillus acidophilus	Longer survival in associated gnotobiote than in conventional host
Rat	*Streptococcus mutans*	Rampant and typical tooth decay
	Actinomyces naeslundii	Periodontal (gum and bone) disease
	Bacillus anthracis	Much greater pathogenicity in associated gnotobiote than in conventional host
	Escherichia coli	Rapid elimination of *Shigella flexneri*
	Salmonella typhimurium	No mortality on one diet, 25% mortality on another
Pig	*Treponema hyodysenteriae*	No infection in associated gnotobiote; severe infection in conventional host
	Lactobacillus species or *Streptococcus* species	More rapid maturation of natural killer cells
Dog	Distemper virus	Much milder symptoms in associated gnotobiote than in conventional host
	Histoplasma capsulatum	Milder disease in associated gnotobiote than in conventional host

Germfree rats have a resting metabolic rate one-fourth less than conventional rats and one-third reductions in heart output and liver blood flow. Thyroid hormone levels do not decline with age in germfree rats and mice, as conventional levels do.

Growth rate. Growth rates are superior for germfree chickens, primates, pigs, and Japanese quail. Most other species show equal germfree and conventional rates, but germfree rabbits and guinea pigs have a lower rate than conventional ones.

Life-span. Germfree rats live one-third longer than conventional rats. Species like the rabbit and guinea pig in which cecal size is not well controlled by diet die earlier in germfree than in conventional status because the cecum may accidentally twist on itself.

Cancers. Spontaneous tumors occur later in germfree than in conventional rats and are usually endocrine-related. The longevity of germfree rats permits spontaneous appearance of new cancer types, such as prostate tumors, to provide animal models for human cancer. Not all types of chemical carcinogens are active in germfree animals.

Cholesterol metabolism. Germfree rats do not excrete cholesterol and bile acids in their feces as efficiently as they would microbially modified forms and therefore accumulate more cholesterol than conventional rats.

Radiation sickness. Germfree animals tolerate larger doses of radiation, live longer, and permit study of radiation effects per se without supravening infection.

Nutrition. Germfree animals need less vitamin A, protein, lysine, and vitamin C than conventional animals, but show poorer utilization of iodine and synthetic vitamin K. Germfree animals fed pure chemical diets have provided the first proof that all essential nutrients are now known.

Immune system. Germfree animals survive thymus removal at birth, permitting long-term study of the organ's role in disease resistance and aging. Germfree animals receiving bone marrow transplants show little reaction of the foreign white cells against their host. This fact has provided the basis for treatment of leukemia with transplantation preceded by antibiotics and isolation.

Germfree animals show conspicuous reductions in circulating antibodies, antibody-producing cells, size of lymph nodes and the number of their active centers, and blood clearance of bacteria by phagocytic cells. These differences can be accentuated by use of antigen-free, ultrafiltered water-soluble diets. Despite reduced ongoing function, germfree animals retain the ability to respond to challenges, and can be used to study natural antibody, natural killer cells, and the nature of a true primary immune response.

Associated gnotobiotes. Otherwise germfree animals associated with one or more defined microbial species supply models for identifying the beneficial or harmful species among the resident microbiota, with the ultimate goal of controlling these in the host's favor. Since germfree animals are also gnotobiotes, several terms to describe associated gnotobiotes have been devised. Gnotophoric indicates the animal's carrying of known microbes, and gnotoxenic indicates the animal's association with known microbes. The **table** shows a selection from the more dramatic effects of associating germfree animals with one or several defined microbial species. Many experiments too complex for the table have shown that combinations of certain microbes from the resident microflora strongly resist the establishment of pathogenic microbes in the gut, a phenomenon called colonization resistance. Bacteria selected for this capacity through gnotobiotic research are now used to reconstitute the microflora of human patients after extensive antibiotic therapy.

Nearly all experiments with parasites (protozoa or worms) added to germfree animals have shown dramatic differences from the parasites' effects in conventional hosts. The parasites either need bacteria to establish themselves, or are dramatically antagonized by bacteria, or provoke less pathogenic effects or immune response when alone.

In terms of numbers of animals involved, the greatest use of gnotobiotes is to upgrade the quality of nearly all experimental rodents. Before the advent of gnotobiotes, all rat and mouse colonies carried ancestral pathogens, such as murine viral pneumonias, which were passed from parents to offspring after birth. These latent infections became active when the animals were stressed or kept for long periods, as occurred in many experiments. Gnotobiotes provided commercial animal breeders with pathogen-free parents or foster parents for new animal colonies which now supply the majority of the millions of rats and mice used annually in research. These are not gnotobiotic but owe their freedom from specific pathogens to the gnotobiotic colonies which commercial breeders maintain. Julian R. Pleasants

Bibliography. M. W. Balk and E. C. Melby, Jr., *The Importance of Laboratory Animal Genetics: Health and Environment in Biomedical Research*, 1984; J. Kawamata and E. C. Melby, Jr. (eds.), *Animal Models: Assessing the Scope of Their Use in Biomedical Research*, 1987; W. F. Loeb and F. W. Quimby (eds.), *The Clinical Chemistry of Laboratory Animals*, 2d ed., 1999; B. S. Wostmann et al., *Germfree Research: Microflora Control and Its Application to the Biomedical Sciences*, 1985.

Goat production

An agricultural business concerned with breeding goats, primarily for meat, milk, and fiber (called mohair and cashmere). The domesticated goat is closely related and similar in size to sheep, but has many anatomical and physiological differences. Domesticated goats of the major breeds have been developed in, and are mainly derived from, five geographical areas: Swiss mountains (for milk); Indian, Arabian, and northeast African drylands (for meat and milk); west African lowland (for meat and milk); south African prairie (for meat); Turkish highland (for mohair production). *See* CASHMERE; MOHAIR; SHEEP.

Accompanying colonizing immigrants and released from the live-food supplies of early merchant ships, goats became established in the Americas, and today make up a small (compared to cattle, swine, poultry, and sheep) but increasingly visible and important part of agriculture in the United States. Mohair production from Angora goats is of economic significance in Texas and ranks first in the world, followed by that of Turkey. Six different dairy goat breeds are recognized in the United States: Alpine, LaMancha, Nubian, Oberhasli, Saanen, and Toggenburg. *See* BEEF CATTLE PRODUCTION; DAIRY CATTLE PRODUCTION; SWINE PRODUCTION.

Breeds. Alpine goats are of Swiss origin. They have many colors, especially faded shades of white into black, a straight face, erect ears, and sickle-shaped horns, when present. Alpine does (mature females) may weigh at least 135 lb (60 kg), measure 30 in. (75 cm) at the withers (top of shoulder), and are (with Saanen and Nubian breeds) the biggest and heaviest dairy goats. Alpine goats are ranked second in milk-producing ability, behind the Saanen breed.

La Mancha goats are of Spanish origin but were developed in California. They have many colors, a straight face, but characteristically no external ears. They are smaller than Alpine goats and produce less milk, but their offspring are fleshier. They are very calm goats, and persistent milk producers in the winter when the other breeds often dry up.

Nubian goats are mostly of Indian origin, but were developed in England. They too have many colors, but characteristically arched faces and long pendulous ears. Nubian goats produce less milk than the Swiss breeds, but their milk-fat level is the highest. They are the most popular in the United States and are most heat-tolerant. They are also more fleshy. Horns in Nubian, as in LaMancha, goats are rarely seen. Although anatomically present, they are routinely and permanently removed in early age to facilitate animal handling. Horns in all Swiss breeds are of the upright sickle shape, while Indian, Spanish, and Turkish breeds have more sideways-growing horns with varying degrees of spiraling. Goats can be selected for the genetic trait of hornlessness, but this is only partially advisable since the trait is genetically linked to sterility, and a certain percentage of intersex or hermaphrodite offspring can be expected when both parents carry the hornless trait.

Oberhasli goats are of Swiss origin. They are related to Alpine goats and similar to them, except that they are often colored solid red or black. They are smaller and lower in milk production than Alpine goats, but very popular in Europe where they are well adapted to high mountain grazing and long-distance traveling. They have also been called

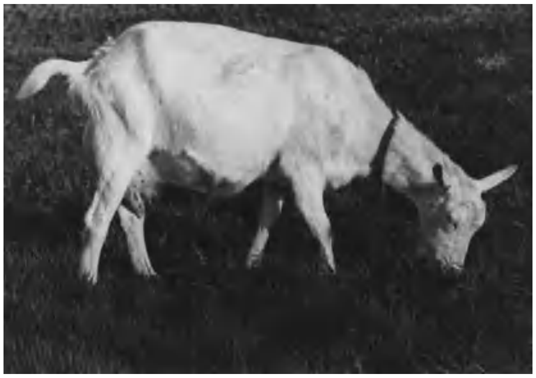


Fig. 1. A typical Saanen doe with good udder.

Chamoisee, Brienz, or Swiss Alpine goats in contrast to the French Alpine and British Alpine breeds which are today just called Alpine goats.

Saanen goats are of Swiss origin. They are only white and have straight faces and erect ears (**Fig. 1**). They are rated as the top milk producers around the world and have been used for the upgrading of many native breeds in underdeveloped countries. All dairy goat breeds in the United States are generally short-haired, although some Swiss strains may have partially long hair, which is of some advantage to them in alpine climates.

Toggenburg goats are also of Swiss origin. They are very popular in the United States and rank in milk-producing ability next to the Alpine and Saanen goats. They have only brown color with characteristic white facial markings and white ear and leg stripes. They are the smallest of the Swiss breeds, approximately 120 lb (54 kg) for the mature does and 26 in. (65 cm) height at withers. They have straight faces and erect ears.

In addition to the six American dairy breeds above (and the Angora), there are also two major goat populations in the United States for mixed purposes, mostly meat, brush clearance, pets, or laboratory use: the Spanish goat, mainly kept on open range land, and the West African Pygmy.

Milk. Milk is the main product of dairy goats in the United States (**Fig. 2**). There are two manufacturing plants for dry goat milk, one in California and one in Arkansas. Interest, mostly of limited local scale, is increasing in goat cheese production: soft curd (similar to cottage-ricotta), cheddar, and feta. Impetus comes from European and Near East traditions and imports, especially from France, which has a thriving, sophisticated Chèvres (soft goat cheese) industry. Liquid raw goat milk can be officially marketed in only 14 states in the United States, those with public health laws permitting such sales. In all other states, goat milk must be pasteurized. There is a strong interest nationwide in raw goat milk by people with malnourished children, by persons with cow milk allergies and gastric problems, and by those interested in natural foods. This situation has led to a wide dispersion of mostly small goat herds throughout the United States to satisfy needs on a family level.

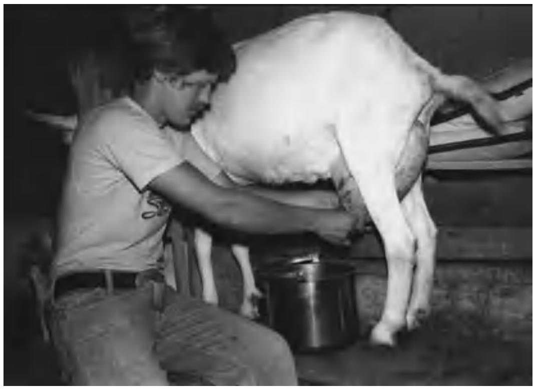


Fig. 2. Milking a dairy goat on an elevated stand.

Goat milk is valued for raising orphaned foals and puppies and for the production of prime veal. Biochemically goat milk differs in several important aspects from cow milk. The major casein in cow milk (alpha-s-1) is found in neither goat nor human milk. The casein type in goat milk makes a softer curd and is more easily digested. The fat in goat milk consists of smaller globules than in cow milk and is more easily digested. The goat milk fat consists characteristically of much more short-chain fatty acids than cow milk.

There are also some consistent differences in vitamin, mineral, and enzyme contents between goat and cow milk; however, both are more similar in gross composition than either is to sheep milk, which has a much higher solids content and smaller daily yield. Daily milk production of good dairy goats in the United States is at least 1 gallon (3.8 liters) with an average content of 2.8% fat, 3.4% protein, 4.6% lactose, and 0.8% minerals, totaling 11.6% solids. *See* CHEESE; MILK.

Breeding. Breeding of dairy goats is similar to that of sheep. Goat kids, born in the spring, are ready for breeding in the fall of the same year and will have their next generation 150 days after conception, usually the next spring. This cycle coincides with the seasonal changes of vegetation; the spring with new pasture growth favors milk production and provides greater survival opportunity. Thus, the dairy goats (for example, the Swiss breeds) of the Northern Hemisphere with its distinct seasons evolved as seasonal breeders. The disadvantage is the consequent shortage of goat milk during winter months. Goats of Mediterranean, Near East, and Indian origin show less seasonality. Seasonal breeding is governed physiologically by day length; and by artificially reducing daylight in controlled housing, dairy goats can be brought into earlier estrus during the summer. This is a practical and successful method for obtaining an even goat milk supply throughout the year.

Breeding of dairy goats is still mostly by natural service. Bucks produce, in the scent glands on top of their head, odors attractive to does but obnoxious to people. Estrus lasts 1–2 days and may recur in 21 days if impregnation failed. Artificial insemination with frozen semen stored in liquid nitrogen tanks at −320°F (−196°C) is gaining popularity since it is the best and most economical avenue of genetic improvement. Pregnancy testing is possible with ultrasonic scanning and radiography, but more practical by routine testing of milk samples for progesterone levels at 3 weeks after breeding.

Kidding in dairy goats is usually easier than in cows or sheep because of the generally lean condition and narrow bone structure of kids. Two or three kids per pregnancy, not necessarily identical, is the rule rather than the exception, including backward presentations. Rebreeding is not inhibited by nursing, as in some other mammalian species, but rather by the increasing day length of spring.

Modern developments are estrus synchronization of does with hormone implants or injections to overcome estrus detection problems and to facilitate artificial insemination, and embryo transfer in conjunction with estrus synchronization to enable wider use and greater number of offspring of genetically superior does. Embryo freezing is possible even now, and the sorting between different sex embryos as well as cloning from single superior embryos will be practical in the near future, thus opening exciting new methods of dairy goat improvements. Considering that in many parts of the world, where the majority of the 450 million goats are found, the productivity per animal usually averages only 10–20% of the average goat milk production in the United States, the modern breeding techniques in developed countries have contributed to the solution of malnutrition problems in the third world. *See* BREEDING (ANIMAL).

Feeding. Feeding of dairy goats is similar to that of cows and sheep, except that goats need and thrive on browse like deer, and some breeds may go without water for days. High-producing dairy goats need scientifically balanced rations as do high-producing dairy cows. *See* ANIMAL FEEDS.

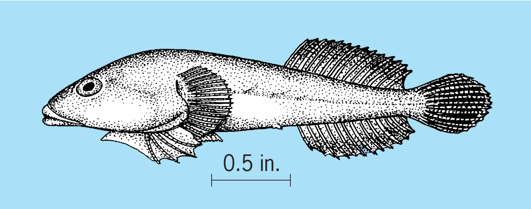
Diseases. Facilities and equipment are not strongly developed for goats specifically; but as with housing, milking machines, and feed storage, medications against parasites, diseases, and metabolic disorders are adapted from, or patterned after, those for cattle. Unlike in dairy cows, the largest problem in dairy goats is internal parasites, which appear to adapt often to certain standard medications. Also, unlike in dairy cows, tuberculosis and brucellosis are not a problem in dairy goats in the United States, although annual testing continues to be advocated for safety reasons. On the other hand, there are viral diseases causing arthritis, bacterial infections producing abscesses, and intestinal bacterial disorders which are more serious in dairy goats than in dairy cows—particularly because veterinary research, education, and care in the United States are more focused on cattle than on goats. *See* AGRICULTURAL SCIENCE (ANIMAL).

George F. W. Haenlein

Bibliography. J. L. Ayers and W. C. Foote, Proceedings of the 3d International Conference on Goat Production and Disease, *Dairy Goat J.*, 1982; G. F. W. Haenlein et al., Proceedings of the International Symposium on Dairy Goats, *J. Dairy Sci.*, 63(10):1591–1781, 1980; G. F. W. Haenlein and D. L. Ace, *Goat Extension Handbook*, University of Delaware, Department of Animal Science and Agricultural Biochemistry, 1983; P. Morand-Fehr et al., Nutrition and systems of goat feeding, *Proceedings of the International Symposium at Tours*, Institut National de Recherche Agronomique, Paris, 2 vols., 1981; National Research Council, *Nutrient Requirements of Goats*: *Angora*, *Dairy*, *and Meat Goats in Temperate and Tropical Countries*, 1981; S. N. Singh and O. P. S. Sengar, *Final Technical Report*: *Studies on the Combining Ability of Desirable Characters of Important Goat Breeds for Meat and Milk Separately and in Combination*, Department of Animal Husbandry and Dairying, Raja Balwant Singh College, Bichpuri (Agra), India, 1979.

Gobiesociformes

An order of bony fishes, also known as the Xenopterygii, or clingfishes, equipped with a thoracic sucking disk which serves for attachment to the substrate. This papillose disk involves the four rays of each pelvic fin, pelvic and pectoral girdles, and dermal flaps. There are single dorsal and anal fins that lack fin spines (see **illus.**). The body is scaleless, and there are no ribs and no swim bladder. The posttemporal is not forked. The order consists of a single family that is classified into 8 subfamilies, 33 genera, and nearly 100 Recent species. A Miocene genus is questionably assigned to the family. The Gobiesocidae are thought to be related to the Batrachoidiformes, or toadfishes. Clingfishes inhabit the intertidal zone of tropical to temperate shores of all continents, and a few occur in swift coastal streams of the American tropics. *See* ACTINOPTERYGII; BATRACHOIDIFORMES; SWIM BLADDER.



Northern clingfish (*Gobiesox maeandricus*). (*After D. S. Jordan and B. W. Evermann, The Fishes of North and Middle America, U.S. Nat. Mus. Bull. 47, 1900*)

Reeve M. Bailey

Bibliography. J. C. Briggs, *A Monograph of the Clingfishes* (*Order Xenopterygii*), Stanford Ichthyol. Bull., vol. 6, 1955.

Gödel's theorem

The result, proved by K. Gödel in 1931, that any sufficiently advanced mathematical system must be incomplete in that there must always be a true sentence that is not provable in the system. Roughly speaking, Gödel showed how, for each such system, a sentence could be constructed that asserted its own nonprovability in the system.

Provability. As an illustration, it is useful to consider the following sentence:

This sentence can never be proved.

If the above sentence is false, then what it says is not the case, which means that it can be proved, but false sentences cannot be proved. Thus it cannot be false; it must be true. Now, it has just been proved that the sentence is true. Since it is true, then what it says really holds, which means that it can never be proved. But the sentence was, in fact, just proved.

Thus, there is an apparent paradox. The fallacy in this reasoning is that the notion of proof is not well defined. No precise definition of proof has ever been given in any absolute sense; only provability within a given axiom system is discussed. It is now useful to consider an axiom system, which may be called system S, in which the notion of provability within the system is clearly defined, and to suppose also that the system is wholly correct in that every sentence provable in the system is really true. Then the following sentence may be considered:

This sentence is not provable in system S.

No paradox now arises, but instead the interesting fact that the above sentence must be true but not provable in system S. Indeed, if the sentence is false, then what it says is not so, which means that the sentence is, in fact, provable in system S, contrary to the given condition that the system proves only true sentences. And so the sentence must be true, hence not provable in system S (as the sentence says).

Self-reference via Gödel numbering. In attempting to construct a sentence that asserted its own nonprovability, Gödel considered mathematical systems that did not involve sentences or provability, but numbers, sets of numbers, and other purely mathematical entities. He avoided this problem by assigning to each sentence a positive whole number, now called the Gödel number of the sentence. Then he constructed an ingenious sentence *G* which asserted that a certain number *g* failed to belong to a certain set *P* of numbers. This set *P* was the set of Gödel numbers of all the provable sentences of the system and, more amazingly, the number *g* was the Gödel number of the very sentence *G*. Thus, *G* asserted that its own Gödel number *g* was not the Gödel number of a provable sentence. If *G* were false, then *g* would, in fact, be the Gödel number of a provable sentence, which would mean that *G* was provable, contrary to the fact that only true sentences are provable in the system. Therefore, *G* must be true, hence *g* really is not the Gödel number of a provable sentence, which means that *G* is not provable in the system. Thus, *G* is true but not provable in the system.

Gödel's second theorem. Gödel demonstrated a still more remarkable result that for these same systems, which include the most comprehensive mathematical systems known. He showed that, if these systems are consistent, then they cannot prove their own consistency. This is known as Gödel's second incompleteness theorem.

Unfortunately, there are widespread misconceptions about this result. For example, Gödel's second theorem is sometimes thought to imply the impossibility of knowing that these systems are consistent. In reality, the fact that a system cannot prove its own consistency does not constitute the slightest grounds for doubting its consistency. Indeed, if a system could, in fact, prove its own consistency, that, of course, would not be any guarantee that the system was consistent, since an inconsistent system can prove anything. To trust the consistency of a system just because it can prove its own consistency would be as foolish as trusting a person's veracity just because he or she claimed to always tell the truth. In reality, whether a system can or cannot prove its own consistency does not have the slightest bearing on whether or not the system is actually consistent. The consistency of the systems considered by Gödel is known rather by self-evident nature of the axioms and the obvious correctness of the rules of reasoning. Still, it is of interest that the systems, though obviously consistent, cannot prove their own consistency.

Impossibility of mechanizing mathematics. In the eighteenth century, G. Leibniz envisioned a universal calculating machine that could solve all mathematical problems. The impossibility of such a device has been conclusively demonstrated by further ramifications of Gödel's work developed by A. Church and A. Turing.

It is useful to consider a type of computing machine that is programmed to generate a set of positive whole numbers. For example, a generating machine might be given the following two instructions: (1) Print out 2. (2) If you print out n, then also print out $n + 2$. Then the machine will successively print out the numbers 2, 4, 6, 8, 10, . . . , in other words, the set of even numbers. A machine can be instructed to generate the set of odd numbers simply by changing the first instruction to "Print out 1." Now, a set S of numbers is called solvable (the more technical term is recursive) if there is a machine M to generate S and another machine M' to generate the set S' of all numbers that are not in S. (For example, the set of even numbers is solvable, since there is a machine that generates the even numbers and another that generates the odd ones.) Given the machines M and M', it is possible to mechanically decide of any given number x whether or not it is in S by simply setting both machines going at the same time and waiting to see whether M prints out x, in which case x is in S, or M' prints out x, in which case x is not in S but in S'. *See* RECURSIVE FUNCTION.

There are infinitely many of these possible generating machines, as many as there are positive whole numbers, and each number x is the label of one and only one machine, which may be denoted M_x. Now, it may or may not happen that the set of numbers generated by M_x will contain the very label x of M_x; if this happens, x is called a special number, and if not, x is called an ordinary number. The interesting thing is that one of the machines C prints out all the special numbers but none of the ordinary ones. This machine C, so to speak, keeps track of the activity of all the machines, and whenever it observes that M_x prints out x, then C also prints out x. Now, the distinction between special and ordinary numbers is of considerable mathematical interest, because all formal mathematical problems can be reduced to determining which numbers are special and which are ordinary. And so if the set of special numbers were solvable, then there would exist a completely mechanical method of solving all formal mathematical problems, and Leibniz's universal calculating machine would be realized. Since there is the machine C that generates all the special numbers, the key question then is whether there is a machine C' that generates the set of ordinary numbers. In fact, there cannot be such a machine C' by the following argument:

If there were, indeed, such a machine, then it would have some label b. Thus M_b would print out all ordinary numbers but no special ones. The question then arises as to whether the number b is special or not. In either case, there is a contradiction. On the one hand, if b, indeed, is special, then M_b prints out b, contrary to the fact that M_b prints out only ordinary numbers. On the other hand, if b is ordinary, then, since M_b prints out all ordinary numbers, M_b must print out b, which makes b special. Thus, b cannot be either ordinary or special without contradiction, and hence there cannot exist a machine M_b that generates the set of ordinary numbers. Therefore, the set of special numbers is not solvable.

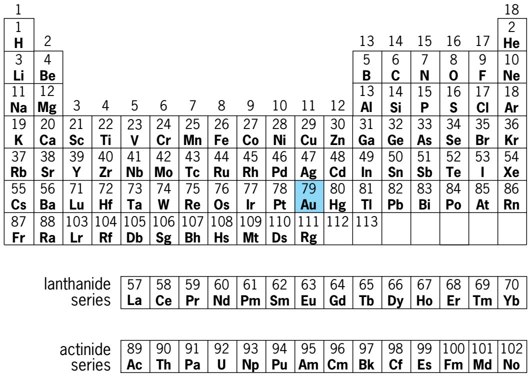
Thus, there is no purely mechanical method of determining which numbers are special and which are not, and so mathematics simply cannot be mechanized. Creativity and ingenuity will always be required. This is perhaps the most important consequence of Gödel's work. Another way of stating this consequence is to say that human beings can never eliminate the necessity of using their own intelligence, regardless of how cleverly they try. *See* LOGIC.

Raymond M. Smullyan

Bibliography. M. Davis, *The Undecidable*, 1973; N. Shankar, *Mathematics, Machines, and Gödel's Proof*, 1994; S. G. Shanker, *Gödel's Theorem in Focus*, 1988; R. Smullyan, *Gödel's Incompleteness Theorems*, 1992.

Gold

A chemical element, Au, atomic number 79 and atomic weight 196.967, a deep yellow, soft, and very dense metal. Gold is classed as a heavy metal and as a noble metal; commercially, it is the most familiar of the precious metals. Copper, silver, and gold are in the same group of the periodic table of elements. The Latin name for gold, *aurum* (glowing dawn), is the source of the chemical symbol Au. There is only one stable isotope of gold, that of mass number 197. *See* PERIODIC TABLE.



1																	18
1 **H**	2											13	14	15	16	17	2 **He**
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**

Uses. Consumption of gold in jewelry accounts for about three-fourths of the world's production of gold. Industrial applications, especially electronic, consume another 10–15%. The remainder is divided among medical and dental uses, coinage, and bar stock for governmental and private holdings. Gold coins and most decorative gold objects are actually gold alloys, because the metal itself is too soft (2.5–3 on Mohs scale) to be useful with frequent handling. *See* GOLD ALLOYS.

Radioactive ^{198}Au is used in medical irradiation, in diagnosis, and in a number of industrial applications as a tracer. Another tracer use is in the study of movement of sediment on the ocean floor in and around harbors. The properties of gold toward radiant energy have led to development of efficient energy reflectors for infrared heaters and cookers and for focusing and retention of heat in industrial processes.

Occurrence. Gold occurs widely throughout the world, but usually very sparsely, so that it is quite a rare element. Sea water contains low concentrations of gold, on the order of 10 μg per ton (10 parts of gold per trillion parts of water). Somewhat higher concentrations accumulate on plankton or on the ocean bottom. At present, no economically feasible process is visualized for extracting gold from the sea. Native, or metallic, gold and various telluride minerals are the only forms of gold found on land. Native gold may occur in veins among rocks and ores of other metals, especially quartz or pyrite, or it may be scattered in sands and gravel (alluvial gold). *See* GOLD METALLURGY.

Properties. The density of gold is 19.3 times that of water at 20°C (68°F), so that 1 ft^3 of gold weighs about 1200 lb (1 m^3, about 19,000 kg). Masses of gold, like those of other precious metals, are measured on the troy scale, which counts 12 oz to the pound. Gold melts at 1064.43°C (1947.97°F) and boils at 2860°C (5180°F). It is somewhat volatile well below its boiling point. Gold is a good conductor of heat and electricity. It is the most malleable and ductile metal. It can easily be made into translucent sheets 0.0000039 in. (0.00001 mm) thick or drawn into wire weighing only 0.00005 oz/ft (0.5 mg/m). The quality of gold is expressed on the fineness scale as parts of pure gold per thousand parts of total metal, or on the karat scale as parts of pure gold per 24 parts of total metal. Gold readily dissolves in mercury to form amalgams. Gold is one of the least active metals chemically. It does not tarnish or burn in air. It is inert to strong alkaline solutions and to all pure acids except selenic acid.

Compounds. Gold may be either unipositive or tripositive in its compounds. So strong is the tendency for gold to form complexes that all the compounds of the 3+ oxidation state are complex. The compounds of the 1+ oxidation state are not very stable and tend to be oxidized to the 3+ state or reduced to metallic gold. All compounds of either oxidation state are easy to reduce to the metal.

In its complex compounds gold forms bonds most readily and stably with halogens and sulfur, less stably with oxygen and phosphorus, and only weakly with nitrogen. Bonds between gold and carbon are fairly stable, as in the cyanide complexes and a variety of organogold compounds. William E. Cooley

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. R. Davis (ed.), *Metals Handbook: Desk Edition*, 2d ed., ASM International, 1998; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; J. Marsden and I House, *Chemistry of Gold Extraction*, 2d ed., 2006; H. Schmidbauer, *Gold: Chemistry, Biochemistry and Technology*, 1999.

Gold alloys

Combinations of gold and other metals. Gold is not oxidized in air, and for thousands of years was the only metal available that remained bright indefinitely. This, coupled with its pleasing yellow color, made it cherished for personal adornment and other decorative uses and valuable as a medium of exchange. The high specific gravity of gold (19.3) also gave some insurance against dilution with cheaper metals of lower density. Pure gold is soft. The addition of copper hardens the gold, and ultimately gold-copper alloys became standard for coinage. Gold coins in the United States contained 10% copper, the balance gold. Prior to 1934 the value of coinage had been based on the price of gold. In that year coinage of gold ceased in the United States, and gold was priced at a market value. The United States Mint sold gold to licensed users for industrial and ornamental uses from 1934 to March 17, 1968. From this date to January 1, 1975, some banks and one broker were licensed to sell gold, but only to licensed users for industrial and artistic purposes. Since 1975, there have been no restrictions on the buying and selling of gold.

Jewelry. Pure gold is weak, having a tensile strength of less than 20,000 lb/in.^{2} (138 megapascals) when annealed; however, by alloying with copper, sometimes in conjunction with silver or nickel, and often a little zinc, gold alloys with strengths of 60,000–100,000 lb/in.^{2} (414–690 MPa) may be made. Addition of these metals changes the color of gold so that red, yellow, greenish, and white golds result. The proportion of gold in solid gold jewelry is designated in karats (k); pure gold is 24 k, 18 k is 18/24 or 75% pure gold, and 14 k is 14/24 or 58.3% pure gold.

Very early it was found that reasonably pure gold could be hammered into extremely thin foil, about 0.000005 in. (125 nanometers) thick, and this was used for architectural and other decorations in pre-Biblical times. Such foil, also known as gold leaf, is still made and used in about the same way.

Because of the relatively high cost of gold it is often used in laminated form, a thin layer of karat gold being welded to a supporting metal, such as nickel or brass, which is rolled or drawn into complex forms without rupturing the gold-alloy surface. This material is called gold-filled or rolled-gold plate, and the ratio of the weight of the gold alloy to the total weight of the composite material is expressed as a fraction, along with the karat of the gold alloy; for example, $^{1}/_{\sqrt{10}}$ 12 k gold means that the article contains 5% of pure gold by weight. An exception to this method of designating quality occurs in watch cases, for which a definite thickness of gold alloy is stipulated.

Properties and uses. Gold is not oxidized in air, even on heating. It is insoluble in all single acids, although it can be dissolved in a mixture of nitric and hydrochloric acids (aqua regia) or in a solution of alkaline cyanides in the presence of air or an electric current. This solubility in cyanide is widely employed in extraction of finely divided gold from ores. The method can recover gold that has escaped amalgamation and many waste heaps from early mining operations have been reworked to extract additional amounts of gold. When an electric current is passed through a cyanide solution, gold can be deposited on other metals as an electroplate; this reaction is often used for protecting and decorating base metals.

Industrial uses of gold depend primarily upon the corrosion resistance and secondarily upon the strength that can be secured by alloying alone or by alloying and heat treating. Many alloys used in dentistry contain gold, silver, and copper, often with small amounts of platinum and palladium; these alloys can be heat-treated to develop strengths above 150,000 lb/in.^{2} (1.0 gigapascal). The latter have good spring properties. Alloys of this type find many electrical uses as contacts, particularly where rubbing is involved. Softer alloys are used for make-and-break electrical contacts, at currents less than 0.5 A, such as those employed in many instruments and in telephone equipment. For the latter application, gold alloys have now largely been replaced by palladium which is more economical and more resistant to electrical erosion. Gold electroplate, often thin, is employed on high-frequency conductors, such as those in radar equipment, because of the high electrical conductivity and tarnish resistance of gold. For the same reason gold is employed in the construction of many transistors, microcircuits, printed circuits, and integrated circuits. In these applications ease of deposition by vacuum evaporation, cathodic sputtering, thermal decomposition, or electrode-position are important. Most such devices are so small that the cost of gold is relatively unimportant.

An excellent thermocouple has been developed that produces nearly the same electromotive force as the widely used Chromel-Alumel thermocouple; one leg is a palladium-platinum-gold alloy and the other leg is a gold-palladium alloy. Thin coatings of gold applied to metallic or glass surfaces may be employed for reflectors, particularly for infrared radiation.

Gold, which melts at a temperature intermediate between the melting points of silver and copper, finds some use, with its alloys, as a corrosion-resistant brazing material in chemical equipment and also in certain electronic vacuum devices. A gold coating is also sometimes applied to the grids of vacuum tubes to minimize electron emission from the grid.

Because gold does not oxidize when heated in air, appropriate gold compounds can be decomposed by heat to liberate the metal. Compounds of this type are used in the decoration of china and also for the production of printed electrical circuits on ceramics. These materials, known as liquid bright golds, are applied in the form of varnish, which is dried and then heated to redness, leaving a thin film of gold firmly attached to the underlying ceramic. This coating may be as thin as 0.00004 in. (1.0 micrometer) but may be made thicker. It is even possible to utilize this technique for applying thin gold coatings to some of the more stable plastics, which are being used for special printed circuits. Certain gold alloys with platinum and palladium can also be applied in this manner to produce stable electrical resistors. *See* GEM; GOLD; GOLD METALLURGY.

Charles R. Marsland; George Sistare

Bibliography. G. Gajda, *Gold Refining*, 2d rev. ed., 1980; K. Von Mueller, *Gold Refiner's Manual*, 1984.

Gold metallurgy

Extracting gold from ores, refining it, and preparing it for use. Total world resources of gold are estimated at about 83,000 tons (75,000 metric tons). South Africa has about half of these resources, and Brazil, Russia, and the United States have about 12% each. The United States produces about 330 tons (300 metric tons) per year. In the United States, gold is used for jewelry and arts (70%), industry and electronics (23%), and dentistry (7%). There are only a few dozen placer mines in the United States, nearly all in Alaska. In such mines, gold is processed with the modern equivalent of gold panning—sluicing, tabling, and jigging. In addition, by-product gold from copper mining is only about 10% (historically, this source used to be much larger). There are several hundred lode mines in the United States, where the ore is mined from solid rock. This gold is often difficult to recover because it is associated with sulfide or carbonaceous minerals. As technology has improved, possibilities for processing the more difficult-to-treat (refractory) ores have expanded. A particular ore is more or less refractory depending on its combination of chemical compounds and minerals. *See* GOLD; PLACER MINING.

Although each ore's treatment process is unique, a number of steps are common to the various methods.

Cyanide leaching. In this process, ore is first crushed dry in a gyratory crusher and ground wet in a semiautogenous grinding mill [a 30-ft-diameter (9-m) rotating cylinder containing steel balls]. During the wet grinding process, cyanide and lime are added. The ore leaves the grinding mill as a slurry of muddy water. At this point, granules of activated carbon are introduced, and the slurry flows into large stirred tanks. The gold gradually leaches out of each tiny ore particle (during its 20–30-h residence time) and dissolves into solution. As the gold dissolves, it is immediately picked up by the carbon, which is then screened from the muddy slurry, rinsed off, and pumped to a gold desorption tank, where gold is stripped from the carbon. Typically, a 20% ethanol stripping solution at 80°C (176°F) is used. The gold has now become concentrated in the stripping solution, and is pumped with the solution into electrowinning tanks containing steel wool as cathodes. There, it is gold-plated onto the steel wool. After several hours, the resulting "golden fleece'' is melted with flux. The flux removes the steel and impurities, and the molten gold is cast into shiny bars. The gold bars are usually sold for further refining. The carbon, now free of its gold, is heated without oxygen to 800°C (1470°F) to reactivate it and to restore its large surface area before it is returned to the slurry tanks to adsorb more gold. In the meantime, the slurry leaves the tanks and, after solid-liquid separation in a thickener, eventually ends up in the tailings pond.

Sometimes lower-grade ore (<0.05 oz/ton or 1.57 g/metric ton of gold) is simply crushed and placed in heaps where it is slowly leached with cyanide solutions. Even though heap leaching gives lower gold recovery (sometimes lower than 60%), the cost is lower. If too much clay is present in the crushed ore, it should be agglomerated by mechanically tumbling the ore with a small amount of water, lime, and portland cement. Agglomeration increases heap permeability and gold recovery and reduces the leaching time. Heap leaching is a form of solution mining that can also be applied to old tailings piles and mined overburden dumps. *See* SOLUTION MINING.

Refractory ores. The vat leaching configuration is known as carbon-in-leach and is one of the simplest ways to overcome preg-robbing—one major characteristic of refractory ores. Preg-robbing ores are difficult to treat because the gold, after first being dissolved into the cyanide solution (now "pregnant" with gold), adsorbs back onto the ore, thereby "robbing" the solution. However, in the carbon-in-leach process the freshly activated carbon is placed directly in the leaching tank. It attracts the gold more strongly than preg-robbing constituents (such as, humic acids, carbonaceous compounds, and certain clay minerals). Furthermore, the carbon flows countercurrent to the slurry, so that the freshest carbon is mixed with the most gold-depleted slurry—an important engineering principle giving maximum efficiency.

If the ore is more refractory, stronger ore oxidation treatments are included in the process. These intense oxidation treatments overcome the other major characteristic of refractory ores, that is, pyrite encapsulation of the gold. The treatments also neutralize or partially neutralize the preg-robbing tendencies. In these treatments, the pyrite, or other encapsulating sulfide mineral, is oxidized. The oxidation of pyrite breaks down the mineral structure, exposes the gold, and makes the gold amenable to subsequent cyanide dissolution. These oxidizing treatments, in order of increasing power and risk, include air or oxygen oxidation of slurries in open tanks, chlorine oxidation of slurries in closed tanks, high-temperature roasting of dry powders in furnaces, high-pressure oxidation of slurries in autoclaves, bacterial oxidation of slurries in open tanks, and bacterial oxidation of coarse crushed ore in heaps.

If the ore is less refractory, the process might be modified to include less powerful alternatives to carbon-in-leach, such as carbon-in-pulp and countercurrent decantation. In the carbon-in-pulp process, the carbon is mixed with the slurry in downstream tanks after leaching. In countercurrent decantation, the carbon is packed in columns. The slurry leaves the leaching tanks and flows through a series of thickeners for solid-liquid separation and thorough washing. Clarified, gold-rich solution overflows from the end thickener and into the carbon columns, where gold is adsorbed as the solution percolates through. Carbon columns are also often used for clarified solutions from heap leaching. Carbon adsorption of silver is less efficient than gold. If silver values in the leach solutions are relatively high [greater than 1.5–3.0 oz/gal (10–20 g/liter)], a zinc cementation process replaces the carbon process. Here, both silver and gold are electrochemically displaced from solution after a countercurrent-decantation-type solid-liquid separation, filtering, vacuum deoxygenation, and the addition of zinc powder. Instead of replacing carbon entirely, another alternative is precipitation with silver sulfide either before or after the carbon adsorption step. Many other alternatives exist for processing a particular gold ore. *See* REFRACTORY.

Chemistry. Gold (Au) dissolution occurs electrochemically in a redox couple: at anodic sites on gold surfaces [reactions (1) and (2)] or at cathodic

$$\mathrm{Au(solid)} \longrightarrow \mathrm{Au^{+}(aqueous)} + e^{-} \qquad (1)$$

$$\mathrm{Au^{+}(aqueous)} + \mathrm{2CN^{-}(aqueous)} \longrightarrow \mathrm{Au(CN)_2{}^{-}(aqueous)} \qquad (2)$$

sites on or near gold surfaces [reactions (3) and (4)].

$$\mathrm{O_2(gas)} + \mathrm{2H^{+}(aqueous)} + 2e^{-} \longrightarrow \mathrm{H_2O_2} \qquad (3)$$

$$\mathrm{H_2O_2} + \mathrm{2H^{+}(aqueous)} + 2e^{-} \longrightarrow \mathrm{2H_2O} \qquad (4)$$

Dissolution rates depend on both cyanide (CN) and dissolved oxygen (O_2) concentration. If cyanide concentrations get low, the anodic reaction is rate limiting and the whole process becomes too slow. Depending on the ore, cyanide concentrations should be kept as low as possible without slowing rates. This helps prevent dissolution of impurities, such as mercury, that must be chemically removed downstream. If oxygen concentrations get too low, the cathode reaction is rate limiting, and the whole process likewise slows down. Sometimes normal stirring and flow is not enough, and compressed air or oxygen must be injected.

The sodium cyanide dissociates into sodium and cyanide ions [reaction (5)]; the cyanide ions react

$$\mathrm{NaCN} \longrightarrow \mathrm{Na^{+}} + \mathrm{CN^{-}} \qquad (5)$$

with water (H_2O) to form cyanide gas [HCN; reaction (6)]. Protective alkalinity must be maintained

$$\mathrm{CN^{-}} + \mathrm{H_2O} \longrightarrow \mathrm{HCN(gas)} + \mathrm{OH^{-}} \qquad (6)$$

(pH 10.5–11.5) to avoid deadly cyanide gas. The hydroxyl ion (OH^{-}) drives reaction (6) to the left if the pH is high enough, but too high pH causes destruction of CN^{-} and unnecessary expense.

Although cyanide is the dominant lixiviant, alternatives such as thiourea, thiosulfate, chlorine, bromine, and iodine have all been proven to work. Their advantages include lower toxicity, more favorable pH ranges, and stronger dissolving power for some refractory ores.

When pyrite is oxidized [one path is shown in reactions (7)–(9), any encapsulated gold is liberated

$$\mathrm{FeS_2} + {}^{7}/_{2}\mathrm{O_2} + \mathrm{H_2O} \longrightarrow \mathrm{FeSO_4} + \mathrm{H_2SO_4} \qquad (7)$$

$$\mathrm{14FeSO_4} + \mathrm{7H_2SO_4} + {}^{7}/_{2}\mathrm{O_2} \longrightarrow \mathrm{7Fe(SO_4)_3} + \mathrm{7H_2O} \qquad (8)$$

$$\mathrm{7Fe_2(SO_4)_3} + \mathrm{FeS_2} + \mathrm{8H_2O} \longrightarrow \mathrm{15FeSO_4} + \mathrm{8H_2SO_4} \qquad (9)$$

and can be dissolved by cyanide. Bacteria (***Thiobacillus ferrooxidans***) in vats and heaps catalyze the conversion of ferrous (Fe^{2+}) to ferric (Fe^{3+}) ion in acidic sulfate [reaction (8)] solutions ($pH < 6$; the optimal growth rate is pH 2.0–2.5). Ferric ion is transported to the pyrite grain, which is oxidized [reaction (9)]. Ferric ion in the solution wetting the ore particles is the key to pyrite oxidation. *See* BIOLEACHING; ELECTROCHEMISTRY; METALLURGY; PH; PYRITE.

Environmental considerations. Government regulations and permit requirements are designed to protect the safety of individual workers, the public, and the environment. Mills must have spill basins. Any storage or tailings ponds must have fences and netting to protect migratory birds and other wildlife, and all accidental wildlife destruction must be reported. Cyanide cannot enter the environment; all cyanide must eventually be destroyed. Cyanide detoxification processes include the sulfur dioxide (SO_2)/air process, the ferrous sulfate process, the alkaline chlorination process, and the bacterial degradation process.

One problem is acid mine drainage. Bacterial action on even tiny amounts of pyrite or other sulfide minerals left in tailings generates acid [reactions (7)–(9)]. Drainage is further contaminated because the acid dissolves toxic metals. The whole process can continue long after the mine is closed. One answer is to continue operating water treatment plants, which typically use lime to neutralize the acid and precipitate metals. *See* LAND RECLAMATION.

Keith A. Prisbrey

Bibliography. R. W. Bartlett, *Solution Mining, Leaching, and Fluid Recovery of Materials*, 2d reprint ed., 1992; M. C. Fuerstenau and J. L. Hendrix (eds.), *Advances in Gold and Silver Processing*, 1989; P. Hayes, *Process Principles in Minerals and Materials Production*, 1993; J. C. Yannapoulos, *The Extractive Metallurgy of Gold*, 1991.

Goldhaber triangle

The phase space triangle, or Goldhaber triangle, corresponds to the kinematically allowed boundary for a high-energy reaction leading to four or more particles. In a high-energy reaction between two particles a and b yielding four particles 1, 2, 3, and 4 in the final state ($a + b \rightarrow 1 + 2 + 3 + 4$), it is convenient to consider the reaction in terms of the production of two intermediate-state quasiparticle composites x and y, which then decay into two particles each, as in expression (1).

$$\begin{array}{llll} a+b\rightarrow & x & & +y \\ & \hookrightarrow 1+2 & & \hookrightarrow 3+4 \end{array} \tag{1}$$

Most high-energy interactions indeed proceed through such intermediate steps, in which, for specific values of the invariant masses $m_x = M_x^*$ and $m_y = M_y^*$, the quasiparticle composites may form resonances.

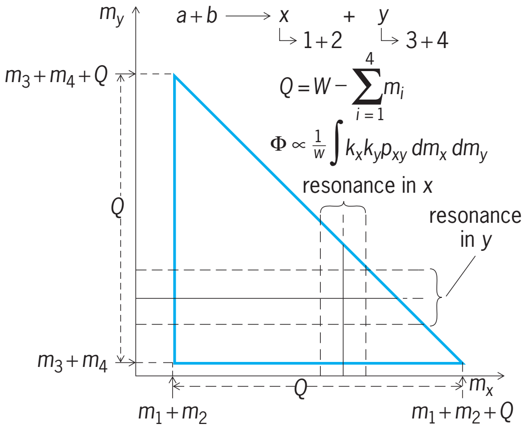
Kinematical limits. The kinematical limits in this representation are particularly simple, namely, they form a right-angle isosceles triangle. A Goldhaber triangle plot corresponds to plotting a point (m_x, m_y) for each event occurring in the above high-energy reaction. Because of the kinematical constraints, these points must all lie inside the triangle.

If one considers the general reaction given in Eq. (1), then the length of each of the two equal sides of the triangle is Q, defined in Eq. (2). Here W

$$Q = W - \sum_{i=1}^{4} m_i \tag{2}$$

is the total energy in the center of mass of particles a and b, and m_i, for $i = 1$ to 4, is the mass of the particles 1 to 4. Hence Q corresponds to the total kinetic energy available in the reaction. All quantities are in energy units of millions or billions of electronvolts.

The values of m_x and m_y run over the intervals $m_1 + m_2 \leq m_x \leq m_1 + m_2 + Q$ and $m_3 + m_4 \leq m_y \leq m_3 + m_4 + Q$, respectively. The effect of changing the incident momentum, and thus Q, is then simply to move the hypotenuse of the triangle, leaving the two sides, as well as the location of any resonances which may occur for certain mass values of the composites x and y, fixed.


m_y
$a+b \longrightarrow x \quad + \quad y$
$\hookrightarrow 1+2 \quad \hookrightarrow 3+4$
$Q = W - \sum_{i=1}^{4} m_i$
$\Phi \propto \frac{1}{W}\int k_x k_y p_{xy}\, dm_x\, dm_y$
resonance in x
resonance in y
m_3+m_4+Q
Q
m_3+m_4
m_1+m_2
Q
m_x
m_1+m_2+Q

Definition of the kinematical boundary of the Goldhaber triangle for four particles.

In the triangle corresponding to the general reaction (see **illus.**), the vertical and horizontal bands indicate resonances at masses M_x^* and M_y^* with full width at half-maximum height or Γ_x and Γ_y, respectively. The bands shown of width 2Γ represent the regions usually chosen if the events corresponding to a given resonance are selected.

Phase space distribution. The phase space is given by $\Phi \propto 1/W \int k_x k_y p_{xy} dm_x dm_y$ where the integral extends over the triangle. Here k_x and k_y are the momenta in the center of mass of the composites x and y, respectively, and p_{xy} is the outgoing momentum of each of the composites in the overall center of mass of particles a and b. It is noteworthy that, along each of the three sides of the triangle, one of the factors in the integrand vanishes.

Comparison with Dalitz plot. Superficially there is a great similarity between the Dalitz plot and the Goldhaber triangle plot. There are, however, several important differences: (1) The Dalitz plot applies to three particles in the final state and corresponds to plotting m_y^2 versus m_x^2, however, in this case the two composites x and y have one particle in common. (2) The Dalitz plot has the advantage that the phase space distribution is uniform but the kinematical boundary is a more complicated function of the variables. In the Goldhaber triangle plot the phase space distribution is more complicated, but the kinematical boundary is very simple—a triangle. (3) If two resonances overlap on the plot, the interpretation is completely different. In the Dalitz plot the overlap corresponds to interference between the two resonances; in the Goldhaber triangle plot the overlap corresponds to double resonance formation, that is, both the x and y composites form essentially independent resonances at the same time. *See* DALITZ PLOT.

Gerson Goldhaber

Bibliography. R. L. Cool and R. E. Marshak (eds.), *Advances in Particle Physics*, vol. 2, 1968.

Golgi apparatus

An organelle found in all eukaryotic cells and absent from prokaryotes such as bacteria. The Golgi apparatus is named after the Italian histologist Camillo Golgi, who visualized an "internal reticular apparatus" by light microscopy in 1898. This organelle consists of flattened membrane-bounded compartments known as cisternae. In most cells, the Golgi cisternae are organized into stacks. Different cell types contain from one to several thousand Golgi stacks. The Golgi apparatus sorts newly synthesized proteins for delivery to various destinations, and modifies the oligosaccharide chains found on glycoproteins and glycolipids. *See* CELL ORGANIZATION.

Structure and location. The detailed morphology of the Golgi apparatus is variable depending upon the cell type. A Golgi stack may contain 4–30 or more cisternae. Usually the stacks are highly ordered (**Fig. 1**), but certain yeasts contain individual Golgi cisternae that are scattered throughout the cell. A typical cisterna measures about a micrometer in diameter. In vertebrate cells, tubules extend from the rims of the cisternae and connect the individual stacks to form a continuous ribbon. During vertebrate cell division, the Golgi ribbon breaks down and then reassembles. By contrast, plant cells contain isolated Golgi stacks that remain intact during cell division. A universal feature of the Golgi apparatus is the presence of multiple transport vesicles of 60–80 nanometers diameter surrounding the cisternae.

Golgi stacks are often located next to endoplasmic reticulum exit sites, which are specialized for protein export. In vertebrate cells, pre-Golgi intermediates are transported away from exit sites by movement along microtubules, thereby generating the Golgi ribbon near the nucleus. When microtubules in vertebrate cells are disrupted, the Golgi ribbon disintegrates to yield isolated Golgi stacks near exit sites.

Protein sorting and secretion. The Golgi apparatus acts at an intermediate stage in the secretory pathway. A subset of the proteins synthesized by the cell are inserted into the endoplasmic reticulum. Most such proteins are then delivered to the Golgi apparatus by means of coat protein II (COPII) transport vesicles, which form at endoplasmic reticulum exit sites (**Fig. 2**). Newly synthesized proteins traverse the Golgi stack until they reach the trans-most Golgi compartment, which is termed the trans-Golgi network to connote its extensive tubulation. The trans-Golgi network sorts the proteins into several types of vesicles. Clathrin-coated vesicles carry certain proteins to lysosomes. Other proteins are packaged into secretory vesicles for immediate delivery

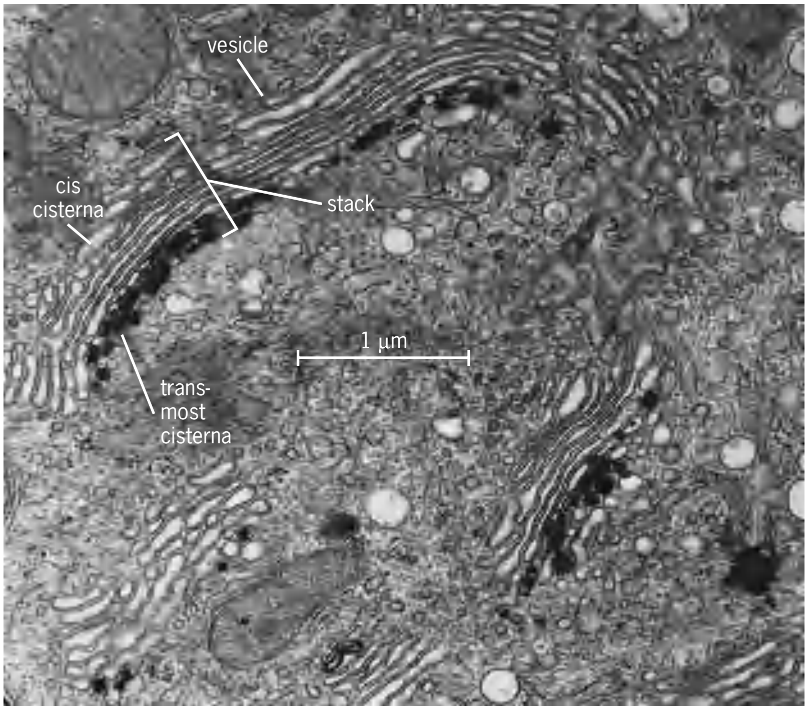


Fig. 1. Electron micrograph of the Golgi apparatus in a rat epididymis cell that was cytochemically reacted for thiamine pyrophosphatase using lead as a capturing agent. The trans-most cisternae are labeled with a black precipitate. (*Photo by Daniel S. Friend*)

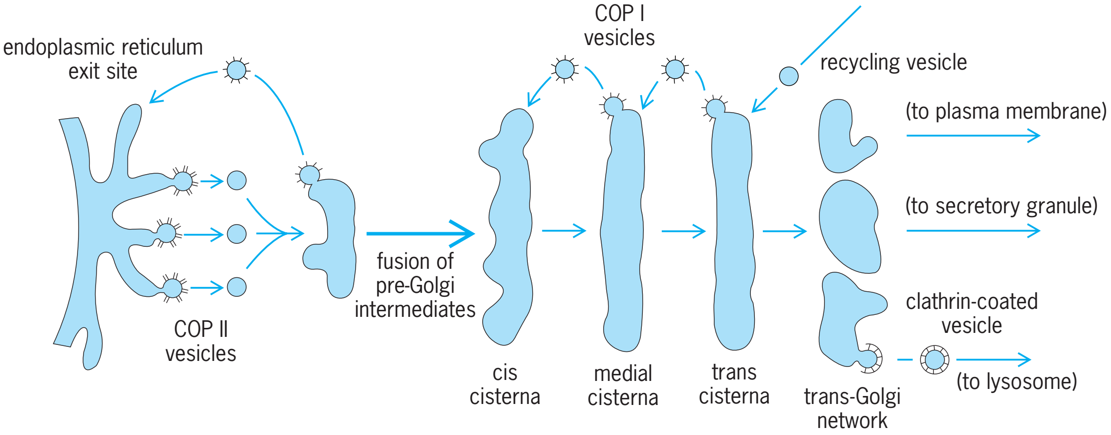


Fig. 2. Cisternal maturation model. At an endoplasmic reticulum exit site, COP II vesicles bud and then fuse with one another to generate pre-Golgi intermediates, which fuse in turn to generate a new cis cisterna. This cisterna then migrates through the stack to the trans face, maturing in the process. Maturation is driven by the recycling of resident Golgi enzymes in retrograde-directed COPI vesicles. At the trans-Golgi network stage, the cisterna is consumed by the formation of vesicles targeted to lysosomes, secretory granules, and the plasma membrane.

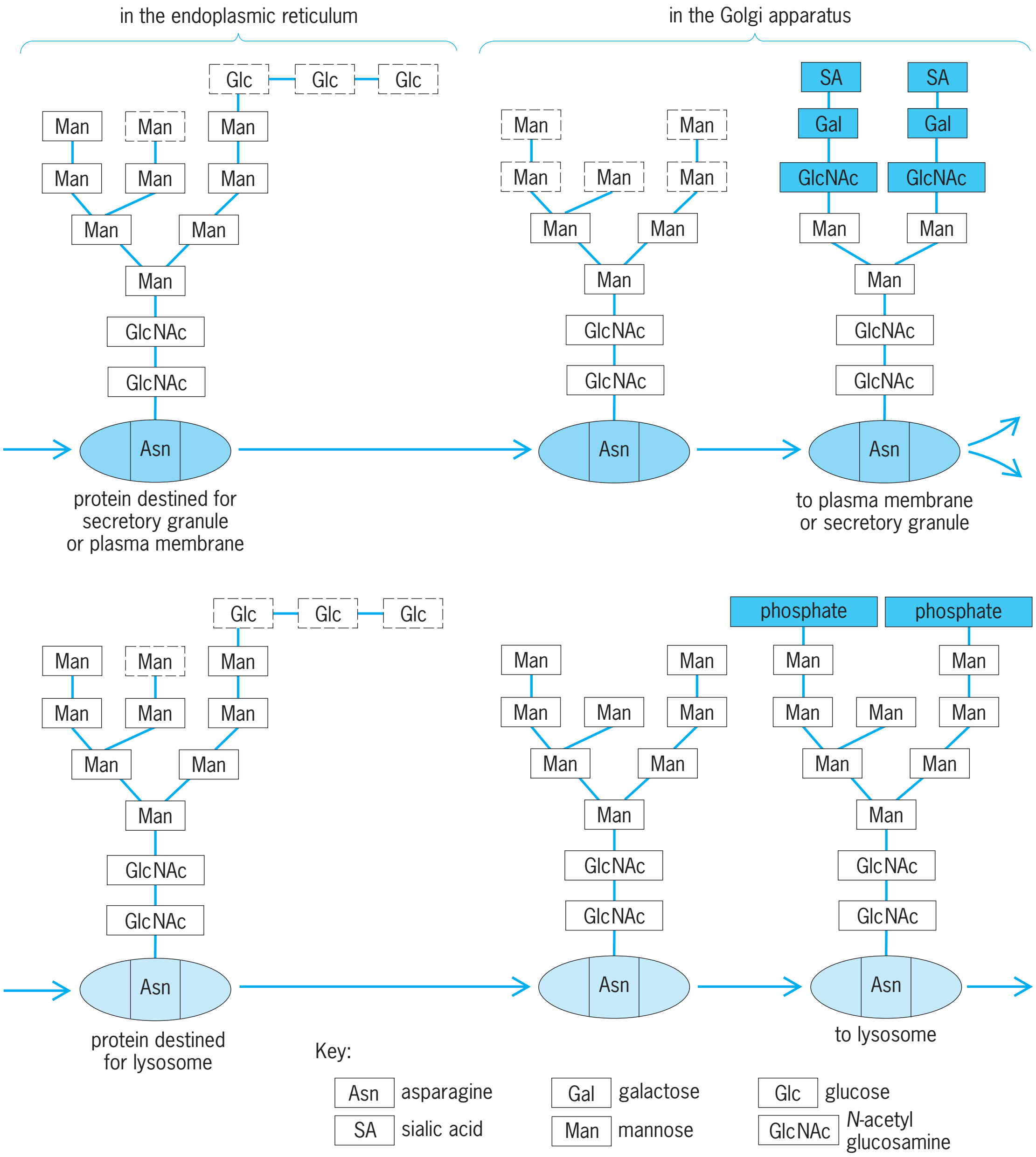

in the endoplasmic reticulum
in the Golgi apparatus
Glc
Man
GlcNAc
Asn
SA
Gal
phosphate
protein destined for secretory granule or plasma membrane
to plasma membrane or secretory granule
protein destined for lysosome
to lysosome
Key:
Asn asparagine
SA sialic acid
Gal galactose
Man mannose
Glc glucose
GlcNAc *N*-acetyl glucosamine

Fig. 3. Remodeling of asparagine-linked oligosaccharides by the vertebrate Golgi apparatus. Boxes with broken outlines represent sugar residues removed in the indicated organelle. (*After J. E. Rothman, Compartmental organization of the Golgi apparatus, Sci. Amer., 253:74–79, 1985*)

to the cell surface. Still other proteins are packaged into secretory granules, which undergo regulated secretion in response to specific signals. This sorting function of the Golgi apparatus allows the various organelles to grow while maintaining their distinct identities. *See* CELL MEMBRANES; ENDOPLASMIC RETICULUM; LYSOSOME.

Transport through the stack. Newly synthesized proteins enter the Golgi apparatus at the cis face and then travel through the medial and trans portions of the stack. The mechanism of protein transport through the Golgi stack is still being debated, but this process somehow involves coat protein I (COPI) vesicles. One theory holds that COPI vesicles carry newly synthesized proteins forward from one cisterna to the next. However, this model does not readily explain how proteins that are too large to fit inside a COPI vesicle can travel through the stack. Most researchers now believe that the cisternae themselves act as forward carriers. According to this cisternal maturation model (Fig. 2), a new cisterna forms at the cis face of the stack by the fusion of COPII vesicles. This cisterna then progresses through the stack to the trans face, where it breaks apart into various types of vesicles. Meanwhile, COPI vesicles move in the retrograde direction to recycle the resident Golgi enzymes. This recycling mechanism allows an individual Golgi enzyme to process a large number of substrate molecules. For example, a newly synthesized protein traverses the vertebrate

Golgi apparatus in about 20 min, whereas a typical Golgi enzyme remains in the organelle for many hours.

Compartmental organization. Each cisterna of the Golgi stack has a slightly different function. The major purpose of the trans-Golgi network is protein sorting. Meanwhile, the other cisternae function to process newly synthesized proteins and lipids. This processing is performed by enzymes that act in ordered pathways. The sequence of processing reactions is reflected in the distribution of Golgi enzymes across the stack. For example, enzymes that catalyze the initial processing steps are concentrated in the cis-most cisternae, whereas enzymes that catalyze the final processing steps are concentrated in the trans-most cisternae. As a result, the various cisternae of the Golgi stack are biochemically distinct (Fig. 1). According to the maturation model, as a cisterna progresses through the stack, it changes its composition by exporting one set of Golgi enzymes to younger cisternae while receiving a different set of Golgi enzymes from older cisternae.

Processing reactions. The best understood of the processing reactions carried out by the Golgi apparatus is the remodeling of oligosaccharides (chains of six-carbon sugars) that are attached to glycoproteins (**Fig. 3**). During insertion of a newly synthesized protein into the endoplasmic reticulum, one or more copies of a 14-sugar oligosaccharide may be attached to the amino acid asparagine at specific locations in the polypeptide chain. As the protein passes through the Golgi stack, the asparagine-linked oligosaccharides are modified to generate a diverse range of oligosaccharide structures. Additional oligosaccharides may become linked to the amino acids serine and threonine. Although the particular oligosaccharide modifications are quite different in animal, plant, and fungal cells, the Golgi apparatus always functions as a "carbohydrate factory." *See* OLIGOSACCHARIDE.

In vertebrate cells, olgiosaccharide remodeling can influence the fate of a newly synthesized protein (Fig. 3). If a protein is destined for either delivery to the plasma membrane or storage in secretory granules, several mannose units are removed from the oligosaccharide. Mannose removal paves the way for addition of variable numbers of *N*-acetylglucosamine, galactose, and sialic acid units. Mannose removal and *N*-acetylglucosamine addition occur in cis and medial cisternae, whereas galactose addition and sialic acid addition occur in trans cisternae. However, if a protein is destined for lysosomes, the mannose units are retained and are modified by the addition of phosphate. When a lysosomal precursor protein reaches the trans-Golgi network, the mannose-phosphate tags are recognized by a specific receptor, and the protein is packaged into a clathrin-coated vesicle for delivery to a lysosome.

The Golgi apparatus also carries out other processing events, including the addition of sulfate groups to the amino acid tyrosine in some proteins, the cleavage of protein precursors to yield mature hormones and neurotransmitters, and the synthesis of certain membrane lipids such as sphingomyelin and glycosphingolipids. *See* LIPID; PROTEIN.

Biochemical and genetic studies. Recent insights into Golgi apparatus function have come from analyzing the molecular components that drive protein transport through the secretory pathway. One experimental strategy was to develop a test tube system that reconstitutes transport through the Golgi stack. This biochemical approach identified COPI, as well as proteins that catalyze the fusion of vesicles with their target membranes. A complementary strategy was to generate yeast mutants with defects in secretion. This genetic approach identified COPII, as well as proteins that guide transport vesicles to their correct destinations. These molecular studies revealed that the principles underlying Golgi apparatus function have been conserved across eukaryotic evolution. *See* CELL (BIOLOGY); METABOLISM.

Benjamin S. Glick

Bibliography. E. G. Berger and J. Roth (eds.), *The Golgi Apparatus*, Birkhäuser Verlag, Basel, 1997; B. S. Glick and V. Malhotra, The curious status of the Golgi apparatus, *Cell*, 95:883–889, 1998; J. E. Rothman, Compartmental organization of the Golgi apparatus, *Sci. Amer.*, 253:74–89, 1985; J. E. Rothman and F. T. Wieland, Protein sorting by transport vesicles, *Science*, 272:227–234, 1996.

Gonorrhea

A common sexually transmitted disease caused by the bacterium *Neisseria gonorrhoeae*. Gonococci are gram-negative microorganisms that usually appear as pairs (diplococci). Humans are the only natural hosts for *N. gonorrhoeae*, which directly infects the epithelium of the mucous membranes of the human genital tract, pharynx, rectum, or conjunctiva. Local epithelial cell destruction usually occurs, but the organisms may spread to adjacent organs or disseminate via the bloodstream. In women, local complications include inflammation of the uterine lining (endometritis), inflammation of the fallopian tube (salpingitis), inflammation of the abdominal wall (peritonitis), and inflammation of Bartholin's glands (bartholinitis); in men, periurethral abscess and inflammation of a duct connected to the testes (epididymitis). Systemic manifestations such as arthritis or dermatitis may develop, and rarely endocarditis or meningitis.

Women are disproportionately affected by the complications of gonorrhea. Acute pelvic inflammatory disease and salpingitis, the most serious complications of gonorrhea, result in ectopic pregnancy and infertility. Gonococcal infection during pregnancy may also predispose women to premature rupture of membranes, delivery in less than full term, and postpartum endometritis. During childbirth, the gonococcus may infect the conjunctiva of the infant and result in the infection ophthalmia neonatorum. This infection is a serious complication that remains common in less developed countries and can lead to permanent damage to the eye and blindness. *See* INFERTILITY; REPRODUCTIVE SYSTEM DISORDERS.

Epidemiology. Gonorrhea continues to be the most commonly reported communicable disease in the United States. The decline in the incidence of gonorrhea in the United States since 1984 has lagged behind that of many European countries such as Sweden where most cases are the result of importation. Data from developing countries are incomplete; however, gonorrhea rates are generally higher than those in western Europe or the United States.

Risk markers and risk factors for gonorrhea are similar to those for other sexually transmitted diseases. Risk markers for gonorrhea, that is, correlates of the probability of encountering an infected partner, include urban residence, young age, race, socioeconomic status, and marital status. Risk factors that may influence the probability of infection include number of sexual partners, lack of barrier contraceptives (for example, condoms), and young age. In the United States, over 71% of the cases reported in males and more than 82% of those in females occur between the ages of 15 and 29 years.

Gonorrhea is an infection spread by physical contact with the mucosal surfaces of an infected person, usually a sexual partner. There is no evidence that natural transmission occurs from toilet seats or similar objects. The risk of infection depends on the anatomic site, the amount of substance containing bacteria, and the number of exposures. Variations in host susceptibility have not been well defined. However, about one-third of men acquire the disease from a single exposure of vaginal intercourse with an infected female; after four exposures, the risk of infection increases to 60–80%. The transmission rate from an infected male to a female after a single exposure is thought to be 50–60%. In a small but significant proportion of infections, there are no symptoms. These individuals are important in the epidemiology of this disease because gonorrhea is usually spread by carriers who have no symptoms or have ignored symptoms.

Control and prevention. Control of gonorrhea depends on early diagnosis, effective treatment, and identification of asymptomatic individuals. The last has been accomplished, in part, through screening programs that have contributed to the detection and treatment of gonorrhea, particularly in women. However, complete control has not been possible because of the emergence and spread of strains that are resistant to less-expensive antimicrobial treatments such as penicillin and tetracycline.

There is no evidence that infected individuals develop long-lasting immunity to reinfection, and vaccination is not available. Thus, the prevention of gonorrhea relies on behavior modification and risk reduction, use of appropriate screening and diagnostic tests, routine use of highly effective antibiotics, early identification and treatment of sexual partners of individuals with gonorrhea, and the appropriate use of barrier methods such as condoms.

Clinical aspects. The usual incubation of gonococcal urethritis in the male is 2 to 7 days from the time of exposure; however, longer intervals are not uncommon. Acute gonococcal urethritis is characterized by discharge of pus, painful urination and swelling and localized redness of the urethral opening. A scant, almost nonpurulent discharge may develop in approximately one-fourth of men, and a few (5–10%) may never develop urethral discharge. Before antibiotic treatment became available, symptoms of urethritis persisted for an average of 8 weeks. Local complications of gonococcal urethritis, which include acute epididymitis, chronic inflammation of the prostate, inflammation of the lymph nodes, inflammation of the seminal vesicles, inflammation or abscess of Cowper's gland, and periurethral abscess, are uncommon occurrences in industrialized countries. Urethral stricture remains a common complication in developing countries.

In women, the glandular mucous membrane of the cervix (endocervix) is the primary site of urogenital infection. Symptomatic and asymptomatic urethral infection occurs in a substantial portion of women with endocervical infection. Although most women are asymptomatic, clinical signs usually develop within 10 days after infection and include increased vaginal discharge due to inflammation of the endocervix, painful urination, abnormal menstrual bleeding due to endometritis, and anorectal discomfort. Pelvic inflammatory disease, which occurs in approximately 15–20% of women with gonorrhea, is the most serious local complication. It typically results from the ascending spread of *N. gonorrhoeae* from the endocervix to the uterus and fallopian tubes. Symptoms may include fever, nausea, vomiting, lower abdominal pain, low back pain, and abnormal vaginal bleeding. Early treatment of acute pelvic inflammatory disease and salpingitis with antibiotics is important in order to prevent further complications. Chronic salpingitis may cause scarring and blockage of the fallopian tubes, which may result in infertility or tubal pregnancy.

Rectal infections, which tend to be asymptomatic, are present in as many as 50% of women with gonorrhea and are common in homosexual men. Infection of the pharynx may occur in approximately 20% of heterosexual women and homosexual men. Pharyngeal gonorrhea may produce an exudative tonsillitis, but characteristically there are no symptoms and the infection usually clears spontaneously over several weeks.

Gonococcal conjunctivitis is rare in adults and is probably due to self-inoculation; if untreated the infection may lead to ulceration of the cornea and blindness. Ophthalmia neonatorum can be prevented by prophylactic use of 1% silver nitrate eyedrops. Additionally, infants born to infected mothers may also develop infections of the pharynx or respiratory tract as well as disseminated infections due to *N. gonorrhoeae*.

Treatment. An increasing proportion of infections are due to antibiotic-resistant strains of

N. gonorrhoeae. Chromosomally mediated resistance to multiple antibiotics as well as plasmid-mediated resistance to beta-lactam antibiotics and tetracycline occurs in strains from both developed and developing countries. Nevertheless, infections can be effectively treated with third-generation cephalosporins (for example, ceftriaxone) or fluoroquinolones (for example, ciprofloxacin or ofloxacin). *See* SEXUALLY TRANSMITTED DISEASES.

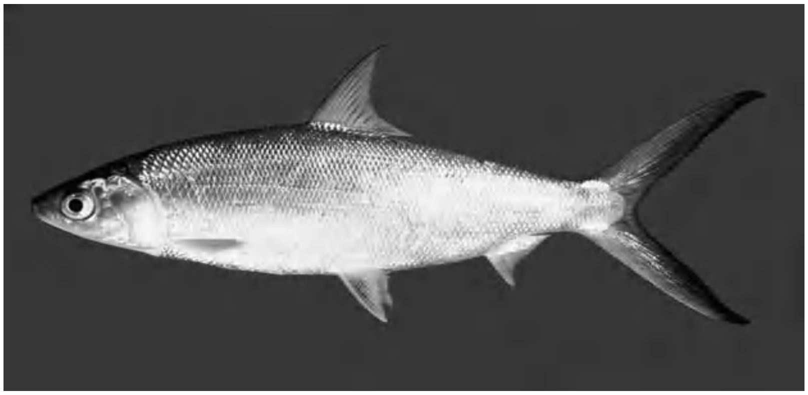
Sandra A. Larsen; Stephen A. Morse

Bibliography. K. K. Holmes et al. (eds.), *Sexually Transmitted Diseases*, 3d ed., 1998; S. A. Morse et al. (eds.), *Atlas of Sexually Transmitted Diseases*, 1990; B. B. Wentworth et al. (eds.), *Laboratory Methods for the Diagnosis of Sexually Transmitted Diseases*, 2d ed., 1991.

Gonorynchiformes

A primitive order of ostariophysan fishes in which the first three vertebrae are specialized and associated with one or more cephalic ribs, thus representing a primitive Weberian apparatus (a series of bony ossicles that form a chain connecting the swim bladder with the inner ear). The order is further characterized by having the following features: an epibranchial (=suprabranchial) organ present, which consists of lateral pouches in the posterior part of the branchial chamber behind the fourth epibranchials; a small mouth and toothless jaws; no postcleithra and orbitosphenoid bones; and small parietal bones. The order comprises four families, treated below. *See* OSTARIOPHYSI; OSTEICHTHYES.

Chanidae (milkfishes). One extant species (see **illustration**) occurs in marine and brackish waters of the tropical and subtropical Indian and Pacific oceans and occasionally inhabits freshwater. The body is fusiform and compressed, and the mouth is terminal. The species feeds on benthic algae and invertebrates. It attains a maximum length of 1.7 m (5.6 ft) but usually is 1 m (3.3 ft). Milkfishes are valued food in Southeast Asia, where they are cultured extensively in ponds.


***Chanos chanos*. (*Photo by Jack E. Randall with permission*)**

Gonorynchidae (beaked sandfishes). One genus and five species are found primarily in the Indo-Pacific region. These fishes possess an elongate body; an inferior mouth with a protrusible upper jaw; a sharply pointed snout with one barbel at the tip; and ctenoid scales on the head and body. Gonorynchidae is the most primitive teleost family with ctenoid scales. It attains a maximum length of 60 cm (24 in.). *See* TELEOSTEI.

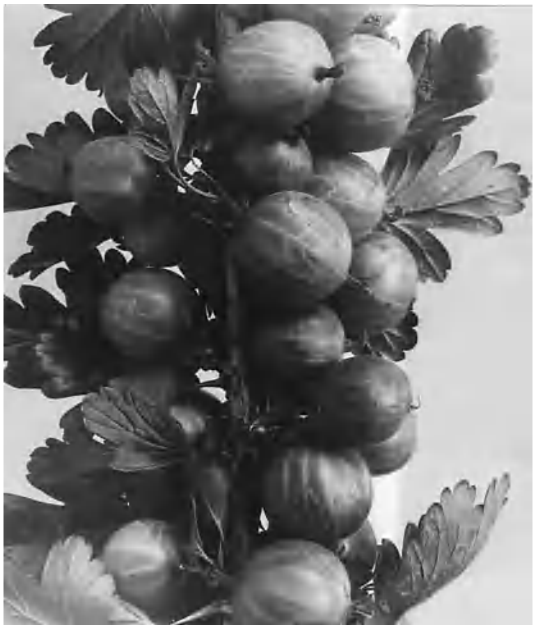
Kneriidae (knerias). Four genera and 30 species inhabit the freshwaters of tropical Africa and the Nile River. The mouth is inferior or subterminal; the upper jaw is protrusible. Some species have cycloid scales and a lateral line, whereas in others (smaller species) the body is naked and lacks a lateral line. Maximum length is about 15 cm (6 in.).

Phractolaemidae (snake mudheads). These fishes comprise a monotypic (pertaining to a taxon that contains only one immediately subordinate taxon) family in the freshwaters of tropical Africa. They have an elongate body, small superior mouth, and a lunglike swim bladder that allows them to survive in poorly oxygenated waters. Maximum length is about 16 cm (6.3 in.). *See* SWIM BLADDER. Herbert Boschung

Bibliography. R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, <www.fishbase.org>, version 05/2005; T. Grande and F. Poyato-Ariza, Phylogenetic relationships of fossil and Recent Gonorynchiform fishes (Teleostei: Ostariophysi), *Zool. J. Linn. Soc.*, 125(2):197–238, 1999; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Gooseberry

A small fruit represented by about six species of the genus *Ribes* of the plant order Rosales. The gooseberry is a thorny, spreading bush which grows to a


Gooseberry branch bearing leaves and fruits. (*USDA*)

height of about 3 ft (0.9 m) and produces red, yellow, or green berries (see **illus.**). The most desirable hardier types in the United States are of American parentage, or are hybrids between American and European species. Commercial culture has declined and is limited to a few states, notably Oregon, Michigan, and Washington, but gooseberries are found in home gardens throughout most of the United States except the far South and Southwest.

The fruit is very acid and only a few European varieties, when fully ripe, are suitable for eating fresh. The fruit may be canned or frozen for use in pies or as preserves.

For diseases of gooseberry *see* CURRANT; ROSALES.

J. Harold Clarke

Gopher

The name for the North American rodents of the family Geomyidae. There are 39 species in 8 genera; 11 species are found in the United States. The distinctive name pocket gopher is applied to these animals, since they have large, fur-lined cheek pouches which open outward on the side of the face. Two common species are the northern pocket gopher (*Thomomys talpoides*), a small burrowing species, most numerous in the western United States and Mexico, and the prairie pocket gopher (*Geomys bursarius*).

Pocket gopher (*Thomomys talpoides*) of North America.

These are small to medium-sized animals with a broad, blunt head, small eyes and ears, and a short, thick tail which has very little hair (see **illus.**) and a tactile sensory function. The gopher's body is stout and its legs are short. The forelimbs of the animal have large feet, terminating in long, strong claws. Gophers have 20 teeth with the dental formula I 1/1 C 0/0 Pm 1/1 M 3/3. The teeth lack roots and grow continuously. One litter, which consists of 2–6 young, is raised each year.

These animals are vegetarians and, although they do not hibernate, hoard food in their burrows for the winter. Apparently they derive sufficient moisture from their food, because they have not been observed to drink. Gophers are pugnacious and defend themselves and their territory against their natural enemies, the hawks, owls, weasels, and the bull snake, and even against dogs and humans. *See* RODENTIA.

Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, The Johns Hopkins University Press, 1999; J. O. Whitaker, Jr. and W. J. Hamilton, Jr., *Mammals of the Eastern United States*, Cornell University Press, Ithaca, 1998; D. E. Wilson and S. Ruff, *The Smithsonian Book of North American Mammals*, 1999.

Gorgonacea

An order of the cnidarian subclass Alcyonaria. The Gorgonacea are the horny corals which often form fanlike or featherlike colonies with branches spread radially or oppositely in one plane. They attach to objects by somewhat enlarged bases of tufts of stolons. They are more widely distributed than the Alcyonacea and extend from the littoral zone to some great depth.

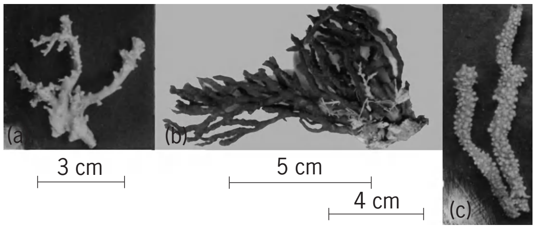


Fig. 1. Gorgonacea colonies. (*a*) Skeleton of *Corallium konojoi*. (*b*) *Melitodes* sp. (preserved specimen). (*c*) *Anthoplexaura dimorpha* (preserved specimen).

The gorgonians are characterized by the presence of an axial rod composed of horny matter or gorgonin. The order is divided into two suborders, the Scleraxonia and Holaxonia, according to the structure of the rod which has an outer cortex and an inner core, or medulla. In the Scleraxonia, which includes such genera as *Paragorgia, Corallium* (**Fig. 1***a*), and *Melitodes* (Fig. 1*b*),

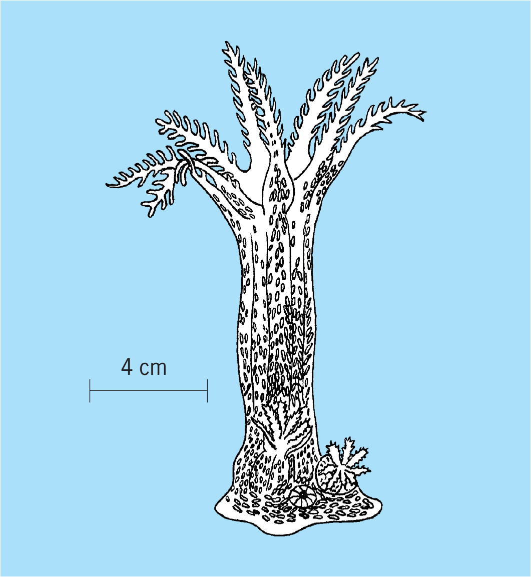


Fig. 2. Primary polyp of *Anthoplexaura dimorpha*.

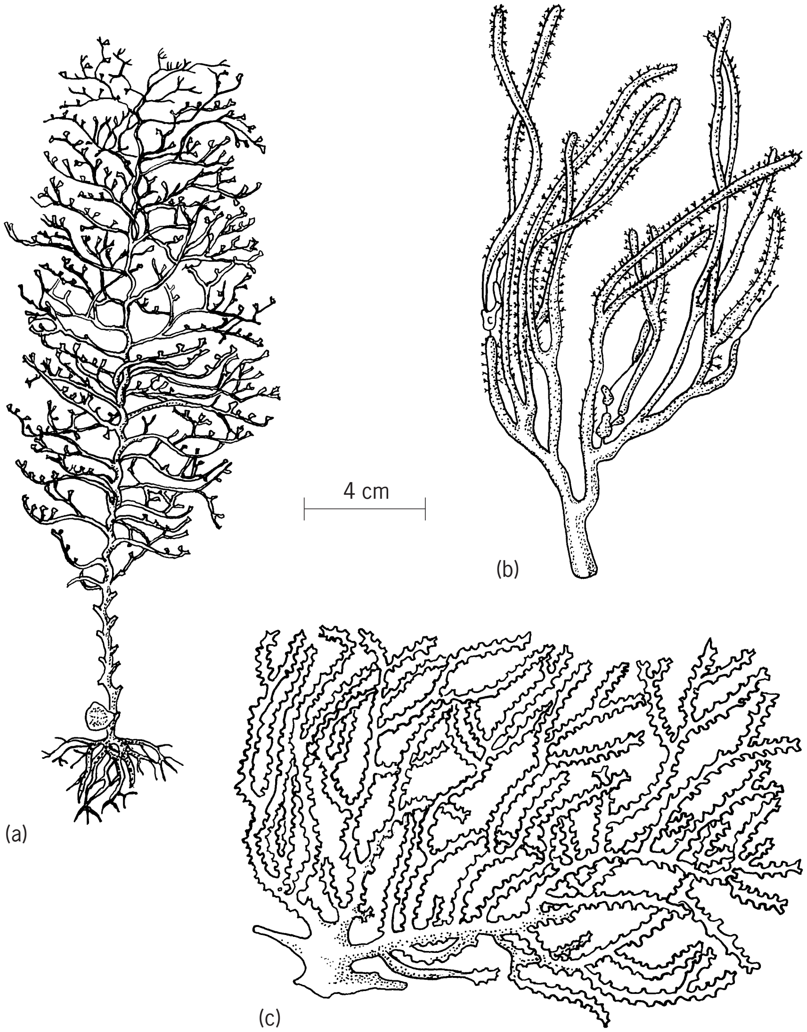


Fig. 3. Holaxonian gorgonians. (*a*) *Chrysogorgia flexilis* var. *africana*. (*b*) *Lophogorgia crista*. (*c*) *Paramuricea kükenthali*.

the axial skeleton contains calcareous spicules. In the Holaxonia, which are the typical gorgonians, a spiculeless skeleton is formed by the gorgonin and a small amount of calcareous matter. The latter group includes many genera: *Gorgonia* or the common sea fan, *Muricella, Anthoplexaura* (Fig. 1*c* and **Fig. 2**), *Chrysogorgia* (**Fig.** 3*a*), *Lophogorgia* (Fig. 3*b*), *Paramuricea* (Fig. 3*c*), and *Keratoisis*. The last has a peculiar segmented axis with alternating calcareous and horny portions.

Most gorgonians are monomorphic. The autozooid (a feeding individual possessing fully developed organs and exoskeleton) resembles that of the Alcyonacea in many respects. Corallidae and *Paragorgia* are dimorphic; the siphonozooid has no tentacle. *See* OCTOCORALLIA (ALCYONARIA).

Kenji Atoda

Gout

A hereditary disease of ancient lineage that is caused by a derangement in purine metabolism, particularly uric acid synthesis. Uric acid is a normal breakdown product of purine metabolism and is derived from either ingested foods or body tissues. The disease is characterized by increased levels of blood uric acid (hyperuricemia), inflammatory arthritis, tophaceous deposits of urate crystals, renal insufficiency, and urolithiasis. Hyperuricemia is caused by increased production or decreased renal excretion of uric acid. *See* ARTHRITIS.

Gouty arthritis may be acute or chronic, but the basic pattern is one of acute arthritis in one joint followed by periods which are completely free of symptoms. These acute attacks are violent; and they frequently occur in conjunction with excessive intake of rich food and drink. The most frequent site of the initial attack is the big toe. Acute arthritis is characterized by the presence of birefringent monosodium urate crystals in synovial fluid leukocytes. Chronic gouty arthritis may be associated with extensive bone erosions and joint destruction.

Primary gout occurs most frequently in middle-aged males and is passed on as a familial or hereditary trait which for some unknown reason does not appear as often in females. Secondary gout refers to the disease when it is associated with some underlying disorder causing an elevation in uric acid production, for example, myeloproliferative disorders.

The uric acid can be precipitated in soft tissues as localized deposits of chalky-white urate crystals called tophi. They occur in regions that are most susceptible to trauma, such as over bony prominences or adjacent to the small joints of hands and feet, in the cartilages of the ear, and in the kidney. Less commonly, they are found in eyelids, muscles, tendons, and heart valves. The mere presence of a tophus does not produce arthritis or pain, but there may be a surrounding inflammatory reaction.

Drugs that have been very helpful in treating gout include colchicine, which controls the acute arthritic attack; probenecid, which increases renal excretion of uric acid; and allopurinol, which inhibits xanthine oxidase, an enzyme required for uric acid formation. The clinical course is marked by long intervals of almost complete remission, but the disease flares up from time to time because of stress, infection, surgery, excessive alcohol intake, or unknown causes. Chronic gout causes functional impairment of the kidneys and greater incidence and earlier onset of atherosclerosis, but the general prognosis is good. *See* ARTERIOSCLEROSIS; PROTEIN METABOLISM; URIC ACID. Robert Searles

Governor

A device used to control the speed of a prime mover. A governor protects the prime mover from overspeeding and keeps the prime mover speed at or near the desired revolutions per minute. When a prime mover drives an alternator supplying electrical power at a given frequency, such as 60 Hz, a governor must be used to hold the prime mover at a speed that will yield this frequency. An unloaded engine will fly to pieces unless its speed is under governor control. *See* PRIME MOVER.

Speed control. A governor regulates the speed of a prime mover by properly varying the flow of energy to or from it. In the case of gas and steam turbines and internal combustion engines, the fuel furnishes the energy to the prime mover. For such applications, the governor usually controls the speed of the unit by regulating the rate at which fuel, and hence energy, is furnished to the prime mover. The governor controls the fuel flow so that the speed of the prime mover remains constant regardless of load and other disturbances, or changes in accordance with such operating conditions as changes in speed setting.

For a diesel engine the governor is connected to the rack which controls the amount of fuel injected. A governor on a gas or gasoline engine is attached to the engine throttle. A steam turbine governor strokes the steam valve or valves which regulate the steam flow to the turbine. *See* DIESEL ENGINE; INTERNAL COMBUSTION ENGINE; STEAM TURBINE.

The output mechanism of a gas turbine governor is connected to the fuel valve with the stroke normally limited in each direction by the allowable combustion chamber temperature and other factors. If the fuel rate is too low, the fire will go out. Compressor stall must be avoided. To give rapid control, combustion chamber temperature is often computed in the governor by measuring other variables and applying the laws of thermodynamics. *See* GAS TURBINE.

A hydraulic turbine governor regulates the flow of water to the turbine by varying the openings of gates or other components. *See* HYDRAULIC TURBINE.

An aircraft propeller governor varies the pitch of the propeller to keep constant the speed of the engine attached to the propeller. This type of governor varies the load on the engine and thus controls the speed by regulating the energy flow from the prime mover. *See* PROPELLER (AIRCRAFT).

Ballhead governor. The speed of a prime mover is usually measured by a ballhead that contains flyweights driven at a speed proportional to the speed of the prime mover. The force from the flyweights is balanced, at least in part, by the force of compression of a speeder spring (**Fig. 1**). The upper end of this spring is positioned according to the speed setting of the governor.

The ballhead toes press against one end of a plunger. In the simplest version of a governor, the plunger position is a function of the engine speed as a result of the balance between the centrifugal and spring forces. The plunger is directly connected to the throttle or other energy controlling means of the prime mover. Because the governor output power is drawn directly from the speed measuring means, the output power and the precision of such a governor are severely limited.

Response. In automatic control theory, the input to the governor is taken to be the difference between the speed setting (reference) and the actual prime mover speed. This difference is the speed error. For the simplest governor, the position of the plunger is the output. The action of the governor is proportional since its output is proportional to its input. In equilibrium, where the engine is running in steady state, the speed of the prime mover depends on both the load and the speed setting. With the mechanical governor and a fixed speed setting, the equilibrium speed decreases as the load increases. A governor with this property is referred to as a droop governor,

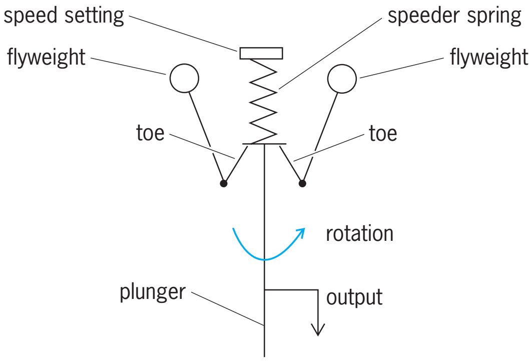


Fig. 1. Ballhead governor.

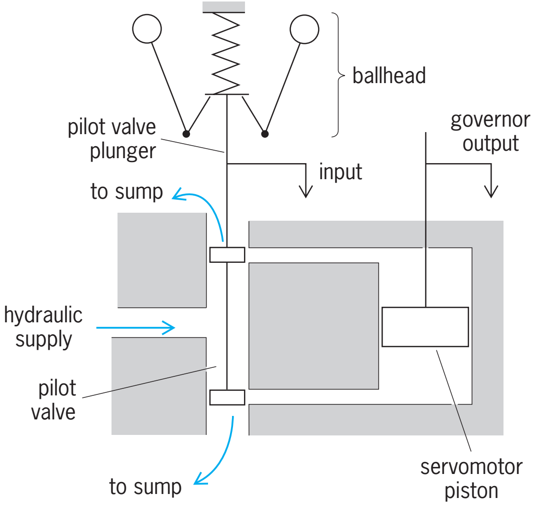


Fig. 2. Isochronous governor.

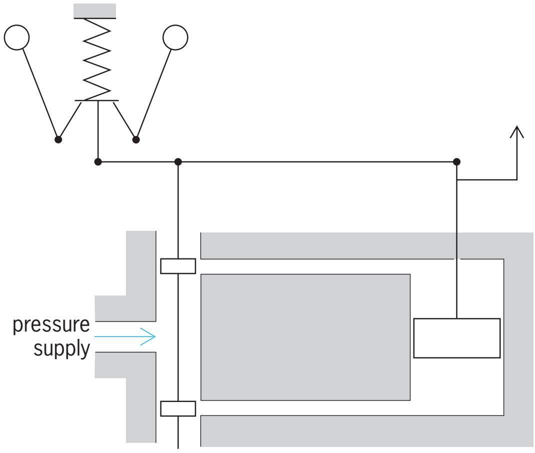


Fig. 3. Hydraulic governor.

or a governor operating on droop. The governor–prime mover unit is then said to be running on droop.

To increase the power output of a governor, a hydraulic amplifier is often employed. A governor that keeps the speed of a prime mover constant is said to be isochronous. In a simple isochronous governor, the ballhead senses the speed and strokes a pilot valve plunger that regulates the flow of fluid to a servomotor (**Fig. 2**). Normally, the fluid is oil. This governor is intended to bring the prime mover speed back to the speed setting after any change in load. *See* SERVOMECHANISM.

A hydraulic droop governor is obtained from the simple isochronous governor by introducing feedback from the governor output to the pilot valve (**Fig. 3**). This governor behaves like a simple mechanical governor except that a smaller ballhead is generally employed, greatly improving the precision of the governor by reducing hysteresis, dead band, and friction in the ballhead. The power output of the governor is much larger so that the effect of the load on governor performance incurred in moving the throttle or equivalent mechanism, is considerably diminished. A hydraulic governor may be sensitive to speed changes of as little as 1/1000 of 1%. For normal disturbances, prime mover speed error may be kept to 0.1% or better.

Use of dashpot. The performance of the simple isochronous governor is often greatly improved by the introduction of a dashpot in the feedback path from the output to the ballhead. If there is little damping in the prime mover, instability often occurs when the simple isochronous governor is used, whereas this instability is removed when the dashpot is incorporated. A system is stable when for each disturbance that dies out, the response of the system settles to an equilibrium condition. When instability occurs, unless some protection means is provided, the prime mover speed oscillates continuously or increases indefinitely until the unit breaks up.

When a dashpot is incorporated into a governor (**Fig. 4**), the governor output becomes a function of the speed error and the integral of this error. Such a governor action is referred to as a proportional plus integral controller. The velocity of the servomotor then depends on prime mover speed and prime mover acceleration. The time lag in the dashpot makes the governor sensitive to prime mover acceleration. As this lag is increased, the response of the governor to prime mover acceleration tends to increase. This increase in lag is accomplished by moving the dashpot needle valve, which controls the orifice area, toward the closed position.

Instead of mechanical feedback from the governor, force feedback is generally preferred (**Fig. 5**). An isochronous dashpot governor is turned into a droop governor by adding direct mechanical feedback from the servomotor to the ballhead.

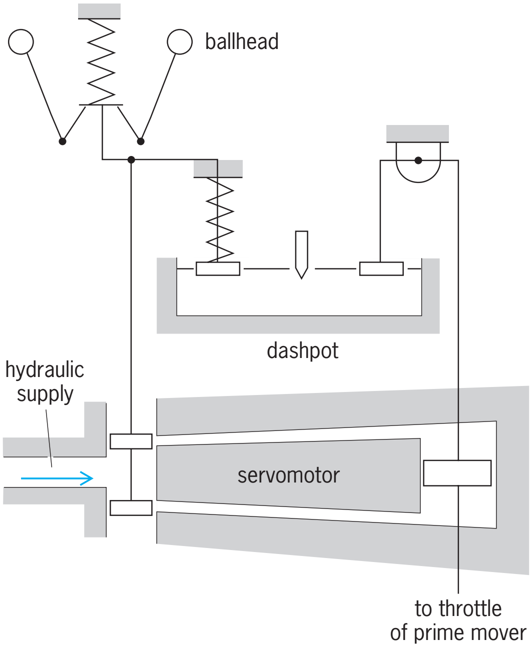


Fig. 4. Isochronous dashpot governor.

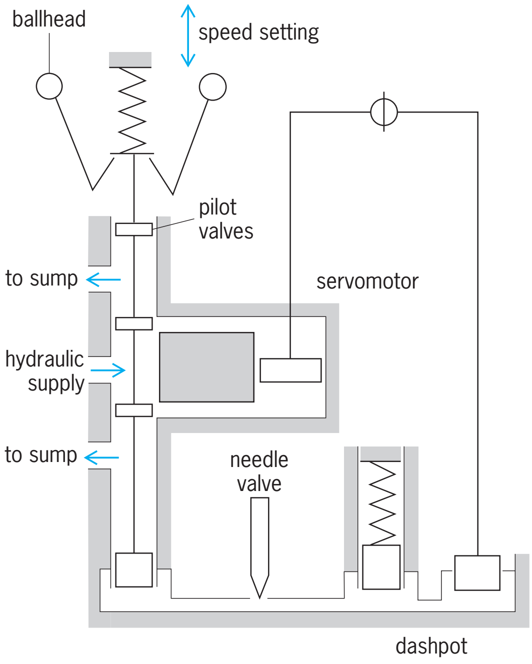


Fig. 5. Governor with force feedback.

Use of flywheel. Acceleration governors are sometimes used in place of governors with dashpots. In such governors a flywheel is employed instead of a dashpot. The prime mover drives the flywheel through a spring. This combination yields a motion proportional, except for a time lag, to the acceleration of the prime mover. *See* FLYWHEEL.

Applications. Governors for large hydraulic turbines require a second stage of hydraulic amplification. The governor servomotor piston is connected to the plunger of a relay valve that regulates the flow of oil to the turbine servomotors. The governor servomotor is then termed the controller. Turbine servomotors often require 10 hp (7500 W) or more.

The speed setting of a governor is often adjusted by an electric, hydraulic, or pneumatic motor as a function of auxiliary variables. Thus in an electrical power system the deviation of the frequency of the system from the desired value is used to position the governor speed setting so as to bring the frequency to the right value. Adjustment of governor speed settings is required to keep electric clocks on time. Oil and gas pipeline governors control pressure rather than speed. The pressure in the line is measured and the speed setting of the governor is adjusted accordingly. This affects the speed of the prime mover, which drives a compressor, raising or lowering the pressure.

Computer control. Traditional mechanical governor systems are being displaced by computer speed control systems. The desired controls are programmed directly into a computer, which reads a speed signal and provides for proportional/integral/differential (PID) control of speed by varying energy input to the prime mover. The computer-controlled governor can be programmed to provide functions similar to traditional mechanical governing systems but without the mechanical failures inherent in such systems. *See* COMPUTER; DIGITAL COMPUTER; DIGITAL CONTROL.

Paralleled prime movers. Prime movers may be paralleled to supply power to the same load. In an electrical power system, prime movers drive alternators electrically connected to the system. At most, one of the governors of paralleled prime movers can be isochronous; the rest must be on droop. The prime movers may all be on droop.

When two or more identical prime mover-generator units are paralleled, and one is controlled by an isochronous governor while the others are on droop, electrical coupling will force all units to run at the same steady-state speed. The alternators then supply electrical power to the electric grid at the same frequency.

Speed control by governor droop is an integral part of frequency control on an electric power system. Governors provide frequency control, and control how increases or decreases in load on the generators (generation pick-up) is shared among the remaining generators in the event of a loss of a generator (or how generation reduction is shared among generators in the event of a loss of load). System coordination of governor droop settings provides for equitable division of replacement generation until system automatic generation control computers assign a new generation dispatch. Plant control computers also participate in controlling the output of individual generators in a plant. In the North American power system, governor droop settings range between 4 and 7%. The droop setting means that for the stated percent frequency deviation the governor calls for the generator's maximum power. For example, on a 60-Hz system, if the droop setting is 5%, during a 5% (3-Hz) frequency deviation governors call for maximum power from the generators. Governors do not return the power system to 60 Hz; automatic generation control computers perform this function. *See* ELECTRIC POWER SYSTEMS.

Aircraft engines are synchronized and synchrophased by using one engine as the master, and adjusting the speed settings of the governors on the other engines to make them follow the master.

Design. Proper design of a governor involves making the governor characteristics fit those of the prime mover and load so as to give acceptable overall performance. Desirable and undesirable nonlinearities occur in governors to complicate design calculations. Thus centrifugal force varies as the square of speed. To compensate for this, nonlinear speeder springs are often employed. Bypass is often used in dashpots to limit the range of the dashpot output.

The output of a generator driven by the prime mover is sometimes used to obtain a voltage proportional to prime mover speed. This voltage, properly filtered, is fed to an electronic, magnetic, or other circuit, and eventually to a transducer to move the pilot valve. A gear with permanent magnet teeth is often employed with a pickup coil, where the gear is driven at a speed proportional to that of the prime mover. The output pressure of a hydraulic pump is sometimes employed as a measure of prime mover speed. Other speed-measuring means are utilized.

Limit or topping governors are often employed to meet fail-safe specifications. The limit governor is usually a mechanical governor that takes over control from the main governor to shut the prime mover down if its speed reaches a fixed overspeed above its rated speed. The flyweight force increases with an increase in the distance of the flyweights from the axis of rotation of the ballhead; this flyweight force opposes the speeder spring force. The flyweight force increases as the square of prime mover speed. When it is sufficiently large, the governor gain becomes infinite, so that a small input to the governor causes an unlimited output. This causes a snap action of the governor at the overspeed for which the governor is set.

In the interests of economy the four-way valve servomotor (Fig. 2) is often replaced by a two-way valve-differential servomotor (**Fig. 6**). The area on one side of the servomotor is half the area on the other side. The supply pressure on the small area is sometimes replaced by a spring.

A load is termed isolated if the power supplying this load comes from one prime mover only. In the design of a governor for a prime mover supplying power to an isolated load, a prime mover differential equation is obtained relating the input to the prime mover, such as throttle position, to the speed of the prime mover, which is taken to be its output. This equation is obtained by mathematically balancing the torques on the prime mover shaft. The equation also involves a delay between a change in the input to the prime mover and the resulting change in the driving output torque. Part of the delay is a dead time. This dead time is normally 0.01–0.5 s. The rest of the delay is usually in the order of 0.1 s.

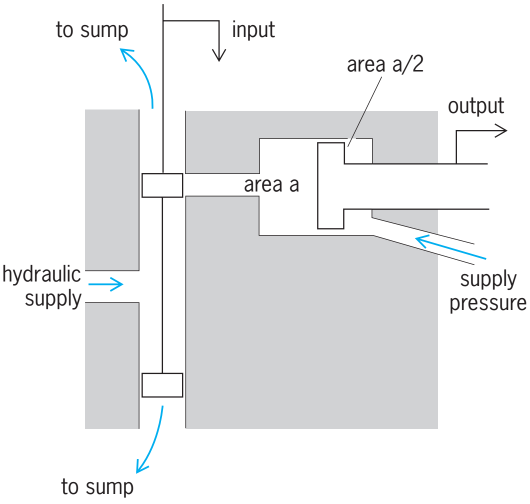


Fig. 6. Two-way valve-differential servomotor.

The nature of a prime mover may introduce other terms in the equation. For example, with hydraulic turbines the equation is complicated because of water hammer; a sudden closing of the gates to decrease the speed of the turbine causes the turbine to speed up initially, resulting in a correction opposite to that desired. As the gates are moved, pressure waves travel up and down the hydraulic conduit from the source of the water to the discharge from the turbine. These waves tend to destabilize the unit. *See* WATER HAMMER.

The equation of the prime mover is combined with the equation of the governor to yield an equation for overall system response from which the governor response is determined. When full load on a diesel engine is rejected, an overspeed of 3% is often considered acceptable with the response dying out in 1 s or less. For a hydraulic turbine this overspeed may be 30–40% and the response may endure as much as 10–20 s. Performance for other prime movers under governor control falls between these extremes. Rufus Oldenburger; Donald Davies

Bibliography. P. M. Anderson and A. A. Fouad, *Power System Control and Stability*, 2d ed., 2003; E. A. Avallone, T. Baumeister III, and A. Sadegh (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 11th ed., 2007; L. L. Grigsby (ed.), *The Electric Power Engineering Handbook*, 2001; H. A. Rothbart (ed.), *Mechanical Design and Systems Handbook*, 2d ed., 1985.

Graben

A block of the Earth's crust, generally with a length much greater than its width, that has been dropped relative to the blocks on either side (see **illus.**). The size of a graben may vary; it may be only a few inches long or it may be hundreds of miles in length. The faults that separate a graben from the adjacent rocks are inclined from 50–70° toward the downthrown block and have displacements ranging from inches to thousands of feet. The direction of slip on these indicates that they are gravity faults. Graben are found in regions where the crust has undergone extension. They may form in the crests of anticlines or domes, or may be related to broad regional warpings. *See* EARTH CRUST; FAULT AND FAULT STRUCTURES; HORST; RIFT VALLEY. Philip H. Osberg

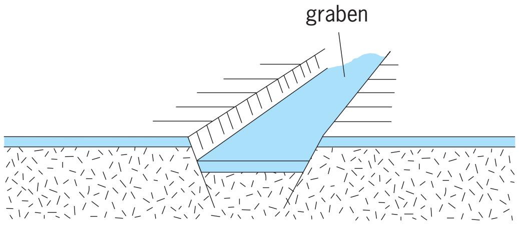


Diagram of simple graben. (*After A. K. Lobeck, Geomorphology*, McGraw-Hill, *1939*)

Gradient of a scalar

The result of the application of the distributive vector differential operator ∇, $\nabla = \mathfrak{i}\,\partial/\partial x + \mathfrak{j}\,\partial/\partial y + \mathfrak{k}\,\partial/\partial z$, to a differentiable scalar function $S(x, y, z)$; thus $\nabla S = \mathfrak{i}\,\partial S/\partial x + \mathfrak{j}\,\partial S/\partial y + \mathfrak{k}\,\partial S/\partial z$. The letters $\mathfrak{i}$, $\mathfrak{j}$, $\mathfrak{k}$ are symbols for the base vectors associated with x, y, z. The gradient of S is also denoted by grad S, while the symbol ∇ is usually called del and less frequently nabla.

The key properties of ∇S are implicit in its relation to the directional derivative of S, which is essentially the rate of change of S with respect to distance in a specified direction. To illustrate, if $S(x, y, z)$ gives the temperature at each point of a room, then from a given point of observation S may increase in certain directions (say toward a heat source) and decrease in others. Specifically, let P be a given point, C a curve given in terms of its arc length s by $\mathfrak{r} = x(s)\mathfrak{i} + y(s)\mathfrak{j} + z(s)\mathfrak{k}$ and further, let C pass through the point P and satisfy the requirement $(d\mathfrak{r}/ds|$ at $P) = \tau$ (a given prescribed unit vector). Along C, $x = x(s)$, $y = y(s)$, $z = z(s)$ and $S = S[x(s), y(s), z(s)]$ and if the functions are of class C'; then the equation below holds.

$$\left.\frac{dS}{ds}\right|_{P,C} = \left.\frac{\partial S}{\partial x}\frac{dx}{ds} + \frac{\partial S}{\partial y}\frac{dy}{ds} + \frac{\partial S}{\partial z}\frac{dz}{ds}\right|_{P}$$
$$= \nabla S|_P \cdot \tau = \ |\nabla S|_P|\cos\,(\nabla S|_P, \tau)|$$

Here P, C indicates that the derivative is taken at the point P and for the curve C. This equation shows that $dS/ds|_{P,C}$ depends on S, P, and τ only; hence, the alternative notation $dS/ds|_{P,\tau}$ is more appropriate. The quantity $dS/ds|_{P,\tau}$ is called the directional derivative.

Evidently, if (i) S and P are fixed, (ii) τ is variable in direction, and (iii) $\nabla S|_P \neq 0$, then $dS/ds|_{P,\tau}$ can vary with the direction of τ and will take on its maximum value when $\cos\,(\nabla S|_P, \tau) = 1$. Thus, $\nabla S|_P$ points in the direction of the greatest rate of increase of S at P and $|\nabla S|_P|$ is this greatest rate of increase. Also, if $S(x, y, z) = S(a, b, c)$, (a, b, c fixed) defines a surface, then by confining C to the surface it may be shown that if $\nabla S|_P \neq 0$, then $\nabla S|_P$ is perpendicular to the surface. *See* CALCULUS OF VECTORS.

Homer V. Craig

Gradient wind

A hypothetical wind based upon the assumption that the sum of the horizontal components of the Coriolis acceleration and the atmospheric pressure gradient force per unit mass is equivalent to a wind acceleration which is normal to the direction of the wind itself (centripetal acceleration), with no viscous forces acting. The direction of the gradient wind is the same as that of the geostrophic wind. Its speed is determined by the equation below, where V_{geo} is

$$V_{\text{grad}} = \frac{V_{\text{geo}}}{\dfrac{1}{2} + \sqrt{\dfrac{1}{4} + \dfrac{V_{\text{geo}}}{2R\Omega\sin\phi}}}$$

the speed of the geostrophic wind, Ω is the angular speed of rotation of the Earth about its axis, ϕ is the latitude, and R is the radius of the curvature of the air trajectory, considered positive when the trajectory is curved in the cyclonic sense (center of curvature on low-pressure side) and negative when the curvature is anticyclonic. The gradient wind speed is less than the geostrophic speed when the air moves in a cyclonically curved path and greater when the air moves in an anticyclonically curved path. The gradient wind is a good approximation of the actual wind and is often superior to the geostrophic wind, particularly when the flow is strongly curved in the cyclonic sense. *See* CORIOLIS ACCELERATION; GEOSTROPHIC WIND; STORM; WIND.

Frederick Sanders

Grain boundaries

The internal interfaces that separate neighboring misoriented single crystals in a polycrystalline solid. Most solids such as metals, ceramics, and semiconductors have a crystalline structure, which means that they are made of atoms which are arranged in a three-dimensional periodic manner within the constituent crystals. Most engineering materials are polycrystalline in nature in that they are made of many small single crystals which are misoriented with respect to each other and meet at internal interfaces which are called grain boundaries. These interfaces, which are frequently planar, have a two-dimensionally periodic atomic structure. A polycrystalline cube 1 cm on edge, with grains 0.0001 cm in diameter, would contain 10^{12} crystals with a grain boundary area of several square meters. Thus, grain boundaries play an important role in controlling the electrical and mechanical properties of the polycrystalline solid. It is believed that the properties are influenced by the detailed atomic structure of the grain boundaries, as well as by the defects that are present, such as dislocations and ledges. Grain boundaries generally have very different atomic configurations and local atomic densities than those of the perfect crystal, and so they act as sinks for impurity atoms which tend to segregate to interfaces. *See* CRYSTAL DEFECTS; CRYSTAL STRUCTURE.

Structure. It was first believed that grain boundaries were made of an amorphous cement which held together the polycrystalline solid. With the advent of modern experimental probes of the atomic and defect structure of solids, such as electron microscopy and x-ray diffraction, it was determined that the grain boundary structure is frequently periodic in two dimensions. The geometry of a grain boundary is described by the rotation axis and angle, θ, that relate the orientations of the two crystals neighboring the interface, and the interface plane (or plane of contact) between the two crystals. Grain boundaries are typically divided into categories characterized by the magnitude of θ and the orientation of the rotation axis with respect to the interface plane. When θ is less than (arbitrarily) 15°, the boundary is called small-angle, and when θ is greater

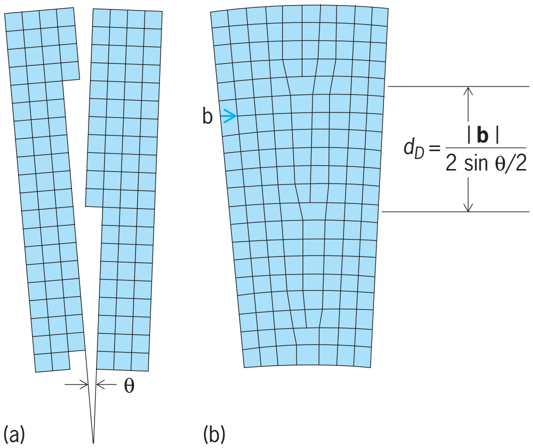


Fig. 1. The formation of a small-angle tilt grain boundary. (*a*) Unrelaxed configuration. (*b*) Relaxed configuration. (*After W. T. Read, Jr., Dislocations in crystals, McGraw-Hill, 1953*)

than 15°, the boundary is large-angle. *See* ELECTRON MICROSCOPE; X-RAY DIFFRACTION.

Small-angle boundaries. When the rotation axis is in the plane of the interface, the boundary is called a tilt boundary. **Figure 1** is a schematic diagram showing the formation of a small-angle tilt boundary in a simple cubic material. Simple cubic means that the atoms are located at the corners of a cube-shaped unit cell, which is translated in three dimensions to build up the crystal. The two crystals have been cut along a vertical plane, the interface plane. It is seen in Fig. 1*a* that the surface of the boundary-plane-to-be in each crystal is stepped in a periodic fashion on the atomic scale. When the two misoriented crystals are brought together and the atoms in the vicinity of the interface are allowed to move (relax) into lower energy configurations, then the result is as shown in Fig. 1*b*. It is seen that the boundary consists of a periodic array of edge dislocations with Burgers vector **b** and spacing d_D, which is given by Eq. (1). For small angles this equation becomes Eq. (2). Between the

$$d_D = \frac{|\mathbf{b}|}{2 \sin \theta/2} \qquad (1)$$

$$d_D = \frac{|\mathbf{b}|}{\theta} \qquad (2)$$

dislocations are regions of good cubic material. Because of the periodic nature of the boundary structure, the strain field associated with the boundary falls off exponentially (rapidly) along the direction normal to the interface. Thus the boundary region can be thought of as a thin layer with a thickness of $2d_D$ (or less), inside of which is disturbed material.

When the rotation axis is normal to the interface plane, then a twist boundary is present. **Figure 2** is a schematic diagram showing the formation of a small-angle twist boundary. In the unrelaxed configuration (Fig. 2*a*), it is possible to pick out regions of good cubic stacking (or good matching) and regions of bad cubic stacking (or bad matching). If the atoms at the interface are allowed to relax (Fig. 2*b*), then the regions of good match tend to expand, and all the mismatch is squeezed into narrow bands, which consist of a square array of screw dislocations with Burgers vector **b**, and spacing d_D given by the same expression obtained for tilt boundaries. **Figure 3** shows a transmission electron micrograph of a small-angle twist boundary in silicon, which contains a square array of screw dislocations, as predicted in Fig. 2*b*.

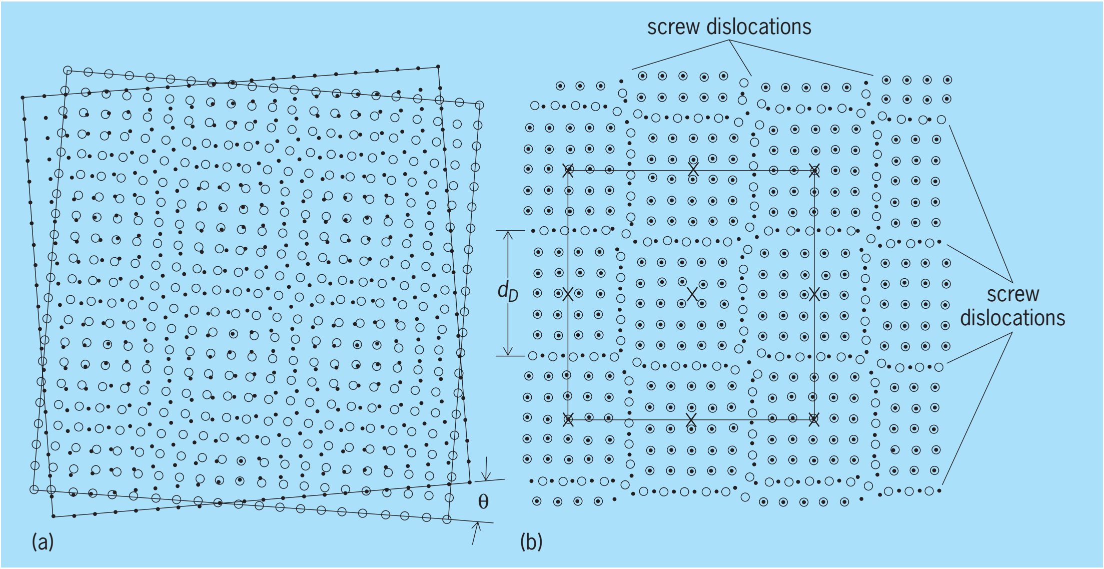


Fig. 2. Atomic configurations in the planes just above and below the boundary plane in a small-angle ⟨001⟩ twist grain boundary (*a*) Unrelaxed configuration. (*b*) Relaxed configuration, illustrating the formation of the screw dislocation network. The periodicity is that of the coincidence site lattice, the unit cell of which is indicated by the square. The crosses indicate the location of O-lattice points. (*After T. R. Schober and R. W. Balluffi, Quantitative observation of misfit dislocation arrays in low and high angle twist boundaries, Phil. Mag., 21:109–148, 1970*)

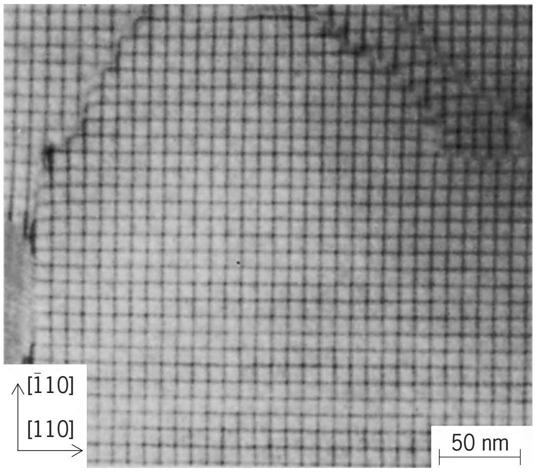


Fig. 3. **Transmission electron micrograph, showing a small-angle ⟨001⟩ twist boundary in silicon. (*From H. Föll and D. Ast, TEM observations on grain boundaries in sintered silicon, Phil. Mag. A, 39:589, 1979*)**

Large-angle boundaries. When the structures of the small-angle boundaries are extended to the large-angle regime, the expected dislocation spacing for a $\theta = 30^\circ$ boundary is $\approx 2\mathbf{b}$, which for most metals is 0.6 nanometer or less. It is unrealistic to postulate the existence of discrete dislocations at such small spacings, since their cores would overlap. A physically appealing approach to the structure of large-angle boundaries is to initially ignore their dislocation character and concentrate on their two-dimensionally periodic nature. This is done by using the concept of the coincidence site lattice (CSL).

The coincidence site lattice and its relation to grain boundary structure can be understood by using the structure of the small-angle twist boundary in Fig. 2*b*. At the corners of the outlined square, atoms in the two crystals are in exact match (or good coincidence); thus, these points define a simplified version of the coincidence site lattice. The plane lattice made up of the points determined by the coincident atoms in the two crystals gives the two-dimensional coincidence site lattice, and the number of such points expressed as a fraction of the lattice points in one plane is defined as $1/\Sigma$. The two-dimensional coincidence site lattice in the boundary plane shown in Fig. 2*b* describes the actual periodicity of the grain boundary structure, even when atoms move from their perfect crystal positions to take on a lower energy configuration.

There are additional locations, besides the points of the coincidence site lattice, where the two crystals are in good coincidence (or matching). These locations lie on a sublattice of the coincidence site lattice which is called the O-lattice, and it is easily seen that the two crystals can rotate with respect to each other at the O-lattice points (marked by crosses in Fig. 2*b*) to achieve good matching on a local scale. In the small-angle case, the O-lattice points are separated by bands of mismatch (with spacing d_D), which are screw dislocations possessing Burgers vectors of the perfect crystal (and hence are called primary dislocations).

For large misorientations and special angles θ_{CSL}, the dimensions of the unit cell of the coincidence site lattice can be relatively small. It is believed that boundaries with small unit cells and low values of Σ (see **table**) have low energy per unit area. **Figure 4*a*** shows four unit cells of a $\Sigma = 13$ twist grain boundary in gold where the atom positions have been determined by x-ray diffraction techniques to an uncertainty of less than 0.01 nm. Careful examination of the structure shows that it can be modeled by using polyhedral arrangements of atoms, as illustrated for inclined octahedra in Fig. 4*a*. The three-dimensional atomic configuration determined by x-ray diffraction is shown in Fig. 4*b*. Arrays of octahedra, tetrahedra, and archimedean antiprisms are interwoven to build up the grain boundary structure. The atomic density across the actual interface plane is approximately 8% less than in the bulk material.

In the discussion of the structure of small-angle boundaries, it has been shown that local regions of good matching occur which are separated by dislocations possessing Burgers vectors of the perfect crystal. In a similar manner, it is expected that in boundaries with misorientations that deviate from the special values of θ_{CSL} there will be local rotations to achieve local regions of low-Σ structure, separated by grain boundary dislocations. These dislocations have Burgers vectors which are not present in the perfect crystal and, therefore, are called secondary dislocations, to distinguish them from the primary dislocations in Figs. 1*b*, 2*b*, and 3. The spacing d_x of the secondary dislocations with Burgers vectors $\mathbf{b_s}$ is given by Eq. (3). Figure 2*b* can also represent the

$$d_s = \frac{|\mathbf{b_s}|}{|\theta - \theta_{\mathrm{CSL}}|} \qquad (3)$$

structure of a large-angle twist boundary slightly deviated from θ_{CSL} with the regions of good matching replaced by regions of low Σ and primary dislocations replaced by secondary dislocations.

Boundaries with inclined rotation axes. The structures of tilt and twist boundaries are the easiest to visualize and thus have been the subject of most research. For the more general type of boundary, where the axis describing the misorientation between the two crystals is inclined to the interface plane, the boundary

Examples of special misorientation angles*

Misorientation angle (θ_{CSL})	Σ
22.62°	13
28.07°	17
36.87°	5

*For grain boundaries with a ⟨001⟩ misorientation axis, in the cubic system.

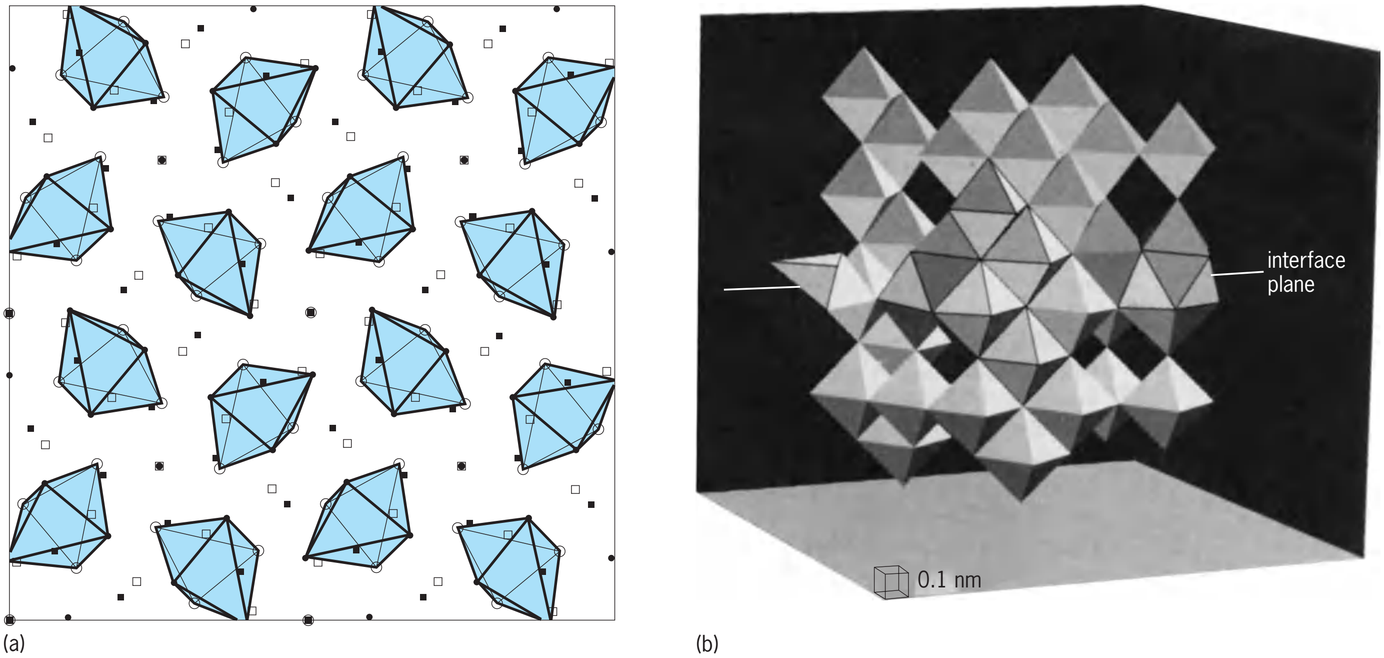


Fig. 4. Structure of the $\Sigma = 13$ ($\theta = 22.62^\circ$) $\langle 001 \rangle$ twist boundary in gold. (*a*) Four unit cells, showing the polyhedral arrangements of atoms, in particular, inclined octahedra. (*b*) Three-dimensional perspective view of the different polyhedral arrangements of atoms that form the $\Sigma = 13$ boundary. (*After M. R. Fitzsimmons and S. L. Sass, The atomic structure of the $\Sigma = 13$ ($\theta = 22.6^\circ$) $\langle 001 \rangle$ twist boundary in gold determined using quantitative x-ray diffraction techniques, Acta Metallurg., 37:1009–1022, 1989*)

structure can be thought of as having tilt and twist components, and so, for example, in the small-angle case, will be made up of a mixture of edge and screw dislocations.

Influence on material properties. Because of the large differences in atomic structure and density between the grain boundary region and the bulk solid, the properties of the boundary are also quite different from those of the bulk, and have a strong influence on the bulk properties of the polycrystalline solid.

Mechanical properties. The mechanical behavior of a solid, that is, its response to an applied stress, often involves the movement of dislocations in the bulk, and the presence of boundaries impedes their motion since, in order for deformation to be transmitted from one crystal to its neighbor, the dislocations must transfer across the boundary and change direction. The detailed structure at the interface influences the ease or difficulty with which the dislocations accomplish this change in direction.

Since grain boundaries in engineering materials are not in a high-purity environment, the presence of impurities dissolved in the solid may have a strong influence on their behavior. The presence of one-half of a monolayer of impurity atoms, such as sulfur or antimony in iron, at the grain boundary, can have a drastic effect on mechanical properties, making iron, which is ductile in the high-purity state, extremely brittle, so that it fractures along grain boundaries. The segregation of the impurity atoms to the boundaries has been well documented by the use of Auger electron spectroscopy, and studies have led to the suggestion that the change in properties may be related to a change in the dislocation structure of the grain boundary induced by the presence of these impurities. *See* METAL, MECHANICAL PROPERTIES OF; PLASTIC DEFORMATION OF METAL.

Diffusion. The movement of atoms in and through solids is important for carrying out a variety of processes in the solid state. Among these processes are the motion of interfaces, which occurs during a variety of deformation and shaping treatments, and processes that occur during phase transformations, which are frequently used to enhance mechanical properties. Again, because the local atomic structure and density of the boundary region are different from that of the bulk, the movement of atoms will also be much different. Atoms move, or diffuse, much faster down grain boundaries than in the bulk. This enhanced motion of atoms down boundaries is called grain boundary diffusion, and is sometimes faster by as much as a factor of 10^6 over the bulk. Atoms also move through the crystal interior in a process called lattice diffusion. The amount of mass transport through a solid is the product of the velocity of individual atoms and the area available for the atoms to move through. At high temperatures, lattice diffusion tends to dominate since the cross-sectional area of the interior of the crystal is huge compared to the cross-sectional area of a grain boundary. At low temperatures, however, where lattice diffusion is slow, grain boundary diffusion tends to dominate. All these effects are of great importance to the electronics industry because, as part of the manufacture of electrical devices, thin films containing many grain boundaries which serve as short-circuits for diffusion must be deposited. *See* DIFFUSION.

Electrical properties. Since modern electronic devices are fabricated from semiconductors, which may be polycrystalline, the presence of grain boundaries and their effect on electrical properties is of great technological interest. In a semiconductor such as silicon, the local change in structure at the interface gives rise to disruption of the normal crystal bonding, or sharing of valence electrons. One consequence can be the charging of the grain boundaries, which produces a barrier to current flowing across them and thus raises the overall resistance of the sample. This polycrystalline effect is exploited in devices such as zinc oxide varistors, which are used as voltage regulators and surge protectors. *See* SURGE SUPPRESSOR; VARISTOR.

Other devices, such as inexpensive solar cells, are also made of polycrystalline semiconductors. Grain boundaries can act as recombination centers in these cells, thus severely limiting their overall efficiency. The disruption of crystal bonding can encourage the segregation of impurities to the grain boundaries during crystal growth and subsequent processing. This effect (known as gettering) can be beneficial to solar cells because the crystalline areas are then left cleaner, so that the devices can be more efficient. *See* PHOTOVOLTAIC CELL; PHOTOVOLTAIC EFFECT; SOLAR CELL.

Stephen L. Sass

Bibliography. M. R. Fitzsimmons and S. L. Sass, The atomic structure of the $\Sigma = 13\,(\theta = 22.6^\circ)\ \langle 001\rangle$ twist boundary in gold determined using quantitative x-ray diffraction techniques, *Acta Metallurg.*, 37:1009–1020, 1989; P. E. J. Flewitt and R. K. Wild, *Grain Boundaries: Their Microstructure and Chemistry*, 2001; G. Gottstein and L. S. Shvindlerman, *Grain Boundary Migration in Metals: Thermodynamics, Kinetics, Applications*, 1998; A. Kelly, G. W. Groves, and P. Kidd, *Crystallography and Crystal Defects*, rev. ed., 2000; R. Raj and S. L. Sass (eds.), Interface Science and Engineering '87, *J. Phys.*, vol. 49, colloque C5, 1989; W. T. Read, Jr., *Dislocations in Crystals*, 1953; M. Ruhle et al. (eds.), International Conference on the Structure and Properties of Internal Interfaces, *J. Phys.*, vol. 46, colloque C4, 1985; T. Schober and R. W. Balluffi, Quantitative observation of misfit dislocation arrays in low and high angle twist boundaries, *Phil. Mag.*, 21:109–148, 1970; H. Zhang, *Grain Boundary Migration in Metals: Molecular Dynamics Simulations*, 2006.

Grain crops

Crop plants that belong to the grass family (Gramineae), generally grown for their edible starchy seeds. They also are referred to as cereal crops and include wheat, rice, maize (corn), barley, rye, oats, sorghum (jowar), and millet. The grain of all these cereals is used directly for human food and also for livestock, especially maize, barley, oats, and sorghum.

These large seeded grasses were among the first domesticated plants. Archeological studies indicate that some of these crops have been cultivated for 12,000–15,000 years. Cereal grains are the cheapest source of calories for human consumption and provide the most important energy source for three-fourths of the world population.

Cultivation. Wheat, barley, rye, and oats are cool season or temperate zone crops and are generally grown in low-rainfall areas (10–30 in. or 25–75 cm) because they are better adapted to cultivation under these conditions than other domesticated crops. However, they can produce large crops under higher rainfall, irrigation, and fertilizer applications. Rice is primarily a tropical or subtropical cereal, but Japanese plant breeders have developed types that grow at 45° latitude. Sorghum and some millets originally were tropical crops, but the area of adaptation has been greatly expanded by breeding new types. Maize also was a subtropical crop but is most productive in a temperate climate. Successful rice production is dependent on an abundance of water. Upland or dryland rice is available but limited in production. Some millets produce crops under dry, low-fertility conditions where other cereals often fail.

Storage. Another important attribute of these grain crops is the easy manner in which they can be stored. The grain often dries naturally before harvest to a safe moisture content (10–12%), or can easily be dried with modern equipment. Grain placed in adequate storage facilities can then be protected against insect infestations and maintained in sound condition for years. *See* CORN; BARLEY; MILLET; OATS; RICE; SORGHUM; WHEAT.

E. G. Heyne

Bibliography. J. E. Pratley, *Principles of Field Crop Production*, 3d ed., 1994.

Gram-equivalent weight

A quantity of a substance that contains the same number (known as the Avogadro number) of molecules as the number of atoms contained in exactly 12.000 g of carbon-12 (^{12}C). This convention stems from the concept that the central principle guiding chemical calculations is the relation of quantities of reacting substances to the numbers of molecules involved. *See* AVOGADRO'S NUMBER.

An added convenience in stoichiometric calculations is to incorporate the combining capacity (n), as well as the number of molecules, so that an equivalence of reacting substances is implicit without the need to examine balanced equations each time. Thus, the gram-equivalent weight of a substance is its gram-molecular weight divided by n. In acid-base reactions, n of the acid or base is given by the number of protons released or consumed in the reaction. For hydrochloric acid (HCl), ammonia (NH_3), acetic acid (CH_3COOH), and the acetate ion [CH_3COO; as in sodium acetate ($NaOOCCH_3$)], $n = 1$. For carbonic acid (H_2CO_3), sodium carbonate (Na_2CO_3), and ethylenediamine ($H_2NCH_2CH_2NH_2$), $n = 2$.

In precipitation and other metathetical reactions, the charge (valence) of the ion involved governs the value of n. Thus, $n = 3$ for ferric chloride ($FeCl_3$) and $n = 6$ for ferric sulfate [$Fe_2(SO_4)_3$] when precipitation yields either ferric hydroxide [$Fe(OH)_3$], barium sulfate, ($BaSO_4$), or silver chloride (AgCl). When oxidation-reduction is involved, the change of valence, rather than the valence itself, defines n. When the ferric ion (Fe^{3+}) acts as an oxidant [with the ferrous ion (Fe^{2+}) as product], $n = 1$ for ferric chloride ($FeCl_3$) and $n = 2$ for ferric sulfate. When potassium permanganate ($KMnO_4$) reacts in acid medium, $n = 5$, but in neutral or basic solution $n = 3$.

Further changes in the definition arise in dealing with metal complex formation, such as $FeF_6{}^{3-}$, $FeCl_4{}^{-}$, $Zn(EDTA)^{2+}$, and $Bi(EDTA)^{-}$; EDTA is the polydentate metal chelating agent ethylenediaminetetraacetate ion. The combining capacity of metals in complex formation depends on their coordination number, which, as these examples demonstrate, differs with different complexing agents, or ligands. *See* ELECTROCHEMICAL EQUIVALENT; EQUIVALENT WEIGHT; ETHYLENEDIAMINETETRAACETIC ACID; MOLE (CHEMISTRY); STOICHIOMETRY; VALENCE.

Henry Freiser

Gram-molecular weight

The molecular weight of an element or compound expressed in grams (g), that is, the molecular weight on a scale on which the atomic weight of the ^{12}C isotope of carbon is taken as 12 exactly. This replaces the earlier scale on which the atomic weight of oxygen was taken as 16.00 g. In the International System of Units, gram-molecular weight is replaced by the mole.

The ratio of the gram-molecular weights of any two elements or compounds must be identical with the ratio of the absolute weights of their individual molecules. Therefore, the gram-molecular weights of all elements or compounds contain the same number of molecules. This number, called the Avogadro number, N, is 6.022×10^{23}.

Since they contain the same number of molecules, the gram-molecular weights of all gases occupy the same volume at the same temperature and pressure. At 0°C and 1 atm (100 kilopascals) pressure this volume, called the gram-molecular volume, is 22.4 liters. *See* AVOGADRO'S NUMBER; GAS; MOLE (CHEMISTRY); MOLECULAR WEIGHT; RELATIVE MOLECULAR MASS.

Thomas C. Waddington

Grand unification theories

Attempts to unify three fundamental interactions—strong, electromagnetic, and weak—with a postulate that the three forces, with the exception of gravity, can be unified into one at some very high energy. The basic idea is motivated by the incompleteness of the electroweak theory of S. Weinberg, A. Salam, and S. Glashow, which has been extremely successful in the energy region presently accessible with the use of accelerators, and by the observation that the coupling constant for strong nuclear forces becomes smaller as energy increases whereas the fine-structure constant ($\alpha = 1/137$) for electromagnetic interactions is expected to increase with energy. The Weinberg-Salam-Glashow theory, which unifies electromagnetic and weak interactions, is incomplete in that strong interactions are not included; two coupling constants in the theory are unrelated; and in spite of many properties shared by them, quarks and leptons are unrelated, and their mass spectra and other properties remain unexplained. *See* GRAVITATION; STRONG NUCLEAR INTERACTIONS; WEAK NUCLEAR INTERACTIONS.

Models and successes. The simplest grand unification theory (GUT), proposed by H. Georgi and Glashow, is based on the assumption that the new symmetry that emerges when the three forces are unified is given by a special unitary group SU(5) of dimension 24. This symmetry is not observable in the low-energy region since it is badly broken. In this model, as in most GUTs, the coupling constants for the three interactions do actually merge into one at an energy of about 10^{14} GeV. Quarks (constituents of the proton, neutron, pion, and so forth) and leptons (the electron, its neutrino, the muon, and so forth) belong to the same multiplets, implying that distinctions between them disappear at the energy of 10^{14} GeV or above. In addition to the known 12 quanta of strong, electromagnetic, and weak interactions (namely 8 gluons, the photon, and the W^{+}, W^{-}, and Z^{0} particles), there appear, in this model, 12 new quanta with the mass of 10^{14} GeV. These generate new but extremely weak interactions that violate baryon- and lepton-number conservation. The most spectacular prediction of GUTs is the instability of the proton, which is a consequence of baryon-number (and lepton-number) violation. In the simplest SU(5) model, the decay $P \rightarrow e^{+} + \pi^{0}$ is a dominant proton-decay mode, with a predicted proton lifetime of $10^{29\pm2}$ years. The baryon-number violation can also induce neutron-antineutron oscillations. *See* GLUONS; INTERMEDIATE VECTOR BOSON; LEPTON; PHOTON; PROTON; QUARKS; SYMMETRY BREAKING; SYMMETRY LAWS (PHYSICS).

GUTs, in general, explain why the charge of the electron is precisely that of the proton with the opposite sign. In some GUTs with left-right symmetry, the lepton-number violation provides a novel way to generate the mass of neutrinos. Thus massive neutrinos are a distinct possibility in GUTs, and the smallness of their mass can also be understood. This leads to a possibility of neutrino oscillations which allow transformation of one neutrino species into another. Neutrinos with the appropriate mass may explain the missing mass in galaxies and clusters and perhaps the formation of galaxies. The prediction of the SU(5) theory of the ratio of the two unrelated constants in the Weinberg-Salam-Glashow model agrees impressively with observation, and the ratios of certain lepton and quark masses can also be calculated with reasonable success. A characteristic of the SU(5) model is the existence of the so-called desert between 100 GeV, which is the mass scale of the Weinberg-Salam-Glashow model, and 10^{14} GeV. This desert may bloom in more general GUTs. *See* NEUTRINO.

When applied to cosmology, GUTs have many implications. According to the scenario based on the GUTs, the universe underwent a phase transition when its temperature cooled to 10^{27} K which corresponds to 10^{14} GeV in energy and to the first 10^{-35} s after the big bang. This was when, for example, the SU(5) symmetry would have been broken. The phase transition caused an exponential expansion (10^{30}-fold in 10^{-32} s) of the universe, which explains why the observed 3 K microwave background radiation is uniform (the horizon problem), and why the universe behaves as if space is practically flat (the flatness problem). Baryon-number–violating interactions that were efficient in the early universe can qualitatively account for the observed ratio of the baryon-number density n_B and the photon-number density n_γ in the present universe, $n_B/n_\gamma \simeq 10^{-10}$, which has been a long-standing puzzle in the standard hot big bang model of the universe. Another prediction common to many popular GUTs is the production of magnetic monopoles with a mass of 10^{16} GeV or less in the early universe. *See* BIG BANG THEORY; COSMOLOGY; PHASE TRANSITIONS; UNIVERSE.

Problems. In spite of its theoretical triumph and spectacular predictions, the simple SU(5) model is practically untested by experiment and appears to be incomplete or even incorrect. No experimental evidence of proton decay has been established. In fact, experiments have set a lower limit for the proton lifetime of 6×10^{31} years, which is somewhat larger than the SU(5) prediction. Although various modifications and generalizations of the simple SU(5) model can solve this problem, they do not, in general, provide significant improvements over the simple version. The problems which GUTS leave unsolved are numerous and include the following: (1) gravity is not included in the unification; (2) the number of unknown parameters turns out to be unsatisfactorily large because of the complexity of symmetry-breaking mechanisms; (3) the mass spectra of the leptons and quarks remain unexplained; and (4) it is difficult to understand why there exist at least two vastly different but stabilized energy scales, one at 100 GeV and the other at 10^{14} GeV.

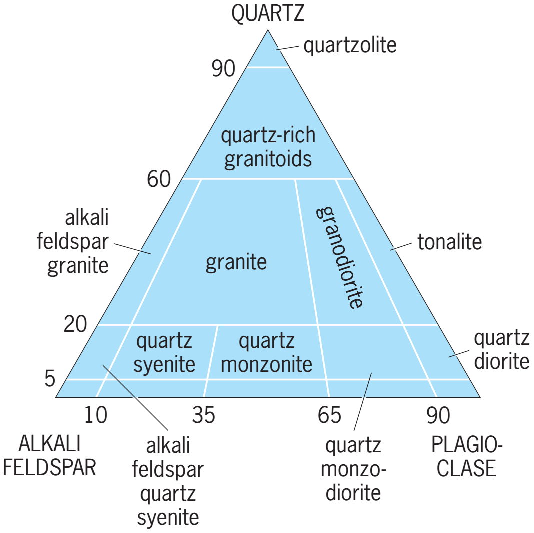
Prospects. Perhaps a true unification theory including gravity will emerge when the nature of symmetry breakings, the possible substructure of the leptons and quarks, and implications of supersymmetry, which is a new kind of symmetry relating matter (quarks and leptons) and quanta (such as the photon and the W^+, W^-, and Z^0 particles), are explored and properly understood. Unification theories with supersymmetry, known as supergravity, have been favored as candidates for such a theory. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; SUPERGRAVITY; SUPERSYMMETRY. Chung W. Kim

Bibliography. A. H. Guth and P. J. Steinhardt, The inflationary universe, *Sci. Amer.*, 250(5):116–128, May 1984; A. Linde, *Particle Physics and Inflationary Cosmology*, 1990; G. Ross, *Grand Unified Theories*, 1985; C. R. Storey, *Grand Unified Theory Made Easy*, 1993.

Granite

A crystalline igneous rock that consists largely of alkali feldspar (typically perthitic microcline or orthoclase), quartz, and plagioclase (commonly calcic albite or oligoclase). Its average grain size is 0.04–1.0 in. (1–25 mm); finer-grained rocks of this composition include rhyolite and aplite, and coarser-grained ones are granite pegmatite. *See* APLITE; PEGMATITE; RHYOLITE.

The revised nomenclature of the International Union of Geological Sciences (IUGS) subcommission defines granite as containing 80–100% by volume quartz, alkali feldspar, and plagioclase in specific proportions (see **illus.**), and 20–0% accessory minerals. The three essential minerals must include 20–60% quartz, and alkali feldspar must constitute 65–90% of the total feldspar. The variety alkali feldspar granite is similar except that alkali feldspar constitutes 90–100% of its total feldspar. The most common accessory minerals are biotite (typically forming 5–15% of the rock), iron-titanium (Fe-Ti) oxides (usually traces to 2% of magnetite or ilmenite), and traces



Classification of granitic rocks by the IUGS scheme. All proportions are by volume percentages.

of apatite and zircon. Hornblende or muscovite also are common accessories. The term granitic rocks includes granodiorite and tonalite as well as granite, and as used by some geologists may include quartz syenite to quartz diorite (see illus.). *See* BIOTITE; FELDSPAR; GRANODIORITE; HORNBLENDE; MUSCOVITE.

Texture. Granites range widely in texture. Some are massive or without structure, and all minerals are randomly oriented. Others show well-oriented tabular feldspars or flaky biotite grains, and are termed foliated or gneissic. Some show uniform grain size, some are seriate, and others exhibit phenocrysts (especially of microcline or orthoclase, rarely of quartz) one to several centimeters long set in a finer-grained matrix. *See* PHENOCRYST.

Occurrence. More than 99% of Earth's granites were emplaced in continental crust. The minor siliceous [silicon dioxide (SiO_2) $\geq$ 70 wt % of the rock] intrusive rocks of the ocean basins—as found in island arcs or mid-ocean ridges—typically are tonalite and contain little or no alkali feldspar. About 80% of all granitic rocks were formed in Precambrian times, and thus are older than 570 million years. Perhaps half of these older rocks are found in Archean (older than 2.5 billion years) greenstone-granite terranes (for example, in northern Ontario and Quebec). The remaining 20% are widely scattered through terranes of Phanerozoic (Paleozoic, Mesozoic, and Cenozoic) age.

Types. Granites may be divided into three major types: calc-alkaline, peraluminous, and alkaline. The peraluminous type is found with either of the other types in many regions. Each type has characteristic chemical and mineralogic features that are related to the geologic environment of its enclosing rocks and to its mode of origin.

Calc-alkaline. These granites, which include 70–80% of all granites, are the major type in orogenic belts of about 2 billion years to modern age, where oceanic crust is known to have been subducted at the margins of continental plates. They are also prominent in Archean crust, whose plate-tectonic mechanism of growth is controversial, but which may be like that of younger crust and occur by accretion of oceanic and other rocks at convergent margins.

Calc-alkaline granites typically are biotite or biotite-hornblende granites, some contain augite, and sphene is a common accessory. This type of granite tends to be more sodic [sodium oxide (Na_2O) = 3.3–5 wt %] than other types, has relatively low initial strontium-87/strontium-86 ($^{87}Sr/^{86}Sr$) isotopic ratios (0.7004–0.7008 for Phanerozoic ones and 0.7001–0.704 for Precambrian ones), and shows primary oxygen-18 ($\delta^{18}O$) values of 7–10 per mille. Because their parental rocks are largely of igneous type, calc-alkaline granites often are termed I-type granites. Those of the oceanic-continental convergent margins occur mostly in continental-margin batholiths, which consist of many tens to several hundreds of individual intrusive bodies of ovoid to irregular plan and of a few to several tens of kilometers in the maximum horizontal dimension. The individual intrusives of these batholiths commonly range from quartz diorite to tonalite, granodiorite, and granite.

Large continental-margin batholiths, like the Sierra Nevada batholith of the Sierra Nevada Range, California, may be several hundred kilometers long by 47–60 mi (75–100 km) broad. They form in the crust directly above subduction zones. Hot (2000–2300°F or 1100–1250°C) basaltic or andesitic magma that originates in the mantle at or above the subducted oceanic plate rises and pools in the lower or intermediate crust below the site of the batholith. The crust at such continental margins ranges from accreted oceanic rocks (basaltic, andesitic, and dacitic volcanic rocks, pelagic sediments, and ophiolites) to continental-margin sediments (especially turbidites) and continental crust (mafic to granitic igneous rocks, metamorphic rocks, and sedimentary rocks). The mantle-derived liquid directly melts and incorporates crustal rocks, while simultaneously precipitating minerals such as pyroxene, hornblende, or plagioclase. By this process the composition of the liquid changes; it becomes more siliceous and of dioritic to tonalitic composition. At the same time the mantle liquid in this lower-crustal chamber heats the enclosing rocks, which partially melt to granodioritic and granitic compositions. At a critical stage the primary magma body breaches its roof and rises as a discrete unit or diapir, and is emplaced in the upper crust. Afterward, the melted portions of the enclosing rocks accrete at the site of the former chamber and then rise up the conduit breached by the primary diapir. Thus magmas of a wide range of compositions are formed from both mantle and lower or intermediate crust. *See* CONTINENTAL MARGIN; DIAPIR; EARTH CRUST; OPHIOLITE; PYROXENE; TURBIDITE.

Calc-alkaline granites and granodiorites also form 25–50% of the granitic and gneissic rocks that contain the scattered Archean greenstone belts. These belts probably are accretionary collages of intra-oceanic island arcs; of submarine fans of graywacke and shale derived from nearby continental masses; of continental-shelf quartzite, carbonate rocks, volcanic rocks, graywacke, and other rocks; of fragments of older continental rocks; and of other rocks. These belts range from 3 to 9 mi (5 to 15 km) thick and cover 4000 to 12,000 mi^2 (10,000 to 30,000 km^2). During or after accretion the greenstone belts were folded and tectonically thickened, and they were intruded by basaltic magma from the underlying mantle. This hot magma was emplaced mostly in the lower crust, which it partially melted, and whose various rocks gave magmas that ranged in composition from tonalitic to granodioritic and granitic. These magmas rose gravitationally and were emplaced in the intermediate to upper crust. *See* GRAYWACKE; PLATE TECTONICS; SHALE; SUBMARINE.

Peraluminous. These granites contain aluminum in excess of that in feldspars. Biotite is the most common accessory mineral. Highly peraluminous granites, known as S-type granites, also may contain muscovite (biotite-muscovite granites are called two-mica granites), garnet, cordierite, sillimanite, and other aluminous minerals. Peraluminous granites are of two general types: one is found in or above accreted masses of graywacke, pelite, and minor basalt, pelagic sediments, and other rocks (that is, accretionary prisms) and consists largely of biotite granite and granodiorite; the other is associated with metamorphosed shale or pelite and is of the highly peraluminous two-mica type. Biotite granite and granodiorite of the first type are common in Archean to modern crust, whereas the biotite-muscovite type is known from the Precambrian but mostly is found in Paleozoic, Mesozoic, and even early Cenozoic fold belts. They occur in relatively heterogeneous plutons of lenticular or elongate plan, of a few to as much as 60 mi (100 km) in maximum dimension, and oriented parallel to regional fold trends. *See* CORDIERITE; GARNET; PLUTON; SILLIMANITE.

Peraluminous granites typically show high ratios of K/Na and Rb/Sr, high abundances of U and Th, initial ^{87}Sr/^{86}Sr ratios higher than about 0.710, and δ^{18}O values greater than 10 per mille. Their mode of origin remains a matter of some debate: one school holds that deep folding and thickening of graywacke or pelite heats them so that 10–50% melting occurs. The magma produced is a mixture of this melt and of pelitic xenoliths and xenocrysts that may be carried in suspension during ascent and emplacement of the magma. In some fold belts other stratified rocks, such as graywacke, arkose, or rhyolitic volcanic rocks, may be interlayered with pelite. Partial melting of such rocks tends to give calc-alkaline liquids, so that magma formed from such a mixed source will give granite intermediate in character between peraluminous and calc-alkaline types. Another school contends that as an oceanic plate is subducted at a continental margin, its covering of turbidites, other sediments, and minor igneous rocks is offscraped to give an accretionary prism. This prism, which may have a maximum thickness of 20–30 km, thus consists largely of graywacke and pelite. Its composition, position just above the Moho, and large mass make it a ready source for the generation of granite. Percolation of very hot basaltic liquid into such a prism from a leaky, underlying and subducting plate may account for many peraluminous granites.

Alkaline. Subalkaline to alkaline granites are characterized by iron-rich mafic minerals and relatively sodic alkali feldspar [molecular Na/(Na + K) from 0.3 to 0.5; K = potassium]. The subalkaline type typically contains ferruginous biotite or hornblende, or both, but varieties containing ferrohedenbergite or fayalite are not uncommon. Allanite and zircon are common accessories. The alkaline type contains the Na-Fe minerals aegirine or riebeckite-arfvedsonite, or both, and may also contain ferruginous biotite or even astrophyllite, eudialyte, or other rare minerals. Both subalkaline and alkaline types show 5–6 wt % K_2O, 3–4% Na_2O, low ratios of Fe^{3+}/Fe^{2+}, magnesium oxide (MgO) less than 0.1%, high abundances of F and light rare-earth elements [lanthanum (La), cerium (Ce), neodymium (Ne), samarium (Sm)], variable initial ^{87}Sr/^{86}Sr ratios, and low contents of water. These granites are found in continental interiors, in most instances along rift zones at moderate distances (60–500 mi or 100–800 km) from the continental margin. Alternatively, some of these granites also may be found along traces of hot spots, or ascending plumes of basaltic magma derived from the deep mantle (for example, the linearly disposed Mesozoic granitic intrusives of Nigeria). Minor dikes or plutons of diabase or gabbro, nepheline syenite, quartz syenite, and anorthosite are associated in many instances, and with the granites form a suite of genetically related rocks. The origin of this suite involves source material from both mantle and crust. During rifting or plume formation, basaltic magma is produced deep in the mantle. Some of this liquid may be extruded or emplaced in the upper crust as dikes, but a major part of it pools in the lower or intermediate crust. Here, as in the model of origin of continental-margin batholiths given above, this very hot liquid reacts with and partially melts the crustal rocks—many of which in this continental crust are granitic—and assimilates a low-melting or granitic component. Liquidus minerals, especially clinopyroxene and plagioclase, and refractory minerals of the crustal rocks are precipitated at the same time. In this process the liquid becomes dioritic, syenitic, and ultimately granitic, and precipitation gives pyroxenite, anorthosite, and mixed rocks. The syenitic and granitic magmas formed rise buoyantly, and are emplaced in the upper crust. In many cases a part of the granitic magma is extruded as rhyolitic volcanic rocks. *See* DOLERITE; GABBRO; IGNEOUS ROCKS; MAGMA; METAMORPHIC ROCKS; METAMORPHISM; NEPHELINE; PETROLOGY; PORPHYRY; QUARTZ.

Fred Barker

Bibliography. M. G. Best, *Igneous and Metamorphic Petrology*, 1982; A. Miyashiro, *Metamorphic Petrology*, 1994; L. A. Raymond, *Petrology: The Study of Igneous, Sedimentary, and Metamorphic Rocks*, 1994; H. J. Stein and J. L. Hannah (eds.), *Ore-Bearing Granite Systems: Petrogenesis and Mineralizing Processes*, 1990.

Granodiorite

A phaneritic (visibly crystalline) plutonic rock composed chiefly of sodic plagioclase (oligoclase or andesine), alkali feldspar (microcline or orthoclase, usually perthitic), quartz, and subordinate dark-colored (mafic) minerals (biotite, amphibole, or pyroxene). Granodiorite is intermediate between granite and quartz diorite (tonalite). Alkali feldspar is dominant over plagioclase in granite but is subordinate to plagioclase in granodiorite. Quartz diorite carries little or no alkali feldspar. For convenience granite and granodiorite are commonly grouped and referred to as granite. *See* GRANITE; IGNEOUS ROCKS.

Carleton A. Chapman

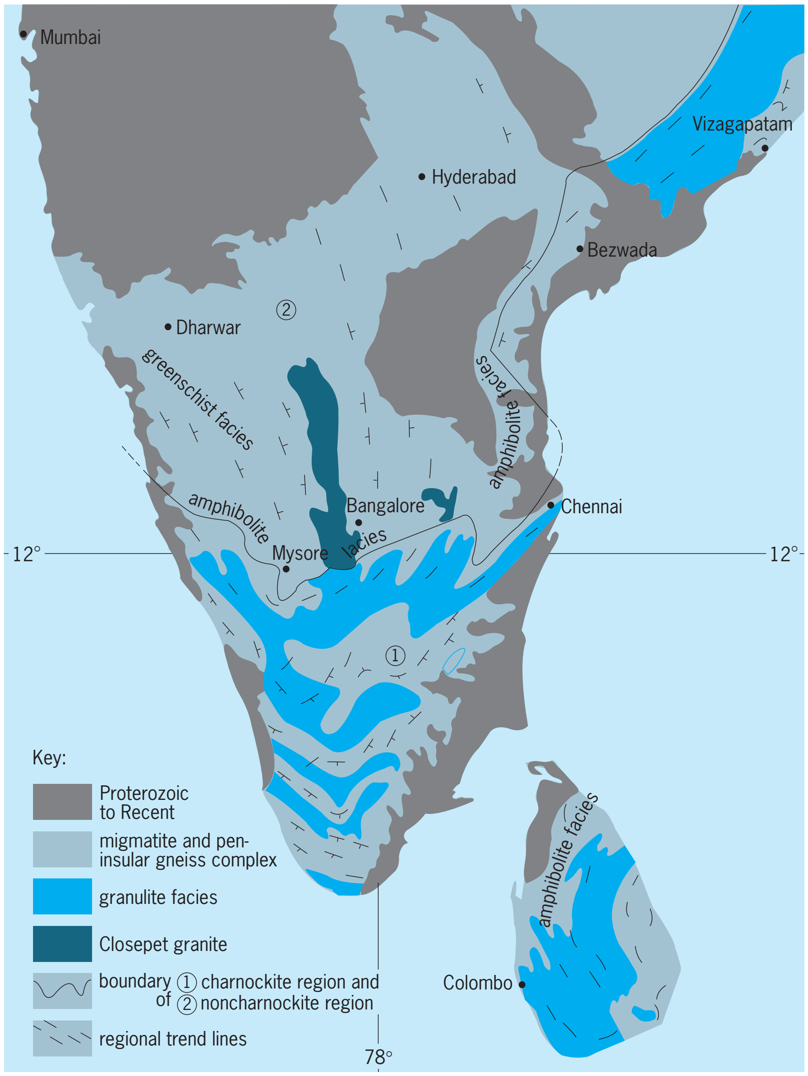
Granulite

An important class of metamorphic rocks exposed at the surface of the Earth's crust, and inferred to make up a large portion of the deeper crust. Granulites are known from experimental petrologic studies to have formed at higher temperatures, and in many cases, higher pressures, than most other crustal rock assemblages. Thus, they are believed to have formed at considerable depths in the crust. The transport of whole granulite terrains, sometimes hundreds of kilometers in lateral dimension, from depths of 12–18 mi (20–30 km), poses a major problem.

Granulites may be of many different bulk compositions, inherited from precursor sedimentary, igneous, or lower-grade metamorphic rocks. The high temperatures of crystallization have resulted in very low water content, reflected in nearly anhydrous mineralogy. Characteristic minerals of granulite metabasalts are plagioclase, orthopyroxene, clinopyroxene, hornblende, and garnet. These minerals are also characteristic of granulites of intermediate to granitic composition, together with progressively greater amounts of quartz and potassium feldspar. The association of potassium feldspar with orthopyroxene is definitive for charnockite, a granulite of approximately granitic composition characteristic of ancient high-grade terrains.

Almost all granulites are Precambrian. The most extensive granulite terrains are late Archean, 2.5–2.9 × 10^9 years old, as in Fennoscandia, southwestern Greenland, southern India, western Australia, and Antarctica, although some extensive granulite areas are younger Precambrian, as in the Grenville Province of southeastern Canada and northern New York. In several well-studied terrains, such as southwestern Greenland, southern Norway, and southern India, there are gradual regional transitions from rocks of lower metamorphic grade, rich in micas and other hydrous minerals, to granulites. The **illustration** shows one such terrain, in southern India and Sri Lanka, in which gradations from lower-grade rocks to granulites take place over distances of tens to a few hundreds of kilometers. These gradations presumably represent crustal gradients of increasing temperature or pressure, related to increasing crustal depth during the Archean metamorphic episode. How the high-temperature and high-pressure mineralogy was "frozen in" without substantial back reaction to lower-grade conditions during slow uplift to the surface after the metamorphic period is poorly understood; granulites share this problem with all other metamorphic rocks. Granulites characteristically contain CO_2-rich fluid inclusions in the mineral grains, in contrast to the more aqueous fluid inclusions of other kinds of rock. This suggests that action of volatiles low in H_2O and rich in CO_2, probably of subcrustal origin, were important in crustal metamorphism early in the Earth's history. *See* BASALT; GRANITE; METAMORPHIC ROCKS; METAMORPHISM.

R. C. Newton



Southern part of peninsular India and Sri Lanka, showing broad tectonic and metamorphic features. (*After A. P. Subramaniam, Charnockites and granulites of southern India*: *A review, Dan. Geol. Foren., 17:473–493, 1967*)

Bibliography. R. A. Howie, Charnockites, *Sci. Progr.*, 52:628–644, 1964; P. R. A. Wells, Chemical and thermal evolution of Archaean sialic crust, southern West Greenland, *J. Petrol.*, 20:187–226, 1979.

Granuloma inguinale

A mildly infectious, chronic, granulomatous disease principally affecting skin and subcutaneous tissues of the genital and rectal areas. Although rare in the United States, the disease is very common in New Guinea, the Caribbean, and other tropical and subtropical areas.

The causative organisms, *Calymmatobacterium granulomatis* (*Donovania granulomatis*), are gram-negative encapsulated bacteria isolated only with difficulty from lesion biopsy material by using chick chorioallantoic membrane or coagulated egg yolk slants. An organism similar to *C. granulomatis* is isolatable from feces, and such organisms are not infrequently detected in vaginal specimens.

The usual diagnostic procedure is the microscopic examination of smears of lesion scrapings that have been stained with Wright-Giemsa stain. There is no uniform serological procedure for either detecting antibodies against *C. granulomatis* or identifying the organism.

Although the method of transmission of granuloma inguinale is controversial, there is a definite correlation with sexual activity and a frequent association with homosexual behavior. Improved personal hygiene and individual awareness could correspondingly reduce morbidity and tissue destruction. *See* SEXUALLY TRANSMITTED DISEASES.

Tetracyclines are the drugs of choice, with streptomycin an effective alternative. Resistance to these drugs is countered by changing to either chloramphenicol or gentamicin. Resolution of the lesions begins in 1 week and is complete within 3–5 weeks, unless treatment is discontinued. *See* ANTIBIOTIC; MEDICAL BACTERIOLOGY. Douglas S. Kellogg

Bibliography. R. F. Dodson et al., Donovanosis: A morphologic study, *J. Invest. Derm.*, 62:611, 1974; G. Hart, Psychological and social aspects of venereal disease in Papua New Guinea, *Brit. J. Vener. Dis.*, 50:453, 1974; S. Lal and C. Nicholas, Epidemiological and clinical features in 165 cases of granuloma inguinale, *Brit. J. Vener. Dis.*, 46:461, 1970; E. H. Lennette et al. (eds.), *Manual of Clinical Microbiology*, 1974.

Grape

The two genera of grapes are *Vitis* and *Muscadinia*. *Vitis vinifera* has intermittent forked tendrils, bark that sheds, a diaphragm at the node, and elongated clusters with berries that adhere to the pedicels at maturity (**Fig. 1**). This species also has thin, smooth, shiny leaves with three, five, or seven lobes. Berries may be round or oval and have edible skins that adhere to the flesh. In the American species, skins slip from the pulp. Many American species have a characteristic musky or foxy odor and taste. *Muscadinia* can be easily distinguished from *Vitis* by bark that does not shed and simple tendrils that do not fork. *See* FRUIT.

Viticulture is the science of grape production. In a broad sense, viticulture includes studies of grape varieties; methods of culture such as trellising, pruning, and training; insect and disease control; propagation; and raisin production.

Cultivation. Countries with the largest acreages are Spain, Italy, France, Russia, and Turkey. California ranks ninth in acreage, and it produces 3% of the world's grapes. Worldwide, about 8000 varieties of grapes have been named and described. Of this number, perhaps about 100 are of major importance.

In the United States, *V. vinifera* is grown on the west coast, and most of the grapes cultivated east of the Rocky Mountains have been derived from American native species such as *V. labrusca* and *V. aestivalis*, or from crosses between them and *V. vinifera*. There is also a native Caribbean species and several Asiatic species. There are three main species of *Muscadinia* that are found mostly in the southeast portion of the United States.

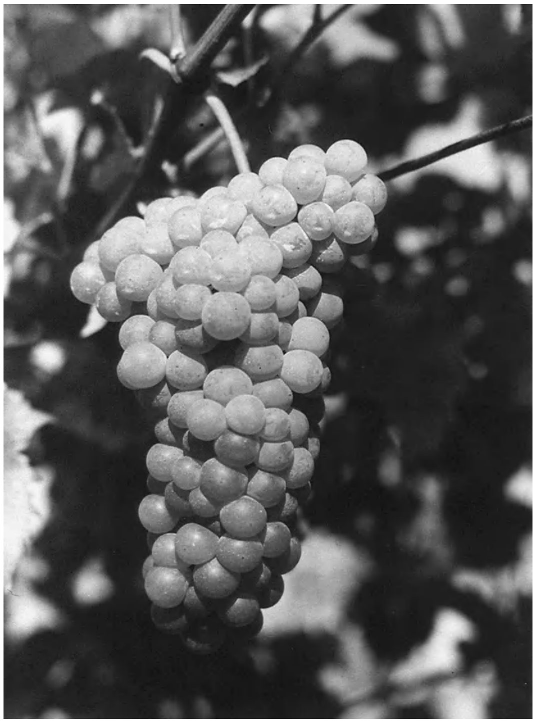


Fig. 1. Seyval grapes. (*Bountiful Ridge Nurseries*)

Grape stems are rather flexible, and so a trellis is usually needed to support the vine. In California the three main systems are the head-, cordon-, and cane-pruned vines (**Fig. 2**). For mechanical harvesting purposes, the vines are usually trained to a vertical, two- or three-wire trellis. In New York and the eastern United States the umbrella Kniffin, the four-cane Kniffin, and the Geneva double curtain are frequently used. The different kinds of training found around the world are virtually innumerable.

Uses. Table grapes are utilized for food and decorative purposes. They must have an attractive appearance, be palatable, have good shipping and storage qualities, and be resistant to injury in handling. Large berries of uniform size that have a firm pulp, a tough skin, a sturdy rachis, and a strong adherence of berries to cap stems are desirable. In the United States there is a marked preference for seedless grapes. Some of the leading table grapes in California are Emperor, Tokay, Thompson Seedless, Cardinal, and Perlette. Some of the principal commercial American varieties are Concord, Catawba, Delaware, and Niagara. Some of the important varieties of *M. rotundifolia*, the Muscadine grape, are Scuppernong, Thomas, and Hunt. The Republic of South Africa and Australia are also important table grape–growing countries.

Important wine grapes in California include Cabernet Sauvignon, Carignane, Chardonnay, Grenache,

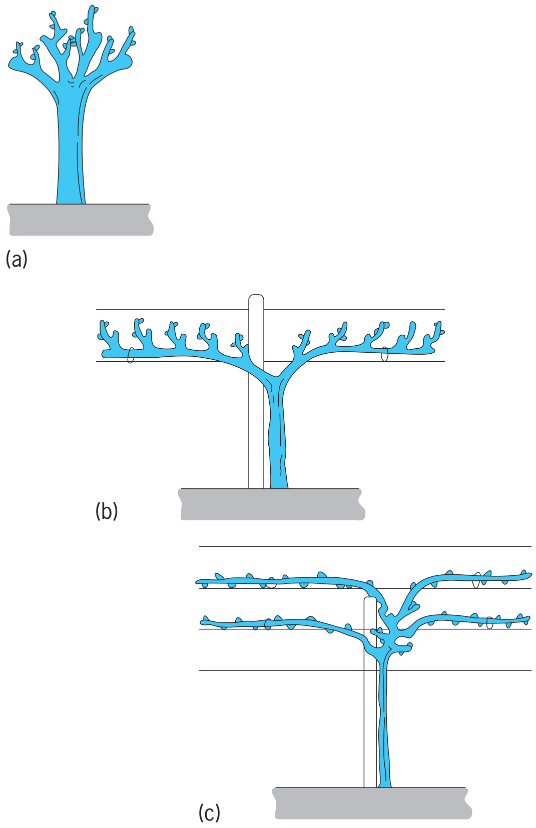


Fig. 2. Three main training-pruning systems in California. (*a*) Head-trained–spur-pruned. (*b*) Cordon-trained–spur-pruned (*c*) Head-trained–cane-pruned. (*After R. J. Weaver, Grape Growing, John Wiley and Sons, 1976*)

French Colombard, and Zinfandel. Many North American and *rotundifolia* species used for eating purposes are also used for wine.

A good dried raisin must be soft, textured, not sticky, early to ripen, and preferably seedless. The leading raisin grapes are Black Corinth, Thompson Seedless, and Muscat of Alexandria. The last variety has seeds that can be removed by machine. California produces 40% of the world's raisins, and there most are dried in the sun between the rows on paper trays. Other important raisin producers are Turkey, Greece, Spain, Iran, and Australia. In Middle Eastern countries the raisins are often dried on flat areas of soil or cement, and in Australia the raisins are dried in the shade on racks. It is important to dry raisin grapes before the rains begin. In rainy areas, facilities for rapidly covering the drying grapes must be available. Raisin grapes can also be dried in dehydrators.

Harvesting, processing, and marketing. Many wine grapes are now harvested mechanically. The slapper-type harvester (**Fig. 3**) has double banks of flexible horizontal rods that shake the vines and remove the clusters. The pulsator type of harvester strikes the posts rapidly so that the berries fall off. Fruit is immediately taken to the winery, where it is crushed and made into wine. Table grapes and most raisin grapes are still picked by hand. Soon after picking, table grapes start to deteriorate. If they are not to be eaten right away, they should be cooled as soon as possible. Most table grapes grown on the west coast of the United States are shipped east by train for marketing. Robert J. Weaver

Diseases. Grapes are susceptible to many diseases, which under certain climatic conditions may limit the kinds of grapes that can be grown. Disease control is particularly difficult in areas with high humidity, frequent summer rains, and high temperatures. Some disease control measures are required in all grape-growing areas.

Fungi. Powdery mildew (caused by *Uncinula necator*), anthracnose (caused by *Elsinoë ampelina*), black rot (caused by *Guignardia bidwellii*), and downy mildew (caused by *Plasmopara viticola*), are important grape diseases worldwide. Unlike the other three, powdery mildew occurs in dry climates. Symptoms include grayish-white powdery growth on the surface of green parts of the vine and dropping, discoloration, or splitting of the fruit (**Fig. 4**).

Anthracnose symptoms include lesions or cankers on all green parts of the vine; fruit lesions are circular, sunken, and ashy gray, and are surrounded by a dark margin (**Fig. 5**). Black rot affects young tissue of all parts of the vine, but damage to the immature fruit is most serious. Symptoms are reddish-brown, circular spots on leaves and, on bunch grapes, a rot which within 7–10 days transforms berries into black, hard, shriveled mummies that remain attached to the bunch. Downy mildew symptoms include light-yellow translucent spots on the upper leaf surfaces, patches of white mildew on the lower surfaces of leaves, and malformation of shoots, tendrils, or berries early in the season. Other fungal diseases include *Eutypa* dieback (caused by *E. armeniacae*), *Botrytis* rot (caused by *B. cinerea*), and bitter rot (caused by *Melanconium fuligineum*).

Vineyard sanitation, proper pruning, and planting on sites with good air drainage help reduce the severity of fungal diseases. Chemical spraying is usually necessary. Sulfur dusts or sprays control powdery

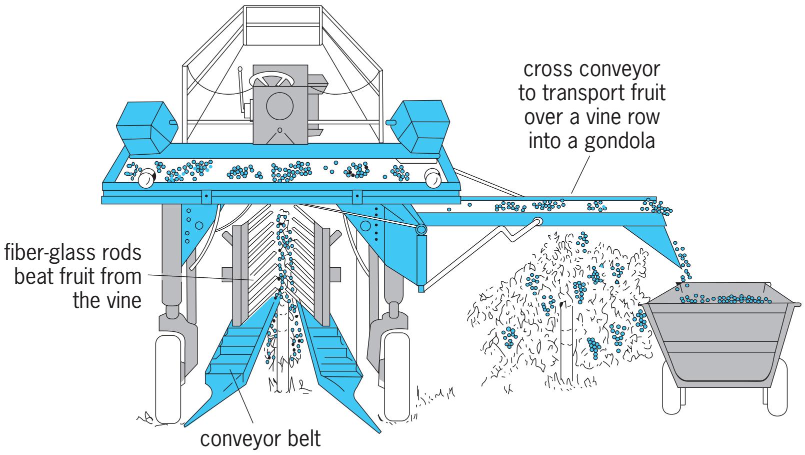


Fig. 3. Slapper-type mechanical harvester. (*After R. J. Weaver, Grape Growing, John Wiley and Sons, 1976*)

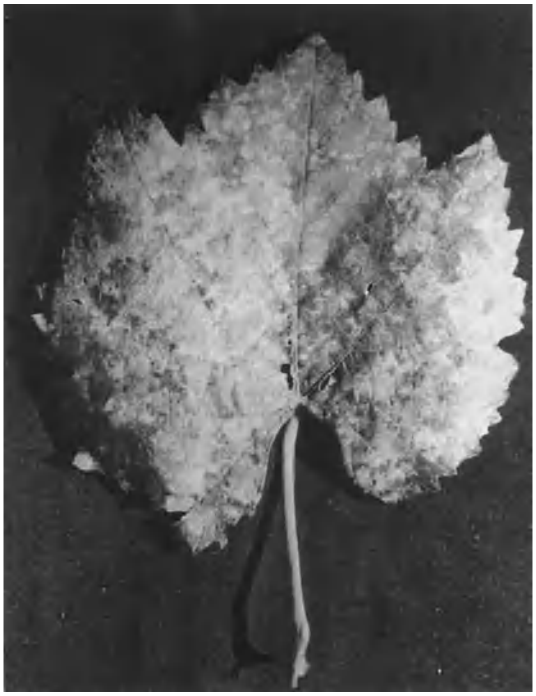


Fig. 4. Powdery mildew on a grape leaf.

mildew but injure the vines at high temperatures. For anthracnose control, a dormant spray of liquid lime-sulfur is beneficial. Bordeaux mixture is effective against most diseases, but injures young grape tissue. Several organic fungicides are effective. These include ferbam, captan, folpet, maneb plus zinc sulfate, zineb, and benomyl. Benomyl and ferbam are not effective against downy mildew.

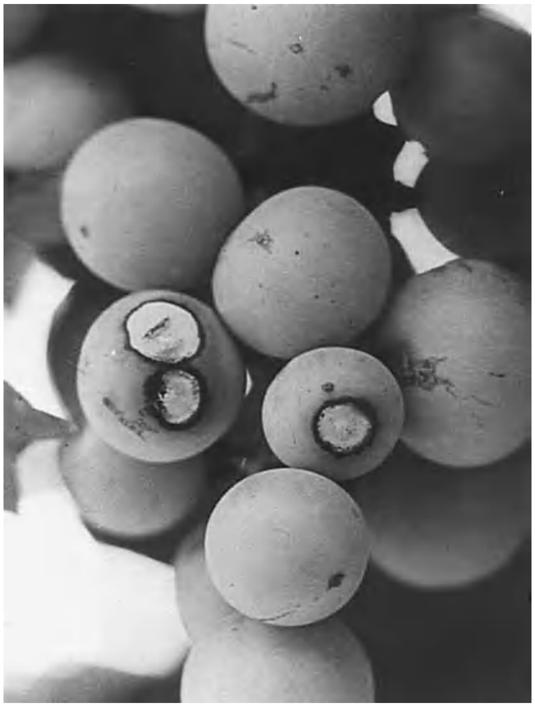


Fig. 5. Anthracnose on the grape fruit, commonly called bird's-eye rot.

Viruses. Grape leafroll, fanleaf, yellow mosaic, veinbanding, and corky bark are important viral diseases worldwide. Many other grape viruses are serious in certain locations. Since the grape viruses either have no vector or are nematode-transmitted, the best control is the use of virus-free planting stocks. Fumigation may also be used to reduce reinfections by the nematode-transmitted viruses such as fanleaf, yellow mosaic, and veinbanding.

Bacteria. Bacterial blight (caused by *Erwinia vitivora*) and crown gall (caused by *Agrobacterium tumefaciens*) are minor problems, but Pierce's disease is devastating to grapes and limits the production of bunch grapes in the southeastern United States. Pierce's disease symptoms include marginal necrosis of leaves, wilting and drying of fruit, decline of vigor, and usually death of the grapevine. The only control for Pierce's disease is the use of resistant or tolerant cultivars. Plants may be freed of the causal agent by hot water treatment (25 min at 50°C), and this is valuable in preventing the movement of the pathogen in propagating wood. *See* PLANT PATHOLOGY.

D. L. Hopkins

Bibliography. D. Flaherty (ed.), *Grape Pest Management*, rev. ed., 1992; R. C. Pearson and A. C. Goheen (eds.), *Compendium of Grape Diseases*, 1988.

Grapefruit

A citrus fruit, *Citrus paradisi*. It apparently arose as a hybrid of shaddock or pummelo and sweet orange in the West Indies. Its first recorded mention was in Barbados in 1750, and the first use of the term grapefruit occurred in Jamaica in 1814. It was thought to have been introduced into Florida by Count Odelle Phillipe near Safety Harbor on Tampa Bay around 1823. The term grapefruit was derived from the tree's tendency to produce large clusters of fruit, as grape vines do.

Plant structure. The tree is a large evergreen, spreading in habit and becoming larger than most other edible citrus species. Grapefruit wood is slightly less cold-tolerant than sweet orange but more resistant to heat. The individual leaves are relatively large compared to those of sweet oranges, and their broadly winged petioles slightly overlap the leaf bases. Flowers are white-petaled and perfect with pollination occurring naturally without the aid of bees. Fruit is relatively large and the peel thick compared to sweet oranges. Fruit shape is oblate or flattened at each end unless grown from off-bloom (not regular bloom) or under growing conditions promoting excessive vigor, in which case the fruit is often pear-shaped or sheep-nosed. The yellow peel color is not related to cool temperature as in the case of sweet oranges, but fruit picked early in the season must be degreened with ethylene to develop a satisfactory peel color. *See* ETHYLENE; FRUIT.

Hedging (shearing back the sides) and topping (shearing off the tops) of trees with large mechanical equipment are used to maintain trees within allotted spaces, providing for better light penetration and more efficient production and harvesting operations. Topping of larger trees reduces the top-to-root system ratio and results in larger fruit of better quality.

Cultivars. The original grapefruit were white-fleshed and extremely seedy; however, current important commercial cultivars are seedless or contain few seeds. Cultivars include the commercially seedless white Marsh, pink Marsh or Thompson, Redblush or Ruby Red, and Star Ruby and the seedy Duncan. Cultivars resembling grapefruit, but of no current commercial interest, include the extremely seedy but tasty Triumph and Royal. Grapefruit cultivars are not divided into early, midseason, and late as are sweet oranges, as there are only slight differences in maturity dates of these cultivars grown commercially.

All important grapefruit cultivars have sterile pollen, and the seedless ones are highly parthenocarpic and largely ovule-sterile. Most grapefruit cultivars have arisen from bud mutations in commercial groves, with their subsequent recognition and propagation by growers rather than as a result of traditional breeding programs. The Star Ruby was developed from seed of the seedy, pink-fleshed Hudson cultivar that was irradiated. A recent product of a breeding program at the University of California is the Oroblanca, a triploid developed from the hybridization of a grapefruit and pummelo.

Composition. Grapefruit ripens slowly over an extended period, storing well on the tree after reaching edible quality, with fruit of a given cultivar harvested from early fall to midsummer. Composition is, therefore, important not only for indicating nutritive values but also for determining proper time of harvest. The composition of fruit varies not only with the degree of maturity, but also with climate, cultivar, cultural practices, and other factors. Grapefruit may frequently be spot-picked for large size with the remainder left on the tree for later harvest.

The fresh weight of grapefruit consists of 35–50% juice, with the remainder made up of peel, pulp, and seeds. Fresh juice consists of 88–93% water, 7–12% soluble solids with sugars (sucrose and reducing) and acids (chiefly citric) constituting 85–90% of these solids. The edible quality of grapefruit depends in large measure upon the ratio of sugars to acids in juice. The nutritive value of juice is in part related to its vitamin C content. The juice also contains a number of other vitamins and mineral elements required in a well-balanced human diet. The principle giving grapefruit its distinctive bitter flavor is naringin, a glucoside not found in its progenitor the pummelo or in other commercial citrus. *See* ASCORBIC ACID; CITRIC ACID.

Uses. Profits from grapefruit production traditionally come from that portion of the crop marketed for fresh fruit. Thus, those factors influencing the percentage of the crop meeting fresh grade (pack-out) are very important. Excessive fertilization which results in poorly shaped fruit and blemishes from wind scarring and pests reduce market grade.

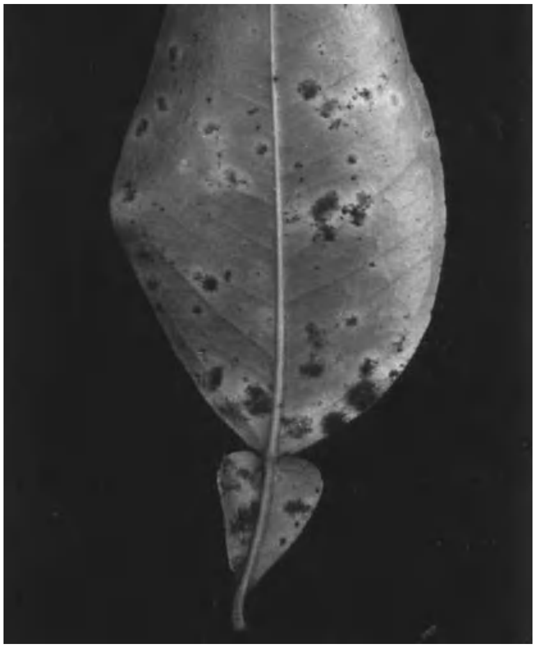


Fig. 1. Greasy spot lesions on lower side of grapefruit leaf.

The United States produces about 75% of the world's grapefruit, and Florida accounts for 75% of this total. The remainder is produced in Texas, California, and Arizona. Florida's grapefruit production has averaged just over 50 million 85-lb (38.5-kg) boxes annually, while combined production of the other United States citrus-producing areas has averaged just over 17 million boxes. Typically, 36% of Florida's crop is marketed as fresh fruit with about 50% of the crop in California-Arizona and Texas utilized in that form. The remainder of the crop in Florida and the other production areas is utilized for

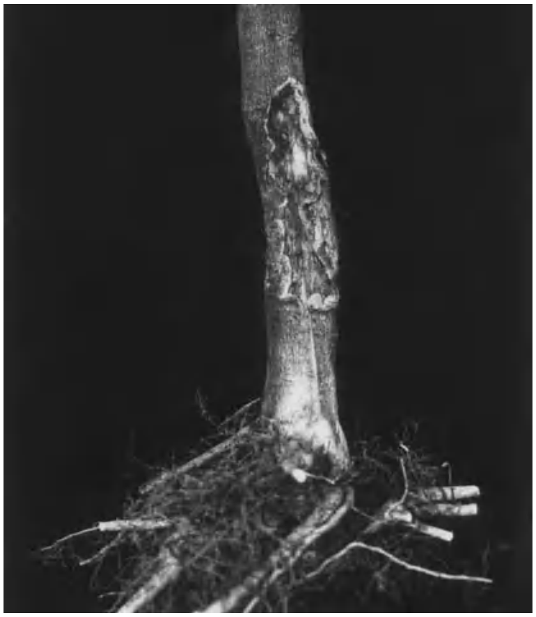


Fig. 2. Rootrot lesion on susceptible grapefruit scion.

processed products, primarily juices. Fresh grapefruit exports from Florida account for a large part of the fresh market shipments. In 1981–1982, for example, one-third of the total shipments were exported. Japan is the major export market for Florida grapefruit, accounting for about 70% of the export shipments. Most of the grapefruit juice production is sold in the United States market. Grapefruit juice is sold as either canned juice, ready-to-serve product, or frozen concentrate. Canned juice sales account for more than 50% of the volume sold through the grocery stores.

Diseases. Fungus diseases of foliage and fruit which must be controlled by spraying include greasy spot, caused by *Mycosphaerella citri* (**Fig. 1**); melanose, caused by *Diaporthe citri*; and scab caused by *Elsinoë fawcetti*. Rootrot, caused by the fungus *Phytophthora parasitica*, can result in a high percentage of young grapefruit tree loss. However, its incidence may be reduced by the use of tolerant rootstocks, high budding, and fungicides. High budding reduces the proximity of the susceptible grapefruit scion to the soil surface, the source of the fungal inoculum (**Fig. 2**). Grapefruit trees are susceptible to the important citrus viruses, including exocortis, psorosis, xyloporosis, and tristeza. However, the economic impact of these viruses is now minimized with the use of virus-free propagation material and tolerant rootstocks. Rust mites, the cause of rind russetting, must also be controlled if fruit is to be marketed fresh. *See* PLANT PATHOLOGY.

David P. H. Tucker

Bibliography. W. Reuther et al., *The Citrus Industry*, 5 vols., 1967–1989; R. W. Ward and R. L. Kilmer, *The Economics of the United States Citrus Industry: A Domestic and International Perspective*, 1988.

Graph theory

A branch of mathematics that belongs partly to combinatorial analysis and partly to topology. Its applications occur (sometimes under other names) in electrical network theory, operations research, organic chemistry, theoretical physics, and statistical mechanics, and in sociological and behavioral research. Both in pure mathematical inquiry and in applications, a graph is customarily depicted as a topological configuration of points and lines, but usually is studied with combinatorial methods. *See* COMBINATORIAL THEORY; TOPOLOGY.

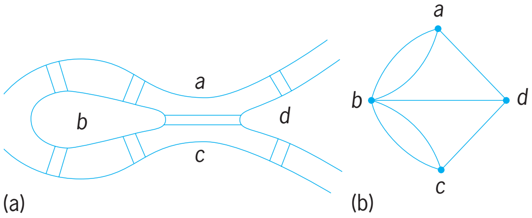


Fig. 1. Königsberg bridge problem. (*a*) The seven bridges of Königsberg. (*b*) Corresponding graph.

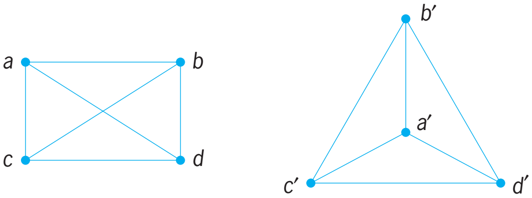


Fig. 2. Two isomorphic graphs.

Origin of graph theory. Graph theory and topology are said to have started simultaneously in 1736 when L. Euler settled the celebrated Königsberg bridge problem. In Königsberg, there were two islands linked to each other and to the banks of the Pregel River by seven bridges. **Figure 1** illustrates both this setting and its topological abstraction as a graph. The points *a*, *b*, *c*, *d* correspond to land areas, and the connecting lines to bridges. The problem is to start at one of the land areas and to cross all seven bridges without ever recrossing a bridge. Euler proved that there is no solution, and he established a rule that applies to any connected graph: such a traversal is possible if and only if at most two points are odd, that is, each is the terminus for an odd number of lines. Euler also proved that the number of odd points in a graph is always an even number. Thus, a complete traversal without recrossing any lines is possible if the number of odd points is zero or two. If zero, the complete traversal ends at the starting point.

In geometry a graph might arise as the set of vertices and edges of a convex, three-dimensional polyhedron, such as a pyramid or a prism. Euler derived an important property of all such polyhedra. Let V, E, and F be the numbers of vertices, edges, and faces of such a polyhedron. Euler proved that $V - E + F = 2$, which is now called the Euler equation. For instance, a cube has $V = 8$, $E = 12$, and $F = 6$, so that $8 - 12 + 6 = 2$. Euler's observations have been extended to a theorem about embeddings of graphs in surfaces and to the Euler-Poincaré characteristic for cell complexes in combinatorial topology.

Definitions. A graph consists of a set of points, a set of lines, and an incidence relation that designates the end points of each line. In many applications no line starts and ends at the same point. (Such a line would be called a loop.) Also, no two lines have the same pair of end points. A graph whose lines satisfy these conditions is called simplicial. The valence of a point is the number of lines incident on it, calculated so that a loop is twice incident on its only end point. Two graphs are isomorphic if there is one-to-one correspondence from the point set and line set of one onto the point set and line set, respectively, of the other that preserves the incidence relation. The point correspondence $a \rightarrow a'$, $b \rightarrow b'$, $c \rightarrow c'$, $d \rightarrow d'$ indicates an ismorphism between the two graphs of **Fig. 2**.

An automorphism of a graph is an isomorphism of a graph with itself. For instance, a plane rotation of 120° would yield an automorphism of either of the two graphs in **Fig. 3**. A plane reflection through

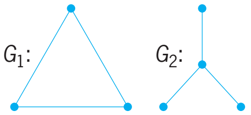


Fig. 3. Two graphs for which either a 120° rotation or a vertical reflection is an automorphism.

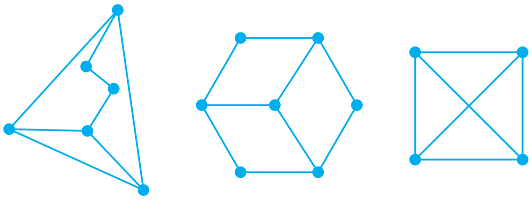


Fig. 4. Homeomorphic but nonisomorphic graphs.

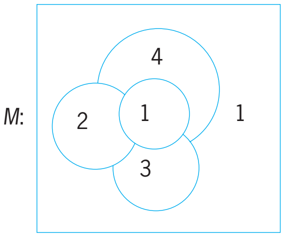


Fig. 5. Plane map requiring four colors.

a vertical axis would also yield an automorphism of either of them. The set of all automorphisms of a graph G forms the automorphism group of G. R. Frucht proved in 1938 that every finite group is the automorphism group of some graph. Two graphs are homeomorphic (**Fig. 4**) if, after smoothing over all points of valence 2, the resulting graphs are isomorphic. *See* GROUP THEORY.

Map coloring problems. Drawing a graph on a surface decomposes the surface into regions. One colors the regions so that no two adjacent regions have the same color, rather like a political map of the world. It is a remarkable fact that for a given surface, there is a single number of colors that will always be enough no matter how many regions occur in a decomposition of the surface. The smallest such number is called the chromatic number of that surface. It is easy to draw a plane map, as in **Fig. 5**, that requires four colors. In 1976 K. Appel and W. Haken settled a problem dating back to about 1850, by showing that four colors are always enough for plane maps.

Some maps on more complicated surfaces require more than four colors. For instance, **Fig. 6** illustrates a map on a torus (the surface of a doughnut) that needs seven. To obtain the toroidal map from the rectangular drawing, first match the top to the bottom to get a cylindrical tube. Then match the left end of the cylinder to the right end to complete the torus. Whereas, before this matching, region 7 meets only regions 1 and 6, after the matching it also meets region 2 along FG, region 3 along GH, region 4 along AB, and region 5 along BC. In fact, after the matching, each of the seven regions borders every other region. It follows that seven colors are necessary. No map on the torus needs more than seven colors, as P. J. Heawood proved in 1890. G. Ringel and J. W. T. Youngs completed a calculation in 1968 of the chromatic numbers of all the surfaces except the plane or sphere.

Planarity. A graph is planar if it can be drawn in the plane so that none of its lines cross each other. Neither of the two graphs in **Fig. 7** can be drawn in the plane. K. Kuratowski proved in 1930 that a graph is planar if and only if it contains no subgraph homeomorphic to either of those two graphs. Testing all the subgraphs might be a very tedious process, even on a fast computer. In 1974 J. Hopcroft and R. Tarjan obtained an extremely fast alternative planarity test. The time it takes a computer to perform the Hopcroft-Tarjan test is linearly proportional to the time it takes to read its point set into the computer.

There are methods to decide for any graph and any surface whether the graph can be drawn on the surface without edge crossings. The time to execute such methods is unfeasibly large for most graphs and most surfaces except the plane or the sphere. Ringel has constructed many important special drawings on higher-genus surfaces.

Variations. In a directed graph, or digraph, each line ab is directed from one end point a to the other end point b. There is at most one line from a to b. The adjacency matrix $M = (m_{ij})$ of a digraph D with points $b_1, b_2, \ldots, b_n$ has the entry $m_{ij} = 1$ if the line b_ib_j occurs in D; otherwise $m_{ij} = 0$ (**Fig. 8**).

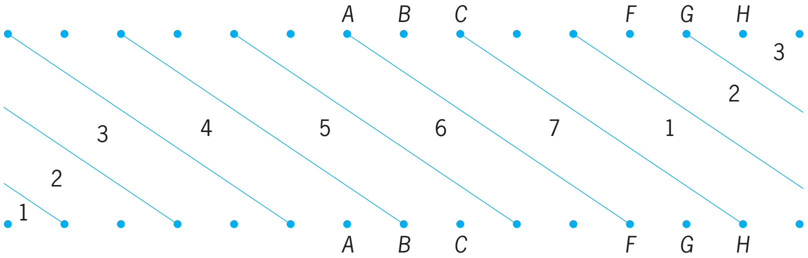


Fig. 6. Map on a torus (doughnut) that requires seven colors. To form the torus, paste opposite sides of the rectangle together.

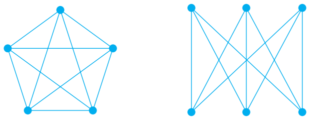
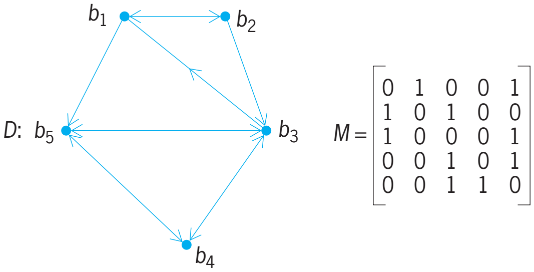


Fig. 7. Prototypes of all nonplanar graphs.



$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Fig. 8. A digraph and its adjacency matrix.

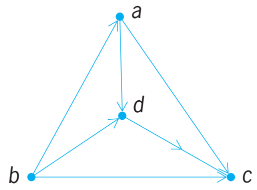


Fig. 9. A tournament.

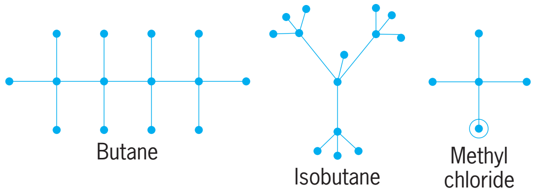
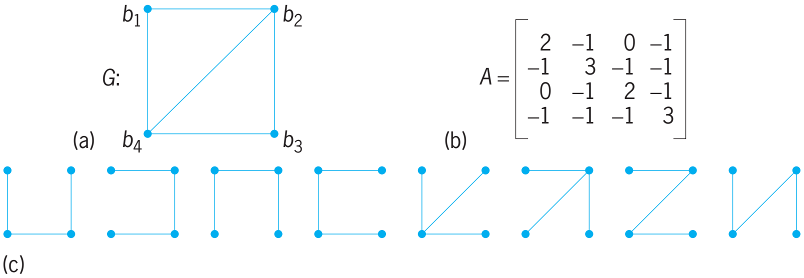


Fig. 10. The two isomers of a saturated hydrocarbon.

An oriented graph is obtained from an ordinary graph by assigning a unique direction to every line. If there is one line between ever pair of points and no loops, an ordinary graph is called complete. An oriented complete graph is called a tournament (**Fig. 9**).

Applications. A. Cayley reformulated the problem of counting the number of isomers of saturated hydrocarbons (C_nH_{2n+2}) in graphical language (**Fig. 10**). Each isomer is a tree all of whose vertices have valence 1 for hydrogen, or 4 for carbon. G. Polya devised a general theorem in 1937 for enumeration to provide a solution to such problems. F. Harary and others have solved many related problems by applying and extending Polya's theorem. Extremely effective use of Polya's theorem occurs in theoretical physics, where G. Ford and G. Uhlenbeck solved several graphical enumeration problems arising in statistical mechanics. *See* MOLECULAR ISOMERISM.

Suppose that some of the points of a graph correspond to workers $x_1, \dots, x_m$, that the rest of the points correspond to jobs $y_1, \dots, y_m$, and the presence of a line between x_i and y_j means that worker x_i is capable of performing job y_j. The personnel assignment problem is to find m lines so that each worker x_i is matched to exactly one possible job. In the optimal assignment problem, labels on the lines tell how well a worker can do a particular job. An algorithm due to H. Kuhn and J. Munkres solves the optimal assignment problem.



$$A = \begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix}$$

Fig. 11. Matrix-tree theorem. (*a*) Graph *G*. (*b*) Corresponding matrix *A*. (*c*) Spanning trees of *G*.

If the points of a graph represent cities and the lines between them are labeled with distances, one might want to find the shortest path from one point to another. An efficient method to determine a shortest path was developed by E. Dijkstra in 1959. K. Menger proved in 1927 that if A and B are disjoint sets of a connected graph G, then the minimum number of points whose deletion separates A from B equals the maximum number of disjoint paths between A and B. L. Ford and D. Fulkerson generalized Menger's theorem into a method for solving network flow problems. *See* LINEAR PROGRAMMING.

According to the physical laws of G. Kirchhoff and G. Ohm, any set of voltages applied to the input nodes of an electrical network determines the voltages at all other nodes and the currents on every branch. Kirchhoff also proved the result known as the matrix-tree theorem: Let G be a connected graph with points $b_1, \dots, b_n$ and let $A = (a_{ij})$ be the matrix such that a_{ij} is the valence of b_i and, for $i \neq j$, $a_{ij} = -1$ if b_i is adjacent to b_j or 0 otherwise. Then the cofactor of each entry a_{ij} equals the number of spanning trees of G, that is, the number of trees in G that includes every point of G (**Fig. 11**). *See* NETWORK THEORY.

Numerous applications of graph theory to social and behavioral science have been developed, many by Harary and his coauthors. If points represent persons and lines represent such interrelationships as communication, linking, or power, then a graph may depict various aspects of social organization. Anthropologists use graphs to describe kinship, and management scientists use them to display corporate hierarchy.

Graph theory is presently in a phase of rapid growth. Two of the major branches not described here are extremal graph theory, founded by P. Turan and developed by P. Erdös, and hypergraph theory, developed by C. Berge. Matroid theory, originated by H. Whitney and expanded by W. Tutte, is closely related to graph theory. Tutte is also responsible for important results in many other areas of graph theory and combinatorial research, including connectivity, decomposition, and chromatic numbers.

Jonathan L. Gross

Bibliography. N. L. Biggs, E. K. Lloyd, and R. J. Wilson, *Graph Theory 1736-1936*, 1986, reprint 1999; B. Bollobas, *Graph Theory: An Introductory Course*, 1979; J. L. Gross and J. Yellen, *Graph Theory and its Applications*, 1999; J. G. Michaels and K. H. Rosen (eds.), *Applications of Discrete Mathematics*, 1991; D. B. West, *Introduction to Graph Theory*, 2nd ed., 2000; R. J. Wilson, *Introduction to Graph Theory*, 1996.

Graphic methods

Procedures and techniques for visually representing on paper or a screen information pertaining to data analysis and decision making or relationships between variables by means of diagrams or charts.

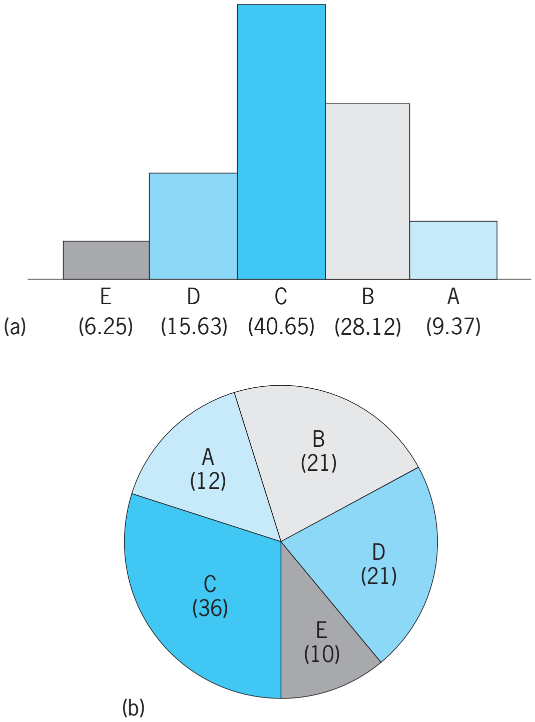


Fig. 1. Percentages of distribution of grades in a class, depicted by (*a*) bar graph and (*b*) circle graph.

Structural information is usually depicted by a bar graph, also known as a histogram (**Fig. 1***a*), or a circle graph, also known as a sectorgram or a pie chart (Fig. 1*b*). Such charts are commonly used to display data from industry, business, international trade, and government. The field of descriptive statistics relies heavily on graphs to simplify the presentation of data.

A graph representing the relation between an independent and a dependent variable is a two-dimensional line, whereas a graph representing the relation between two independent variables and a dependent variable is a three-dimensional surface. **Figure 2**, for example, is a line graph representing the functional relation $y = 5 - x$ between the nonnegative independent variable x and the dependent variable y.

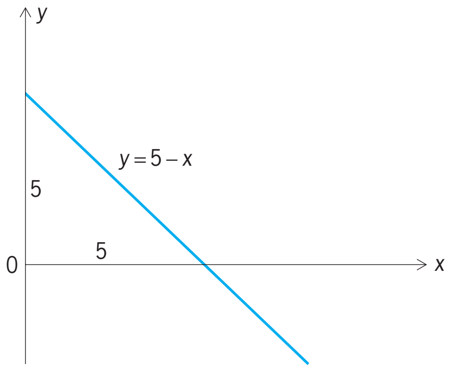


Fig. 2. Line graph of the functional relation $y = 5 - x$.

Graphical methods are frequently used to solve problems of curve fitting, correlation and regression analysis, nomographs and alignment charts, numerical integration, areas under curves, and interpolation and extrapolation. The bell-shaped curve representing the normal probability distribution is a line graph which is extensively used in problems involving sampling theory, estimation, and hypothesis testing in statistical analysis.

Line graphs are by far the most important type because they can be used not only to represent functional relationships but also to solve problems. For example, the roots of an equation in one variable can be found with the help of a graph. If $f(x)$ is a function which becomes 0 for a certain value of x, that value is a root of the equation $f(x) = 0$. The equation is solved by finding the x coordinates of points where the graph of $f(x)$ crosses the x axis. Two further examples of the solution of problems by line graphs will be briefly discussed: determination of equilibrium price and linear programming problems. *See* ROOT (MATHEMATICS).

Equilibrium price. If the price for a commodity is considered as the independent variable and the demand for the commodity as the dependent variable, the demand curve is a line graph illustrating the fact that as the price falls the demand increases. Likewise, the supply curve illustrates the fact that as price goes up the supply of the commodity is increased. If demand and supply curves are shown together, the point of intersection defines the equilibrium price under the law of supply and demand.

Linear programming. The line graphs of first-degree equations in two variables x and y of the type $ax + by = c$ (where a, b, and c are constants) are straight lines. A line divides the plane into two halves, one of which is represented by the inequality $ax + by \leq c$. The region enclosed by a finite number of half-planes and the two coordinate axes is a convex polygon called the feasible region, assumed to be nonempty. The problem of standard linear programming is that of finding a point in the feasible region at which the value of the objective function (a first-degree function of x and y) is a maximum. Even though there is an infinite number of points in the polygon, there is a vertex of the polygon at which the objective function attains its maximum value. Thus a graphical method can be used to solve this linear programming problem in two variables by drawing the polygon and computing the coordinates of all its vertices (a finite number) from which an optimal vertex can be easily chosen. *See* COORDINATE SYSTEMS; CURVE FITTING; EXTRAPOLATION; INTERPOLATION; LEAST-SQUARES METHOD; LINEAR PROGRAMMING; NOMOGRAPH; NUMERICAL ANALYSIS; STATISTICS.

V. K. Balakrishnan

Bibliography. R. W. Bowen, *Graph It!: How to Make, Read and Interpret Graphs*, 1991; R. D. Gustafson and P. D. Frisk, *Functions and Graphs*, 2d ed., 1991; S. M. Kosslyn, *Elements of Graph Design*, 1993; E. M. Mikhail, *Observations and Least Squares*, 1982; J. Molnar, *Nomographs: What They Are and How to Use Them*, 1981.

Graphic recording instruments

Instruments that make a graphic record of one or more quantities as a function of another variable, usually time. Signals representing information, such as the shape of time-varying electronic waveforms or the movements of a machine, are presented in graphical form by these devices. *See* TRANSDUCER.

The name of an instrument with a recording device incorporated often includes "-graph," such as a barograph for recording barometric pressure data.

Although there are still many graphic recording instruments in use, the trend now is to interface the sensor providing the signal to a computer. The signal can be processed immediately in any desired way and presented as a graph on the computer display or by a printer attached to the computer. This practice is much more versatile and often less expensive than the use of graphic recording instruments.

Classification. Recorders are of two forms: those that plot an input variable with respect to time (denoted x-t) and those that plot two different variables (denoted x-y). Graphic recorders can also be classified as being either directly or indirectly driven by the input signal. In addition, they can be classified by their exhibiting means, recording means, number of marking devices or channels, and marking means.

Direct- and indirect-acting recorders. Direct-acting units (**Fig. 1**) are suitable when the primary variable exerts sufficient and appropriate force to overcome the frictional loads of the bearings and marking means. Direct-drive recorders lack general-purpose usefulness and are, therefore, more costly to produce than indirect alternatives. In some circumstances, however, they are essential.

In the indirect form an intermediate stage is used to convert the input signal into the equivalent mechanical movements needed. Indirect-acting recorders generally accept voltage or current input signals, converting these into equivalent mechanical positions on a suitable exhibiting medium. Other input variables, such as pressure and temperature, are first transformed into electrical signals. Such signals may range from microvolts to megavolts or from picoamperes to kiloamperes. Recorders are often supplied with adjustable input stages so that the input signal can be matched to the scale size needed for the graph.

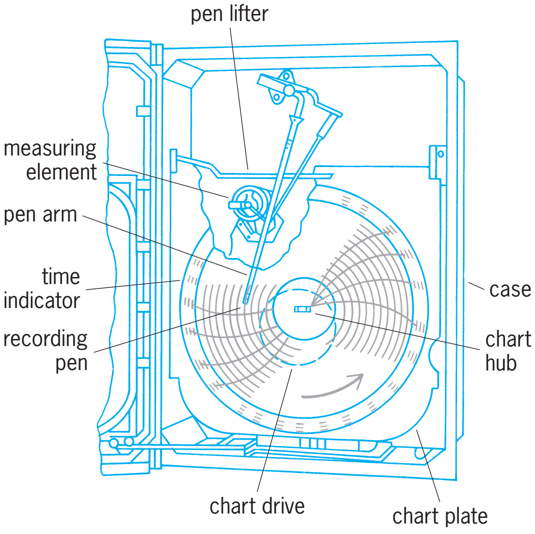


Fig. 1. Single-pen, direct-acting, circular-chart pressure recorder. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

Exhibiting means. Exhibiting means can be of either circular form or linear form, the latter being continuous strip or a single sheet. Circular charts vary from 4 to 12 in. (100 to 300 mm) in diameter. Strip-chart widths range from 2 to 12 in. (50 to 300 mm), with flat-bed sheets being the commonly used paper sheet sizes. Roll-chart lengths are available from 65 ft (20 m) to more than 330 ft (100 m). The paper is often fan-folded for convenience. The medium used can be either plain or specially treated paper or film. The recording means may be continuous, intermittent (when the marking device is retracted from the chart between measurements), or sequential (when more than one variable is intermittently recorded).

Multichannel recording. Multichannel recorders can record several signals simultaneously. This can be done by using several recording mechanisms combined in a single unit (**Fig. 2**). Alternatively a single marking mechanism can sequentially record values of each channel. One popular method, the dotting recorder, makes dots on the paper of a different color for each channel. Being closely placed, the dots produce a set of continuous, differently colored traces.

Marking methods. Many different methods are used to produce the permanent trace on the recording medium. They include the use of fluid inks and fiber pens that mark in direct contact with the paper; pressure-forced ink jetting; marking with a heated stylus on heat-sensitive paper; pressure- or voltage-sensitive paper on which the stylus moves in contact; pressure of an inked ribbon onto the paper after the marking head has been positioned; exposure of heat-sensitive or photographically sensitive paper or film; and electrostatically charged images in a laser printer.

Drive systems. Motive power is required to move the marking device relative to the exhibiting medium. In the x-t recorder the marker is driven across an axis mounted perpendicular to the medium that is moved under the medium at a steady rate with time (Fig. 2). In the x-y version the marker can be driven in two perpendicular directions, the medium being fixed (**Fig. 3**). Some systems move a single sheet backward and forward under the markers.

Two basic kinds of drive are used in recorders. The first is a speed-controlled drive that moves a strip of exhibiting medium under the marker; the other positions a marker along the axis at a distance that is proportional to the input magnitude.

Spring-motor chart drives are sometimes used, but the majority of x-t recorders are driven by electric motors of various kinds. Stepping motors have become prolific. Gearing is used to obtain the required chart speeds which can vary from several feet per second to a fraction of an inch per day. Speed is regulated by the use of servo velocity-controlled drives or synchronous motors when suitable alternating-current (ac) supplies are available. *See* MOTOR; SERVOMECHANISM; STEPPING MOTOR; SYNCHRONOUS MOTOR.

Marker positioning systems make use of open- or closed-loop positioning systems. For these, rotary or linear drive motors are used. *See* CONTROL SYSTEMS.

A simple, extensively used way to move a marker to a position proportional to an input electrical signal is to place it on the end of a needle driven by an electromagnetic meter movement. The disadvantage of this method is that the friction of the contacting pen and the circular sweep of the pen can introduce error. Ink pen systems are also not suitable when very rapid following of the response is required. In the ultraviolet recorder and film oscillographs the meter movement (galvanometer) is used to rotate a small mirror that deflects a beam of radiation onto a sensitive exhibiting medium. These instruments can record kilohertz waveforms. *See* AMMETER; ELECTRICAL MEASUREMENTS; GALVANOMETER; VOLTMETERS.

For more robust operation, feedback control systems are used to control the speed of chart drives and the position of marker units. Each uses a similar principle of control.

As the marker unit of a typical position-control feedback system (**Fig. 4**) moves across the chart, it also moves the connection point made on a resistance potentiometer unit that is fed with electric current from a biasing dc voltage source. The input voltage (e) and the potentiometer output voltage (e_s) are fed to the two inputs of a differential electronic amplifier. If they are not equal, the amplifier produces a difference signal that is applied to the drive motor (M). Depending on the polarity of this difference, the motor moves the marker (as a smooth and continuous process) to a point where the difference is kept close to zero. At that point the marker is at a position on the chart that is proportional to the input signal magnitude. Similar means are used to control speed, but there a velocity sensor is used instead of the potentiometer. *See* DIFFERENTIAL AMPLIFIER.

Analog and digital forms. Graphic recorders (commonly referred to as plotters) are designed for use with either analog or digital electrical input signals. In the analog variation the input circuitry accepts signals that have the information to be displayed carried in terms of the variable amplitude of a voltage or current. A variable-gain input stage allows input signal magnitudes to be matched, by amplification or attenuation, to the set sensitivity of the positioning system.

In the digital plotter the digital signal (having only two states of existence for each of the bits forming the digital word equivalent to the analog alternative) can be of either serial format or parallel format. In the former a time sequence of serially occurring binary bits represents the signal amplitude. In the latter it is the state of binary signals simultaneously existing on several parallel lines that represents the amplitude. The signal lines carrying these digital signals, for purposes of interconnection, are called the communication interface.

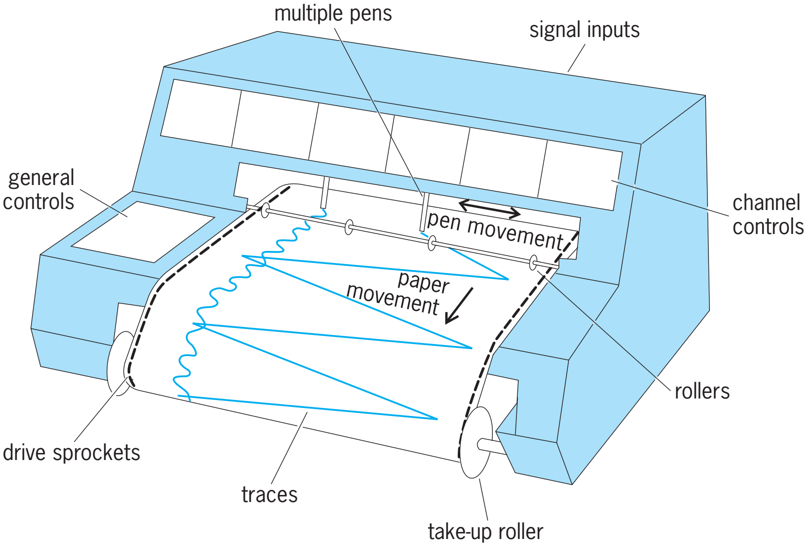


Fig. 2. Multichannel, x-t recorder with several recording mechanisms combined in one unit. Signal inputs feed channels at rear, and each channel has one set of controls. Multiple pens are staggered to miss each other as they traverse exhibiting medium, and produce traces which may be in different colors. Paper is held down by rollers and moved forward by sprocket-driven perforations.

As the digital interface is capable of transmitting multiple channels of information to the plotter, and because digital plotters contain a microprocessor computer that provides processing power and a measure of intelligence, such plotters offer more facilities than analog models. For example, in addition to plotting more than one input signal, they can be instructed (using a command language) to generate common geometric shapes, select axis sensitivities, change pens to provide different colors or line widths, draw text of required size, provide annotated axes to charts, add data point markers, lift the pen, set the drawing speed, and set the plot to a chosen paper or plotted size. Storage buffers are often provided that store data of the signal to be recorded.

Dynamic performance. Inadequate dynamic response from a recorder will give traces that have less

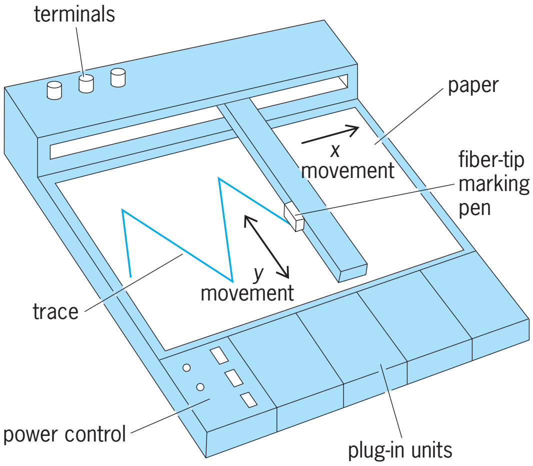


Fig. 3. One form of flat-bed, x-y chart recorder. Plug-in units can be used to configure recorder range, filter signals, and drive the x movement from time.

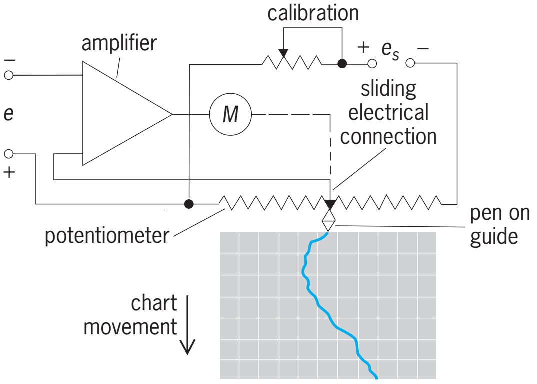


Fig. 4. System for potentiometric feedback positioning of a marker.

amplitude and possibly a different time shape than the actual signal. For multiple-trace systems, it can also result in incorrect phase information between the channels. Dynamic response is described by the time taken to trace a stated amplitude on the exhibiting medium. A high-speed oscillograph can trace a full-scale waveform in a millisecond, whereas a typical ink pen recorder will need 250 ms to cover the width. The buffered digital plotter largely avoids dynamic plotter response problems by recording the data in a high-speed electronic buffer, plotting output at a speed suited to the observer.

Transient recorders. Recording details of fast-changing signals requires considerable chart length. If the events desired only rarely arise, a transient recorder can be used. Here incoming data are continuously captured by using digital electronic storage. As new data are entered, the first part of earlier recorded data is destroyed. If the set of data held at any time contains a recognizable event of interest, the output is fed to a conventional recorder. Transient recorders are used to capture transients on electrical power lines and to record large amplitudes of seismic vibrations.

Application. The extensive acceptance of computer-based data logging and the general availability of hard-copy printers have contributed to a shift in usage from dedicated chart recorders to digital printers that provide permanent displays in the form of computer printouts. The plotting mechanisms incorporated within printers are similar to those of conventional graphic recording instruments, the majority being driven with stepping motors and using thermal- or laser-based printing. Dedicated graphic recording devices are still in use because it is often not economical to replace them with modern equivalents. They also find application where records need longer paper lengths than a computer printer can provide. *See* COMPUTER PERIPHERAL DEVICES; STEPPING MOTOR.

Peter H. Sydenham

Bibliography. G. K. McMillan (ed.), *Process/Industrial Instruments and Controls Handbook*, 5th ed., 1999; L. Michalski et al., *Temperature Measurement*, 2d ed., 2001; R. L. Moore, *Process Analyzers and Recorders*, vol. 2 of *Basic Instrumentation Lecture Notes and Study Guide*, ISA, rev. ed., 1989; B. E. Noltingk, *Instrumentation Reference Book*, 2d ed. 1995; J. G. Webster (ed.), *The Measurement, Instrumentation, and Sensors Handbook*, 1999.

Graphite

A low-pressure polymorph of carbon, the common high-pressure polymorph being diamond. Several other rare polymorphs have been synthesized or discovered in meteorites. The contrast in physical properties between these two polymorphs is remarkable: Graphite is metallic in appearance and very soft, whereas diamond is transparent and one of the hardest substances known.

Graphite is hexagonal, space group P 6_3/m mc, $a = 0.248$ nanometer, $c = 0.680$ nm, with 4C in the unit cell. Its atomic arrangement consists of sheets of carbon atoms at the vertices of a planar network of hexagons (**Fig. 1**). Thus, each carbon atom has three nearest-neighbor carbon atoms. The layer distance between the sheets is $c/2 = 0.340$ nm. Diamond, on the other hand, is a three-dimensional framework structure with the carbon atoms in tetrahedral (fourfold) coordination. Crystals of graphite are infrequently encountered since the mineral usually occurs as earthy, foliated, or columnar aggregates often mixed with iron oxide, quartz, and other minerals (**Fig. 2**). *See* CRYSTAL STRUCTURE; DIAMOND.

Properties. The sheetlike character of the graphite atomic arrangement results in distinctive physical properties. The mineral is very soft, with hardness $^1/_2$; it soils the fingers and leaves a black streak on paper, hence its use in pencils. The specific gravity is 2.23, often less because of the presence of pore spaces and impurities. The color is black in earthy material to steel-gray in plates, and thin flakes are deep blue in transmitted light. One perfect cleavage is parallel to the hexagonal sheets, allowing the mineral to be split into thin flexible but nonelastic folia.

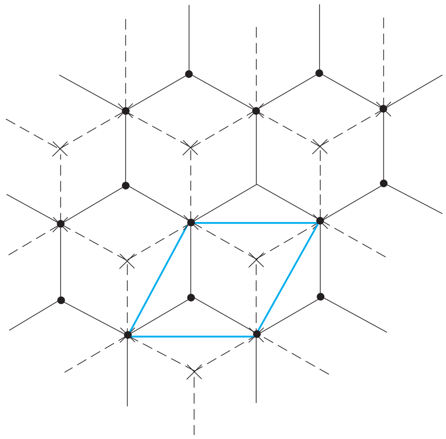


Fig. 1. The graphite atomic arrangement down the *c* axis. There are two distinct layers of carbon atoms, related by symmetry. The first level consists of carbon atoms (black circles) at the vertices of a network of hexagons. The second level, also a hexagonal network with carbon atoms at the vertices (crosses), is shifted relative to the first level.

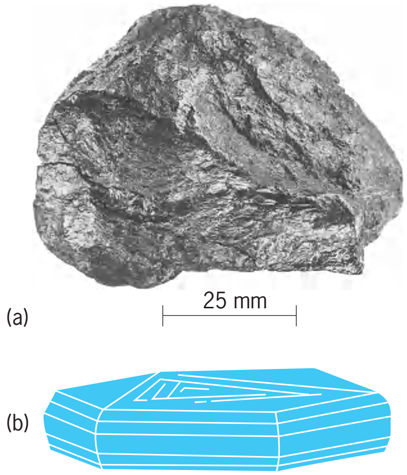


Fig. 2. Graphite. (*a*) Earthy aggregate (*American Museum of Natural History specimen*). (*b*) Hexagonal crystal of graphite with triangular markings on face (*after C. Palache, H. Berman, and C. Frondel, Dana's System of Mineralogy, vol. 1, 7th ed., John Wiley and Sons, 1944*)

Graphite is a conductor of electricity, distinguishing it from amorphous carbon (lampblack).

Graphite closely resembles molybdenite, MoS_2, a mineral with similar crystal structure, but the two can be distinguished by the greenish streak and much higher (4.70) specific gravity of the latter mineral. Early confusion with the brittle gray lead sulfide, galena, resulted in the synonymous trade names plumbago and black lead for graphite. In a similar manner, the misleading term "lead pencil" has persisted and is still in common usage.

Occurrence. Graphite arises from the thermal and regional metamorphism of rocks such as sandstones, shales, coals, and limestones which contained organic products not exposed to an oxidizing environment. It also can form in a strongly reducing environment, such as in serpentinites and limestones where hydrogen gas may reduce carbon dioxide. Platy graphite showing crude crystal surfaces often occurs speckled in coarsely crystallized marbles. The major sources of graphite are in gneisses and schists, where the mineral occurs in foliated masses mixed with quartz, mica, and so on. Noteworthy localities include the Adirondack region of New York, Korea, and Sri Lanka. In Sonora, Mexico, graphite occurs as a product of metamorphosed coal beds. Graphite is also observed in meteorites, where the mineral was formed under strongly reducing conditions, usually in association with metallic iron. Paul B. Moore

Synthetic Graphite

Graphite has a highly developed crystalline structure, and its softness, high thermal and electrical conductivity, and self-lubricating qualities differentiate it from other forms of carbon.

Carbon in graphitic form has both metallic and nonmetallic properties. Commercially produced synthetic graphite is a mixture of crystalline graphite and cross-linking intercrystalline carbon. Its physical properties are the result of contributions from both sources. Thus, among engineering materials, synthetic graphite is unusual because a wide variation in measurable properties can occur without significant change in chemical composition.

At room temperature the thermal conductivity of synthetic graphite is comparable to that of aluminum or brass. An unusual property of graphite is its increased strength at high temperature. The crushing strength is about 20% higher at 1600°C (2900°F) and the tensile strength is 50–100% higher at 2500°C (4500°F) than at room temperature.

Graphite is resistant to thermal shock because of its high thermal conductivity and low elastic modulus. It is one of the most inert materials with respect to chemical reaction with other elements and compounds. It is subject only to oxidation, reaction with and solution in some metals, and formation of lamellar compounds with certain alkali metals and metal halides.

Uses. Graphite has many uses in the electrical, chemical, metallurgical, nuclear, and rocket fields: electrodes in electric furnaces producing carbon steel, alloy steel, and ferroalloys (**Fig. 3**); anodes for the electrolytic production of chlorine, chlorates, magnesium, and sodium; motor and generator brushes; sleeve-type bearings and seal rings; rocket motor nozzles; missile nose cones; metallurgical molds and crucibles; linings for chemical reaction vessels; and, in a resin-impregnated impervious form, for heat exchangers, pumps, pipings, valves, and other process equipment.

Preparation. Synthetic graphite can be made from almost any organic material that leaves a high carbon residue on heating to 2500–3200°C (4500–5800°F). In commercial operations, raw materials are carefully selected because not all substances with high carbon content undergo a suitably complete transformation to graphite at these temperatures. Petroleum coke is raw material for the most commonly used production process. After calcining and sizing, the coke is mixed with coal tar pitch, heated to about 165°C (329°F), and formed by extrusion or molding to so-called green shapes. Baking to 750–1400°C (1400–2600°F) in gas- or oil-fired kilns follows the forming operation. Graphite is produced by heating the baked shapes to 2600–3000°C (4700–5400°F) by passing electricity amounting to 1.6–3.0 kW per pound of graphite through the bed of a furnace

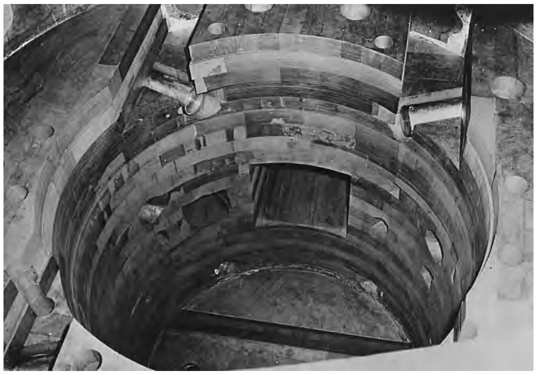


Fig. 3. Graphite reflector for a test reactor.

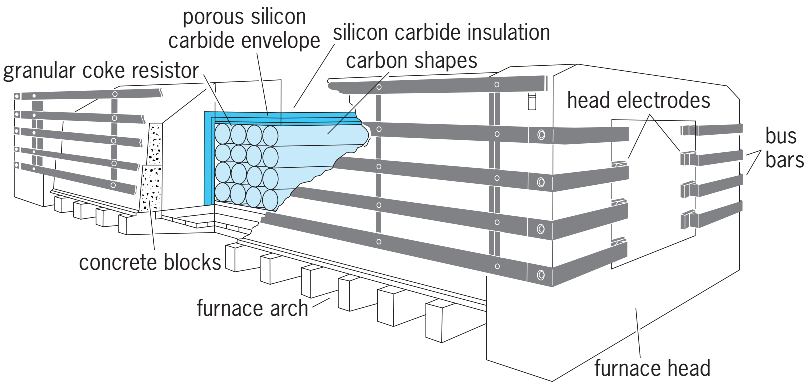


Fig. 4. Graphitizing furnace, for making synthetic graphite.

made of the shapes laid in granular coke (**Fig. 4**). The whole bed is covered by an insulating blanket of silicon carbide, coke, and sand. Higher-density synthetic graphite can be obtained by impregnating the baked carbon with pitch prior to graphitization. Graphite with total ash content less than 20 parts per million is needed for a number of nuclear and electrolytic uses, and is obtained by heating the graphite shapes electrically to about 2500°C (4500°F) while bathing them in a purifying gas. *See* CARBON; COKING (PETROLEUM); ELECTROCHEMICAL PROCESS.

Highly ordered crystalline graphite can be produced up to about $^{1}/_{4}$-in. (0.6-cm) thickness by pyrolyzing organic gases under controlled conditions at 1400–2000°C (2600–3600°F). Pyrolytic graphite exhibits a high degree of anisotropy (varying properties in different directions). Parallel to the thickness the thermal conductivity is comparable to copper, but perpendicular to the thickness conductivity is about 1/200 the conductivity of copper. Tensile strength and thermal expansion also vary greatly with orientation. The room temperature density of pyrolytic graphite reaches up to 2.22, about 98% of the 2.26 density of the graphite single crystal, whereas the density of graphite electrodes ranges from 1.5 to 1.7. Pyrolytic graphite has been formed into rocket nose cones.

Approaching the density limit, very highly oriented graphite can be prepared by stress-annealing pyrolytic carbons. Though polycrystalline, the highly oriented pyrolytic graphite exhibits most of the properties of the ideal graphite crystal and, therefore, has been very useful in studies of the properties of graphite and its intercalation compounds. Because of its high reflectivity, the highly oriented pyrolytic graphite has important applications as monochromators for x-rays and thermal neutrons.

Graphite Fibers

Carbon fibers are filamentary forms of carbon, with a fiber diameter normally in the 6–10-micrometer range. The product is offered in the form of yarns or tows containing from 1000 to 500,000 filaments per strand. The fibers offer a unique combination of properties. They are flexible, lightweight, thermally and to a large extent chemically inert, and are good thermal and electrical conductors. In their high-performance varieties, carbon fibers are very strong and can be extremely stiff. The fibers are made by carbonizing (charring) an organic polymer yarn. Many precursor materials have been proposed, but only three, rayon, polyacrylonitrile (PAN), and pitch, are being used in commercial production. The descriptive terms carbon and graphite are sometimes used to indicate that the material has been carbonized at a temperature below approximately 2500°C (4500°F, carbon) or above (graphite). Quite frequently, however, they are used synonymously. With the exception of some very-high-modulus (over 80,000,000 lb/in.^{2} or 552 gigapascals) pitch-based fibers, the commercial products do not have the three-dimensional order characteristic of graphite, and should properly be called carbon.

Preferred orientation. The most important parameter to characterize carbon fibers is the preferred orientation, that is, the extent to which carbon (or graphite) microcrystallites are oriented parallel to the fiber axis. **Figure 5** shows the Young's modulus E_c, corrected for fiber porosity, as a function of the orientation q. Fibers from all three raw materials fall on the same curve, in excellent agreement with theory. The thermal and electrical conductivity of carbon fibers have a similar increase with greater orientation. The longitudinal (in the direction of the fiber axis) coefficient of thermal expansion in the −100 to +100°C (−148 to +212°F) temperature range becomes slightly negative when the fiber modulus approaches 30,000,000 lb/in.^{2} (207 GPa), and decreases further for fibers with higher preferred orientation. This property is important because it permits the design of composite structures, such as large space antennas and mirrors, having a geometry that is not affected by temperature changes.

The intrinsic tensile strength of carbon fibers also increases with increased preferred orientation. For some fibers, this value probably exceeds 1,000,000 lb/in.^{2} (6.9 GPa). However, the practically

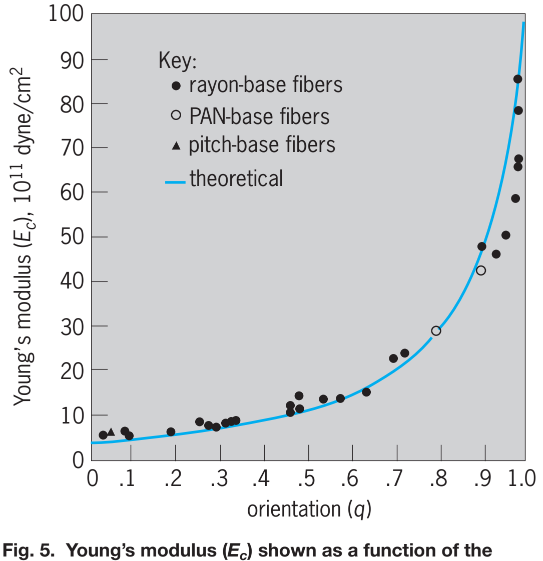


Fig. 5. Young's modulus (E_c) shown as a function of the orientation q; 1 dyne/cm^2 = 0.1 pascal. (*After R. Bacon, Carbon fibers from rayon precursors, Chem. Phys. Carbon, 9:1–102, 1973*)

useful fiber strength depends on the frequency and the severity of strength-limiting flaws such as surface irregularities, holes, or foreign particle inclusions in the fibers, and rarely exceeds 500,000 lb/in.^{2} (3.4 GPa).

Categories. Carbon fibers fall into two categories: low-modulus (under 20,000,000 lb/in.^{2} or 138 GPa) and high-performance (moduli above 25,000,000 lb/in.^{2} or 172 GPa).

Low-modulus fibers. These products are intended primarily for nonstructural applications. They are made from rayon or pitch and are offered as yarns, cloth, felt, and mats. The yarns are impregnated with fluoropolymer resin or graphite powder, braided, and used as packing materials for pumps and valves. Formerly limited to some high-temperature uses where the performance of asbestos packings was inadequate, they are expanding because of environmental concerns about asbestos usage. Rayon-based cloth products are used as reinforcement in resin composites where short-time high-temperature strength and good ablative characteristics are desirable. Cloth carbonized to temperatures above 2100°C (3800°F) is used in rocket nozzle throats and ablation chambers. Low-temperature (under 1800°C or 3300°F) cloth is used where good strength and low thermal conductivity are required, such as in reentry vehicle heat shields, rocket nozzle entrance sections, and exit cones of spacecraft. Rayon-based felts are also used in phenolic matrices as heat shield materials. Other applications are as thermal insulations for high-temperature vacuum furnaces and as the electrical conductor for the sulfur side of sodium/sulfur batteries. Carbon fiber mats are made predominantly from a pitch precursor. The carbon is used in the form of a thin veil mat (50–100 g/m^2 or 0.16–0.32 oz/ft^2) as a surface layer in automotive sheet-molded components (for example, hoods), where it provides sufficient electrical conductivity for electrostatic shielding and electrostatic painting without a conductive primer. Other applications are as thermal insulation similar to carbon felt, and, in a chopped form, as reinforcement for injection moldings to provide high stiffness, good electrical conductivity, and resistance to abrasion. Automotive brake formulations containing short mat fibers as a partial replacement for asbestos are being evaluated.

High-performance fibers. These were originally also made from rayon. To obtain the high modulus and high tensile strength, it was necessary to stretch these fibers to several times their original length at carbonizing temperatures above 2800°C (5100°F). The high cost of this stretching process has made these rayon-based fibers obsolete. High-performance carbon fibers are made from either a PAN or pitch precursor. Polyacrylonitrile is wet- or dry-spun and stretched 500–2000% at approximately 100°C (212°F) to increase the polymer orientation. Ordinary pitch is isotropic and forms a fiber with little or no preferred orientation. By first converting the pitch to a mesophase or liquid crystal state, it is possible to spin very highly oriented pitch fibers whose orientation is maintained and further enhanced during carbonization. After spinning, the PAN and pitch fibers are still thermoplastic and are converted into thermosets by oxidative cross-linking. During this operation, PAN fibers must be kept under tension to maintain the preferred orientation; mesophase pitch fibers can be oxidized without tension at significant savings in process costs. After cross-linking, both types of fibers are carbonized in an inert (nitrogen) atmosphere. The chemical carbon yield from PAN is about 50%, and from mesophase pitch approximately 80%. The carbonized fibers normally receive a surface treatment to enhance the resin-fiber bonding and, hence, the shear strength and transverse tensile

Typical properties of high–modulus carbon yarns and their composites*

Yarns	Units	Thornel† 300	Thornel† 50 (PAN)	Thornel‡ P55S	Thornel‡ 75S
		PAN and pitch			
Tensile strength (10 in. gage length)	10^3 lb/in.^{2}	500	350	300	300
Tensile modulus	10^6 lb/in.^{2}	33.5	55	55	75
Density	g/cm^3, Mg/m^3	1.75	1.81	2.02	2.06
Electrical resistance	μohm-m	18	9.5	7.5	4.6
Thermal conductivity	Cal/s-cm K	.02	.16	.26	.44
		Unidirectional laminates (60% fiber volume in high-performance epoxy)			
Tensile strength (coupon)	10^3 lb/in.^{2}	250	195	150	115
Tensile modulus	10^6 lb/in.^{2}	20	33	34	44
Flexural strength	10^3 lb/in.^{2}	260	190	120	100
Flexural modulus	10^6 lb/in.^{2}	19	29	28	38
Compressive strength	10^3 lb/in.^{2}	230	120	70	60
Compressive modulus	10^6 lb/in.^{2}	20	30	28	35
Short beam shear	10^3 lb/in.^{2}	17	10	5	9
Coefficient of thermal expansion	$\Delta l/l^\circ F \times 10^{-6}$				
Longitudinal $\times 10^{-6}$		$-.3$	$-.4$	$-.5$	$-.7$
Transverse $\times 10^{-6}$		15	15	1515	15

*1 in. = 2.54 cm; 10^3 lb/in.^{2} = 6.895 MPa; 10^6 lb/in.^{2} = 6.895 GPa; $\Delta l/l^\circ C \times 10^{-6} = 1.8 \Delta l/l^\circ F \times 10^{-6}$.

† Typical of intermediate and high-modulus PAN-based fibers. Thornel is a registered trademark of Union Carbide Corporation.

‡ Pitch-based fibers.

strength of the composite. Most, if not all, carbon fibers also receive a thin coating of a polymer finish or sizing for the primary purpose of protecting the brittle fibers from self-abrasion during shipping and handling.

The principal use of high-performance carbon fibers is as the reinforcing component in structural composites. The **table** lists the representative properties of some PAN and pitch-based carbon fibers and of their epoxy matrix composites. Due to initially high cost, the original applications were almost exclusively for lightweight, high-stiffness, and high-strength composites for the aerospace industry.

The second major usage of high-performance carbon fibers is in sporting goods, such as golf club shafts, tennis rackets, fishing rods, and sailboat structures. The major matrix material for both aerospace and sporting goods applications is epoxy; polyimides are used where higher temperature requirements (350–650°F or 177–343°C) exist. Other resins, including polyesters and thermoplastics such as nylon and polyphenylenesulfide, are increasingly used to lower fabrication costs for industrial and automotive applications. Metal matrices, primarily aluminum and magnesium, are used in development programs for space structures such as antennas. These metal matrix composites require the use of extremely high-modulus (over 75,000,000 lb/in.^{2} or 517 GPa) carbon fibers. *See* COMPOSITE MATERIAL; MANUFACTURED FIBER; METAL MATRIX COMPOSITE.

H. F. Volk

Bibliography. D. B. Chung, *Carbon Fiber Composites*, 1994; J. Delmonte, *Technology of Carbon and Graphite Fiber Composites*, 1981, reprint 1987; M. S. Dresselhaus et al. (eds.), *Graphite Fibers and Filaments*, 1988; L. H. Peebles, Jr., *Carbon Fibers: Information, Structure, and Properties*, 1994.

Graptolithina

A group of marine organisms that were common in the early Paleozoic (the Late Cambrian to Early Devonian periods). Graptolites became extinct in the late Paleozoic Era. They were minute animals that built communal skeletons. Each graptolite colony contained from two to many hundreds of separate graptolite animals (referred to individually as zooids). All of the zooids within a single colony (a rhabdosome) were formed by asexual budding from the founder zooid, which probably grew from a fertilized egg. Thus, each colony started with a sexually produced animal and then enlarged by budding new zooids from one another. The result was a spreading rhabdosome composed of a few to many hundreds of minute tubes (thecae), each containing its own zooid (**Fig. 1**). Graptolite thecae exhibit a range of diameters (30 μm–2 mm), but are most commonly about 0.5–1.0 mm in diameter and several times this length. The thecae consist of the protein collagen, which is organized into two main building blocks. The most prominent are the fuselli, which have the form of narrow half-rings stacked one upon the other like a set of semicircular bricks. Collectively, the fuselli formed the main substance of the thecal tube. The second building block was deposited on the outside of the theca in the form of bandagelike strips or thin continuous sheets (**Fig. 2**). Also built of collagen fibers, this second layer (the cortex) is much thinner, but probably added greatly to the strength of the theca.

With the exception of a small group of living organisms (for example, *Rhabdopleura* of the class Pterobranchia; Fig. 1) that may be relatives of the graptolites, little is known for sure of the detailed anatomy and ecology of these enigmatic organisms. Although graptolites are common as fossils in dark shales that were deposited under conditions of low oxygen concentration, the original soft body parts are not preserved—only the empty tubes remain. Thus, many basic features (such as exactly how they built their colonies) are still subjects of debate among specialists. Most specialists think that graptolites and pterobranchs are closely related to one another, and that graptolites employed a specialized organ similar to that used by *Rhabdopleura* zooids to secrete the fuselli and cortex from their body (externally) in a mortaring fashion (**Fig. 3**). A smaller group of specialists interpret the evidence very differently. They suggest that the thecae were formed internally beneath an enveloping tissue layer by zooids unlike those of *Rhabdopleura*, to which graptolites, therefore, would not be closely related.

Based on the shape and patterns of occurrence of their colonies, graptolites were probably suspension feeders that extracted food particles such as bacteria, algae, and other organic matter from the surrounding ocean water. Many graptolite species were very widely distributed throughout the world's oceans during the Ordovician, Silurian, and Early Devonian periods (an interval of about 110 million years in total). Graptolites evolved very rapidly. Most species occupied a relatively short segment of this portion of the Earth's history. The average duration for individual graptoloid graptolite species was approximately a million years. As a result of this combination of wide occurrence and short duration, graptolite species make very useful index fossils. For example, finding specimens of a particular species in both China and Australia indicates that the rocks containing these two fossils were formed during the same short time of the Earth's history when this particular species lived. Using relations like these, paleontologists and geologists have untangled the complex history of the Earth. *See* DEVONIAN; FACIES (GEOLOGY); INDEX FOSSIL; ORDOVICIAN; SILURIAN.

Types. Graptolites are mainly of two sorts: bottom-living (benthic) and free-floating (planktic). Benthic graptolites generally either formed horizontal, sheet-like colonies attached to the sea floor and other hard objects like shells, or grew as upright, bushy or cone-shaped colonies. Some of these erect benthic colonies became relatively large, longer than 10–15 cm (4–6 in.) from base to tip. Many of these benthic graptolites are grouped by graptolite taxonomists as the Dendroidea. Benthic graptolites also include several other graptolite orders (such as the Tuboidea); however, these are very small and very rare fossils. Dendroids are characterized by long, tubular thecae of two distinctly different sizes that are arranged in pairs within their branches. The larger thecae are known as autothecae, and the smaller as bithecae. Thecae within the dendroid colony were interconnected through a special chord called the stolon, which probably permitted coordinated activity and the sharing of nutrients among zooids. Benthic graptolites are less common than planktic graptolites and appear to be of limited value as index fossils. Thus, they are less well studied and less well understood than the planktic graptolites. For instance, it is unclear why the thecae have a regular division into autothecae and bithecae. It may reflect some division of labor among different-sized zooids, or sexual dimorphism, or possibly some other factor. Dendroids first appeared in the Late Cambrian Period (around the same time as most other marine invertebrate animals, including arthropods, brachiopods, mollusks, and echinoderms) and became extinct in the Carboniferous Period.

In contrast to the benthic graptolites, species of the Graptoloidea were planktic. The founding individual of the colony in these graptolites did not settle to the ocean floor to complete its maturation as did the founders of dendroid rhabdosomes (Fig. 1*a*). Instead, a graptoloid larva matured while in the ocean currents, and there it initiated formation of its rhabdosome. This change in habit was accompanied by a corresponding change in the form of the tube constructed by this zooid (the sicula). The sicula of graptoloid graptolites is conical with a fine threadlike extension (nema) on the closed end (Fig. 1*b*). The earliest planktic graptolites retained many of the features of their dendroid ancestors, including the many-branched colony form and differentiation of the thecae into autothecae and bithecae. Thus, these early graptoloids are not easily distinguished from some dendroid species except by the form of the sicula and their planktic habit. The evolutionary appearance of planktic graptolites is one of the key events that marks the beginning of the Ordovician Period.

Early Ordovician graptoloids evolved simpler colonies over the next several tens of millions of years. Their rhabdosomes came to possess fewer branches and to consist of only large thecae comparable to the autothecae of dendroids. Whatever purpose bithecae served in benthic graptolites, they evidently were not beneficial to planktic graptolites. These changes, which resulted in lighter, more streamlined, and possibly more mobile rhabdosomes, are presumed to be adaptations to life in the plankton, but once again the exact nature of the advance afforded by these changes is the subject of continued study.

During the Ordovician the evolutionarily advanced (but simpler) graptoloids diversified into a great many different species with a wide range of colony forms and sizes (**Fig. 4**). At the same time that the overall rhabdosome form became simpler, the form of the individual zooid tubes (thecae) became much more varied and complex. Some graptoloids exhibit hooked, sharply flexed, hooded, or spiny forms. In some species the rhabdosome was reduced to a scaffoldinglike set of rods without any continuous skeletal material between them. The colony form of graptoloids shows numerous evolutionary trends, but overall quite different forms dominate different intervals during the Ordovician to Early Devonian. Graptoloids became extinct in the Early Devonian.

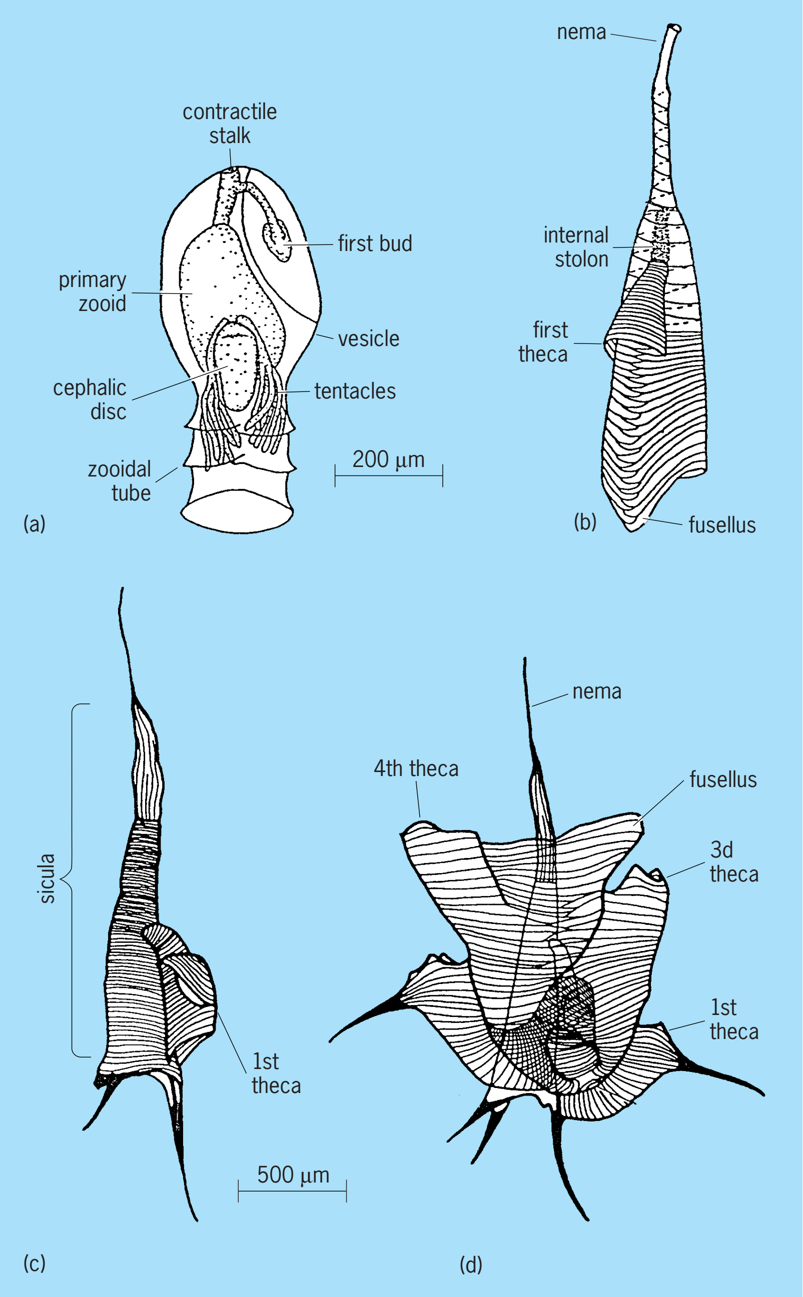


Fig. 1. Sketches of portions of pterobranch and graptolite colonies showing the dwelling tubes and the constructional units of the graptolite skeleton. (*a*) Young *Rhabdopleura compacta* colony with founder zooid and first asexual bud (Recent). Benthic graptolites may have been similar to this pterobranch. (*b*) Comparable graptoloid graptolite sicula (Early Ordovician). (*c, d*) Later growth stages of the more derived graptoloid *Hustedograptus uplandicus* (Middle Ordovician).

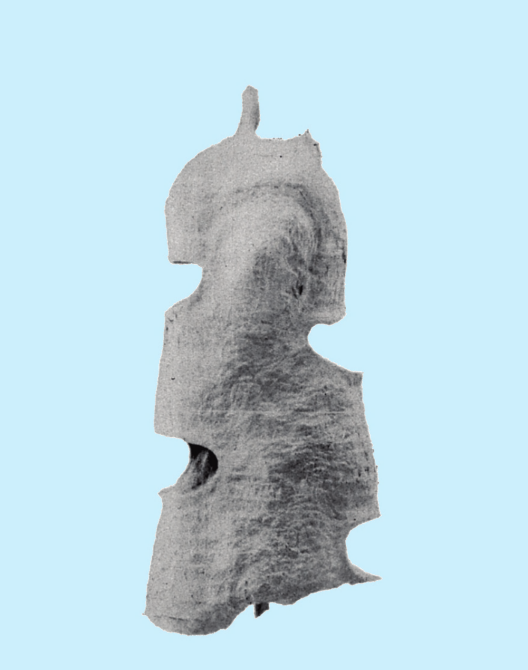


Fig. 2. Scanning electron micrograph of a graptolite (*Pseudoclimacograptus scharengergi*) showing the fuselli (horizontal ridges) that compose graptolite thecae, and the cortical bandages (oblique strips) that were deposited on top of the fuselli (Middle Ordovician); specimen is approximately 1 mm in width.

Preservation. Graptoloids lived mostly in oxygen-minimum zone waters at shelf margins. Thus, they are most common in rocks deposited in outer-shelf and slope environments. Most graptoloid graptolites are preserved in gray to black shales as black or silvery films flattened on the surface of the rock. Rough handling or careless washing of the rock may completely remove these delicate fossils. Occasionally, the mineral pyrite may form within the hollow graptolite rhabdosomes prior to compression of the original mud into shale, and an exquisite three-dimensional fossil may result. Rarely, graptolites may be preserved in limestones. Often graptolites in these rocks are preserved uncompressed, and with the use of hydrochloric or acetic acid, they can sometimes be liberated from their limestone matrix (Fig. 2). Most understanding of graptolites' detailed structure comes from material of this kind.

Life history and extinction. The evolutionary history of planktic graptolites was complex, and studies of their patterns of morphological change and fluctuations in species diversity have provided important examples for evolutionary paleobiology about the nature of adaptation to this unusual lifestyle and to the causes of changes in organic diversity in the world's oceans. For instance, the biosphere experienced a catastrophic decline in species number among a wide array of marine invertebrates, but especially among trilobites, brachiopods, conodonts, corals, and graptolites near the end of the Ordovician Period. This mass extinction has been estimated to have been the second or third most severe extinction event in the history of marine metazoans. During this event, planktic graptolites almost became entirely extinct.

Many planktic graptolite species became extinct during the Late Ordovician due to environmental changes that have been linked to global climate changes. Environmental changes in the oceans included a sea-level fall of as much as 100 m (330 ft) and the loss of many oceanic upwelling sites located near continental shelf margins. These changes in ocean conditions developed as a consequence of global climate change from an ice-free, "greenhouse" world climate to an "icehouse" global climate lasting approximately 1 million years during which ice covered the South Pole. Lowered sea level during glaciation drained most marine environments that were spread broadly across shelves of the time. When the ice melted, global climates and marine environments returned to preglaciation conditions.

The majority of preglaciation Late Ordovician graptolites lived in hypoxic (oxygen-deficient) waters, most of which developed in oxygen-minimum zones under ocean upwelling waters at shelf margins. Some planktic graptolites lived in hypoxic environments in basins within shelves. Locally endemic species occur in a number of these shelf basins. A few

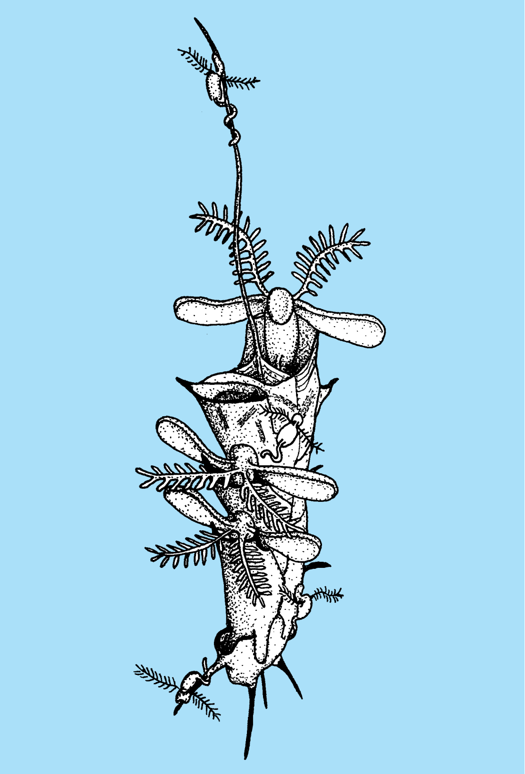


Fig. 3. Hypothetical form of zooids that inhabited colony tubes, based on a planktic graptolite. Large zooids are shown with possible oarlike swimming organs used to move the colony through the ocean water and tentacles for feeding. Also shown are two possible modes of graptolite colony construction: large zooids linked together built the whole colony, or smaller free-roaming zooids may have helped the larger zooids by adding to the cortex and spines. A third hypothesis (not shown) involves an entirely different zoid reconstruction (see text).

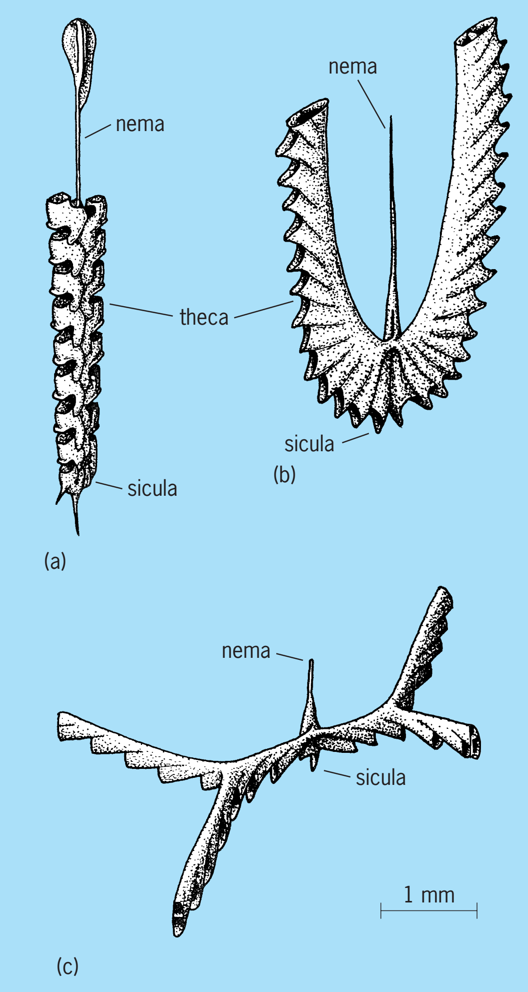


Fig. 4. Examples of some planktic graptolites, reconstructed to show three-dimensional form.

Late Ordovician graptolites (primarily those called normalograptids) lived in relatively oxic (oxygen-containing) waters near or at the ocean surface. Some of these species occur in both shelf basin and shelf marginal rock sequences.

Hypoxic environments preferred by many graptolites were lost during glaciation as a consequence of the draining of shelf basins and the loss of shelf margin oxygen-minimum zones. Most planktic graptolites living in hypoxic environments became extinct during the glacial episode. In contrast, those graptolites living in oxic waters thrived during glaciation. They even radiated modestly with a number of new species appearing in certain areas.

A Late Ordovician–Early Silurian planktic graptolite refugium (an area that has escaped the great changes that occurred in the region as a whole, often providing conditions in which relic colonies can survive) developed in a shelf basin (Yangtze Basin) in modern south China. Freshwater runoff from surrounding land areas flowed into that basin. Because freshwater is less dense than the deeper basinal waters, they most likely served as a seal that enabled the deep, hypoxic basin waters to be density-stratified. Preglaciation Late Ordovician planktic graptolites living in hypoxic waters beneath the freshwater seal of that basin survived through the interval during which other graptolites living in sites where hypoxic waters were diminished or had disappeared became extinct. Turnover among species and rates of extinction of species living in this basin remained relatively similar to preclimate change conditions throughout the climate change interval.

Deglaciation and resultant sea-level rise gradually led to a return of the preglaciation ocean conditions. Many new graptolite species appeared during the early phases of sea-level rise and redevelopment of ocean upwelling sites with oxygen-minimum zones and shallow shelf basins in which hypoxic conditions redeveloped.

Viruses, bacteria, and other microorganisms as well as floating larvae of many organisms are abundant in many modern hypoxic environments. Molecular biologists have discovered an extensive system of gene exchanges mediated by plasmids, episomes, and viruses among bacteria. Viruses in hypoxic environments may invade bacteria and obtain a portion of the host genetic material. When the host wall ruptures, viruses containing new genetic information may spread out to infect another host and transfer genetic information from the old to new host.

Late Ordovician graptolites living in hypoxic to anoxic (oxygen-absent) waters likely became crowded together in small enclaves as their environments diminished during the glaciation-related sea-level fall and the loss or near-loss of the hypoxic environments in which they lived. Thus, viral transfer of genetic information could have taken place between graptolite larvae in Late Ordovician diminished hypoxic environments. Lateral transfer of genetic information by viruses, if it did take place, could have been one mechanism for the relatively rapid species originations seen in strata deposited during post-glacial sea-level rise and redevelopment of relatively broader hypoxic environments.

Although planktic graptolites radiated relatively rapidly to a number of lineages during the early part of the Silurian, they appear to have undergone as many as eight significant reductions in species number during the Silurian. During one event that took place in the mid-Silurian (Late Wenlock), planktic graptolites almost became extinct. Nearly all of the survivors of this near-extinction lived in oxic waters. Some of the other severe reductions in graptolite taxa that took place during the Silurian have been linked to climate-induced sea-level changes and related environmental changes. The majority of pre-Late Wenlock Silurian planktic graptolites appear to have lived in or near hypoxic waters. The marked reductions in planktic graptolite diversity at intervals within the pre-Late Wenlock Silurian appear to coincide with diminished hypoxic environments in a pattern similar to that of the Late Ordovician extinction.

The majority of post-Late Wenlock graptolites probably lived in oxic waters where they no longer had the protection from predators afforded by life in hypoxic environments. Many taxa appear to have lived in waters over the shelves of the time.

Graptolite species diversity diminished significantly during the Late Silurian into the Early Devonian. Finally, planktic graptolites became extinct late in the early part of the Devonian. *See* EXTINCTION (BIOLOGY). William B. N. Berry; Charles E. Mitchell

Bibliography. D. E. B. Bates, Graptolites: Strange plankton of the past, *Endeavor, New Series*, 13:54–62, 1989; W. B. N. Berry and H. Hartman, Graptolite parallel evolution and lateral gene transfer, pp. 397–404 in M. Syvanen and C. L. Kado (eds.), *Horizontal Gene Transfer*, 2d ed., Academic Press, San Diego, 2002; W. B. N. Berry, Late Ordovician climate change: Impact on graptolite biodiversity. *Earth Systems Processes 2, Abstracts with Programs* 55, 2005; O. M. B. Bullman, in R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, Pt. V: *Graptolithina*, Kansas University Press, 1970; P. R. Crowther, The Fine Structure of Graptolite Periderm, *Spec. Pap. Palaeont.*, no. 26, 1981; S. C. Finney and W. B. N. Berry, New perspectives on graptolite distributions and their use as indicators of platform margin dynamics, *Geology*, 25:919–922, 1997; A. Hallam and P. B. Wingnall, *Mass Extinctions and Their Aftermath*, Oxford University Press, 1997; C. E. Mitchell et al., Was the Yangtze Platform a refugium for graptolites during the Hirnantian (Late Ordovician) mass extinction?, pp. 523–526 in G. L. Albanese, M. S. Beresi, and S. H. Peralta (eds.), *Ordovician from the Andes*, Instituto Superior de Correlacion Geologica (INSUEGO) Serie Correlacion Geologica 17, 2003; R. B. Rickards, Palaeoecology of the Graptolithina, an extinct class of the phylum Hemichordata, *Biol. Rev.*, 50:397–436, 1975; P. Storch, Biotic crises and post-crisis recoveries recorded by Silurian planktonic graptolite faunas of the Barrandian area (Czech Republic), *Geolines*, 3:59–70, 1995.

Grass crops

Members of the family Gramineae cultivated as forage and grain for consumption. The grasses are the most useful of all the plants that cover the Earth. The cereal grasses (rice, wheat, maize, rye, barley, oats, sorghum, and the millets) supply directly three-fourths of the energy and over half of the protein in food consumed by humans. Indirectly, these cereals together with the forage grasses supply most of the food for the domestic animals that provide milk, meat, eggs, and much of the draft power required to grow crops. Deer, antelope, rabbits, and many other wild game depend on grasses for much of their sustenance. Sugarcane produces more than half of the world supply of sugar. Starch, and most of the alcohol for beverage and industrial uses, comes from the cereal grasses. *See* BARLEY; CEREAL; CORN; MILLET; OATS; RICE; RYE; SORGHUM; SUGARCANE; WHEAT.

The bamboos are of vast importance in the Indo-Malay region, where they are used in building houses, bridges, furniture, rafts, water pipes, vessels for holding water, and so forth. Slender bamboo stems make excellent fishing poles, whereas the giant bamboos, measuring up to 12 in. (30 cm) in diameter and over 100 ft (30 m) high, are important substitutes for wood. *See* BAMBOO.

The grasses protect soil from erosion and help conserve water resources. More than any other family of plants, the sod-forming grasses blanket golf courses, athletic fields, lawns, parks, and cemeteries with a protective covering that beautifies and enhances the environment. No other family of plants in the vast plant kingdom is so useful to humans. *See* EROSION; LAWN AND TURF GRASSES; SOIL CONSERVATION.

Structure. Grass stems have solid joints (nodes) and leaves arranged in two rows, with one leaf at each joint (**Fig. 1**). The leaves consist of the sheath, which fits around the stem like a split tube, and the blade, which is commonly long and narrow. Seed heads are made up of minute flowers on tiny branchlets, often several crowded together, but always two-ranked like the leaves. The flowers are generally wind-pollinated. The seeds are enclosed between two bracts, or glumes, which remain on the seed when ripe.

Growth characteristics. The 600 genera grouped into 14 tribes that make up the grass family may be annual or perennial. In the tropics, elephant grass (*Pennisetum purpureum*) can produce up to 37 tons of dry matter per acre (75 tons per hectare per year) whereas Tifdwarf Bermuda grass (*Cynodon* spp.) on a golf green will produce only a few pounds of matter in the same period. Annuals and some perennial grasses are bunch grasses which spread only by seeding. Others, mostly perennials, also spread by creeping stems called stolons when above ground and rhizomes when below the soil surface. Stoloniferous and rhizomatous species usually are most tolerant of continuous defoliation by animals. The creeping grasses form the best sods and surpass others for soil conservation; they are also the best turf grasses. All grasses have fine fibrous root systems that permeate the soil extensively to depths ranging from much less than 3 ft to more than 10 ft (1 to 3 m). The roots are short-lived, are continually being replaced, and in the process increase the organic matter content of the soil.

Distribution. Grasses are distributed throughout the world. A few profuse seeders with a very short life-cycle such as annual bluegrass (*Poa annua*) can be found growing from the Equator to the Arctic Circle. Annual species predominate in the adverse environments found in the deserts and the arctic areas. Temperature is the principal factor determining the distribution of perennial grasses. Most tropical grasses are killed when temperatures drop much below 32°F (0°C), whereas temperate perennial grasses can tolerate much lower temperatures. However, there is a great deal of overlapping of tropical and temperate grasses that may differ in temperature hardiness among varieties within some species. *See* BERMUDA GRASS.

Perennial grasses are frequently classed as cool- or warm-season grasses depending upon the season in which they make their best growth. Generally, cool-season grasses such as bluegrass (*Poa pratense*) are also temperate grasses and fail to survive the long hot

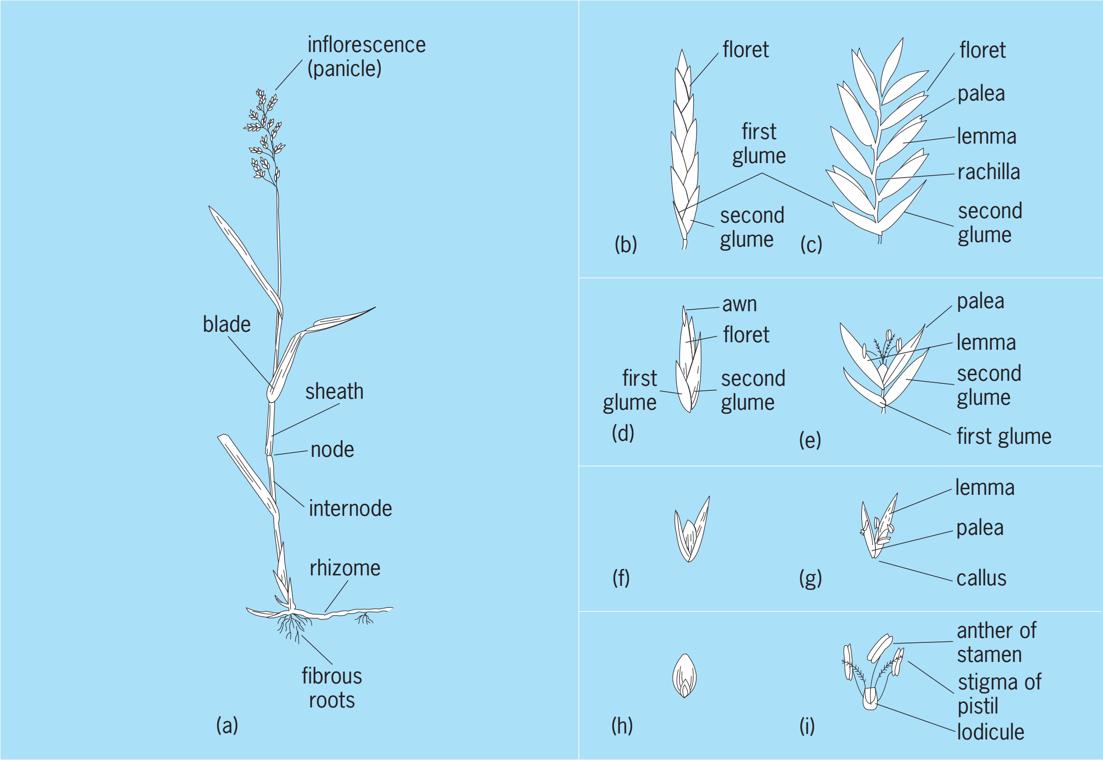


Fig. 1. Parts of a typical grass plant. (*a*) Complete plant. (*b*) Many-flowered spikelet. (*c*) Generalized spikelet. (*d*) One-flowered spikelet. (*e*) Spikelet at flowering. (*f*) Floret in fruit. (*g*) Floret in flower. (*h*) A fruit. (*i*) A flower. (*After P. D. Strausbaugh and E. L. Core, Flora of West Virginia, West Va. Univ. Bull., ser. 52, no. 12–2, pt. 1, p. 67, 1952*)

summers in the tropics. Usually diseases contribute to their death. Warm-season grasses taken into the temperate zone usually die during winter. *See* BLUEGRASS.

Some 10% of the 1400 or more species of grasses found in the United States originated in other countries. Many of the weedy ones were introduced accidentally with seed or feed brought in with livestock. Johnson grass, one of the South's worst weeds, was introduced from the Middle East as a forage grass for cattle. Pensacola bahia grass (*Paspalum notatum*) came to Pensacola, Florida, in the digestive tracts of cattle shipped from Santa Fe, Argentina, in the early 1900s. Today Pensacola bahia grass is widely grown in pastures and roadsides in Florida and the southern half of the states bordering the Gulf of Mexico. Most of the cool-season grasses planted in the United States for forage and turf originated in Europe. Among these, the bluegrasses (*Poa* spp.), fescues (*Festuca* spp.), bents (*Agrostic* spp.), cocksfoots (*Dactylis* spp.), and timothy (*Phleum pratense*) are the most common. Africa and South America are the original homes of many of the warm-season grasses such as the Bermuda (*Cynodon* spp.), the bahia and dallis grasses (*Paspalum* spp.), and Pangola grass (*Digitaria* spp.) that are planted in the southern United States. *See* BERMUDA GRASS; TIMOTHY.

Reproduction and propagation. Annuals and most perennial grasses reproduce sexually and are propagated by seed. Many of the grasses are cross-pollinated largely by the wind. In some species, cross-pollination is facilitated by self-incompatibility that occurs at variable frequencies. *See* POLLINATION.

Most grasses produce perfect flowers containing both male and female organs. The male organs (anthers) must be carefully removed before they shed pollen to make controlled hybrids. In maize, the male (in the tassel) and the female (in the ear) organs are separated, a characteristic that greatly facilitates hybrid production. Controlled hybridization without emasculation in the *Pennisetum* spp. is made possible by their characteristic exsertion of the female pollen-receptive stigmas at least 24 h before their pollen-shedding anthers. Cytoplasmic male sterility has been discovered in a few species and is used to produce commercial F_1 hybrid seed of maize, sorghum, and pearl millet. *See* FLOWER.

A few species, largely tropical perennials, reproduce by apomixis, simply defined as vegetative reproduction through the seed. Apomictic seeds produce the same genotype as the plant that produced them. An obligate apomict such as tetraploid bahia grass rarely produces a sexual seed, whereas facultative apomicts such as Kentucky bluegrass produce both sexual and apomictic seeds. Apomixis permits the increase and use of superior genotypes with the added advantage of seed propagation. Some perennial bunch grasses, such as elephant grass and sugarcane, are important enough to warrant the added cost of vegetative propagation from stem cuttings. A few superior genotypes of grasses that spread rapidly by stolons or rhizomes are propagated vegetatively. Usually these grasses are sterile or produce too few seeds to permit seed propagation. They are also heterozygous F_1 hybrids that would lose most of their superiority if they could be propagated by seed. Examples are Pangola grass and Coastal and Midland Bermuda grasses that have been planted on several million acres. The sterile interspecific Bermuda grass hybrids developed for turf and widely grown across the southern United States are planted vegetatively. Machines designed to harvest and plant vegetative parts of these grasses greatly facilitate vegetative planting. *See* PLANT PROPAGATION; REPRODUCTION (PLANT).

Nutritive value. Cereal grasses are excellent sources of carbohydrates, but tend to be low in protein and deficient in lysine when used as a sole source of food for humans and animals. Protein and lysine content in cereal grain has been improved by breeding for these traits. The quality of the young foliage of forage grasses is usually adequate for ruminants. As the leaves and stems age, digestibility and protein content decreases, indigestible lignin increases, and the performance of animals consuming them declines. The protein and mineral content of grasses growing on most soils can be increased by fertilizing the soil with nitrogen and minerals. The nutritive value of forages can be improved by breeding. *See* BREEDING (PLANT); GRAIN CROPS; GRASSLAND ECOSYSTEM. Glenn W. Burton

Diseases. Although grasses are the oldest agricultural crop, it has only been in recent years that the diagnosis and control of their diseases have received major attention. Of the 95 diseases known to occur on the various cultivated grass species, not all affect the same grass, nor do they all occur in the same region of the world. Kentucky bluegrass (*Poa pratensis*), for example, is susceptible to 52 known diseases, while tall fescue (*Festuca arundinacea*) is subject to only 6. With each of the cultivated grass species, however, there is at least one known disease for each season of the year.

Pathogens. Fungi are the most common of the biotic agents that cause diseases of grasses. The types of pathogenic fungi range from those that infect and remain primarily inside the plant tissue and are visible only with the aid of a microscope, to those that can be seen with the naked eye as they grow freely over the surface of the leaves, to the highly conspicuous mushrooms and toadstools that cause fairy rings. Fungi that parasitize the leaves of grasses cause the most conspicuous symptoms. Depending on the disease, these symptoms vary from distinct leaf spots to indiscriminate blotches to complete blighting of the entire leaf (**Figs. 2** and **3**). Many of these fungi can also cause destructive rots of the root systems of the same plants. Several of the fungal diseases are very severe, and can cause significant reductions in yields and nutritive value of harvested forage, the longevity of the productive life of the stand of forage or pasture grass, or the landscape or recreational value of turfgrass.

Next in importance as grass pathogens are nematodes. While some species attack the floral tissue of grass plants, the most important parasitic nematodes are those that feed on the roots. When nematode populations are high, their feeding activity destroys large amounts of root tissue. As a result, the plants recover more slowly from mowing or grazing. Also, the overall yield of forage is reduced, and plant resistance to various environmental and use stresses is lowered.

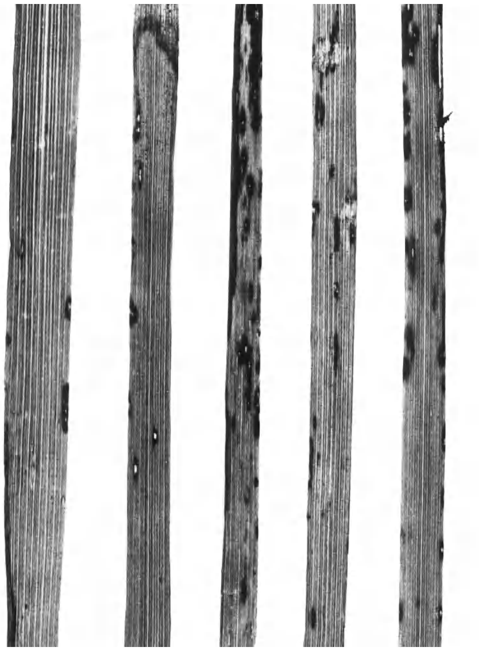


Fig. 2. Cerospora leaf spot of tall fescue, an example of a leaf spot-producing disease. (*From H. B. Couch, Diseases of Turfgrasses, 2d ed., Robert Krieger Publishing Co., 1973*)

Bacteria, mycoplasmas, spiroplasmas, and viruses also cause diseases of grasses. In some instances these diseases are of major importance, but collectively the total impact on forage yield or turfgrass quality is not as great as those of fungi and nematodes.

Control measures. Control of grass diseases is achieved by the use of disease-resistant varieties

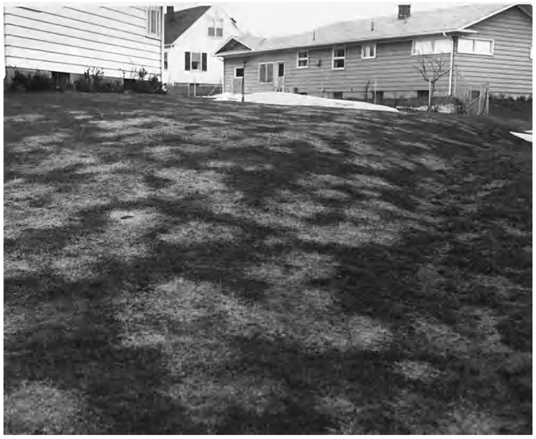


Fig. 3. Fusarium patch of Kentucky bluegrass, a winter disease of grasses and an example of a leaf-blighting disease. (*From H. B. Couch, Diseases of Turfgrasses, 2d ed., Robert Krieger Publishing Co., 1973*)

(especially with grasses grown for pasture and harvested forage) and by the application of pesticides. Many grass varieties have both multiple resistance to several pathogens and high growth potential. In general, it is easier to develop varieties resistant to foliar diseases than to obtain resistance to diseases of the root system. Diseases that are caused by fungi with a high capacity for hybridization, for example, the rusts and powdery mildews, are also difficult to control by the development of resistant varieties. In these instances it is not uncommon for new races of the pathogen to appear that are highly pathogenic to a variety that was formerly resistant to the disease. Application of fungicides and nematicides for the control of grass diseases is generally considered practical only in the cultivation of landscape and recreation turf. Commercially available pre- and postplant nematicides are very effective in reducing the populations of soil-inhabiting nematodes that parasitize grass roots. Also, there are several commercial fungicides that provide high levels of control of parasitic fungi on the leaves of grasses. Both nematicides and fungicides are formulated as granules for application with a spreader, and as powders and liquids for use with sprayers. *See* FUNGISTAT AND FUNGICIDE; PESTICIDE; PLANT PATHOLOGY. H. B. Couch

Bibliography. H. B. Couch, *Diseases of Turfgrasses*, 3d ed., 1995; A. S. Hitchcock, *Manual of Grasses of the United States*, 2d ed., 1983; M. B. Jones and A. Lazenby, *The Grass Crop: The Physiological Basis of Production*, 1988; H. B. Sprague (ed.), *Grasslands of the United States: Their Economic and Ecological Significance*, 1974.

Grassland ecosystem

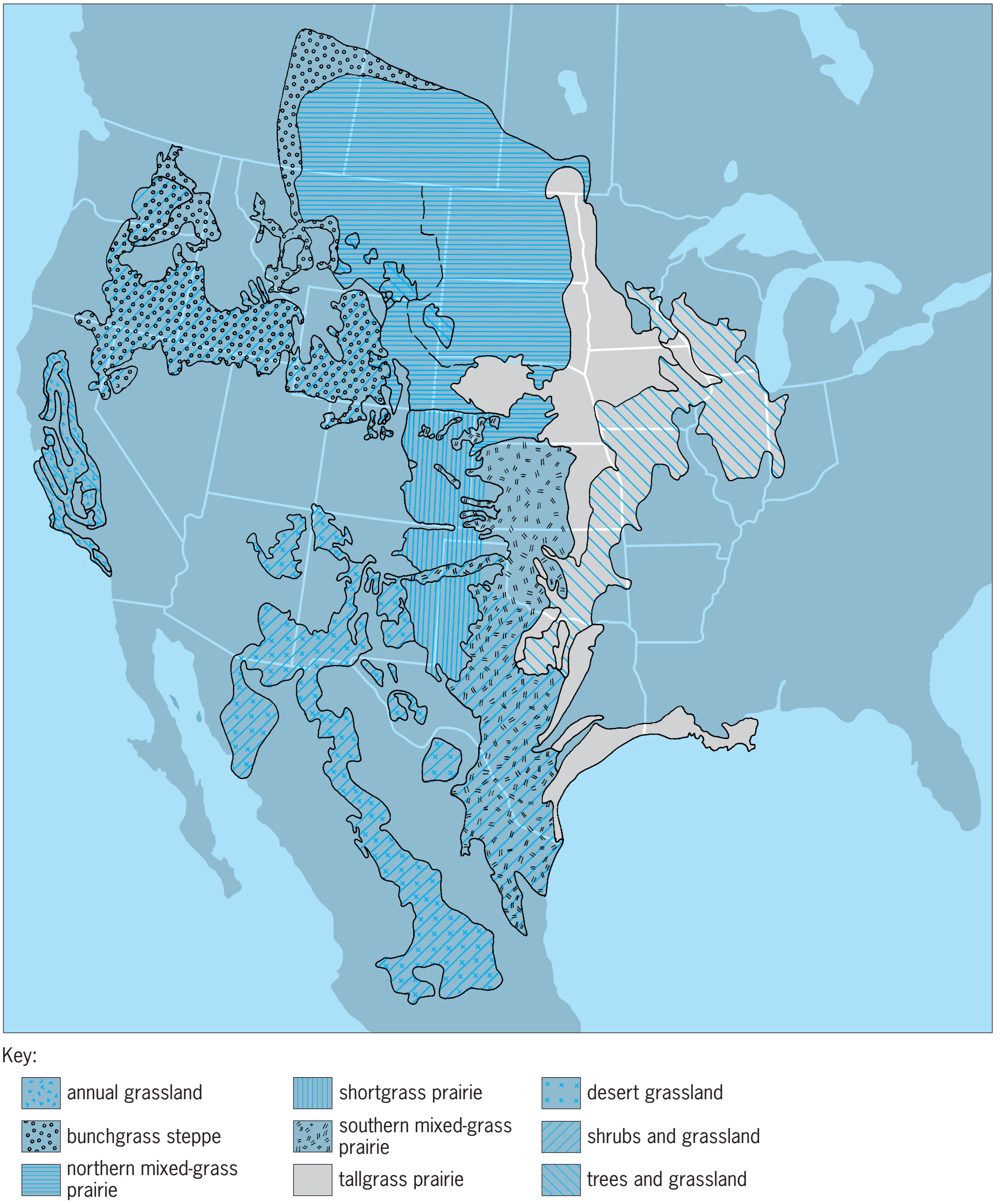
A Biological community that contains few trees or shrubs, is characterized by mixed herbaceous (nonwoody) vegetation cover, and is dominated by grasses or grasslike plants. Grassland ecosystems range from the dense bamboo of the Amazonian tropics to the cool northern steppes of Russia, and from dry plains in the western United States to Canadian arctic grasslands. Mixtures of trees and grasslands occur as savannas at transition zones with forests, as in the east-central United States, or where rainfall is marginal for trees, such as in south-central Africa and Australia. About 1.2 × 10^8 mi^2 (4.6 × 10^7 km^2) of the Earth's surface is covered with grasslands, which make up about 32% of the plant cover of the world. The proportion of original grasslands varies widely among continents with about 44% remaining in Europe and less than 10% in Australia, although the latter has vast expanses of savannas. In North America, grasslands include the Great Plains, which extend from southern Texas into Canada. The European meadows cross the subcontinent, and the Eurasian steppe ranges from Hungary eastward through Russia to Mongolia; the pampas cover much of the interior of Argentina and Uruguay. Vast and varied savannas and velds can be found in central and southern Africa and throughout much of Australia. *See* SAVANNA.

Most civilizations have developed in grassland and savanna regions, and were it not for the abundance and widespread distribution of grasses, the human population of the world would not have attained its present level. Significant portions of the world's grasslands have been modified by grazing or tillage or have been converted to other uses. The most fertile and productive soils in the world have developed under grassland, and in many cases the natural species have been replaced by cultivated grasses (cereals). *See* CEREAL; GRASS CROPS.

Grasslands occur in regions that are too dry for forests but that have sufficient soil water to support a closed herbaceous plant canopy that is lacking in deserts. Thus, temperate grasslands usually develop in areas with 10–40 in. (25–100 cm) of annual precipitation, although tropical grasslands may receive up to 60 in. (150 cm). Grasslands are found primarily on plains or rolling topography in the interiors of great land masses, and from sea level to elevations of nearly 16,400 ft (5000 m) in the Andes. Because of their continental location they experience large differences in seasonal climate and wide ranges in diurnal conditions. In general, there is at least one dry season during the year, and drought conditions occur periodically. *See* DROUGHT; PLANT-WATER RELATIONS.

Classification. Different kinds of grasslands develop within continents, and their classification is based on similarity of dominant vegetation, presence or absence of specific dominant species, or prevailing climate conditions (see **illus.**). In North America, the tallgrass prairie lies between the eastern deciduous forest and the Central Plains. Annual rainfall is 30–40 in. (75–100 cm), and under optimum conditions the dominant bluestem (*Andropogon*) grass species may exceed a height of 6 ft (2 m). The mixed-grass prairie contains species of the tallgrass prairie and the shortgrass prairie, which dominates farther west. Annual precipitation amounts range from 20 to 30 in. (50 to 75 cm), and the height of the dominant species has a similar range. Shortgrass prairive precipitation amounts of 10–20 in. (25–50 cm) and are dominated by grasses adapted to dry conditions, such as grama grasses (*Bouteloua*) and buffalo grass (*Buchloe dactyloides*). Desert grasslands have annual rainfalls of 10–18 in. (25–45 cm), which usually fall in summer (July–August) and winter (December–January). Between those wet periods, the desert grasslands are subjected to extreme drought.

Other North American grasslands include the annual grassland in the central valley of California. Originally, these grasslands were dominated by perennial species, but with grazing pressure most of them have been replaced by native and exotic annual species that are more resistant to grazing. Mountain grasslands occupy higher elevations and are frequently mixed in a parklike appearance with trees up to elevations of 8200 ft (2500 m). Shrub steppe grassland vegetation can be found in the northwestern



Map of grassland types in central and western North America.

United States and includes many herbaceous grassland species in combination with woody species such as sagebrush (*Artemisia*).

Climate. The climate of grasslands is one of daily and seasonal extremes. Deep winter cold does not preclude grasslands since they occur in some of the coldest regions of the world. However, the success of grasslands in the Mediterranean climate shows that marked summer drought is not prohibitive either. In North America, the rainfall gradient decreases from an annual precipitation of about 40 in. (100 cm) along the eastern border of the tallgrass prairie at the deciduous forest to only about 8 in. (20 cm) in the shortgrass prairies at the foothills of the Rocky Mountains. A similar pattern exists in Europe, with the taller grasses along the northwestern coast and decreasing plant height and rainfall toward the south and east into the plains of Hungary. Growing-season length is determined by temperature in the north latitudes and by available soil moisture in many regions, especially those adjacent to deserts. Plants are frequently subjected to hot and dry weather conditions, which are often exacerbated by windy conditions that increase transpirational water loss from the plant leaves.

Soils. Soils of mesic temperate grasslands are usually deep, about 3 ft (1 m), are neutral to basic, have high amounts of organic matter, contain large amounts of exchangeable bases, and are highly fertile, with well-developed profiles. The soils are rich because rainfall is inadequate for excessive leaching of minerals and because plant roots produce large amounts of organic material. Humus, partially decomposed organic material that may constitute up to 10% of the soil, expands its capability to retain water by as much as 20% and binds soil particles into clumps, increasing the effectiveness of the soil to make nutrients and water available to the plants. With less rainfall, grassland soils are shallow, contain less organic matter, frequently are lighter colored, and may be more basic. Tropical and subtropical soils are highly leached, have lower amounts of organic material because of rapid decomposition and more leaching from higher rainfall, and are frequently red to yellow. Almost all true grasslands soils are classified as Mollisols and Aridisols, although a few from drier climates are Alfisols and some coastal grasslands are Vertisols. *See* HUMUS.

Grassland soils are dry throughout the profile for a portion of the year. If the soil profile is relatively shallow over a rock subsurface or an impervious layer, the total water stored may be too small to support trees even in geographical areas of relatively high rainfall, which accounts for the presence of grasslands in regions whose climate would support forests. Sandy soils reduce loss of water from runoff and increase water storage in the soil profile because of increased percolation. As a result, sandy soils support more vigorous grassland vegetation, or permit tree growth in a relatively arid region, when compared with soils containing higher amounts of silt and clay particles. Because of their dense fibrous root system in the upper layers of the soil, grasses are better adapted than trees to make use of light rainfall showers during the growing season. When compared with forest soils, grassland soils are generally subjected to higher temperatures, greater evaporation, periodic drought, and more transpiration per unit of total plant biomass. *See* BIOMASS; SOIL.

Vegetation. Worldwide, there are approximately 600 genera and 7500 species of grasses. Temperate North American dominant grass genera include the bluestems (*Andropogon*, *Schizachyrium*), grama grasses (*Bouteloua*), switchgrasses (*Panicum*), wheatgrasses (*Agrypyron*), wire grasses (*Artistida*), fescues (*Festuca*), and bluegrasses (*Poa*). However, many other herbaceous grassland species are broad-leaved forbs, which are plants other than grasses, and like most of the grasses, are perennial. Throughout the year, flowering plants bloom in the grasslands with moderate precipitation, and flowers bloom after rainfall in the drier grasslands. The number of plant species in any one grassland is relatively small, usually 50 to 350 species. Dominant species that flower late in the growing season tend to be taller than the early, cool-season species. In general, the dominant species persists from year to year with relatively small changes in biomass production or changes in importance. The subdominant species demonstrate much greater fluctuations from year to year, depending upon the weather conditions and especially in relation to periods of seed germination and growth.

Just as the above-ground canopy is layered, so are the roots below ground. In mesic grasslands, the roots of perennial grasses have rooting depths greater than 5 ft (1.5 m); many forbs have even deeper roots. Usually, however, 75% or more of the root biomass occurs in the top 10 in. (25 cm) of the soil profile.

With increasing aridity and temperature, grasslands tend to become less diverse in the number of species; they support more warm-season species (which mature late in the growing season); the complexity of the vegetation decreases; the total above-ground and below-ground production decreases; but the ratio of above-ground to below-ground biomass becomes smaller.

Many grassland plant species demonstrate adaptations to minimize damage from grazing. Unlike forbs and woody species, the growing tissues in grasses are at the base of the leaves near the soil surface, so when the leaves are grazed, the meristematic region is protected and can still produce new leaves. Some forbs also have antigrazing mechanisms, such as spines or tough structural material, chemicals that discourage grazing because of taste, supine growth forms so that the leaves stay near the soil surface, or growth flushes of plentiful forage under advantageous environmental conditions so that some of the plants are likely to be ungrazed.

Animal life. There are many more invertebrate species than any other taxonomic group in the grassland ecosystem. For example, more than 1600 different species of insects have been identified from a shortgrass prairie in Colorado. In North America, each prairie state has at least 100 species of grasshopper; Kansas has nearly 300 species. Invertebrate species occur both above and below the soil surface, and depending upon their life cycle, some are found in both locations. Invertebrates play several roles in the ecosystem. For example, many are herbivorous, such as grasshoppers, and eat leaves and stems, whereas others, like cutworms, feed on the roots of plants. Earthworms process organic matter into small fragments that decompose rapidly, scarab beetles process animal dung on the soil surface, flies feed on plants and are pests to cattle, and many species of invertebrates are predaceous and feed on other invertebrates. Soil nematodes, small nonarthropod invertebrates, include forms that are herbivorous, predaceous, or saprophagous, feeding on decaying organic matter. In a South Dakota mixed-grass prairie, $2–6 \times 10^6$ herbivorous nematode forms were found within a soil sample 3 ft (1 m) square and 8 in. (20 cm) deep. *See* SOIL ECOLOGY.

Most of the reptiles and amphibians in grassland ecosystems are predators. For example, lizards and box turtles prey on insects, and snakes prey on rodents and small invertebrates.

Relatively few bird species inhabit the grassland ecosystem, although many more species are found in the flooding pampas of Argentina than in the dry grasslands of the western United States. In North American prairies, perhaps only about a dozen bird species are restricted to the grasslands, and another two dozen are characteristic of but not limited to the grasslands. The small number of species is related to the uniform habitat. Typical groups of birds include hawks, grouse, meadowlarks, longspurs, and sparrows. Their role in the grassland ecosystem involves consumption of seeds, invertebrates, and vertebrates; seed dispersal; and scavenging of dead animals.

Small mammals of the North American grassland include moles, shrews, gophers, ground squirrels, and various species of mice. Among intermediate-size animals are the opossum, fox, coyote, badger, skunk, rabbit, and prairie dog; large animals include the mule deer (*Odocoileus hemionus*), white-tailed deer (*O. virginianus*), pronghorn (*Antilocapra americana*), and elk (*Cervus canadensis*). The most characteristic large mammal species of the North American grassland is the bison (*Bison bison*), but the 60–75 × 10^6 of these animals were largely eliminated in the late 1800s and are now mostly confined to reserves. Mammals include both ruminant (pronghorns) and nonruminant (prairie dogs) herbivores, omnivores (opossum), and predators (wolves).

Except for large mammals and birds, the animals found in the grassland ecosystem undergo relatively large population variations from year to year. These variations, some of which are cyclical and others more episodic, are not entirely understood and may extend over several years. Many depend upon predator–prey relationships, parasite or disease dynamics, or weather conditions that influence the organisms themselves or the availability of food, water, and shelter. *See* POPULATION ECOLOGY.

Microorganisms. Within the grassland ecosystem are enormous numbers of very small organisms, including bacteria, fungi, algae, and viruses. One study found 9 × 10^6 bacteria in a gram of tallgrass prairie soil; another study found about 6560 ft (2000 m) of fungal hyphae (strands of fungus) in 0.04 oz (1 g) of short-grass prairie soil. From a systems perspective, the hundreds of species of bacteria and fungi are particularly important because they decompose organic material, releasing carbon dioxide and other gases into the atmosphere and making nutrients available for recycling. Bacteria and some algae also capture nitrogen from the atmosphere and fix it into forms available to plants. *See* NITROGEN FIXATION; SOIL MICROBIOLOGY.

System functions. The grassland ecosystem consists of several components. Producers are plants that use the Sun's energy to capture carbon as carbon dioxide from the atmosphere and, with available nutrients from the soils, produce more plant material. Consumers consist of animals and microorganisms that feed upon plant parts (herbivores) and other animals (carnivores). Decomposer microorganisms and invertebrates convert dead organic matter to released carbon dioxide and available nutrients in the soil. The carbon cycle involves the transfer of carbon from the atmosphere into plants, through various animals and microorganisms, and back into the atmosphere. Energy is captured first from the Sun and then cycled through the system as organic material until this organic material is eventually decomposed to carbon dioxide. The nutrients are then released again to the soil. Only about 1% of the total solar radiation is captured by the vegetation. *See* BIOGEOCHEMISTRY.

The amount of vegetation produced depends on the type of grassland and the level of water and nutrients in the soil. In the tallgrass prairie, annual aboveground plant production is 1.6–2.0 oz/ft^2 (500–600 g/m^2), about 1.1 oz/ft^2 (350 g/m^2) in the mixed-grass prairie, 0.5–0.6 oz/ft^2 (150–200 g/m^2) in the shortgrass prairie, and as much as 13.1 oz/ft^2 (4000 g/m^2) in tropical grasslands. Total annual primary production is two or more times those amounts, because much of the energy is translocated below ground. *See* BIOLOGICAL PRODUCTIVITY; ECOLOGICAL ENERGETICS.

Grasslands have been used for grazing for the last 1.5–2.0 × 10^7 years, probably beginning in the Miocene. For all practical purposes, it can be assumed that grasslands have evolved under the influence of herbivores. In North America, the bison roamed across the landscape, heavily grazing the grassland they traversed. As the bison moved on, however, the rangelands recovered. Where the grassland is grazed heavily, as in prairie dog towns, around waterholes, and in fenced-in livestock pastures, grazing exerts enormous pressure on the vegetation. Because many of the native species are palatable, they are selectively grazed, and weedy, less palatable species remain. Moreover, those influences may extend to other components of the grassland ecosystem and increase the vulnerability of the soil to erosion or the susceptibility of the rangeland to insect herbivory or diseases. *See* ECOLOGICAL SUCCESSION.

Much of the grassland ecosystem has been burned naturally, probably from fires sparked by lightning. Human inhabitants have also routinely started fires intentionally to remove predators and undesirable insects, to improve the condition of the rangeland, and to reduce cover for predators and enemies; or unintentionally. Thus, grasslands have evolved under the influences of grazing and periodic burning, and the species have adapted to withstand these conditions. If burning or grazing is coupled with drought, however, the grassland will sustain damage that may require long periods of time for recovery by successional processes. Because of the prehistoric and historic burning of prairies, some rangeland management strategies have included periodic fires to remove woody species and old or dead herbaceous vegetation. *See* AGROECOSYSTEM; ECOSYSTEM.

Paul G. Risser

Bibliography. G. P. Chapman, *Desertified Grasslands: Their Biology and Management*, 1992; A. Langley, *Grasslands*, 1993; S. P. Long, M. B. Jones, and M. J. Roberts (eds.), *Primary Productivity of Grass Ecosystems*, 1991.

Gravel

An unconsolidated sedimentary aggregate containing more than 50% by weight of gravel-sized particles (mean diameter greater than 0.08 in. or 2 mm). The gravel-sized particles are termed the framework; those less than 0.08 in. (2 mm) in diameter are the matrix. There is an important distinction between framework-supported and matrix-supported gravels. The latter may possess a muddy matrix, in which case they are termed diamictons. Typically, diamictons are unstratified internally, contain subangular framework clasts, and are deposited by mass-flow processes such as debris flow or glacial-ice transport. Water-laid gravels are typically stratified or cross-stratified and are framework-supported, with subangular to rounded clasts in a sand matrix. Less commonly, they may be sand-matrix–supported, or they may lack matrix and then are termed openwork gravels. Water-worn gravel clasts tend to conform to the shape of triaxial ellipsoids and develop preferred orientation; with long axes normal to stream flow and intermediate axes dipping gently upstream.

The consolidated equivalents of gravels are conglomerates and breccias, the latter including only angular particles. Paleoenvironmental indicators for conglomerates include stratification, size grading, particle roundness, particle orientation, and matrix: framework relations. *See* BRECCIA; CONGLOMERATE; SEDIMENTARY ROCKS. Brian R. Rust

Bibliography. W. C. Krumbein and F. J. Pettijohn, *Manual of Sedimentary Petrography*, 1988; F. J. Pettijohn and P. E. Potter, *Sand and Sandstone,* 2d ed., 1987.

Gravimetric analysis

That branch of quantitative chemical analysis in which a desired constituent is converted (usually by precipitation) to a pure compound or element of definite, known composition, and is weighed. In a few cases, a compound or element is formed which does not contain the constituent but bears a definite mathematical relationship to it. In either case, the amount of desired constituent can be determined from the weight and composition of the precipitate. The following are the essential steps in a conventional gravimetric analysis.

Dissolution. A sample is weighed and dissolved in a suitable solvent. Water and dilute mineral acids dissolve most inorganic substances, but occasionally concentrated acids or the specific effect of hydrofluoric acid are required. Some refractories require fusions to convert them to acid-soluble products. *See* BALANCE; SOLUBILIZING OF SAMPLES.

Precipitation. After removal of any interfering substances, the desired constituent is precipitated by the addition of the appropriate reactant to the properly prepared solution. Conditions are regulated so that coprecipitation of foreign substances is minimized. In most cases, favorable conditions are attained by precipitating the constituent from a well-stirred, highly diluted, hot solution by the drop-by-drop addition of a dilute reagent in only slight excess over the amount theoretically required. In order to obtain a more nearly pure product, it is usually desirable to dissolve the precipitate in a suitable solvent and to form it a second time. *See* PRECIPITATION (CHEMISTRY).

Digestion. In order to be readily filterable, a suspension of a precipitate is ordinarily allowed to stand at high temperature for the period of time necessary to permit amorphous particles to clot, or crystalline particles to increase in size. This digestion is usually carried out on an electric hot plate so regulated as to maintain a temperature just below the boiling point of the liquid. The addition of paper pulp is helpful in aiding subsequent filtration, but may be used only in cases where the precipitate is to be ignited at a temperature that causes the pulp to burn off.

Filtration. Filtration is accomplished by pouring the suspension of a precipitate through a suitable filtering medium. Whatever the medium, as much of the supernatant liquid as possible is decanted through the filter, and transference of the precipitate is delayed as long as possible. Common filters are as follows.

Filter paper for gravimetric use is specially prepared and has undergone a treatment with hydrofluoric acid and with other acids so that on ignition it gives an ash of known, and usually negligible, weight. Papers of different degrees of porosity are available, and in a given filtration that grade is chosen which gives as rapid a filtration as possible and yet retains the precipitate completely. In general, gelatinous precipitates require a coarse-mesh paper; fine crystalline precipitates, a fine-mesh paper. Suction is almost never used in paper filtration.

A Gooch crucible is a perforated-base porcelain crucible, and the filtering medium is a pad of asbestos produced by pouring into the crucible a suspension of asbestos fibers which are matted by applying suction. A perforated plate on top of the pad holds it in place. Fiber-glass disks may be used instead of asbestos. Suction is applied during filtration; the crucible is dried to constant weight and is weighed before and after the filtration. Gooch crucibles are usually used for precipitates that attain a definite composition at the moderate temperature of a drying oven.

A Munroe crucible differs from a Gooch crucible in that it is made of platinum and uses a platinum sponge (produced by igniting alcohol-moistened ammonium chloroplatinate) as the filtering medium. It retains even fine precipitates and can be heated to a high temperature. It is too expensive for routine use.

An Alundum crucible is made from aluminum oxide and is porous throughout. It therefore needs no additional filtering medium and can be heated to a high temperature.

A Selas crucible has glazed porcelain sides and an unglazed porcelain base which serves as the filtering medium.

A glass filtering crucible has a sintered glass base for the filtering medium and is favored in cases where the precipitate needs to be heated only to the moderate temperatures of a drying oven.

Washing. Precipitates are washed (usually with hot water) until essentially free from soluble foreign matter. Occasionally an aqueous solution of an electrolyte is used to prevent peptization of the precipitate, with resulting conversion to a colloidal solution. Ammonium nitrate is favored for this purpose since it is removed by volatilization when the precipitate is subsequently ignited. Washing is more efficient if the wash water is added in several portions with intermediate drainage.

Drying and ignition. Some substances can be dried to constant weight by heating to relatively low temperatures (110–275°C or 230–530°F) in drying ovens. High-temperature ignition is usually carried out on a precipitate that has been filtered on paper. First the paper is smoked off at a low temperature, and the residue is ignited either in an electric muffle furnace or over a free flame. In the latter case, one of the following types of burners is used: a bunsen burner, a simple tube with an opening at the base to permit air to be mixed with the illuminating gas used; a tirrill burner, a modification of the bunsen burner which allows greater flexibility in the adjustment of the air-gas mixture; a meker burner, which is larger in diameter than the tirrill burner and has a grid at the top to give broad, hot flame; a blast lamp, which supplies air (or oxygen) under pressure to the illuminating gas used. The temperature delivered to the contents of a platinum crucible by a bunsen or tirrill burner is about 850°C (1600°F); that by a grid-top burner is about 1000°C (1800°F); that by a gas-air blast lamp about 1150°C (2100°F).

Cooling. A dried or ignited precipitate is cooled in a desiccator, which is a jarlike receptacle containing a desiccant, and which, in analytical chemistry, is used principally to allow a heated crucible and its contents to come to room temperature without taking on moisture from the air. Anhydrous calcium chloride, although not a very efficient desiccant, is commonly used for this purpose because of its low cost. Other desiccants in the order of increasing effectiveness are anhydrous barium perchlorate, sodium hydroxide sticks, silica gel, aluminum oxide, anhydrous calcium sulfate, calcium oxide, anhydrous magnesium perchlorate, barium oxide, and phosphorus pentoxide.

Calculations. At least two weighings are required for each analysis—the original sample, and the dried or ignited residue. From these weights, the percentage or proportion of the desired constitutent may be calculated from the equation below.

$$\%\mathrm{A} = \frac{\text{wt of residue} \times \text{factor} \times 100}{\text{wt of sample}}$$

The factor is determined from a knowledge of the chemical relationships between the weight of substance A contained in, or equivalent to, a fixed weight of residue of known composition. *See* ANALYTICAL CHEMISTRY; QUANTITATIVE CHEMICAL ANALYSIS; STOICHIOMETRY. Stephen G. Simpson

Bibliography. G. D. Christian, *Analytical Chemistry*, 5th ed., 1993; A. I. Vogel et al., *Vogel's Textbook of Quantitative Chemical Analysis*, 5th ed., 1989.

Gravitation

The mutual attraction between all masses and particles of matter in the universe. In a sense this is one of the best-known physical phenomena. During the eighteenth and nineteenth centuries gravitational astronomy, based on Newton's laws, attracted many of the leading mathematicians and was brought to such a pitch that it seemed that only extra numerical refinements would be needed in order to account in detail for the motions of all celestial bodies. In the twentieth century, however, Albert Einstein with his general theory of relativity and the concurrent development of quantum mechanics shattered this complacency. The subject is currently in a healthy state of flux.

Until the seventeenth century, the sole recognized evidence of this phenomenon was the gravitational attraction at the surface of the Earth. Only vague speculation existed that some force emanating from the Sun kept the planets in their orbits. Such a view was expressed by Johannes Kepler, the author of the laws of planetary motion, to justify his empirical laws. But a proper formulation for such a force had to wait until Isaac Newton founded Newtonian mechanics, with his three laws of motion, and discovered, in calculus, the necessary mathematical tool. *See* CELESTIAL MECHANICS; NEWTON'S LAWS OF MOTION; PLANET.

Newton's law of gravitation. Newton's law of universal gravitation states that every two particles of matter in the universe attract each other with a force that acts in the line joining them, the intensity of which varies as the product of their masses and inversely as the square of the distance between them. Or, the gravitational force F exerted between two particles with masses m_1 and m_2 separated by a distance d is given by Eq. (1), where G is called the constant of gravitation.

$$F = \frac{Gm_1m_2}{d^2} \tag{1}$$

A force varying with the inverse-square power of the distance from the Sun had been already suggested—notably by R. Hooke but also by other contemporaries of Newton, such as E. Halley and C. Wren—but this had been applied only to circular planetary motion. The credit for accounting for, and partially correcting, Kepler's laws and for setting gravitational astronomy on a proper mathematical basis is wholly Newton's.

Newton's theory was first published in the *Principia* in 1686. According to Newton, it was formulated in principle in 1666 when the problem of elliptic motion in the inverse-square force field was solved. But publication was delayed in part because of the difficulty of proceeding from the "particles" of the law to extended bodies such as the Earth. This difficulty was overcome when Newton established that, under his law, bodies having spherically symmetrical distribution of mass attract each other as if all their mass were concentrated at their respective centers.

Newton verified that the gravitational force between the Earth and the Moon, necessary to maintain the Moon in its orbit, and the gravitational attraction at the surface of the Earth were related by an inverse-square law of force. Let E be the mass of the Earth, assumed to be spherically symmetrical with radius R. Then the force exerted by the Earth on a small mass m near the Earth's surface is given by Eq. (2), and the acceleration of gravity on the Earth's surface, g, by Eq. (3).

$$F = \frac{GEm}{R^2} \qquad (2)$$

$$g = \frac{GE}{R^2} \qquad (3)$$

Let a be the mean distance of the Moon from the Earth, M the Moon's mass, and P the Moon's sidereal period of revolution around the Earth. If the motions in the Earth-Moon system are considered to be unaffected by external forces (principally those caused by the Sun's attraction), Kepler's third law applied to this system is given by Eq. (4).

$$\frac{4\pi^2 a^3}{P^2} = G(E + M) \qquad (4)$$

Equations (3) and (4), on elimination of G, give Eq. (5).

$$g = 4\pi^2 \frac{E}{E + M} \frac{a^2}{R^2} \frac{a}{P^2} \qquad (5)$$

Now the Moon's mean distance from the Earth is $a = 60.27R = 3.84 \times 10^8$ m (2.39×10^5 mi), and the sidereal period of revolution is $P = 27.32$ days $= 2.361 \times 10^6$ s. These data give, with $E/M = 81.35$, $g = 9.77$ m/s^2 (32.1 ft/s^2), which is close to the observed value.

This calculation corresponds in essence to that made by Newton in 1666. At that time the ratio a/R was known to be about 60, but the Moon's distance in miles was not well known because the Earth's radius R was erroneously taken to correspond to 60 mi (97 km) per degree of latitude instead of 69 mi (111 km). As a consequence, the first test was unsatisfactory. But the discordance was removed in 1671 when the measurement of an arc of meridian in France provided a reliable value for the Earth's radius.

Gravitational constant. Equation (3) shows that the measurement of the acceleration due to gravity at the surface of the Earth is equivalent to finding the product G and the mass of the Earth. Determining the gravitational constant by a suitable experiment is therefore equivalent to "weighing the Earth."

In 1774, G was determined by measuring the deflection of the vertical by the attraction of a mountain. This method is much inferior to the laboratory method in which the gravitational force between known masses is measured. In the torsion balance two small spheres, each of mass m, are connected by a light rod, suspended in the middle by a thin wire. The deflection caused by bringing two large spheres each of mass M near the small ones on opposite sides of the rod is measured, and the force is evaluated by observing the period of oscillation of the rod under the influence of the torsion of the wire (see **illus.**). This is known as the Cavendish experiment, in honor of H. Cavendish, who achieved the first reliable results by this method in 1797–1798. More recent determinations using various refinements yield the results: constant of gravitation $G = 6.67 \times 10^{-11}$ SI (mks) units; mass of Earth $= 5.98 \times 10^{24}$ kg. The result of the best available laboratory measurements, announced in 2000, is $G = (6.674215 \pm 0.000092) \times 10^{-11}$ in SI (mks) units. As a result, the mass of the Earth is $(5.972254 \pm 0.000082) \times 10^{24}$ kg and the Sun's mass is $(1.988435 \pm 0.000027) \times 10^{30}$ kg.

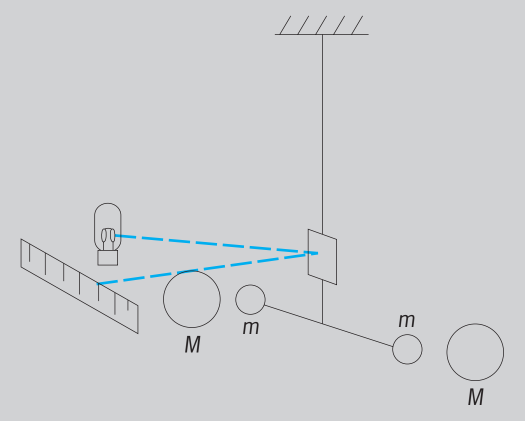


Diagram of the torsion balance.

In a more basic sense, G is a conversion factor between the units used for mass and distance and the units used for force. This definition must give the same force unit as the electromagnetic methods. These other methods are much more accurate than our current knowledge of G. Hopefully, improvements in the value of G will give the same size for the unit of force. Thus, knowing G more accurately will check the consistency of our definition of force.

In Newtonian gravitation, G is an absolute constant, independent of time, place, and the chemical composition of the masses involved. Partial confirmation of this was provided before Newton's time by the experiment attributed to Galileo in which different weights released simultaneously from the top of the Tower of Pisa reached the ground at the same time. Newton found further confirmation, experimenting with pendulums made out of different materials. Early in this century, R. Eötvös found that different materials fall with the same acceleration to within 1 part in 10^7. The accuracy of this figure has been extended to 1 part in 10^{11}, using aluminum and gold, and to 0.9×10^{-12} with a confidence of 95%, using aluminum and platinum.

With the discovery of antimatter, there was speculation that matter and antimatter would exert a mutual gravitational repulsion. But experimental results indicate that they attract one another according to the same laws as apply to matter of the same kind. *See* ANTIMATTER.

A cosmology with changing physical "constants" was first proposed in 1937 by P. A. M. Dirac. Field theories applying this principle have since been proposed by P. Jordan and D. W. Sciama and, in 1961, by C. Brans and R. H. Dicke. In these theories G is diminishing; for instance, Brans and Dicke suggest a change of about 2×10^{-11} per year. This would have profound effects on phenomena ranging from the evolution of the universe to the evolution of the Earth. There is no firm evidence at present to support a time variation of G. For instance, detailed analyses of the motion of the Moon using the lunar laser ranging data, the analysis of solar system data, and especially the ranging data using Viking landers on Mars, and the timing analysis of the binary pulsar PSR B1913 + 16 have all resulted in only upper limits on a possible temporal variation of G. The present upper limit on the relative rate of time variation of G is a few parts in 10^{14} per year. This result is theory-independent; that is, it is based purely on observations without relying on any particular theory in which G is presumed to vary with time. *See* PULSAR.

Mass and weight. In the equations of motion of Newtonian mechanics, the mass of a body appears as inertial mass, a measure of resistance to acceleration, and as gravitational mass in the expression of the gravitational force. The equality of these masses is confirmed by the Eötvös experiment. It justifies the assumption that the motion of a particle in a gravitational field does not depend on its physical composition. In Newton's theory the equality can be said to be a coincidence, but not in Einstein's theory, where this equivalence becomes a cornerstone of relativistic gravitation.

While mass in Newtonian mechanics is an intrinsic property of a body, its weight depends on certain forces acting on it. For example, the weight of a body on the Earth depends on the gravitational attraction of the Earth on the body and also on the centrifugal forces due to the Earth's rotation. The body would have lower weight on the Moon, even though its mass would remain the same. *See* CENTRIFUGAL FORCE.

Gravity. This should not be confused with the term gravitation. Gravity is the older term, meaning the quality of having weight, and so came to be applied to the tendency of downward motion on the Earth. Gravity or the force of gravity is today used to describe the intensity of gravitational forces, usually on the surface of the Earth or another celestial body. So gravitation refers to a universal phenomenon, while gravity refers to its local manifestation.

A rotating planet is oblate (or flattened at the poles) to a degree depending on the ratio of the centrifugal to the gravitational forces on its surface and on the distribution of mass in its interior. The variation of gravity on the surface of the Earth depends on these factors and is further complicated by irregular features such as oceans, continents, and mountains. It is investigated by gravity surveys and also through the analysis of the motion of artificial satellites. Because of the irregularities, no mathematical formula has been found that satisfactorily represents the gravitational field of the Earth, even though formulas involving hundreds of terms are used. The problem of representing the gravitational field of the Moon is even harder because the surface irregularities are proportionately much larger. *See* EARTH, GRAVITY FIELD OF.

In describing gravity on the surface of the Earth, a smoothed-out theoretical model is used, to which are added gravity anomalies, produced in the main by the surface irregularities.

Gravity waves are waves in the oceans or atmosphere of the Earth whose motion is dynamically governed by the Earth's gravitational field. They should not be confused with gravitational waves, which are discussed below.

Gravitational potential energy. This describes the energy that a body has by virtue of its position in a gravitational field. If two particles with masses m_1 and m_2 are a distance r apart and if this distance is slightly increased to $r + \Delta r$, then the work done against the gravitational attraction is $Gm_1m_2\Delta r/r^2$. If the distance is increased by a finite amount, say from r_1 to r_2, the work done is given by Eq. (6). If $r_2 \to \infty$, Eq. (7) holds.

$$W_{r_1,r_2} = Gm_1m_2 \int_{r_1}^{r_2} \frac{dr}{r^2} = Gm_1m_2 \left(\frac{1}{r_1} - \frac{1}{r_2}\right) \qquad (6)$$

$$W_{r_1,\infty} = \frac{Gm_1m_2}{r_1} \qquad (7)$$

If one particle is kept fixed and the other brought to a distance r from a very great distance (infinity), then the work done is given by Eq. (8). This is called

$$-U = \frac{-Gm_1m_2}{r} \qquad (8)$$

the gravitational potential energy; it is (arbitrarily) put to zero for infinite separation between the particles. Similarly, for a system of n particles with masses $m_1, m_2, \ldots, m_n$ and mutual distance r_{ij} between m_i and m_j, the gravitational potential energy $-U$ is the work done to assemble the system from infinite separation (or the negative of the work done to bring about an infinite separation), as shown in Eq. (9).

$$-U = -G \sum_{i<j} \frac{m_im_j}{r_{ij}} \qquad (9)$$

A closely related quantity is gravitational potential. The gravitational potential of a particle of mass m is given by Eq. (10), where r is distance measured from

$$V = \frac{-Gm}{r} \qquad (10)$$

the mass. The gravitational force exerted on another mass M is M times the gradient of V. If the first body is extended or irregular, the formula for V may be extremely complicated, but the latter relation still applies. *See* POTENTIALS.

A good illustration of gravitational potential energy occurs in the motion of an artificial satellite in a nearly circular orbit around the Earth which is affected by atmospheric drag. Because of the frictional drag the total energy of the satellite in its orbit is reduced, but the satellite actually moves faster. The explanation for this is that it moves closer to the Earth and loses more in gravitational potential energy than it gains in kinetic energy. *See* ENERGY.

Similarly, in its early evolution a star contracts, with the gravitational potential energy being transformed partly into radiation, so that it shines, and partly into kinetic energy of the atoms, so that the star heats up until it is hot enough for thermonuclear reactions to start. *See* STELLAR EVOLUTION.

Another related phenomenon is that of speed of escape. A projectile launched from the surface of the Earth with speed less than the speed of escape will return to the surface of the Earth; but it will not return if its initial speed is greater (atmospheric drag is neglected). For a spherical body with mass M and radius R, the speed of escape from its surface is given by Eq. (11). For the Earth, V_e is 11.2 km/s

$$V_e = \left(\frac{2MG}{R}\right)^{1/2} \qquad (11)$$

(7.0 mi/s); for the Moon, it is 2.4 km/s (1.5 mi/s), which explains why the Moon cannot retain an atmosphere such as the Earth's. By analogy, a black hole can be considered a body for which the speed of escape from the surface is greater than the speed of light, so that light cannot escape; however, the analogy is not really exact since Newtonian mechanics is not valid on the scale of the universe. *See* BLACK HOLE; ESCAPE VELOCITY.

Application of Newton's law. In modern times, Eq. (5)—in a modified form with appropriate refinements to allow for the Earth's oblateness and for external forces acting on the Earth-Moon system—has been used to compute the distance to the Moon. The results have been superseded in accuracy only by radar measurements and observations of corner reflectors placed on the lunar surface.

Newton's theory passed a much more stringent test than the one described above when he was able to account for the principal departures from Kepler's laws in the motion of the Moon. Such departures are called perturbations. A notable triumph of the theory occurred when the observed perturbations in the motion of the planet Uranus enabled J. C. Adams in 1845 and U. J. Leverrier in 1846 independently to predict the existence and calculate the position of a hitherto-unobserved planet, later called Neptune. When yet another planet, Pluto, was discovered in 1930, its position and orbit were strikingly similar to predictions based on the method used to discover Neptune. But the discovery of Pluto must be ascribed to the perseverance of the observing astronomers; it is not massive enough to have revealed itself through the perturbations of Uranus and Neptune. *See* PERTURBATION (ASTRONOMY).

F. W. Bessel observed nonuniform proper motions of Sirius and Procyon and inferred that each was gravitationally deflected by an unseen companion. It was only after his death that these bodies were telescopically observed, and they both later proved to be white dwarfs. More recently evidence has been accumulated for the existence of some planetary masses around stars. The discovery of black holes (which will never be directly observed) hinges in part on a visible star showing evidence for having a companion of sufficiently high mass (so that its gravitational collapse can never be arrested) and on observations of the motions of stars near galactic centers. *See* BINARY STAR.

Newton's theory supplies the link between the observed motion of celestial bodies and certain physical properties, such as mass and sometimes shape. Knowledge of stellar masses depends basically on the application of the theory to binary-star systems. Analysis of the motions of artificial satellites placed in orbit around the Earth has revealed refined information about the gravitational field of the Earth and of the Earth's atmosphere. Similarly, satellites placed in orbit around the Moon have yielded information about its gravitational field, and other space vehicles have yielded the best information to date on the masses and gravitational fields of other planets. *See* SATELLITE (SPACECRAFT); SPACE PROBE.

Newtonian gravitation has been applied to the motion of stars in galaxies as well as the motion of galaxies in clusters of galaxies. Careful observations of the motions of stars and gas (using the 21-cm line of neutral hydrogen) in the outer reaches of spiral galaxies have shown that the speed of rotation far from the center of a galaxy does not diminish with distance as would be expected on the basis of the gravitational influence of the rest of the stars as implied by the light distribution in the galaxy, but stays essentially constant. This circumstance leads to a rotation curve (the plot of the rotational speed versus the radius away from the center of a disk-shaped spiral galaxy) that is flat at large distances from the nucleus of the galaxy. The phenomenon associated with flat rotation curves for the outer regions of spiral galaxies has been well established, and is now generally interpreted in terms of the existence of the so-called dark matter, that is, excess matter that is invisible. Of course, the discrepancy with the prediction of Newtonian gravitation could be due to the breakdown of the theory. Newtonian gravitation and its relativistic generalization (general relativity) have been well tested in the solar system, but significant deviations over larger scales cannot be ruled out in principle. Though phenomenological modifications of the theory have been proposed to explain the discrepancy, no alternative theory that is based on fundamental physical principles exists. Therefore, the discrepancy is generally attributed to the existence of dark matter halos in which the visible galaxies are presumed to be embedded. Dark halo models together with rotation curves are used to calculate the density of dark matter in galaxies. It is estimated that the visible mass of a galaxy is only about 10% of its total mass. Furthermore, the discrepancy between the luminous and dynamical masses is thought to be larger for clusters of galaxies. Even more dark matter is needed if certain models of cosmology, such as the inflationary scenario, are taken into account. The result is that more than 90% of the mass of the universe could be in the form of dark matter. Proposals for the nature of dark matter range from exotic subatomic particles to black holes. The determination of the nature of dark matter constitutes a fundamental problem in astrophysics and cosmology. *See* COSMOLOGY; DARK MATTER; GALAXY, EXTERNAL; INFLATIONARY UNIVERSE COSMOLOGY; MILKY WAY GALAXY; UNIVERSE.

Accuracy of Newtonian gravitation. Newton was the first to doubt the accuracy of his law when he was unable to account fully for the motion of the perigee in the motion of the Moon. In this case he eventually found that the discrepancy was largely removed if the solution of the equations was more accurately developed. Further difficulties to do with the motion of the Moon were noted in the nineteenth century, but these were eventually resolved when it was found that there were appreciable fluctuations in the rate of rotation of the Earth, so that it was the system of timekeeping and not the gravitational theory that was at fault. *See* EARTH ROTATION AND ORBITAL MOTION.

A more serious discrepancy was discovered by Leverrier in the orbit of Mercury. Because of the action of the other planets, the perihelion of Mercury's orbit advances. But allowance for all known gravitational effects still left an observed motion of about 43 seconds of arc per century unaccounted for by Newton's theory. Attempts to account for this by adding an unknown planet or by drag with an interplanetary medium were unsatisfactory, and a very small change was suggested in the exponent of the inverse square of force. This particular discordance was accounted for by Einstein's general theory of relativity in 1916.

Newtonian mechanics considered gravity as a direct action-at-a-distance and in the early days did not properly account for the time delay between the action of gravity experienced by a body and the motion of the source. General relativity, on the other hand, considers gravitational effects as being geometric and time-related. *See* RELATIVITY.

Gravitational lens. From a relativistic point of view, light is expected to be deflected when it passes through a gravitational field. An analogy can be made to the refraction of light passing through a lens. For example, a galaxy situated between an observer and a more distant source will have a focusing effect, and this accounts for some of the observed properties of quasistellar objects. The multiple images of the quasar (Q0957 + 561 A,B) are almost certainly caused by the light from a single body passing through a gravitational lens. While this is the best-studied gravitational lens, many other examples of this phenomenon have been discovered. *See* GRAVITATIONAL LENS; QUASAR.

Testing of gravitational theories. One of the greatest difficulties in investigating gravitational theories is the weakness of the gravitational coupling of matter. For instance, the gravitational interaction between a proton and an electron is weaker by a factor of about 5×10^{-40} than the electrostatic interaction. (If gravitation alone bound the hydrogen atom, then the radius of the first Bohr orbit would be 10^{13} light-years, or about 1000 times the Hubble radius!) Of the four basic interactions (that is, strong, electromagnetic, weak, and gravitational) that are known at present, gravitation is by far the weakest force in nature. [A unification of the electromagnetic and weak interactions into the so-called electroweak interaction has been carried out, and further unification of this interaction with the strong (nuclear) force exists. So far, no sucessful realistic unification with gravitational interactions has been accepted.] *See* ELECTROWEAK INTERACTION; FUNDAMENTAL INTERACTIONS.

Many attempts have been made to discover Yukawa-type forces that would couple to matter with strengths close to that of gravity. These forces are expected to have finite ranges, however, in contrast to the infinite range of universal gravitation. Thus, deviations from Newton's inverse-square law of gravitation could reveal the presence of such interactions. These forces also appear in certain theories that attempt to unify gravity with the other fundamental interactions. Compelling evidence for the existence of any anomalous macroscopic force has not been found.

Results of experiments to search for new spin-dependent interactions place strict upper limits on anomalous spin-dependent couplings. Furthermore, torsion-balance experiments using ordinary matter have been interpreted as indicating the universality of free fall for matter as well as antimatter.

There is no firm evidence at present for any deviation from the Newtonian law of gravitation in the nonrelativistic regime.

Relativistic theories. Before Newton, detailed descriptions were available of the motions of celestial bodies—not just Kepler's laws but also empirical formulas capable of representing with fair accuracy, for their times, the motion of the Moon. Newton replaced description by theory, but in spite of his success and the absence of a reasonable alternative, the theory was heavily criticized, not least with regard to its requirement of "action at a distance" (that is, through a vacuum). Newton himself considered this to be "an absurdity," and he recognized the weaknesses in postulating in his system of mechanics the existence of preferred reference systems (that is, inertial reference systems) and an absolute time. Newton's theory is a superb mathematical one that represents the observed phenomena with remarkable accuracy.

The investigation of electric and magnetic phenomena culminated in the second half of the nineteenth century in the complete formulation of the laws of electromagnetism by J. C. Maxwell. Maxwell based his theory on M. Faraday's field concept. The electromagnetic field propagates with the speed of light; in fact, Maxwell's theory unified the science of optics with electricity and magnetism. Maxwell's theory of the electromagnetic field was extended and strengthened with the subsequent observations of electromagnetic waves by H. Hertz and the successes of the theory of electrons developed by H. A. Lorentz. *See* ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS.

The theory of relativity grew from attempts to describe electromagnetic phenomena in moving systems. No physical effect can propagate with a speed exceeding that of light in vacuum; therefore, Newton's theory must be the limiting case of a field theory in which the speed of propagation approaches infinity. Einstein's field theory of gravitation (general relativity) is based on the identification of the gravitational field with the curvature of space-time. The geometry of space-time is affected by the presence of matter and radiation. The relationship between mass-energy and the space-time curvature is therefore a relativistic generalization of the Newtonian law of gravitation. The relativistic theory is mathematically far more complicated than Newton's. Instead of the single Newtonian potential described above, Einstein worked with 10 quantities that form a tensor. *See* TENSOR ANALYSIS.

Principle of equivalence. An important step in Einstein's reasoning is his principle of equivalence, saying that a uniformly accelerated reference system imitates completely the behavior of a uniform gravitational field. Imagine, for instance, a scientist in a space capsule infinitely far out in empty space so that the gravitational force on the capsule is negligible. Everything would be weightless; bodies would not fall; and a pendulum clock would not work. But now imagine the capsule to be accelerated by some agency at the uniform rate of 981 cm/s^2 (32.2 ft/s). Everything in the capsule would then behave as if the capsule were stationary on its launching pad on the surface of the Earth and therefore subject to the Earth's gravitational field. But after its original launching, when the capsule is in free flight under the action of gravitational forces exerted by the various bodies in the solar system, its contents will behave as if it were in the complete isolation suggested above. This principle requires that all bodies fall in a gravitational field with precisely the same acceleration, a result that is confirmed by the Eötvös experiment mentioned earlier. Also, if matter and antimatter were to repel one another, it would be a violation of the principle. *See* FREE FALL.

Einstein's theory requires that experiments should have the same results irrespective of the location or time. This has been said to amount to the "strong" principle of equivalence.

Classical tests. The ordinary differential equations of motion of Newtonian gravitation are replaced in general relativity by a nonlinear system of partial differential equations for which general solutions are not known. Apart from a few special cases, knowledge of solutions comes from methods of approximation. For instance, in the solar system, speeds are low so that the quantity v/c (v is the orbital speed and c is the speed of light) will be small (about 10^{-4} for the Earth). The equations and solutions are expanded in powers of this quantity; for instance, the relativistic correction for the motion of the perihelion of Mercury's orbit is adequately found by considering no terms smaller than $(v/c)^2$. This is called the post-newtonian approximation. (Another approach is the weak-field approximation.)

Einstein's theory has appeared to pass three famous tests. First, it accounted for the full motion of the perihelion of the orbit of Mercury. (Mercury is the most suitable planet, because it is the fastest-moving of the major planets and has a high eccentricity, so that its perihelion is relatively easily studied.) Second, the prediction that light passing a massive body would be deflected has been confirmed with an accuracy of about 5%. Third, Einstein's theory predicted that clocks would run more slowly in strong gravitational fields compared to weak ones; interpreting atoms as clocks, spectral lines would be shifted to the red in a gravitational field. This, again, has been confirmed with moderate accuracy. Locations on the Earth's surface can be determined with great accuracy using the artificial satellites that make up the Global Positioning System (GPS). The clocks in that system must make relativistic corrections for both their motions relative to the Earth (special theory) and their gravitational potential relative to the Earth's surface (general theory). *See* SATELLITE NAVIGATION SYSTEMS.

Predictions of the theory have been confirmed in an experiment in which radar waves were bounced off Mercury; the theory predicts a delay of about 2×10^{-4} s in the arrival time of a radar echo when Mercury is on the far side of the Sun and close to the solar limb. Tests, similar in principle, have been conducted using observations of the Mariner space vehicles, the accuracy of confirmation being in the region of 4%. A greater level of accuracy has been achieved, by better than an order of magnitude, using data from transponders on Viking orbiters and landers on Mars. Furthermore, the deflection of microwave radiation passing close to the Sun has been observed using radio interferometry with a baseline of 22 mi (35 km). The amount of bending that has been found is 1.015 ± 0.011 times the amount predicted by general relativity. In another test, the precession of a gyroscope in orbit around the Earth is to be studied for evidence of the so-called geodetic precession as well as the precession due to the gravitomagnetic field of the rotating Earth. The lunar laser-ranging data have been used to measure the de Sitter precession of the Moon's orbit. This is due to the geodetic precession of the Earth-Moon orbital angular momentum in the gravitational field of the Sun. It amounts to an advance in the lunar node and perigee by about 2 seconds of arc per century, a prediction first made by W. de Sitter in 1916 soon after the advent of general relativity. Such small secular effects are suitable for study since they accumulate in time. Other periodic (noncumulative) orbital effects have until recently been too small to observe. But the current revolution in observational techniques and accuracy has changed the situation; post-newtonian terms are now routinely included in many calculations of the orbits of planets and space vehicles, and comparison with observations will furnish tests of the theory.

Mach's principle. One of the most penetrating critiques of mechanics is due to E. Mach, toward the end of the nineteenth century. Some of his ideas can be traced back to Bishop G. Berkeley early in the eighteenth century. Out of Mach's work there has arisen Mach's principle; this is philosophical in nature and cannot be stated in precise terms. The idea is that the motion of a particle is meaningful only when referred to the rest of the matter in the universe. Geometrical and inertial properties are meaningless for an empty space, and the motion of a particle in such space is devoid of physical significance. Thus the behavior of a test particle should be determined by the total matter distribution in the universe and should not appear as an intrinsic property of an absolute space. Mach's principle suggests that gravitation and inertia are equivalent. This idea strongly influenced the development of general relativity; however, general relativity, having an absolute concept of rotation, does not fully satisfy Mach's principle.

Brans-Dicke theory. This is a classical field theory of gravitation that was developed in 1961 by Brans and Dicke on the basis of an interpretation of Mach's principle. In this theory the gravitational field is described by a tensor and a scalar, the equations of motion being the same as those in general relativity. The addition of a scalar field leads to the appearance of an arbitrary constant, whose value is not known exactly. The Brans-Dicke theory predicts that the relativistic motion of the perihelion of Mercury's orbit is reduced compared with Einstein's value, and also that the light deflection should be less. With regard to the orbit of Mercury, Dicke pointed out that if the Sun were oblate, this might account for some of the motion of the perihelion. In 1967 he announced that measurements showed a solar oblateness of about 5 parts in 100,000 (or a difference in the polar and equatorial radii of about 21 mi or 34 km). His observations and discussion are still subject to some controversy. The difference between the theory and that of general relativity can be parametrized by the number ω, where $\gamma = (1 + \omega)/(2 + \omega)$; for general relativity, $\gamma = 1$. Dicke has proposed $\omega \approx 7.5$; but the results of measurement of deflection of radiation by the Sun indicate a value of ω greater than 23, for which the predictions of the two theories would not be greatly different. Subsequent data from the Viking spacecraft imply that this constant must be greater than about 500, thus rendering the Brans-Dicke theory almost indistinguishable from general relativity.

There are, of course, many other theories not mentioned here.

Supergravity. This is the term applied to highly mathematical theories of gravitation attempting to form a part of a unified field theory in which all types of forces are included. String, loop, and M-brane theories are under investigation. *See* FUNDAMENTAL INTERACTIONS; QUANTUM GRAVITATION; SUPERGRAVITY; SUPERSTRING THEORY.

Gravitational waves. The existence of gravitational waves, or gravitational "radiation," was predicted by Einstein shortly after he formulated his general theory of relativity. They are now a feature of any relativity theory. Gravitational waves are "ripples in the curvature of space-time." In other words, they are propagating gravitational fields, or propagating patterns of strain, traveling at the speed of light. They carry energy and can exert forces on matter in their path, producing, for instance, very small vibrations in elastic bodies. The gravitational wave is produced by change in the distribution of some matter. It is not produced by a rotating sphere, but would result from a rotating body, or pairs of bodies, not having symmetry about their axis of rotation or from pulsars and supernova explosions. In spite of the relatively weak interaction between gravitational radiation and matter, the measurement of this radiation is now technically possible. *See* SUPERNOVA.

There are currently (2006) five experiments attempting to detect gravitational radiation by interferrometric methods: the two LIGO (Laser Interferometer Gravitational-Wave Observatory) detectors located in the United States (near Livingston, Louisiana, and at the Washington Hanford Nuclear Research Area), VIRGO (France and Italy), GEO 600 (Germany), and TAMA (Japan). These instruments can detect a shift of the order of 1 part in 10^{21}. It is possible, if these instruments meet their design standards, that they may be sensitive enough to detect nearby wave sources. LIGO 2, which is under design, should increase sensitivity by a factor of 10. *See* LIGO (LASER INTERFEROMETER GRAVITATIONAL-WAVE OBSERVATORY.

The National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA) are planning a space-based interferometer, the Laser Interferometer Space Antenna (LISA). It will consist of three spacecraft in an equilateral triangle fromation with arms 5×10^6 km (3×10^6 mi) long. It will be located in the Earth's orbital path about 20° behind the Earth. Communication lasers will be located in each spacecraft. This design eliminates Earth noise and avoids the problems of folded light paths. *See* GRAVITATIONAL RADIATION.

J. M. A. Danby; Bahram Mashhoon; John L. Safko

Bibliography. H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; J. B. Hartle, *Gravity: An Introduction to Einstein's General Relativity*, 2003; W. Kopczynski and A. Trautman, *Spacetime and Gravitation*, 1992; C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, 1973; H. C. Ohanian and R. Ruffini, *Gravitation and Spacetime*, 2d ed., 1994; K. S. Thorne, *Black Holes and Time Warps: Einstein's Outrageous Legacy*, 1994; C. M. Will, *Theory and Experiment in Gravitational Physics*, rev. ed., 1993.

Gravitational collapse

The stage in the evolution of a star in which the pressure in the star is insufficient to maintain the star at a stable size. The material in the star or in the core of the star then falls inward under its own gravitational attraction. Depending on the mass, composition, and spin of the star, the collapse may proceed to the formation of a neutron star or black hole, possibly accompanied by a supernova explosion. *See* STELLAR EVOLUTION.

Stability of stars. Stars similar to the Sun maintain a stable size by continually burning nuclear fuel. In most stars, as for the Sun, this burning involves the conversion of hydrogen to helium with a release of energy as the nuclear reactions take place. This energy supplies heat to the interior of the star, which in turn keeps the core of the star hot enough so that nuclear reactions can continue. Heat is also continually transferred from the interior to the exterior of the star, where it is lost, primarily in the form of radiation.

Just as the temperature decreases from the interior to the exterior of the star, so does the pressure. The change in the pressure over an interval of distance is called a pressure gradient. A pressure gradient is required to keep the star in equilibrium, as can be seen by the following argument. First, any chunk of gas in a stable star is in equilibrium, on the average, if the star is stable. Second, there is a net gravitational force acting on any chunk of gas as a result of the attraction to the rest of the matter in the star, and this force, which is normally called the weight of the gas, must be directed toward the center of the star. Finally, the pressure gradient supplies a force to balance the gravitational force, since the pressure on the bottom of the chunk of gas (bottom meaning closer to the center of the star) is larger than the pressure on the top of the chunk. In equilibrium, the variation of pressure in the star is just that for which all chunks of gas are in equilibrium.

Contraction and the onset of collapse. Suppose now that the nuclear burning is turned off, as is the case if the hydrogen in the core is used up and only helium remains. Then, no more heat is supplied to the core of the star, and its temperature drops as remaining heat is transported to the surface of the star and is lost. Likewise, since the pressure in the gas depends on the temperature, the pressure in the core also drops, and more importantly, the pressure gradient decreases throughout the star. The various chunks of gas are no longer in equilibrium, and they move inward. This compression of the gas then causes a rise in temperature and the temporary reestablishment of the required temperature gradient. However, as heat is transported outward, the temperature again drops, the gas is further compressed, and so on. In this stage, the star is undergoing contraction, the heating coming from the gravitational potential energy of the star rather than nuclear reactions.

Eventually, the star condenses, and the temperature rises to a sufficiently high value that helium burning can take place. Once again, the star is in stable equilibrium until the reservoir of fuel in the core of the star is used up, after which contraction takes place until conditions are right for nuclear burning of higher elements. The contraction discussed here refers to what is going on in the core of the star; the atmosphere may actually be expanding in the process.

When the star has a core which is composed of iron, no further nuclear burning can take place, since iron is the most stable element. Continued contraction must then take place. In fact, theoretical calculations indicate that under normal conditions the temperature does not rise high enough for nuclear burning to proceed beyond carbon, at which point the star effectively runs out of fuel, and continued contraction also takes place. If, after ejection of material by the star, the mass of the star is less than about 1.2–1.4 solar masses, the limiting value known as the Chandrasekhar limit, the contraction takes place to a white dwarf. Theoretical studies indicate that normal stars smaller than about 4 solar masses will eject enough matter in the process of evolution to end up as white dwarfs. Observations suggest that stars of up to 8 solar masses can eject sufficient matter to become white dwarfs. A white dwarf is stable without any nuclear burning taking place. The pressure gradient is produced by the same kind of quantum interactions among the electrons as those which make atoms stable. The white dwarf then can only cool off with time, since it has no source of heat. The cooling period is very long, however, up to 1 billion years. As a result, many white dwarfs are observed in the sky, since the radiation continues long after the star, in some sense, has died. *See* WHITE DWARF STAR.

The mass loss by stars in the range of around 4 to 8 solar masses is not well understood. However, theoretical studies indicate that stars of more than about 8 solar masses will not lose enough mass in their evolution to become white dwarfs. What happens then is that the contracting star at some point becomes unstable; that is, the heating as a result of contraction is insufficient to produce the required pressure gradient to support the gas against gravity. This instability occurs first in the core of the star, and the core of the star starts to fall inward on itself. This is the stage of gravitational collapse. During collapse, the released gravitational potential energy goes into kinetic energy of motion of the core rather than into heating the core, since a near-equilibrium situation is required for efficient transfer of gravitational energy into heat energy. In essence, the core is then freely falling inward under its own gravitational attraction.

Collapse to a neutron star. There are several possible end points to the gravitational collapse of a star. White dwarfs have already been mentioned as a possible end point of stellar evolution, but a collapsing core will be too condensed to form a white dwarf, even if its mass is less than the maximum allowed mass. The only other known possibilities are collapse to a neutron star and collapse to a black hole. A neutron star is a highly condensed star in which the predominant constituent of matter is in

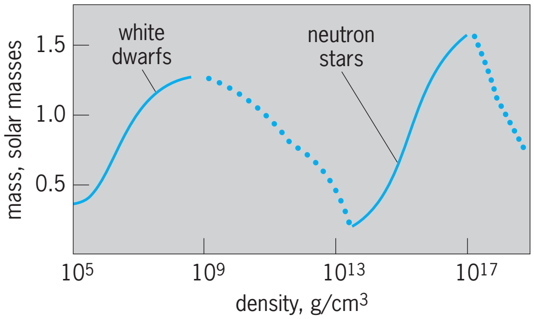


Fig. 1. Masses of white dwarfs and neutron stars as a function of central density. The solid lines indicate possible stable configurations.

the form of neutrons. The stability of such stars is a result of the quantum-mechanical interaction of neutrons, the same kind of interaction which, for electrons, leads to stability of white dwarfs. *See* NEUTRON STAR.

A neutron star is so condensed that the density of matter in the neutron star is comparable with the density of matter in the nucleus of an atom. For example, a neutron star of 1 solar mass has a radius of about 6 mi (10 km). As with white dwarfs, there is a maximum mass for stable neutron stars, in the range of 1.4 to 3 solar masses, the actual value being somewhat uncertain because nuclear forces are important at such a high density. The upper limit of 3 solar masses is fairly certain, however, since beyond that limit gravity will make the neutron star unstable, independent of the nature of the nuclear forces. **Figure 1** shows how the mass of white dwarfs and neutron stars depends on the central density of the star. The actual curve for white dwarfs depends on the composition of the star, and the actual curve for neutron stars is uncertain because of lack of knowledge of nuclear forces at high density.

Theoretical studies of collapsing stars indicate that the star develops a collapsing iron core as result of nuclear reactions induced in the compressed matter. When the core is sufficiently compressed, protons in the nuclei of the ions in the core are converted to neutrons, with the emission of neutrinos in the process. This takes place because electrons, due to quantum-mechanical interactions, are being forced into higher and higher energy states as the core of the star gets smaller and smaller. Protons can then absorb electrons to become neutrons. This is just the inverse of the process by which a free neutron decays to a proton plus an electron (and an antineutrino), a process which is prohibited in the interior of a neutron star because the electron does not have sufficient energy to be emitted. *See* RADIOACTIVITY.

Once the core reaches nuclear densities, repulsive nuclear forces become important and, if the mass of the core is small enough, the collapse is stopped and the core "bounces." The current thinking is that the bounce produces a shock wave which blows off the outer core and envelope of the star while a neutron star may be formed at the center. The large kinetic energy of the collapsing core is then converted to heat which ends up in radiation and in the kinetic energy of the material blown off the star. This process is thought to be the origin of some supernovae, although current models of collapsing stars do not follow this scenario in full three-dimensional detail. The problems with the models are thought to arise from simplifying assumptions, for example, no rotation or magnetic fields, which are needed to make the calculations tractable. *See* SHOCK WAVE; SUPERNOVA.

Collapse to a black hole. Some collapsing stars are expected to end up with a core that has a mass larger than the maximum mass of a neutron star. In addition, neutron stars (and white dwarfs as well) that had been formed earlier could accrete enough matter to become more massive than the maximum mass of a neutron star. In these cases, the only known alternative is for such stars to collapse to a black hole. A black hole is a region in space in which gravity is so strong that even light cannot escape from its surface. Although black holes had been conjectured earlier, Karl Schwarzschild found the first black-hole solution of general relativity in 1916, although the significance of the solution as a black hole was not realized at the time. After suggestions that black holes might be an end point of stellar evolution, J. R. Oppenheimer and H. Snyder showed in 1939 that a black hole must result from spherically symmetric gravitational collapse if the mass of the collapsing body is large enough. At present, a black hole is believed to be the only result of a collapsing star that cannot lose enough matter to become a white dwarf or neutron star. Black holes are even more condensed than neutron stars. For example, a 1-solar-mass black hole would have a radius of about 2 mi (3 km). *See* BLACK HOLE.

The collapse of a star to a black hole could take place without a supernova occurring, in which case the disappearance of the star might not be seen as a visible event. However, one potentially observable characteristic of such a collapse would be the emission of gravitational waves (discussed further below) as the black hole settled down to its equilibrium configuration. Gravitational waves interact with matter even more weakly than neutrinos, but detectors being planned and built would be sensitive enough to measure the gravitational radiation coming from collapse of a nearby star to a black hole. *See* GRAVITATION; GRAVITATIONAL RADIATION.

Because black holes emit no light, they are intrinsically dark and directly unobservable. However, they can still have an effect on nearby matter because their gravitational field is still present. In fact, there is strong evidence that at least one condensed body in a binary star system, Cygnus X-1, is a black hole. Its gravitational field results in matter heating up sufficiently to emit x-rays, which can only be done by a highly condensed body, and its mass is determined from the orbital motion to be larger than 5 solar masses. From theory, the only known body which has these properties is a black hole. Black holes have also been identified by similar evidence in several other x-ray binaries, including LMC X-3 and A0620-00. There is no proof that such black holes came from the gravitational collapse of a star, but that is the most reasonable hypothesis. *See* ASTROPHYSICS, HIGH-ENERGY; BINARY STAR; X-RAY ASTRONOMY.

One might think that the range of masses of black holes formed from collapse is limited, since a normal star can burn nuclear fuel under stable conditions only if its mass is less than about 60 solar masses. Of course, black holes of larger than 60 solar masses could exist if sufficient additional matter fell into existing black holes or if many black holes coalesced to form a larger black hole. However, it has been conjectured that supermassive stars could exist, with masses between 10^3 and 10^7 solar masses, in which the pressure gradient arises from radiation pressure, and the release of gravitational energy during contraction provides the major or, for the more massive stars, the only energy source. Theoretical studies have shown that the more massive stars would become, at some stage in their evolution, unstable to gravitational collapse, ending up as supermassive black holes. There is strong evidence for supermassive black holes at the centers of some galaxies, including the Milky Way Galaxy, and supermassive black holes are at the centers of quasars. Even if the existence of such large black holes is confirmed, it would not prove that the black holes necessarily came from the collapse of a supermassive star. *See* GALAXY, EXTERNAL; MILKY WAY GALAXY; QUASAR; SUPERMASSIVE STARS.

General relativistic collapse. All scenarios of gravitational collapse involve Einstein's theory of gravitation, general relativity, in an intimate way. In some cases, general relativity is responsible for the instability in the star which starts the collapse. In other cases, general relativity prevents the formation of any final state of matter other than a black hole. While these effects are important, they do not illustrate the dramatic way in which general relativity changes perception of the geometry of space and time. This is best illustrated by imagining what gravitational collapse to a black hole would look like, both to an observer looking at the event from outside the system and to an observer riding down on the surface of the collapsing star. In order to lessen the effect of the gravitational tidal forces, the same kind of forces responsible for the tides on the Earth, it is assumed that the collapse is of a supermassive star.

The outside observer sees the star initially falling inward, not unlike what one would expect from newtonian gravity. However, as the radius of the star decreases to a value close to the Schwarzschild radius (the radius of the black hole that will be formed), the rate of falling inward slows. In fact, the star never quite visibly reaches the Schwarzschild radius, even after an infinite time, as far as the outside observer is concerned. Another important effect comes about because of the gravitational redshift, the shifting of light toward the red part of the spectrum when the emitting atoms are in a region of large gravitational potential. The light from the star is shifted more and more toward longer wavelengths as the star approaches the Schwarzschild radius. To the outside observer, any light emitted by the star becomes shifted quickly outside the visible range of the spectrum and the star becomes dark. Finally, since the redshifting of the light just reflects the apparent slowing down of the atomic clocks in the atoms in the star, other clocks, including the life processes of the observer riding down on the surface of the star, will be similarly slowed down. As the images of the star and the observer on the surface fade out of view, there will be left a picture of the star and the observer at some definite time, as when a motion picture slows to a halt. *See* GRAVITATIONAL REDSHIFT.

The observer riding down on the surface has a quite different view of the process. Nothing particularly unusual is observed as the star collapses past its Schwarzschild radius, which occurs at the same time that the outside observer sees the star frozen in place as it fades out of view. However, crossing that radius does limit the options of the observer on the star; no matter what the observer does, escape past the Schwarzschild radius into the outside universe is impossible. Moreover, no observer, riding on the surface of a collapsing star, can escape the fate that awaits him or her. Initially, as the star gets smaller, unusual optical effects are noticed—the surface of the star appears to rise around the observer and the sky becomes smaller. More ominous is the increase in the tidal force on the observer. At the same time, this force stretches the observer from head to toe and squeezes the observer from the sides. After a short time this force rises to an infinite value, crushing both the observer and the star to infinite density. At this point the star is said to have reached a singularity. Once inside the Schwarzschild radius, the observer has no way of avoiding the singularity. Even if the observer were able to move outward at the speed of light, he or she would be crushed by the singularity in a short time. *See* RELATIVITY. Philip C. Peters

Observation of collapse. Although gravitational collapse does not necessarily imply a violent astronomical event, the existence of supernovae in the Milky Way Galaxy and in other galaxies is taken to be evidence for the collapse of the cores of stars. Moreover, in the Milky Way Galaxy the Crab and Vela supernovae now show the remnants of a dramatic explosion. Both remnants also have pulsars approximately at their center, pulsars being identified as rapidly rotating neutron stars. The Crab supernova was noted when it occurred in 1054, so there is no question that the remnant seen today came from an explosive event. Prior to 1987, these examples provided the best evidence that at least some stars undergo gravitational collapse to a neutron star, in the process generating a supernova explosion. However, many questions could not be answered, in particular the nature of the star before collapse and the details of the collapse itself. *See* PULSAR.

On February 24, 1987, a supernova was observed in the Large Magellanic Cloud, a small galaxy associated with the Milky Way Galaxy, and therefore relatively close to Earth as far as supernovae are concerned. This supernova, named 1987A, must be considered one of the most important astronomical events of the twentieth century. For the first time a supernova occurred in a star that could be identified and studied on existing plates, and the record of the evolution of the supernova has been obtained with remarkable detail, using all parts of the spectrum from radio waves to x-rays and gamma rays. Moreover, prior to the appearance of the supernova optically, neutrinos from the supernova were observed in two widely separated detectors, confirming for the first time the predicted production of neutrinos in supernova explosions. Information from the light curve yields information on the radioactive elements produced in the explosion. The gravitational collapse of the core of the star was expected to end up as a neutron star. However, not all the details corresponded to existing models. For example, the progenitor star was a blue supergiant, not a red supergiant, which is the kind of star, based on models, thought most likely to explode. Perhaps connected with this is the fact that the supernova was initially much dimmer than a normal supernova would have been at that distance. Substantial mass loss of the progenitor stars played an important role in its evolution. The effects of this mass loss have been seen in the complex cocoon-like structure that developed around 1987 as the rapidly expanding stellar ejecta from the explosion interacted with the more slowly expanding ejecta from the progenitor star. *See* SUPERNOVA.

Events nowadays widely believed to be connected with the core collapse of stars are gamma-ray bursts. These bursts have durations from a few milliseconds to a few minutes and emit gamma rays, the most energetic photons in the universe. Gamma-ray bursts shine hundreds of times brighter than typical supernova and are observed roughly once a day from random positions on the sky. It is believed that gamma-ray bursts are created in extremely energetic supernovae called hypernovae. Only about one out of 100,000 supernovae is a hypernova, and the cores of stars collapsing in hypernovae are believed to be massive enough to produce a black hole, connecting gamma-ray bursts with the formation of black holes. While the core collapses to form a black hole, a blast wave is created which moves through the star at a speed close to the speed of light. Once the blast wave collides with the stellar material inside the star, the energy of the blast wave is partially converted into the production of gamma rays, which leave the star's surface with the speed of light. Roughly 2700 gamma-ray bursts have been observed with the *Compton Gamma Ray Observatory*'s Burst And Transient Source Experiment (BATSE). *See* GAMMA-RAY BURSTS.

Future measurements of gravitational collapse will focus on measuring gravitational wave signals. In the case of a nonspherical symmetric collapse, the general theory of relatively perdicts the emission of gravitational waves. These waves are ripples in space-time, moving through it with the speed of light and perturbing it (**Fig. 2**). The basic idea is to measure the perturbation to space-time that the gravitational wave induces. The best way to do so is in interferometer experiments. An interference pattern is generated by a laser beam, which is split into two beams initially in phase with each other and moving perpendicularly away from each other. These beams are reflected several times between a pair of mirrors. When a gravitational wave passes the experiment, the space between the mirrors in each perpendicular arm of the experiment will be changed differently, resulting in a change in the interference pattern that can be measured. Several experiments—ground-based such as the Laser Interferometer Gravitational-Wave Observatory (LIGO), and space-based such as the Laser Interferometer Space Antenna (LISA)—will be trying to measure these signatures. *See* LIGO (LASER INTERFEROMETER GRAVITATIONAL-WAVE OBSERVATORY).

Fig. 2. Artist's conception of gravitational waves. (*LIGO*)

Sadegh Khochfar; Philip C. Peters; Joseph Silk

Bibliography. D. Arnett, *Supernovae and Nucleosynthesis*, 1996; H. A. Bethe and G. Brown, How a supernova explodes, *Sci. Amer.*, 252(5):60–68, 1985; T. Piran, S. Weinberg, and J. C. Wheeler (eds.), *Supernovae*, 1991; S. L. Shapiro and S. A. Teukolsky, *Black Holes*, *White Dwarfs*, *and Neutron Stars*, 1983; J. C. Wheeler, *Cosmic Catastrophes: Supernovae, Gamma-ray Bursts, and Adventures in Hyperspace*, 2000.

Gravitational lens

A massive body producing distorted, magnified, or multiple images of more distant objects when its gravitational fields bend the paths of light rays. Lenses have been observed when the light from very distant quasars is affected by intervening galaxies and clusters of galaxies, producing several different images of the same quasar. A. Einstein predicted the occurrence of this phenomenon in 1936, but the discovery of real gravitational lenses did not occur until 1979. Gravitational lenses, in addition to being intrinsically interesting, can reveal the intrinsic properties of galaxies, active galaxies, and quasars, and provide information on the universe and its contents, including dark matter.

Action of gravity. The lens phenomenon exists because gravity bends the paths of light rays, which is

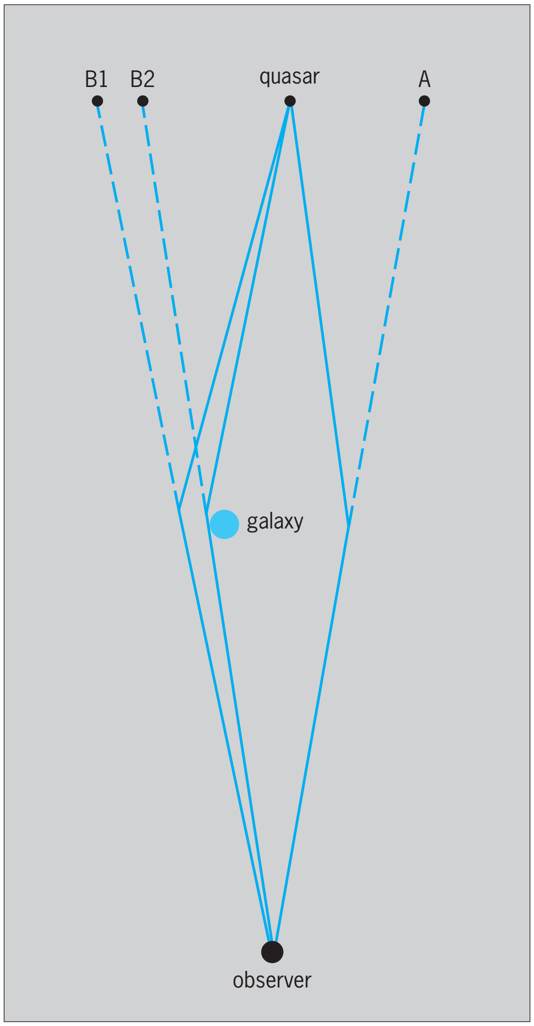


Fig. 1. Schematic illustration of a gravitational lens. The angles are exaggerated for clarity. Here, the lens action produces three images of a quasar (A, B1, and B2), since light from the image can travel along three different curved paths and still reach the observer.

predicted by Einstein's general theory of relativity. Since photons, the carriers of light energy, have no mass, Newton's theory of gravity indicates that light would always travel in a straight line even if there were heavy, massive objects between the source and the observer. (Even if photons are given mass in Newton's theory, the predicted bending of light is different from the result in general relativity.) But in general relativity, gravity acts by producing curvature in space-time, and the paths of all objects, whether or not they have mass, are also curved if they pass near a massive body. *See* GRAVITATION; RELATIVITY.

Numerous eclipse observations have confirmed Einstein's prediction with modest accuracies of 20–30%. Radio astronomers have measured changes in the positions of quasars that occur when the Sun passes near them in the sky. The precision of these experiments, which fit Einstein's predictions, is now at the level of tenths of a percent.

A massive object acts as a gravitational lens when light rays from a distant quasar are bent around or through it and are focused to form an image, which can be seen or photographed by an astronomer on Earth (**Fig. 1**).

Sometimes a lens can amplify the total intensity of light in a quasar image, making it considerably brighter than the quasar would appear to be in the absence of a lens. If the galaxy and quasar are sufficiently well aligned, several images of the same quasar will appear, since light can travel on many different paths and still arrive at the detecting telescope. In some cases the lensed image is part of a ring. It has been shown that if there are multiple images of the same quasar, there must be an odd number of them, as long as the galaxy is big enough so that it does not act as a point mass.

Discovery of lenses. Astronomers treated gravitational lenses as curiosities for a long period of time. Space is so empty that the probability of two stars being aligned accurately enough is extremely small. But the discovery of quasars, hyperactive galaxies which are bright enough to be visible even though they are nearly 10^{10} light-years (1 light-year is equal to 5.88×10^{12} mi or 9.46×10^{12} km) away, made it possible to probe a much larger volume of space. Now, there was a reasonable chance that a galaxy might lie in the path of light traveling from a quasar to the Earth. *See* QUASAR.

A survey of faint blue objects in the northern sky by Richard Green and Maarten Schmidt, completed in 1977, was one of the keys to establishing gravitational lenses as actual (as opposed to hypothetical) astrophysical objects. This Palomar-Green survey isolated several pairs of quasars, where two very similar faint blue objects were very close to each other in the sky. Subsequent investigation demonstrated that these two objects were not only very similar; they were two distinct images of the same object that were being produced when light from the object passed around a galaxy that was between the quasar and the Earth. In some cases, there were three images; **Fig. 2** shows one of the early multiple

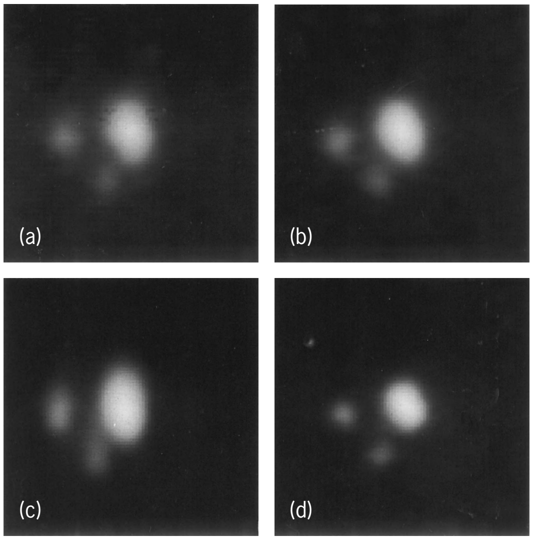


Fig. 2. Images of the triple quasar PG 1115 + 08. (*a*) Blue filter. (*b*) *V* band (yellow) filter. (*c*) Unfiltered image. (*d*) Red filter. (*From E. K. Hege et al., Morphology of the triple QSO PG1115 + 08, Nature, 287:416–417, 1980*)

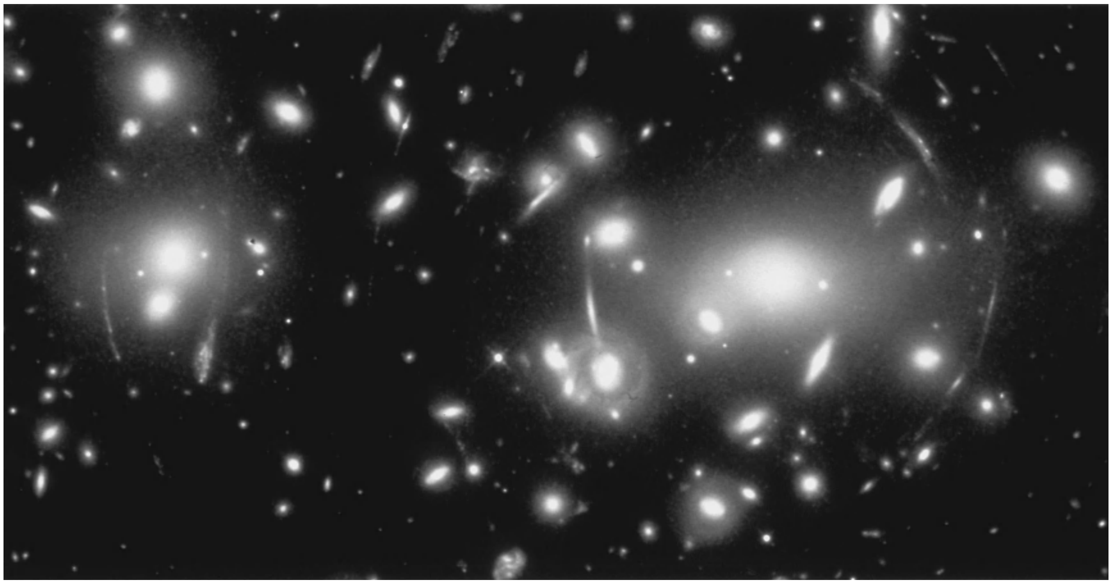


Fig. 3. Image of the galaxy cluster Abell 2218 obtained from the *Hubble Space Telescope*. The ring of light is an illusion, light from a very distant object that is bent by the gravity from the galaxy cluster. (*NASA*)

quasars, PG1115 + 08. (PG designates the Palomar-Green survey, and the numbers provide an approximate position in the sky in terms of the astronomical coordinates of right ascension and declination.)

More striking are some ringlike arcs produced when the distant object, the lensing galaxy or galaxy cluster, and the observer are exactly in a straight line. **Figure 3** is a picture, taken with the *Hubble Space Telescope*, of one of the most impressive arcs. When light passes through the cluster, it is bent. The distant object could be either a galaxy or a quasar, but it is definitely much farther away than the cluster. The cluster acts like a little telescope of its own, focusing the light from the distant object toward the Earth and making it brighter than it otherwise would be. *See* SATELLITE (ASTRONOMY).

Implications. The discovery of gravitational lenses affects astronomers' understanding of the universe on the very largest scales. The very existence of this phenomenon indicates that nearly a dozen quasars—the ones that are being lensed—are more distant than the galaxies that are focusing their light. When quasars were first discovered, some astrophysicists argued that their redshifts were produced by exotic new physics and the quasars were just beyond the boundary of the Milky Way Galaxy. This controversy has largely subsided but has not been completely resolved. The lens phenomenon shows that at least some quasars are billions of light-years away, well beyond the edge of the Milky Way. If some quasars are billions of light-years away and others look like them, it is reasonable to conclude that all quasars are billions of light-years away at the edge of the observable universe.

Gravitational lenses can be used to determine the distance scale of the universe. Most quasars change the amount of light that they produce. In the case of a multiply imaged quasar like PG 1115+080 (Fig. 2), observers on Earth could see that change occur at different times because light travels on different paths to get here. The image where light travels on a more direct path would brighten first, and the one taking a more roundabout route would brighten later. The differences in the two path lengths can be used to deduce the distance to the quasar and the lensing object. Astronomers can then measure the redshifts of these distant objects and use the lens as another way to determine how fast the universe is expanding.

This seemingly easy idea is hard to implement in practice. Only one of 500 quasars is lined up in exactly the right way, and the mass distribution in the lensing object must be fully understood in order to interpret the data correctly. Sharp infrared pictures from the *Hubble Space Telescope* severely constrained possible models of the lens, making the interpretation much more secure. The data indicate a Hubble constant of 70 kilometers per second per megaparsec (the conventional units for measuring the Hubble constant), meaning that, if the universe has a very low density, its age is 14 billion years. *See* COSMOLOGY; HUBBLE CONSTANT.

Dark matter. Gravitational lenses also enable the discovery of invisible objects. The speed with which stars move in galaxies and galaxies move in galaxy clusters indicates that many galaxies may be surrounded by massive dark halos. Since the matter that composes these halos cannot be seen, the name "dark matter" has been used to describe it. The dark matter could be brown dwarfs (objects not massive enough to be stars), dead stars, Jupiter-sized objects, or subnuclear particles. The more massive forms of dark matter are termed MACHOs (massive compact halo objects). *See* BROWN DWARF.

If a MACHO passed directly between a distant star and the Earth, the light from the star could be temporarily brightened as the MACHO focused the starlight toward the Earth. Precise calculations of this event indicate that the brightening should last about a week. Several teams of astronomers have made repeated observations of a nearby galaxy, the Large Magellanic Cloud, in search of this phenomenon, and seven such events have been detected. These events indicate that MACHOs, which are probably low-mass white dwarf stars, make up a sizable fraction of the mass of the halo of the Milky Way Galaxy, probably at least 20% of the dark matter and possibly as much as 100%. *See* MAGELLANIC CLOUDS; MILKY WAY GALAXY.

Harry L. Shipman

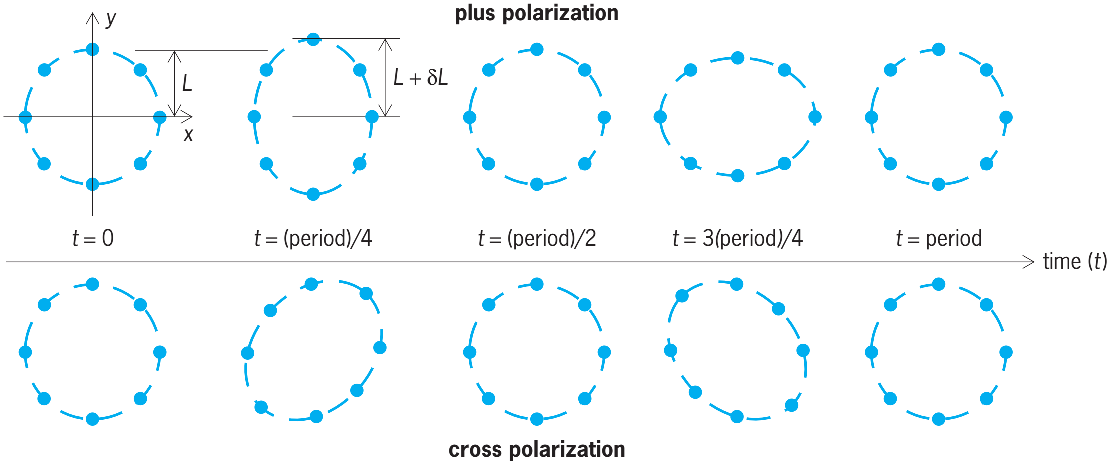
Bibliography. N. Cohen, *Gravity's Lens: Views of the New Cosmology*, 1988; J. Glanz, Is the dark matter mystery solved?, *Science*, 271:595–596, 1996; R. Schild, Gravity is my telescope, *Sky Telesc.*, 81(4):375–379, April 1991; G. Taubes, OGLEing, MACHOs, and the search for dark matter, *Science*, 260:492–493, 1993; K. Thorne, *Black Holes and Time Warps: Einstein's Outrageous Legacy*, 1994.

Gravitational radiation

A wave that is generated by the acceleration of mass and travels at the speed of light. Gravitational waves are an implicit outcome of the special theory of relativity, and were explicitly put forward in Albert Einstein's theory of general relativity in 1916. Einstein showed that the acceleration of masses generates time-dependent gravitational fields which can carry energy away from their source at the speed of light. *See* RELATIVITY.

Properties. The essence of general relativity is that mass and energy produce a curvature of four- dimensional space-time and that matter moves in response to this curvature. The Einstein field equations proscribe the interaction between mass and space-time curvature, much as Maxwell's equations proscribe the relationship between electric charge and electromagnetic fields. One solution to the field equations is a weak, oscillating perturbation to space-time curvature, that is, a gravitational wave, just as electromagnetic waves, in the form of light, microwaves, or x-rays, are solutions to Maxwell's equations. Gravitational waves can be thought of equivalently in several ways: as an oscillating perturbation to a flat, or Minkowski, space-time metric; as an oscillating strain in space-time; as an oscillating tidal force between free test masses. *See* ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS; SPACE-TIME.

As with electromagnetic waves, gravitational waves travel at the speed of light and are transverse in character; that is, the strain oscillations occur in directions orthogonal to the direction in which the wave is propagating. Whereas electromagnetic waves are dipolar in nature, gravitational waves are quadrupolar: the strain pattern contracts space along one transverse dimension, while expanding it along the orthogonal direction in the transverse plane (see **illustration**). This characteristic arises from the fact that gravitational radiation is produced by oscillating multipole moments of the mass distribution of a system, just as electromagnetic waves are produced by oscillating moments of the charge distribution. The principle of mass conservation rules out monopole radiation, just as charge conservation does so in electromagnetism, and the principles of linear and angular momentum conservation rule out gravitational dipole radiation. Quadrupole radiation is the lowest allowed form and is thus usually the dominant form. In this case, the gravitational-wave field strength is proportional to the second time derivative of the quadrupole moment of the source, and if the gravitational wave is to carry away energy, the field strength must fall off in amplitude inversely with distance from the source. The tensor character of gravity—the hypothetical graviton is a spin-2 particle—means that the transverse strain field comes in two orthogonal polarizations, rotated with respect to each other in the transverse plane by 45°. *See* ANGULAR MOMENTUM; CONSERVATION LAWS (PHYSICS); GRAVITON; LINEAR



Effect of a gravitational wave, passing in a direction perpendicular to the plane of the paper, on a ring of test particles. The ring is distorted into an ellipse, elongated in one direction in one half-cycle of the wave, and elongated in the orthogonal direction in the next half-cycle. The waves come in two independent polarization states, denoted plus polarization and cross polarization; the strain patterns for the two states are rotated 45° from each other.

MOMENTUM; MOMENTUM; MULTIPOLE RADIATION; POLARIZATION OF WAVES.

Gravitational waves differ from electromagnetic waves in that they propagate essentially unperturbed through space, as they interact only very weakly with matter. Furthermore, gravitational waves are intrinsically nonlinear, because the wave-energy density itself generates additional curvature of space-time. However, this phenomenon is significant only very close to strong sources of waves, where the wave amplitude is relatively large. More usually, gravitational waves distinguish themselves from electromagnetic waves by the fact that they are very weak. One cannot hope to detect any waves of terrestrial origin, whether naturally occurring or artificial; instead one must look to very massive astrophysical objects moving at very high velocities. For example, strong sources of gravitational waves that may exist in our galaxy or nearby galaxies are expected to produce wave strengths on Earth that do not exceed strain levels of one part in 10^{21}.

Sources. The prototypical example of a gravitational-wave source is a pair of compact astrophysical objects such as stars in mutual orbit about their center of mass, otherwise known as a binary star system. The pair could consist of two neutron stars, two black holes, or one of each type of object, both of which are remnants of once-giant, massive stars. Both types of objects are extremely dense, which is key to generating strong gravitational waves. A binary system will radiate waves with a period equal to half the orbital period. This period, and the physical separation between the objects, will decrease over time as gravitational waves carry energy away, a process known as an inspiral. Eventually, the objects will come close enough to each other that they merge into a single object, during which gravitational waves will continue to be emitted. This coalescence of two neutron stars or a neutron star and a black hole is now thought to be the source of the so-called short gamma-ray bursts that have been detected by various gamma-ray observatories. Finally, the merged objects may form a black hole, which itself would oscillate in a normal mode, emitting further gravitational energy in heavily damped sinusoidal waves. *See* BINARY STAR; BLACK HOLE; GAMMA-RAY BURSTS; NEUTRON STAR.

Another example of a source is an isolated, spinning neutron star, which deviates in shape from a perfect sphere (so that it has a quadrupole moment). Such a star would emit nearly monochromatic gravitational waves, with the wave frequency changing only very slowly as energy is lost. On the other hand, short bursts of gravitational waves could be emitted by supernovae and other stellar collapses or by the infall of a massive object into a black hole. Finally, there may be a stochastic background of gravitational waves, analogous to the cosmic microwave background, that originates early in the evolution of the universe. Indeed, the inflationary theory of cosmology predicts that such a background should exist, though it would be very weak in amplitude. Other posited cosmological sources of a stochastic background include a network of "cosmic strings," and rapidly expanding temperature bubbles, formed by a phase transition that may have taken place as the universe cooled and expanded. *See* COSMIC BACKGROUND RADIATION; COSMIC STRING; COSMOLOGY; GRAVITATIONAL COLLAPSE; INFLATIONARY UNIVERSE COSMOLOGY; PHASE TRANSITIONS; SUPERNOVA; UNIVERSE.

Detection. Because gravitational waves are inherently so weak—very much weaker than electromagnetic waves—Einstein concluded that they would never be detected. Nevertheless, efforts to directly detect gravitational waves have been made since the 1960s, though as yet with no detection. Even indirect evidence of their existence did not come until many decades after their prediction, with the 1974 discovery and subsequent observations of a binary pulsar by Russell Hulse and Joseph Taylor. By precisely monitoring the orbital period of this binary star system, thought to be made up of two neutron stars, they were able to confirm that the orbit was speeding up at just the rate predicted by the general-relativistic emission of gravitational waves. *See* PULSAR.

Resonant-mass detectors. The effort to detect gravitational waves directly started in the early 1960s with Joseph Weber's experiments with resonant acoustic bar antennas. This type of detector consists of a large mass equipped with highly sensitive vibration detectors. The mass is typically a cylinder of aluminum weighing several tons and having a resonant vibrational mode around 1000 Hz. A passing gravitational-wave burst would resonantly excite the cylinder's mode, and the mode amplitude could be measured using some type of electromechanical transducer that is tuned to the mode. This line of research expanded considerably after 1969 in response to Weber's apparently claimed detection of gravitational waves (though Weber himself later said that he had not actually claimed a detection), as over a dozen groups tried to construct similar bar detectors to see if Weber's results could be reproduced. None were able to confirm the detection. The sensitivity of bar detectors has increased significantly since the initial experiments, with the detectors being cooled to cryogenic temperatures to reduce thermal noise and isolated from terrestrial vibrations by a series of mechanical filters. Present acoustic bar detectors have achieved a strain sensitivity to short bursts of several parts in 10^{20}.

Free-mass detectors. The second, newer method of detecting gravitational waves on Earth uses "free masses" separated from each other by long baselines whose lengths are precisely monitored using a laser interferometer. The basic detection idea was first suggested by Mikhail E. Gertsenshtein and V. I. Pustovoit in 1962 and was independently conceived of by Rainier Weiss in 1972. Robert Forward built the first prototype gravitational-wave interferometer in the 1970s, and this early work was quickly followed by more prototypes at numerous research laboratories. As with the bar detectors, the sensitivities of the prototype interferometers improved by several orders of magnitude over the first detectors. But it was clear from the start that interferometers would need to be greatly scaled up in size before they could hope to achieve a sensitivity sufficient to detect gravitational waves. The laboratory prototype research eventually led to the design and construction of several long-baseline interferometric detectors around the world: the LIGO (Laser Interferometer Gravitational-Wave Observatory) detectors at two locations in the United States, constructed with 4-km-long (2.5-mi) arms; the Virgo detector in Cascina, Italy, with 3-km-long (1.9-mi) arms; the GEO detector outside Hannover, Germany, with 600-m-long (2000-ft) arms; and the TAMA detector near Tokyo, Japan, with 300-m (1000-ft) arms. Presently the LIGO detectors are the most sensitive and are able to detect strains as small as one part in 10^{21}. *See* LIGO (LASER INTERFEROMETER GRAVITATIONAL-WAVE OBSERVATORY).

All of these detectors are L-shaped Michelson interferometers, a configuration that takes advantage of the quadrupolar field pattern of a gravitational wave. The test-mass mirrors of the interferometer are suspended as low-frequency pendula, so that they are free to be moved by a passing gravitational wave. The interferometer is illuminated with a powerful and extremely well-stabilized laser, and the arms and main components are contained in a high-vacuum enclosure to reduce fluctuations from air molecules. When a gravitational wave passes through the interferometer, the light travel time down and back one arm will become shorter or longer relative to the round-trip time in the other arm; this will result in a light signal at the interferometer output, proportional to the wave amplitude. The optical beams are usually folded many times within the arms, typically using resonant optical cavities, to increase the effective length of the arms and thus the change in light travel time. This type of detector is more broadband than the resonant-bar type, and is sensitive to gravitational waves in a frequency band starting around 10 Hz and extending to nearly 10 kHz. At frequencies below 10 Hz, terrestrial vibrations and forces simply become too large, and at the upper end of the band the interferometers lose sensitivity as the gravitational wavelength becomes shorter than the effective arm length. *See* INTERFEROMETRY.

Detectors in space. Searching for gravitational waves at frequencies below 10 Hz requires detectors in space, several of which have been conceived. For gravitational-wave periods between a few minutes and a few hours, the best detection method at present is the Doppler tracking of spacecraft. In this method, a highly coherent microwave signal is generated on Earth and transmitted to the spacecraft. The microwaves are received by the spacecraft and transponded back to Earth. The frequencies of the transmitted and received signals are compared, and from this the relative velocity between Earth and the spacecraft is determined. A gravitational wave would produce fluctuations in the measured Doppler shift, proportional to the wave amplitude. Interestingly, each feature in the gravitational waveform shows up three times in the Doppler shift, a fact that can be used to help distinguish gravitational waves from noise in the data. In December 2001, the *Cassini* space probe was tracked in this way as it cruised toward Mars. An upper bound was set on the amplitude of gravitational waves at frequencies around a milliHertz, at a level of 2 parts in 10^{15}.

Monitoring pulsar signals. At even lower frequencies, below 10^{-5} Hz, very good limits on gravitational waves have come from observations that monitor the timing signals from pulsars. The basic idea here is that the rotation period of the pulsar's neutron star is extremely stable. By comparing the timing of the received pulsar signal with the most stable atomic clocks on Earth, one can look for fluctuations in the pulsar timing that would occur if gravitational waves pass through the intervening space. These measurements have set tight bounds on the strength of gravitational waves having periods of around 1 year.

LISA project. The proposed LISA project—the Laser Interferometer Space Antenna—would bridge some of the frequency gap between the terrestrial interferometers and spacecraft tracking. The LISA concept is to construct an interferometric detector in space, where it would be free of terrestrial background noise and the measurement baseline can be made extremely long. The plan calls for a trio of satellites to be launched into orbits around the Sun, separated from each other by 5×10^6 km (3×10^6 mi). With such a long baseline, the instrument would be sensitive to gravitational waves in the frequency band from 1 mHz to 0.1 Hz. The technology for this mission is under development.

Cosmic background measurements. Finally, there is hope that future measurements of the cosmic microwave background will show signs of gravitational radiation arising from inflation, the hypothesized epoch when the universe underwent nearly exponential expansion. Gravitational waves produced during inflation would leave their imprint on the polarization of the microwave background, which may still be detectable today.

Gravitational-wave astronomy. Physicists and astrophysicists are hopeful that these detectors will soon provide direct observations of gravitational waves, opening up the field of gravitational-wave astronomy. Our present knowledge of the universe, what it is made of and its history, has been acquired largely through observations of electromagnetic radiation. As technology progressed and each new region in the electromagnetic spectrum was opened up for observation, unexpected and important information was revealed. But for many phenomena, gravitational radiation carries information that cannot be found in the electromagnetic spectrum. Probing the universe with gravitational radiation is thus likely to bring exciting surprises and unanticipated astrophysics. *See* GRAVITATION.

Peter K. Fritschel

Bibliography. B. Barish and R. Weiss, LIGO and the detection of gravitational waves, *Phys. Today*, 52(10):44–50, October 1999; M. Bartusiak, Catch a gravity wave, *Astronomy*, 28(10):54–59, October 2000; M. Bartusiak, *Einstein's Unfinished Symphony: Listening to the Sounds of Space-Time*, Joseph Henry Press, Washington, DC, 2000; K. S. Thorne, Gravitational radiation, in S. Hawking and W. Isrrael (eds.), *3000 Years of Gravitation*, chap. 9, pp. 330–458, Cambridge University Press, 1987.

Gravitational redshift

A shift toward longer wavelengths of spectral lines emitted by atoms in strong gravitational fields. It is also known as the Einstein shift. One of three famous predictions of the general theory of relativity, this shift results from the slowing down of all periodic processes in a gravitational field. The amount of the shift is proportional to the difference in gravitational potential between the source and the receiver. For starlight received at the Earth the shift is proportional to the mass of the star divided by its radius. In the solar spectrum the shift amounts to about 0.001 nanometer at a wavelength of 500 nanometers. In the spectra of white dwarfs, whose ratio of mass to radius is about 30 times that of the Sun, the shift is about 0.03 nm, which can easily be measured if it can be separated from the Doppler effect. This was first done by W. S. Adams for the companion of Sirius, a white dwarf whose true velocity relative to the Earth can be deduced from the observed Doppler effect in the spectrum of Sirius. The measured shift agreed with the prediction based on Einstein's theory and on independent determinations of the mass and radius of Sirius B. A more accurate measurement was carried out in 1954 by D. M. Popper, who measured the gravitational redshift in the spectrum of the white dwarf 40 Eridani B. Similar measurements, all confirming Einstein's theory, have since been carried out for other white dwarfs. Attempts to demonstrate the gravitational redshift in the solar spectrum have thus far proved inconclusive, because it is difficult to distinguish the gravitational redshift from so-called pressure shifts resulting from perturbations of the emitting atoms by neighboring atoms. *See* WHITE DWARF STAR.

In 1960 R. V. Pound and G. A. Rebka, Jr., succeeded in measuring the gravitational redshift in a laboratory experiment. ^{57}Fe nuclei bound in a solid emit and absorb gamma rays in an exceedingly narrow frequency range ($\Delta\nu/\nu = 3 \times 10^{-13}$) centered on a sharply defined frequency near 14.4 keV. A source and an absorber of this accurately monochromatic radiation were separated by a vertical distance h of 74 ft (22.5 m), so that the radiation incident on the absorber was shifted in frequency by an amount given by the equation below, according to Einstein's theory, where g is the acceleration of gravity and c is the speed of light.

$$\frac{\Delta\nu}{\nu} = \frac{gh}{c^2} = 2.5 \times 10^{-15}$$

The rate of absorption was measurably diminished relative to the rate that would have been observed in the absence of the gravitational redshift. A suitable Doppler shift introduced by a small, accurately measurable relative motion of the source and the absorber restored the measured absorption rate to its normal value. The Doppler shift needed to accomplish this restoration agreed with the value predicted by Einstein's theory to within 10%. Subsequently, the error has been reduced to 1%. *See* MÖSSBAUER EFFECT.

In 1976 R. F. C. Vessot and collaborators used a rocket-borne hydrogen maser with a frequency stability of 1 part in 10^{14} to measure a gravitational redshift $\Delta\nu/\nu = 4 \times 10^{-10}$, corresponding to the difference in gravitational potential between the ground and the apogee of the rocket. This experiment confirmed the predicted redshift to better than 1 part in 10^4. *See* MASER.

Attempts have also been made to deduce stellar masses from measurements of the gravitational redshift, but the difficulty of allowing properly for the Doppler effect and for pressure shifts renders these determinations very uncertain.

Gravitational redshift has also been detected as temperature fluctuations in the observed cosmic microwave background. In 1967 R. K. Sachs and A. M. Wolfe showed that large-scale fluctuations in cosmic density should lead to different gravitational redshifts of observed microwave photons; the measurement of angular temperature variations therefore gives direct information about the mass distribution. This effect was observed in 1992 by the *Cosmic Background Explorer (COBE)* satellite, and has become a key tool in measuring the initial density perturbations from which cosmological structures are believed to have formed by gravitational instability. *See* COSMIC BACKGROUND RADIATION; COSMOLOGY; RELATIVITY.

David Layzer; Anthony Aguirre

Bibliography. R. d'Inverno, *Introducing Einstein's Relativity*, Clarendon Press, Oxford, 1996; C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, W. H. Freeman, San Francisco, 1973.

Graviton

A theoretically deduced particle postulated as the quantum of the gravitational field. According to Einstein's theory of general relativity, accelerated masses (or other distributions of energy) should emit gravitational waves, just as accelerated charges emit electromagnetic waves. And according to quantum field theory, such a radiation field should be quantized; that is, its energy should appear in discrete quanta, called gravitons, just as the energy of light appears in discrete quanta, namely photons. *See* PHOTON; QUANTUM FIELD THEORY; RELATIVITY.

The properties of the graviton follow from the properties of the classical gravitational field. That is, its rest mass and charge are zero (like the photon); it has spin 2 in units of $h/2\pi$, where h is Planck's constant, and is therefore a boson, which is a particle that obeys Bose-Einstein statistics. Because of its vanishing rest mass, its spin is restricted to be parallel to its motion, so that a graviton has only two independent spin states (again like a photon). *See* BOSE-EINSTEIN STATISTICS.

Observation of gravitational radiation are difficult, because matter is very weakly coupled to the gravitational field (the gravitational force between an electron and a proton is only 10^{-39} times the electrical force between them), so that the rate of emission and absorption of gravitational radiation is very small. The observation of quanta of the gravitational field, gravitons, is practically impossible according to present knowledge. As far as is known, the only physical situations in which the quantization of gravitation plays any appreciable quantitative role are the evaporation of black holes (light ones particularly), and the very high-temperature phase early in the history of the universe, namely within 10^{-33} s after the big bang. *See* BIG BANG THEORY; BLACK HOLE; COSMOLOGY; ELEMENTARY PARTICLE; GRAVITATION; GRAVITATIONAL RADIATION; UNIVERSE. Charles J. Goebel

Bibliography. G. Kane, *Modern Elementary Particle Physics*, updated ed., 1993; M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*, 1995.

Gravity

The gravitational attraction at the surface of a planet or other celestial body. The quantity *g* is often referred to simply as "gravity" or "the force of gravity" of Earth, both of which are incorrect. The force of gravity means the force with which a celestial body attracts an object, that is, the weight of the object. The letter *g* represents the acceleration caused by the gravitational force and, of course, has the dimensions of acceleration. *See* EARTH, GRAVITY FIELD OF; GRAVITATION. Dirk Brouwer; G. M. Clemence

Gravity meter

A device that measures local acceleration due to the Earth's gravity; it is also called a gravimeter. Such instruments fall into two categories: relative gravity meters, which are used to determine gravity differences among a number of geographic locations or changes in gravity that occur at a single location over time; and absolute gravity meters, which can measure the true value of the acceleration due to gravity at a given location and time.

The local value of gravity is the acceleration undergone by a freely falling mass upon which gravity is the only force acting. Because the value of gravity at any particular position depends on the distribution of mass throughout the Earth (and also slightly on the Earth's rotation), measurements of the gravity field can yield information on the density of underlying rock. Thus, gravity meters are used for geologic studies and for oil and mineral exploration. Local gravity also depends on the shape of the Earth; the observation of gravity over time, then, provides a measure of deformations in the Earth that can be caused by a wide variety of phenomena, including tides, tectonic activity, and volcanism.

Nowhere on the surface of the Earth does the value of gravity differ from the nominal value of 980 Gal by more than about 0.5%. (The SI unit for gravity is the meter/second^2; the more commonly used unit is the Gal, defined as 1 cm/s^2, or the milliGal, which equals 0.001 Gal.) Values of gravity predicted with a latitude-and-height-dependent Earth model usually agree with observed values to within about 30 mGal. The gravitational acceleration produced by the mass of a 1-m-thick (3-ft) sheet of water (having infinite lateral extent) is 0.043 mGal.

A number of different instruments are available. One of two methods is used in all gravity meters. The first, employed by absolute gravity meters, is the direct determination of the acceleration of a test mass falling inside a vacuum chamber by using optical interferometry. The second is the observation of variations in the position of a mass supported by a mechanical or magnetic spring. This method, applied in relative gravity meters and shipboard gravity meters, is usually used in conjunction with an additional applied force (nulling force) that maintains the mass at a null position. The small nulling force is a relative measure of gravity.

Absolute gravity meters. In this type of gravity meter a mechanical system inside a high-vacuum chamber effects the release of a test mass (**Fig. 1**). A corner-cube prism mounted to the test mass

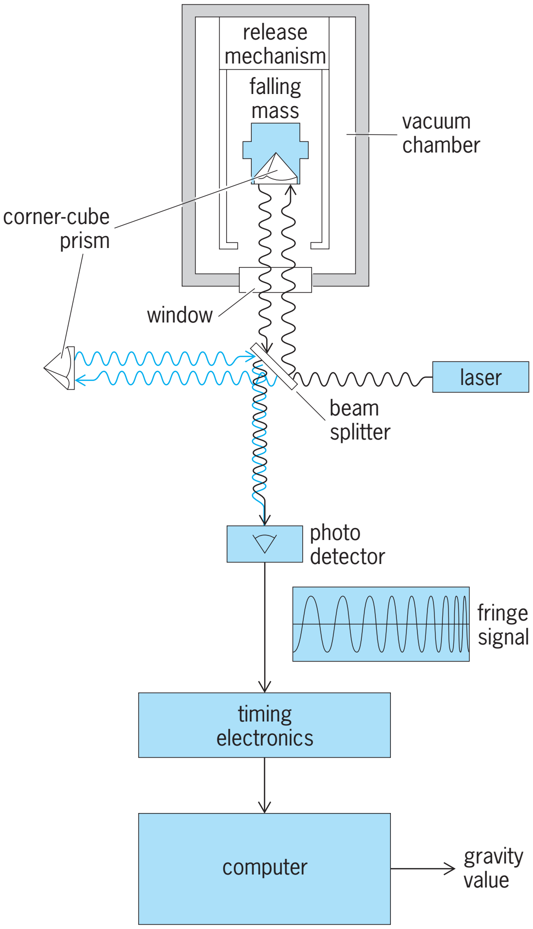


Fig. 1. Operation of an absolute gravity meter.

reflects laser light entering the chamber through a window in the bottom. A beamsplitter (or half-silvered mirror) combines this reflected light with that similarly reflected from a fixed corner-cube. The two reflections interfere with one another, producing successive light and dark fringes. A photodetector senses the fringes and generates an electrical signal from which the local value of gravity (or the rate at which the falling mass accelerates) can be calculated through appropriate timing analysis. Typically, the mass falls a distance of 10–20 cm (4–8 in.). Because the measurement is based on calibrated length and time standards—the wavelength of the laser light and the frequency of the timing oscillator—the measurement is absolute. *See* INTERFEROMETRY.

The mechanical systems vary. In some, the test mass's fall is initiated by a small elevator inside the vacuum chamber, which also surrounds the mass as it falls and acts as a shield against minute nongravitational forces (Fig. 1). In others, the mechanical system resides at the bottom of the vacuum chamber and launches the test mass upward, providing a mass in free flight, the acceleration of which can be measured during both its rise and fall. Another important variation is the use of a spring suspension system to isolate the reference reflector from vibrations.

The best instruments achieve an accuracy approaching 0.002 mGal. A complete gravity determination at a single location requires 1–2 days, the time required to make a thousand measurements (drops), the average of which is required to smooth out imperfectly corrected environmental and tidal effects. These instruments weigh several hundred pounds but are fairly portable.

A very recent technique for making an absolute measurement of gravity involves measuring the acceleration of individual atoms in free fall. While largely a laboratory exercise (but with promise for a portable instrument eventually), this technique relies on the laser-cooling and trapping of atoms. The accelerations of falling cesium atoms are determined with an atom interferometer; accuracy is similiar to that of the macroscopic free-fall method.

Relative gravity meters. These are more common than absolute gravity meters. In a relative gravity meter, a mass is suspended by a spring and its position is monitored electronically with respect to some reference position (**Fig. 2***a*). As the force of gravity on the mass changes (because, for example, the meter has been moved to a different location), the length of the spring changes slightly in order to maintain equilibrium. An additional force applied to the mass to restore its position to the original reference point is a measure of the difference in gravity between the two observation sites. In some cases, the mechanical spring is replaced by a magnetic suspension system.

Spring gravity meters. For a given change in gravity, the change in length of the spring in a relative gravity meter is proportional to the square of the period of the spring mass system. Maximizing the period enhances the meter's sensitivity; this, however, involves softening the spring and increasing the length of the suspension system. In one gravity meter of this type, a hinged beam connects the test mass to the instrument frame and the spring is oriented diagonally (Fig. 2*b*). This suspension, while requiring only a few centimeters of length, achieves a natural period in the spring-mass system of about 15 s (a vertically oriented spring having this period would stretch to a length of 55 m or 180 ft). The meter is read by noting the amount of rotation in a spring-adjusting screw (nulling screw) that is required to restore the beam to its original equilibrium position. The instrument can sense gravity differences as small as 0.003 mGal. This type of instrument is the most common type of gravity meter. Because the instrument weighs only a few pounds, can be easily transported, and requires only battery power, it is used extensively in geophysical surveying and oil exploration. *See* GEOPHYSICAL EXPLORATION.

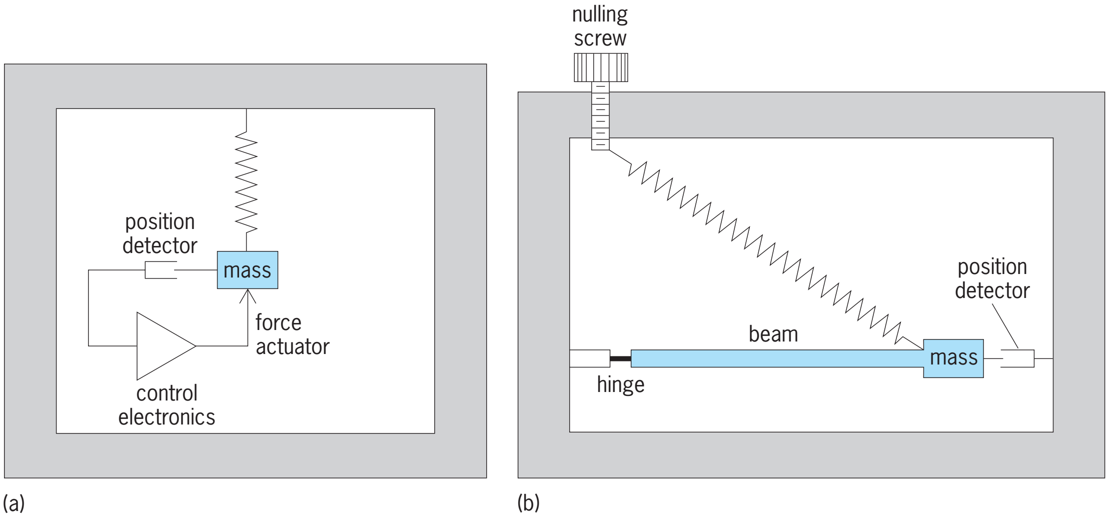


Fig. 2. Relative gravity meters. (*a*) Mass suspended by a spring whose position is monitored electronically. (*b*) Mass connected to the instrument frame by a hinged beam, with the spring oriented diagonally. (*LaCoste & Romberg*)

Superconducting gravity meters. Another type of relative gravity instrument is the superconducting gravity meter. In this instrument a 2.5-cm-diameter (1-in.) superconducting sphere is levitated by persistent currents in a pair of superconducting coils immersed in a liquid helium bath. The sphere's position in sensed capacitively by electrodes, which are also employed to apply an electrostatic force on the sphere to maintain it at a constant position. The voltage required to maintain the null position of the sphere changes as the local value of gravity changes. Superconducting gravity meters are the most sensitive gravity recording devices available, capable of detecting changes in gravitational acceleration as small as 0.0001 mGal. Because these meters must be operated inside liquid helium dewars, they are not easily moved; generally, they are used to monitor changes in gravity that occur over time at fixed locations. *See* DEWAR FLASK; LIQUID HELIUM; SUPERCONDUCTING DEVICES.

Shipboard gravity meters. Gravity variations are frequently mapped from ships or airplanes. The accelerations encountered by a moving ship are significant compared to the minute variations in gravitational acceleration produced by the suboceanic density structures of interest. By averaging the gravity field over several minutes, however, changes of 1 mGal can be detected. The gravity sensor must be mounted on a stabilized platform to keep it aligned with the gravity vector (local vertical). This is accomplished by means of a double-gimbal motorized platform, the position of which is controlled electronically by using gyroscopes and accelerometers. *See* ACCELEROMETER; GYROSCOPE.

Two types of shipboard gravity systems are in wide use. One is similar to the land meter that uses a diagonally oriented spring (Fig. 2*b*), except that the motion of the beam-mass pair is heavily overdamped. In another commonly used shipboard gravity meter, a wire coil is wound around a pivoted mass held between the poles of a permanent magnet. The current in the coil required to maintain the mass at a null position is proportional to gravity. *See* GRAVITY.

Mark A. Zumberge

Bibliography. O. Francis et al., Calibration of a superconducting gravimeter by comparison with an absolute gravimeter FG5 in Boulder, *Geophys. Res. Lett.*, 25(7):1075–1078, 1998; I. Marson and J. E. Faller, *g*, the acceleration of gravity: Its measurement and its importance, *J. Phys. E: Sci. Instrum.*, 19:22–32, 1986; A. Peters, K. Y. Chung, and S. Chu, Measurement of gravitational acceleration by dropping atoms, *Nature*, 400(6747):849–852, 1999; W. M. Telford, L. P. Geldart, and R. E. Sheriff, *Applied Geophysics*, 2d ed., 1990; W. Torge, *Gravimetry*, 1989.

Graybody

An energy radiator which has a blackbody energy distribution, reduced by a constant factor, throughout the radiation spectrum or within a certain wavelength interval. The designation "gray" has no relation to the visual appearance of a body but only to its similarity in energy distribution to a blackbody. Most metals, for example, have a constant emissivity within the visible region of the spectrum and thus are graybodies in that region. The graybody concept allows the calculation of the total radiation intensity of certain substances by multiplying the total radiated energy (as given by the Stefan-Boltzmann law) by the emissivity. The concept is also quite useful in determining the true temperatures of bodies by measuring the color temperature. For a discussion of the Stefan-Boltzmann law and color temperature *see* HEAT RADIATION; BLACKBODY.

Heinz G. Sell; Peter J. Walsh

Graywacke

A well-indurated dark gray sandstone that is characterized by abundant dark-colored detrital rock fragments and more than 15% clay matrix minerals between sand grains. Graywacke sands were deposited chiefly in marine basins near the edge of continental margins where plate subduction was taking place. Subsequent compressional deformation and uplift of rocks in the sedimentary basins results in the occurrence of most graywackes in Alpine-type (compressional) mountain ranges. *See* CLAY MINERALS; CONTINENTAL MARGIN.

The term graywacke evolved from initially informal usage for sandstones examined as hand specimens in the field to later formal usage for sandstones defined rigorously on microscopic criteria. Graywacke was applied initially to sandstones of Late Devonian to Early Carboniferous age in the Harz Mountains of Germany that are distinctive in that they are nonporous, strongly indurated, clayey, and dark gray. Geologists who had seen the Harz Mountain exposures later applied the term graywacke to sandstones of similar appearance elsewhere in the world. It was not until the 1950s and 1960s that descriptions in English were made of the microscopic texture and mineral composition of graywackes. P. D. Krynine, F. J. Pettijohn, and R. L. Folk provided important descriptions of graywacke and incorporated the term into formal sandstone classifications, but their definitions all differed. Krynine defined graywacke as being composed of more than 25% detrital micas and micaceous rock fragments; Folk defined it as being composed of more than 25% detrital grains derived from (exclusive of micas) metamorphic rocks; and Pettijohn defined it as dark gray, well-indurated sandstone that contained more than 15% detrital clay matrix. All three experts studied the same type of sandstone, but they based their definitions on different criteria. Several additional definitions have been proposed since these studies were completed. Because of the confusion that developed from its use, the term graywacke has been omitted from most formal classification schemes. It is used, however, as a field term referring to sandstone with the characteristic detrital rock fragments and clay matrix minerals.

Mineral composition. Graywackes have a wide range in mineral composition, which reflects the varied source rocks from which the detritus in them was derived. They tend to be quartz-poor (10–50%), to be rich in both feldspar and unstable rock fragments, and to contain several percent of unstable accessory minerals such as micas, pyroxenes, and amphiboles. Feldspathic graywackes (those in which feldspar exceeds rock fragments) are derived chiefly from plutonic cores of denuded island arcs. Lithic graywackes (those in which rock fragments exceed feldspar) are derived either from volcanic island arcs or from sedimentary rocks in adjacent basins that were deformed and uplifted. Volcanic rock fragments characterize the former type of lithic graywackes, whereas sandstone, shale, and their weakly metamorphosed equivalents characterize the latter type.

Matrix origin. The term matrix is applied to mineral particles which are less than 0.001 in. (0.03 mm) in diameter and which are interstitial to coarser, sand-size grains. In graywacke, such matrix minerals are chiefly illite and chlorite. In the 1950s it was believed that matrix was composed of detrital particles deposited simultaneously with the sand grains. The poor sorting reported for many graywackes was interpreted as the result of deposition of the sediment from muddy marine turbidity currents. Subsequent work demonstrated that "matrix clays" are either (1) sand grains of claystone and shale that were mashed between more resistant quartz grains so that evidence of this detrital origin is obscured or obliterated; (2) chemically precipitated cements (such as chlorite, smectite, or other clay species) that formed when water in pores in the sand reacted with detrital grains, chiefly volcanic rock fragments; or (3) clay minerals of origins 1 and 2 that underwent recrystallization upon encountering higher temperatures and pressure upon deep burial.

Depositional origin. Most graywackes were deposited in submarine fans and adjacent basin-plain environments by turbidity currents. They commonly display graded bedding, Bouma sequences, and current-formed and biogenic sole marks. The term Bouma sequence refers to five divisions of a single, ideal turbidity current deposit. Graywackes are interbedded with shale beds that were deposited by dilute turbidity currents and other marine processes. Thicknesses of several miles of interbedded turbidite graywacke and shale accumulated in many basins. Burial and subsequent compressional deformation of these sequences resulted in the generation of clay matrix, loss of porosity, and strong induration. The gray color of the sandstone is derived from rock fragments and organic-stained clay minerals. *See* ARKOSE; SANDSTONE; SEDIMENTARY ROCKS; SHALE; TURBIDITE; TURBIDITY CURRENT. Earle F. McBride

Bibliography. R. L. Folk, *Petrology of Sedimentary Rocks*, 1982; P. D. Krynine, *The Megascopic Study and Field Classification of Sedimentary Rocks*, 1948; F. J. Pettijohn and P. E. Potter, *Sand and Sandstone*, 2d ed., 1987.

Great Basin

A semiarid region of the western United States characterized by internal drainage and a distinctive topography of narrow, elongate, predominantly north-trending mountain ranges with intervening deep valleys. The Great Basin is about 400,000 km^2 (144,000 mi^2) in area and includes most of Nevada and western Utah, as well as parts of eastern California, southeastern Oregon, and southern Idaho (**Fig. 1**). The region is bounded by the Sierra Nevada mountains to the west, the Columbia Plateau to the north, the Wasatch Range and Colorado Plateau to the east, and the Mojave Desert to the south. The province was named by John Fremont, who explored the area in 1843–1845 and recognized that the region is internally drained, with no rivers having outlets to the sea.

Landscape features. The Great Basin forms the northern part of the Basin and Range physiographic province, named for its characteristic topography of narrow fault-block mountain ranges separated by broad flat valleys, many of which form closed basins (**Fig. 2**). Extensional block faulting has resulted in uplift of the mountains relative to the valley floors along range-bounding normal faults. Generally, east-west extension and stretching of the continental crust began about 20 million years ago and is still active in much of the Great Basin.

Faulting and localized volcanism has blocked drainages and contributed to the formation of the internal drainage system that defines the region today. Within the Great Basin, rivers have no outlet to the sea but terminate in local depressions, where they form playas or saline lakes. The longest river is the Humboldt, which flows west and then south across Nevada, eventually terminating in the Humboldt Sink. Major lakes include the Great Salt Lake,

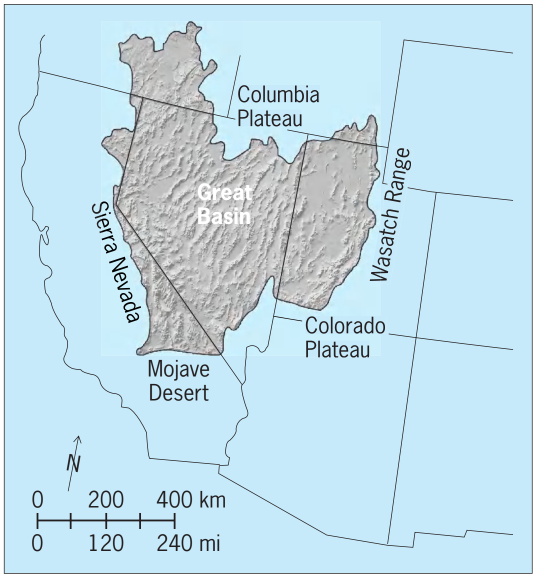


Fig. 1. Location of the Great Basin and bounding topographic features.

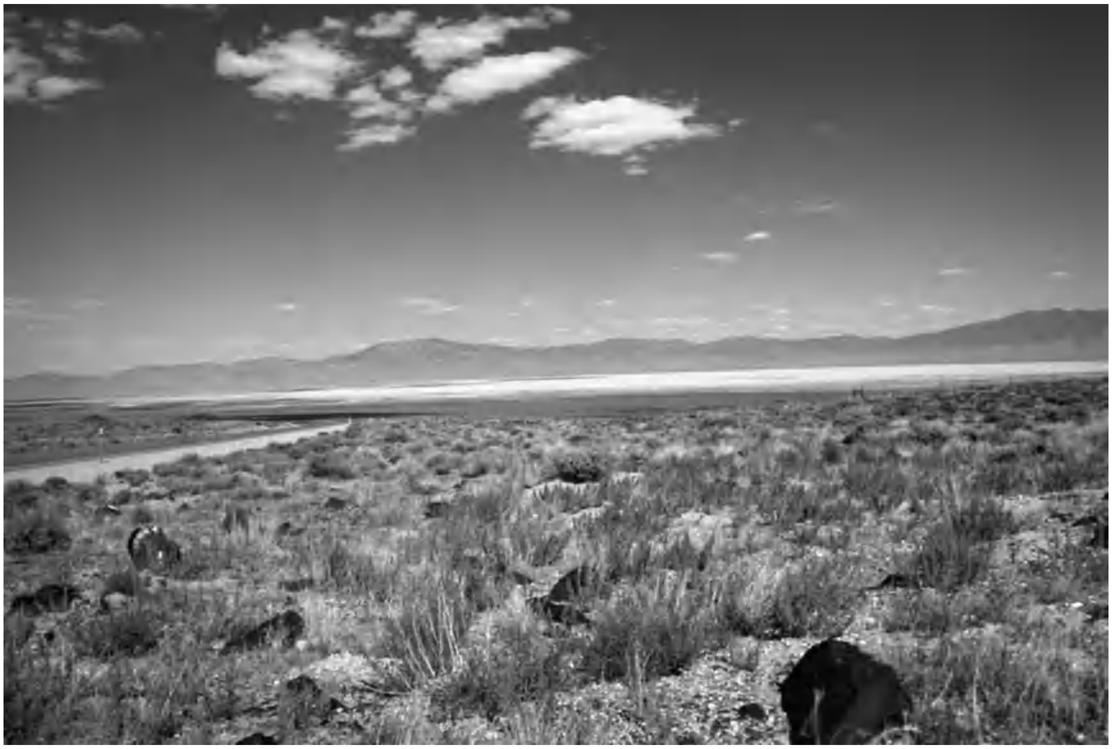


Fig. 2. Railroad Valley and the Grant Range, a typical basin and range in eastern Nevada.

fed by runoff from the Wasatch Mountains, and Mono and Pyramid lakes, fed by runoff from the Sierra Nevada. *See* BASIN; DESERT; DESERT EROSION FEATURES; FAULT AND FAULT STRUCTURES; PLATE TECTONICS; PLAYA; VOLCANO.

Climate and vegetation. The Great Basin is generally arid, as rainshadow effects caused by the bounding mountain ranges block the flow of moisture into the region from both the Pacific Ocean to the west and the Gulf of Mexico to the southeast. Precipitation averages between 100 and 300 mm (4 and 12 in.) per year, but is highly variable from location to location and from year to year. Rainfall increases with elevation, and the mountain ranges receive significantly more precipitation than adjacent valleys. The Great Basin is considered a "cold" desert, with mean winter temperatures below 0°C (32°F) and most winter precipitation falling as snow.

Currently, vegetation in the Great Basin is dominated by sagebrush (*Artemisia* spp.), with few trees except for some pinion-juniper woodlands at higher elevations and cottonwoods along perennial streams. The ecology has been significantly altered by the spread of introduced invasive species, including cheatgrass, Russian thistle, and tamarisk that thrive in the disturbed areas created by fires and overgrazing.

Human impact. The Great Basin remains one of the least populated regions of the continental United States. Rapid population growth and urbanization have centered on the cities of Las Vegas, Reno, and Salt Lake City, all located on the borders of the region. Outside of these urban areas, cattle grazing and mining remain the major economic activities. Human environmental impact on the region includes overgrazing and resulting vegetation changes; land disturbance and waste rock from mining operations; lowered water tables and loss of surface stream flows due to heavy use of water resources for irrigation; and the interbasin transfer of water to supply rapidly growing populations in cities such as Las Vegas and Los Angeles. David C. Greene

Bibliography. J. Hudson, *Across This land: A Regional Geography of the United States and Canada*, Johns Hopkins University Press, 2002; J. Laity, Desert environments, in A. Orme (ed.), *The Physical Geography of North America*, pp. 380–401, Oxford University Press, 2002; S. Trimble, *The Sagebrush Ocean: A Natural History of the Great Basin*, University of Nevada Press, 1989.

Great circle, terrestrial

A circle or near-circle representing a trace on the Earth's surface of a plane that passes through the center of the Earth and divides it into equal halves (**Fig. 1**). The Equator is a great circle, the trace of the plane that bisects and is perpendicular to the Earth's axis. Planes through the Poles cut the Earth along meridians. All meridians are great circles; actually, they are not quite circular because of the slightly flattened Earth. The equatorial diameter is 1.0034 times the size of the polar diameter. All parallels other than the Equator are called small circles, being smaller than a great circle.

The shortest distance between two points on the Earth's surface follows a great circle arc. It is therefore advantageous for commercial airlines to plan long-distance, nonstop flights along great circle routes called orthodromes. Also important in navigation are rhumb lines or loxodromes, which are lines of constant compass directions. Except where they coincide with meridians or the Equator, loxodromes are not great circle arcs.

By using two map projections—the gnomonic and the Mercator—one can navigate by compass and follow great circle routes. The gnomonic is the only map projection on which any straight line is a great circle arc. The Mercator projection shows rhumb lines as straight lines. Navigators plot the origin and destination of trips on a gnomonic projection to obtain shortest routes. They then select from the gnomonic projection points along a route for transfer to a Mercator projection. A series of line segments or rhumb lines connecting points on the Mercator projection thus approximate the shortest route. Using the charted route on the Mercator projection, navigators can follow the compass bearing except when changing it is required at intermediate points.

Although sophisticated instruments have replaced the compass as a navigational tool, intercontinental air routes are still laid out as nearly as possible to true great circles. For example, a trip from Seattle to Tokyo passes near Anchorage, Alaska, and the Aleutian Islands, and a direct flight from Seattle to Amsterdam crosses Canada's Northwest Territories, Iceland, and Scotland. Both are great circle routes.

Despite its desirable property of showing great circle arcs as straight lines, the gnomonic projection cannot show an entire hemisphere and greatly exaggerates shapes and sizes away from its center. Moreover, direct distance measures on the projection are unreliable.

Two common methods can be used to calculate the distance of a great circle arc. One method uses trigonometric functions: $\cos D = \sin a \sin b + \cos a \cos b \cos c$. Here D is the arc distance between points A and B in degrees, a is the latitude of A, b is the latitude of B, and c is the difference in longitude between A and B. After D is calculated, it can be converted to a linear distance measure by multiplying D by the length of one degree of the Equator, which is 111.32 km or 69.17 mi.

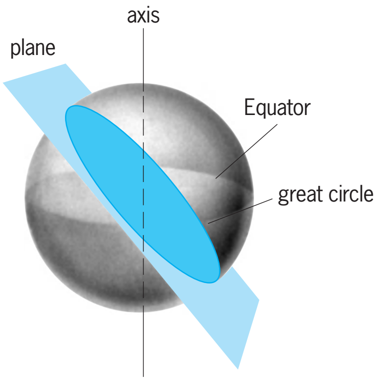


Fig. 1. Diagram of a great circle described by a plane through the center of the Earth.

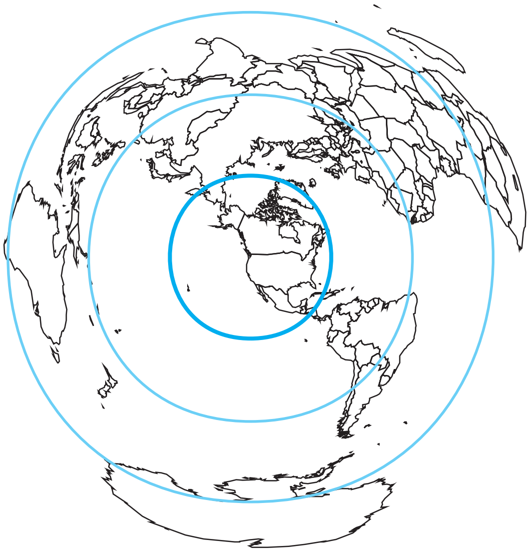


Fig. 2. Azimuthal equidistant projection centered at Seattle, Washington. Circles away from Seattle have an interval of 5000 km (3120 mi).

The other method uses the azimuthal equidistant projection (**Fig. 2**). Unlike the gnomonic projection, the azimuthal equidistant projection can be centered at any point on the Earth's surface and can show the entire sphere. More importantly, a straight line from the center of the projection to any other point is a great circle route and the distances are at a comparable (consistent) scale between the two points. The azimuthal equidistant projection is therefore useful in showing any movement directed toward or away from a center, such as seismic waves, radio transmissions, missiles, and aircraft flights. Kang-tsung Chang

Bibliography. B. D. Dent, *Cartography: Thematic Map Design*, 5th ed., William C. Brown Publishers, 1998; A. H. Robinson et al., *Elements of Cartography*, 6th ed., John Wiley, 1995.

Green chemistry

The discovery, development, and application of sustainable chemical reactions that create no waste or generate only benign wastes. The idea that toxic chemicals, once created, can be contained or prevented from entering the environment is unrealistic. Since some portion of the toxins will always enter the environment, whether through use, processing, disposal, or unforeseen circumstances such as accidents, terrorism, or natural disasters, green chemistry solutions are applied at the molecular level to avoid creating toxins in the first place. Using green chemistry to prevent pollution provides a more efficient approach to preserving human health and the environment than waste processing or remediation.

Developing practical approaches for sustainable manufacturing is one of the major goals of green

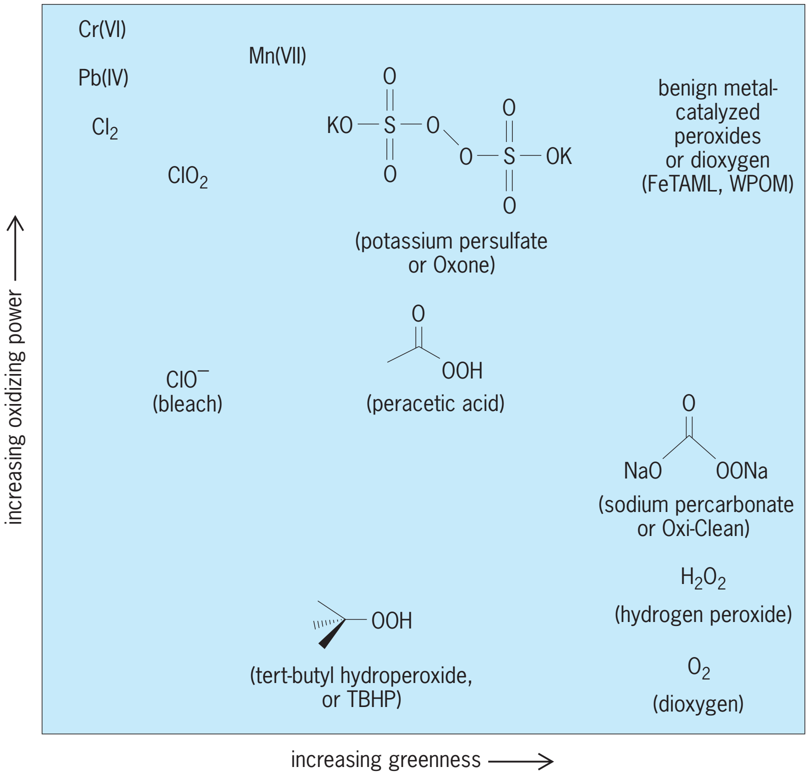

Cr(VI)
Pb(IV)
Cl$_2$
Mn(VII)
ClO$_2$
KO—S—O—O—S—OK
(potassium persulfate or Oxone)
benign metal-catalyzed peroxides or dioxygen (FeTAML, WPOM)
OOH
(peracetic acid)
ClO$^-$
(bleach)
NaO
OONa
(sodium percarbonate or Oxi-Clean)
H$_2$O$_2$
(hydrogen peroxide)
OOH
(tert-butyl hydroperoxide, or TBHP)
O$_2$
(dioxygen)
increasing oxidizing power ⟶
increasing greenness ⟶

Fig. 1. Comparison of some common oxidants in terms of oxidizing power versus greenness.

chemistry. Green chemistry faces many challenges, including social, business, political, and technological barriers. Even within the field of chemistry, a major challenge exists to shift the focus of mainstream chemical research away from just developing compounds and processes to solving short-term (1–5 years) problems of practical and economic importance. Green chemistry is acting now to ameliorate the long-term environmental impact (50–200 years) of chemical technologies through molecular-level design, acting in concert with green principles and a sustainable vision of the future.

Roots of green chemistry. The 12 principles of green chemistry put forth by P. T. Anastas and J. C. Warner provide an effective framework for developing greener methods. In summary, the principles are (1) prevent waste before it is created; (2) be atom-economical; (3) use benign substances; (4) make benign products; (5) use less solvents; (6) use less energy; (7) use renewables; (8) avoid protecting groups; (9) use catalysts; (10) make things biodegradable; (11) analyze in real time; (12) be safe.

When every atom in the reactants appears in the products, a maximum of atom economy is reached. A similar concept is the efficiency or E factor, which is equal to the mass of inputs for a reaction process (including solvents and purification steps) divided by the mass of useful products obtained from a reaction. The lowest possible E factor is 1. Most industrial processes have E factors ranging from about 3–5 for high-volume chemicals such as acetic acid or monomers/polymers and 20–100 for high-value chemicals such as pharmaceuticals or semiconductors. Improving atom economy or lowering the E factor simultaneously saves money and protects the environment, unless doing so substantially increases the energy requirements for the reaction.

Examples. Two representative examples of green chemistry are oxidative catalysis and alternative solvents.

Oxidative catalysis. Few oxidants are simultaneously inexpensive and reactive enough to be commercially viable. **Figure 1** shows the main commercial oxidants and ranks them qualitatively in terms of oxidizing power versus greenness. In order to replace environmentally undesirable oxidizing systems such as heavy-metal ions or active chlorine, switching to oxygen-based oxidants is not enough. Additional activation is necessary, preferably catalytic activation with a benign metal ion via an enzymatic or synthetic catalyst that satisfies green principles.

T. J. Collins and coworkers have developed a viable green hydrogen peroxide (H_2O_2)-activating catalyst using an iron (Fe) complex of an oxidation-resistant tetra amide macrocyclic ligand (TAML). FeTAML complexes are constructed from benign components and activate H_2O_2 in water over a wide pH range. Long-lived under oxidizing conditions, FeTAML complexes efficiently bleach lignins, dyes, or chlorophenols at very low catalyst concentrations. The main commercial application of FeTAML complexes is for totally chlorine-free (TCF) paper bleaching. TCF paper bleaching avoids the formation of dioxins and absorbable organic halogens, unlike paper bleaching with chlorine dioxide (ClO_2).

N. Mizuno and coworkers have developed a tungsten polyoxometallate (WPOM) catalyst that activates hydrogen peroxide to perform selective olefin (double-bond) oxidation reactions. The catalyst uses hydrogen peroxide efficiently and exhibits very high selectivities, useful features for eventual commercial development as an alternative to the use of lead-, chromium-, manganese-, or organic peroxide-based oxidants. Molybdenum polyoxometallate (MoPOM) complexes, developed by C. Hill and others, activate oxygen (O_2) for oxidizing a variety of different substrates, including lignin bleaching for TCF paper manufacture and in-situ detoxification of chemical warfare agents. *See* CATALYSIS; HYDROGEN PEROXIDE; OXIDATION PROCESS.

VOC alternatives. Most solvents used today are volatile organic compounds (VOCs). VOCs readily escape to the atmosphere when used, causing a substantial fraction of all air pollution. Eliminating VOCs is environmentally desirable but requires that practical and economical VOC solvent alternatives be developed.

Supercritical water and supercritical carbon dioxide (CO_2) continue to provide successful green approaches for replacing VOCs in chemical processes such as decaffeinating coffee, dry cleaning, and demanding chemical reactions.

R. Rogers and others have developed a new class of solvents called room-temperature ionic liquids (RTILs). Many RTILs are based on chloroaluminate anions or alkyl imidazolium cations. Most RTILs exhibit a low melting point, a high boiling point, and a high viscosity. As solvents, RTILs have extremely low vapor pressures, an important green feature for replacing VOCs and decreasing atmospheric pollution. Many of the chemical properties of ionic liquids such as tunable polarity, good dissolving power for organic molecules, and easy drying are identical to those of VOCs. The major hurdles for commercialization of RTILs are cost, limited toxicological data, and difficulties associated with removing solutes from the RTIL to allow the ionic liquid to be recycled or reused.

C. J. Li and coworkers have developed an arsenal of useful synthetic organic reactions that take place in water instead of VOCs. Controlling chemical reactions in water is an important green objective; however, care must be taken that water-soluble reagents do not leak out of the chemical process and become widely dispersed in the environment.

IBM has pioneered a technique that uses soap and water to degrease silicon wafers, thereby avoiding the use of chlorinated fluorocarbons (CFCs, or Freons) during semiconductor manufacturing. Another CFC-free wafer-cleaning technology has been developed by J. DeSimone that uses surfactants and supercritical CO_2 as the solvent. *See* IONIC LIQUIDS; SOLVENT; SUPERCRITICAL FLUID.

New approaches. Exciting developments in green chemistry are coming from new ideas and research projects that use the green framework as a starting point instead of trying to retrofit green solutions onto existing technologies.

Supramolecular synthesis. J. C. Warner and coworkers have used supramolecular chemistry (which aims to generate complex chemical systems from components bound together by noncovalent intermolecular forces) to challenge the fundamental idea of forming covalent bonds as a means of controlling chemical reaction pathways. When an undesirable covalent bond is formed, it is difficult to cleave and often leads to the production of chemical waste or to the use of high-energy conditions to try to control the process of covalent bond formation. Supramolecular interactions are much weaker than covalent bonds and therefore more reversible. Warner has used weak supramolecular interactions to direct the formation of titanium dioxide (TiO_2) at room temperature, thereby avoiding the use of temperatures as high as 1000°C (1832°F), to synthesize the next generation of dye-sensitized TiO_2 solar photovoltaic cells. In contrast, current methods for manufacturing solar cells are very energy intensive, thus rendering solar technology expensive and compromising most, if not all, of the benefit derived from obtaining energy from the sun. Currently, most energy used in manufacturing processes is derived from combustion, including the synthesis of solar cells. *See* SOLAR CELL; SUPRAMOLECULAR CHEMISTRY; TITANIUM OXIDES.

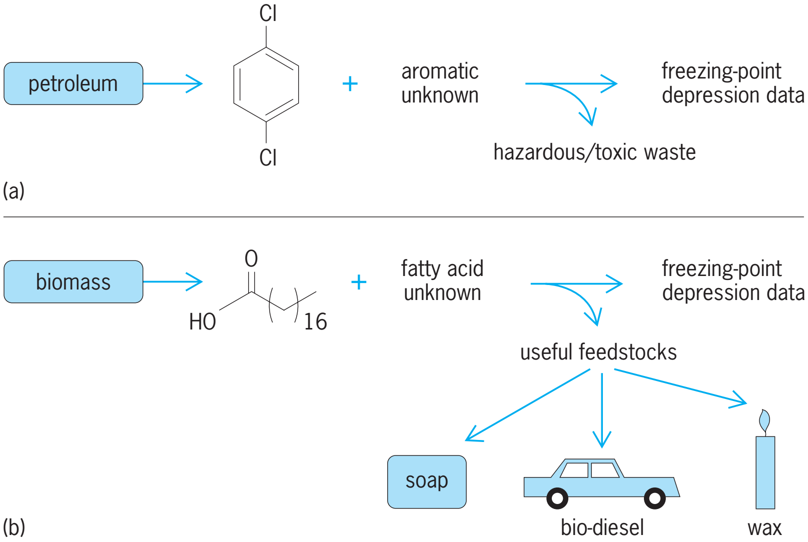


Fig. 2. Comparison of laboratory experiments for measuring colligative properties (which depend on the number of atoms or molecules but not their nature). (*a*) Old laboratory approach uses an exhaustible feedstock, uses hazardous reagents, creates toxic waste, and has an E factor of ∞. (*b*) New greener laboratory approach uses a renewable feedstock, uses benign reagents, creates almost zero waste, and approaches an E factor of 1.

Copper complexes. In the United States, most exterior wood is pressure-treated with chromated copper arsenate (CCA). Pressure-treated wood is green in color because of the inclusion of copper, but its antifungal and anti-insect properties are typically attributed to the use of chromium and arsenic. CCA-treated wood is used for playground lumber, exposing children to hexavalent chromium or arsenic. In addition, as the CCA leaches out of the treated wood into waterways, background levels of chromium and arsenic in drinking water are increased. Quaternary alkylammonium copper complexes have been developed by Chemical Specialties Inc. for pressure-treating wood that exhibit lifetime, antifungal, and anti-insect properties comparable to CCA-treated wood, eliminating the need for chromium or arsenic. This technology provides a tremendous step forward using some relatively simple green chemistry. *See* COPPER.

Poly(lactic acid). P. Gruber at Cargill-Dow has applied a green vision to establish a large-scale poly(lactic acid) plant that makes plastic from corn. The process depends on a renewable feedstock (corn) instead of an exhaustible one (petroleum). Variation in the chemical and physical properties of the resulting plastics is achieved by controlling the separation and chemical modification steps used on the corn feedstock. The process steps have been designed to be as green as possible, including incentives for the corn suppliers to use sustainable agricultural practices.

Education. Education plays a key role in green chemistry. How people learn chemistry affects how they will solve real-world problems. If chemical education fails to incorporate green principles, the next generation of chemists will continue with waste-generating practices. **Figure 2** compares a conventional chemical laboratory experiment with one designed to incorporate green principles. The new, greener procedure preserves all of the pedagogical elements of the older laboratory, but also teaches the powerful lesson that green chemistry can create knowledge and compounds without harming the natural world or creating toxic waste.

S. W. Gordon-Wylie

Bibliography. P. T. Anastas and J. C. Warner, *Green Chemistry: Theory and Practice*, Oxford University Press, 1998; M. C. Cann and M. E. Connelly, *Real World Cases in Green Chemistry*, American Chemical Society, 2000; S. S. Gupta et al., Rapid total destruction of chlorophenol pollutants by activated hydrogen peroxide, *Science*, 296:326–328, 2002; K. Kamata et al., Efficient epoxidation of olefins with ≥99% selectivity and use of hydrogen peroxide, *Science*, 300:964–966, 2003; R. J. Lempert et al., *Next Generation Environmental Technologies: Benefits and Barriers*, RAND Science and Technology Policy Institute, 2003; W. M. Nelson, *Green Solvents for Chemistry: Perspectives and Practice*, Oxford University Press, 2003; R. D. Rogers and K. R. Seddon, *Ionic Liquids: Industrial Applications for Green Chemistry*, American Chemical Society, 2002; B. M. Trost, The atom economy: A search for synthetic efficiency, *Science*, 254:1471–1477, 1991.

Green fluorescent protein

A protein produced by the bioluminescent jellyfish *Aequorea victoria* that fluoresces in the lower green portion of the visible spectrum. The gene for green fluorescent protein (GFP) has been isolated, and the protein has been extensively characterized biochemically and found useful in biological research. After synthesis of GFP, three of its amino acids undergo a self-catalyzed reaction that produces an internal fluorophore (fluorescent molecule). The protein does not require a specific cellular environment to become fluorescent and can be artificially expressed in cell types as diverse as bacterial, plant, and animal cells. It can be attached to other proteins of interest using recombinant DNA techniques, making it possible to easily trace the synthesis, location, and movement of these proteins in single living cells using conventional fluorescence microscopy. *See* BIOLUMINESCENCE; FLUORESCENCE; FLUORESCENCE MICROSCOPY; HYDROIDA.

Biological role. The natural function of the GFP in the jellyfish is to convert the blue light emitted by a calcium-regulated protein called aequorin into green light. When the animal is disturbed (for example, when it is touched lightly), calcium stores inside some of its cells are released, causing excitation of the aequorin. The excitation energy is then transferred to GFP (probably by a quantum-mechanical process) and released as a brief green glow. The biological significance of this process is still not known. When expressed alone, the protein (in the absence of aequorin) is fluorescent and emits green light when excited with ultraviolet light, and to some extent when excited with blue light.

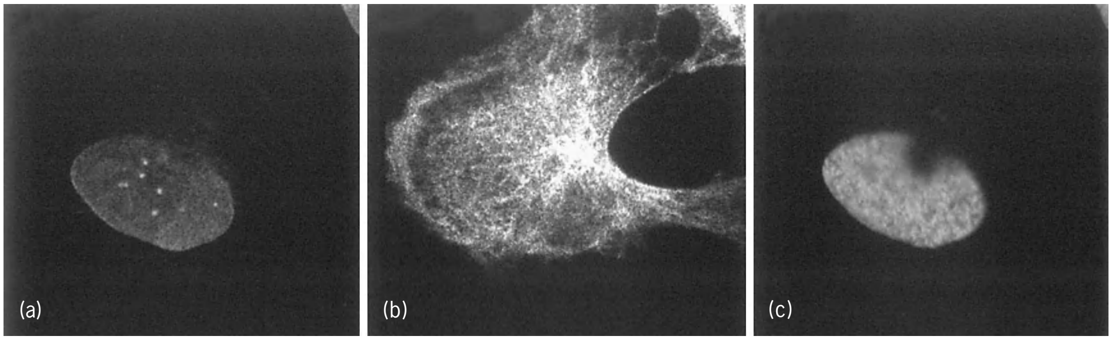
Tracking and quantitation. Green fluorescent protein technology, by allowing observations of processes in living cells over time, has begun to provide new biological insights. GFP can be attached to a protein marker for specific organelles, such as mitochondria or the Golgi apparatus, allowing the organelle's location and morphology to be followed over time. Changes in organelle structure in response to drug treatments, or natural processes such as differentiation or mitosis can be studied.

Green fluorescent protein tagging also makes it possible to follow the trafficking of proteins between different parts of the cell or between different organelles (via video microscopy). Moreover, the exact levels of the tagged protein in different organelles can be quantitated, since the intensity of light emission from GFP is directly proportional to the number of molecules. It is possible to quantitate the absolute number of GFP molecules in an organelle of interest. *See* VIDEO MICROSCOPY.

A laser scanning confocal microscope, which scans a laser over the sample, can be used for imaging GFP within cells or within small organisms, and unlike conventional microscopes it can map the location of GFP in three dimensions. GFP, like other fluorophores, can be photobleached (rendered nonfluorescent) when exposed to intense light. The laser in a confocal microscope provides an ideal tool for selective bleaching of GFP in a particular region of an organelle or of an entire organelle. Recovery by diffusion of GFP into the bleached area or by transport from other organelles can be monitored, providing important information about the mobility of the protein not previously available. This technique, known as fluorescence recovery after photobleaching (FRAP), has been utilized to show that resident proteins of the Golgi apparatus are freely mobile (rather than anchored to fixed positions), and to track movement of proteins between organelles. *See* CONFOCAL MICROSCOPY; CYTOCHEMISTRY.

Other uses. Green fluorescent protein can also be placed under the control of promoters (regulatory DNA sequences) that allow its selective expression only in specific cell types, for example, neurons. This has proved highly useful in studies of developmental biology for examining the differentiation and migration of a wide variety of cell types in development, particularly in small, relatively transparent organisms such as *Drosophila* embryos, nematodes, and zebrafish. It is also possible to do genetic studies of the promoter sequences themselves (for example, by examining the effects on GFP expression of changes of the DNA sequence at different regions of the promoters) to better determine which specific DNA sequences within the promoter contribute to its capacity to regulate expression of a linked gene. *See* DEOXYRIBONUCLEIC ACID (DNA).

Cells expressing GFP-tagged proteins can be separated from nonexpressing cells or cells expressing different levels of GFP using devices called cell sorters. Cells are sent one at a time past a laser beam set at a frequency that excites fluorescent molecules. The cells can then be scored for fluorescence and deposited into distinct pools. One application is identification of transfected cells after foreign DNA coding for a gene of interest and also for GFP is introduced. The GFP identifies the transfected cells, and



Labeling of multiple organelles in the same cell with spectrally distinguishable green fluorescent protein variants. (*a*) Lamin-binding receptor (a protein found in the nuclear envelope) tagged with CFP. (*b*) Microtubules containing tubulin tagged with YFP. (*c*) Histone 2b (a protein that binds to DNA in the nucleus) tagged with RFP. (*Jan Ellenberg, European Molecular Biology Laboratory*)

the cell sorter can then provide a pure population of them.

Spectral variants. New improved variants of the protein have been devised that are brighter than the original and show a preference for blue light (which is nontoxic to cells) rather than ultraviolet (which can rapidly kill cells being observed). Also, variants have been engineered which excite and emit at different wavelengths, allowing tagging and following two or more different proteins in the same cell. Two fluorescent proteins, a cyan-emitting variety (CFP) and a yellow-green-emitting variety (YFP), have proved particularly useful in double-label experiments since they can be visualized separately in a fluorescence microscope. Recently, a related red fluorescent protein (RFP) has been identified from other coelenterates, so it is now possible to tag three different proteins in the same cell or in a cell population (see **illustration**). John Presley

Bibliography. M. Chalfie and S. R. Kain (eds.), *Green Fluorescent Protein: Properties, Applications and Protocols*, 2d ed., Wiley, New York, 2005; J. Ellenberg, J. Lippincott-Schwartz, and J. F. Presley, Dual-color imaging with GFP variants, *Trends Cell Biol.*, 9:52–56, 1999; B. W. Hicks, *Green Fluorescent Protein: Applications & Protocols (Methods in Molecular Biology)*, Humana Press, Totowa, NJ, 2002; J. R. Lakowicz and C. D. Geddes, *Green Fluorescent Protein (Topics in Fluorescence Spectroscopy)*, Springer, New York, 2006; J. F. Presley et al., ER-to-Golgi transport visualized in living cells, *Nature*, 389:81–85, 1997; K. F. Sullivan and S. A. Kay, *Green Fluorescent Proteins*, Academic Press, New York, 1999.

Greenhouse effect

The ability of a planetary atmosphere to inhibit heat loss from the planet's surface, thereby enhancing the surface warming that is produced by the absorption of solar radiation. For the greenhouse effect to work efficiently, the planet's atmosphere must be relatively transparent to sunlight at visible wavelengths so that significant amounts of solar radiation can penetrate to the ground. Also, the atmosphere must be opaque at thermal wavelengths to prevent thermal radiation emitted by the ground from escaping directly to space. The principle is similar to a thermal blanket, which also limits heat loss by conduction and convection. In recent decades the term has also become associated with the issues of global warming and climate change induced by human activity. *See* ATMOSPHERE; SOLAR RADIATION.

Of the terrestrial planets, Venus has by far the strongest greenhouse effect. Though only about 1% of the incident solar radiation penetrates to the ground, the thermal opacity of Venus's atmosphere, which is 100 times more massive than the Earth's and composed mostly of carbon dioxide, is exceedingly large. As a result, the trapped solar radiation on Venus generates a surface temperature of nearly 460°C (860°F), which is hot enough to melt lead and vaporize mercury. This is 500°C (900°F) hotter than the surface would be if it were simply in thermal equilibrium with the global mean solar energy absorbed by Venus (−40°C, −40°F), without benefit of the greenhouse effect. For the Earth, the strength of the greenhouse warming effect is a more modest 33°C (60°F), due primarily to the thermal opacity of atmospheric water vapor, clouds, and carbon dioxide. The trapped sunlight makes the global mean surface temperature of the Earth a relatively comfortable 15°C (59°F) instead of an otherwise frigid −18°C (−1°F). On Mars, the very tenuous carbon dioxide atmosphere can muster only a few degrees (≈3°C, 5°F) of greenhouse warming. Meanwhile, on Mercury, which has no tangible atmosphere, there is no greenhouse effect at all. Mercury's surface temperature is determined solely by local thermal equilibrium with the absorbed sunlight. *See* MARS; MERCURY (PLANET); VENUS.

Mechanism. Basic understanding of the greenhouse effect dates back to the 1820s, when the French mathematician and physicist Joseph Fourier performed experiments on atmospheric heat flow and pondered the question of how the Earth stays warm enough for plant and animal life to thrive; and to the 1860s, when the Irish physicist John Tyndall demonstrated by means of quantitative spectroscopy that common atmospheric trace gases, such as water vapor, ozone, and carbon dioxide, are strong absorbers and emitters of thermal radiant energy but are transparent to visible sunlight. It was clear to Tyndall that water vapor was the strongest absorber of thermal radiation and, therefore, the most influential atmospheric gas controlling the Earth's surface temperature. The principal components of air, nitrogen and oxygen, were found to be radiatively inactive, serving instead as the atmospheric framework where water vapor and carbon dioxide can exert their influence. Based on his understanding of the radiative properties of absorptive gases in the atmosphere, Tyndall speculated in 1861 that a reduction in atmospheric carbon dioxide could induce an ice age climate.

Relying on the work of Tyndall and on careful measurements of heat transmission through the atmosphere compiled by the American astronomer Samuel Langley, the Swedish chemist and physicist Svante Arrhenius was the first to develop a quantitative mathematical model of the Earth's greenhouse effect. In 1896 he published his heat-balance calculations of the Earth's sensitivity to carbon dioxide change. Given that Arrhenius's basic interest was to explain the likely causes of ice age climate, his greenhouse model was successful. He showed that reducing atmospheric carbon dioxide by a third would cool the global surface temperatures by −3°C (−5.5°F), and that doubling carbon dioxide would cause the tropical latitudes to warm by 5°C (9°F), with somewhat larger warming in polar regions. These results are in remarkably close agreement with current expectations for global climate change in response to carbon dioxide forcing. *See* SPECTROSCOPY.

Large seasonal and epochal (ice age) variability is evident in the terrestrial climate record. This implies that the greenhouse effect on Earth operates rather differently from the static, or at least slowly evolving, carbon dioxide greenhouses on Venus and Mars. Partly, this is the result of changes in the atmospheric concentration of carbon dioxide over geological time scales. More importantly, it is due to the dominant strength of water vapor as the largest contributor to the terrestrial greenhouse effect, and the fact that the amount of atmospheric water vapor is strongly dependent on temperature. Because of this, and also because condensing water vapor produces clouds which typically cool the Earth, there is strong coupling and interaction between the amount of atmospheric water vapor and temperature.

This aspect of water vapor behavior was noted by the American geologist Thomas Chamberlin who, in 1905, described the greenhouse contribution by water vapor as a positive feedback mechanism. Surface heating due to another agent, such as carbon dioxide or solar radiation, raises the surface temperature and evaporates more water vapor which, in turn, produces additional heating and further evaporation. When the heat source is taken away, excess water vapor precipitates from the atmosphere, reducing its contribution to the greenhouse effect to produce further cooling. This feedback interaction converges and, in the process, achieves a significantly larger temperature change than would be the case if the amount of atmospheric water vapor had remained constant. The net result is that carbon dioxide becomes the controlling factor of long-term change in the terrestrial greenhouse effect, but the resulting change in temperature is magnified by the positive feedback action of water vapor.

It is believed that in the early stages of the solar system, when the Sun had only about 75% of its present luminosity, Venus may also have had an ocean and an atmosphere much like that of the Earth. However, in its early days Venus was apparently unable to sequester its store of carbon in the form of limestone, as did the Earth. As a result, carbon dioxide emitted from volcanic eruptions continued to accumulate in the Venusian atmosphere, increasing its greenhouse effect and evaporating more and more water vapor into the atmosphere to further magnify the growing strength of its greenhouse. Eventually all ocean water was vaporized, and the water vapor was carried high into the atmosphere where most of it was photolyzed and lost to space. Perhaps the climate disaster on Venus can be attributed to the lack of appropriate life forms that might have limited the buildup of carbon dioxide. In any case, Venus stands as an example of a runaway greenhouse effect that has rendered the planet inhospitable to life.

Besides water vapor, many other feedback mechanisms operate in the Earth's climate system and impact the sensitivity of the climate response to an applied radiative forcing. Determining the relative strengths of feedback interactions between clouds, aerosols, snow, ice, and vegetation, including the effects of energy exchange between the atmosphere and ocean, is an actively pursued research topic in current climate modeling. Current best estimates from climate modeling results show that without feedback effects, a doubling of atmospheric carbon dioxide will raise the global surface temperature by 1.2–1.3°C (2.2–2.4°F), but that the net feedback magnification due to water vapor, snow/ice melting, and cloud changes magnifies the no-feedback result by approximately a factor of 3. Interestingly, carbon dioxide accounts for approximately 7°C (13°F) of the total 33°C (60°F) terrestrial greenhouse effect, and the other nonvolatile greenhouse gases such as ozone, methane, nitrous oxide, and anthropogenic chlorofluorocarbons add another 3–4°C (5–7°F) of greenhouse strength. In this context, the nonvolatile greenhouse gases can be thought of as providing the core forcing for the Earth's greenhouse effect. The larger greenhouse contribution by water vapor and clouds may then be considered to represent the feedback response that magnifies the core forcing by the atmospheric feedback factor of 3. This distinction helps to identify more clearly the causes and effects in an otherwise highly interactive climate system. *See* CLIMATE MODIFICATION.

The modern approach to studying Earth's climate and the human impact on climate change began in the 1930s with the work of Guy Callendar, a British engineer who systematically documented the time trend of anthropogenic fossil fuel use and linked it to the corresponding increase in atmospheric carbon dioxide. Having had available to him the precise spectroscopic measurements of the absorption characteristics of water vapor, carbon dioxide, and other heat-absorbing gases, Callendar could link his estimates of carbon dioxide increase with the 0.4°C (0.7°F) temperature increase recorded in the 50 years prior to 1950. In 1958 Charles Keeling, a research chemist at Scripps Institute in California, began making high-precision measurements of carbon dioxide accumulation in the atmosphere. This ushered in a new era of precise measurement and documentation of the atmospheric increases of carbon dioxide and other greenhouse gases such as methane and nitrous oxide. The new measurement techniques were subsequently applied to measuring the precise gaseous composition of air bubbles trapped in glacial ice, thereby extending knowledge of atmospheric composition over time scales going back to the ice ages. The measurements show that atmospheric concentrations of carbon dioxide increased from a preindustrial 280 parts per million in 1850 to the present value of 370 ppm. In geological context, during the last ice age when the global temperature was 5°C (9°F) colder, atmospheric carbon dioxide was at levels near 150 ppm.

Given accurate knowledge of the atmospheric concentration of greenhouse gases, their impact on the strength of the atmospheric greenhouse effect can be accurately calculated using the extensive spectroscopic databases that are available now. Inasmuch as the term "greenhouse effect" refers to determining the radiative efficiency of trapping solar heat at the ground surface by the absorption and emission of thermal radiation, a moderate computational effort can provide the answer. Insofar as human activity induces global warming and climate change, it is necessary to incorporate the uncertainties that are associated with atmospheric feedback processes and with atmosphere/ocean dynamical interactions that contribute to greenhouse efficiency. At present, it is necessary to rely on general circulation climate models to determine the feedback contribution (due mostly to clouds and water vapor) to the total atmospheric greenhouse effect, since remote sensing measurements capable of measuring feedback-related water vapor and cloud changes are not currently available. *See* ATMOSPHERIC GENERAL CIRCULATION; CLIMATE MODELING; REMOTE SENSING.

Human impact. Both Arrhenius and Callendar considered the projected global temperature increases due to anthropogenic carbon dioxide buildup as likely to be beneficial in warming the cold regions of the Earth. But human-induced contributions to climate change are not necessarily benign. For example, discovery of an ozone hole in the Antarctic in the early 1980s pushed the world community to the 1987 Montreal agreement to phase out production of chlorofluorocarbons to reverse ozone destruction in the stratosphere caused by these compounds. As a by-product, this also helped to reduce the rate of increase of the anthropogenic greenhouse forcing. In late 1997 the world community met in Kyoto to consider what could be done to help stabilize and curtail the steady increase in fossil fuel use and the resulting increase in the strength of greenhouse forcing. There are good reasons to be concerned: (1) The projected global warming is much larger than anything previously experienced in human history. (2) The climate response to a globally uniform forcing is not necessarily uniform and may include large regional fluctuations. (3) Extreme events, both droughts and floods, are more likely with the stronger hydrological cycle of a warmer Earth. (4) The inevitable rise in sea level will put low-lying coastal areas at increased risk. (5) Unanticipated changes in global climate could occur if the climate system were to exceed some critical threshold that is beyond the modeling capability of current climate models. *See* STRATOSPHERE.

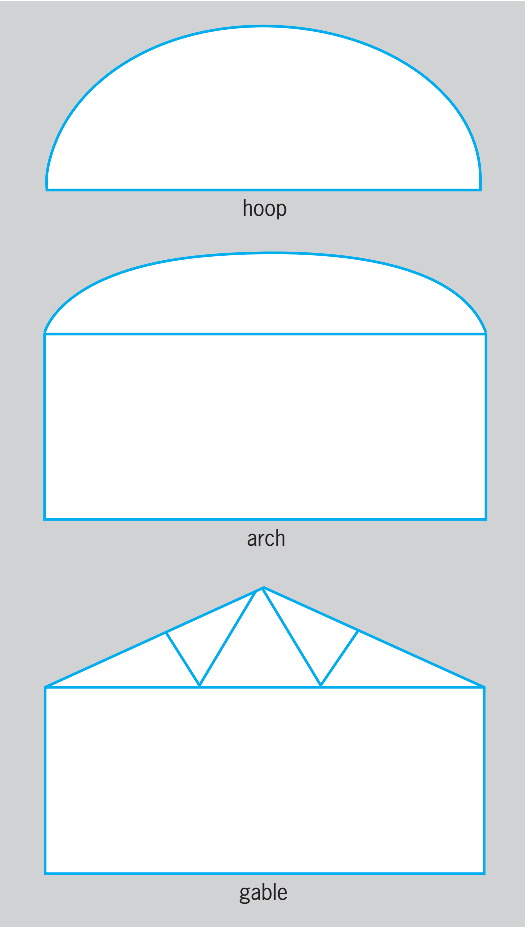
Andrew A. Lacis

Bibliography. G. E. Christianson, *Greenhouse*, 1999; J. R. Fleming, *Historical Perspectives on Climate Change*, 1998; J. T. Houghton et al., *IPCC WGI 1995: Climate Change 1995*.

Greenhouse technology

Low tunnels, high tunnels, and greenhouses are structures used to grow plants under protected conditions. The progression of terms shows the level (low to high) of technical sophistication in the plant-growing systems. Low tunnels, also called row covers, primarily advance the growing season for outdoor crops (for example, tomatoes, melons, strawberries, and sweet corn). Low tunnels are created using long, narrow strips of transparent plastic material (often polyethylene) buried in the ground along their outer edges to cover one or several adjacent rows of plants grown directly in the soil. French cloches and traditional hot caps are variations of this technology. High tunnels are large versions of low tunnels, raised sufficiently above the ground that people can walk within them.

Types of greenhouses. Greenhouses (or glasshouses) are relatively permanent structures (glass or plastic with aluminum or steel frames) equipped with several means of environmental modification. Free-standing greenhouses are the most basic structural type; they are typically 7–10 m (23–33 ft) wide and 5–10 times as long as wide. Cross-sectional shapes can be classified as arch, hoop, or gable (see **illus.**). The hoop shape is common for high tunnels. Multispan greenhouses are typically connected by a series of roof gutters to create a single air space. Large multispan greenhouses can cover several hectares under one roof, and they are the design of choice for larger commercial greenhouse operators. Eave height continues to increase, and modern structures 5 m (16 ft) tall at the eaves are not uncommon. The greater height improves environmental control, particularly ventilation.



Common commercial greenhouse shapes.

Construction. Aluminum, steel, wood, and concrete have been used to build greenhouse frames, but steel and aluminum are most common. The most prevalent shape has a straight sidewall and a gable roof, although arched roof designs are frequently seen (see illus.). The greenhouse can be a single freestanding unit, or multiple units can be joined at the eaves to cover several hectares. Floors are frequently made of concrete, although gravel floors with concrete walkways may be used to reduce cost.

Light transmittance is important when selecting a covering material. Glass provides the most light to the plants and retains its light transmittance; however, various rigid and film plastic glazing materials are used because of initial lower costs. Plastic film glazing provides a greater ratio of diffused to direct light, which may be of some benefit for plant growth. *See* GLASS.

Recommended greenhouse orientation depends on latitude. If the latitude is greater than 35°, an east-west orientation of the ridge line is preferred to admit more natural light in winter when the Sun does not rise far above the horizon, even at midday. If the latitude is less than 35°, the preferred orientation is north-south to reduce midday solar heating and prevent shade lines caused by long structural elements oriented east-west from casting shadows that remain in one place for a long time, which retards plant growth in the shadow areas.

Environment control. Environment control typically encompasses air temperature, supplemental light, air movement (circulation and mixing), and carbon dioxide concentration. Some degree of relative humidity control may also be included. Natural ventilation may be used to provide less stringent control, but mechanical ventilation is used when more precise temperature control is needed and evaporative cooling is required during hot weather.

Integrated control by computer provides the flexibility of zoned control of each environmental parameter without conflicting control signals (for example, ventilating while supplementing carbon dioxide). Most modern greenhouse facilities have installed environment control computers. Additionally, computerized control usually provides data summaries useful for management decisions. Continuing developments in sensor technology will undoubtedly lead to increasingly sophisticated control of aerial and root environments and fault detection capabilities.

Heating. The function of a greenhouse is to grow plants; structural insulation opportunities are minimal and heat requirements are high in comparison to most other types of buildings. Efforts to conserve heat, such as insulating the north wall and part of the north roof, have often shown negative benefits by reducing natural light and degrading plant growth and quality. Movable, horizontal, indoor curtain systems (which also double as movable shade systems) can save approximately one-quarter of the yearly heat in a cold climate.

Greenhouses can be heated by oil or natural gas. Electric heat is uneconomical. Large greenhouses may be heated using heavier industrial heating oils; propane and wood have been used in specialized situations. Heat delivery is either hydronic (hot water) or by steam. In some cases, particularly in smaller greenhouses, hot air systems can be used if care is taken not to direct the hot air directly onto the plants. No matter what the heat source, care is required to prevent incompletely burned flue gases from entering the greenhouse and causing plant damage (phytotoxicity). *See* HEAT.

Cooling. Solar loads in greenhouses are so great that mechanical cooling, as by air conditioning, is prohibitively expensive. Options for greenhouse cooling are thus limited. If natural light is generally high, simple shading (preferably a movable shade, but it can also be semipermanent) can cool a greenhouse several degrees. The most typical cooling mechanism is ventilation, either natural or mechanical. Natural ventilation is not controlled, but mechanical ventilation is generally designed to keep the temperature rise of the ventilation air to no more than 4–5°C (7–9°F). This may require an air exchange rate as high as one total greenhouse volume each minute. The next step of cooling is to use evaporative means. Evaporative cooling can be used in all but the most hot and humid climates to reduce midday temperature extremes. Cooling can be obtained by spraying a fine mist into the ventilation air or by pulling outside air through matrices (such as aspen fiber pads or manufactured units having large surface areas but limited airflow restriction) or structures that are wetted to cool the air flowing past them. A less common evaporative cooling means is to spray the outside of the greenhouse with water, or perhaps to wet a shade cloth covering the greenhouse.

Lighting. Greenhouse lighting may be used for photoperiodic reasons, or for enhancing growth. Photoperiodic lighting is a very low intensity light during the night to break the darkness period and induce plant responses representative of summer (short nights and long days). Plants use light within the wavelength band of 400–700 nanometers for growth, and grow proportionally to the number of photons within that energy band. This is in contrast to the human eye which has highly variable sensitivity over the visible light band. The photon flux is measured in moles of photons (one mole of photons is equal to Avogadro's constant, approximately 6.02×10^{23}). The average daily solar photon flux over the Earth is approximately 26 mol/m^2, but greenhouse lighting generally provides less due to the cost of electricity. Greenhouse supplemental lighting is usually provided by high-pressure sodium (HPS) lights because of their relatively high energy efficiency. *See* LIGHT; PHOTOPERIODISM; PHOTOSYNTHESIS.

Carbon dioxide. Ambient carbon dioxide concentrations range from 360 parts per million in rural areas to nearly 400 ppm in urbanized areas. Plants, however, are able to use carbon dioxide efficiently at concentrations up to 2000 ppm. Optimum concentrations are often in the range 800–1000 ppm, which can lead to 25% greater growth provided that other inputs are not limited. Carbon dioxide can be added through carefully controlled flue gases from the greenhouse heating system, or from tanks of liquid carbon dioxide.

Solar applications. Greenhouses are, by their nature, solar collectors, but commercial greenhouses generally do not include specialized or enhanced solar features. Home (and hobby) greenhouses, in contrast, frequently include additional thermal mass and other features to enhance their passive solar collection potential.

Mechanization and automation. Many greenhouse operations that were formerly done by hand are now mechanized or automated. Root medium is mixed, fertilized, and placed directly into flats and pots by machine. Seeding can be totally by machine. Transplanting seedlings into larger containers can be by machine. Plant watering and fertilizing (termed "fertigation" when combined) can now be automated using overhead watering booms, drip irrigation emitters, ebb-and-flow benches, or flooded (and then drained) floors. Automatic material movement at harvest, coordinated by a computer, is no longer unusual. Although the human worker is unlikely to be totally replaced, mechanization and automation have reduced labor requirements to five workers per hectare, or fewer, in the most modern greenhouse operations.

Plant nutrition management. Sixteen elements are believed to be essential for plant growth, including carbon, oxygen, and hydrogen. Minerals required for plant nutrition have traditionally been grouped into two categories: macronutrients and micronutrients. The difference is based on the quantity of the nutrient required for good plant growth.

Plant fertilizers are composed of a mix of salts that are electrically conducting when dissolved into water. This characteristic leads to the use of electrical conductivity as a measure of fertilizer concentration. Computer programs have been developed that are suitable for balancing a nutrient mix to achieve close approximations to the desired molar ratios of elements. *See* PLANT MINERAL NUTRITION.

Hydroponics. Hydroponics is defined as growing plants without using soil. However, a root medium such as sand, gravel, or rockwool may be used. Two common hydroponics systems that use no root medium are the nutrient film technique (NFT) and deep flow troughs (DFT). *See* HYDROPONICS.

The nutrient film technique is a closed system for growing plants so their roots remain in a shallow stream of recirculating nutrient solution. There is no root medium other than the nutrient solution. However, seedlings are typically started in a small cube of root medium, which is then transferred to the nutrient film technique system when roots begin to emerge from the cube. All required nutritional elements are dissolved in the nutrient solution. The system typically uses shallow troughs or channels to support the plant roots and contain the flowing nutrient solution. The defining characteristic of true nutrient film technique is that the nutrient stream is very shallow. Plant roots rest partly within the stream and partly within the air above. However, nutrient solution moves by capillary action around the roots that extend into the air to bathe them in water and nutrients, while permitting free access by the roots to oxygen. The nutrient stream may flow continuously or may be intermittent with the roots alternately submerged and exposed to air (on a rapid cycle of numerous times each hour so even the smallest roots do not desiccate).

The deep flow troughs system for plant production in closed systems is also identified as ponds, deep-flow hydroponics, and raft systems. The plants float on a raft (typically a sheet of foam plastic such as polystyrene) in a shallow tank less than 0.3 m (1 ft) deep. The tank, or pond, is filled with nutrient solution, and the plant roots hang down into the solution. The nutrient solution is monitored, oxygenated, replenished, and recirculated as required. The mass of water in the troughs provides several distinct advantages. First, it provides significantly greater buffering than do other hydroponic techniques such as the nutrient film technique. Changes of nutrient levels, pH, temperature, and dissolved oxygen (DO) concentration (as examples) are significantly damped. Additionally, the ponds can be used for material movement with a minimum of mechanization or physical effort.

Economics. Growing plants in greenhouses is another form of farming and is subject to many of the risks inherent in agriculture. Modern greenhouse technologies have mirrored developments in most of agriculture in that increased labor efficiency, larger sizes of greenhouse operations, and mass production of a few crops, or even a single crop, have become the rule to be profitable. Large buyers of greenhouse products (such as discount chains) can deal best with large producers, which has led to considerable industry consolidation. There remain, of course, many niche opportunities for small grower operations. The current dynamic in the greenhouse industry in the United States is characterized by the entry of many growers in small, specialized operations, and consolidations and mergers of large operations. *See* FLORICULTURE; PLANT GROWTH; PLANT-WATER RELATIONS.

Louis D. Albright

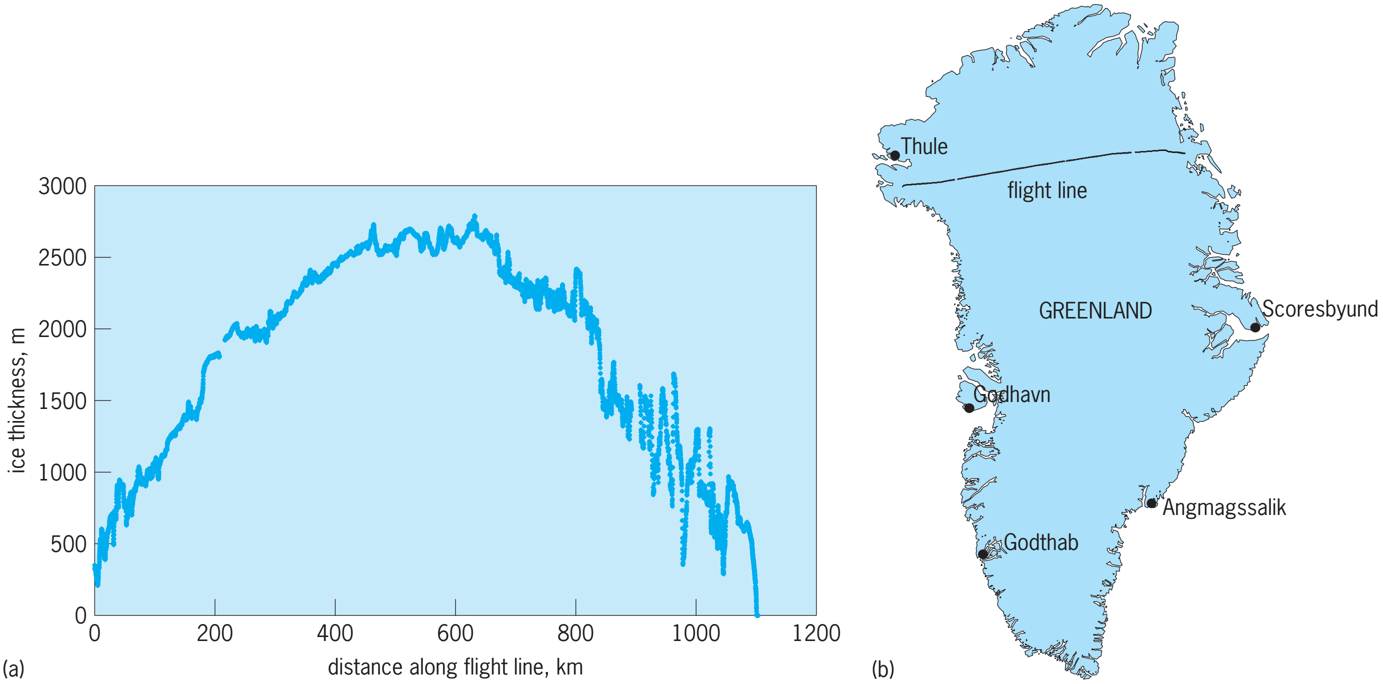
Bibliography. R. A. Aldrich and J. W. Bartok, Jr., *Greenhouse Engineering*, Natural Resource, Agriculture and Engineering Service (NRAES), Cornell University, Ithaca, NY, 1994; E. Goto et al. (eds.), *Plant Production in Closed Ecosystems*, Kluwer Academic, Dordrecht, The Netherlands, 1997; J. J. Hanan, *Greenhouses: Advanced Technology for Protected Horticulture*, CRC Press, Boca Raton, FL, 1998; R. W. Langhans, *Greenhouse Management*, Halcyon Press, Ithaca, NY, 1990; H. M. Resh, *Hydroponic Food Production*, Woodbridge Press, Santa Barbara, CA, 1995.

Greenland

The world's largest island. It is a major land and ice mass in the North Atlantic Ocean between North America and Europe, bounded by the Arctic Ocean to the north. The total area of Greenland is 2,166,086 km^2 (836,330 mi^2), or a little more than three times the size of Texas. Its most distinctive feature is an ice sheet that covers about 80% of the island like a dome, with elevations toward the center of the island exceeding 3000 m (9800 ft). The Greenland Ice Sheet is the largest ice mass in the Northern Hemisphere and is the second largest ice sheet in the world. (The Antarctic Ice Sheet is about ten times larger.) The Greenland Ice Sheet affects weather systems in the North Atlantic Ocean and northern Europe, and has the potential to affect the environment by contributing to global sea-level rise. Greenland has a total ice volume of 2,600,000 km^3 (624,000 mi^3) which, if melted, would cause global sea level to rise about 6 m (20 ft). Significant melting would also introduce a large volume of freshwater into the North Atlantic Ocean, which could disrupt the Gulf Stream current. Recent studies, using ice thickness measurements from airborne ice-penetrating radar and ice surface movement from satellite measurements, show accelerated ice discharge into the ocean. An example of an ice thickness profile along a flight line from the west to the east coast is shown in the **illustration**. *See* ARCTIC AND SUBARCTIC ISLANDS; ARCTIC OCEAN; GLACIOLOGY; GULF STREAM.

Greenland is also a country that has a population of about 56,000, most of whom live in small villages along the coast, with about 25% of the population living in the capital Godthab (Nuuk). Emblematic of its polar location, the summer solstice is a national holiday. Greenland is a self-governing parliamentary democracy but is part of the Kingdom of Denmark and uses the Danish krone as its currency. Denmark handles its foreign affairs and defense responsibilities, and provides financial support that is equivalent to about half the Greenland government revenues.

Greenland's climate is classified (Köppen system) as polar tundra along the coastal margins, with exception of the northern regions bordering the Arctic



Aerial survey using ice-penetrating radar, showing (*a*) the ice thickness profile along (*b*) a flight line. (*University of Kansas*)

Ocean. For this region and the interior region of Greenland beginning at the ice-sheet edge, the climate is classified as polar ice cap. Along the coastal margins, the average temperature and precipitation generally decrease from south to north. There are profound climatic differences between the east and the west coasts of Greenland because of differences in the ocean current direction. Along the east coast, the East Greenland Current transports cold arctic water southward, reducing the air temperature. Along the west coast, a northward-flowing current warmed by a branch of the Gulf Stream helps to moderate the air temperature. *See* KÖPPEN CLIMATE CLASSIFICATION SYSTEM; POLAR METEOROLOGY; TUNDRA.

The eastern part of Greenland is dominated by the East Greenland Mountains, much of which is buried under the Greenland Ice Sheet. Exposed mountain peaks near the ice-sheet margin range from 1800 to 3700 m (5900 to 12,000 ft). The terrain along the west coast ranges from rolling hills to mountainous, with peaks of 700 to 1000 m (2300 to 3300 ft). The East Greenland Mountains strongly influence the ice flow of the Greenland Ice Sheet, with the main ice flow toward the west. David A. Braaten

Bibliography. Central Intelligence Agency, *The World Factbook*, 2006; S. Gogineni et al., Coherent radar ice thickness measurements over the Greenland Ice Sheet, *J. Geophys. Res.*, 106(D24):33,761–33,772, 2001; E. Rignot and P. Kanaratnam, Changes in the velocity structure of the Greenland Ice Sheet, *Science*, 311:5763, 986–990, 2006; A. Weidick, *Greenland: Satellite Image Atlas of Glaciers of the World*, U.S. Geol. Surv. Prof. Pap. 1386-C, 1995.

Greenockite

A mineral having composition CdS (cadmium sulfide) and crystallizing in the hexagonal system (dihexagonal pyramidal class). Pyramidal crystals are rare, and greenockite usually occurs as earthy coatings with resinous luster and yellow-to-orange color. There is good prismatic cleavage; the hardness is 3 (Mohs scale) and specific gravity is 4.9. Greenockite and wurtzite, ZnS, are isostructural, and a complete solid-solution series exists between the two minerals. Although greenockite is the most common cadmium mineral, no deposits of it are sufficiently large to warrant mining it solely as a source of cadmium. It is commonly associated with sphalerite, and thus the supply of cadmium comes as a by-product from the treating of zinc ores. *See* CADMIUM; SPHALERITE.

Cornelius S. Hurlbut, Jr.

Green's function

A solution of a partial differential equation for the case of a point source of unit strength within the region under examination. The Green's function is an important mathematical tool that has application in many areas of theoretical physics including mechanics, electromagnetism, acoustics, solid-state physics, thermal physics, and the theory of elementary particles. The underlying physics in each of these areas is generally described by some linear partial differential equation which relates the physical variable of interest (electrostatic potential or pressure amplitude in a sound wave, for example) to a source function. For present purposes the source may be regarded as an independent entity, although in some applications (for example, particle physics) this view masks an inherent nonlinearity. The source may be physically located within the region of interest, it may be simulated by certain boundary conditions on the surface of that region, or it may consist of both possibilities. A Green's function is a solution to the relevant partial differential equation for the particular case of a point source of unit strength in the interior of the region and some designated boundary condition on the surface of the region. Solutions to the partial differential equation for a general source function and appropriate boundary condition can then be written in terms of certain volume and surface integrals of the Green's function.

Helmholtz equation. As a primary example, consider the inhomogeneous Helmholtz equation with parameter k, Eq. (1), where ρ_k is the source, ∇^2 is

$$(\nabla^2 + k^2)\,\phi_k(\vec{r}) = -4\pi\rho_k(\vec{r}) \qquad (1)$$

the laplacian operator, and ϕ_k is the function of the independent variable $\vec{r}$ whose solution is sought. This is an important partial differential equation whose solutions can be related through Fourier transformation to solutions of the inhomogeneous wave equation (that is, the wave equation with a source). *See* CALCULUS OF VECTORS; FOURIER SERIES AND TRANSFORMS; LAPLACIAN; WAVE EQUATION.

The special case of $k = 0$ gives Poisson's equation which describes both electrostatics and newtonian gravitation. The Green's function for the Helmholtz equation obeys Eq. (1) with the source replaced by a Dirac delta function $\delta^3(\vec{r} - \vec{r}')$ as appropriate for a point source. The Dirac delta function is zero except at $\vec{r} = \vec{r}'$ where its value is infinite. This infinity is such that the volume integral of $\delta^3(\vec{r} - \vec{r}')$ over $\vec{r}$ (for fixed $\vec{r}'$) is one. While it is convenient to think of $\vec{r}'$ as the location of the source and $\vec{r}$ as the location at which the strength of the Green's function (due to the source) is determined, the inherent symmetry property of Eq. (2) implies that source and field

$$G_k(\vec{r}, \vec{r}') = G_k(\vec{r}', \vec{r}) \qquad (2)$$

points can be interchanged. A unit source at $\vec{r}'$ produces the strength $G_k(\vec{r},\vec{r}')$ at $\vec{r}$, and a unit source at $\vec{n}$ produces the strength $G_k(\vec{r},\vec{r}')$ at $\vec{r}'$. Green's theorem enables the general solution to be determined from the Green's function, using Eq. (3), where $\vec{n}'$ is the

$$\begin{aligned}\phi_k(\vec{r}) = &\int_V dV'\,G_k(\vec{r}, \vec{r}')\rho_k(\vec{r}) \\ &+ \oint_s dS'\;G(\vec{r}, \vec{r}')\vec{n}' \cdot \vec{\nabla}'\phi_k(\vec{r}') \\ &- \oint_s dS'\vec{n}' \cdot \vec{\nabla}'G(\vec{r}, \vec{r}')\phi_k(\vec{r}') \qquad (3)\end{aligned}$$

outward normal to the surface *S* which bounds the region *V*, and $\vec{\nabla}'$ is the gradient with respect to $\vec{r}'$. If $\phi_k(\vec{r})$ is specified on the boundary (the Dirichlet condition), it is required that $G_k(\vec{r},\vec{r}') = 0$ for $\vec{r}'$ on the boundary. If $\vec{n} \cdot \vec{\pi}_k(\vec{r})$ is specified on the boundary (the Neumann condition), then it is required that $\vec{n}' \cdot \vec{\nabla}' G(\vec{r},\vec{r}') = 0$ for $\vec{r}'$ on the boundary. In principle, a more complicated boundary condition can be imposed. The physical content of Eq. (3) is that the solution $\phi_k(\vec{r})$ arises from the sources within *V* (the volume integral) and the effective sources on the surface *S* (the surface integrals). *See* GREEN'S THEOREM; POTENTIALS.

Boundary effects. If the boundary recedes to infinity, $G_k(\vec{r},\vec{r}')$ approaches 0 as $|\vec{r} - \vec{r}'|$ approaches infinity, and the Green's function for the Helmholtz equation is given in closed form by Eq. (4) for gen-

$$G_k(\vec{r}, \vec{r}') = \frac{e^{\pm ik|\vec{r}-\vec{r}'|}}{|\vec{r} - \vec{r}'|} \qquad (4)$$

erally complex values of *k* the top sign is taken for the imaginary part of *k* (Im *k*) greater than 0 and the bottom is taken for Im *k* less than 0. The need for complex values of *k* comes about when the Green's function for thewave equation, Eq. (5), is found by

$$\left(\nabla^2 - \frac{1}{S^2}\frac{\partial^2}{\partial t^2}\right)\phi(\vec{r}, t) = -4\pi\rho(\vec{r}, t) \qquad (5)$$

Fourier transformation of $G_k(\vec{r},\vec{r}')$. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

For finite domains the determination of $G_k(\vec{r},\vec{r}')$ is difficult and usually analytically impossible, but for a number of boundary shapes the Green's function can be expressed as the infinite series in Eq. (6), where $u_n(\vec{r})$ is an eigenfunction of the ho-

$$G_k(\vec{r}, \vec{r}') = \sum_{n=1}^{\infty} \frac{u_n(\vec{r})u_n^*(\vec{r}')}{k^2 - k_n^2} \qquad (6)$$

mogeneous Hemholtz equation (for the specified boundary condition) and k_n^2 is its associated eigenvalue. In principle, it is possible to explicitly carry out some of the summation implicit in Eq. (6), and in practice this procedure is often very useful. For parallelepipeds, right circular cylinders, and spheres, the eigenfunctions and eigenvalues can be determined in terms of functions such as sines and cosines, Bessel functions and Legendre functions. In some cases the Green's function can be found in closed form. For the half-space $z > 0$ the Green's function is given by Eq. (7), where R_- and R_+ are given by Eqs. (8).

$$G_k(\vec{r}, \vec{r}') = \frac{e^{ikR_-}}{R_-} \mp \frac{e^{ikR_+}}{R_+} \qquad (\text{Im } k > 0) \qquad (7)$$

$$\begin{aligned} R_- &= [(x-x')^2 + (y-y')^2 + (z-z')^2]^{1/2} \\ rR_+ &= [(x-x')^2 + (y-y')^2 + (z+z')^2]^{1/2} \end{aligned} \qquad (8)$$

In Eq. (7) the minus sign between the two terms is for the Dirichlet condition on the surface $z = 0$ and the plus sign is for the Neumann condition. The case Im $k < 0$ can be treated as in Eq. (4). For Poisson's equation ($k = 0$), closed-form expressions can also be found for several other geometries by the technique referred to as the method of images. An image source is one placed outside the region *V* in such a way that the boundary condition is satisfied while leaving the source function unchanged in *V*. Equation (7) is an example of the method of images for the Helmholtz equation with the image source placed at the point (x', y', $-z'$). The second term in Eq. (7) is the contribution from the image source and the first term is the contribution from the source. *See* BESSEL FUNCTIONS; EIGENFUNCTION; LEGENDRE FUNCTIONS; SPECIAL FUNCTIONS; TRIGONOMETRY.

Spectral property. An important property of the Helmholtz Green's function is that it provides the eigenvalue spectrum of the Helmholtz equation. It is evident from the eigenfunction expansion [Eq. (6)] that $G_k(\vec{r},\vec{r}')$ has a pole in the complex *k* plane at $k^2 = k_n^2$ with residue related to the eigenfunction u_k. This property is particularly useful if it is possible to find the Green's function by some means other than the eigenfunction expansion. The property is also of great importance in applications to complicated many-particle systems.

Quantum-mechanical applications. The time-independent Schröodinger equation for a single particle in an external potential can be cast in the form of Eq. (1), but the source function depends linearly on the wave function $\phi_r(\vec{r})$. In this case Eq. (1) becomes a linear integral equation for the wave function whose kernel is the Helmholtz Green's function. Alternatively, a Green's function can be introduced directly for the Schröodinger equation which has an eigenfunction expansion similar to Eq. (6) in which $u_n(\vec{r})$ is an eigenfunction of the Schröodinger equation and the k^2 is its energy eigenvalue (to within a scale factor). *See* EIGENVALUE (QUANTUM MECHANICS); INTEGRAL EQUATION; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

It is also possible to describe many-particle quantum-mechanical systems in terms of a Green's function formalism. Such systems occur in condensed matter physics (notably, solid-state physics), nuclear physics, and elementary particle physics. Unfortunately, the equations obeyed by these Green's functions are generally nonlinear and approximation methods are required. But the spectral property of Eq. (6) remains an important feature, and it is sometimes possible to detect effective single-particle–like behavior in complicated many-particle motion. These quasi-particles appear as (approximate) poles in the Green's function for the many-particle system.

If the interactions between particles are relatively unimportant, it is possible to develop an approximation scheme in which the Green's function for an individual free particle plays an essential role. These Green's functions are called propagators and form the basis for many calculations in modern relativistic quantum field theory. *See* DIFFERENTIAL EQUATION; PROPAGATOR (FIELD THEORY); QUANTUM CHROMODYNAMICS; QUANTUM ELECTRODYNAMICS; QUANTUM FIELD THEORY. Peter Shaw

Bibliography. G. B. Arfken and H.-J. Weber (eds.), *Mathematical Methods for Physicists*, 5th ed., 2000; G. Barton, *Elements of Green's Functions and Propagation*, 1989; E. N. Economou, *Green's Functions in Quantum Physics*, 1990; G. D. Mahan, *Many-Particle Physics*, 3d ed., 2000; R. D. Mattuck, *A Guide to Feynman Diagrams in the Many-Body Problem*, 2d ed., 1976, reprint 1992.

Green's theorem

A term used variously in mathematical literature to denote either (1) the Gauss divergence theorem (shown below) or some one of several forms or

$$\iiint \nabla \cdot \Im \, dV = \iint \Im \cdot \boldsymbol{\nu} \, dS$$

immediate consequences of their theorem, or (2) the plane case of Stokes' theorem. *See* GAUSS' THEOREM; STOKES' THEOREM.

The variation of the Gauss divergence theorem in which $\Im$ is the product $u \ \nabla v$ of a scalar function and the gradient of a scalar function is sometimes called Green's first identity. In this case Eq. (1) holds, and Gauss' theorem assumes the form of Eq. (2), where

$$\nabla \cdot \Im = \nabla \cdot (u \ \nabla v) = u \ \nabla^2 v + \nabla u \cdot \nabla v \qquad (1)$$

$$\iiint u \ \nabla^2 v \, dV + \iiint \nabla u \cdot \nabla v \, dV = \iint u \frac{\partial v}{\partial n} \, dS \qquad (2)$$

$\partial v/\partial n = \nabla \ v \cdot \boldsymbol{\nu}$ and is therefore the directional derivative of v in the direction of $\boldsymbol{\nu}$ the unit normal to the surface; thus n may be taken to be arc length along any curve of class C' which is tangent to $\boldsymbol{\nu}$ at the point in question. The derivative $\partial v/\partial n$ is a function of the surface coordinates.

By interchanging u and v in Eq. (2) with subsequent subtraction of the new form from the original, there results the relationship shown as Eq. (3),

$$\iiint u \ \nabla^2 v \, dV - \iiint v \ \nabla^2 u \, dV = \iint \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) dS \qquad (3)$$

which is known as Green's second identity as well as Green's theorem. Two additional equalities associated with Green may be obtained by taking $u = v$ and $u = 1$ in Eq. (2); thus Eqs. (4) and (5) result.

$$\iiint u \ \nabla^2 u \, dV + \iiint \nabla u \cdot \nabla u \, dV = \iint u \frac{\partial u}{\partial n} \, dS \qquad (4)$$

$$\iiint \nabla^2 v \, dV = \iint \frac{\partial v}{\partial n} \, dS \qquad (5)$$

The formulas (2)–(5) provide powerful devices for the investigation of differential equations involving the laplacian operator ∇^2, such as Poisson's equation and the wave equations. *See* CALCULUS OF VECTORS.

Homer V. Craig

Gregarinia

A subclass of the class Telosporea. These protozoans occur principally as extracellular parasites in the digestive tracts and body cavities of invertebrates. Their spores are formed directly by the zygote. There are three orders: the Archigregarinida, whose life cycle embraces both sexual and asexual phases; the Eugregarinida, which increases only by sporogony; and the Neogregarinida, whose life cycle involves schizogony and gamont formation. The most familiar gregarines belong to the Eugregarinida and are represented by two types: cephaline, whose trophozites (sporadins) are divided into an anterior protomerite and a larger posterior deutomerite by a transverse septum; and acephaline, which lack the septum. *See* ARCHIGREGARINIDA; NEOGREGARINIDA; PROTOZOA; SPOROZOA; TELOSPOREA.

Elery R. Becker; Norman Levine

Greisen

A type of hydrothermal wall-rock alteration and a class of tin-tungsten deposits (so-called greisen deposits). Hydrothermal wall-rock alteration is the process whereby rocks on the margins of hydrothermal flow channels are changed from an original assemblage of minerals to a different one. This change occurs because of heat and mass exchange between water and rock. The term greisen was originally used by miners in Saxony with reference to relatively coarse-grained aggregates of quartz and muscovite found on the borders of tin veins in granite of the Erzegebirge.

Granitic rocks altered to greisen are known as apogranites. They are composed mainly of quartz, topaz (fluor-aluminosilicate), and muscovite (white mica), accompanied by accessory minerals such as tourmaline (complex boro-aluminosilicate), fluorite (calcium fluoride), and zinnwaldite (iron-lithium mica). Abundant veins of quartz-topaz are characteristic of intensely greisenized zones. Skarn and limestone on the margins of apogranites may also be altered to greisen (aposkarn greisen and apocarbonate greisen, respectively) with abundant fluorite. Apogranite greisen commonly is accompanied by other types of hydrothermal wall-rock alteration, including early feldspathic and late sericitic and lesser argillic.

Tin-tungsten-(beryllium-molybdenum) deposits in peraluminous granites commonly are accompanied by greisen. Ore minerals may include cassiterite, wolframite, scheelite, molybdenite, bismuth, and bismuthinite, accompanied in some deposits by pyrrhotite and sphalerite, in addition to chalcopyrite and other sulfides. *See* GRANITE.

Tin greisens represent the dominant world source of lode tin, with examples in Southeast Asia (Malaya, Indonesia, Burma, Thailand); southeast China; Tasmania, Australia; Zinnwald and Altenberg (Erzegebirge), Germany; and Cornwall-Devon, southwest England. Greisenized skarn and apocarbonate greisen, also mostly tin deposits (but including beryllium and tungsten), are found in western Tasmania, Australia; Seward Peninsula, Alaska; and Yunnan (tungsten), China. Many of the tungsten deposits of southeast China, the richest tungsten province in the world, occur in greisenized granite. *See* TIN; TUNGSTEN.

Although a continuum likely exists between greisen and lower-temperature sericitic and argillic alteration, a distinction is made between these different alteration types on the basis of fluoride-, boron-, and lithium-enrichment in greisen. Additional mass changes relative to fresh granite include loss of sodium and calcium, gain or loss of potassium and silica, and gain of volatile substances (especially fluorine and sulfur). The chemical components added during greisenization originate in a magmatic aqueous fluid given off during the final stages of crystallization of granitic magmas. Greisenization and ore deposition results from the interaction of these fluorine-rich fluids with granite and its wall rocks as temperatures decline below 600°C (1110°F). *See* METASOMATISM; ORE AND MINERAL DEPOSITS; PNEUMATOLYSIS.

Marco T. Einaudi

Bibliography. R. G. Taylor, *Geology of Tin Deposits*, 1979.

Grignard reaction

A reaction between an alkyl or aryl halide and magnesium metal in a suitable solvent, usually absolute ether. This is represented by reaction (1), where R

$$\mathrm{RX + Mg \xrightarrow{ether} RMgX} \qquad (1)$$

stands for alkyl or aryl, and X for chlorine, bromine, or iodine. The organomagnesium halides produced by this reaction are known as Grignard reagents and are useful in many chemical syntheses. They are named after Victor Grignard, who discovered them and developed their use as synthetic reagents, for which he received a Nobel prize in 1912.

The scope of the Grignard reaction is extremely broad, and Grignard reagents have been prepared from many kinds of alkyl and aryl halides. In general, alkyl chlorides, bromides, and iodides and aryl bromides and iodides react readily. A few halides, such as aryl chlorides, react very sluggishly, and require specially activated magnesium, modified reaction techniques, the use of high-boiling solvents, and long reaction periods. Vinylmagnesium halides can be prepared by using tetrahydrofuran as the solvent.

The structure of a Grignard reagent is usually written RMgX, where X represents a halogen. However, the actual structure in ether solution is more complex. Equilibrium is rapidly established between the organomagnesium halide (RMgX) and the corresponding dialkylmagnesium (RMgR), as in reaction (2). These two species (RMgX and R_2Mg) are

$$\mathrm{2RMgX \rightleftharpoons R_2Mg + MgX_2} \qquad (2)$$

reactive, and are solvated in ether solvents by coordination of the ether oxygen to magnesium. In solution they further associate as dimers or higher polymers. Thus, while oversimplified, a Grignard reagent can be considered as RMgX.

Probably the most common synthetic use of the Grignard reaction involves the reaction of Grignard reagents with carbonyl compounds, followed by hydrolysis to produce a variety of products containing new carbon-carbon bonds. For example, in reaction (3), a Grignard reagent with carbon dioxide at

$$\mathrm{RMgX + CO_2 \rightarrow RCOOMgX \xrightarrow{H_2O} RCOOH + MgXOH} \qquad (3)$$

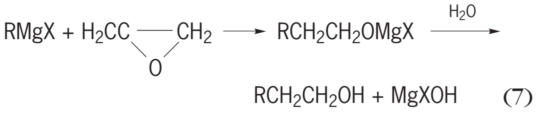
low temperatures produces salts of carboxylic acids, from which the acids can be readily isolated. Additon of Grignard reagents to the carbonyl group of many aldehydes and ketones produces primary, secondary, or tertiary alcohols. The production of these alcohols is shown in reactions (4)–(6).

$$\mathrm{RMgX + CH_2{=}O \longrightarrow RCH_2OMgX \xrightarrow{H_2O} \underset{\text{Primary alcohol}}{RCH_2OH} + MgXOH} \qquad (4)$$

$$\mathrm{RMgX + R'CH{=}O \longrightarrow RR'CHOMgX \xrightarrow{H_2O} \underset{\text{Secondary alcohol}}{RR'CHOH} + MgXOH} \qquad (5)$$

$$\mathrm{RMgX + R'R''C{=}O \longrightarrow RR'R''COMgX \xrightarrow{H_2O} \underset{\text{Tertiary alcohol}}{RR'R''COH} + MgXOH} \qquad (6)$$

The reaction of Grignard reagents with carbon dioxide, aldehydes, and ketones serves as a convenient means of lengthening the carbon chain. Further, a hydrocarbon chain can be increased by two carbon atoms by the reaction of a Grignard reagent with ethylene oxide, as in reaction (7). The resulting alcohol can be converted to the corresponding halide, and the process repeated.



(7)

Grignard reagents react with esters to produce ketones initially. The reaction as a rule does not stop at this stage, however, since additional Grignard reagent reacts with the ketone to yield a tertiary alcohol.

Acid chlorides and anhydrides react in an analogous manner to produce ketones, which react further to give tertiary alcohols. A valuable modification of this procedure involves prior reaction of the Grignard reagent with cadmium or zinc chloride. The organocadmium or organozinc compounds are less reactive than those of magnesium and react so slowly with ketones that the conversion of the ketone to the tertiary alcohol can usually be avoided.

When aliphatic Grignard reagents are exposed to air or oxygen, they are oxidized to alkoxides, which on hydrolysis produce alcohols. Aromatic Grignard reagents react poorly in this way, and only unsatisfactory yields of phenols can be obtained.

The reaction of Grignard reagents with alkyl halides which contain reactive halogen atoms produces hydrocarbons by a process of alkylation, as in reaction (8). In a similar manner, the addition

$$\mathrm{RMgX + XR \longrightarrow R{-}R + MgX_2} \qquad (8)$$

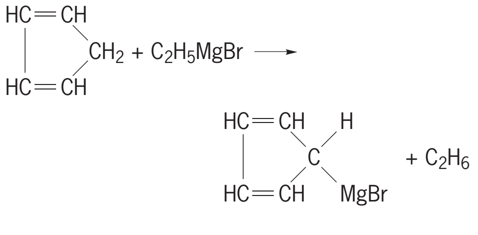
of a variety of metal halides, such as silver bromide or cobaltous chloride, tends to couple Grignard reagents, as in reaction (9).

$$\mathrm{2RMgX + 2AgBr \longrightarrow R{-}R + 2Ag + 2MgXBr} \qquad (9)$$

Compounds containing active hydrogen atoms react readily with Grignard reagents. Reaction with water brings about immediate decomposition and formation of the corresponding hydrocarbon, as in reaction (10). In certain instances, hydrogen attached to carbon is sufficiently acidic to react with Grignard reagents. The reactions of ethyl-magnesium bromide with acetylene, reaction (11), and cyclopentadiene, reaction (12), are examples. These reactions

$$\mathrm{RMgX + H_2O \longrightarrow RH + MgXOH} \qquad (10)$$

$$\mathrm{2C_2H_5MgBr + HC{\equiv}CH \longrightarrow 2C_2H_6 + BrMgC{\equiv}CMgBr} \qquad (11)$$



(12)

enable formation of additional Grignard reagents which could otherwise not be prepared by conventional methods.

Grignard reagents react with a variety of nitrogen- and sulfur-containing compounds, such as imines, nitriles, and sulfoxides. Furthermore, the reaction of Grignard reagents with halides of boron, phosphorous, silicon, and tin is a very convenient procedure for producing organometallic compounds of these elements. An example, reaction (13), is the

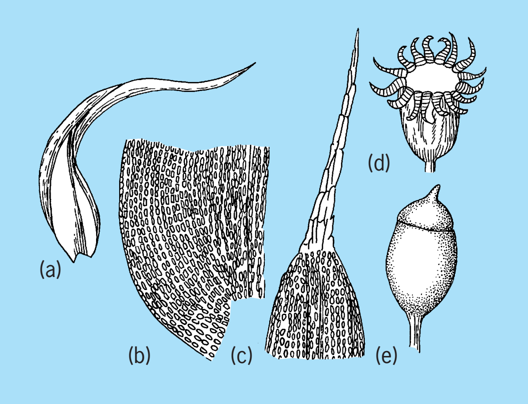
$$\mathrm{4C_6H_5MgBr + SiCl_4 \rightarrow (C_6H_5)_4Si + 4MgBrCl} \qquad (13)$$

preparation of tetraphenylsilane from phenylmagnesium bromide and silicon tetrachloride. *See* ORGANOMETALLIC COMPOUND. Paul E. Fanta

Bibliography. L. Kurti and B. Czako, *Strategic Applications of Named Reactions in Organic Synthesis*, 2005; J. J. Li, *Name Reactions: A Collection of Detailed Reaction Mechanisms*, 2d ed., 2003; M. B. Smith and J. March, *March's Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*, 5th ed., 2001; T. W. G. Solomons, *Organic Chemistry*, 8th ed., 2003.

Grimmiales

An order of the true mosses (subclass Bryidae) which consist of two families and about five genera and generally grow in dry, exposed places, especially on rock. They are often dark, even blackish, with erect-ascending and simple or forked stems, or prostrate and freely branched stems. The leaves, in many rows, are often hair-pointed (see **illus.**) with a single and well-developed costa. The cells are generally smooth but commonly have side walls that are nodulose-thickened. The capsules are immersed to exserted (sometimes on curved setae), and are usually erect, symmetric, and sometimes ribbed. The operculum is differentiated, and the single peristome (rarely lacking) is composed of 16 teeth which are commonly perforated or variously cleft. The calyptrae are cucullate or mitrate. The chromosome numbers are generally 12 and 13.

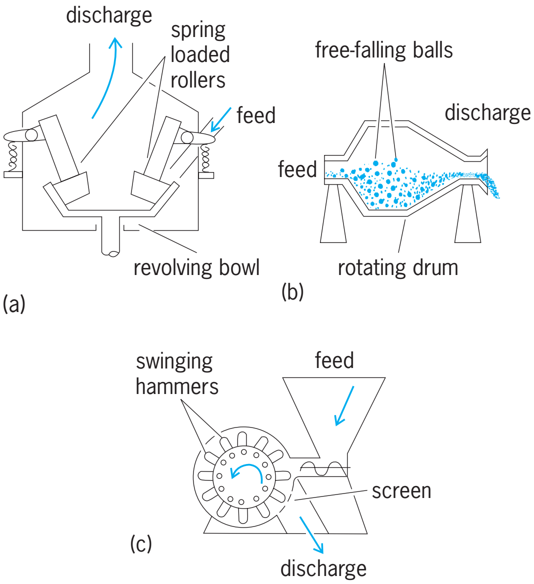


***Grimmia apocarpa*; grows generally on rock. (*a*) Leaf. (*b*) Portion of base of leaf. (*c*) Apex of leaf. (*d*) Urn and peristome. (*e*) Capsule. (*After W. H. Welch, Mosses of Indiana, Indiana Department of Conservation, 1957*)**

Grimmia has many resemblances to the genus *Orthotrichum*, which belongs to a double-peristome series. Both Hypnales and Orthotrichales seem intermediate between the acrocarpous and pleurocarpous groups of mosses and grow in similar, dry habitats. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA; ORTHOTRICHALES. Howard Crum

Grinding mill

A machine that reduces the size of particles of raw material fed into it. The size reduction may be to facilitate removal of valuable constituents from an ore or to prepare the material for industrial use, as in preparing clay for pottery making or coal for furnace firing. Coarse material is first crushed. The moderate-sized crushings may be reduced further by grinding or pulverizing.



Basic grinding mills. (*a*) Ring-roller mill. (*b*) Tumbling mill. (*c*) Hammer mill.

Grinding mills are of three principal types, as shown in the **illustration**. In ring-roller pulverizers, the material is fed past spring-loaded rollers. The rolling surfaces apply a slow large force to the material as the bowl or other container revolves. The fine particles may be swept by an airstream up out of the mill. In tumbling mills the material is fed into a shell or drum that rotates about its horizontal axis. The attrition or abrasion between particles grinds the material. The grinding bodies may be flint pebbles, steel balls, metal rods lying parallel to the axis of the drum, or simply larger pieces of the material itself. In hammer mills, driven swinging hammers reduce the material by sudden impacts. *See* BALL-AND-RACE-TYPE PULVERIZER; PEBBLE MILL; TUMBLING MILL.

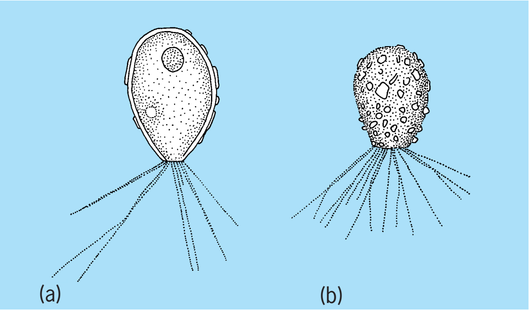
Depending on the required fineness and uniformity of the finished particles, the discharge may or may not be classified by size. Oversized particles may be returned in a closed circuit to the grinding mill for further reduction. Material may be ground dry or wet, in batches or continuously. *See* CRUSHING AND PULVERIZING; MECHANICAL CLASSIFICATION.

Ralph M. Hardgrove

Bibliography. F. T. Farago, *Abrasive Methods Engineering*, vols. 1–2, 1976–1980; R. H. Perry and D. Green (eds.), *Perry's Chemical Engineers' Handbook*, 7th ed., 1997.

Gromiida

An order of the subclass Filosia. The test of these protozoa is mostly chitinous in some species, rather thin and a bit flexible in others. Siliceous particles of endogenous origin, or sometimes minute platelets, may be embedded in the chitinous layer; some species typically reinforce the test with sand grains or other extraneous particles. In addition to *Gromia*, the genera *Chlamydophrys*, *Clypeolina*, *Plagiophrys* (**illus.** *a*), and *Pseudodifflugia* (illus. *b*), among others, have been assigned to the order. A few species are marine, the rest fresh-water types. Little is known about most Gromiida, although mitosis and encystment have been described in *Chlamydophrys*. *See* PROTOZOA; RHIZOPODEA; SARCODINA; SARCOMASTIGOPHORA.



Representative Gromiida. (*a*) *Plagiophrys parvipunctata*. (*b*) *Pseudodifflugia fulva*. (*After R. P. Hall, Protozoology, Prentice-Hall, 1953*)

Richard P. Hall

Ground-penetrating radar

A nondestructive technique using electromagnetic waves to locate objects or interfaces buried beneath the Earth's surface or located within a visually opaque structure; also termed ground-probing, surface-penetrating, or subsurface radar.

Principle of operation. Ground-penetrating radar (GPR) transmits a regular sequence of low power bursts of electromagnetic energy into the ground and detects the weak reflected signal from the buried target. The energy is in the form of either a very short duration impulse or a sweep over a range of frequencies. Most GPR systems, which are regulated and licensed by the countries in which they are used to comply with radio spectrum requirements, operate within the range of frequencies from 10 MHz to 10 GHz and can have a bandwidth of several gigahertz. The average radiated power is in the order of a thousandth of a watt. The receiver is highly sensitive and can detect reflected signals of less than 10^{-12} W. *See* MICROWAVE; RADAR.

The buried target can be a conductor of electricity, a nonconducting dielectric, or combinations of both; the surrounding host material can be soil, earth materials, rocks, ice, fresh water, or human-made materials such as concrete or brick. A typical GPR achieves

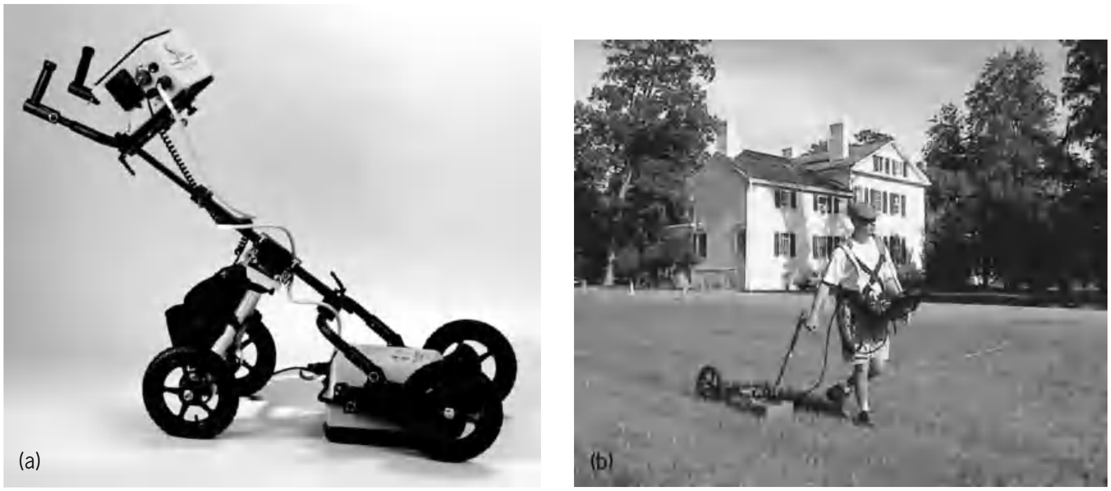


Fig. 1. Ground-penetrating radar equipment. (*a*) Portable equipment with 500-MHz antenna and controller that displays scrolling vertical slices of the soil in real time (*USRADAR*). (*b*) Portable equipment with 400-MHz antenna being used in archeological survey at Franklin D. Roosevelt's Hyde Park Estate in New York (*Geological Survey Systems, Inc.*).

a range of up to a few meters in depth, but some special systems can penetrate up to hundreds of meters. A few GPR systems have been operated from aircraft and even from satellites. The range of GPR in the ground is limited because of the absorption the signal undergoes while it travels on its two-way path through the ground. GPR works well through materials such as granite, dry sand, snow, ice, and fresh water, but it will not penetrate clay or salt water because of the high absorption of electromagnetic energy by these materials. In air the GPR signal travels at the speed of light, but it is slowed down in materials in the ground by their permittivities (dielectric constants). Hence true range requires calibration for each material. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; PERMITTIVITY.

GPR can be used on vehicles for rapid survey by means of an array of antennas. Other GPR systems are designed to be inserted into boreholes to provide tomographic images of the intervening rock. Most GPR systems use separate, portable, transmit-and-receive antennas which are placed on the surface of the ground and moved linearly to provide an image of the cross section of the ground traversed (**Fig. 1**). By systematically surveying the area in a regular grid pattern, a radar image of the ground can be built up. GPR images are displayed either as two-dimensional representations, using horizontal (x or y) and depth (z) axes or a horizontal plane representation (x,y) at a given depth (z), or as a three-dimensional reconstruction.

The image of a buried target generated by a GPR does not correspond to its geometrical equivalent and is generally of much lower resolution. Unprocessed GPR images often show bright spots caused by multiple internal reflection, as well as a distortion of the aspect ratio of the image of the target caused by variations in the velocity of propagation. Symmetrical targets, such as spheres or pipes, cause migration of the reflected energy to a hyperbolic pattern. GPR images can be processed to compensate for these effects. This is usually done off-line. GPR can be

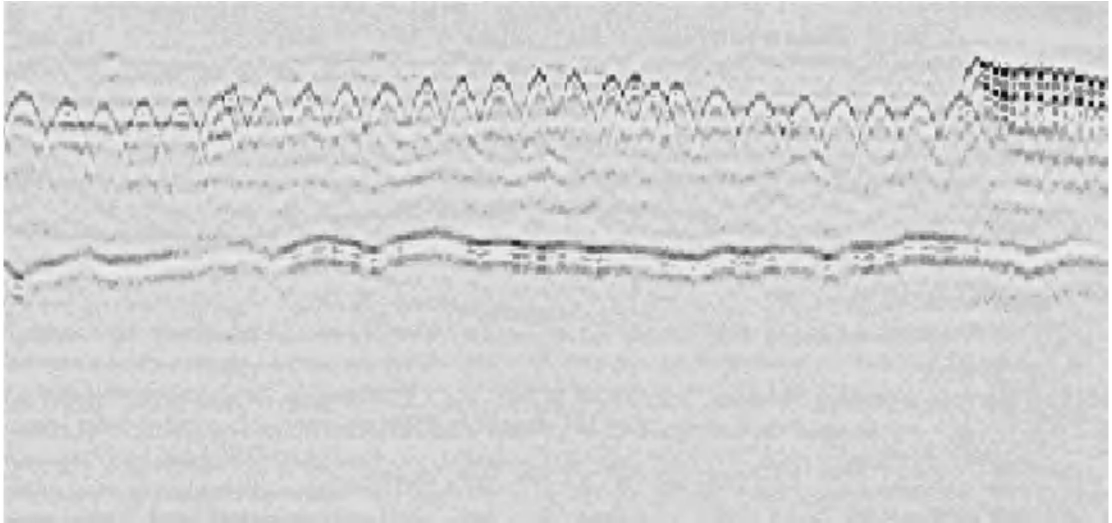


Fig. 2. Radar cross-sectional image of reinforced concrete road. (*ERA Technology UK*)

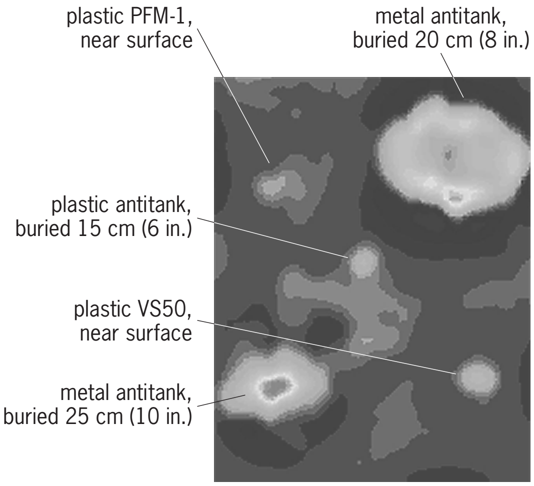

plastic PFM-1, near surface
metal antitank, buried 20 cm (8 in.)
plastic antitank, buried 15 cm (6 in.)
plastic VS50, near surface
metal antitank, buried 25 cm (10 in.)

Fig. 3. Radar plan image of mines, area 0.9 by 1 m (3.0 by 3.3 ft). (*ERA Technology UK*)

designed to detect specific targets such as interfaces in roads, pipes, and cables; and localized objects such as cubes, spheres, and cylinders. It is capable of detecting features many hundreds of years old; hence a prospective site should remain unexcavated prior to survey, to preserve its information.

Applications. GPR technology has many applications, some well established and some still undergoing research and development. A typical radar cross-sectional image is shown in **Fig. 2**; it represents a survey line 4.5 m (14.9 ft) long over a reinforced concrete road. The positions of the steel reinforcing mesh rods can be seen at the top of the image, while the interface between the concrete and the road subbase can be seen at a depth of 0.5 m (1.7 ft) half way down the image. Plastic pipes and cables can also be detected by GPR, and this is one of the routine applications of modern GPR equipment.

Typical applications for GPR are as follows:

Archeological investigations
Bridge deck analysis
Borehole inspection
Building condition assessment
Contaminated land investigation
Detection of buried mines
(antipersonnel and antitank)
Evaluation of reinforced concrete
Forensic investigations
Geophysical investigations
Medical imaging
Pipes and cable detection
Planetary exploration
Rail track and bed inspection
Remote sensing from aircraft
and satellites
Road condition survey
Security applications
Snow, ice, and glacier
Timber condition
Tunnel linings
Wall condition

GPR has been used very successfully in forensic investigations. The most notorious cases occurred in the United Kingdom in 1994 when the grave sites, under concrete, of the victims of a serial murderer were pinpointed. Similar use of GPR was made in Belgium in 1996. GPR in the hands of an expert allows rapid, nondestructive investigation of suspected sites and saves much fruitless excavation.

Abandoned antipersonnel land mines and unexploded ordnance are a major hindrance to the recovery of many countries from war. Their effect on the civilian population is catastrophic, and major efforts are being made by the international community to eliminate the problem. Mines are located mostly with metal detectors, which respond to the large amount of metallic debris in abandoned battlefield areas and hence have difficulty in detecting minimum metal or plastic mines. GPR technology is being applied to this problem (**Fig. 3**).

Archeological applications of GPR have been numerous, ranging from attempts to detect Noah's Ark to the exploration of Egyptian and Indian sites as well as castles and monasteries in Europe. The quality of the radar image can be exceptionally good, although correct understanding normally requires

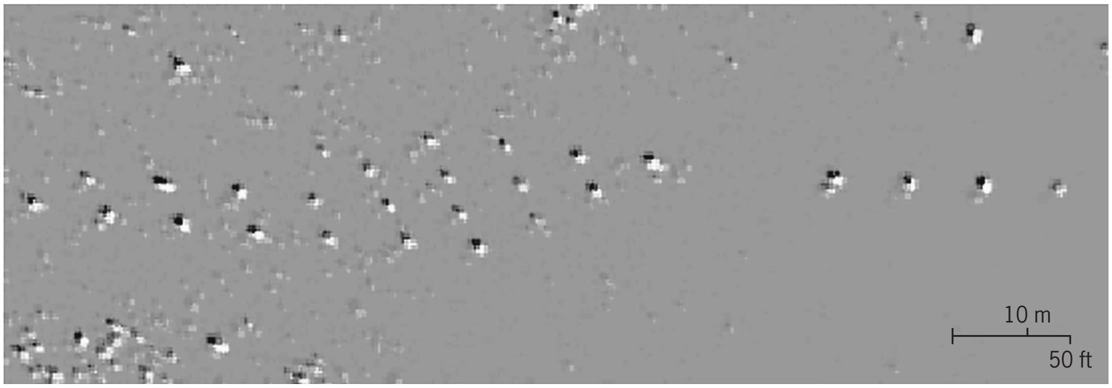


Fig. 4. Synthetic aperture radar (SAR) image of buried metal mines taken from an altitude of 400 m (1300 ft) above the Yuma desert. (*Courtesy of R. Vickers, SRI, USA*)

joint interpretation by the archeologists and radar specialists. GPR has been used for many different types of geological surveys ranging from exploration of the Arctic and Antarctic icecaps and the permafrost regions of North America, to mapping of granite, limestone, marble, and other hard rocks as well as geophysical strata. While most GPR systems are used close to the ground, airborne systems have been able to map ice formations and glaciers, and penetrate through forest canopy. Airborne GPR, processed using synthetic aperture techniques, has been used to detect buried metallic mines from a height of several hundred meters (**Fig. 4**). *See* ARCHEOLOGY; REMOTE SENSING. David J. Daniels

Bibliography. C. E. Baum (ed.), *Detection and Identification of Visually Obscured Targets*, Taylor and Francis, 1998; S. Cloude, *An Introduction to Electromagnetic Wave Propagation and Antennas*, UCL Press, 1995; D. J. Daniels (ed.), *Ground Penetrating Radar*, 2d ed., IET Publishing, 2004; J. D. Taylor (ed.), *Introduction to Ultra-wideband Radar Systems*, CRC Press, 1995.

Ground proximity warning system

A system carried on many aircraft to warn the pilot that the aircraft may be in danger of inadvertent contact with the ground. It is intended to reduce the occurrence of controlled-flight-into-terrain (CFIT) accidents, in which aircraft with no apparent mechanical difficulty or defect strike the ground while under the direct or indirect control of the pilot. These accidents usually occur in conditions of poor visibility (due to atmospheric obscuration such as fog or rain, or darkness of night) where external visual cues are not available to provide added assurance of terrain clearance. Since 1975, federal aviation regulations have required installation of the system on large turbine-powered aircraft in commercial service.

Components. The heart of the ground proximity warning system (GPWS) is a computer which receives inputs from several sensors on the aircraft and issues warnings to the pilot through visual and aural alerting devices. The primary sensors providing inputs to the computer are (1) the radio altimeter, which measures height of the aircraft above the terrain; (2) the barometric altimeter, which measures altitude with respect to an atmospheric pressure datum; (3) the electronic glideslope of the Instrument Landing System (ILS), which provides the pilot with vertical descent guidance during the final portions of an ILS approach and landing; and (4) sensors which indicate aircraft configuration such as the position of flaps and landing gear. When a hazardous condition is detected by the system, the pilot is warned by illumination of caution lights on the instrument panel, accompanied by an aural alarm and a synthetic voice message indicating the type of hazard. *See* ALTIMETER; INSTRUMENT LANDING SYSTEM (ILS).

Operation. The system is designed to detect and warn the pilot of the following conditions: excessive descent rate near the ground; excessive terrain closure rate; approaching the ground with landing gear or flaps not in the landing configuration; and descending significantly below the ILS electronic glideslope when on approach to landing. Also, during takeoff and immediately after initiating a missed-approach go-around, the system warns the pilot if the aircraft is descending when it should normally be climbing. The Radio Technical Commission for Aeronautics (RTCA), in its *Minimum Performance Standards for Airborne Ground Proximity Warning Equipment*, prescribes the parameters that must be monitored and the thresholds at which the system should declare a hazardous condition. There are provisions for the pilot to test the operation of the system, and also to inhibit certain warnings in unusual or emergency conditions (if, for instance, an approach and touchdown is planned with the landing gear or flaps retracted due to mechanical failure). However, the operation of the system generally requires no action on the part of the pilot.

Limitations. While GPWS has been credited with significantly reducing the incidence of CFIT accidents, it does have limitations. Its sensors cannot detect hazards which may be ahead of the aircraft such as steeply rising terrain or artificial obstacles. In addition, the effectiveness of GPWS is largely dependent on the pilot's prompt reaction to the system's warnings. However, these limitations have been surmounted in some respects with the availability of enhanced ground proximity warning systems (EGPWS) since 1996. The enhancement adds a detailed terrain database that can be interrogated to identify hazards along the current course in time for the pilot to take evasive action. Consequences of the tendency of pilots to delay taking evasive action are also mitigated by the additional time to react that is provided by an EGPWS. *See* AIR NAVIGATION. George W. Flathers II

Bibliography. R. Hurst and L. Hurst (eds.), *Pilot Error: The Human Factors*, 1982; *Minimum Performance Standards: Airborne Ground Proximity Warning Equipment*, Radio Technical Commission for Aeronautics, DO-161A, May 1976.

Ground state

In quantum mechanics, the stationary state of lowest energy of a particle or a system of particles. The ground state may be bound or unbound; when bound, its energy generally is a finite amount less than the energy of the next higher or first excited state. In the typical circumstances that the potential energy is zero at infinite separation, the magnitude of the negative ground-state energy is the binding energy, that is, the energy required to separate all the particles infinitely. *See* ENERGY LEVEL (QUANTUM MECHANICS); EXCITED STATE; NONRELATIVISTIC QUANTUM THEORY; NUCLEAR BINDING ENERGY. Edward Gerjuoy

Ground-water hydrology

The occurrence, circulation, distribution, and properties of any liquid water residing beneath the surface of the earth. This article excludes any water molecules which constitute mineral species such as clays. Generally, ground water is that fraction of precipitation which infiltrates the land surface and subsequently moves, in response to various hydrodynamic forces, to reappear once again as seeps or in a more obvious fashion as springs. Most of ground-water discharge is not evident because it occurs through the bottoms of surface water bodies; in fact, large fresh-water springs are relatively common on the ocean bottom off the eastern coast of Florida.

Ground water can be found, at least in theory, in any geological horizon containing interconnected pore space. Thus a ground-water reservoir (an analogy to an oil reservoir) can be a classical porous medium, such as sand or sandstone; a fractured, relatively impermeable rock, such as granite; or a cavernous geologic horizon, such as certain limestone beds. Ground-water reservoirs which readily yield water to wells are known as aquifers; in contrast, aquitards are formations which do not normally provide adequate water supplies, and aquicludes are considered, for all practical purposes, to be impermeable. These terms are, of course, subjective descriptions; the flow of water which constitutes an economically viable supply depends upon the intended use and the availability of alternative sources. *See* AQUIFER.

Physics of flow. Ground water and its dissolved constituents move according to mathematical relationships derived from the physics of flow through porous media. These relationships are simplest in the case of the saturated flow of a homogeneous fluid in a granular formation. There are a number of governing equations for this family of problems, and many solutions are available for problems of practical importance. There are also equations describing the behavior of dissolved constituents in flowing ground water. This class of problems has gained prominence because of the recognition of existing and potential instances of ground-water contamination. However, these equations are more difficult to solve, and consequently the simulation of most field situations involving contaminant transport requires the use of large digital computers. When the unsaturated zone must be considered, the governing equations are still available, but there is a paucity of solutions. The major research advances in unsaturated flow can be attributed to soil physicists, who are primarily concerned about water movement through the root zone during the irrigation of crops.

Regional flow. Ground-water flow in a regional sense is concerned with movement attributable to gravitational forces. Ground water moves from points of high potential energy to points of low potential energy. Water in transit from regional topographic highs to regional topographic lows dissipates energy through friction. The flow path may be relatively short or extend for hundreds of miles, depending on the physical system. The generally observed correspondence between changes in the elevation of the surface of the saturated zone and changes in the land topography is maintained because of a continual source of water via precipitation. When precipitation patterns change, the ground-water surface responds accordingly, rising in wet years and falling in dry years. The topographic lowlands often correspond to areas of ground-water discharge. They may be characterized by springs, wetlands, or surface water bodies such as lakes, rivers, and oceans. *See* WETLANDS.

Artesian systems. The water level in a drilled well is a measure of the potential energy of the ground water it encounters. In a recharge area, where water moves downward under gravitational force, the fluid potential energy, and therefore the water level in wells, decreases with depth. On the other hand, in a discharge area where ground water moves toward the surface, the fluid potential increases with depth. This latter situation results in a well-water elevation which is higher than that observed for the upper surface of the saturated zone in the vicinity of the well. This phenomenon has been designated an artesian head. In some instances the well-water-level elevation may actually be higher than the land surface elevation. In this situation the ground water flows naturally from the well, and no pump is required. Such wells are called flowing artesian wells. It has often been observed that the discharge of flowing wells decreases with time. This is a natural consequence of providing a ground-water flow path to the surface which is less energy-dissipating than the original porous medium flow. *See* ARTESIAN SYSTEMS.

Ground-water mining. It is important to recognize that when ground water is used consumptively, that is, when it is removed permanently from the hydrologic system, the ground-water reservoir must be affected. The most immediate impact is a lowering of the saturated zone elevation. As the influence of the pumping reaches surface water bodies, water is generally induced from them. The rate of loss from the surface water body will increase until an equilibrium is established between water withdrawn from the well and that entering the aquifer. If no surface water bodies are available, the ground-water surface will fall as long as ground water is utilized: this is called ground-water mining. Practically speaking, the process is self-limiting, because the cost of pumping ground water increases as the water levels in the wells decline.

Chemical and physical characteristics. As ground water moves through a regional flow system, its physical and chemical characteristics are modified by the environments encountered. It dissolves soluble ions, and thereby changes its chemical composition. Thus the chemistry of ground water can sometimes be used to decipher its complex flow history. During transit through the subsurface, when ground water encounters elevated rock temperatures, the water is

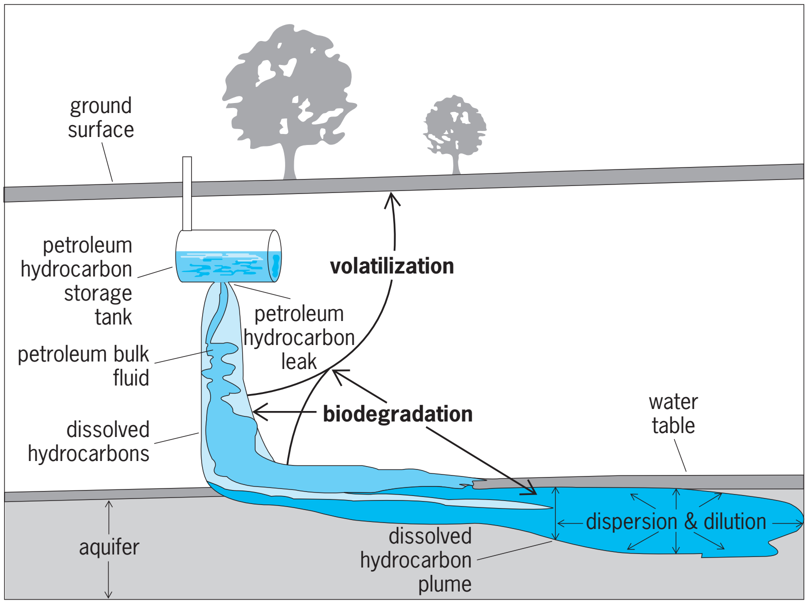


Fig. 1. Movement of a lighter-than-water petroleum hydrocarbon liquid in a flowing ground-water system. (*After D. F. Pope and J. N. Jones, Monitored Natural Attenuation of Petroleum Hydrocarbons, EPA/600/F-98/021, May 1999*)

heated. Under certain circumstances, the water may reach the surface as hot springs or geysers. When surface manifestations are absent, hot water and occasionally steam may be tapped by wells to provide a source of energy. High-temperature water can be used to produce electricity; lower-temperature water may heat homes, or may be employed effectively in selected industries. *See* GEOTHERMAL POWER.

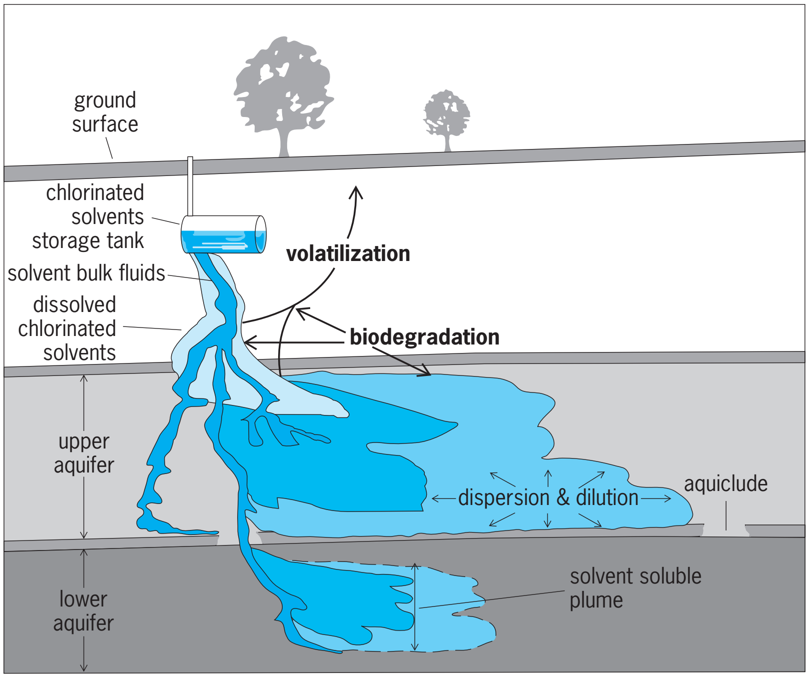


Fig. 2. Movement of a heavier-than-water chlorinated solvent in a flowing ground-water system. (*After D. F. Pope and J. N. Jones, Monitored Natural Attenuation of Chlorinated Solvents, EPA/600/F-98/022, May 1999*)

Reservoir behavior. To effectively utilize ground water as a natural resource, it is necessary to be able to forecast the impact of exploitation on water availability. When ground water is used for water supply, a concern is the potential energy in the aquifer as reflected in the water level in the producing well or neighboring wells. When a ground-water reservoir that does not readily transmit water is tapped, the energy loss associated with flow to the well can be such that the well must be drilled to prohibitively great depths to provide adequate supplies. On the other hand, in a formation able to transmit fluid easily, water levels may drop because the reservoir is being depleted of water. This is generally encountered in reservoirs of limited areal extent or those in which natural infiltration has been reduced either naturally or through human activities.

To forecast reservoir behavior, some type of model of the ground-water system must be employed. Such models may be statistical or deterministic. When problems involving a change in pumping demand are encountered, deterministic models founded on physical principles are generally employed. They can be physical (scale models of the prototype), electrical (based on the analogy between porous media flow and electric current flow), or mathematical (involving analytical or numerical solutions of the governing physical equations).

Early models focused on forecasting well performance. From the point of view of water supply, it was important to forecast how the water level in the pumping well and neighboring wells would respond to pumping. Analytical solutions to the appropriate governing equations were developed for a number of field situations. The solution for the long-term equilibrium case is known as the Thiem equation. The Theis equation is similar, but expresses transient behavior. These, and related expressions, can be readily modified to accommodate a number of possible hydrologic complications. For example, when a well is located in the vicinity of a stream, the stream will, unless hydraulically disconnected from the aquifer, eventually contribute some of its flow to satisfying the well discharge. On the other extreme, when a well is located near the aquifer boundary, the well water levels will be lower than would be encountered had the well been located away from any such impermeable barrier.

These analytical solutions have, in fact, a dual purpose. It is evident that, given appropriate aquifer parameters, these equations can be used to forecast ground-water reservoir response to pumping. A less evident application involves the determination of the aquifer parameters. This normally requires a carefully controlled field experiment called a pumping test, which normally involves a central pumping well and a series of nearby observation wells. Analysis of the response of the ground-water system after the cessation of pumping is known as the recovery test. Both the pumping and recovery tests are analyzed by using the analytical formulas discussed above.

Ground-water quality. Problems involving ground-water quantity were once the primary concern of hydrologists; interest is now focused on ground-water quality. Ground-water contamination is a serious problem, particularly in the highly urbanized areas of the United States. Legislation restricting the design and construction of waste-disposal facilities has accelerated and expanded ground-water quality studies. Predictive models which describe the convective and dispersive transport of contaminants have been developed and applied to field situations. Such models can be used to investigate the effectiveness of remedial measures designed to contain or remove contaminated ground water. *See* WATER POLLUTION.

Nonaqueous-phase contamination. The recent discovery of ground-water contamination by toxic liquids that are only slightly soluble in water has focused attention on the characteristics and behavior of non-aqueous-phase liquids in the subsurface. A lighter-than-water non-aqueous-phase liquid reaches the water table, whereupon it begins to spread and extend its area (**Fig. 1**). A denser-than-water intruding liquid moves downward through the water table and continues vertically until it encounters an impeding geological horizon (**Fig. 2**). It then begins to move in the direction of the slope of the top of this horizon. Due to the fact that many non-aqueous-phase liquids have low solubility, they present a significant threat to ground-water quality over long periods of time.

Applications. Although this discussion has considered the ground-water phenomenon in a rather narrow sense, the same fundamental principles are applicable to a number of related but distinctly different disciplines. Soil mechanics, the study of foundation engineering, relies heavily upon porous flow physics for its governing equations. This is also true of excavation dewatering in construction and mining, drainage problems in agricultural engineering, and reservoir simulation problems in the oil, gas, and geothermal industries. Ground-water hydrology has played a historically important role in the exploitation of water resources in the United States, and continues to contribute to their protection and effective utilization. *See* HYDROLOGY; SOIL MECHANICS.

George F. Pinder

Access to ground-water resources by wells has enabled the growth of agriculture, population, and commerce throughout the world. The types and uses of wells vary substantially, reflecting the widespread dependence on water supply as well as the efforts to assess and protect the quality and quantity of water for future uses. *See* WELL.

Water supply wells. These serve a variety of water needs ranging from those of an individual home or farmstead to those of urban population centers and industries. Water supply wells are usually designed, constructed, and operated in order to maximize the yield of water. The design factors should include both durable materials and pumping strategies that will provide long-term sustained yield from the aquifer (**Fig. 3**). The diameter of these wells may range from 4 to greater than 24 in. (10 to > 60 cm) depending on the aquifer properties and the pump required for lifting the amount of water needed. The well screen interval is normally placed in the most transmissive (or high-water-yielding) geologic formation available. Above the screen, a well casing made of iron, steel, or a thermoplastic material completes the subsurface portion of the well. Surface completion methods vary; however, the casings are carefully sealed with clay and cement grouting materials to ensure the surface water does not enter the well. In this way, the infiltration of microbial organisms or runoff chemicals from the land surface are prevented from entering the produced water. A pump and water discharge line are contained within the well and are fitted with appropriate valves at the wellhead.

The hydraulics of well operation (Fig. 1) relate to the response of the static water level in the aquifer to pumping. Pumping causes a decline in the water level (drawdown; Fig. 1) to the pumping water level. In certain hydrogeologic settings, the water level

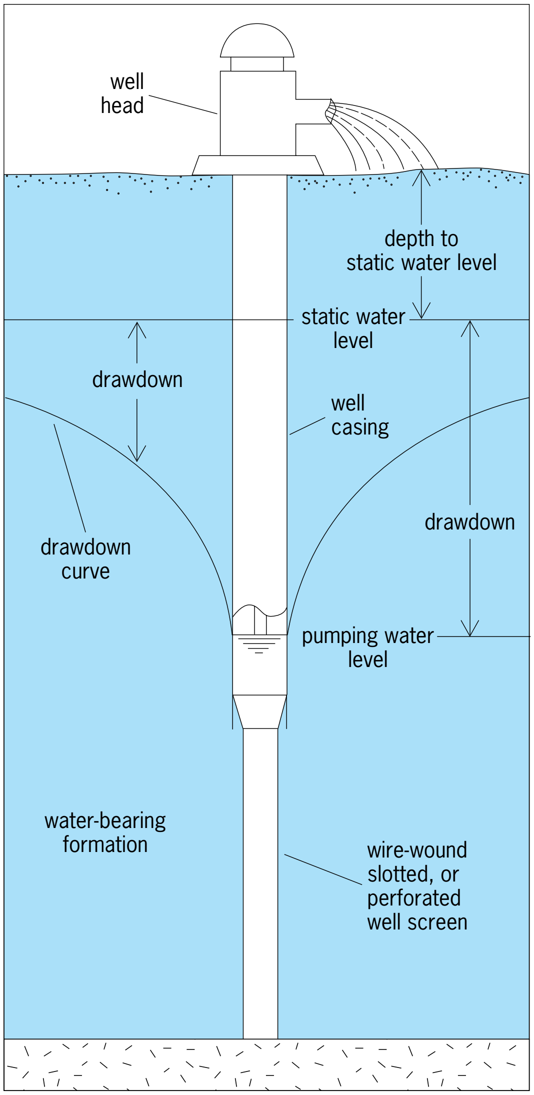


Fig. 3. Diagram of a production well, showing aspects of design that relate to well hydraulic performance. (*After F. G. Driscoll, Groundwater and Wells, 2d ed., Signal Environmental Systems, Johnson Division, 1986*)

in the aquifer may be above the land surface (because the water in the geologic formation is confined or is at higher pressure than overlying formations), and the well may yield water without pumping. Such wells are termed artesian. Once completed, the well would provide water without pumping as long as the pressure differential between the land surface and the screened formation was maintained.

The long-term yield and hydraulic performance of production wells are usually evaluated on the basis of a pumping test, which provides information on the hydraulic properties of the aquifer under pumping conditions. In this type of test, an array of water-level-observation wells are located within the estimated zone of drawdown of the production well. In a typical configuration, there are several small-diameter observation wells screened in the same formation as the production well (**Fig. 4**). One observation well (no. 4 in Fig. 4) would be used to determine if the clay layer shown between 46 and 50 ft (14 and 15 m) below land surface acts as a confining layer between the upper water-table aquifer and the lower aquifer. Water levels would be collected before pumping of the production well was begun, and then at successive time intervals over several days to weeks as the pumping water levels declined from static (nonpumping) levels. The amount of water in storage in the aquifer and the transmissive properties of the geologic formation can then be determined in order to establish a pumping strategy that would provide long-term yield of water without causing excessive drawdown (so-called mining) of the resource.

Monitoring or observation wells. These wells, also known as piezometers, are often used to permit the observation of water levels and to collect samples of ground water for microbiological or chemical analyses. They have many similarities to water supply wells in design and construction. The principal difference between monitoring and production wells is that monitoring wells are most often screened in discrete formations to provide specific hydrogeologic and chemical information rather than to produce large volumes of water. In general, these wells have diameters from 1 to 4 in. (2.5 to 10 cm), and the screened intervals are of the order 2 to 10 ft (0.6 to 3 m). The construction materials may be similar to those used for water supply wells, but designs often call for the use of stainless steel or chemical-resistant thermoplastics for durability under potentially contaminated subsurface conditions (**Fig. 5**). The precautions to seal the monitoring well bore from surface infiltration with bentonite clay and cement may be far more involved than in the construction of a production well. The need to achieve a proper seal becomes all the more important when surface soils or overlying formations may be contaminated. Percolation of contaminated water down the well bore would otherwise bias the analytical results from the well.

Since monitoring or observation wells are designed to provide depth-discrete information, they are frequently built in nested installations. In this situation, two or more wells are finished at different depth ranges to provide information on vertical hydraulic-head (that is, pressure) differences or concentration distributions of chemical or microbial parameters.

Water sample collection is conducted in monitoring wells by small-diameter submersible pumps, surface pumps (if the static water level is within approximately 18 ft or 5.4 m of land surface), or grab samplers. Water-level measurements are made during sampling events to record water levels relative to a surveyed benchmark for analysis of ground-water flow velocity and direction.

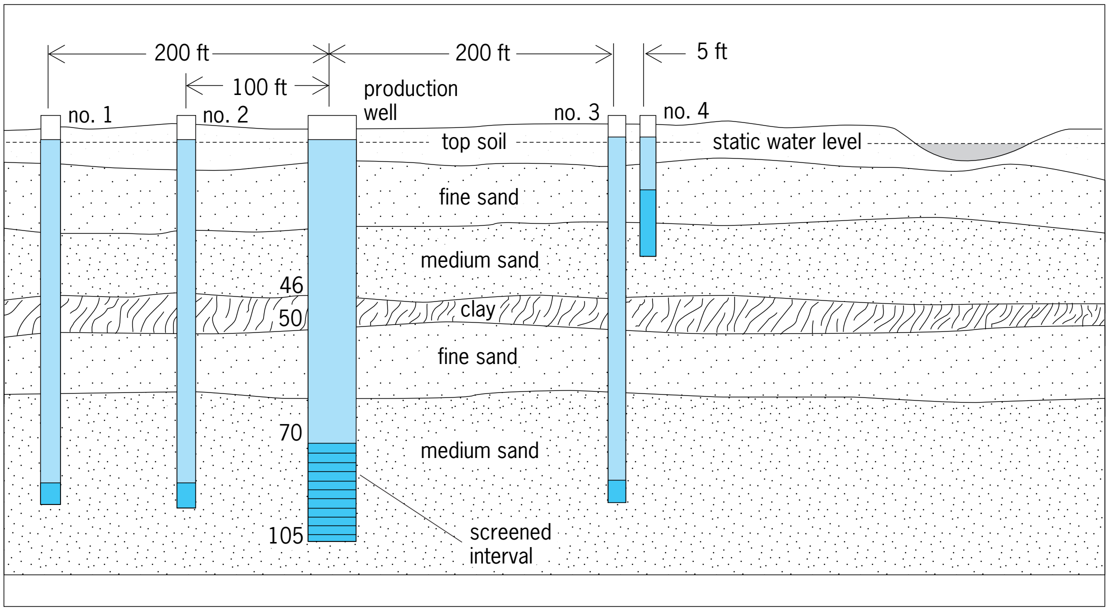
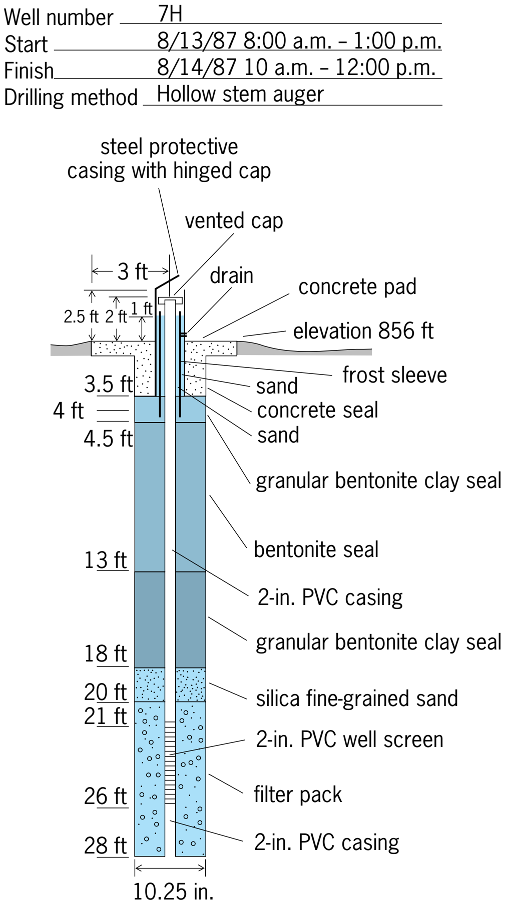


Fig. 4. Geologic section on a line through a test well and observation wells. Numbers along the production well indicate depth in feet. 1 ft = 0.3 m. (*After F. G. Driscoll, Groundwater and Wells, Signal Environmental Systems, Johnson Division, 2d ed., 1986*)



Well number 7H

Start 8/13/87 8:00 a.m. – 1:00 p.m.

Finish 8/14/87 10 a.m. – 12:00 p.m.

Drilling method Hollow stem auger

Fig. 5. Design detail for a typical monitoring well. (*After L. Aller, Handbook of Suggested Practices for the Design and Installation of Ground-Water Monitoring Wells, National Water Well Association, 1989*)

Alternative designs for nested monitoring installations include multilevel sampling arrays that employ discrete sampling ports linked to the wellhead by tubing. In these cases, mechanical samplers or surface pumps are used to retrieve water samples. These types of installations can provide significantly more vertical detail in chemical distributions than nested, conventionally screened monitoring wells. However, their use has been linked mainly to shallow depths (less than 100 ft or 30 m below land surface) because of constraints on the size of the sampling mechanism and the lift capabilities. *See* WATER SUPPLY ENGINEERING. Michael J. Barcelona

Bibliography. M. J. Barcelona et al. (eds.), *Handbook of Groundwater Protection*, 1988; J. Bear, *Hydraulics of Groundwater*, 1980; P. B. Bedient, H. S. Rifai, and C. J. Newell, *Groundwater Contamination*, 1999; R. J. Charbeneau, *Groundwater Hydraulics and Pollutant Transport*, 1999; S. N. Davis and R. J. M. De Wiest, *Hydrogeology*, 1966, reprint 1991; L. Dingman, *Physical Hydrogeology*, 1993; F. G. Driscoll, *Groundwater and Wells*, 2d ed., 1986; C. W. Fetter, Jr., *Applied Hydrogeology*, 1980; R. A. Freeze and J. A. Cherry, *Groundwater*, 1979; M. E. Harr, *Groundwater and Seepage*, 1962, reprint 1992; D. M. Nielsen (ed.), *Practical Handbook of Ground Water Monitoring*, 1991; M. E. Renz, *Practical Groundwater Hydrology*, 1995; U.S. Environmental Protection Agency, *Ground Water Handbook*, vol. 1: *Ground Water and Contamination*, EPA 625/6-90/016a, and vol. 2: *Methodology*, EPA 625/6-90/016b, 1991; F. van der Leeden, F. L. Troise, and D. K. Todd, *The Water Encyclopedia*, 2d ed., 1990.

Grounding

Intentional electric connections to a reference conducting plane, which is typically earth (hence the use of the term “ground,” of “earth” in British usage), but which more generally consists of a specific array of interconnected electrical conductors, referred to as the grounding electrode conductor. The symbol that denotes a connection to the grounding conductor is three parallel horizontal lines, the lower two being shorter than the one above (**Fig. 1**). The electric system of an airplane, spacecraft, satellite, or ship observes specific grounding practices with prescribed points of grounding, but no connection to earth is involved in this scheme. A connection to such a reference grounding conductor which is independent of earth is denoted by the symbol shown in **Fig. 2**.

The subject of grounding may be divided into two categories: system grounding and equipment

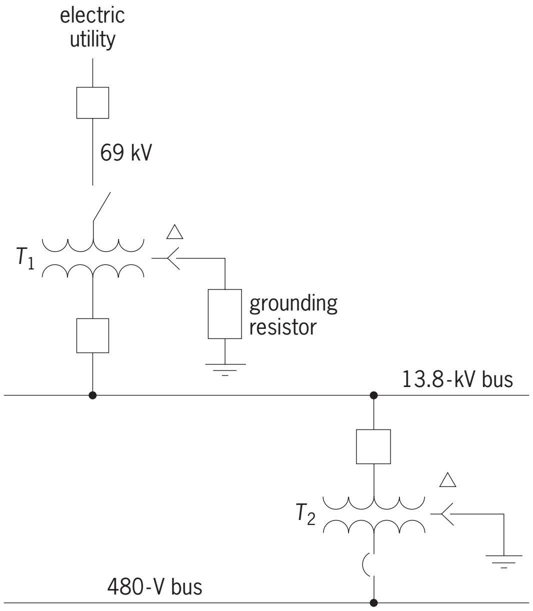


Fig. 1. Each conductively isolated portion of a distribution system requires its ground.

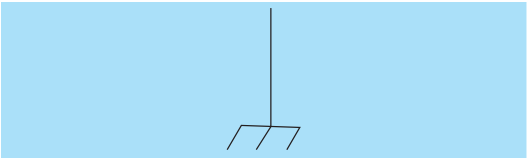


Fig. 2. Symbol to denote connection to a reference ground that is independent of earth.

or "safety" grounding. System grounding refers to an intentional connection from the electric power system conductors to earth for the purpose of securing superior performance qualities in the electrical system. The term "equipment or safety grounding" relates to a grounding connection from the various electric-machine frames, equipment housing, metal raceways containing energized electrical conductors, and closely adjacent conducting structures judged to be vulnerable to contact by an energized conductor. The purpose of such equipment (safety) grounding is to avoid environment hazards such as electric shock to area occupants, fire ignition hazard to the building or contents, and sparking or arcing between building interior metallic members which may be in loose contact with one another. Such ground connections provide a low-impedance path for the return of fault current, which will cause the overcurrent protective device on the affected circuit to "open," thereby removing the hazard. The design of outdoor open-type installations presents special grounding problems.

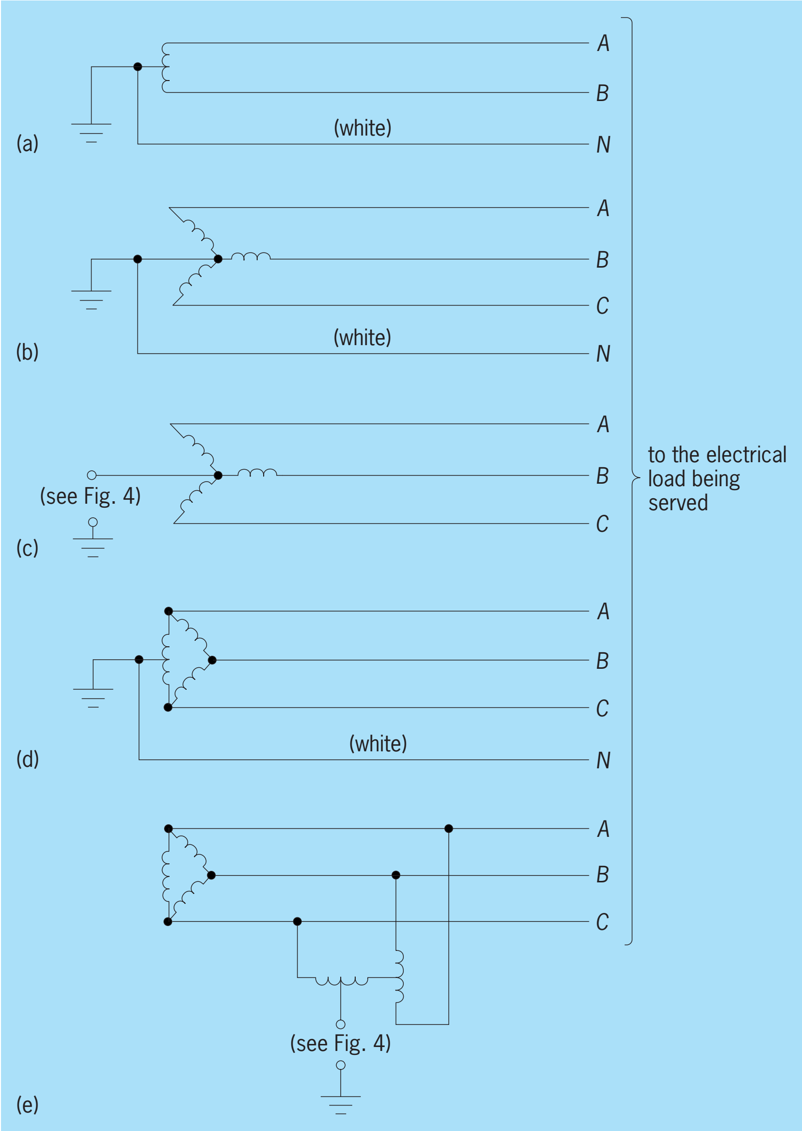


Fig. 3. Commonly used patterns of grounding. (*a*) Single-phase three-wire 240/120-V service. (*b*) Three-phase four-wire 208/120- or 480/277-V service. (*c*) Three-phase three-wire pattern. (*d*) Three-phase four-wire Δ 240-V line-to-line pattern. (*e*) Three-phase Δ with derived neutral.

System grounding. Appropriately applied, system grounding can (1) avoid excess voltage stress on electrical insulation within the system, leading to longer apparatus life and less frequent breakdown; (2) improve substantially the operating quality of the overcurrent protection system; (3) greatly diminish the magnitude of arc fault heat energy released at an insulation breakdown point, lessening arc burning damage and fire ignition possibility; and (4) provide for lightning discharge and minimize damage should high voltages be accidentally impressed on lower-voltage system conductors and equipment.

Sectionalization. Each voltage transformation point employing an insulating transformer interrupts the continuity of the system grounding circuit, and is defined as a "separately derived system" by the National Electrical Code (NFPA-70-2005) [Fig. 1]. System grounding connections made on the 69-kV electric service companies' lines extend their influence only to the 69-kV winding of transformer T_1. The grounding connections established at the 13.8-kV winding of transformer T_1 extend its influence only to the 13.8-kV winding of transformer T_2. The grounding connections established at the 480-V secondary winding of transformer T_2 apply to the 480-V conductor system only. Two distinct advantages result from this. First, the system grounding arrangement of each voltage-level electric system is independent of all others. Second, the type and pattern of system grounding to be used with any individual voltage-level electric system can be tailored to optimize the performance of that particular electric-system section. *See* ELECTRIC POWER SYSTEMS.

It is mandatory to locate the grounding connection at the source-point electrical neutral of the particular voltage-level separately derived system, and mandatory to do so at the service entrance point if the point of origin is outside the local building.

Common patterns. The great majority of system grounding patterns fall into one of the varieties that are shown in **Fig. 3**. The most used varieties of system grounding impedance are illustrated in **Fig. 4**.

The use of solid grounding exclusively for grounding patterns of Fig. 3*a*, *b*, and *d* are influenced by two considerations: (1) Overcurrent protection is present in only the phase conductor of single-phase, one-side grounded circuits. (2) The National Electrical Code (NEC) requires any electric system that can be solidly grounded to limit phase-to-ground potential to not more than 150 V.

Solid grounding is also used almost exclusively in the case of operating voltages of 69 kV and higher, as in Fig. 3*c*, to achieve the most rigid control of overvoltage stress. Such control allows reduced-rating lightning arresters, which in turn permits the successful use of reduced insulation level on station apparatus and equipment.

The high, unrestricted magnitude of short-circuit current created by a line-to-ground (L-G) fault on

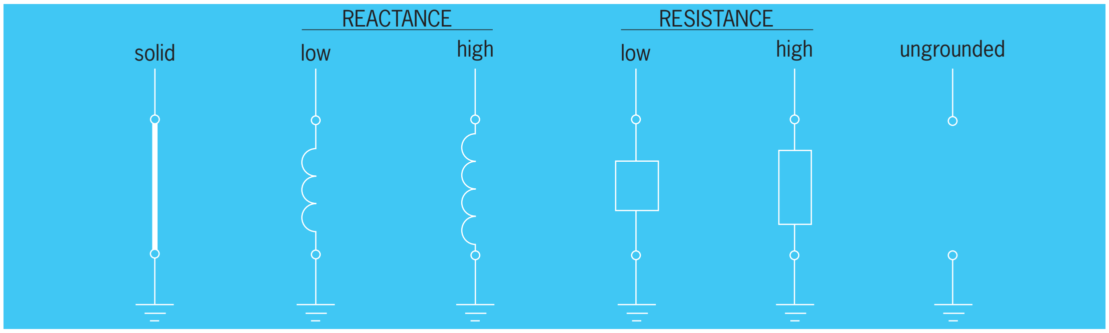


Fig. 4. Varieties of system grounding impedances in common use.

a solidly grounded system can pose severe design problems with costly solutions. The desire to artificially reduce the magnitude of L-G fault current is the chief reason for the use of other grounding impedances.

A low-reactance (inductive) grounding connection impedance can be used to cause a moderate reduction in the L-G short-circuit current, particularly for the purpose of avoiding excess short-circuit current flow in a phase winding of a rotating machine or of accomplishing a desired distribution of neutral unbalanced load current among source machines. Using reactance to achieve the reduction in L-G fault current to below about 40% of the three-phase value enters the high-reactance region, which is subject to the generation of damaging transient overvoltages as a result of a ground fault condition, unless appropriate resistive damping circuits are added. Only one special case of high-reactance grounding is free of overvoltage trouble: the ground-fault neutralizer case, in which the reactance magnitude is carefully matched with the electric system L-G capacitance to create a natural frequency of oscillation almost exactly equal to power system operating frequency. It is critical, however, because a small deviation from the resonant value will destroy the overvoltage immunity. The need for impedance grounding, in general, has been somewhat reduced with the introduction of overcurrent protective devices with ratings of up to 300,000 AIR (ampere interrupting rating).

Resistance grounding. By substituting a resistive grounding impedance (a totally dissipative impedance), much greater reductions in the L-G fault current can be intentionally created without danger of transitory overvoltages.

The low-resistance region is characterized by an established level of available ground-fault current well below the three-phase fault value yet ample to properly operate protective devices responsive to ground-fault current flow. Typical current values in use range downward from a few thousand amperes. Present-day protective practices allow the current value to be set at 400 A for general purpose medium-voltage electric systems widely used in industry. The far more critical electric-shock hazards incident to electric power supply to portable excavating machinery have led to the selection of a much lower level of available ground-fault current, typically in the 25–50-A region.

Most electrical breakdowns occur as between line and ground, and many remain so throughout the interval of detection and isolation. A summary of the operating advantages achieved by intentional reduction of available ground-fault current is given below and illustrated in **Fig. 5**. (1) Low heat-energy release at the fault location because of the low current magnitude. (2) No noticeable dip in the system line-to-line voltages, which means no disturbance to the operation of all healthy load circuits. (3) Diminished interrupting duty on the circuit interrupter (low current and high power factor), which contributes to infrequent maintenance requirements. (4) Diminished duty imposed on the equipment grounding conductor network, which allows superior performance achievement at lower cost.

High-resistance grounding relates to a mode of operation in which the fault location and subsequent corrective action are undertaken manually by skilled maintenance personnel. It is used principally in electric systems that serve critical continuous-process machines.

The available ground-fault current is reduced to the same order as the electric system charging current to ground (generally less than 5 A). This resistive component of fault current is sufficient to arrest the generation of transient overvoltages by L-G

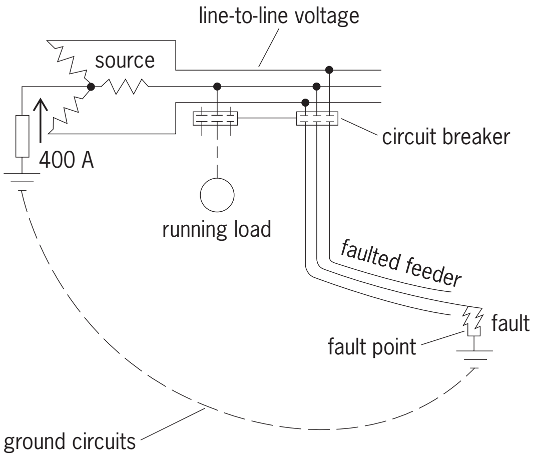


Fig. 5. Resistance grounding. (*General Electric Co.*)

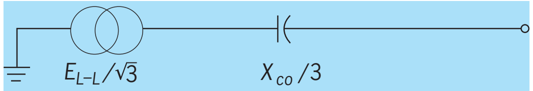


Fig. 6. Equivalent supply circuit to one phase conductor (relative to ground) of an "ungrounded" electric power system. $E_{L\text{-}L}$ = line-to-line voltage; $X_{CO}/3$ = capacitive reactance coupling to ground.

fault disturbances and provides a positive signal for identifying the presence of an L-G fault somewhere on the system. Successful results depend upon the presence of a skilled maintenance crew who can respond quickly to the ground-fault alarm, promptly locate the faulty circuit element, and take effective action to remove this faulty circuit from the system before a second insulation failure is induced.

Ungrounded system. Although a system may have no intentional grounding connection, and hence is named an ungrounded system, it is in fact unavoidably capacitively grounded, which is the reason that a fatal shock can be received when a person contacts an equipment enclosure and a phase conductor simultaneously. The layer of insulation surrounding every energized conductor metal constitutes the dielectric film of a minute distributed capacitor between power conductors and ground. In the aggregate this capacitance can amount to a substantial fraction of a microfarad for the complete metallically connected system. Surge voltage suppression filters typically include line-to-ground-connected capacitors which add to the distributed capacitors inherent in the system proper. Unless modified by the stabilizing qualities of grounding connections previously described, L-G fault disturbances can create dangerous overvoltage transients (impressed on system insulation) in about the same ways possible had the grounding connection impedance of Fig. 4 been a capacitor. The equivalent circuit of one phase conductor of an ungrounded three-phase power supply (relative to ground) takes the form illustrated in **Fig. 6**.

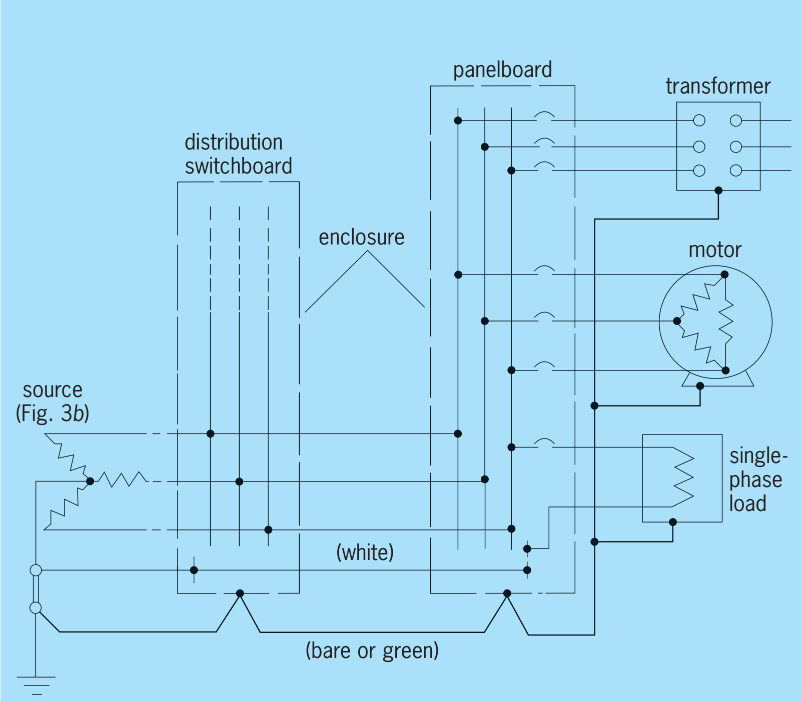


Fig. 7. Heavy line shows typical equipment grounding conductor.

Equipment (safety) grounding. At each unit of electrical equipment, safety grounding serves to establish a near-zero potential reference plane (even during L-G fault conditions). This reference extends to the outer reaches of the particular voltage-level electric system to which a solid equipment (safety) grounding connection can be made from the metal frames of served electric machines, the metal housings that contain switching equipment or other electric-system apparatus, and the metal enclosures containing energized power conductors; these enclosures may be metal cable sheaths or metal raceways.

The purposes of this interconnected mesh conducting network, drawn as the heavy lines in **Fig. 7**, are listed below.

1. To avoid electric shock hazard to any occupant of the area who may be making bodily contact with a metallic structure containing energized conductors, one of which has made an electric fault connection to the mentioned structure.

2. To provide an adequately low impedance to the return path of L-G fault current so as to cause automatic operation of the affected circuit's overcurrent protective device and not to interfere with the operation of system overcurrent protectors of unaffected circuits.

3. To provide ample conductivity (cross-sectional area) to carry the possible magnitude of ground-fault current for the duration controlled by the overcurrent protectors in the electric system.

4. To avoid, by installation, with appropriate geometric spacing (relative to phase conductors) dangerous amounts of ground-fault current diverted into paralleling conductive paths.

The fast-growing use of low-signal-level input high-speed electronic information technology systems places added emphasis on the need to minimize the transmission of stray electrical noise from the electric power circuits to the surrounding space in which these critical equipments are located. The presence of fast-acting solid-state switching devices among the electric-system switching components can aggravate the problem by intensifying the amount of high-frequency disturbance present with the electric-system voltage carried by the power conductors.

The complete metal enclosure of the electric power system, as shown in **Fig. 8**, contributes immensely to the elimination of electrical noise by reducing radiated interference. The NEC requires that all such metal enclosures be interconnected to form an adequate continuous electrical circuit, and also that they be grounded (with some exceptions). The effect of this construction is to enclose the entire electric power system conductor array within a continuous shell of grounded metal that functions

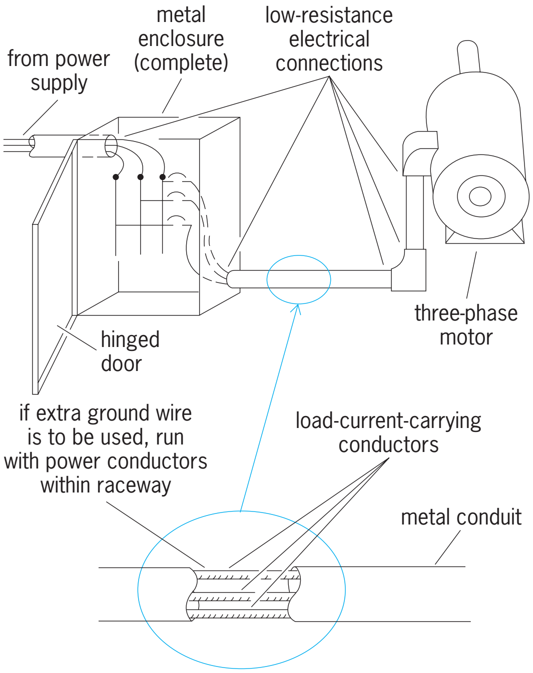

from power supply
metal enclosure (complete)
low-resistance electrical connections
hinged door
three-phase motor
if extra ground wire is to be used, run with power conductors within raceway
load-current-carrying conductors
metal conduit

Fig. 8. Complete enclosure of power conductors within a continuous grounded steel shell confines electric and magnetic effects of power currents.

as a Faraday shield to confine the electrostatic and electromagnetic fields associated with the power conductors to the space within the metallic shell. The contribution of electrical noise external to the enclosures is reduced to almost zero. *See* ELECTROMAGNETIC COMPATIBILITY.

To avoid a by-pass circuit by which power-conductor noise voltages might be conductively transmitted to the outside of the metallic enclosure, a careful check of the integrity of the insulation of the grounded power conductor (white wire) throughout the building interior is warranted. The NEC prescribes that the power-system grounded conductor be connected to the grounding conductor at the point of service entrance to the building, and at no other point within the building beyond the service equipment, with some exceptions (such as a separately derived system). Only if a supply feeder extends from one building into another is it permissible to reground the white wire, and then only at the point of entrance to the second building. *See* ELECTRICAL CODES; WIRING.

Protection considerations. The importance of limited magnitude L-G fault current in easing the problem of electric shock exposure control will be evident from the discussion below. The NEC contains a mandatory requirement for installation of automatic ground-fault-responsive tripping of the power supply at the service equipment for all solidly grounded wye-connected electric services of more than 150 V L-G but not those exceeding 600 V line to line (L-L), for each service, main building, or feeder disconnecting means rated 1000 A or more. In the interest of assuring superior electric shock protection, Section 210.8 of the NEC mandates the installation of ultrasensitive ground-fault-responsive tripping features for personal protection (type-GFCI personal protectors) on certain 120-V grounded circuits. Protection is accomplished by deenergizing the supply power to the receptacle if more than 5 milliamperes of the circuit current becomes diverted to a return path other than the grounded power conductor of the circuit.

Grounding of outdoor stations. Installations where earth is used as a reference ground plane present special problems. It is difficult to design an earth "floor surface" for an outdoor open-type substation that will be free of dangerous electric shock voltage exposure to persons around the station.

Personal electric shock danger. A beginning step in dealing with the shock hazard is to establish the magnitude of electric shock exposure which can be accepted by a person without ill effect. The evaluation of dangers from electric shock can be aided by the simple "person model" (**Fig. 9**). The organs which can be vitally affected are the heart and lungs. Lung muscles can be propelled to a tight closed spasm, shutting off the respiratory action. However, the lungs resume normal action once the severe shock voltage is removed. On the other hand, a single short-time excess-value electric shock incident may throw the heart muscles into a state of fibrillation from which they are unable to automatically recover normal action, resulting in death within a few minutes unless defibrillation is accomplished.

Electric shock is communicated to the body via electrical contact to the body extremities (Fig. 9), hand to hand, hand to foot, or foot to foot, each of which creates nearly the same shock intensity at the central chest area. The hand-to-foot case applies to a person standing on the substation floor and touching nearby conducting parts of the station with the extended arms; this applied shock hazard is named the E_{touch} exposure. The foot-to-foot case applies to a person walking across the substation floor during which activity the person's two feet, with considerable separation distance between, contact the substation surface; this shock hazard is named the E_{step} exposure. The effect of a surface layer of higher-resistance material over the active area of the station may be represented by an added resistance element R_{fc} (Fig. 9).

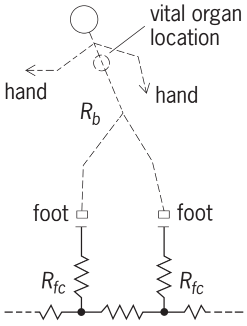


Fig. 9. A simple "person model" to be used in the evaluation of electric shock voltage-exposure intensity.

The basic limits of tolerable electric shock exposure are expressed by Eq. (1), where i is the current

$$i = \frac{0.116 \text{ ampere}}{\sqrt{t}} \qquad (1)$$

magnitude through the body and t the elapsed time in seconds, or, in similar form, by Eq. (2).

$$i^2 t = 0.0134 \text{ ampere}^2 \text{ seconds} \qquad (2)$$

The sources of potential electric shock hazard are commonly expressed in terms of volts (rather than amperes), but the limiting shock exposure tolerance can be converted to terms of voltage and time if the body resistance R_b (Fig. 9) is known. The value of R_b may go up into the megohm region in the absence of excess moisture or perspiration, but drops drastically when these substances are present on the skin. Assuming the lower limit of the body resistance to be 1000 ohms, the maximum allowable impressed body voltage V_s is given by Eq. (3).

$$V_s = i(R_b) = i(1000) \text{ volts} = \frac{116 \text{ volts}}{\sqrt{t}} \qquad (3)$$

Use of ground rods. The voltage exposure values of both E_{touch} and E_{step} depend heavily on the magnitude of voltage gradients along the surface of the outdoor station floor. An outdoor station might receive incoming power at 69 kV and an L-G circuit fault at this terminal would result in the injection of 3000 A into the grounding conductor at the station. During this current flow the station grounding conductor system could rise above mean earth potential by as much as 2500 V or more. The resulting potential distribution pattern across the station floor must be evaluated and tailored to achieve the limiting values of E_{touch} and E_{step}.

As a first step, one may consider driving a standard ground rod (8 ft or 2.4 m long, 0.75 in. or 19 mm in diameter) in the center of the substation floor area. In a typical soil makeup this might be found to establish a 25-ohm connection to earth. It takes only 100 A injected into such a connection to produce a voltage drop of 2500 V, about half of which takes place within a radius of about 2.5 ft (0.8 m). As a voltage probe is moved away from the rod on the station surface, the potential drops rapidly in a conical pyramidal fashion.

The earth-connection properties of a single driven ground rod are thus controlled, almost totally, by a small cylinder of earth, about 10 ft (3 m) in diameter, immediately surrounding it. The resistance value is usually substantial (on the order of 25 ohms) and is reduced only slightly by a second ground rod within the influence zone of the first. A low-value-resistance connection to earth therefore requires a multiplicity of distributed ground rods, spaced to be nearly independent of the influence fields of each other.

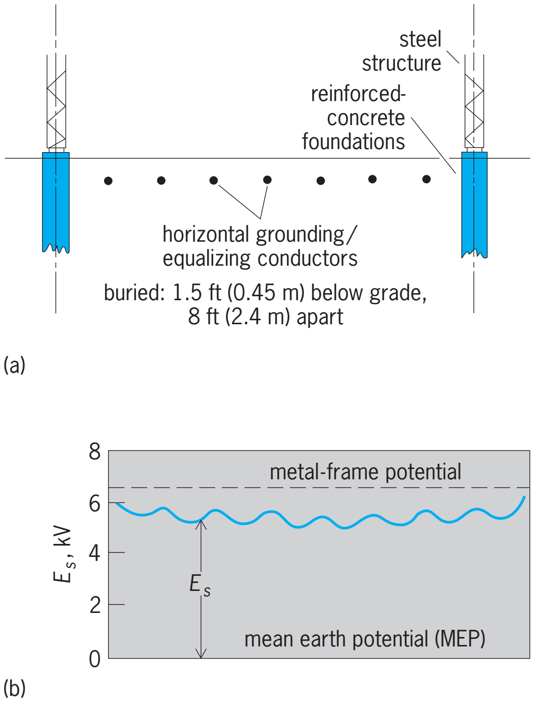

steel structure
reinforced-concrete foundations
horizontal grounding/ equalizing conductors
buried: 1.5 ft (0.45 m) below grade, 8 ft (2.4 m) apart
(a)
8
6
4
2
0
E_s, kV
metal-frame potential
E_s
mean earth potential (MEP)
(b)

Fig. 10. Surface-potential control in an earth-floor electrical station obtained by a mesh grid of buried grounding equalizing conductors. (*a*) Physical pattern of conductors in one bay. (*b*) Station surface potential E_s above mean earth potential during a local L-G fault.

To meet the electric shock safety requirements in open-type outdoor stations usually requires an elaborate array of metallic grounding conductors buried in the surface soil of the station area. The array is commonly composed of parallel horizontal conductors with perhaps 8 ft (2.4 m) horizontal spacing at a depth of some 1 to 2 ft (0.3 to 0.6 m) below the surface (**Fig. 10*a***). The geometry of this conductor system usually matches that of the station structure and makes use of the below-grade concrete reinforcing steel involved in the construction of the station. The ground-surface voltage, within one specific bay, relative to the station grounding conductor potential and above the mean earth potential, resembles that shown in Fig. 10*b*. An artificially controlled level of available ground-fault current at 400 A is common in industrial electric power system design, in contrast to a value of 3000 A, not uncommon elsewhere, representing a 7-to-1 advantage.

Hazards near station boundary. There is the possibility that unacceptable levels of electric shock voltage exposure E_{step} and E_{touch} may be found in some places external to the outdoor station area. To ensure the absence of electric shock danger to persons who may frequent these areas adjoining the substation, it is important to check out suspect spots and institute corrective measures as necessary.

R. H. Kaufmann; Brian J. M. Partland

Bibliography. *IEEE Guide for Safety in AC Substation Grounding*, IEEE Stand. 80–2000, 2000; *IEEE Recommended Practice for Electric Power Systems in Commercial Buildings*, IEEE Stand. 241–1990, 1990; *IEEE Recommended Practice for Grounding of Industrial and Commercial Power Systems*, IEEE Stand. 142–1991, 1991; National Fire Protection Association, *National Electrical Code*, 2005; R. P. O'Riley, *Electrical Grounding*, 6th ed., 2002.

Group theory

Any set of elements which is equipped with an operation (called multiplication) satisfying the requirements (1), (2), (3), and (4) is called a group. Group theory is the branch of mathematics devoted to the properties of groups. Group theory has applications in many branches of physical sciences, and in some branches of algebra and in analytic function theory.

Four requirements on group operation. The operation of multiplication is supposed to satisfy the following requirements.

Closure. The product of two elements, for example, g_1 and g_2, in the group G gives another element g_3 (called their product) in the group. That is, requirement (1) is satisfied, where $g \in G$ means that g is an element of G.

$$\text{For any } g_1 g_2 \in G, \qquad g_1g_2 = g_3, g_3 \in G \qquad (1)$$

Associative law. If g_1, g_2, and g_3 are any elements of the group, then Eq. (2) is satisfied; that is, in forming a

$$(g_1g_2)g_3 = g_1(g_2g_3) \qquad (2)$$

product of three elements the same result is obtained by first multiplying g_1 and g_2 and then multiplying the result and g_3, as first multiplying g_2 and g_3 and then multiplying g_1 and the result.

Existence of an identity element. There is in the group an element e (called the identity) which satisfies Eq. (3) for every element g of the group.

$$eg = ge = g \qquad (3)$$

Existence of inverses. For each element g of the group there is an element g^{-1} (called the inverse of g) which satisfies Eq. (4).

$$g^{-1}g = gg^{-1} = e \qquad (4)$$

Examples. An example of a group is the set of positive real numbers equipped with the operation of ordinary multiplication. The identity element is then the number 1 and the inverse of an element is its reciprocal. Another example of a group is the set of all real numbers equipped with the operation of ordinary addition. The identity element is then the number 0, and the inverse of an element is its negative.

Both these examples are commutative groups (also called abelian groups after the mathematician Niels Abel) because their multiplication law satisfies Eq. (5) for every g and h in the group.

$$gh = hg \qquad (5)$$

An example of a noncommutative group (also called nonabelian) is the group of rotations of a three-dimensional rigid body around a point. Here an element of the group is a rotation, that is, the act of rotating the body around a certain axis in space by a certain angle. The group operation means merely successive transformation; g_1g_2 stands for the act of rotation obtained by carrying out the acts of rotation g_2 and g_1 successively in that order. The identity element is the act of rotating through zero angle or making no rotation. The inverse of an element is the rotation around the same axis by the same angle but in the opposite sense. It is easy to check that if g_1 is a rotation around the vertical by 90° and g_2 a rotation around some horizontal axis by 90°, that $g_1g_2 \neq g_2g_1$, so the group is noncommutative.

Groups may have a finite or infinite number of elements. Furthermore, the elements may be discrete, or form a continuum in the topological sense described below. A simple example of a discrete finite group is the group with two elements, the identity e and another e_1 satisfying $e_1e_1 = e$. All other examples given above are of groups with an infinite number of elements, that is, infinite groups; the first two are examples of discrete infinite groups, while the third (the group of three-dimensional rotations) is an example of a continuous group.

Topological groups. For infinite groups, it often occurs that a group has a natural geometry. For example, the group of positive real numbers described above has the geometry of the real numbers. This geometry is compatible with the group operations in the sense that the product gh is a continuous function of g and h, and the inverse g^{-1} is a continuous function of g. Generalizing this scheme to arbitrary groups leads to the notion of a topological group (or less frequently a continuous group) which is a set of elements equipped not only with a group operation but also with a topology, the two notions being required to be compatible in the above sense. A topology can be specified in a number of ways; the net effect of any of them is to give meaning to the phrase: Two group elements are close to one another. For example, in the case of the rotation group discussed above, a topology can be defined by specifying that two rotations are close to one another when their axes have nearly the same direction and their angles of rotation differ by very little. *See* TOPOLOGY.

There is a particular class of topological groups which is of great importance. These are the so-called Lie groups (named after the mathematician Sophus Lie). These are topological groups in which it is possible to label the group elements by a finite number of coordinates in such a way that the coordinates of gh and g^{-1} are analytic functions of the coordinates of g and h and of g, respectively. That means that they should be representable as convergent power series in those coordinates. Most of the topological groups occurring in applications are Lie groups. One important example of a group which is not a Lie group (because its elements cannot even be labeled by a finite number of coordinates in such a way as to make the group operation continuous) is the group of all permissible coordinate transformations in the general theory of relativity. The significant transformations in the special theory of relativity, the so-called Lorentz transformations, do form a Lie group. *See* LIE GROUP; RELATIVITY. Arthur S. Wightman

Applications to physical sciences. Group theory, the mathematical theory of symmetry, has many applications in physics and allied areas of chemistry and materials science. Some major ways in which group theory is used in physics will be discussed in conceptual terms. For a more systematic survey of the physically important symmetry groups *see* SYMMETRY LAWS (PHYSICS)

Classification. A symmetry may be so powerful that its possible realizations can be classified. To accomplish such classifications is the task of the theory of transformation groups. An outstanding example is the classification of possible forms of crystals. In the nineteenth century N. Fedorov demonstrated that there are exactly 230 essentially different possible transformation groups of crystal lattices. The two-dimensional version of this problem (which is also of physical interest) can be interpreted as the problem of classifying the essentially different periodic patterns of planar ornaments, and there are exactly 17 such patterns. *See* CRYSTAL STRUCTURE; CRYSTALLOGRAPHY.

Reduction of variables. If a physical system is invariant under some group of transformations, then its behavior generally depends upon fewer variables than might appear in general. For instance, if the potential energy due to the interaction between two particles depends on their positions, then in general it depends upon six variables, namely three coordinates for the position of each particle. However, if this energy does not depend upon the absolute position of these particles in space, but only on their relative position—that is, it is invariant under the group of spatial translations—then this energy can depend only on three parameters, namely the three coordinates of the relative position or, equivalently, the three components of the vector connecting the particles. Further, this energy does not depend on the absolute orientation of the particles—that is, it is invariant under the group of spatial rotations—then it can depend on only one parameter, namely the distance between the particles. Thus, exploiting symmetry can drastically simplify the description of the problem.

In this simple example, the precise reduction in complexity entailed by the symmetries can be inferred rather easily. In more complex cases, it becomes all but essential to use systematic mathematical procedures to carry out similar reductions. The general idea is that physical behavior can depend only on quantities that are invariant under the symmetry transformations; and the relevant branch of group theory, which provides systematic procedures for identifying such quantities, is called invariant theory.

Spectroscopy. A major problem in many branches of quantum mechanics is to find the energy levels of the system under study. Group theory is invaluable in predicting such energy levels. For if the fundamental laws describing a physical system have a certain symmetry, then states related by symmetry operations will all have the same energy; and if the fundamental laws are only approximately symmetric, or if some external condition breaks the symmetry, then these states will have nearly but not quite the same energy. The transitions between a cluster of states having nearly but not quite the same energy, and some other given state, will then give rise to a cluster of closely spaced spectral lines. The laws of group theory make it possible to anticipate the patterns of the closely spaced clusters of spectral lines that result. *See* ENERGY LEVEL (QUANTUM MECHANICS); NONRELATIVISTIC QUANTUM THEORY.

For example, the fundamental laws governing a hydrogen atom are very nearly invariant under transformations that rotate the electron's spatial position, while leaving the direction of its spin unchanged. However, a more accurate treatment shows that there are a number of relatively small effects which are not invariant under this transformation. Spectral features that at low resolution appear to be single lines will therefore, when examined more closely, exhibit fine structure, whose details depend in very direct and characteristic fashion upon the symmetries of the states involved. *See* ATOMIC STRUCTURE AND SPECTRA; FINE STRUCTURE (SPECTRAL LINES).

Another important situation that frequently occurs is that a transition between two specific energy levels is forbidden because they differ in symmetry. Such selection rules were very important historically in establishing the laws of quantum mechanics and testing their validity. In modern applications the procedure is often reversed, and the symmetry of the states is inferred from the observed selection rules. This information, in turn, is invaluable in forming models of the structure and dynamics of the system under consideration. *See* SELECTION RULES (PHYSICS).

Important examples of this procedure abound. In elementary particle physics, the systems of interest were, by a complicated set of indirect inferences, proposed to be composites of three quarks (for baryons) or a quark and an antiquark (for mesons). For these systems, states of definite energy appear as short-lived particles, or resonances, in high-energy collisions. The quark hypothesis, augmented with group-theoretic arguments, enabled deductions of the kind described above for the number of closely spaced resonances and the selection rules for transitions among them. Experiments that confirmed these deductions gave great support to the quark idea. *See* BARYON; ELEMENTARY PARTICLE; MESON; QUARKS.

Similarly, a definite model of the spatial structure of a molecule allows predictions of the possible energy levels of electrons in that molecule, which can be compared with observed spectra. Again, the shell model allows prediction of the possible energies of systems consisting of a given number of protons and neutrons, that is, an atomic nucleus. *See* MOLECULAR STRUCTURE AND SPECTRA; NUCLEAR STRUCTURE.

In applications of group theory to spectroscopy (in the broad sense used above), the mathematical theory of unitary representations plays a central role. *See* SPECTROSCOPY; UNITARY SYMMETRY.

Guide to new laws. In modern physics, symmetry, or group theory, is increasingly used as a guide to formulating new laws. For example, the discovery that the weak interaction is not invariant under spatial inversion or parity (*P*) was an important clue for formulating the proper theory of this interaction. The operation of combined parity (*CP*), a transformation involving both changing particles into their antiparticles and spatial inversion, is an approximate symmetry. However, some tiny violations of *CP* have also been observed, whose inner meaning has not been completely elucidated. *See* PARITY (QUANTUM MECHANICS); WEAK NUCLEAR INTERACTIONS.

A more abstract and wide-ranging symmetry postulate, gauge invariance, plays a central role in the modern theory of fundamental interactions. Roughly speaking, gauge invariance postulates not only an overall symmetry among different entities, but that the symmetry transformations can be made independently at every point of space-time. When applied to the color symmetry of quarks, this powerful hypothesis leads directly to quantum chromodynamics, which is presently accepted as the fundamental theory of the strong interaction. *See* GAUGE THEORY; QUANTUM CHROMODYNAMICS.

A new form of group, a supergroup, allows more kinds of symmetry transformations than were previously considered. Specifically, supersymmetry makes it possible to transform particles of different spin into one another. Demanding supersymmetry leads to models of elementary particle interactions that are especially attractive in several ways. At present there is no direct evidence that these models are correct, but supersymmetry provides a striking example of how group theory can be used as a guide to suggesting possible new physical laws. *See* SUPERSYMMETRY. Frank Wilczek

Applications within mathematics. The applications of group theory within mathematics itself are numerous and important. The part of mathematics in which the notion of group was first clearly isolated was the theory of algebraic equations.

Algebra. Here, the basic idea is to associate a group (the so-called Galois group, named after the mathematician E. Galois) with algebraic equation (6) in

$$a_0 + a_1x + \cdots + a_nx^n = 0 \qquad (6)$$

such a way that the structure of the group reflects the type of operations necessary to compute the roots from the coefficients $a_0, \ldots, a_n$. For example, this method is used to obtain a simple proof that for n greater than 4, no solution in roots of the general equation exists. *See* ALGEBRA; EQUATIONS, THEORY OF.

Topology. In topology, groups are used to classify the structure of various geometric objects. The homology and homotopy groups serve this purpose for a manifold. As an example consider the homotopy group. To define it, a point *P* of the manifold is chosen and all continuous closed paths starting and ending at *P* are considered. A notion of product of such paths is defined: If x_1 and x_2 are two paths, the product x_1x_2 is the closed path which is obtained by tracing out x_2 and then x_1. The inverse of a path is then the same curve traversed in the opposite direction and the identity path consists of the point *P* alone. Paths are divided into equivalence classes, two paths lying in the same equivalence class if one can be deformed continuously into the other. The notion of product then extends to equivalence classes, the product of two equivalence classes being obtained as the equivalence class of the product of any two representative paths. With this definition of product, the set of equivalence classes of continuous closed paths starting at *P* is a group, the (one-dimensional) homotopy group (also called fundamental group). Provided the manifold under discussion is connected (that is, provided any two points of the manifold can be connected by a continuous curve lying in the manifold), the fundamental group is essentially the same, independent of which point is taken as the starting point for the paths. If the fundamental group of the manifold consists of one element only, the manifold is said to be simply connected. *See* MANIFOLD (MATHEMATICS).

Analytic functions. Group theory plays a fundamental role in the theory of analytic functions. This is true not only for the thoroughly studied case of one complex variable, but also for the case of several complex variables, which is less well understood. Only the former will be discussed here. The starting point is the fact that each analytic function has a natural domain of definition, its Riemann surface. There is another Riemann surface associated with it, its universal covering surface, which is simply connected. The universal covering surface may have several sheets for each sheet of its underlying Riemann surface, each point of the latter being replaced by several "lying above" it. Any function which is analytic on the underlying Riemann surface can be regarded as analytic on the universal covering surface; it will take the same values at those points which lie above a given point. The analytic function regarded as defined on the universal covering surface is invariant under transformations which carry the points lying above any point of the Riemann surface into themselves. These transformations form a group. Now a second basic fact is that any simply connected Riemann surface can be mapped in a one-to-one and analytic fashion onto one of three regions: the interior of the unit circle, the complex plane, or the complex plane closed by adding a point at infinity. Thus, an analytic function on an arbitrary Riemann surface can be regarded as an analytic function defined in one of these three regions and invariant under a certain group of transformations. Such functions are called automorphic functions. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES; RIEMANN SURFACE.

Structure of groups. If G is a group and a subset G' of its elements is a group under the same law of multiplication, then G' is called a subgroup of G. Especially significant are the invariant subgroups which have the property that if g is any element of G' and h any element of G, then hgh^{-1} is an element of G'. Let G and H be two groups and suppose that f is a mapping of G into H [that is, a function which for each element, g, of G yields an element $f(g)$ in H]

with the property $f(gh) = f(g)f(h)$. Then f is called a homomorphism of G into H and G is homomorphic to H. The subset of G consisting of those elements which f maps onto the identity in H form an invariant subgroup, the kernel of the homomorphism f. If $H = G$ so that f maps G into itself, it is called an endomorphism. If f maps G one-to-one onto H, then it is called an isomorphism. Finally, if f is an isomorphism of G onto G, it is called an automorphism of G. The idea of isomorphism gives precise meaning to the vague notion that two groups have the same structure.

An important concept, also used in applications of group theory to the physical sciences, is that of a representation of a group. This is a correspondence between group elements g and linear operators $M(g)$ satisfying Eq. (7). Evidently, a representation of a

$$M(g_1)M(g_2) = M(g_1g_2) \qquad (7)$$

group is a homomorphism of the group into the group of linear transformations of a vector space.

For topological groups it is natural to require of a homomorphism that it preserve not only the group structure but also the topological structure; for a topological group, a homomorphism is defined as a continuous mapping, f, of G into H satisfying $f(gh) = f(g)f(h)$. The group structure and the topological structure of a topological group impose severe restrictions on each other. Evidence for this statement is the solution of Hilbert's fifth problem (posed in 1900 and solved about 50 years later), which may be stated roughly: Every topological group whose elements can be labeled by a finite number of coordinates so that the group operations are continuous is isomorphic to a Lie group.

For some topological groups it is possible to define a notion of integration on the group which has invariance properties under group multiplication. Namely, if F is a complex-valued function defined on the group and $\int F(g)\, d\mu(g)$ is the integral of F over the group with respect to a measure $d\mu(g)$, then $d\mu(g)$ is said to be left-invariant if Eq. (8) holds for all in-

$$\int F(g)\, d\mu(g) = \int F(hg)\, d\mu(g) \qquad (8)$$

tegrable F and all h in the group. Analogously, right-invariance is defined by Eq. (9). When a left-invariant

$$\int F(g)\, d\mu(g) = \int F(gh)\, d\mu(g) \qquad (9)$$

measure exists so does a right-invariant, and they are unique up to a multiplicative factor. For a compact group, that is, one for which $\int d\mu(g) < \infty$, left- and right-invariant measures coincide. *See* MEASURE THEORY.

Invariant integration is a powerful tool in the study of representations of groups. Consider, for example, two irreducible representations of a compact group, G, by unitary matrices. (A representation is irreducible if there is no proper subspace of vectors carried into itself by all matrices of the representation. It is unitary if all the matrices are unitary.) Then Eq. (10) holds, where the first alternative holds if

$$\int M^{(1)}(g)_{\kappa\lambda}\overline{M^{(2)}(g)}_{\mu\nu}\, d\mu(g) = \left\{\begin{matrix}1\\0\end{matrix}\right\} \frac{\delta_{\kappa\mu}\delta_{\lambda\nu}\int d\mu(g)}{\dim M^{(1)}(g)} \qquad (10)$$

$M^{(1)} = M^{(2)}$ and the second if they are not equivalent. $M^{(1)}$ and $M^{(2)}$ are equivalent if there exists a unitary operator U such that $M^{(1)}(g) = UM^{(2)}(g)U^{-1}$ for all g in G; dim $M^{(1)}(g)$ stands for the number of rows (or columns) in $M^{(1)}(g)$. Angular momentum selection rules of quantum-mechanical systems are consequences of Eq. (10) in the case that G is the three-dimensional rotation group. *See* ANGULAR MOMENTUM; MATRIX THEORY.

For Lie groups there is an important method of analysis, the so-called infinitesimal method. Consider, for example, a Lie group of matrices with matrix multiplication as group multiplication. (Not all Lie groups are isomorphic to such groups, but many are.) A one-parameter subgroup is a subset of elements, $g(t)$, labeled continuously by a real parameter, t, and satisfying Eq. (11) for all real t_1 and t_2. Such a subgroup can be written in the form of Eq. (12).

$$g(t_1)g(t_2) = g(t_1 + t_2) \qquad (11)$$

$$g(t) = \exp tx \qquad (12)$$

The matrix x is called the infinitesimal element of the one-parameter subgroup. If the group elements are labeled by n parameters, n distinct one-parameter subgroups can be chosen with linearly independent infinitesimal elements $x_1, \ldots, x_n$. These matrices will then satisfy commutation relations (13).

$$x_jx_k - x_kx_j = \sum_{l=1}^{n} C^l_{jk}x_l \qquad (13)$$

The constants C^l_{kj} are called the structure constants of the Lie group, and they largely determine the structure of the group. *See* GRAPH THEORY; RING THEORY; SET THEORY.

Arthur S. Wightman

Bibliography. D. Bailin and A. Love, *Supersymmetric Gauge Field Theory and String Theory*, 1994; T. P. Cheng and L.-F. Li, *Gauge Theory of Elementary Particle Physics*, 1988; D. Marris and M. Bertolucci, *Symmetry and Spectroscopy*, 1989; D. J. Robinson, *A Course in the Theory of Groups*, 2d ed., 1996; J. Rotman, *An Introduction to the Theory of Groups*, 4th ed., 1995; D. H. Sattinger and O. L. Weaver, *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*, 1993; H. Weyl, *Symmetry*, 1952.

Group velocity

The velocity of propagation of a group of waves forming a wave packet; also, the velocity of energy flow in a traveling wave or wave packet. The pure sine waves used to define phase velocity v_p do not ever really exist, for they would require infinite extent. What do exist are groups of waves, wave packets, which are combined disturbances of a group of sine waves having a range of frequencies and wavelengths. Good approximations to pure sine waves exist, provided the extent of the media is very large in comparison with the wavelength of the sine wave. In nondispersive media, pure sine waves of different frequencies all travel at the same speed υ_f, and any wave packet retains its shape as it propagates. In this case, the group velocity υ_g is the same as υ_p. But if there is dispersion, the wave packet changes shape as it moves, because each different frequency which makes up the packet moves with a different phase velocity. If υ_p is frequency-dependent, then υ_g is not equal to υ_p. *See* PHASE VELOCITY; SINE WAVE; WAVE MOTION.

The relationship between υ_g and υ_p is very easily derived in one dimension. Consider a wave packet made of two waves: the first has wavelength λ and phase velocity υ_p; the second has wavelength $\lambda + \Delta\lambda$ and phase velocity $\upsilon_p + \Delta\upsilon_p$. The wave packet is the combined disturbance which is just the sum of the two waves taken over their infinite extent. The "position" of the wave packet may be taken as the position of maximum disturbance which occurs at some point x_0 at time $t = 0$, as shown in **illus.** ***a*** at the position of the crests labeled 2. For clarity, the waves are drawn separated. After some time t_0, crest 2 of the second wave will have moved ahead of crest 2 of the first wave, and the crests labeled 1 will be aligned as shown in illus. *b*. The position of the wave packet has moved forward a distance D in time t_0, so the group velocity is $\upsilon_g = D/t_0$. But $D = \upsilon_p t_0 - \lambda$ is evident from the figure, as is $\lambda + \Delta\lambda - \lambda = \Delta\lambda = (\upsilon_p + \Delta\upsilon_p)t_0 - \upsilon_p t_0 = \Delta\upsilon_p t_0$. Therefore, $t_0 = \Delta\lambda_p/\Delta\upsilon_p$, and upon combining these one has $\upsilon_g = \upsilon_p - \lambda(\Delta\upsilon/\Delta\lambda)$. A calculus derivation would yield Eq. (1), which is usually given as the basic relationship.

$$\upsilon_g = \upsilon_p - \lambda\,\frac{d\upsilon_p}{d\lambda} \qquad (1)$$

A more convenient form utilizes the equation $\upsilon_p = \lambda f$, where f is the frequency associated with wavelength λ. By the chain rule of differentiation, $d\upsilon_p/d\lambda = f + \lambda(df/d\lambda)$ so $\upsilon_g = \upsilon_p - \lambda f - \lambda^2(df/d\lambda) = -\lambda^2(df/d\lambda)$. But then one uses the wave number $k = 2\pi/\lambda$ to finally write Eq. (2), where $\omega = 2\pi f$ is the

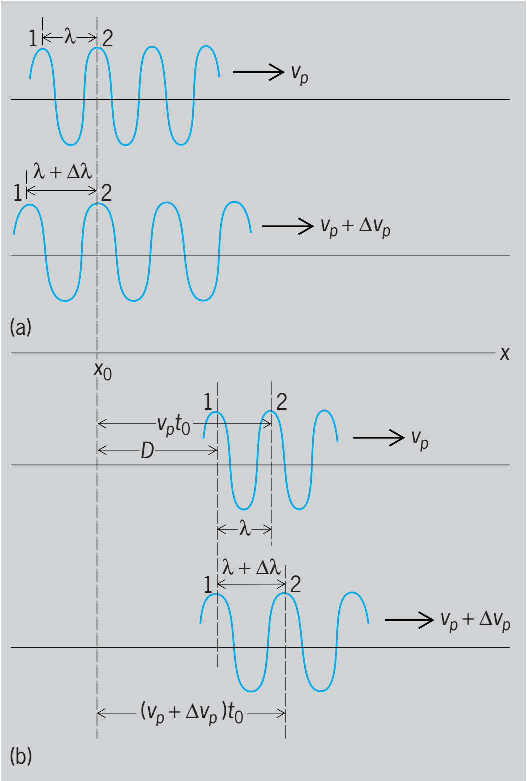
$$\upsilon_g = \frac{d\omega}{dk} \qquad (2)$$

angular frequency and $\omega/k = \upsilon_p$. This last form is readily generalized to more than one dimension.

For waves in deep water, $\upsilon_p = ck^{-1/2}$, where c is a constant. Thus $\omega = ck^{1/2}$ and $\upsilon_g = d\omega/dk = {}^1\!/_2\,\upsilon_p$. These waves are very strongly dispersive, and the individual wave crests slide through the group at twice the group speed. *See* LIGHT; SOUND; WAVE EQUATION; SURFACE WAVES.

S. A. Williams

Bibliography. A. P. French, *Vibrations and Waves*, 1977; K. U. Ingard, *Fundamentals of Waves and Oscillations*, 1988; H. J. Pain, *Physics of Vibrations and Waves*, 5th ed., 1999.



One-dimensional wave packet made of two waves. (*a*) Maximum disturbance of wave packet at time t_0 is located at x_0, where crests labeled 2 coincide. (*b*) After time t_0, wave packet has moved distance *D* to a position at which crests labeled 1 now coincide.

Grout

A binding or structural agent used in construction and engineering applications. Grout is typically a mixture of hydraulic cement and water, with or without fine aggregate; however, chemical grouts are also produced.

The type most commonly specified in construction and engineering is cementitious grout, which is used where its more conventional sister material, concrete, is less suited because of placing limitations or restrictions on coarse-aggregate contents.

Grout can be formulated from a variety of cements and minerals and proportioned for specific applications. Neat cement grout refers to formulations without aggregate, containing only hydraulic cement, water, and possibly admixtures. Sanded grout is any mix containing fine aggregate and it is formulated much like masonry mortar. Whether neat or sanded, cementitious grouts derive their strength and other properties from the same calcium silicate-based binding chemistry as concrete.

Through manual application or pumping, cementitious grouts are used to fill voids and cracks in pavements, building and dam foundations, and brick and concrete masonry wall assemblies; to construct floor toppings or provide flooring underlayment; to place ceramic tile, and to bind preplaced-aggregate concrete. Pressure grouting is employed in raising or leveling uneven slabs, foundations, or structures.

Grouting practice also includes design and placement of flowable fill, a product used for trenches and backfilling. Grouts mixed with special expansive cements are employed for demolition as alternatives to techniques relying on explosive devices or heavy equipment.

Cementitious grouts are either batched in hoppers on site or delivered in truck mixers. Specialty grouts for small-scale applications are prepackaged. Grouts are formulated from either ASTM type I portland cement or other hydraulic cements, depending upon project conditions. Mineral and chemical admixtures, which impart special properties to concrete mixes, are similarly specified for grouts. Mixes are designed for requirements or factors such as minimum compressive strength, placing conditions and consistency, workability and working time, and setting and hardening. Physical and structural properties of the hardened product are predicated on water-cement ratio, cement particle hydration, and the subsequent bonding with fine aggregate or surfaces in contact with grout. *See* CEMENT; CONCRETE.

Don Marsh

Bibliography. S. H. Kosmatka, *Cementitious Grouts and Grouting*, Portland Cement Association, 1990.

Growth factor

Any of a group of biologically active polypeptides which function as hormonelike regulatory signals, controlling the growth and differentiation of responsive cells. Indeed, the distinction between growth factors and hormones is frequently arbitrary and stems more from the manner of their discovery than from a clear difference in function. *See* CELL DIFFERENTIATION; HORMONE.

The discovery of growth factors has been primarily a direct result of the development of techniques for animal-cell culture. Modern tissue-culture techniques permit the removal of a variety of cell types from animals and the growing of the cells for indefinite periods in artificial growth media. However, supplying cells with the nutritional essentials (such as glucose, vitamins, minerals, and amino acids) is not sufficient for growth; cells also require a number of hormones and growth factors. The particular hormone–growth factor combination that is most favorable differs for various cell types. *See* TISSUE CULTURE.

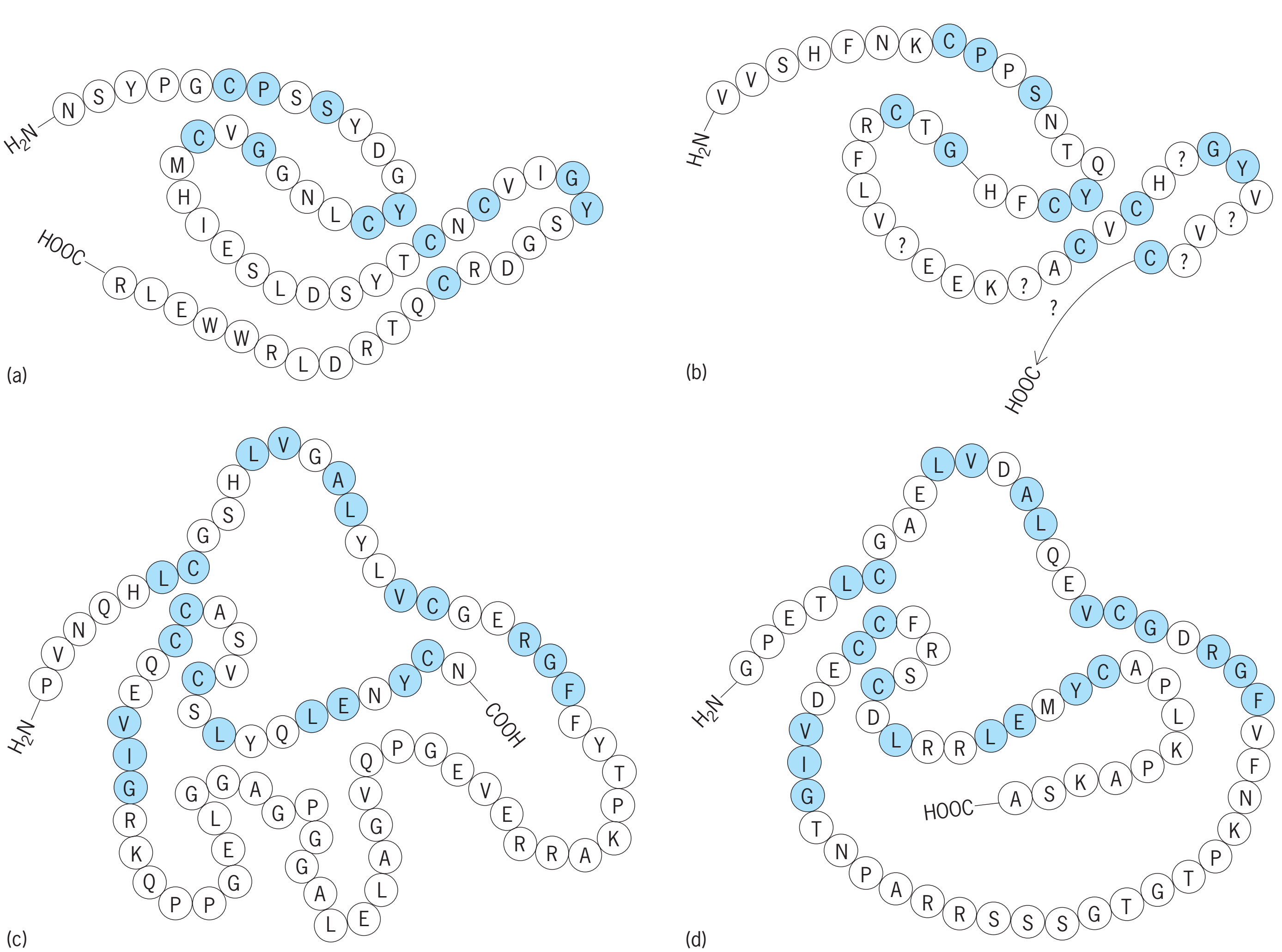
Growth factors are first identified merely as growth-promoting agents; subsequently, they are identified as highly purified discrete molecular species. Generally, growth factors discovered in this way have been polypeptides. Like polypeptide hormones, the action of growth factors on cells depends on binding to specific cell-surface receptors (proteins).

The physiological functions of growth factors can be inferred from their activities in tissue cultures and probably involve the regulation of cell proliferation and differentiation during embryonal development and during wound healing. However, in many cases there has been no direct demonstration of the physiological importance of particular growth factors. Some of the most fully characterized growth factors are described below.

Families. The sequence of amino acids has been determined for several growth-factor polypeptides (see **illus.**). This information permits a number of growth factors to be placed into families, members of which have related amino acid sequences, suggesting that they evolved from a single ancestral protein. The insulin family comprises somatemedins A and C, insulin, insulinlike growth factor (IGF), and multiplication-stimulating factor (MSF). A second family consists of sarcoma growth factor (SGF), transforming growth factors (TGFs), and epidermal growth factor (EGF). In addition, there are growth factors, such as nerve growth factor (NGF), fibroblast growth factor (FGF), and platelet-derived growth factor (PDGF), for which structural homologs have not been identified. *See* INSULIN; PROTEIN.

Nerve growth factor. Nerve growth factor is produced in large amounts in mouse submaxillary gland and in the prostate gland of several mammals; it is probably produced in smaller amounts in other tissues. Nerve growth factor appears to function in promoting survival of, and extension of, axons from nerve cells of the peripheral nervous system. For example, injection of nerve growth factor into newborn mice causes excessive axon growth from sympathetic and sensory nerve cells, while injection of antiserum to nerve growth factor causes these nerve cells to die. The mouse-gland protein which has been most fully characterized exists as a 136,000-molecular-weight aggregate of α, β, and γ subunits. The β subunit, which consists of two noncovalently associated 13,000-molecular-weight peptide chains, contains the full nerve growth factor activity, while the γ subunit is a proteolytic enzyme which is closely related to glandular kallikrein. *See* NERVOUS SYSTEM (VERTEBRATE).

Epidermal and transforming growth factors. Epidermal growth factor, like nerve growth factor, is produced in large amounts in the submaxillary gland of male mice (a fact of uncertain physiological consequence) but is apparently produced by other tissues also. Submaxillary-gland epidermal growth factor, like nerve growth factor, is found in a high-molecular-weight aggregate. The 60,000-molecular-weight epidermal growth factor aggregate contains the 6000-molecular-weight biologically active subunit, as well as a kallikrein-like proteolytic enzyme which is very similar to, but distinct from, the γ subunit of nerve growth factor. Both epidermal growth factor and nerve growth factor are produced by proteolytic cleavage of large precursor proteins, and it appears that the kallikrein-like enzymes perform this function. Epidermal growth factor injected into newborn mice causes premature opening of eyelids and eruption of incisors, suggesting that epidermal growth factor normally functions in promoting growth of epithelial tissue. Transforming growth factors are a family of factors with similar biological activity. Several transforming growth factors have been shown



Comparison of the amino acid sequences of (*a*) mouse epidermal growth factor, (*b*) rat transforming growth factor, (*c*) bovine proinsulin, and (*d*) human insulinlike growth factor I. Shaded circles depict positions of structural homology between epidermal growth factor and transforming growth factor and between proinsulin and insulinlike growth factor. Amino acid abbreviations are: A, alanine; C, cysteine; D, aspartate; E, glutamate; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, trytophan; Y, tyrosine.

to be structurally related to epidermal growth factor (illus. *a* and *b*). Transforming growth factors are secreted by a variety of cultured cells, particularly tumor cells, and have the property of causing normal fibroblasts to behave like tumor cells. Consistent with their structural similarity to epidermal growth factor, some transforming growth factors will interact with the epidermal growth factor cell-surface receptor.

Somatemedin growth factors. Insulinlike growth factors I and II are distinct but similar, differing only in a few amino acids. Somatemedins A and C and multiplication-stimulating factor are themselves quite similar to the insulinlike growth factors. They share a number of biological activities, including stimulation of sulfation of cartilage by chondrocytes, stimulation of proliferation of fibroblasts, and insulinlike stimulation of glucose metabolism in fat cells. Insulin is a more distantly related member of this family (illus. *c*). Nevertheless, the relationship is close enough to permit insulin to bind to somatemedin receptors and, at high concentrations, to stimulate proliferation of fibroblasts.

Platelet-derived growth factor. This growth factor is stored in blood platelets and is released during blood clotting. Since platelet-derived growth factor stimulates fibroblast cell proliferation tissue cultures, it is likely that release of platelet-derived growth factor at wound sites may be important for stimulation of wound healing. Fibroblast growth factor has not been so fully characterized as platelet-derived growth factor, but while they share many similarities, it appears to be a distinct protein.

A broader family. While there are no readily apparent structural similarities among epidermal growth factor, platelet-derived growth factor, and insulin, the cell-surface receptor proteins for these growth factors do have remarkable similarities, which suggests that epidermal growth factor, insulin, and platelet-derived growth factor families may themselves be related. The similarity of the receptors lies in the fact that each is a protein kinase (an enzyme which phosphorylates proteins), and the preferred substrate of the receptor kinase activity is the receptor itself. The receptor protein kinase activities are unusual in preferring tyrosine residues in proteins as substrates (most known kinases prefer serine or threonine). The only other tyrosine-specific protein kinases known are tumor-causing proteins produced by ribonucleic acid (RNA) tumor viruses. *See* TUMOR VIRUSES.

Growth factors and cancer. The stimulation of cell proliferation by several growth factors is similar in some ways to the rapid cell proliferation characteristic of tumor cells. Furthermore, the growth factor receptors are similar to the tumor-causing proteins produced by several RNA tumor viruses. It has been demonstrated that platelet-derived growth factor is virtually identical to the tumor-causing protein of the RNA tumor virus, simian sarcoma virus. Thus, what has long been a subject of speculation can no longer be doubted: some forms of cancer involve improper function of growth factors. *See* CANCER (MEDICINE); ONCOLOGY.

Mark Bothwell

Bibliography. H. N. Antoniades and A. J. Owen, Growth factors and regulation of cell growth, *Annu. Rev. Med.*, 33:445–463, 1982; G. Carpenter and S. Cohen, Epidermal growth factor, *Annu. Rev. Biochem.*, 48:193–216, 1979; D. Gospodarowicz and J. S. Moran, Growth factors in mammalian cell culture, *Annu. Rev. Biochem.*, 45:531–558, 1976; L. A. Greene and E. M. Shooter, The nerve growth factor, *Annu. Rev. Neurosci.*, 3:353–402, 1980.

Gruiformes

A highly heterogeneous, worldwide order of field, marsh, and aquatic birds that may be closely related to the shorebirds and their allies (see **illustration**). Diagnosis of this order is difficult because of its great diversity; some taxonomists divide the gruiforms into a number of separate and possibly unrelated orders. Some families, such as the Turnicidae and the Pedionomidae, are considered by some workers to be members of the Charadriiformes. Many of the families contain a small number of species and live in restricted areas. *See* CHARADRIIFORMES.

American coot (*Fulica americana*). (*Photo by Antonio J. Ferreira, © California Academy of Sciences*)

Classification. The order Gruiformes is arranged into 10 suborders and 20 families, as listed below.

Order Gruiformes
Suborder Mesoenatides
Family Mesoenatidae (mesites, 3 species; Madagascar)
Suborder Turnices
Family: Turnicidae (button quail, 14 species; Old World from Europe to Australia)
Pedionomidae (plains wanderer; 1 species; Australia)
Suborder Grues
Family: Geranoididae (fossil; Eocene of North America)
Eogruidae (fossil; Eocent to Oligocene of Mongolia)
Ergilornithidae (fossil; Oligocene to Miocene of Asia)
Gruidae (canes; 15 species; worldwide, except South America)
Aramidae (limpkin; 1 species; New World tropics)
Psophiidae (trumpeters; 3 species; South America)
Suborder Ralli
Family Rallidae (rails, gallinules, coots; 142 species; worldwide, including numerous oceanic islands)
Suborder Heliornithes
Family Heliornithidae (sun grebes; 3 species; pantropics)
Suborder Apterornithes
Family Apterornithidae (fossil; Quaternary of New Zealand)
Suborder Rhtnocheti
Family Rhynochetidae (kagu; 1 species; New Caledonia)
Suborder Eurypygae
Family Eurypygidae (sun bittern; 1 species; New World tropics)
Suborder Cariamae
Family Cunampaiidae (fossil; Eocene of Argentina)
Cariamidae (seriemas; 2 species; South America)
Bathornithidae (fossil; Oligocene of the United States)
Idiornithidae (fossil; Eocene and Oligocene of France)
Phorusrhacidae (fossil; Eocene to late Pliocene of Europe to South America, with the greatest radiation from the Oligocene to the late Pliocene of South America)
Suborder Otides
Family Otididae (bustards; 24 species: Old World, including Australia)

Fossil record. The fossil record of some groups within the Gruiformes is well known from the early

Eocene, including the cranes, rails, and seriemas. Other living families have either no fossil record or a poor one. The Apterornithidae of New Zealand are large, flightless birds only distantly related to any other gruiform group. Rails are common as recent fossils on many oceanic islands; many are flightless, as are some living island species. Cranes are well represented as fossils from the early Eocene of the Northern Hemisphere. Some of the Ergilornithidae possess only two toes, which led a few workers to conclude that this family is ancestral to the ostriches, but that idea has been almost universally rejected. *See* STRUTHIONIFORMES.

The most fascinating group of gruiform fossils is found in the Cariamae and includes the large radiation of the giant, flightless, predatory Phorusrhacidae and their relatives. It was centered in South America from early Oligocene to late Pliocene times, but with a late Pliocene form found in Florida and an Eo-Oligocene species discovered in France. The North American Bathornithidae and the European Idiornithidae, known from the late Eocene to the early Miocene, represent a separate northern radiation of the Cariamae. All that remains of these large, diverse radiations are the two living species of seriemas of Brazil and Argentina.

Characteristics. The Gruiformes range from small to very large. Some have short legs and others long, and the bill varies from short and straight to long and decurved. Some have long wings and are strong fliers, whereas others are flightless. Wading, terrestrial, or aquatic, they are usually found in open country, but some inhabit dense forests. The sexes in most species are similar in plumage, and most species are monogamous. The downy young usually leave the nest shortly after hatching and are cared for by both parents. Northern species are migratory; the migratory concentrations and flights of cranes are spectacular. Some forms, including the kagu and a number of rail species, are flightless. Many gruiforms are secretive and fly only when other forms of escape are impossible. Although rails are reluctant to fly, many species undertake long migrations and have colonized many isolated oceanic islands, where they tend to evolve flightlessness.

Despite the fact that cranes are revered as symbols of good fortune and luck by many cultures, most species have been seriously reduced in numbers, mainly by habitat destruction but also by hunting. Most of their 15 species are seriously endangered and are barely maintaining their low numbers despite rigorous conservation efforts. Several species of larger bustards are equally endangered because of habitat loss and high-power lines which are hazardous to the flying birds. Many island rails have become extinct primarily as a result of the introduction of small mammalian predators. *See* AVES; ENDANGERED SPECIES.

Walter J. Bock

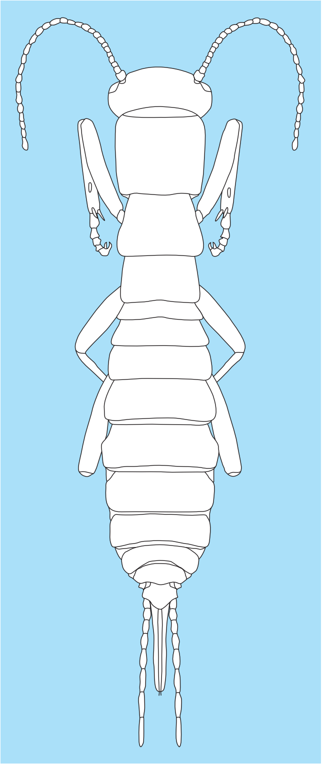
Bibliography. J. del Hoyo et al. (eds), Order Gruiformes, *Handbook of the Birds of the World*, vol. 3, pp. 34–273, Lynx Edicions, 1996; P. A. Johnsgard, *Bustards, Hemipodes and Sandgrouse: Birds of Dry Places*, Oxford University Press, 1992; P. A. Johnsgard, *The Cranes of the World*, Indiana University Press, 1983; B. Taylor, *Rails: A Guide to the Rails, Crakes, Gallinules and Coots of the World*, Yale University Press, 1998.

Grylloblattodea

A relatively new order of flightless insects first described from Sulphur Mountain near Banff, Canada, in 1914 by entomologist E. M. Walker. There are now 26 described species of living Grylloblattodea in five genera. Called ice crawlers, they are typically found at high elevations in East Asia, Japan, Siberia, and western North America. *See* INSECTA; ORTHOPTERA.

Description. Ice crawlers are found under rocks, in leaf litter, and occasionally in caves. Some Asian cave dwellers are blind. *Grylloblatta* species were one of the first groups of insects to colonize Mount St. Helens (Washington) within a few years after the 1980 eruption.

Some species appear to require a narrow temperature range: a *Grylloblatta* living on Mount Rainier (Washington) undergoes heat convulsions above about 14°C (57°F), and has a metabolism that functions best in a thermal window between −6 and 12°C (between 21 and 54°F). These insects appear to survive by occupying moderate habitats such



Dorsal view of the first ice crawler species to be described: *Grylloblatta campodeiformis* as. (*Redrawn from the original paper by E. M. Walker, A new species of Orthoptera, forming a new genus and family, Can. Entomol., 46(3):93–99, 1914*)

as those with temperatures moderated by snow cover.

Living Grylloblattodea are defined by a number of unique characters, including an eversible sac on the underside of the first abdominal segment, a spine on the underside of the meta (third) thoracic segment, and distinct genitalic and ovipositor structures (see **illustration**).

Phylogeny. There is some debate about the relationships of Grylloblattodea with both living and extinct insects. With regard to living species, some analyses of molecular (DNA) characters and analyses of a combination of molecular and structural characters show Grylloblattodea as sharing a common ancestry with earwigs (order Dermaptera). More recent molecular analysis pairs Grylloblattodea with the flightless African Mantophasmatodea, named in 2002 and displacing Grylloblattodea as the newest described insect order. One analysis places stick insects (Phasmatodea) in the same group as Grylloblattodea and Mantophasmatodea. This grouping of two to three orders shares a common ancestor with the lineage (Dictyoptera) comprising cockroaches [order Blattaria (Blattodea)] and praying mantids (order Mantodea). *See* BLATTARIA (BLATTODEA); DERMAPTERA.

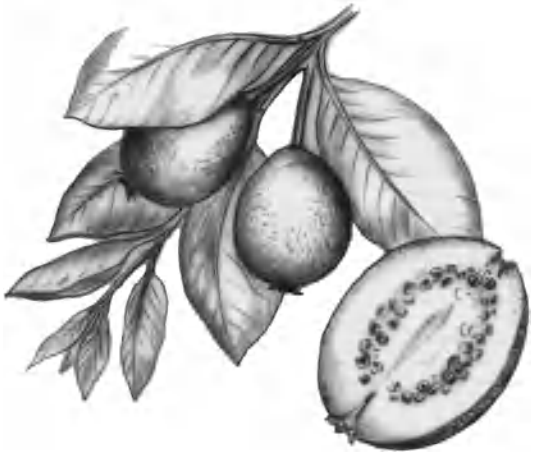
Likely Jurassic and Permian ancestors of modern Grylloblattodea include fossil insects (some with fully formed wings) that share characters such as ovipositor structure with living species. However, some investigators argue that the order Grylloblattodea includes a much broader diversity of fossil insects representing a group that was once very widespread and common in the Paleozoic and early Mesozoic and whose habitats included much warmer temperate and tropical climes than occupied by modern ice crawlers. This broad view of Grylloblattodea suggests that they were ancestral to orthopteroid orders as well as Dermaptera and Plecoptera (stoneflies). These extinct species included within Grylloblattodea comprised about half of all insects from some Permian sites. Many were consumers of pollen and spores. Other authorities argue that this broad view of the past diversity of Grylloblattodea is based on an inappropriate grouping of unrelated extinct families. *See* PLECOPTERA.

Darryl T. Gwynne; Glenn K. Morris

Bibliography. D. Grimaldi and M. S. Engel, *Evolution of the Insects*, Cambridge University Press, 2005; V. R. Vickery and D. K. M. Kevan, *The Grasshoppers, Crickets and Related Insects of Canada and Adjacent Regions*, Canadian Government Services, Ottawa, 1985.

Guava

A plant, *Psidium guajava*, of tropical America that has long been in cultivation. It is a shrub or low tree which belongs to the myrtle family (Myrtaceae). The fruit (see **illus.**) is a berry, yellow when ripe, and quite variable in size depending on variety and growing conditions, the average being about $2^1/_2$ in. (7.2 cm) long. The guava is quite aromatic, sweet, and juicy. It is used mostly for jellies and preserves, but also as a fresh fruit. *See* MYRTALES.

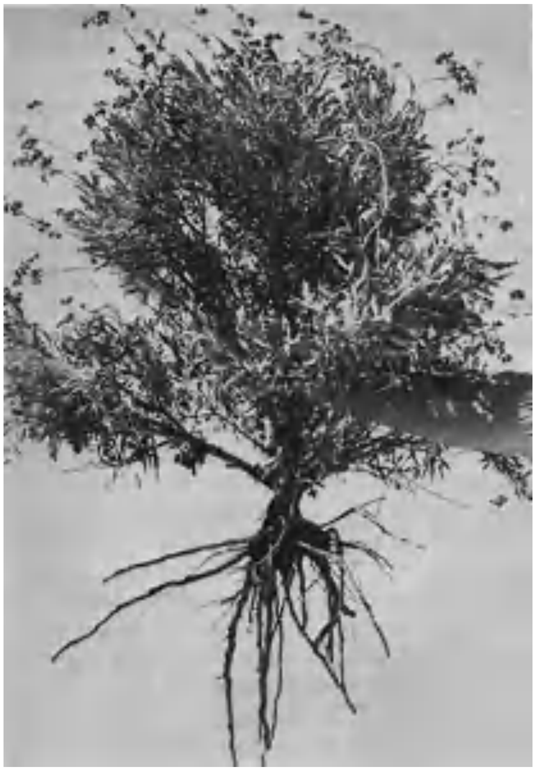


***Psidium guajava*, showing a branch with leaves and two berries and a berry cut in half.**

Perry D. Strausbaugh; Earl L. Core

Guayule

A desert plant, *Parthenium argentatum*, of the composite family (Compositae), which produces rubber. It is a native perennial shrub growing in the Chihuahuan Desert of north-central Mexico and southwestern Texas. The plant is bushy with dense branches, thick clusters of silverlike leaves, a strong taproot, and a thick crown (see **illus.**). At maturity



Guayule plant, maximum height 40 in. (102 cm).

the plant varies in height from 30 to 40 in. (76 to 102 cm). *See* CAMPANULALES.

Early history. The presence of rubber in guayule was known by the Aztecs and certain North American Indians for several centuries. However, the first published account of guayule was not until 1852, when J. M. Bigelow, a medical doctor, discovered the plant while working with the Mexican Boundary Survey Team. His specimens were sent to the Gray Herbarium of Harvard University, where guayule was first described taxonomically.

Guayule was widely exposed to the American public for the first time in 1876, when the Mexican government sent an exhibit to the Centennial Exposition at Philadelphia. During the next 25 years several companies researched guayule rubber. In 1904 W. A. Lawrence developed the pebble mill extraction method. The Mexicans were very enthusiastic about guayule, and within 5 years over a dozen extraction factories were built to harvest and process guayule from wild stands. By 1910 one-half of the United States supply of natural rubber came from guayule in Mexico.

Breeding and cultivation. The wild stands of guayule were quickly depleted, and with the exhaustion of the natural supply, the Continental-Mexican and intercontinental rubber companies undertook studies to domestically cultivate guayule. Extensive commercial development began in Salinas, California, in 1925, and between 1931 and 1941, 3,068,630 lb (1,391,907 kg) of rubber were milled. During World War II the domestication of guayule became a national issue when the supply of *Hevea* rubber from the Far East was cut off. With the development of synthetic rubber and the end of the war, economic and political incentives for guayule production disappeared, and all projects were terminated by 1959.

Guayule is a tetraploid of 72 chromosomes and reproduces by apomixis. Diploids of 36 chromosomes are also found, and they reproduce sexually. Breeding programs in guayule seek improved varieties with larger rubber yields, increased disease resistance, and greater cold tolerance. There are several closely related species of *Parthenium*, some treelike and some very weedy. Crossing these species with guayule gives wide variation for selection and breeding.

In 1976 research was reinitiated in Arizona and California. The rising price of oil and the accompanying political difficulties with oil-producing countries rekindled interest in guayule as a source of natural rubber. Also, guayule could be very important for agricultural economy in certain parts of Texas, Arizona, and California, where the high summer temperatures, low rain fall, and absence of freezing winters are favorable for guayule cultivation. As water becomes a diminishing resource, a drought-tolerant crop such as guayule could be very important for farmers. It seems possible that if certain problems in cultivation and processing are successfully resolved, guayule could become a major cash crop. *See* RUBBER; RUBBER TREE. David D. Rubis; Linda J. Cassens

Guidance systems

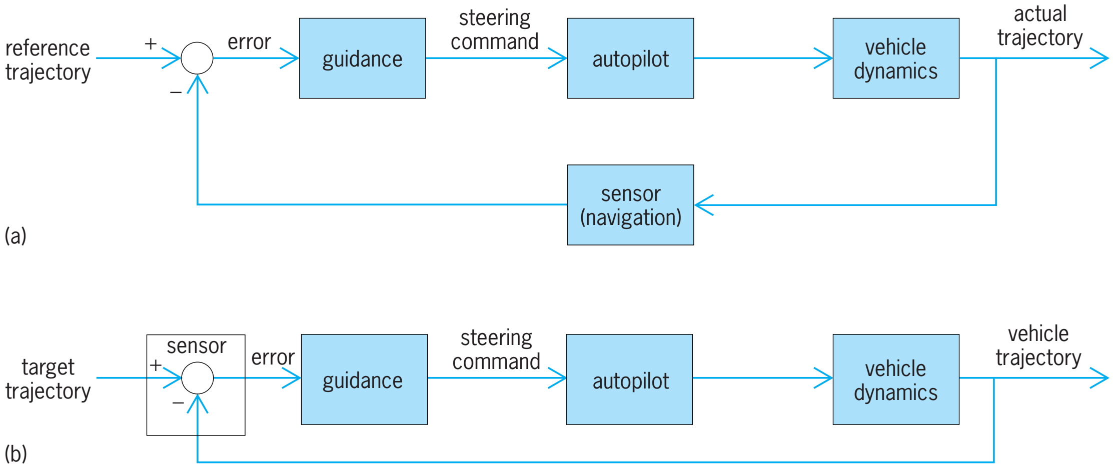
The algorithms and computers utilized to steer a vehicle along a path. The types of vehicles include airplanes, rockets, missiles, ships, torpedoes, drones, and material transport vehicles within factories and so forth. The means of steering depend on the vehicle and can be the rudder, elevators, and other control surfaces on an airplane, the rudder on a ship, the control surfaces on a missile or on a torpedo, the gimbal angle of the motor on a rocket, and others. In every case the guidance system utilizes knowledge of the difference between where the vehicle should be and where it is. The difference between these two vectors is processed by the guidance algorithm. The output is a steering command intended to reduce the error between the desired and the actual paths. *See* DRONE; ELEVATOR (AIRCRAFT); FLIGHT CONTROLS; SHIP POWERING, MANEUVERING, AND SEAKEEPING.

The configuration shown in **illus.** *a* applies where a path or trajectory for the vehicle has been computed and stored in the computer. As time passes, the corresponding values for the trajectory, desired position, and velocity are made available at the summation device. Meanwhile, sensors measure the actual position and velocity of the vehicle. This set of signals is fed back to the summation device and is subtracted from the reference trajectory. The output of the summation device is the difference between these two signals, that is, the error signal. This configuration would be utilized, for example, to boost a satellite into orbit. Inertial navigation utilizing gyroscopes could provide the sensing operation. The role of the guidance system is to process this error signal and use the results to steer the vehicle along the reference trajectory. *See* AUTOPILOT; CONTROL SYSTEMS.

The configuration shown in illus.*b* applies when the vehicle is attempting to track a target which may or may not be moving. The absolute position of the target is not sensed, but rather the difference between the vector from the vehicle to the target and the vector describing the vehicle orientation. Thus the sensor is different from that used in the configuration of illus. *a*. A common example would be a radar rigidly mounted to the nose of a vehicle. Here the sensed target bearing would automatically be measured relative to the attitude or orientation of the vehicle. The role of the guidance system is to process this error signal and use the result to steer the vehicle toward the target. *See* HOMING.

Several important performance attributes contribute to the effectiveness of the system. These attributes are governed by the guidance system and by the other system components, including the vehicle itself and its dynamic behavior.

A primary concern is accuracy, which is certainly a function of the guidance system but under the best of circumstances can be no better than the quality of the sensor. Whether the goal is to insert a satellite into synchronous orbit or to try to intercept an



Guidance system for (*a*) following a specified trajectory and (*b*) homing on a target.

enemy aircraft with an air-to-air missile, the accuracy of the sensor and the properties of the guidance system are the principal factors in overall accuracy.

Another concern is speed of response. Here the dynamics of the vehicle itself can be a limiting factor. The guidance system must compensate to the extent possible in providing a fast, responsive system. The system should be able to recover from errors as quickly as possible and return to the desired path. In the case of homing on a target, this is crucial if the target can maneuver. Coupled with the need for a quick response is the simultaneous need for a stable response. The system must not respond so fast as to overshoot its proper heading excessively, causing possible growing oscillations. Consideration of stability places certain limitations on the guidance system and on the speed of response of the overall systems.

Another important feature of the system is its robustness. The guidance system design is based on a mathematical model of the vehicle, the autopilot, and the sensor. In fact, none of these behaves exactly like its model. The guidance system must provide good overall performance despite these modeling errors. Environmental effects (winds, ocean currents, and so forth) may also be different than what was expected. A robust guidance system provides good performance despite these uncertainties.

Reliability is also important. Efforts must be made to accommodate failures of various kinds. In many cases, backup components are provided for redundancy. This is frequently the case for the digital computer of the guidance system, especially for crewed space flight. *See* RELIABILITY, AVAILABILITY, AND MAINTAINABILITY.

Gerald Cook

Bibliography. R. H. Battin, *An Introduction to the Mathematics and Methods of Astrodynamics*, 1987, revised, 1999; M. Kayton, Navigation: Ships to space, *IEEE Trans. Aerosp. Electr.*, 24(5):474–519, 1988; C.-F. Lin, *Modern Navigation, Guidance, and Control*, 1991; G. W. Stimson, *Introduction to Airborne Radar*, 2d ed., 1998.

Guided missile

An uncrewed, controlled-flight vehicle that is guided to a target by guidance and control equipment. This equipment may be carried in the missile vehicle itself, or guidance may be directed from the launch site. The term is generally reserved for aerodynamic maneuverable missiles that may be guided to predetermined targets for military purposes.

Classification. Guided missiles are classified by launch or target mode such as air to air (AAM), air to surface (ASM), surface to air (SAM), surface to surface (SSM), and other possible modes such as subsurface launch to air, surface, or subsurface targets, and surface or air launch to subsurface targets. Missiles may be classified by range (short, medium, long) or by techniques related to tracking and guidance (radar, infrared heat seeker, optical or television, laser, radio, wire control command, fiber optics, inertial, acoustical). Some missiles make use of terrain following, which permits the missile to look at the terrain, compare it with a predetermined mapped route, and in effect fly a course as if by following a road map. Particularly for long-range flight, missiles may also make use of celestial navigation (star tracking) or may use electronic guidance assistance provided by satellites or midcourse surface stations. With optical missiles, a two-way data link may also be incorporated that provides the launcher with a continuous display from the missile video after launch and egress, and permits target updating or even reselection from a safer standoff distance (**Fig. 1**). *See* ELECTRONIC NAVIGATION SYSTEMS; GUIDANCE SYSTEMS; SATELLITE NAVIGATION SYSTEMS; STAR TRACKER.

Propulsion. Guided missiles are generally self-propelled, and may use rocket motors (liquid or solid), air-breathing turbojet engines, ramjets, or various types of combined-cycle engines such as air-augmented rockets, ducted rockets, turboramjets, or hybrid rockets (for example, solid fuel and liquid oxidizer). Future use may be made of nuclear rocket engines or various types of electrical rockets, such as charged particles or ion rockets. Advances

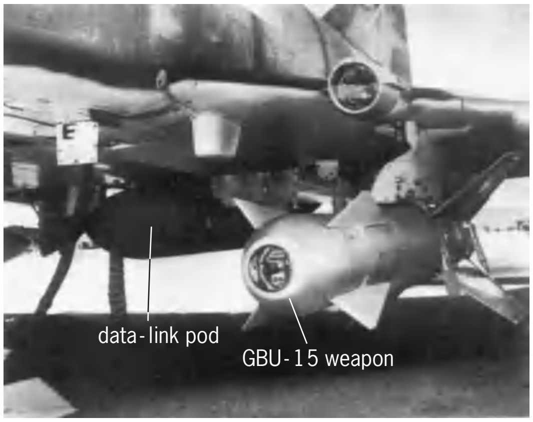


Fig. 1. GBU-15 with data link system, mounted on F-4 Phantom II fighter. (*Hughes Aircraft Co.*)

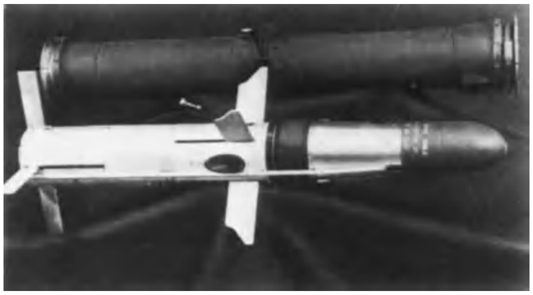


Fig. 2. TOW (tube-launched, optically tracked, wire-guided) antitank missile and launch tube. Wings and tails fold to permit tube mounting. (*Hughes Aircraft Co.*)

in the development of scramjet engines could provide the propulsion for a hypersonic cruise missile in the Mach number speed range from about 4 up to about 7. For some missions, particularly air-to-surface missions, unpowered, gliding, guided missiles may be used. *See* ION PROPULSION; ROCKET PROPULSION; TURBINE ENGINE SUBSYSTEMS; SCRAMJET.

Warhead. The kill mechanism for a missile consists of some form of explosive warhead and a system for detonation (fusing and arming). Warheads are typically either a high explosive or nuclear, and vary according to the type of fragment that is ejected and the spray pattern of the fragment. Warheads may be exploded upon contact with the target, by command from an external source, by a proximity fuse that senses the target, by preset timers, and so on.

Capabilities and constraints. Guided missiles are intended to be a type of munition extension with capabilities beyond that of commonly used, pure ballistic, free-flying projectiles. The capability extension may be in accuracy, warhead size, mobility, speed, range, altitude, endurance, or target selection. By the nature of the increased capability, missiles can be complicated and costly. Functions required of a missile system require the combined efforts of specialists in fields of aerodynamics, propulsion, electronics, materials, structures, warheads, sensors, fuses, reliability, safety, packaging, carriage, launching, and testing and evaluation. A severe constraint is imposed by the fact that a missile is a "throwaway" item, and its cost versus effectiveness must be carefully weighed. It is necessary that the required mission be achievable, but at an affordable cost. That is to say, a missile cost that exceeds the target cost would be undesirable; hence an effective missile with simplicity and low cost is a system to be sought.

Early development. During and shortly after World War I, the British tested radio-controlled airplanes and rockets, while the Americans tested an airplane controlled by on-board, preset pneumatic and electrical controls. The Germans became interested in missiles in the mid-1930s and undertook experiments with large guided rockets. This led to the initiation of guided missile warfare in 1943, with the use of radio- controlled, rocket-powered glide bombs against ships, and to the launching of several thousand V-1 and V-2 surface-to-surface missiles during 1944 and 1945. At the close of World War II, the Germans were developing a surface-to-air guided missile designed to carry out complex pursuit and evasion

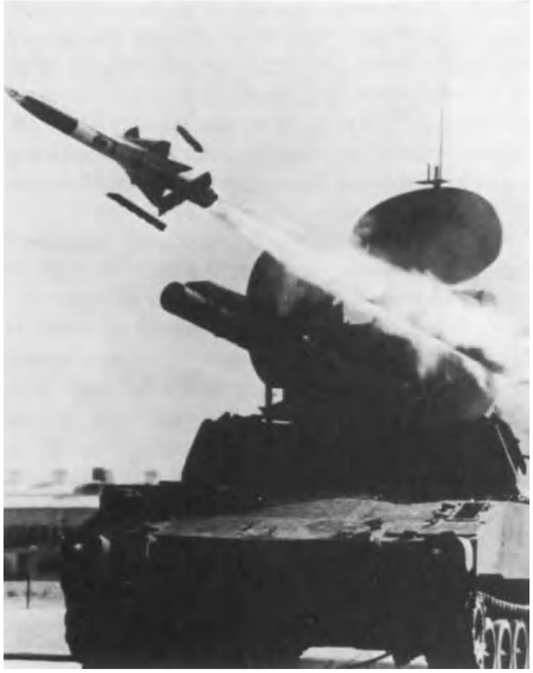


Fig. 3. Launching of a field-mobile, short-range, all-weather surface-to-air Roland missile. The German–French Roland has been licensed for production in the United States for use as a U.S. Army air defense weapon. (*U.S. Army*)

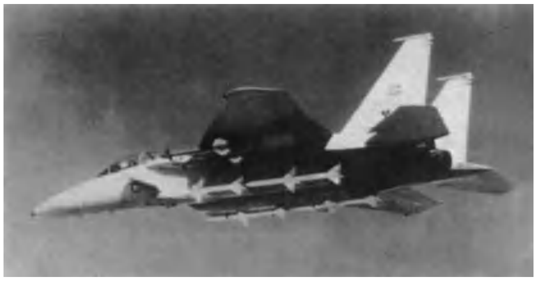


Fig. 4. F-15 Eagle fighter aircraft equipped with a mixture of short-range, infrared Sidewinder and medium-range, radar Sparrow air-to-air missiles. (*McDonnell-Douglas*)

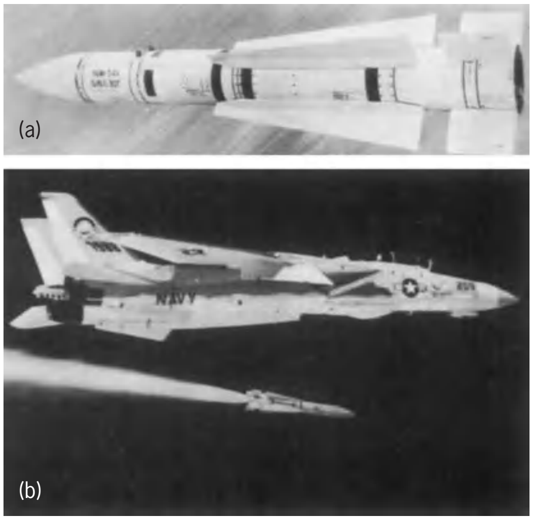


Fig. 5. Phoenix long-range air-to-air supersonic missile. (*a*) Side view (*Hughes Aircraft Co.*). (*b*) Launch from F-14 fighter aircraft (*U.S. Navy*).

maneuvers, and research based on this system continued in the United States.

Use. Since World War II, guided missiles have experienced only limited use in combat, primarily in Southeast Asia, in the Middle East wars, and in the Falklands; thus a measure of combat effectiveness is difficult to assess. However, missiles have changed the nature of warfare in many ways. Ground troops have an increased firepower through the use of human-portable antiair and antitank missiles (**Fig. 2**). Ground troop protection is increased through the use of mobile air defense missiles (**Fig. 3**). Ground warfare has also been changed through the use of helicopter-launched missiles for both ground and air target attack.

Navy vessels have an increased firepower through the use of antishipping-type missiles, as well as increased protection with air defense missiles. Some unique missile deployment systems have appeared in the Soviet navy. The Kirov cruiser, for example, carries a variety of missiles suggestive of both defensive and offensive capability, including antisubmarine and two antiair types, as well as long-range cruise surface-to-surface antishipping or land-attack missiles. The Slava cruiser, as another example, is equipped with long-range surface-to-surface cruise missiles and two types of surface-to-air missiles. Newer ships, such as these, make use of vertical launch systems which provide certain advantages when launching from a moving platform in that launcher elevation and azimuth settings are not critical to the missile trajectory, and target acquisition is performed by the missile itself after launch.

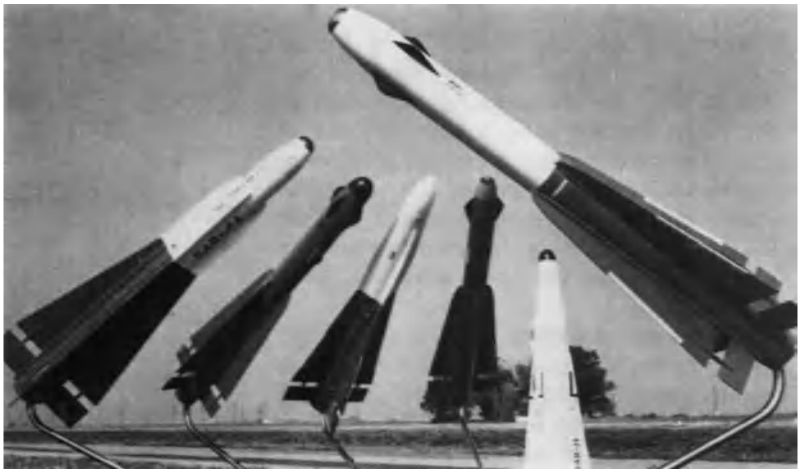


Fig. 6. Falcon air-to-air missile family, including missiles with infrared and radar guidance, and high-explosive and nuclear warheads. (*Hughes Aircraft Co.*)

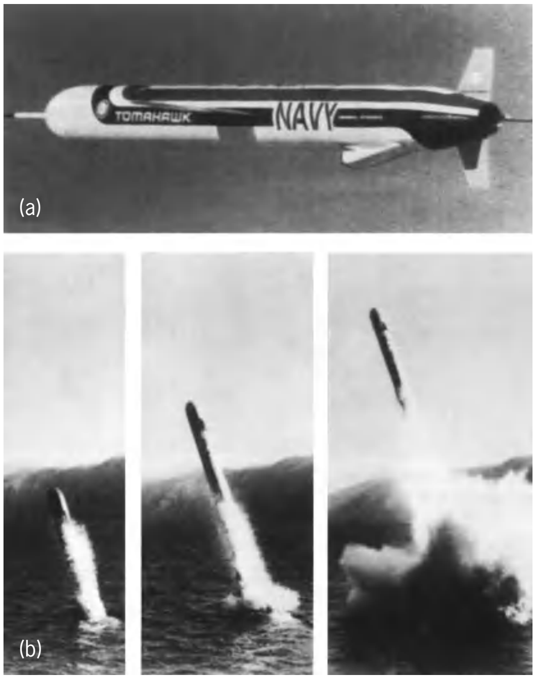


Fig. 7. Tomahawk long-range, subsonic cruise missile. (*a*) Side view. (*b*) Submerged launch. Missile may also be launched from surface or air. (*General Dynamics*)

Aircraft have an increased effectiveness through the use of a variety of air-to-air combat missiles (**Figs. 4, 5,** and **6**) and air-to-surface missiles ranging from battlefield ground support to long-range cruise stand-off missiles.

The cruise missile (**Figs. 7** and **8**), either subsurface-, surface-, or air-launched, provides increased flexibility for tactical and strategic missions from stand-off distances that provide improved safety for the launch platform. Various degrees of stand-off distances are achieved with ballistic missiles by using a variety of types of midcourse and terminal guidance, types of warheads, propulsion schemes, and launch techniques. *See* AIR ARMAMENT; ANTISUBMARINE WARFARE; ARMY ARMAMENT; MILITARY AIRCRAFT; NAVAL ARMAMENT.

Guided missile systems are essential elements of the arsenals of many nations. Some of the world missile systems are listed in the **table**.

Some omissions and uncertainties are probably due to incomplete sources of information. The listing

World missile systems

United States

AAM

Falcon
Sparrow
Oriole
Meteor
Ding-Dong
Sidewinder
Eagle
Genie
Diamondback
Phoenix
Agile
Brazo
CLAW
SHAG
Seek Bat
AMRAAM
ASAT (ALMV)

SAM

Lark
Terrior
Nike Ajax
Bomarc
Loki
Hawk
Talos
Tartar
Nike Hercules
Mauler
Nike Zeus
Redeye
Davy Crockett
Bull Goose
Typhoon
Nike X/Sprint
Chaparral
SAM-D
Standard
Spartan
Stinger
Patriot
RAM
ADATS

ASM

Rascal
Bullpup
Bulldog
Dove
Gorgon
Bold Orion
Corvus
Hound Dog
Skybolt
Wagtail
Quail
Crossbow
Green Quail
Shrike
Condor
SRAM
WASP
Standard ARM
Teton
Cobra
LASRM
Maverick
Hornet
Walleye
Viper
ZAP
SCAD
Harpoon
ALCM
HARM
Hellfire
HOBOS
Paveway
ASALM
SLAM
GBU-15

SSM

Matador
Corporal
Regulus I
Regulus II
Honest John

SSM (*cont.*)

Snark
Navaho
Redstone
Lacross
Hermas
Rigel
Atlas
Dart
Little John
Sergeant
Jupiter
Polaris
Thor
Titan
Triton
Clam
Slam
Cobra
Lobber
Mace
MAW
Minuteman
Pershing
Shillelagh
Law
Taurus
TOW
Lance
Poseidon
Dragon
Trident
Tomahawk
Viper
Petrel
Able
Asroc
Subroc
Astor
Rat
Able-Alfa
MX
Copperhead

Russia

AAM

AA-1
AA-2
AA-3
AA-4
AA-5
AA-6
AA-7
AA-8
AA-9
AA-XP-1
AA-XP-2

SAM

SA-1
SA-2
SA-2/3
SA-2/4
SA-3
SA-4
SA-5
SA-6
SA-7
SA-8
SA-9
SA-10
SA-11
SA-12
SA-13
SA-N-3
SA-N-4
SA-N-5
SA-N-6
SA-N-7
SA-N-8
Galosh
ABM-X-3
SH-4
SH-8

ASM

AS-1
AS-2
AS-3
AS-4
AS-5
AS-6
AS-7
AS-8
AS-9
AS-X-10
AS-11
AS-X-15
BL-10

SSM

AT-1
AT-2
AT-3
AT-4
AT-5
AT-6
Salish
Samlet
Scrubber
Shaddock
Styx
Sark
Serb
Scud
Shyster
Sandal
Skean
Sapwood
Saddler
Sa-sin
Scarp
Scrag
Sego
Scaleboard
Savage
Scapegoat
Scrooge

SS-16
SS-17
SS-18
SS-19
SS-20
SS-21
SS-22
SS-23
SS-X-24
SS-X-25
SS-N-6
SS-N-7
SS-N-8
SS-N-9
SS-N-10
SS-N-11
SS-N-12
SS-NX-13
SS-N-14
SS-N-15
SS-N-16
SS-N-17
SS-N-18
SS-NX-19
SS-N-20
SS-NX-21
SS-NX-22
SS-NX-23
SSC-X-4

Argentina

Martin Pescador
Mathogo

Australia

Ikara

Brazil

Carcara
MSS-1

China

CSS-2
CSS-3
CSS-X-4

France

ACRA
Entac
SS-11 (Harpoon)
SS-12
Pluton
Magic R550
R-511
R-530
AS-12
AS-20
AS-30
Martel
Crotale R-440
Super 530
Masurca
Exocet
Malafon
HOT
Milan
Otomat
Mercure
Hirondelle
MICA
Fakir-H
ANS
AS-15
ASMP
ARMAT
SATCP
AS-11
SSBS S-3
MSBS M-20
MSBS M-4
Sica/Shahine

Germany

Cobra
Roland
Viper
Kormoran
Armburst
Mamba
Jumbo
Tiralleur

Israel

Shafrir
Gabriel
Picket
Python
PDM

Italy

Airtos
Aspide
Indigo
Sea Killer
Marte
Mosquito
Sparviero

Japan

KAM-3D
KAM-9
ASM-1
AAM-1
AAM-2
Tan-SAM

Norway

Penguin
Teme

Republic of South Africa

Kukri

Sweden

Bantam RB53
RB04
RB05
RB08
RB70
RB56
RB15

Switzerland

Micon
Mosquito

Taiwan

Kun Wu

United Kingdom

Blue Steel
Swingfire
Vigilant
Blowpipe
SLAM
Rapier
Firestreak
Red Top
SRAAM
Bloodhound
Thunderbird
Tiger Cat
Hellcat
Seacat
Sea Dart
Seaslug
Seawolf
Sea Skua
SAMM10
Hawkswing
Beeswing
Skyflash
ASRAAM
ALARM
Javelin
Sea Eagle

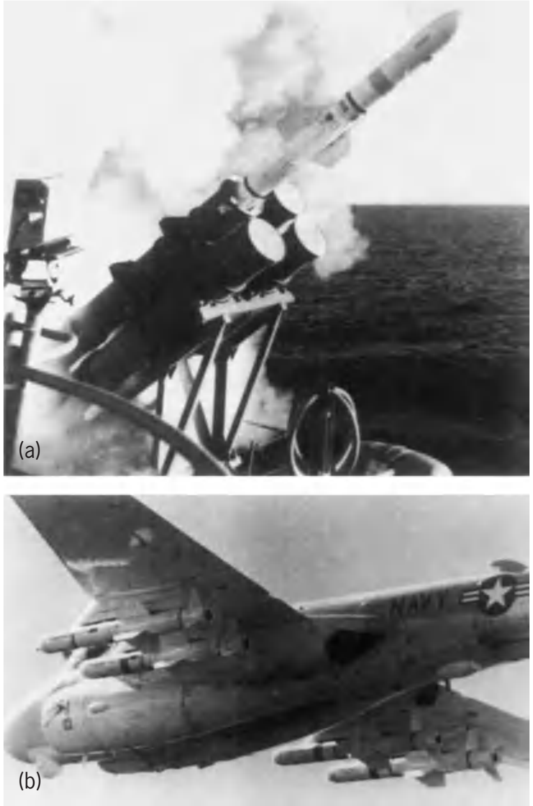


Fig. 8. Harpoon antiship cruise missile. (*a*) Ship-launched. (*b*) Air launch version on A-6 intruder attack aircraft. (*McDonnell-Douglas*)

also includes many experimental and defunct systems.

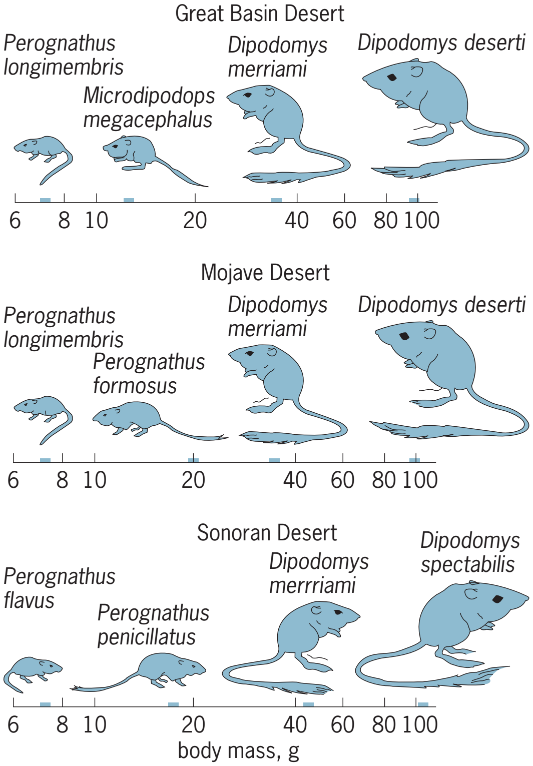
M. Leroy Spearman

Bibliography. Air Force Magazine, *Air Force Almanac*, annually; Aviation Week and Space Technology, *Aerospace Forecast and Inventory*, annually; J. R. Chambers, *Partners in Freedom: Contributions of the Langley Research Center to U. S. Military Aircraft of the 1990s*, National Aeronautics and Space Administration, Washington, DC, 2000; E. Fleeman, *Tactical Missile Design*, 2d ed., American Institute of Aeronautics & Astronautics, Reston, VA, 2006; Flight International, *World Missile Directory*, annually; General Dynamics—Pomona Division, *The World's Missile Systems*, revised periodically; B. Gunston, *The Illustrated Encyclopedia of the World's Rockets and Missiles*, Smithmark Pub, 1987; M. J. Hemsch (ed.), *Tactical Missile Aerodynamics: General Topics*, American Institute of Aeronautics & Astronautics, Reston, VA, 1992; M. L. Spearman, *Historical Development of World Wide Guided Missiles*, NASA TM 85658, 1983; M. L. Spearman, *A Historical Review of Missile Airframe Designs*, AIAA Pap. 97–2277, 1997.

Guild

A group of species that utilize the same kinds of resources, such as food, nesting sites, or places to live, in a similar manner. Emphasis is on ecologically associated groups that are most likely to compete because of similarity in ecological niches, even though species can be taxonomically unrelated. The term was derived from the guild in human society composed of people engaged in an activity or trade held in common.

Richard B. Root first used the term guild in 1967, specifying the "foliage-gleaning guild" as a group of birds that habitually harvest arthropods (mainly insects and spiders) from tree leaves. In 1973 he distinguished three guilds based on feeding methods of herbivorous invertebrate animals in a community utilizing one plant species: the sap-feeding guild, the strip-feeding guild, and the pit-feeding guild. Members of the sap-feeding guild sucked juices from the plant with tubular mouthparts and included plant lice (Aphidae), frog hoppers (Cercopidae), leaf hoppers (Cicadellidae), and plant bugs (Miridae). Members of the strip-feeding guild chewed off strips of leaves with cutting mandibles or similar mechanical shears, and included caterpillars (Lepidoptera), crickets and grasshoppers (Orthoptera), and snails and slugs (Mollusca). Members of the pit-feeding guild used chewing or sucking mouthparts but concentrated feeding in small areas, resulting in many pits of consumed or damaged tissue; they included leaf beetles (Chrysomelidae), weevils (Curculionidae), and thrips (Thripidae). This was the first time that unrelated organisms were grouped into ecologically meaningful units smaller than the whole community; these groups represent the major functional roles in the community. *See* ECOLOGICAL COMMUNITIES.



Guilds of granivorous rodents in three deserts in the southwest United States. Mean weight (in grams) of each species is plotted on the body mass axis. The body mass scale is logarithmic, so that equal spacing shows equal ratios. (*After M. A. Bowers and J. H. Brown, Body size and coexistence in desert rodents: Chance or community structure?, Ecology, 63:391–400, 1982*)

The guild concept focuses attention on the ways in which ecologically related species differ enough to permit coexistence, or avoid competitive displacement. For example, new places to live for some plants are provided by badger mounds in dense tall-grass prairie vegetation. Many species were associated with these disturbances, and since mounds were relatively small and rare, the potential for competition in this guild was high. One important difference between guild members was dispersal of seeds, so that some colonized new mounds quickly and others slowly, the result being a temporal divergence in utilization. The other major difference to which plants responded was the soil moisture in mounds. The combination of these characteristics permitted many species of plants to coexist on a population of badger mounds.

Seeds provide the major source of food for many desert rodents. Studies on such granivorous rodent guilds have shown that they are highly structured in relation to body sizes of member species. In one study it was found that although species assemblages differ between the Great Basin, Mojave, and Sonoran deserts, four species per guild coexist, with distinct body sizes at about 100, 40, 20, and 7 g (3.5, 1.4, 0.7, and 0.2 oz, respectively; see **illus.**). This uniformity in guild composition suggests the strong influence of competition organizing the guild. *See* POPULATION ECOLOGY.

Not all guilds are as clearly structured as the granivorous rodent guild. Often the hypothesis of random assemblage of species cannot be refuted in many cases based on size differences. Much research interest is focused by ecologists on the role of competition in structuring guilds and communities.

The guild is also commonly used as the smallest unit in an ecosystem in studies relating to environmental impact, wildlife management, and habitat classification. A representative species of a guild may be selected for study involving the uncertain assumption that environmental impact will influence this species in the same way as other guild members. *See* ECOSYSTEM.

Peter W. Price

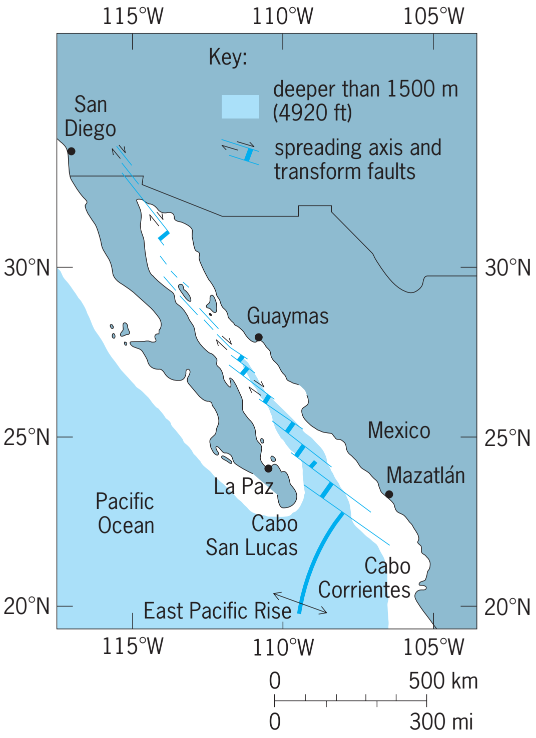
Bibliography. M. A. Bowers and J. H. Brown, Body size and coexistence in desert rodents: Chance or community structure?, *Ecology*, 63:391–400, 1982; P. W. Price, *Insect Ecology*, 3d ed., 1997; R. B. Root, Organization of a plant-arthropod association in simple and diverse habitats: The fauna of collards (*Brassica oleracea*), *Ecol. Monogr.*, 43:95–124, 1973; D. Simberloff and W. Boecklen, Santa Rosalia reconsidered: Size ratios and competition, *Evolution*, 35:1206–1228, 1981.

Gulf of California

A young, elongate ocean basin on the west coast of Mexico. It is flanked on the west by the narrow mountainous peninsula and continental shelf of Baja California, while the eastern margin has a wide continental shelf and coastal plain. The floor of the gulf consists of a series of basins 3300–12,000 ft (1000–3600 m) deep, whereas the northern gulf is dominated by a broad shelf which is the result of deltaic deposition from the Colorado River. The structural depression of the gulf continues northward into the Imperial Valley of California, which is cut off from the ocean by the delta of the Colorado River. *See* CONTINENTAL MARGIN.

The gulf formed along the plate edge between the Pacific and the North American lithospheric plates (see **illus.**). The plate edge enters the gulf at the north as the San Andreas Fault System and passes through a series of short segments of transform faults and spreading axes to where it becomes the East Pacific Rise entering the mouth of the gulf at the south. Opening of the gulf is believed to have been initiated between 5 and 6 million years before present, contemporaneous with an eastward jump of the trend of the San Andreas system in southern California, and the gulf was created by approximately 180 mi (300 km) of northwestward displacement of the peninsula of Baja California as a part of the Pacific plate. Thus, the floor of the gulf consists of young oceanic crust beneath the basins, adjacent to thin subsided continental crust along the continental margins. *See* FAULT AND FAULT STRUCTURES; PLATE TECTONICS.

Most of the gulf lies within an arid climate, with 4–6 in. (10–15 cm) of annual rainfall over Baja California and ranging on the eastern side from 4 in. (10 cm) in the north to about 34 in. (85 cm) in the southeast. No year-round streams enter the gulf on the west; a series of intermediate-size rivers flow in on the east side; and the major source of freshwater sediment came from the Colorado River at the



Simplified tectonics of the Gulf of California. (*After J. R. Curray et al., Tectonics and geological history of the passive continental margin at the tip of Baja California, in Initial Reports of the Deep Sea Drilling Project*, *vol. 64*, *pt. 2*, *U.S. Government Printing Office*, *1982*)

north prior to damming it upstream in the United States.

Water circulation is driven by seasonal wind patterns. Surface water is blown into the gulf in the summer by the southwesterly wind regime. In the winter, surface water is driven out of the gulf by the northwesterly wind regime, and upwelling occurs along the eastern margin, resulting in high organic productivity. Bottom sediments of the gulf range from deltaic sediments of the Colorado River at the north and coalesced deltas of the intermediate-size rivers on the east. A strong oxygen minimum occurs between 990 and 3000 ft (300 and 900 m) water depth, where seasonal influx of terrigenous sediments and blooms of diatoms due to upwelling produce varved sediments consisting of alternating diatom-rich and clay-rich layers. Rates of sediment accumulation are high, and total sediment fill beneath the Colorado River delta at the north may attain thicknesses of greater than 6 mi (10 km), even though the structural depression and the underlying crust are geologically young. *See* BACILLARIOPHYCEAE; DELTA; MARINE SEDIMENTS; OCEAN CIRCULATION; UPWELLING; VARVE. Joseph R. Curray

Bibliography. J. K. Crouch and S. B. Bachman (eds.), *Tectonics and Sedimentation Along the California Margin*, Pacific Section, Society of Economic Paleontologists and Mineralogists, 1984; J. R. Curray et al., *Initial Reports of the Deep Sea Drilling Project*, vol. 64, pts. 1 and 2, 1982; R. A. Schwartzlose and J. R. Hendrickson, *Bibliography of the Gulf of California: Marine Sciences (Through 1981)*, Inst. Cienc. Mar Limnol., Univ. Nal. Auton. Mex., vol. 7, 1983; Tj. H. van Andel and G. G. Shor, Jr. (eds.), *Marine Geology of the Gulf of California: A Symposium*, Amer. Ass. Petrol. Geol. Mem. 3, 1964.

Gulf of Mexico

A subtropical semienclosed sea bordering the western North Atlantic Ocean. It connects to the Caribbean Sea on the south through the Yucatán Channel and with the Atlantic on the east through the Straits of Florida. To the north, it is bounded by North America, to the west and south by Mexico and Central America, and on the east and southeast by Florida and Cuba respectively. The Gulf of Mexico Basin lies wholly within the boundaries of the North American tectonic plate. *See* BASIN; INTRA-AMERICAS SEA; PLATE TECTONICS.

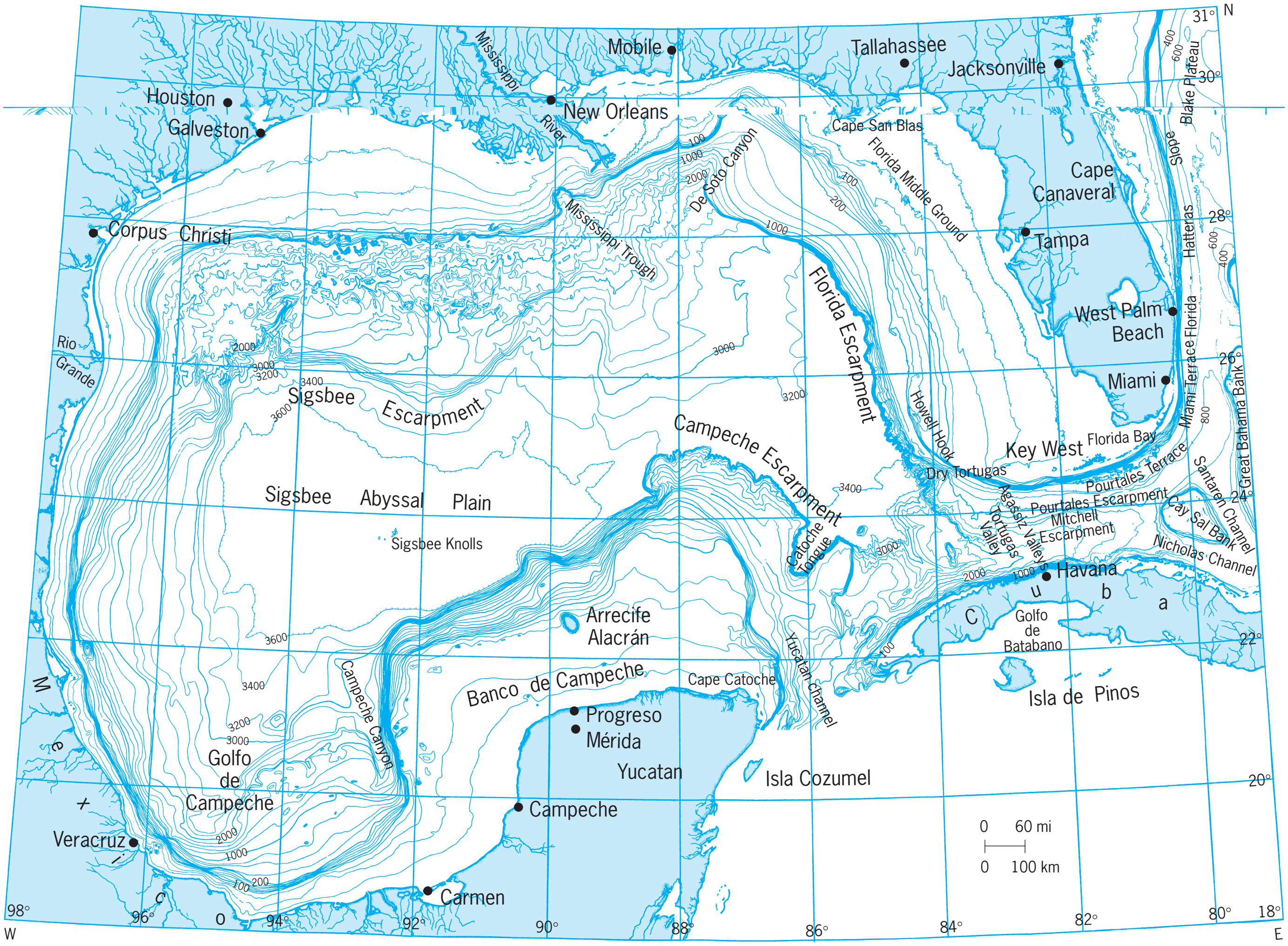
Marine geology. The continental shelves surrounding the gulf are very broad along the eastern (Florida), northern (Texas, Louisiana, Mississippi, Alabama), and southern (Campeche) area, averaging 125–186 mi (200–300 km) wide (see **illus.**). The continental shelves along the western and southwestern (Mexico) and southeastern (Cuba) boundaries of the gulf are narrow, often being less than 12 mi (20 km) wide. Between the continental shelves and the Sigsbee Abyssal Plain are three steep continental slopes: the Florida Escarpment off west Florida, the Campeche Escarpment off Yucatán, and the Sigsbee Escarpment south of Texas and Louisiana. Two major submarine canyons crease the gulf's shelf areas: the De Soto Canyon near the Florida-Alabama border, and the Campeche Canyon west of the Yucatán Peninsula. *See* CONTINENTAL MARGIN; ESCARPMENT; HOLOCENE; MARINE GEOLOGY.

Meteorology. Compared with the North American rivers, the Mexican rivers are short, but they still provide approximately 20% of the fresh-water input to the gulf because of extensive orographic rainfall from the trade winds that dominate the southern flank of the basin. Meteorologically, the Gulf of Mexico is a transition zone between the tropical wind system (easterlies) and the westerly frontal-passage-dominated weather (in winter particularly) to the north, punctuated with intense tropical storms in summer/autumn called the West Indian Hurricane. Much of the atmospheric moisture supplied to the North American heartland during spring and summer has its origin over the gulf, and thus it is a vital element in the so-called North American Monsoon. *See* HURRICANE; MONSOON METEOROLOGY; STORM SURGE; TROPICAL METEOROLOGY.

Circulation. The Gulf Stream System dominates the oceanic circulation in the Gulf of Mexico. The Yucatán Current, flowing northward into the eastern Gulf of Mexico, is the first recognizable western boundary current in the Gulf Stream System. North of the Yucatán Peninsula, the flow penetrates into the eastern gulf (where it is called the Gulf Loop Current) at varying distances with a distinctive chronology, loops around clockwise, and finally exits through the Straits of Florida, where it is called the Florida Current. This intense current reaches to more than 3300 ft (1000 m) depth, and transports 1.1×10^9 ft^3/s (3×10^7 m^3/s) of water, an amount 1800 times that of the Mississippi River. *See* GULF STREAM; MEDITERRANEAN SEA; OCEAN CIRCULATION.

Over a nearly annual cycle, the Gulf Loop Current flows between Mexico and Cuba, moving northward to the Mississippi Delta, then southward along the Florida Escarpment, and finally eastward into the Straits of Florida between Florida and Cuba. After reaching its maximum northerly position, an anticyclonic eddy, approximately 186 mi (300 km) in diameter, separates. The main flow then reforms between the Yucatan Channel and the Florida Keys, and the process starts again. *See* OCEAN CIRCULATION; SEAWATER.

Biogeography. Surrounding the Gulf of Mexico are many population centers that exploit the numerous estuaries, lagoons, and oil and gas fields. Coral reefs off Yucatán, Cuba, and Florida provide important fishery and recreational activities. There are extensive wetlands along most coastal boundaries with ecological connections to many seagrass beds nearshore and coastal mangrove forests of Mexico, Cuba, and Florida. This biogeographic confluence creates one of the most productive marine areas on Earth, providing the food web for commercially important species such as lobster, demersal (bottom-dwelling) fish, and shrimp; this same ecology supports large populations of sea turtles and marine mammals. The coastal and nearshore waters also support large phytoplankton populations.



Bottom topography and physiographic provinces in the Gulf of Mexico. Contours in meters.

The juxtaposition of these enormous marine resources and human activities has led to a distinctive anthropogenic impact on the health of the marine ecosystem. *See* BIOGEOGRAPHY; ESTUARINE OCEANOGRAPHY; FOOD WEB; MANGROVE; MARINE ECOLOGY; REEF; WETLANDS. George A. Maul

Bibliography. L. R. A. Capurro and J. L. Reid (eds.), *Contributions on the Physical Oceanography of the Gulf of Mexico*, 1972; G. A. Maul and F. M. Vukovich, The relationship between variations in the Gulf of Mexico Loop Current and Straits of Florida volume transport, *J. Phys. Oceanog.*, 23(5):785–796, 1993; H. Stommel, *The Gulf Stream*, 2d ed., 1965.

Gulf Stream

A great ocean current transporting about 7 × 10^7 tons (6.3 × 10^7 metric tons) of water per second (1000 times the discharge of the Mississippi River) northward from the latitude of Florida to the Grand Banks off Newfoundland. Before the days of scientific oceanography, it was supposed that the origin of the water in the Gulf Stream was the Gulf of Mexico. The origin has since been traced farther upstream, through the Gulf of Mexico and Caribbean, to the Great North and South Equatorial currents of the Atlantic. The Gulf Stream is thought of now as a portion of a great horizontal circulation in the ocean, where particles of water execute closed circuits, sometimes moving slowly in mid-ocean regions and other times rapidly in strong currents like the Gulf Stream. Thus the beginning and end of the Stream have arbitrary geographical limits. *See* ATLANTIC OCEAN; GULF OF MEXICO.

The Gulf Stream is a narrow (60 mi or 100 km) and swift (up to 5 knots or 250 cm/s) eastward-flowing current jet which is surrounded by intense eddies. As it leaves the coast at Cape Hatteras, the Stream becomes unstable and meanders from side to side like a river. The meanders usually move downstream with a speed of 0.2 knot (10 cm/s), becoming progressively larger; they have a particularly large amplitude near the New England seamounts near 60°W. South of New England, meanders are 60–120 mi (100–200 km) in length and width, and change position with periods of several days to several weeks. Meanders pinch off from the Stream 10–15 times per year, forming energetic eddies, typically 120 mi (200 km) in diameter (**Fig. 1**). These eddies are called rings because they consist of a ring of Gulf Stream water encircling water in the center of the eddy. Rings frequently drift westward and eventually coalesce with the Stream after a lifetime of months to years.

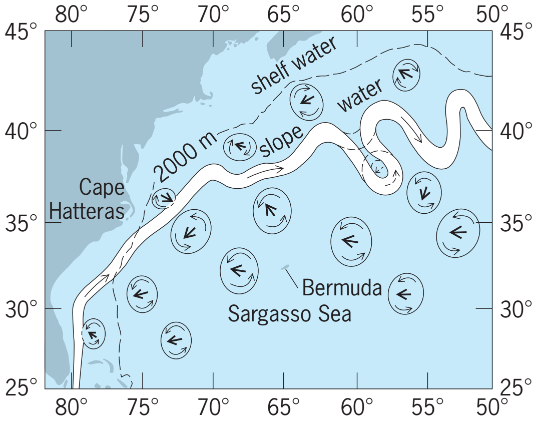


Fig. 1. The path of the Gulf Stream and the distribution and number of rings. Warm rings are formed to its north and cold rings to its south. 2000 m = 6600 ft. (*After P. L. Richardson, Gulf Stream rings, Oceanus, 19(3):65–68, 1976*)

The meanders and rings are thought to force counterrotating gyres that recirculate water from the Gulf Stream on both sides of it (**Fig. 2**).

The near-surface Gulf Stream transports warm water from southern latitudes eastward to the Grand Banks, where the flow becomes broader and weaker, separating into several branches and eddies. About half the near-surface flow continues eastward across the Mid-Atlantic Ridge, and half recirculates southwestward.

The subsurface Gulf Stream coincides with a strong horizontal temperature gradient marking the juncture of cold slope water on the north and warm Sargasso water on the south. Although the thermal gradient extends deeper than 6600 ft (2000 m) under the Stream, the deep current structure is complex and is not a simple downward extension of the near-surface flow. The mean velocity of the Gulf Stream decreases with depth and extends to the sea floor even in depths of 4000 m. The deep Gulf Stream gyre including most of the recirculation is located west of the Mid-Atlantic Ridge and north of Bermuda. *See* MID-OCEANIC RIDGE.

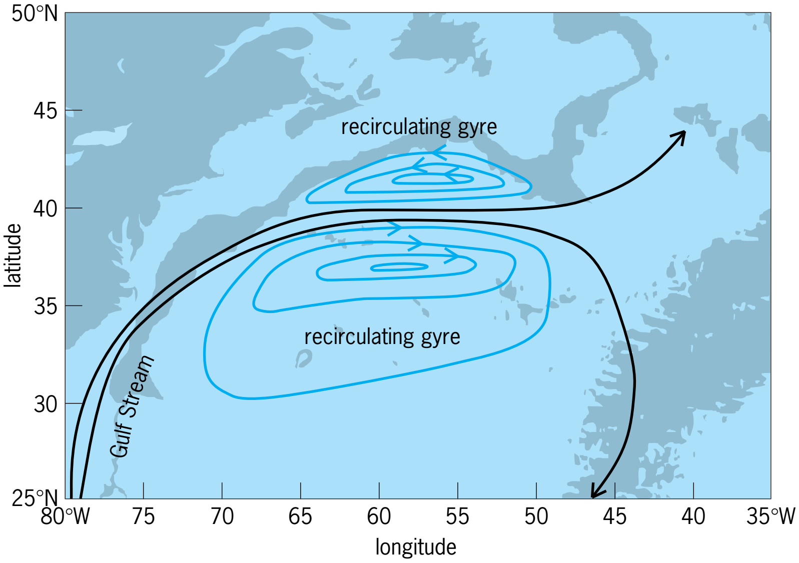


Fig. 2. Average transport of the Gulf Stream and recirculating gyres. Each line represents approximately 15 Sv of transport. (*After N. G. Hogg, On the transport of the Gulf Stream between Cape Hatteras and the Grand Banks, Deep-Sea Res., 39:1231–1246, 1992*)

The volume transport of the Gulf Stream increases from 30 sverdrups (1 Sv = 10^6 m^3/s or 3.5 × 10^7 ft^3/s) off Florida to 150 Sv north of Bermuda, and then decreases eastward (Fig. 3). A large part of the increase in transport is caused by eastward components of the recirculating gyre which extend deeply to the sea floor.

The Gulf Stream is driven predominantly by the large-scale wind pattern, the westerlies in the north and the trades in the south. The winds exert a torque on the ocean that, because of the shape and rotation of the Earth, causes a large western-intensified gyre that transports 30 Sv. Part of the Atlantic circulation is driven by the large-scale density differences caused mainly by heating in low latitudes and cooling in high latitudes. Cold, deep water is formed in northern seas and flows southward as a deep western boundary current transporting roughly 15 Sv; warm upper layer water flows northward and replaces it. This flow is called the meridional overturning circulation. The deep western boundary current which augments the westward-flowing portion of the northern recirculating gyre encounters the Gulf Stream off Cape Hatteras, North Carolina. Some of the deeper water in the deep boundary current continues southward along the boundary passing under the mean axis of the Stream. The rest of the deeper water is entrained into the Gulf Stream, some to eventually cross the mean axis of the Stream, recirculate westward toward the western boundary, and continue southward. Energetic fluctuation of the deep currents and the Gulf Stream make it much more complex than shown in schematics which attempt to portray the mean flow. *See* OCEAN CIRCULATION; WIND STRESS.

Almost half of the Gulf Stream transport in the Straits of Florida originates in the South Atlantic. The rest comes from the large-scale recirculation or flow in the subtropical gyre of the North Atlantic. Near the Grand Banks of Newfoundland, the Gulf Stream divides into several branches. The first branch flows northward as a western boundary current east of Newfoundland, turns eastward near 50°N, and flows across the Mid-Atlantic Ridge as the North Atlantic Current. The second branch flows eastward across the Mid-Atlantic Ridge south of the Azores, and is concentrated into the relatively swift eastward-flowing Azores Current centered near 34°N. The third branch recirculates water south of the Gulf Stream.

Philip Richardson

Bibliography. W. E. Johns et al., Gulf Stream structure, transport, and recirculation near 68°W, *J. Geophys. Res.*, 1995; A. R. Robinson (ed.), *Eddies in Marine Science*, 1983; W. J. Schmitz, Jr., and M. S. McCartney, On the North Atlantic circulation, *Rev. Geophys.*, 1993; H. Stommel, *The Gulf Stream*: *A Physical and Dynamical Description*, 2d ed., 1965; B. A. Warren and C. Wunsch (eds.), *Evolution of Physical Oceanography*, 1981; L. V. Worthington, *On the North Atlantic Circulation*, Johns Hopkins Oceanographic Series, vol. 6, 1976.

Gum

A class of high-molecular-weight molecules, usually with colloidal properties, which in an appropriate solvent or swelling agent are able to produce gels (highly viscous suspensions or solutions) at low dry-substance content. The molecules are either hydrophilic or hydrophobic; that is, they do or do not have an affinity for water. The term gum is applied to a wide variety of substances of gummy characteristics, and therefore cannot be precisely defined.

Various rubbers are considered to be gums, as are many synthetic polymers, high-molecular-weight hydrocarbons, or other petroleum products. Chicle for chewing gum is an example of a hydrophobic polymer which is termed a gum but is not frequently classified among the gums. Quite often listed among the gums are the hydrophobic resinous saps that often exude from plants and are commercially tapped in balsam (gum balsam) and other evergreen trees (gum resin). Incense gums such as myrrh and frankincense are likewise fragrant plant exudates.

Usually, however, the term gum, as technically employed in industry, refers to plant polysaccharides or their derivatives. These are dispersible in either cold or hot water to produce viscous mixtures. Modern usage of the term includes water-soluble derivatives of cellulose and derivatives and modifications of other polysaccharides which in the natural form are insoluble. Usage, therefore, also includes with gums the ill-defined group of plant slimes called mucilages. *See* CELLULOSE; COLLOID; POLYSACCHARIDE.

Viscosity. Gums are important in that they impart viscosity to aqueous solutions. Their physical properties are manifestations of their chemical structure, the kind and amount of solvent, and the kind and concentration of ions and other substances dissolved in the solvent. Because gums are commonly composed of several different kinds of monomer units with many possible variations in regard to degree of branching, length of branches, and types of linkages, an almost infinite number of structures is possible. Forces act between molecules, between different parts of the same molecule, and between polymer and solvent. These forces include hydrogen bonding, ionic charges, dipole and induced dipole interactions, and van der Waals forces.

All of these forces affect such properties as gel-forming tendency, viscosity, and adhesiveness. The types of linkage due to their effects on chain flexibility are important also in determining physical properties. For example, it is known that linear molecules make more viscous solutions than do long-branch molecules of similar molecular weights, but they have a tendency to precipitate because of association of the chains. If this association is prevented, stability can be achieved without much sacrifice of viscosity. This can be done by introducing groups with ionic charges that repel one another, or by attaching many very short side branches to prevent

Sources and sugars of important gums

Gum	Source	Sugars present and linkages
Seaweeds		
Agar	Red algae (*Gelidium* sp.)	D-Galactose β-(1 → 4), 3,6-anhydro-L-galactose α-(1 → 3) + sulfate acid ester groups
Algin	Brown algae (*Macrocystis pyrifera*)	D-Mannuronic acid β-(1 → 4), L-guluronic acid β-(1 → 4), Na salt
Carrageenan	Red algae (*Chondrus crispus, Gigartina stellata*)	D-Galactose, 3,6-anhydro-D-galactose + sulfate acid ester groups
Fucoidan	Brown algae (*Fucus* sp., *Laminaria* sp.)	L-Fucose + sulfate acid ester groups
Laminaran	Brown algae (*Laminaria* sp.)	D-Glucose, D-mannitol, β-(1 → 3) chain and β-(1 → 6) branches
Plant exudates		
Gum arabic	*Acacia* sp.	L-Arabinose, D-galactose, L-rhamnose, D-glucuronic acid
Ghatti	*Anogeissus latifolia*	L-Arabinose, D-xylose, D-galactose, D-mannose, D-glucuronic acid
Karaya	*Sterculia urens*	D-Galactose, D-rhamnose, D-galacturonic acid
Tragacanth	*Astragalus* sp.	D-Galactose, D-xylose, D-glucuronic acid
Plant extracts		
Pectin	Cell walls and intracellular spaces of all plants (commercial source, citrus waste)	D-Galacturonic acid α-(1 → 4) partially esterified, L-arabinose α-(1 → 5) and α-(1 → 3) branches, D-galactose β-(1 → 4)
Larch arabinogalactan	Western larch	D-Galactose, L-arabinose
Ti	Tubers of *Cordyline terminalis*	D-Fructose, D-glucose
Plant seeds		
Corn-hull gum	Corn seed coat	D-Xylose, L-arabinose, D-galactose, L-galactose, D-glucuronic acid
Guar	*Camposia teragonolobus* endosperm	D-Mannose β-(1 → 4), D-galactose α-(1 → 6) branches
Locust bean	Carob tree (*Ceratonia siliqua*) endosperm	D-Mannose β-(1 → 4), D-galactose α-(1 → 6) branches
Quince seed	*Cydonia vulgaris*	L-Arabinose, D-xylose, hexuronic acid, monomethyl hexuronic acid
Psyllium seed	*Plantago* sp.	D-Xylose, L-arabinose, D-galacturonic acid, L-rhamnose, D-galactose
Flax seed	*Linum usitatissimum*	D-Galacturonic acid, D-xylose, L-rhamnose, L-arabinose, L-galactose, D-glucose
Tamarind	*Tamarindus indica*	D-Glucose, D-galactose, D-xylose
Wheat gum	Wheat	D-Xylose β-(1 → 4), L-arabinose branches
Miscellaneous		
Cellulose derivatives	Plant cell walls, wood pulp, and cotton	D-Glucose β-(1 → 4)
Starch	Cereal grains and tubers	D-Glucose α-(1 → 4), D-glucose α-(1 → 6) at branch points
Amylose		
Amylopectin		
Dextran	Bacterial action on sucrose	D-Glucose α-(1 → 6) and α-(1 → 3)
Chitin	Exoskeleton of animals of the phylum Arthropoda	*N*-Acetyl-D-glucosamine β-(1 → 4)

close approach of the chains. Much work remains to be done in this area. While it is known that solutions of some gums are slimy or mucilaginous, whereas those of other gums are tacky, the reasons for these differences are unknown. Rheological properties of the different gum solutions also differ. *See* HYDROGEN BOND; POLARIZATION OF DIELECTRICS; RHEOLOGY.

Source and structure. Many plant gums originally used by humans still are important items of commerce. Most of these are dried exudates from trees and shrubs, produced with or without artificial stimulation by injury. These gums are collected by hand, usually in the hot, semiarid regions of the world, often from wild plants, but sometimes from cultivated plants. For the most part, these gums have highly branched structures containing two to five different sugar units. They frequently are acidic in that they contain carboxyl groups which are portions of glycuronic acids. Commercial gums often are mixtures of two or more different polysaccharides.

Seed gums, which also have been used for many centuries, likewise are important items of commerce today. The more ancient gums were extracted from quince, psyllium, or flax seeds. The ground endosperm of locust trees grown in the Mediterranean area also has been an item of commerce for many years. It is still important today, along with a similar gum obtained from the endosperm of the guar plant which is native to India and Pakistan but is grown to a small extent in the southwestern United States.

Seaweed gums produced by hot water or alkaline extraction of seaweed are other important industrial raw materials.

Industrially important or potentially useful gums are listed in the **table**, together with the plant source, the sugar units, and the most prevalent glycosidic linkage when known.

Use. Gums are used in foods as stabilizers and thickeners. They form viscous solutions which prevent aggregation of the small particles of the dispersed phase. In this way they aid in keeping solids dispersed in chocolate milk, air in whipping cream, and fats in salad dressings. Gum solutions also retard crystal growth in ice cream (ice crystals) and in confections (sugar crystals). Their thickening and stabilizing properties make them useful in water-base paints, printing inks, and drilling muds. Because of these properties they also are used in cosmetics and pharmaceuticals as emulsifiers or bases for ointments, greaseless creams, toothpastes, lotions, demulcents, and emollients. Gums are used to modify texture and increase moisture-holding capacity, for example, as gelling agents in canned meats or fish, marshmallows, jellied candies, and fruit jellies. Their adhesive properties make them useful in the production of cardboard, postage stamps, gummed envelopes, and as pill binders. Other applications include the production of dental impression molds, fibers (alginate rayon), soluble surgery films and gauze, blood anticoagulants, plasma extenders, beverage-clarifying agents, bacteriological culture media, half-cell bridges, and tungsten-wire-drawing lubricants. *See* FOOD MANUFACTURING.

Roy L. Whistler

Bibliography. A. A. Lawrence, *Natural Gums for Edible Purposes*, 1977; C. L. Mantell, *The Water Soluble Gums*, 1947, reprint 1965; W. Pigman and M. L. Wolfrom (eds.), *Advances in Carbohydrate Chemistry*, vol. 13, 1958; K. R. Stauffer, *Handbook of Edible Gums,* 1988; R. L. Whistler and J. N. BeMiller (eds.), *Industrial Gums*: *Polysaccharides and Their Derivatives*, 3d ed., 1993; R. L. Whistler and M. L. Wolfrom (eds.), *Methods in Carbohydrate Chemistry*, vol. 5: *General Polysaccharides*, 1965.

Gunsights

Optical instruments which establish an optical line or axis for the purpose of aiming a weapon. The axis includes the observer's eye, a suitable mark in the instrument, and the target. Most gunsights employ as their basis a telescope or a partially reflective mirror.

Rifle sights. A typical rifle sight (**Fig. 1**) consists of a terrestrial telescopic system having an objective, an eyepiece, an erector lens, and a reticle. Sometimes a field lens is employed to ensure uniform illumination. The magnifying powers customarily range from unity to 12 diameters, magnifications of $2^1/_2$ to 9 being most common. Instruments are available in which the power may be varied optically over a wide range to suit existing conditions. Large eyepieces are used to provide a wide field of view and a substantial eye relief, the latter so that the eyepiece will not strike the user's eye in recoil. In order to adjust for elevation or windage, either the reticle may be moved, usually by means of screws, or the entire sight may be tilted relative to the gun barrel. Machine-gun sights are also telescopic in nature and customarily employ a prism for erecting the image. Magnifying powers range between $1^1/_2$ and 3, and the reticle may be illuminated by a small lamp to permit night use. *See* EYEPIECE; TELESCOPE.

Aircraft gunsights. These are usually of the reflector type (also known as reflex sights) and employ in their simplest form a lamp, a reticle, a collimating lens, and a glass plate or partially reflecting mirror (**Fig. 2**). The collimator images the reticle pattern at infinity, and the mirror superimposes this image over the target area. The collimator's pupil is so large that the eye can be positioned at any point within a large cylindrical volume in back of the sight. The collimating system may employ a simple spherical mirror, a Schmidt lens, or a Mangin mirror. A Mangin mirror consists of a negative meniscus lens, the shallower

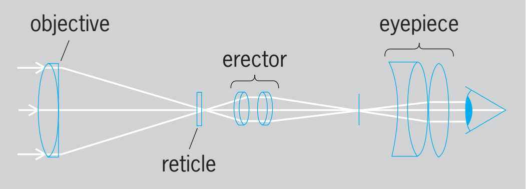


Fig. 1. Rifle sight.

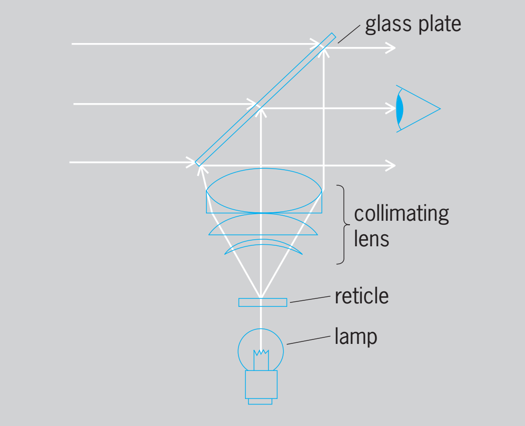


Fig. 2. Reflex sight.

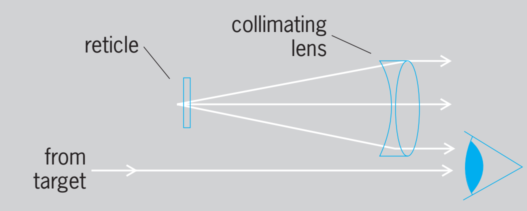


Fig. 3. Collimator sight.

surface of which is silvered to act as a spherical mirror, while the other surface corrects for the spherical aberration of the reflecting surface. The magnification of this type of sight is unity. *See* MIRROR OPTICS.

Artillery sights. These can assume various forms, the simplest of which is the collimator sight (**Fig. 3**), consisting of an objective having a reticle at its focus. When the eye is so placed as to receive light simultaneously from the target and the reticle, the latter appears superimposed on the former and a line of sight is established. Other forms of artillery sights include the elbow telescope, which is a conventional terrestrial telescope containing a prism for producing a right-angle bend, and the panoramic sight. For a discussion of panoramic sights *see* LENS (OPTICS); PERISCOPE. Edward K. Kaprelian

Bibliography. R. Kingslake, *Optical System Design*, 1983; Optical Society of America, *Handbook of Optics*, 2d ed., 1994; W. J. Smith, *Modern Optical Engineering*, 3d ed., 2000; B. H. Walker, *Optical Engineering Fundamentals*, 1995.

Gymnolaemata

A class of predominantly marine, colonial animals belonging to the phylum Bryozoa. Gymnolaemate zooids (individual colony members) are characterized by their short, wide, vaselike or boxlike zooecia (skeletal outer structures) and circular lophophores (food-gathering organs).

Classification and history. The gymnolaemates include several thousand species—mostly marine, some brackish, and a very few freshwater—belonging to the orders Ctenostomata and Cheilostomata. Appearing early in the Early Ordovician, when they included the probable stem group of the Ectoprocta, gymnolaemates remained quite inconspicuous until the Cretaceous, when they rose to the position of dominance among bryozoans which they still maintain. *See* BRYOZOA; CHEILOSTOMATA; CTENOSTOMATA.

Morphology. Highly varied in size and shape, most gymnolaemate colonies are small and delicate, but a few are large, conspicuous growths. Characteristics of these colonies remain essentially uniform throughout their extent (that is, they are not divisible into distinct endozone and exozone regions). The individual zooids may be relatively isolated, loosely grouped side by side, or tightly packed together.

The boxlike gymnolaemate zooecia lack internal calcareous cross-partitions like diaphragms. The zooecial wall may be a thin chitinous membrane (rarely coated externally with gelatinous material), a moderately thick and firm chitinous wall, or a thin to thick calcareous wall (covered externally by a thin chitinous cuticle). Just inside the zooecial wall is an epidermis, and immediately interior to that is a peritoneum, both composed of flat, thin cells; however, the body wall lacks a muscle layer. The zooecial aperture may be as wide as, but is often narrower than, the zooecium below.

An epistome (flap covering the mouth) is not present. The visceral cavities of adjacent zooids are not directly interconnected. The lophophore bears few to moderate numbers of tentacles (8–34, where counted). Individual gymnolaemate zooids are usually hermaphroditic, sometimes male or sometimes female. Individual colonies are always hermaphroditic, usually because they bear hermaphroditic zooids but sometimes because they bear both male and female zooids.

Life cycle. Asexual budding, which produces a new zooid within a gymnolaemate colony, begins by separating off a small sac- or boxlike portion of the parent zooid's visceral or body cavity. Cells along the inside of the body wall (cystids) then develop into a new set of internal soft-part organs (polypides) suspended in the newly separated sac or box. Periodically, the polypides of gymnolaemate zooids degenerate into brown bodies; the new succeeding polypides are regenerated from the original zooids' body walls. Only the very few freshwater gymnolaemates produce resistant resting bodies, known as hibernacula. Roger J. Cuffey

Bibliography. A. H. Cheetham and P. L. Cook, General features of the class Gymnolaemata, pp. 138–207 in R. A. Robison (ed.), *Treatise on Invertebrate Paleontology*, part G, *Bryozoa*, revised, vol. 1, Geological Society of America & University of Kansas, Boulder & Lawrence, 1983; R. J. Cuffey and J. E. Utgaard, *Bryozoans*, pp. 204–216, vol. 1, in R. Singer (ed.), *Encyclopedia of Paleontology*, Fitzroy Dearborn, Chicago, 1999; F. K. McKinney and J. B. C. Jackson, *Bryozoan Evolution*, Unwin Hyman, Boston, 1989; J. S. Ryland, *Bryozoans*, Hutchinson University Library, London, esp. pp. 27, 29, 43–85, 1970; P. D. Taylor, *Bryozoa*, pp. 465–488 in M. J.

Benton (ed.), *The Fossil Record 2*, Chapman & Hall, London, 1993.

Gymnostomatida

An order of the Holotrichia which contains a large, widely distributed group of what are believed to be the most primitive ciliate protozoa. These organisms occur abundantly in sands of intertidal zones, as well as in the more usual fresh- and salt-water habitats. The body size is frequently large; ciliation is simple and plentiful; and the oral area lacks buccal ciliature, as the name of the order implies. The cytopharynx, however, is reinforced by fibrillar rods known as trichites, or nemadesmata. Members of two families exist as harmless commensals in such herbivores as horses and camels and in a few other mammals. *Prorodon*, *Holophrya*, *Didinium*, and *Dileptus* are examples of carnivorous gymnostomes. *Chilodonella* and *Nassula* are herbivorous and anatomically more complex. *Ichthyophthirius*, a fish parasite, is often classified erroneously as a gymnostome, but belongs in the order Hymenostomatida. It is found on fish in home aquaria and causes the disease known as the ich. *See* CILIOPHORA; HOLOTRICHIA; HYMENOSTOMATIDA; PROTOZOA.

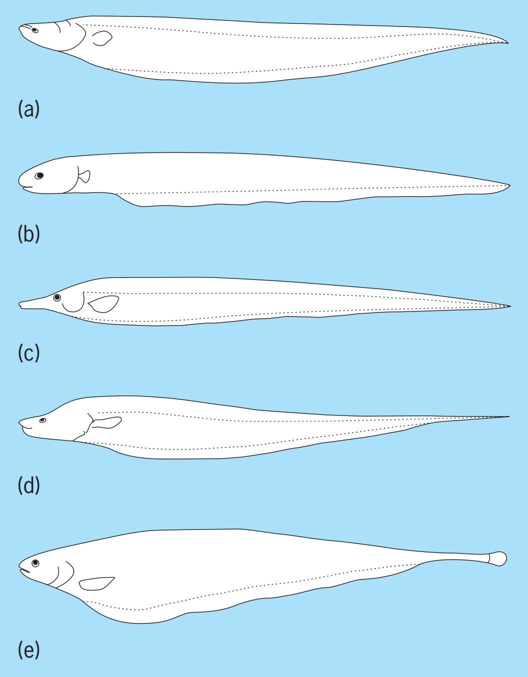
John O. Corliss

Gymnotiformes

An order of teleost fishes in the superorder Ostariophysi, series Otophysi, characterized by a combination of the following characters: Weberian ossicles (a series of small bones that form a chain connecting the swim bladder with the inner ear); an eel-like body, usually compressed but sometimes rounded; absence of a dorsal fin and absence of a pelvic girdle and pelvic fins; a very long anal fin; a caudal fin that is absent or very small; and electric organs. The anal rays are articulated in a way that allows for circular motion; thus undulation of the anal fin can move the gymnotiform forward or backward. The order Gymnotiformes, commonly known as American knifefishes, comprise five families (see **illustration**) described below. *See* ELECTRIC ORGAN (BIOLOGY); OSTARIOPHYSI; SWIM BLADDER.

Gymnotidae (nakedback knifefishes). Nakedback knifefishes are represented by two genera. The principal genus, *Gymnotus*, with about 32 species, ranges in freshwater through southern Mexico, Central America, South America, to Argentina, as well as Trinidad. It has a superior mouth, small scales, and a weak electrical discharge, and attains a total length of 60 cm (24 in.).

The other genus, *Electrophorus*, with one species, *E. electricus* (electric knifefish, commonly known as the electric eel), was formerly in the family Electrophoridae. It occurs primarily in the Orinoco and Amazon river basins, and differs from *Gymnotus* in having a subterminal mouth, no scales, and a strong electrical discharge (up to 600 volts). It attains a total length of 230 cm (90 in.). Gymnotids are nocturnal fishes, usually occurring in quiet waters from deep rivers to swamps. *See* EELS.



Gymnotiform fishes. (*a*) Nakedback knifefish. (*b*) Electric knifefish. (*c*) Sand knifefish. (*d*) Glass knifefish. (*e*) Ghost knifefish. (*From J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006*)

Rhamphichthyidae (sand knifefishes). Sand knifefishes, with three genera and 12 species, are known only from freshwaters of South America. The following combination of characters distinguishes them from other gymnotids: an elongated compressed body, no caudal fin, a long tubular snout, nostrils close together, no teeth on lower jaws, and myogenic (muscular) electric organs. They are nocturnal fishes that burrow in the sand during daylight hours. The maximum recorded length is 100 cm (39 in.).

Hypopomidae (bluntnose knifefishes). Hypopomids range from the Río de la Plata of Argentina (35°S) to the Río Tuira of Panama (8°N) and are known from all South American countries except Chile. They have their greatest diversity in the Amazon River basin. The family, consisting of about 35 species in six genera, is distinguished from other gymnotiforms by the absence of teeth from both jaws, a moderate to short snout length, well-separated nostrils, small eyes, and no caudal fin. The smallest species attains a total length of only 8 cm (3 in.), and the largest a moderate length of 35 cm (14 in.).

Sternopygidae (glass knifefishes). The range of glass knifefishes is essentially the same as that of the bluntnose knifefishes. The family comprises six generia and 30 valid species, with 11 more species assigned manuscript names and awaiting descriptions. The family possesses the following unique combination of characters among gymnotiforms: multiple rows of small villiform (brushlike) teeth in both jaws, large eyes [diameter equal to or greater than the distance between the nares (nostrils)], large infraorbital bones with expanded bony arches, anterior nares located outside the gape, no urogenital papilla, and no caudal fin. Some sternopygid species are specialized for life in main river channels, and many species are highly compressed laterally and translucent. Sternopygid species are medium- to large-sized, varying in total length from 12 to 140 cm (4.7 to 55 in.).

Apteronotidae (ghost knifefishes). The range of ghost knifefishes is essentially the same as that of the bluntnose knifefishes and glass knifefishes. Apteronotids are the only gymnotiform fishes with a caudal fin and a dorsal organ (a longitudinal strip of fleshy tissue firmly attached to the posterodorsal midline). Apteronotid species are further characterized by one or two rows of conical teeth in both jaws, infraorbital bones ossified as slender tubes, anterior nares located outside the gape, and no urogenital papilla. Apteronotids also possess a high-frequency tone-type electric organ discharge (more than 750 Hz at maturity). There are 52 valid species in 13 genera, and 12 additional species assigned manuscript names and awaiting descriptions. Apteronotid species range in total length from 16 to 130 cm (6 to 51 in.). Herbert Borschung

Bibliography. J. S. Albert, *Species Diversity and Phylogenetic Systematics of American Knifefish (Gymnotiformes, Teleostei)*, Misc. Publ. Mus. Zool. Univ. Mich. 190, 2001; J. S. Albert, Sternopygidae (Glass knifefishes, Rattail knifefishes), pp. 487–491, and Hypopomidae (Bluntnose knifefishes), pp. 494–496 in R. E. Reis, S. O. Kullander, and C. J. Ferraris, Jr. (eds.), *Checklist of the Freshwater Fishes of South and Central America*, Edipucrs (Editoria Universitaria da PUCRS), Porto Alegre, Brasil, 2003; J. S. Albert and R. Campos-da-Paz, Phylogenetic systematics of Gymnotiformes with diagnoses of 58 clades: A review of the available data, pp. 419–446 in L. R. Malabarba et al. (eds.), *Phylogeny and Classification of Neotropical Fishes*, Edipucrs, 1998; J. S. Albert and W. G. R. Crampton, Family Hypopomidae (bluntnose knifefishes), pp. 494–496 in R. E. Reis et al., (eds.), *Checklist of the Freshwater Fishes of South and Central America*, Edipucrs, 2003; C. J. Ferraris, Jr., Family Rhamphichthyidae (sand knifefishes), pp. 492–493 in R. E. Reis et al. (eds.), *Checklist of the Freshwater Fishes of South and Central America*, Edipucrs, 2003; R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, www.fishbase.org, version 05/2005; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Gypsum

The most common sulfate mineral, characterized by the chemical formula $CaSO_4 \cdot 2H_2O$; it shows little variation from this composition.

Gypsum is one of the several evaporite minerals. This mineral group includes chlorides, carbonates, borates, nitrates, and sulfates. These minerals precipitate in seas, lakes, caves, and salt flats due to concentration of ions by evaporation. When heated or subjected to solutions with very large salinities, gypsum converts to bassanite ($CaSO_4 \cdot H_2O$) or anhydrite ($CaSO_4$). Under equilibrium conditions, this conversion to anhydrite is direct. The conversion occurs above 108°F (42°C) in pure water. The presence of halite (NaCl) or other sulfates in the solution lowers this temperature, although metastable gypsum exists at higher temperatures. *See* ANHYDRITE; HALITE.

Crystals of gypsum are commonly tabular, diamond-shaped, or lenticular; swallow-tailed twins are also common. The mineral is monoclinic with symmetry $2/m$. The common colors displayed are white, gray, brown, yellow, and clear. Cleavage surfaces show a pearly to vitreous luster. Gypsum is the index mineral chosen for hardness 2 on Mohs scale with a specific gravity of 2.32. In addition to free crystals, the common forms of gypsum are satin spar (fibrous), alabaster (finely crystalline), and selenite (massive crystalline).

Gypsum is used for a variety of purposes, but chiefly in the manufacture of plaster of paris, in the production of wallboard, in agriculture to loosen clay-rich soils, and in the manufacture of fertilizer. Plaster of paris is made by heating gypsum to 392°F (200°C) in air. A hemihydrate is formed as part of the water of crystallization is driven off. Later, when water is added, rehydration occurs. The interlocking, finely crystalline texture that results forms a uniform hardened mass. The slightly increased volume of the set plaster serves to fill the mold into which it has been poured. *See* PLASTER OF PARIS.

Gypsum deposits play an important role in the petroleum industry. The organic material commonly associated with its formation is considered a source of hydrocarbon (oil and gas) generation. In addition, these deposits act as a seal for many petroleum reservoirs, preventing the escape of gas and oil. *See* PETROLEUM.

Because most gypsum deposits form by evaporation from solution, they are indicators of former arid and semiarid climates. The worldwide distribution of gypsum is an indicator of climatic change through geologic time. Large-scale gypsum deposits are found in strata of nearly all ages, the oldest recognized being 3.4×10^9 years old (North Pole, Pilbara Block, Australia). The largest known deposits formed in Silurian (400×10^6 years ago), Pennsylvanian-Permian (295–245×10^6 years), Jurassic (200×10^6 years), and Miocene (5×10^6 years) time.

Gypsum deposits are mined throughout the world, with the United States being a world leader in gypsum production. The majority of United States gypsum is mined in Michigan, Iowa, Texas, California, and Oklahoma. Canada is the world's second largest producer. Most Canadian production is in the province of Nova Scotia. Among the other leading producers are France, Japan, Iran, Russia, Italy, Spain, and the United Kingdom.

Marc L. Helman; Charlotte Schreiber

Bibliography. American Society for Testing and Materials, *Annual Book of ASTM Standards*, vol. 04.01: *Cement; Lime; Gypsum*, 1993; Bureau

of Mines–U.S. Department of the Interior, *Minerals Yearbook*, centennial ed., vol. 1: *Metals and Minerals*, 1982; L. L. Y. Chang, R.A. Howie and J. Zussman, *Rock-Forming Minerals*, vol. 5B: *Non-silicates: Sulphates, Carbonates, Halides*, 2d ed., 1996; C. J. Dixon, *Atlas of Economic Mineral Deposits*, 1979; C. Klein and C. S. Hurlbut, Jr., *Manual of Mineralogy,* 21st ed., 1993, revised 1998; C. C. Plummer and D. McGeary, *Physical Geology,* 8th ed., 1999.

Gyrator

A linear, passive, two-port electric circuit element whose transmission properties are such that it is effectively a half wavelength longer for one direction of transmission than for the other direction of transmission. Thus a gyrator is a device that causes a reversal of signal polarity for one direction of propagation but not for the other. (A two-port element has a pair of input terminals and a pair of output terminals.) This device is novel, since it violates the theorem of reciprocity. *See* RECIPROCITY PRINCIPLE.

Until the early 1950s, all known linear passive electrical networks obeyed the theorem of reciprocity. However, several different types of nonreciprocal networks are now widely applied, principally at microwave frequencies. These devices are used to control the direction of signal flow and to protect or isolate components from undesired signals. One common application of a three-port nonreciprocal network, called a circulator, is to permit connection of a transmitter and a receiver to the same antenna. This is accomplished with minimum interference and virtually no power loss of either transmitted or received signal. *See* CONTINUOUS-WAVE RADAR; NETWORK THEORY.

Reciprocity. The theorem of reciprocity can be stated in many equivalent forms, and perhaps the simplest to consider and understand is the particular form it takes when it is expressed especially for a two-port microwave network, as shown in **Fig. 1***a*. Here d_1 represents a complex signal wave propagating toward the network input port, r_1 represents the wave propagating away from the network on the same side, and d_2 and r_2 represent the same quantities on the other side. In a linear network the relationship between the waves can be expressed as shown in Eqs. (1). Thus r_1 consists of a part which

$$\begin{aligned} r_1 &= S_{11}d_1 + S_{12}d_2 \\ r_2 &= S_{21}d_1 + S_{22}d_2 \end{aligned} \qquad (1)$$

arises from the partial reflection of d_1 at the input port and a part contributed by the fraction of d_2 which is transmitted through the network. The reflection and transmission properties of the network are described by the complex coefficients S_{ij}, which are termed the scattering parameters of the network.

The network can be characterized by selectively setting one of the input waves to zero and measuring the resulting response. Thus when d_2 is zero, Eq. (2)

$$S_{21} = \frac{r_2}{d_1} \qquad (2)$$

holds. The quantity S_{21} gives the amplitude and phase of the wave emerging from the right-hand side of the network when a unit wave, with zero phase angle, is incident on the left-hand side. The significance of the other scattering parameters can be deduced by similar arguments. If there is no reflection of the incident waves, then S_{11} and S_{22} are zero, and the network is said to be matched.

The theorem of reciprocity is stated by Eq. (3).

$$S_{12} = S_{21} \qquad (3)$$

That is, the network has the same transfer characteristics for one direction of propagation as it has for the other. Thus, if a matched network causes a particular insertion loss and phase shift for signals transmitted from left to right, signals transmitted from right to left must suffer the same loss and phase shift. An ideal gyrator is one which has the scattering matrix given in Eq. (4), and hence violates reciprocity.

$$[S] = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} 0 & -S_{21} \\ S_{21} & 0 \end{bmatrix} \qquad (4)$$

The terminal currents i_1, and i_2 and voltages of V_1, and V_2 of the network (Fig. 1*b*) are related by impedances Z_{ij} as shown in Eqs. (5). The theorem of

$$\begin{aligned} V_1 &= Z_{11}i_1 + Z_{12}i_2 \\ V_2 &= Z_{21}i_1 + Z_{22}i_2 \end{aligned} \qquad (5)$$

reciprocity now requires the condition of Eq. (6).

$$Z_{12} = Z_{21} \qquad (6)$$

Theoretical gyrators. The first comprehensive treatment of nonreciprocal two-port networks was given in 1948 by B. D. H. Tellegen, who applied the word gyrator to describe such networks. Tellegen restricted his analysis to an ideal gyrator whose impedance matrix has the form given in Eq. (7). It

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} 0 & -R \\ R & 0 \end{bmatrix} \qquad (7)$$

may be easily shown that for such a gyrator to be

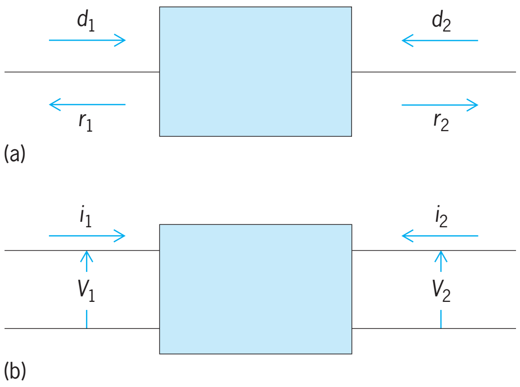


Fig. 1. Two-port microwave network. (*a*) Scattering matrix representation. (*b*) Currents and voltages.

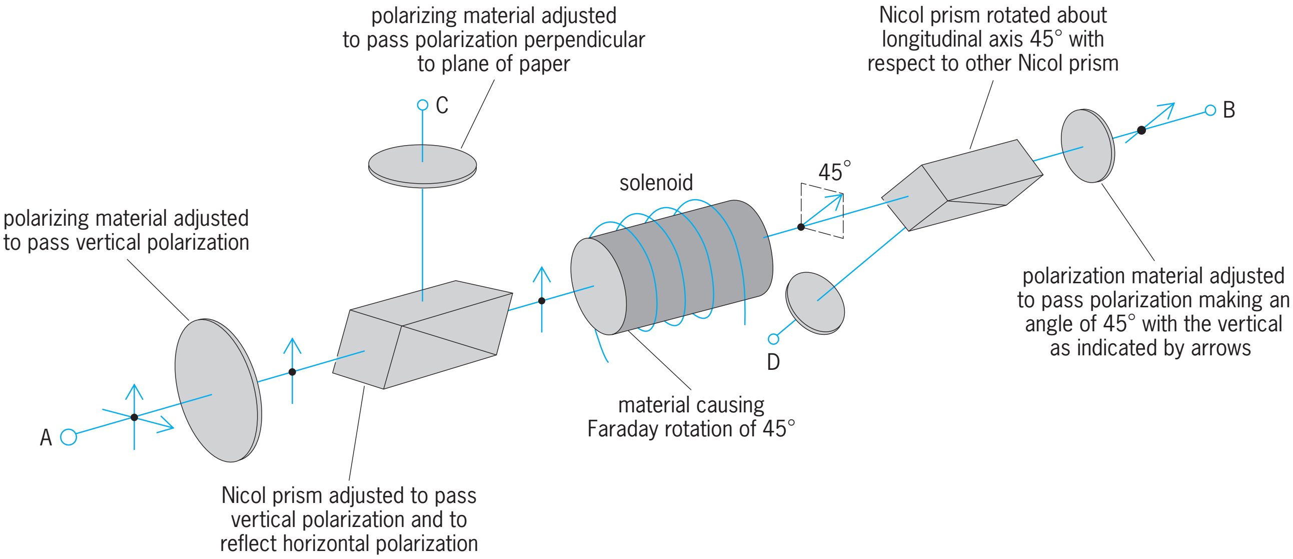

polarizing material adjusted
to pass polarization perpendicular
to plane of paper
Nicol prism rotated about
longitudinal axis 45° with
respect to other Nicol prism
C
B
45°
solenoid
polarizing material adjusted
to pass vertical polarization
polarization material adjusted
to pass polarization making an
angle of 45° with the vertical
as indicated by arrows
D
material causing
Faraday rotation of 45°
A
Nicol prism adjusted to pass
vertical polarization and to
reflect horizontal polarization

Fig. 2. Passive nonreciprocal optical device based on Faraday rotation.

nondissipative, the Z_{ij} must be real. However, it is not necessary that the diagonal terms, Z_{11} and Z_{22}, be zero. Furthermore, this restriction on the diagonal terms is not needed in order to realize two of the most important properties of the gyrator, namely, the construction of one-way transmission systems and the phase inversion of signal polarity for one direction of transmission.

Tellegen's ideal gyrator has the additional property of impedance inversion. If the gyrator is terminated by an impedance Z_L, Eq. (8), the input impedance

$$Z_L = \frac{V_2}{i_2} \tag{8}$$

Z_{in} of the gyrator for Eq. (5) is given by Eq. (9). This property of impedance inversion is not, how-

$$Z_{\mathrm{in}} = \frac{V_1}{i_1} = \frac{R^2}{Z_L} \tag{9}$$

ever, unique to nonreciprocal networks. Indeed, if one makes the same requirement on a nondissipative reciprocal two-port, that $Z_{11} = Z_{22} = 0$. the input impedance Z_{in} of the terminated network is that given by Eq. (10), where X_{12} is the transfer reac-

$$Z_{\mathrm{in}} = \frac{V_1}{i_1} = \frac{X_{12}^2}{Z_L} \tag{10}$$

tance of a nondissipative network, $Z_{12} = jX_{12}$. Such a reciprocal impedance-inverting network can be easily realized.

Practical gyrators. The theorem of reciprocity has in the past been considered so universally valid that present-day textbooks still make the statement that if the condition stated in Eq. (6) is valid, the two-port is passive, and that if the condition does not hold for a particular network, it cannot be a passive one.

Although reciprocity is as universally valid in mechanical, acoustical, or optical systems as it is in electrical ones, there are passive systems that can be constructed in each of these areas which are nonreciprocal. For example, it has been long known that a mechanical system which contains a gyroscopic coupler does not obey the theorem of reciprocity.

Perhaps the first passive nonreciprocal system was an optical one proposed by Lord Rayleigh, making use of the rotation of the plane of polarization of light when it passed through a transparent material in the presence of a magnetic field. This phenomenon is called Faraday rotation. If polarized light is propagated through a transparent medium along the direction of the magnetic field, the plane of polarization of the light is rotated through some angle θ per unit length, which is determined by the properties of the medium and the strength of the magnetic field. Faraday rotation is unusual in that it is nonreciprocal. Thus the sense (clockwise or counterclockwise) of the rotation is the same whether the light travels parallel to the applied magnetic field direction or contraparallel to it. Hence, if the plane of polarization is rotated through an angle θ in traversing the Faraday cell and the ray is reflected back through the cell toward its source, it will again be rotated through an angle θ, so that when it arrives back at the source, the plane of polarization will have been rotated through a total angle 2θ.

Lord Rayleigh's one-way system, shown in **Fig. 2**, consisted of two polarizing Nicol prisms oriented so that their planes of acceptance made an angle of 45° with each other. The material causing the Faraday rotation was placed between them. If the Faraday rotator is adjusted to cause a 45° rotation, light which is passed by the first crystal is passed by the second also. In the reverse direction, however, the 45° rotation added by the Faraday cell produces light rays polarized horizontally which are reflected by the Canada balsam cement in the first Nicol prism and directed toward point C. Thus light admitted to the device at point A is transmitted to point B; light admitted at point B is transmitted to point C; light admitted at point C is transmitted to point D; and light admitted at point D is transmitted to point A.

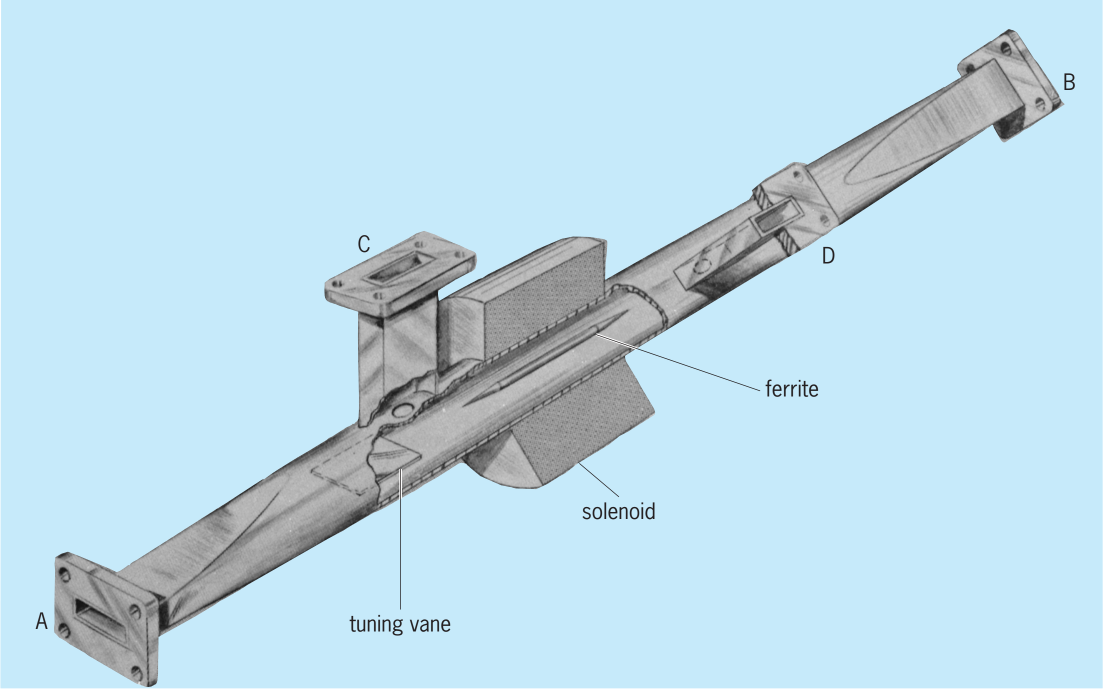


Fig. 3. Microwave analog of nonreciprocal optical device.

See FARADAY EFFECT; POLARIZED LIGHT.

The microwave analogy of Lord Rayleigh's device was proposed by C. L. Hogan and is shown in **Fig. 3**. Since it circulates microwave power from waveguide A to B, from B to C, from C to D, and from D to A, it has been called a (four-port) circulator. The nonreciprocal medium used here is a ferrimagnetic material called ferrite. In such a material, infinitesimal magnetic dipole moments which arise from the electronic structure of the material act gyroscopically when a steady magnetic field is applied, as shown in **Fig. 4**. They precess about the applied field direction in a counterclockwise sense, thus permitting strong coupling to the component of a microwave-frequency magnetic field which is circularly polarized in the same sense. The component with the opposite sense of polarization is weakly coupled. Thus energy exchange between the magnetic dipoles and the microwave field is polarization-sensitive. *See* FERRIMAGNETISM; FERRITE; FERRITE DEVICES.

Present-day circulators utilize the properties of electromagnetic fields in ferrite loaded microwave circuits. Consider the three-port circulator shown in **Fig. 5**, which comprises a circular disk resonator filled with ferrite connected to three transmission lines. When microwave energy is transmitted to the resonator along one of the transmission lines, an electromagnetic field is established in the resonator which is stationary in space, as shown in **Fig. 6**. The application of a dc magnetic field perpendicular to the plane of the disk resonator will rotate this stationary pattern through an angle dependent on the strength of the applied field. The field pattern is dipolar in nature, and hence has a region where the microwave magnetic field intensity is low. If the pattern is oriented to position this low-intensity region at one of the output transmission line ports, very little microwave power will leave the resonator via this port: it is isolated from the input. From the symmetry of the junction, this stationary pattern will advance 30°

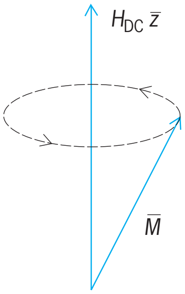


Fig. 4. Precessional motion of the ferrite internal magnetization vector, $\bar{M}$, with a static magnetic field H_{DC} applied in the $\bar{z}$ direction.

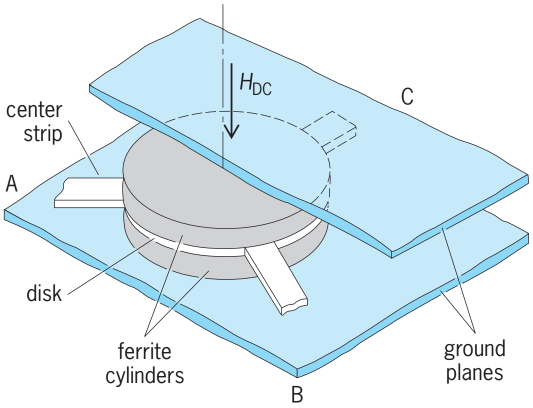


Fig. 5. Stripline Y-junction circulator. (*After C. E. Fay and R. L. Comstock, Operation of ferrite junction circulator, IEEE Trans. Microwave Theory Tech., MTT-13:1–13, January 1965*)

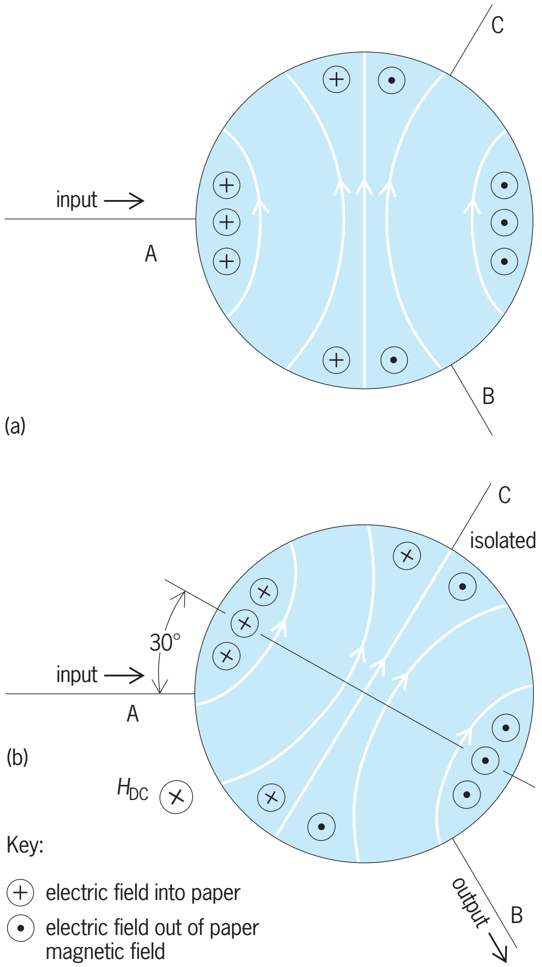


Fig. 6. Stationary electromagnetic field pattern in disk resonator of stripline circulator. (*a*) Junction is not magnetized, resulting in symmetric coupling to ports B and C. (*b*) Junction is magnetized to rotate pattern 30°, for circulation. Port C is isolated. (*After C. E. Fay and R. L. Comstock, Operation of ferrite junction circulator, IEEE Trans. Microwave Theory Tech., MTT-13:1–13, January 1965*)

if the ports are excited sequentially. Thus energy at port A will be transmitted to port B, energy incident on port B will be transmitted to port C, and energy incident on port C will be transmitted to port A. The ideal three-port scattering matrix of this junction circulator is given in Eq. (11).

$$[S] = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad (11)$$

Devices based on this pattern rotation principle have been realized in a variety of transmission line geometries, including rectangular waveguide, stripline, and microstrip. High performance can be obtained over wide microwave-frequency bandwidths (with ratio of upper to lower frequency limit less than or equal to 4:1) with isolation in the order of 0.01 of the incident power and with about 95% of the power transmitted to the desired port. These devices are found in most microwave systems, where they are employed to control the flow of microwave signals. For example, their use permits a radar transmitter and a radar receiver to share the same antenna.

In another use, a transmission line–matched termination connected to one of the ports of a three-port circulator produces a one-way transmission device, known as an insolator. (Any signal incident on this termination is absorbed.) Such two-port devices are widely used to prevent unwanted reflections from altering the properties of oscillators or amplifiers and to eliminate frequency-dependent transmission behavior caused by reflections in networks. Isolators realized from terminated circulators are compact, operate over wide bandwidths, and can be made to dissipate high microwave power levels through the use of an external high-power termination. Commercial low-power devices are in "drop-in" form compatible with microwave integrated (printed) circuit technology. *See* INTEGRATED CIRCUITS.

Fred J. Rosenbaum

Bibliography. R. E. Collin, *Fundamentals for Microwave Engineering*, 1992; J. Helszajn, *Microwave Engineering: Passive, Active, and Non-Reciprocal Circuits*, 1992; S. Y. Liao, *Microwave Devices and Circuits*, 3d ed., 1990.

Gyrocompass

A north-seeking form of gyroscope used as a directional reference in navigation. The first practical gyrocompasses were developed by H. Anschütz (Germany) in 1908, E. A. Sperry (United States) in 1911, and S. G. Brown (England) in 1916. Modern gyrocompasses are so reliable and so much more accurate than magnetic compasses that they are now used as the prime navigational instrument on nearly every ship and on major aircraft and missiles. *See* MAGNETIC COMPASS.

Basic operation. A gyrocompass combines the action of two devices, a pendulum and a gyroscope, to produce alignment with the Earth's spin axis. The principle is demonstrated with the model of **Fig. 1**, which consists of a rapidly spinning, heavy gyro

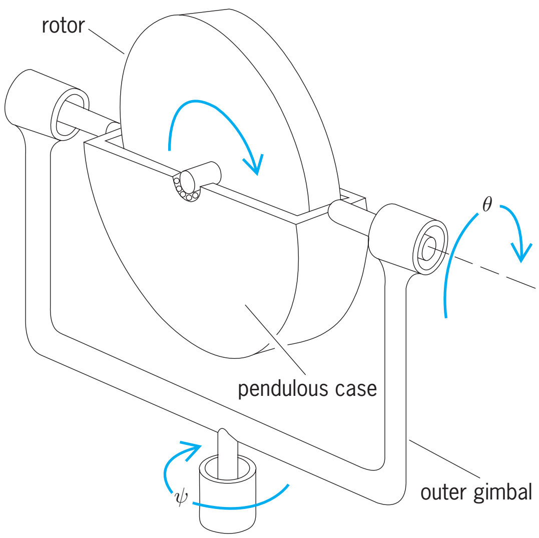


Fig. 1. Gyrocompass model.

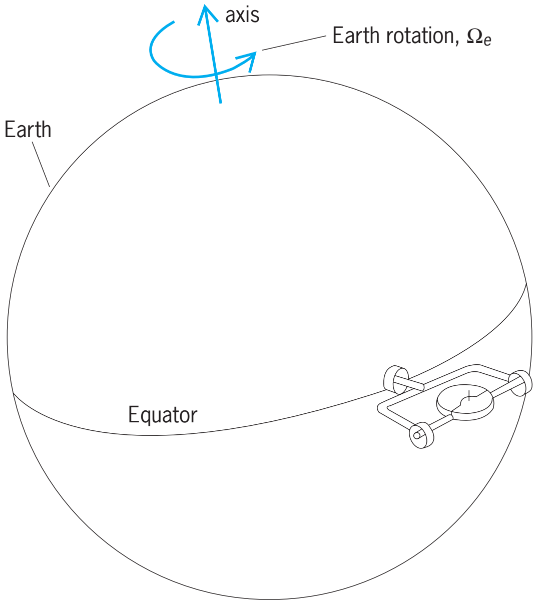


Fig. 2. Precession due to Earth's rate of rotation.

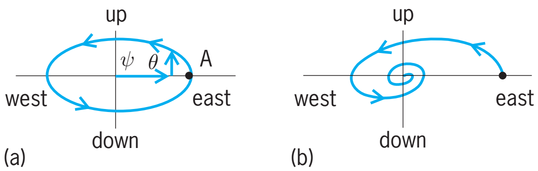


Fig. 3. Path of rotor axle of gyrocompass. (*a*) Typical undamped pattern. (*b*) Pattern with damping added.

rotor, a pendulous case which permits the rotor axle to nod up and down (angle θ), and an outer gimbal which permits the axle to rotate in azimuth (angle ψ).

In **Fig. 2** the model is positioned at the Equator of the Earth. As the Earth rotates, the gimbal moves with it. So long as the rotor's spin axis is aligned with the Earth's axis, the gyro experiences no torque from Earth rotation. If there is misalignment, however, a sequence of restoring torques is initiated.

The restoring action can be seen by imagining a beam of light to be projected by the gyro rotor axle onto a vertical piece of paper held just at the north end of the axle. **Figure 3*a*** shows a typical pattern traced on the paper by the light spot. At point A the rotor axle has been displaced east by $\psi = 5^\circ$. The spin axis of the gyro is no longer parallel to that of the Earth. Therefore, as the Earth rotates, the north end of the gyro rotor tilts upward. The ensuing dip angle results in a pendulum torque, down at the south end and up at the north end of the rotor axis. This torque causes the north end of the gyro axis to precess westward. As soon as the gyro axis passes through the north-south meridian, the action reverses. The north end of the rotor tilts down and the pendulum torque causes the gyro to precess eastward. Thus, the axis traces the elliptic path in Fig. 3*a*. In practical instruments damping is added and the path converges to true north in a spiral motion, as in Fig. 3*b*. *See* GYROSCOPE.

Shipboard installation. In a shipboard installation the system must be mounted in a complete set of gimbals to isolate it from rolling, pitching, and yawing motions of the ship. Friction must be minimized. Moreover, Schuler tuning is employed to keep horizontal accelerations of the ship from producing false torques on the pendulum; the unique combination of gyro spin speed and pendulosity is chosen so that no acceleration of the instrument can disturb its vertical reference. Then the light spot in Fig. 3 encircles the true north point once every 84 min. *See* SCHULER PENDULUM.

If the ship travels north or south, the ship's velocity V over the Earth creates an apparent Earth rate of $\Omega_s = V/R$ (R is the Earth's radius) about an axis perpendicular to the Earth's spin axis. The gyrocompass aligns to the vector sum of Ω_e and Ω_s, which results in a north steaming error angle whose tangent is Ω_s/Ω_e. This error is corrected by using V from the ship's pitometer to precess the gyro in the opposite direction at an equal rate.

Practical gyrocompass design is shown by the typical system depicted in **Fig. 4**. The rotor and case of Fig. 4 replace the rotor and case of Fig. 1, except that in Fig. 4 this assembly is not itself pendulous. Instead the case carries a coupling pin which rides in a slot on the ballistic ring, which is pendulous by virtue of the two mercury bottles and connecting tube known as a mercury ballistic. Gyroscope damping is introduced by making the coupling pin slightly eccentric.

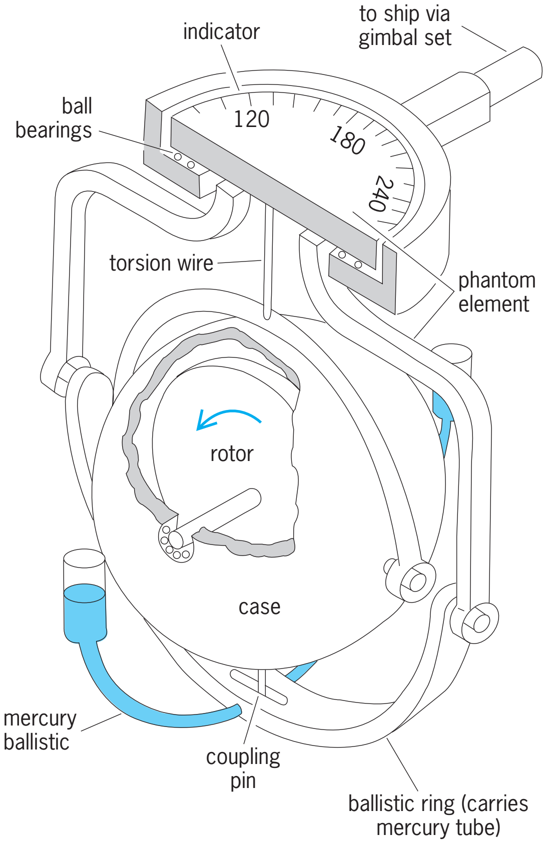


Fig. 4. Typical gyrocompass design.

The ballistic ring is supported by bearings from the phantom element, which replaces the outer gimbal of Fig. 1. The gyro assembly is also supported from the phantom element by a torsion wire which provides frictionless support. A servomotor is arranged to drive the phantom element whenever there is any twist in the wire. Thus the phantom element always stays aligned with the gyro. The phantom element carries an indicator from which ship's heading is read. Robert H. Cannon, Jr.

Aircraft application. For many years the use of gyrocompasses in aircraft was impractical because of their high speed and large, rapid changes in attitude. As mentioned above, the north-south component of vehicle velocity produces an error which depends on the velocity magnitude. Aircraft applications of gyrocompasses therefore use a modified version of the marine gyrocompass. The gyroscopes are mounted on a platform that is stabilized by signals from the gyroscopes. The platform is aligned to the local vertical and to north prior to takeoff by using essentially the same technique as for a marine gyrocompass. The 84-min Schuler period is shortened by amplifying signals from tilt sensors or accelerometers on the platform to rapidly remove platform tilt and align to north. Alignment times range from 5 to 30 min, depending upon the desired accuracy. The heading and vertical, once established, are "remembered" by the gyroscopes during flight. Vehicle velocity can be computed from the accelerometer data and used to correct for vehicle velocity and for dead-reckoning navigation. The need for preflight heading and vertical alignment can be eliminated by "gyrocompassing" this system in-flight by using an independent velocity sensor such as a Doppler radar or the Global Positioning System. *See* DOPPLER RADAR; INERTIAL GUIDANCE SYSTEM; SATELLITE NAVIGATION SYSTEMS.

The combination of the stable inertial platform and the radar allows accurate determination of local vertical and velocity of the aircraft. These data are put into the platform azimuth gyro servo loop to allow it to perform the gyrocompassing function correctly during flight. The Doppler radar must be capable of very high accuracy. To achieve 0.1° compass accuracy, the aircraft velocity must be determined to about 1 knot (0.5 m/s) at a latitude of 45°.

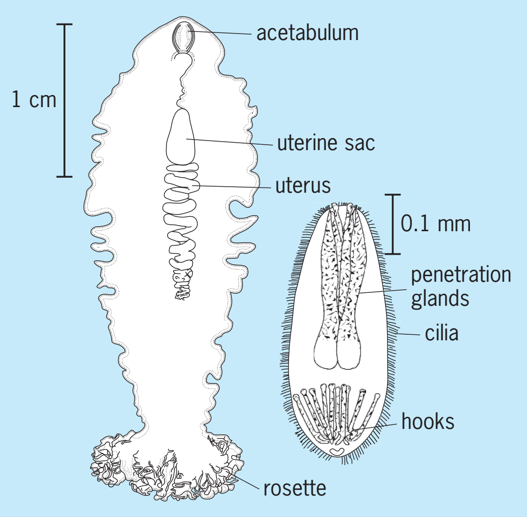
The complex gimbals that support the platform have been eliminated in newer gyrocompasses by the use of body-mounted or strapdown gyroscopes and accelerometers. The development of the ring-laser gyroscope has resulted in small lightweight gyrocompasses of comparable accuracy but lower cost than earlier gimbaled systems. Doppler radar velocity data can be used for in-flight alignment. Ring-laser gyroscopes are very reliable and retain their accuracy over very long periods of time. They are applied extensively in both commercial and military applications.

A gyrocompass must not be confused with a directional gyro. The latter is a gyroscope which is slaved to a magnetic compass and is not itself a north-seeker. Most conventional aircraft are equipped with north-slaved directional gyros. Heinz Buell

Bibliography. A. Frost, *Marine Gyro Compasses for Ships' Officers*, 1982; E. H. Pallett, *Automatic Flight Control*, 4th ed., 1994.

Gyrocotylidea

An order of flatworms, 2–3 cm (1 in.) long, in the intestine of Holocephali (the chimaeras, a subclass of the Chondrichthyes fishes). There are at least 10 species in two genera, *Gyrocotyle* and *Gyrocotyloides*. (The Cestodaria, in which they were formerly included along with the Amphilinidea, is probably not a valid taxon.) They lack proglottids (the segments found in tapeworms) and an intestine, and have a posterior attachment organ, the rosette (modified to form a cuplike sucker at the end of a caudal stalk in *Gyrocotyloides*) [see **illustration**]. Their body margins undulate. Most testes follicles are located in groups near the anterior end. The follicular vitellarium (the part of the female reproductive system that produces nutritive cells filled with yolk) is lateral, from the anterior to posterior body end. The ovary is located in the posterior third of the body. The uterus is sac-shaped or (in one species) branched. A ciliated lycophora larva about 0.3–0.5 mm (0.012–0.02 in.) long, with two separate anterior nephridiopores (external excretory structure openings) and five pairs of posterior hooks, hatches from the egg. The life cycles are unknown. Young hosts harbor many worms; large ones harbor usually no more than two. No pathological effects are known except in high-intensity infections. *See* AMPHILINIDEA; CESTODA; PLATYHELMINTHES. Klaus Rohde



Gyrocotylidean adult and larva.

Bibliography. K. Rohde, The minor groups of parasitic Platyhelminthes, *Adv. Parasitol.*, 33:145–234, 1994; W. Xylander, Gyrocotylidea (unsegmented tapeworms), in K. Rohde (ed.), *Marine Parasitology*, pp. 89–92, CSIRO Publishing, Melbourne, and CABI Publishing, Wallingford, England, 2005.

Gyromagnetic effect

An effect arising from the relation between the angular momentum and the magnetization of a magnetic substance. It is the effect which is exploited in the measurement of the gyromagnetic ratio of magnetic materials. The gyromagnetic effect is demonstrated by a simple experiment in which a freely suspended magnetic substance is subjected to a magnetic field. Upon a change in direction of the magnetic field, the magnetization of the substance must change. In order for this to happen, the atoms must change their angular momentum. Since there are no external torques acting on the system, the total angular momentum must remain constant. Thus the sample must acquire a mass rotation which may be measured. In this way, the gyromagnetic ratio may be determined. Two common methods of determination are the Einstein-de Haas method and the Barnett method. *See* GYROMAGNETIC RATIO.

Einstein-de Haas method. This is usually used to determine the gyromagnetic ratio of ferromagnetic materials. Imagine a ferromagnetic substance in the shape of a cylinder suspended on one end by a torsion fiber, forming thereby a torsional pendulum. A magnetic field is applied parallel to the axis of the cylinder, for example, by means of a coil surrounding the cylinder. The ferromagnetic cylinder is now magnetized in the direction of the magnetic field. If the field is suddenly reversed, the magnetization will reverse, and the accompanying change in angular momentum of the atoms will be balanced by a mass rotation of the cylinder which can be measured by noting the change in amplitude of displacement of the torsional pendulum. A measurement is made of the change in magnetization by means of a magnetometer. The ratio of magnetization change to angular momentum change is the magnetomechanical ratio, the reciprocal of the gyromagnetic ratio.

Barnett method. This is an alternative to the Einstein-de Haas technique. In this experiment, a ferromagnetic bar with a coil surrounding it is spun rapidly about its axis. The atoms, therefore, acquire an angular momentum in the direction of rotation. The magnetization thereby produced is measured by stopping the rotation abruptly and measuring the voltage induced in the surrounding coil which is part of a fluxmeter circuit. Now the same increase in magnetization in the direction of the rotation could be achieved by applying a magnetic field in that direction instead of rotating the bar. In the first case, if an angular momentum L makes an angle θ with the axis of rotation, it will experience a torque $L\omega \sin\theta$, where ω is the angular velocity of rotation, tending to turn it into the axis of rotation. In the second case, the magnetic field H would produce a torque on L in the same direction of $(1/\gamma)\,LH\sin\theta$ where $1/\gamma$ is the magnetomechanical ratio (L/γ = magnetic moment). If the magnetizations in the two cases are the same, then the torques are the same also, and Eq. (1)

$$L\omega\sin\theta = \frac{1}{\gamma}LH\sin\theta \qquad (1)$$

holds. Thus Eq. (2) is obtained.

$$\gamma = \frac{H}{\omega} \qquad (2)$$

In this way the magnetomechanical ratio can be measured. In an experiment similar to the one described, S. J. Barnett showed in 1914 that the magnetomechanical ratio for a ferromagnet was close to e/mc, the ratio of charge to mass of an electron.

Elihu Abrahams; Frederic Keffer

Bibliography. C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996.

Gyromagnetic ratio

The ratio of angular momentum to magnetic moment for atomic systems. This ratio is usually expressed in terms of the magnetomechanical factor g', as in Eq. (1). The ratio is written here in electromagnetic

$$\frac{\text{Angular momentum}}{\text{Magnetic moment}} = \frac{2mc}{g'e} \qquad (1)$$

units; thus, e/c and m are the charge and mass of the electron. The factor g' is sometimes loosely called the gyromagnetic ratio.

Magnetomechanical ratio. This quantity is the inverse of the gyromagnetic ratio. It is usually denoted by γ and is equal to $g'e/2\,mc$.

The magnetomechanical ratio of a substance identifies the origin of the magnetic moment. For example, for electron spin the angular momentum is $^1/_2\,\hbar$, where $\hbar$ is Planck's constant divided by 2π. The magnetic moment is the Bohr magneton $e\hbar/2mc$. Thus, the magnetomechanical ratio is given by Eq. (2).

$$\gamma = \frac{e\hbar/2mc}{\hbar/2} = \frac{e}{mc} \qquad (2)$$

Since $\gamma = g'e/2mc$, for electron spin $g' = 2$. For orbital angular momentum, $\gamma = e/mc$ and $g' = 1$. The experimental values of g' for most ferromagnetic materials are in the neighborhood of 2, showing that the major contribution to the magnetization comes from the electron spin. Deviations of g' from 2 show the extent to which the orbital motion contributes to the magnetization. In superconductors, the fact that $g' = 1$ shows that the diamagnetic currents which cause the Meissner effect are caused by electrons. For discussion of the measurement of the magnetomechanical ratio *see* ELECTRON SPIN; GYROMAGNETIC EFFECT; MEISSNER EFFECT; SUPERCONDUCTIVITY.

Spectroscopic splitting factor. This is the measure of the energy-level splittings of an atomic system in a magnetic field. For a free electron spinning in a magnetic field H, the energy levels are split according to $\Delta E = g\mu_B H$ where μ_B is the Bohr magneton and g is the spectroscopic splitting factor, or g factor, and is equal to 2.00. For electron spins in paramagnetic salts, there are complicated energy-level schemes because of the spin-orbit coupling and crystalline field interactions; the g factor is different from case to case and may depend upon the orientation of the magnetic field with respect to the crystal axes. Under these circumstances, the *g* factor is defined quantum mechanically. For ferromagnetic materials, the spectroscopic splitting factor is defined similarly; it is the factor in the ferromagnetic resonance condition which gives the splitting of the energy levels in the magnetic field. It is not generally equal to the magnetomechanical factor g', as the two are affected differently by the effects of orbital angular momentum. For free atoms, the spectroscopic splitting factor is identical to the Landé *g* factor. For further discussion of *g* factors. *See* MAGNETIC RESONANCE.

Elihu Abrahams; Frederic Keffer

Gyroscope

A device that is used to define a fixed direction in space or to determine the change in angle or the angular rate of its carrying vehicle with respect to a reference frame. Gyroscopes (also called gyros) respond to vehicle angular rates, that is, rates of change of angles between vehicle axes and reference axes, from which these angles can be computed. Gyros are used for guidance, navigation, and stabilization, for example, to measure the angular deviation of a guided missile from its desired flight trajectory; to determine the heading of a vessel for steering; to determine the heading of an automobile as it turns through city streets; to indicate the heading and orientation of aircraft during and after a series of maneuvers; and to stabilize and point radar dishes and satellites. *See* AIR NAVIGATION; GYROCOMPASS; INERTIAL GUIDANCE SYSTEM; MARINE NAVIGATION; NAVIGATION.

Gyros can be utilized either mounted on a stabilized platform, whose orientation in a moving vehicle remains fixed in space by means of two, three, or four gimbals, or directly attached to the vehicle's body, so that the gyro experiences the same maneuvering as the vehicle, an operating mode referred to as strapdown. Strapdown operation is desirable because it enables a much less expensive system; it became feasible only in the 1970s and 1980s, when the very high digital computing speed required to mechanize the strapdown algorithms became available.

Gyros can be operated closed-loop or open-loop. Closed-loop means that feedback from the gyro output introduces a restoring mechanism either inside the gyro (for example, torquing in mechanical gyros) or counterrotating platform motions to maintain the gyro at its null (initial) setting. In open-loop operation, the gyro is allowed to operate off its null position as it responds to the input angular rates. *See* CONTROL SYSTEMS; SERVOMECHANISM.

Gyro performance is usually characterized by the drift-rate parameter, which is routinely reported in degrees per hour. A drift rate of 0.015°/h corresponds to an accumulated position error of 1 nautical mile (1.85 km) per hour on the Earth's surface at the Equator. Another important performance parameter is angle random walk, which is reported in degrees per square root hour. This is an error which results from the integration of white noise into the angular rate measurement. The most precise navigation or guidance systems require performance to better than 0.001°/h; aircraft navigation requires 0.01°/h; tactical missiles and automobile navigation require 0.1–10°/h; autopilots, seeker heads, and so forth, require 10–100°/h; and automobile chassis control and virtual-reality games require only a few thousand degrees-per-hour performance. Incorporating Global Position System (GPS) receivers in an integrated navigation system can reduce the requirements on gyro accuracy by an order of magnitude. Many gyro error sources such as bias (that is, the output even when there is no input rate), sensitivity to temperature, and acceleration can be calibrated or modeled and compensated for, although not completely, in the navigation or control-system computer. *See* AUTOPILOT; SATELLITE NAVIGATION SYSTEMS.

Gyroscopes use different physical phenomena to respond to input angular rates; for example, spinning-mass gyros sense changes in angular momentum from Coriolis acceleration; resonator gyroscopes sense deflections from Coriolis acceleration; and optical gyros sense phase shifts (the Sagnac effect) between counterpropagating beams of light. Instruments that do not have spinning masses are not technically gyros but angular rate sensors. However, the term "gyro" is commonly used for all rate-sensing devices. *See* CORIOLIS ACCELERATION.

Gyroscopic precession. The classical spinning-mass gyroscope is based upon the phenomenon that the spin axis of a spinning mass points in a fixed direction in space unless acted upon by an external influence. This phenomenon can be observed with the toy gyroscope, whose orientation remains generally fixed (that is, defines a fixed reference) no matter how its base is moved. (Such a body is called a free gyro.) However, the spin axis can be made to rotate if a torque (or rotational rate) is applied at right angles to the spin vector. The spin vector then begins to rotate (precess) about a third axis, perpendicular to the spin axis and the applied torque; that is, the spin axis tries to align itself with the applied torque. This is the law of gyroscopic precession; measurement of precession is what makes the spinning-mass gyro useful for knowing the changes in direction of the carrying vehicle.

A mass, of inertia I, spinning at rate ω_s, defines an angular momentum vector $\mathbf{H}$ along the axis of rotation or spin axis (**Fig. 1**). When a torque $\mathbf{M}$ is applied about a perpendicular axis, a mass particle at a point A on the rim is forced downward and reaches its maximum downward velocity when it has rotated through 90° to a point A'. After rotation through another 90° to a point A'', the mass particle has moved back up to zero displacement; however, the net force pattern was down. On the other side, a similar particle is experiencing an upward force pattern. Thus, the spinning mass is twisted (precesses) about the third orthogonal axis, called the output axis. The precession rate Ω is given by Eq. (1).

$$\Omega = \frac{M}{H} = \frac{M}{I\omega_s} \qquad (1)$$

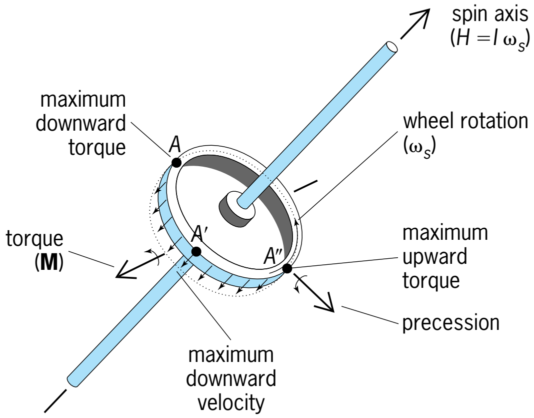


Fig. 1. Gyroscopic precession.

See ANGULAR MOMENTUM; MOMENT OF INERTIA; TORQUE.

A simple experiment, requiring a bicycle wheel with handles on the hub, can be used to demonstrate precession. The wheel is spun up at fairly high speed and supported vertically by a string tied to one of the handles. The wheel will not fall but will remain vertical, precessing slowly around the string because of the torque applied by the wheel's weight. If a freely rotatable base is available, a person standing on the base can cause self-precession by holding the wheel with spin axis horizontal and twisting the handles clockwise or counterclockwise.

Conversely, if a spinning mass experiences an input angular rate Ω about an axis orthogonal to the spin axis, an output torque **M** appears about the third orthogonal axis. For gyroscopic instruments, since vehicle rotations are to be measured, the torque produced by an applied rate is the quantity of interest. *See* PRECESSION; RIGID-BODY DYNAMICS.

Spinning-mass gyroscopes. The free gyroscope's spinning mass is isolated from the rotations of the case or the carrying vehicle so that it remains fixed in space. The relative position of the spinning mass to the gyro case is proportional to the vehicle's angle of rotation. Low-performance versions are used for artificial horizons, turn indicators, and directional (heading) indicators.

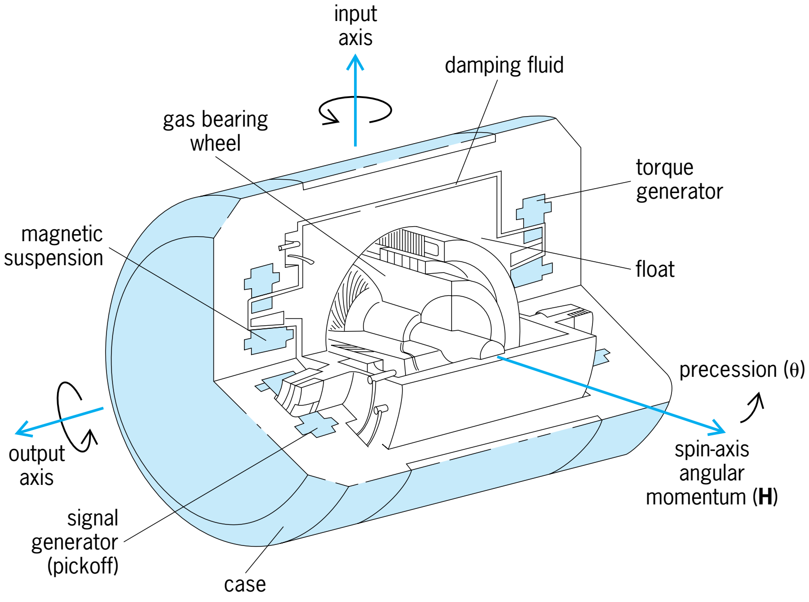


Fig. 2. Single-degree-of-freedom, rate-integrating floated gyroscope. (*After L. Finkelstein and I. V. Grattan, eds., Concise Encyclopedia of Measurement and Instrumentation, Pergamon Press, 1994*)

Rate gyro. The rate gyro's spinning mass is forced to rotate with the vehicle rotation rate Ω about one particular axis (for example, pitch or roll). The output torque **M** causing the spinning mass to turn (precess) is opposed by an elastic restraint of spring constant K. The angle θ, measured by pickoffs, that the spinning mass turns through is proportional to the vehicle rotation rate Ω, as given in Eq. (2).

$$\Omega = \frac{M}{H} = \frac{K\theta}{H} \tag{2}$$

See AUTOMATIC HORIZON.

Rate gyros are used in applications in which errors greater than 10°/h can be tolerated, such as autopilots and stability-augmentation flight control. For measuring rapid rate changes or when higher accuracy is required, a precision gyroscope may be required. *See* STABILITY AUGMENTATION.

Floated gyro. The floated, single-degree-of-freedom, rate-integrating gyro, or floated gyro as it is commonly known, is basically a rate gyro in which the spring restraint is replaced by viscous damping. The spinning mass is contained inside a "float" which is isolated from the case by means of the viscous flotation fluid, low-restraint electromagnetic suspensions, pickoffs, and torque generators (**Fig. 2**). Under an applied vehicle rotation rate Ω, the spinning mass (float) begins to precess (rotate) at rate given by Eq. (3), where C is the damping coefficient. Integrating this equation with respect to time gives Eq. (4) for the change in the carrying vehicle's angular

$$\Omega = \frac{M}{H} = \frac{C\dot{\theta}}{H} \tag{3}$$

$$\theta_v = \frac{C\theta}{H} \tag{4}$$

direction θ_v. The integration is done mechanically inside the gyro via the fluid, hence the term "integrating;" pickoffs (usually electromagnetic) measure θ directly. *See* VISCOSITY.

Floated, integrating gyros went from revolutionizing military aircraft navigation in the 1950s to enabling strategic missile guidance, submarine navigation, space flight (for example, the Apollo spacecraft), and satellite stabilization (for example, the *Hubble Space Telescope*). However, floated gyros have relatively high costs, are labor-intensive with respect to their overall life cycle, and have long warmup times—factors which motivated the development of other spinning-mass and non-spinning-mass gyroscopes.

DTG. The dynamically (or dry-) tuned gyroscope (DTG) was invented in the early 1960s (**Fig. 3**). It is a two-degrees-of-freedom gyro whose spinning mass is suspended by a universal, or Hooke's, joint. By selecting, or tuning, the spin speed, the inertia of the gimbal portion (typically 15–40 mm or 0.6–1.6 in. in diameter) of the universal joint cancels the elastic restraint of the joint flexures; hence, the rotor is decoupled from the drive shaft, and the spinning mass operates as a free gyro, but with limited angular displacement. By symmetry, the DTG senses two axes of angular rate simultaneously, while avoiding the use of flotation fluids (hence the term "dry") and electromagnetic suspensions. Torquing or platform motions are used to keep the rotor aligned with the case. DTGs have been used for aircraft and missile navigation, north finders, heading and reference systems, seeker heads, and spacecraft stabilization. *See* UNIVERSAL JOINT.

Suspended gyros. The electrostatically suspended gyroscope (ESG) was developed in the 1960s and 1970s. Its spherical rotor (10–40 mm or 0.4–1.6 in. in diameter) is suspended in a spherical chamber in vacuum by electrostatic forces. External motor windings spin the rotor to the desired speed and are then turned off. The rotor will spin for days before requiring motor excitation. Optical or electrical techniques are used to pick off rotor position. The ESG is basically a free gyroscope and does not require a torquer to keep the rotor and case aligned. ESGs are very accurate (0.0001°/h) and are used for long-term navigation of submarines and aircraft and for land surveying. Cryogenic ESGs, with accuracies three orders better than the best floated gyros, have been developed for measurements testing relativity theory in space. A similar concept is the magnetically suspended gyro (MSG), which uses a magnetic field to suspend the rotor. The performance of MSGs is not quite as good as that of ESGs, but they are more rugged. *See* ELECTROSTATICS; MAGNETIC FIELDS; RELATIVITY.

Sagnac effect. Optical gyros use the Sagnac effect to detect rotation. The Sagnac effect pertains to the postulate of the theory of relativity that the speed of light c is constant, and independent of the motion of the source. If two identical light waves (wavelength λ) circulate in opposite directions along a closed path undergoing a rotation Ω, then the light beam traveling in the same direction as the rotation takes longer to travel around the path than the other beam, resulting in a changed interference pattern. Thus, the beams become phase-shifted with respect to each other by an angle ϕ given by Eq. (5), where A is the

$$\phi = \left(\frac{8\pi AN}{\lambda c}\right)\Omega \qquad (5)$$

area enclosed by the path, and N is the number of times the light travels around the path. The quantity in parentheses, called the scale factor, relates the gyro's output to its input. It is apparent that the larger the enclosed area A (that is, the bigger the device), the better the performance.

Optical gyroscopes. These include the ring laser gyro (RLG) and fiber-optic gyros (FOGs).

Ring laser gyro. This gyro was invented in the 1960s and is now widely used in tactical and navigation systems (**Fig. 4**). It comprises a closed optical cavity (usually a three- or four-sided block of low-expansion-coefficient material, such as modified quartz or lithium aluminum silicate), whose light path is defined by mirrors mounted at the corners. The light travels through holes bored in the block containing a low-pressure gas, usually a helium-neon mixture which lases when the anode and cathode are excited. Thus, the RLG is itself actually a laser (that is, it does not require an external light source), and is thus said to be an active device. *See* LASER.

The lased light propagates clockwise and counterclockwise so that there are two optical beams resonating in opposite directions in the cavity (typically with path lengths of 80–400 mm or 3–16 in.). Each optical beam is maintained in resonance (that is, each beam contains an integral number of wavelengths). Under a rotation rate about the gyro input

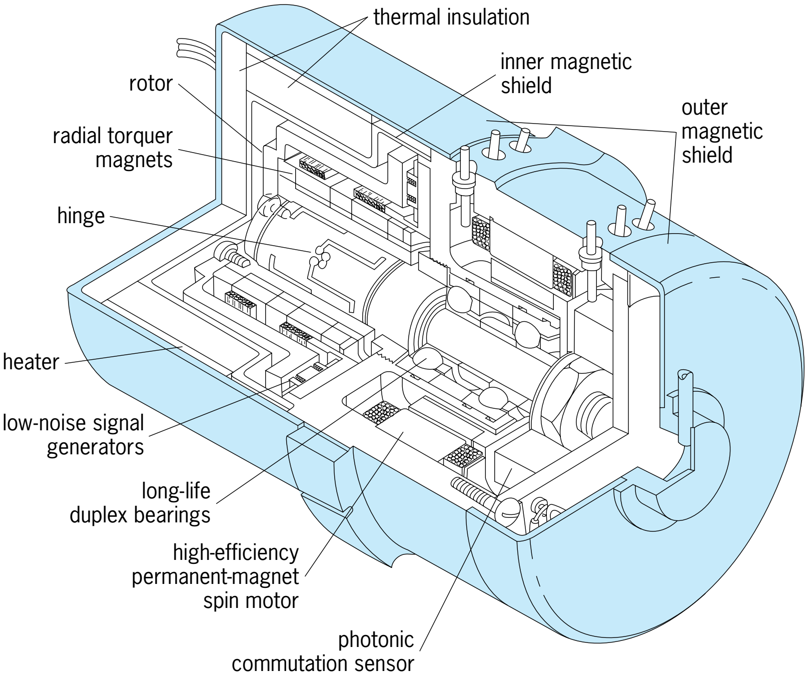

thermal insulation
inner magnetic shield
rotor
outer magnetic shield
radial torquer magnets
hinge
heater
low-noise signal generators
long-life duplex bearings
high-efficiency permanent-magnet spin motor
photonic commutation sensor

Fig. 3. Dynamically tuned gyroscope.

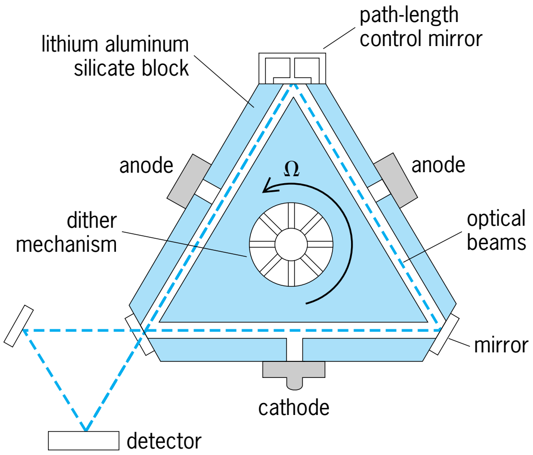


Fig. 4. Ring laser gyroscope. Input axis is perpendicular to plane of page.

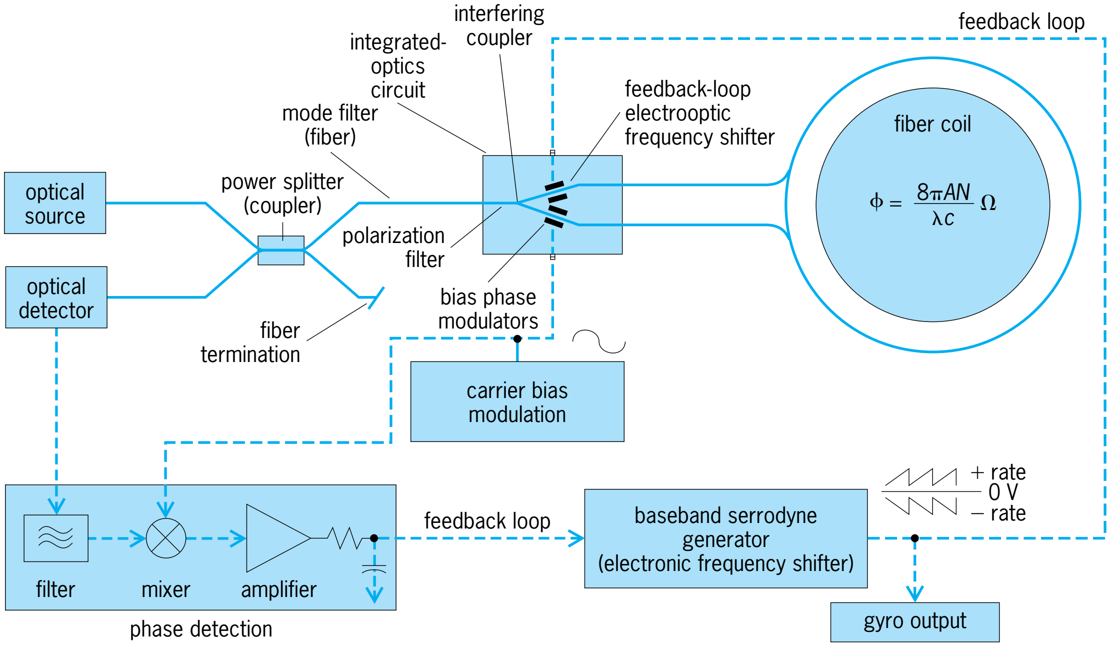


Fig. 5. **Interferometric fiber-optic gyroscope. Solid lines show optical path; broken lines show electrical path.**

axis, the resonant frequencies of the clockwise and counterclockwise beams change. Some light from both beams is transmitted through one of the mirrors to impinge on a detector. Because the beams have different frequencies, a changing interference pattern (fringes) appears. The detector has two sensitive elements which allow measurement of the fringe pattern changes across the detector in both the positive and negative direction, thereby determining the external rotation rate and direction. Each fringe is equivalent to an angle input, so that counting the number of fringes produces a measurement of the angle. Therefore, the RLG (together with its detector) is an open-loop integrating gyro.

The key element in the RLG is the mirrors, which must be very precise to minimize backscatter between the two beams. Backscatter causes the two light beams to lock frequencies so that low input rates cannot be detected. This phenomenon, known as lock-in, can be avoided by dithering the RLG about the input axis at variable frequencies by a piezoelectrically driven dither motor. Another way to avoid lock-in is to use four light beams in a slightly bent optical path and either Faraday or Zeeman rotators. These rotators alter the propagation constant of the light beams by rotating the polarization, resulting in a phase change, and thus separating the beams to prevent lock-in. The Faraday rotator performs this operation as the light beams travel through a magnetically active crystal, whereas the Zeeman rotator applies a magnetic field bias directly to the gas medium. High-performance RLGs also need a path-length control device integral with one of the mirrors to make the gyro insensitive to changes in the optical cavity length. *See* FARADAY EFFECT; OPTICAL ISOLATOR; ZEEMAN EFFECT.

Since 1980, RLGs have advanced rapidly into many applications, taking over from mechanical instruments. The RLG is an exceptional device for strapdown applications because it has excellent scale-factor stability and linearity, and almost negligible sensitivity to acceleration. It also has digital output, fast turn-on, excellent dormancy, and no moving parts, and it requires much less maintenance than mechanical devices.

Fiber-optic gyros. These gyros use optical fibers, in place of a lasing block, to define the optical path. The light source is external, and its light is split by a beam splitter or optical coupler to produce clockwise and counterclockwise light beams in the fiber-optic coil. FOGs are called passive devices because the optical source (a laser) is external. FOG development began in the 1970s. There are two principal types: interferometric and resonant.

The interferometric fiber-optic gyro (IFOG) [**Fig. 5**] has up to 1 km (0.6 mi) of optical fiber wound into a coil with both ends brought into a coupler. Light from the optical source passes through a power splitter (coupler) and a mode filter (fiber), and into the integrated optics circuit containing an interfering coupler (Y-junction). The interfering coupler splits the light into counterpropagating beams, and then recombines them after they have propagated through the fiber coil. The recombined beam then retraces the path through the mode filter, and is guided into an optical detector via the power splitter, thereby ensuring that both beams have traveled identical paths. The integrated optics circuit is an electrooptic crystal (lithium niobate) comprising optical waveguides fabricated by a proton-exchange process or a titanium-diffused process. It is used for polarization filtering, phase and frequency modulation, and precision interfering (optical phase–to–optical intensity conversion) of the recombined beams. Bias modulation ensures operation of the IFOG in the regime most sensitive to input rate. Backscatter errors are eliminated by using a wide-band laser source. *See* ELECTROOPTICS; INTEGRATED OPTICS.

The basic open-loop IFOG is not an integrating gyro like the RLG, nor is its output (scale factor) linear with input rate except for a limited range. However, the addition of a feedback loop, including a frequency shifter via serrodyne (sawtooth-wave) phase modulation, can make the IFOG closed-loop and linear with rate. As the input rotation rate changes and is detected at the detector, the feedback loop provides a signal to the frequency shifter which changes the frequency in the opposite direction so that the detector reads at the null position. The sawtooth frequency is proportional to rate. Counting the cycles in the sawtooth provides angle, so that the gyro is now an integrator. Diameters of IFOGs vary from 180 mm (7 in.) for high-performance gyros down to 25 mm (1 in.) for gyros devoted to tactical applications.

The resonant fiber-optic gyro (RFOG) maintains the counterpropagating light beams in resonance, recirculating them in a short fiber-optic coil. Light from an external source enters into the coil by a coupler. In the RFOG, unlike the IFOG, the light continues to travel around the coil while constantly being balanced by the source. The coupler allows light to leave the coil after the beams become phase-shifted under a rotation of the coil. The clockwise beam and the counterclockwise beam are diverted onto separate detectors that measure the frequency shifts; the difference in clockwise and counterclockwise frequencies is proportional to rate. The frequency of the light entering the coil is shifted to maintain both beams in resonance and to provide closed-loop operation. A very narrow band laser is required to resolve coil resonances; therefore, RFOGs are susceptible to backscatter. RFOGs are lower-performing devices but offer size advantages over IFOGs. *See* FIBER-OPTIC SENSOR.

Vibrating gyroscopes. These gyros use an oscillating mass in place of a spinning mass to sense rates. The mass oscillates (sinusoidally) back and forth through a fixed angle; the amplitude of the oscillation is restrained by the elastic (spring) stiffness of the vibrating structure. Nearly all such gyros oscillate at the resonant frequency of the mass-spring system since the gyro output is maximized at this frequency; hence these gyroscopes are also called resonator gyros. *See* RESONANCE (ACOUSTICS AND MECHANICS); VIBRATION.

In the tuning-fork resonator gyro (**Fig. 6***a*), the two tines oscillate from side to side ($\dot{x}$) at a frequency ω, 180° out of phase with each other (that is, as one moves to the right, the other moves to the left). If an angular rate Ω is applied about the base, the Coriolis force causes the tines to move up and down, also at the frequency ω and also 180° out of phase with each other (Fig. 6*b*). The overall motion is thus an ellipse, and the up-and-down component of the vibration ($\dot{x}x$) can be used to sense Ω. *See* TUNING FORK.

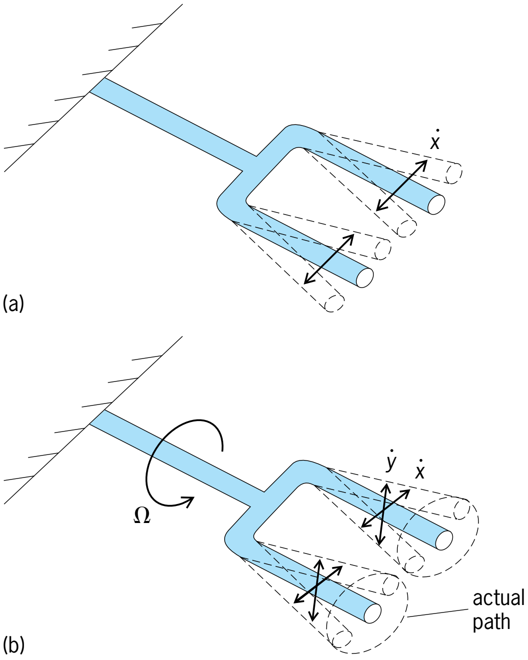


Fig. 6. Coriolis effect on a vibrating (resonator) structure. (*a*) Drive vibration ($\dot{x}$). (*b*) Vibration in presence of rotation rate Ω, which includes output (sense) vibration ($\dot{y}$).

There are many configurations of resonator gyros, many of which are made using micromachining techniques developed for the solid-state electronics industry, such as photolithography, masking, and etching, to carve mechanical shapes in quartz and silicon. Quartz devices use piezoelectric effects, via metallized electrodes deposited on the quartz, to drive the oscillation as well as to sense the response. Silicon devices use electrostatic forces to drive the oscillating mass and capacitive plates to sense the response. These devices can be made very small (300 micrometers × 600 micrometers). Micromechanical resonators offer the very inexpensive, high-production capability required for many commercial markets such as toys and automobiles. Because they are small, they also have the ability to withstand incredibly high acceleration forces (over 100,000 *g*), such as those experienced by artillery shells.

Another resonator gyro offering high performance is the hemispherical resonator gyro (HRG), which consists of a hemispherical quartz shell (20–50 mm or 0.8–2 in. in diameter), electrostatically forced into resonance so that the rim of the shell becomes elliptical. As the gyro is rotated about the stem, the node shape of the vibrating rim changes, and this change is sensed capacitively by sensors adjacent to the rim. This type of gyro has been used in commercial aircraft navigation systems and for satellite attitude control.

Other types. While the most widely used gyroscope techniques have been described, there are several other ways of sensing rate. Gyros that have been developed are the fluidic type and the magnetohydrodynamic angular motion sensor. In the fluidic (or laminar jet) type, a fluid jet is deflected under a rate to impinge on temperature-sensitive elements; the power change due to differential cooling leads to a measure of rate. The magnetohydrodynamic angular motion sensor measures angular acceleration by magnetohydrodynamic effects in a mercury-filled ring, and the angular acceleration is then integrated to obtain the angular rate. *See* MAGNETOHYDRODYNAMICS.

Gyros under development are the nuclear magnetic resonance (NMR) type, which uses the angular motion of atomic nuclei to respond to rate, and the Josephson junction gyro (JJG), which measures rotation by measuring a phase-induced change in the current in a superconducting ring. Because electrons have a much higher phase sensitivity than photons, the JJG is theoretically capable of being two orders of magnitude smaller than an RLG for equivalent performance. A promising technology which is in its infancy stages is inertial sensing based upon atom interferometry (sometimes known as cold atom sensors). A typical atom de Broglie wavelength is 30,000 times smaller than an optical wavelength, and because atoms have mass and internal structure, atom interferometers are extremely sensitive. Accelerations, rotations, electromagnetic fields, and interactions with other atoms change the atom interferometric fringes. In theory, this means that atom interferometers could make the most accurate gyroscopes, accelerometers, gravity gradiometers, and precision clocks, by orders of magnitude. *See* DE BROGLIE WAVELENGTH; INTERFEROMETRY; JOSEPHSON EFFECT; NUCLEAR MAGNETIC RESONANCE (NMR). Neil Barbour

Bibliography. A. Lawrence, *Modern Inertial Technology: Navigation, Guidance, and Control*, 2d ed., 1998; H. Lefèvre, *The Fiber-Optic Gyroscope*, 1993; *Optical Fiber Sensors*, vol. 4: *Applications, Analysis, and Future Trends*, ed. by J. Dakin and B. Culshaw, Artech House, 1997; W. Wrigley, W. Hollister, and W. Denhard, *Gyroscopic Theory, Design, and Instrumentation*, 1969.

Gyrotron

One of a family of microwave generators, also called cyclotron resonance masers, in which cyclotron resonance coupling between microwave fields and an electron beam in vacuum is the basis of operation. This type of coupling has the advantage that both the electron beam and the associated microwave structures can have dimensions which are large compared with a wavelength. Thus, cyclotron resonance masers are potentially greatly superior to conventional microwave tubes with respect to power capability at short wavelengths.

The development of these power sources is particularly significant for magnetically confined plasma fusion experiments. Microwave heating is considered an attractive method of supplying the energy needed to bring a reactor to ignition temperature, and gyrotrons provide a potential means of producing sufficient microwave power at the very short wavelength required. Gyrotrons also have potential application in millimeter-wave radar and communications systems. *See* NUCLEAR FUSION.

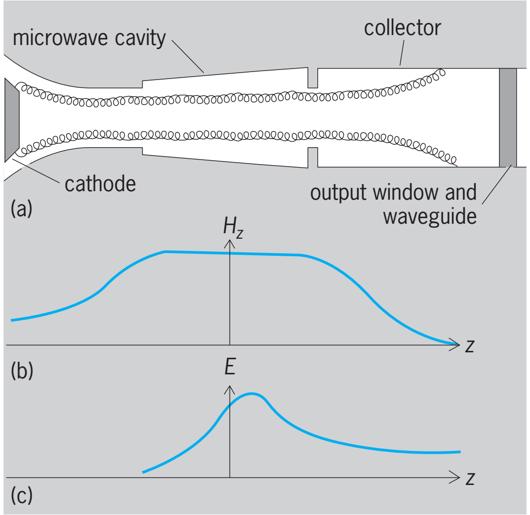


Fig. 1. Gyrotron. (*a*) Schematic diagram, showing elements. (*b*) Plot of typical dc magnetic field H_z as function of distance *z* along axis. (*c*) Plot of representative microwave electric field intensity $|E|$.

Basic characteristics. The basic cyclotron resonance condition is given by Eq. (1), where ω is the

$$\omega = n\omega_c \tag{1}$$

operating frequency, n is an integer, and ω_c is the cyclotron frequency or angular velocity of the electron given by Eq. (2). Here, B is the dc magnetic

$$\omega_c = \frac{eB}{\gamma m_0} \tag{2}$$

field, e is the electron charge, m_o is the rest mass, and γ is the relativistic mass factor. The fundamental cyclotron resonance occurs when $n = 1$. This is the strongest and most useful interaction. The resonance condition requires that very high magnetic fields be used for high-frequency devices. For example, a frequency of 120 GHz requires a magnetic field of about 45 teslas. Generally, the very high frequency gyrotrons have used superconducting magnets.

Larger values of n allow corresponding reductions in the required dc magnetic field. Practical devices have generally been limited to values of n no larger than 2 (second-harmonic operation).

The most important microwave field component in the gyrotron is the electric field tangential to the orbit of the electron. With the fundamental cyclotron resonance interaction, any spatial variation of the microwave fields is of little importance. It is this property that allows the gyrotron to use cross-section areas which are large compared with a wavelength.

Electron bunching in the gyrotron occurs by virtue of the relativistic mass effect included in Eq. (2). The transverse microwave electric field introduces a sinusoidal modulation of γ depending on the angular position of the electron in its orbit relative to the direction of the electric field. The modulation of γ results in a modulation of angular velocity as given by Eq. (2). As the beam drifts, this converts to

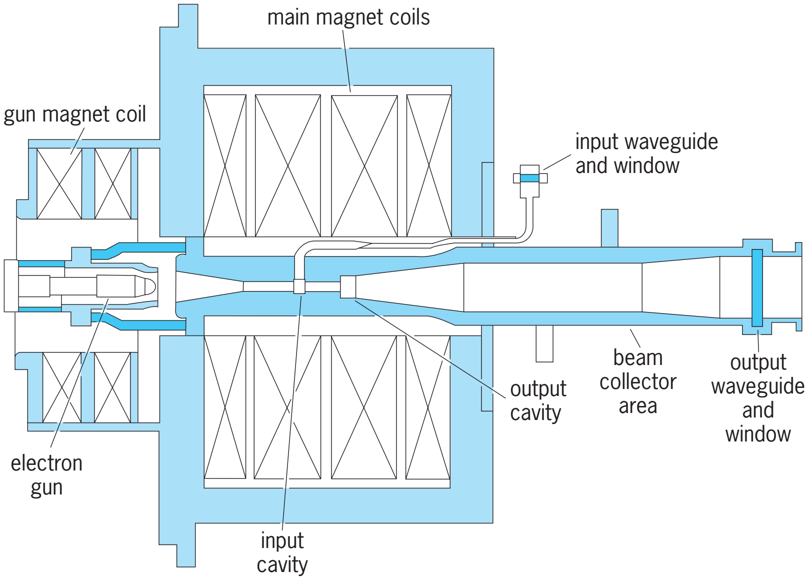


Fig. 2. Simplified cross section of pulsed gyroklystron amplifier.

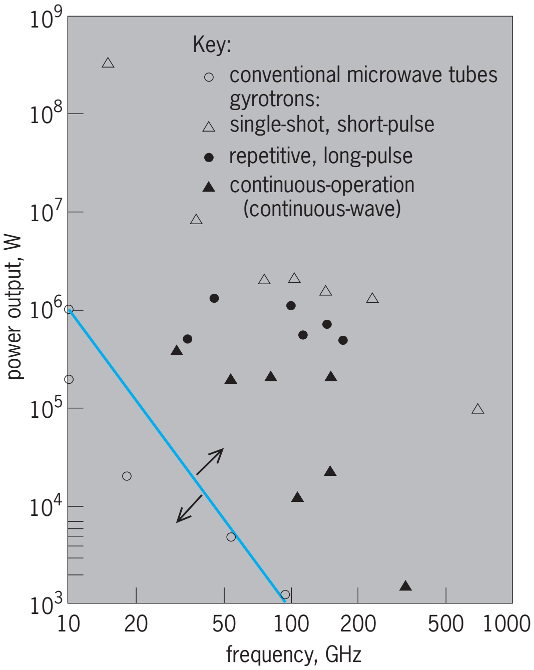


Fig. 3. Power output versus frequency typical of gyrotrons compared with that of conventional microwave tubes. For pulsed devices, peak power output is shown.

angular bunching in the coordinate system centered on each electron orbit. By proper adjustment of phase conditions, the bunched beam can give up most of its energy to microwave energy.

A number of tube configurations are possible using the cyclotron resonance interaction. The simplest form, and that used for most practical gyrotrons to date, is an oscillator using a single resonant cavity (**Fig. 1**). Here, the electron beam is a hollow beam with all electrons having helical motion. For efficient operation, all electrons must have a large fraction of their total energy contained in motion transverse to the device axis. A gyroklystron amplifier employing two or more cavities is another alternative (**Fig. 2**); and in a third variation a traveling-wave circuit, in analogy to a traveling-wave tube, is used.

Capabilities. Gyrotrons can produce higher power at high frequency than conventional klystrons and traveling-wave tubes (**Fig. 3**). Single-shot, short-pulse devices use intense relativistic beams which are not suitable for repetitive pulsing. *See* KLYSTRON; MICROWAVE TUBE; TRAVELING-WAVE TUBE. Howard R. Jory

Bibliography. J. Benford and J. Swegle, *High-Power Microwaves*, 1991; C. Edgecomb (ed.), *Gyrotron Oscillators*, 1993; R. J. Temkin (ed.), *17th International Conference on Infrared and Millimeter Waves*, 1992.

H

Hackberry — Hysteriales (lichenized)

Hackberry

A medium-sized to large tree, *Celtis occidentalis*, occasionally growing to 120 ft (36 m). It occurs in the eastern half of the United States, except the extreme south, and has corky or warty bark; alternate, long-pointed serrate leaves unequal at the base; and a small drupaceous fruit with thin, sweet, edible flesh. The pith of the twigs is chambered. The wood is used for furniture, boxes, and baskets. It is a shade tree and is also used for shelterbelts. Sugarberry (*C. laevigata*) is similar to hackberry. It grows in the southeast United States and has narrower leaves with entire margins and smaller fruit. It is used for furniture, boxes, and baskets; shelterbelts; and shade. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves/Kenneth P. Davis

Hadean

The portion, referred to as an eon, of geological time that extends for several hundred million years from the end of the accretion of the Earth to the formation of the oldest recognized rocks. According to current models, the inner planets formed by the accretion of planetesimals in an environment where gas and volatiles had been swept away by early intense solar activity. The accretion of the Earth appears to have been completed between 50 and 100 million years (m.y.) after the beginning of the solar system (T_0) as recorded in the oldest refractory inclusions in the Allende Meteorite, whose age of 4566 ± 2 m.y., ascertained by lead isotope dating, is taken as T_0. Core formation on the Earth appears to have been coeval with accretion and so preceded the Hadean. Any primitive atmosphere was removed by early collisional events, and the present atmosphere has arisen by a combination of degassing and additions from comets. *See* EARTH, AGE OF; LEAD ISOTOPES (GEOCHEMISTRY).

The Acasta Gneiss in the Northwest Territories of Canada, dated at 3960 m.y., is often regarded as the oldest rock. However, that date refers to relict zircon crystals in the rock rather than the age of formation of the rock itself. The oldest definitely dated rocks are at Isua, Greenland, with an age of 3650–3700 m.y. Thus the Hadean Eon begins around 4500–4450 m.y. ago and extends to between 3960 and 3650 m.y. ago depending on the age assigned to the oldest rock.

Lunar record. There is almost no evidence on Earth of the Hadean Eon, but the Nectarian (3850–?4200 m.y.) and Pre-Nectarian Systems (?4200–?4450 m.y.) on the Moon cover this missing period of time in the terrestrial rock record. The formation of the Moon is conventionally ascribed to a giant collision between an impacting body about 0.20 earth mass with the Earth about 50 m.y. after T_0, predating the Hadean. Part of the rocky mantle of the impactor (but not that of the Earth) was spun out into orbit, melted, and crystallized to form a feldspathic crust. The formation of this lunar highland crust is dated at 4440 ± 20 m.y. and represents the oldest recorded event in the Earth-Moon system. Accretion of the Moon was essentially complete at that time, but the Moon was subjected to a continuing bombardment that lasted until 3850 m.y. ago. Over 40 ringed basins over 300 km in diameter and 10,000 craters with diameters of 30–300 km were formed during this period. A similar but more intense bombardment must have struck the Earth during this period, causing chaotic interruptions to geological processes. A consensus is emerging that this bombardment was not continuous but mostly due to a spike or cataclysm around 3850–4000 m.y. Probably several spikes in the cratering flux were involved. The presence of a population of large impactors for 500 m.y. after the formation of the inner planets may be explained by the collisional disruption of a large body in the asteroid belt that provided a swarm of impacting bodies.

Magma oceans. Apart from the impacts of other major planetesimals during the formation of the Earth, the Moon-forming collision was sufficiently energetic to melt the Earth. However, there is little evidence of the sort of mineralogical or geochemical zoning in the upper mantle that typified the crystallization of the Moon and terrestrial layered intrusions and that might have been expected to result from the crystallization of a terrestrial magma ocean. Since it seems inevitable that the Earth was melted, probably the rapid crystallization of a magma ocean frustrated crystal settling and the formation of a differentiated mantle. Mantle temperatures during the Hadean were probably significantly higher due both to thermal inputs from impacts of large bodies and to a much higher heat flow. Over half the heat produced by the decay of ^{235}U to ^{207}Pb (isotopes of uranium and lead) in the Earth was released during Hadean time alone, adding as much as several hundred degrees to the Earth's internal temperature. *See* EARTH, HEAT FLOW IN; RADIOACTIVE MINERALS.

Early crusts. Observations from Mars, the Moon, Venus, and Vesta suggest that the crust of the Hadean Earth was basaltic. Enigmatic evidence suggests that there was some extraction of the relatively large elements (such as potassium, rubidium, thorium, and uranium) from the mantle at this time, but the scale of this extraction was probably minor as the evidence is restricted to very small areas of outcrop of early Archean rocks. Relict zircon crystals up to 4200 m.y. in age are the only surviving remnants of surface rocks that have been identified. The zircons appear to be derived from felsic igneous rocks. If so, small areas of such rocks presumably formed from remelting of basaltic crust that sank back into the mantle. *See* EARTH CRUST.

It is often thought that, by analogy with the Moon, the Earth formed an early plagioclase-rich (anorthositic) crust. Several reasons make this unlikely. First, the composition of the Moon is richer in calcium and aluminum than that of the terrestrial mantle, leading to the early appearance of plagioclase during crystallization of the lunar magma ocean. Second, plagioclase is unstable at shallow depths (40 km) in the Earth and will transform to garnet, thus locking up calcium and aluminum in a dense phase. In contrast, plagioclase will be stable in the Moon to depths of several hundred kilometers. Plagioclase will sink in a wet terrestrial basaltic magma. Finally, there is no sign of an ancient geochemical reservoir of europium or of primitive ^{87}Sr/^{86}Sr (strontium isotopes) that would have resided in an early strontium-rich rubidium-poor anorthostic crust.

Was there an early world-encircling crust of granite on the Hadean Earth? A primitive "sialic" crust is refuted, among other evidence, by the virtual absence of an old zircon population in younger sediments. The significant feature about the Earth, in contrast to the other terrestrial planets, appears to be the presence of liquid water at the surface that enables recycling of subducted basaltic crust through the mantle. It is this process that permits the slow production of the continental crust throughout geologic time. The conditions for the production of massive granitic crusts are probably unique to the Earth and require three or more stages of derivation from a primitive mantle composition. In other planets, the absence of subduction leads to the persistence of barren basaltic plains. Thus no good evidence exists for that enduring geological myth of a primordial crust of "sial" or granite. Such models originated through false analogies between the production of a silicic residuum during crystallization of basaltic magmas and conditions in an early molten Earth, and a lack of appreciation of the difficulties of producing granite. *See* SILICATE MINERALS; ZIRCONIUM.

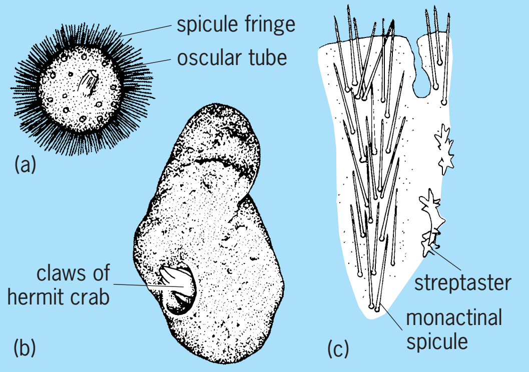
Different picture of Earth. Conditions on the Hadean Earth bore little resemblance to more recent times. A picture dimly appears of a hot young Earth with a thick basaltic crust, covered by an ocean. Dry land was rare. Plate tectonics had not yet begun. A few remnant zircon crystals indicate the formation of an occasional felsic rock, produced by remelting of the basalt. Sporadic disruption of the surface was caused by the collisions of basin-forming impactors that probably culminated in a spike or cataclysm around 3850–4000 m.y. ago. Such events must have frustrated the origin and development of life, which emerged in post-Hadean time.

Stuart Ross Taylor

Bibliography. C. J. Allègre and S. H. Schneider, The evolution of the Earth, *Sci. Amer.*, 271:44–51, 1994; P. Cloud, The Hadean Earth, in *Oasis in Space: Earth History from the Beginning*, chap. 2, Norton, New York, 1988; H. E. Newsom and J. H. Jones, *Origin of the Earth*, Oxford, 1990; P. J. Thomas, C. F. Chyba, and C. P. McKay, *Comets and the Origin of Life*, Springer, 1997.

Hadromerida

An order of sponges of the class Demospongiae with monactinal megascleres that usually have a terminal knob at one end. Microscleres in the form of streptasters, asters, sigmas, or small spined diactinals may occur. Megascleres tend to be arranged in tracts radiating from the center of the sponge; spongin is usually sparse in occurrence.



Hadromerine sponges. (*a*) *Radiella sol*, deep-sea species. (*b*) *Suberites ficus* living on shell occupied by hermit crab. (*c*) Spicule arrangement of *Spirastrella*.

In shape, hadromeridan sponges include radially symmetrical forms (see **illus.**) as well as encrusting, massive, or branching types. They occur in tidal and shallow waters of all seas and extend down to depths of at least 18,000 ft (5500 m). Included in the group are the limestone-excavating clionids. *See* BIOERODING SPONGES; DEMOSPONGIAE. Willard D. Hartman

Hadron

The generic name of a class of particles which interact strongly with one another. Examples of hadrons are protons, neutrons, the π, K, and D mesons, and their antiparticles. Protons and neutrons, which are the constituents of ordinary nuclei, are members of a hadronic subclass called baryons, as are strange and charmed baryons, for example, $\Lambda^0{}_s$ and $\Lambda^0{}_c$. Baryons have half-integral spin, obey Fermi-Dirac statistics, and are known as fermions. Mesons, the other subclass of hadrons, have zero or integral spin, obey Bose-Einstein statistics, and are known as bosons. The electric charges of baryons and mesons are either zero or ± 1 times the charge on the electron. Masses of the known mesons and baryons cover a wide range, extending from the pi meson, with a mass approximately one-seventh that of the proton, to values of the order of 10 times the proton mass. The spectrum of meson and baryon masses is not understood. *See* BARYON; BOSE-EINSTEIN STATISTICS; FERMI-DIRAC STATISTICS; MESON; NEUTRON; PROTON.

It is believed that the net number of baryons in the universe is a conserved quantity. In making this count, baryons and antibaryons are arbitrarily assigned baryon numbers $+1$ and -1, respectively, so that production or annihilation of a baryon-antibaryon pair in a given reaction has no effect on the net baryon number. (Mesons are assigned baryon number zero.) In this scheme, the least massive baryon, the proton, is stable, and all other baryons unstable. However, it has been suggested, as a result of some theoretical attempts to unify the strong, weak, and electromagnetic interactions, that protons may be unstable with a lifetime many orders of magnitude greater than the age of the universe. *See* FUNDAMENTAL INTERACTIONS; GRAND UNIFICATION THEORIES; SYMMETRY LAWS (PHYSICS).

In addition to baryon number, hadrons and their antiparticles are assigned quantum numbers to represent other properties of hadronic matter. Among these are *i*-spin (related to electric charge), strangeness, charm, and two others. Strong interactions conserve *i*-spin, and hence hadrons may form *i*-spin multiplets whose members differ in mass by only a small fraction of the proton mass, for example, the proton-neutron doublet. Hadrons with nonzero values of the strangeness and charm quantum numbers, as well as hadrons without a decay mode that conserves *i*-spin, are quasistable; their decays take place either through the weak interaction, which does not conserve strangeness or charm, or through the electromagnetic interaction, which does not conserve *i*-spin. *See* CHARM; I-SPIN; STRANGE PARTICLES; WEAK NUCLEAR INTERACTIONS.

Based on an enormous body of data, hadrons, are now thought of as consisting of elementary, fermion constituents known as quarks which have electric charges of $+\frac{2}{3}|e|$ and $\frac{1}{3}|e|$, where $|e|$ is the absolute value of the electron charge. A quark-antiquark pair makes up a meson, while three quarks constitute a baryon. To satisfy a well-tested principle concerning the assignment of quantum numbers in a closed fermion system, the quarks are assigned a quantum number (color) with three possible values, in addition to the quantum numbers signifying other quark properties such as strangeness and charm. Massless, spin-1 gluons are considered to be the propagators of the force between quarks. A quantitative quantum theory based on these ideas, called quantum chromodynamics (QCD), has been developed. *See* ELEMENTARY PARTICLE; GLUONS; QUANTUM CHROMODYNAMICS; QUARKS. A. K. Mann

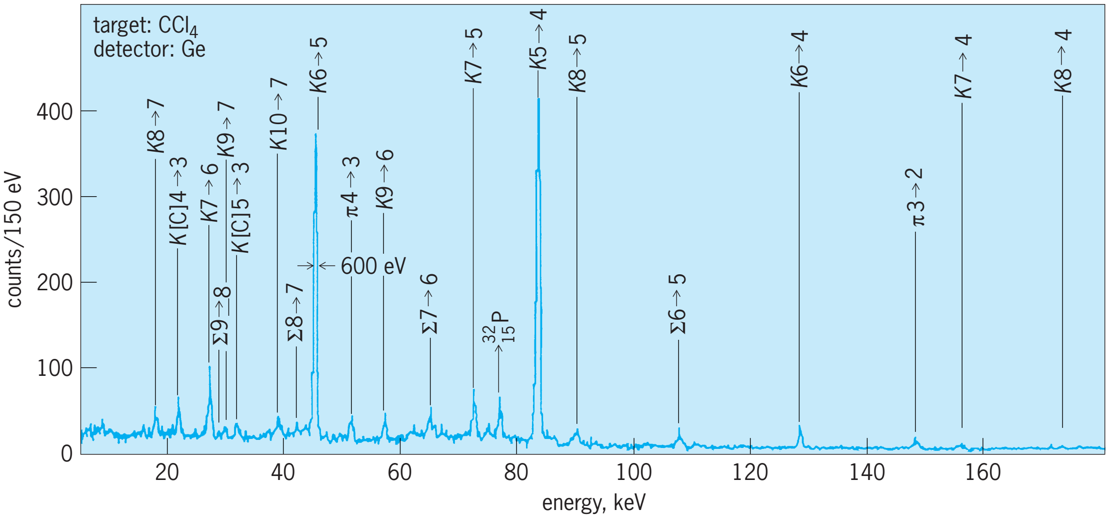
Bibliography. L. B. Okun, *Particle Physics: The Quest for the Substance of Substance*, 1985; D. H. Perkins, *Introduction to High Energy Physics*, 4th ed., 2000; C. Quigg, Elementary particles and forces, *Sci. Amer.*, 252(4):84–95, April 1985.

Hadronic atom

A hydrogenlike system that consists of a strongly interacting particle (hadron) bound in the Coulomb field and in orbit around any ordinary nucleus. The kinds of hadronic atoms that have been made and the years in which they were first identified include pionic (1952), kaonic (1966), Σ^- hyperonic (1968), and antiprotonic (1970). They were made by stopping beams of negatively charged hadrons in suitable targets of various elements, for example, potassium, zinc, or lead. The lifetime of these atoms is of the order of 10^{-12} s, but this is long enough to identify them and study their characteristics by means of their x-ray spectra. They are available for study only in the beams of particle accelerators. Pionic atoms can be made by synchrocyclotrons and linear accelerators in the 500-MeV range. The others can be generated only at accelerators where the energies are greater than about 6 GeV. *See* ELEMENTARY PARTICLE; HADRON; PARTICLE ACCELERATOR.

The hadronic atoms are smaller in size than their electronic counterparts by the ratio of electron to hadron mass. For example, in pionic calcium, atomic number $Z = 20$, the Bohr radius of the ground state is about 10 fermis (1 fermi $= 10^{-15}$ m), and in ordinary calcium it is about 2500 fermis. Thus the atomic electrons are practically not involved in the hadronic atoms, and the equations of the hydrogen atom are applicable. The close approach of the hadrons to their host nuclei suggests that hadron-nucleon and hadron-nucleus forces will be in evidence, and this is one of the motivations for studying these relatively new types of atoms.

X-ray emissions. Negative hadrons are captured into orbits of large principal quantum number n by



X-ray spectrum resulting from kaons stopped in carbon tetrachloride. Lines from Σ^- hyperonic atoms are seen along the kaonic x-rays. The pionic lines came from decay products. Nuclear gamma rays of the first excited state of phosphorus-32 are seen at 78 keV. (*Lawrence Berkeley Laboratory*)

the attraction of the positively charged nuclei. As the hadrons fall through successively smaller Bohr orbits (cascade), electrons are ejected from the cloud of atomic electrons (Auger effect). When a hadron has reached about the same radius as that of the electronic atom ground state, x-ray emission becomes the dominant method for the system to shed its excitation energy. X-rays, whose energy increases with each successive jump, are emitted until the hadron reaches the ground state ($n = 1$) of the hadronic atom or is absorbed by the nucleus in a strong interaction. *See* ATOMIC STRUCTURE AND SPECTRA; AUGER EFFECT; X-RAYS.

The precise measurement of the energies and shapes of the x-ray transitions not perturbed by the strong interaction is an excellent method to determine masses and magnetic moments of the hadrons. This is the case for the masses of negative pions, kaons, antiprotons, and Σ^-, as well as the magnetic moments of the last two.

The lines of the spectra of special interest are due to transitions between the lowest quantum levels because the hadrons are then closest to the nuclei and the nuclear forces perturb the orbits. The effects expected and observed are that some of the lines are slightly broadened (energy indefinite) and the average transition energy is different from that predicted solely on the basis of Coulomb effects.

The series of x-ray lines generally cuts off rather abruptly at $n > 1$. However, in light pionic atoms the ground state is reached. Kaons, Σ^-, and antiprotons ($\bar{p}$) can probably reach the ground state only when the nucleus is singly charged. In kaonic chlorine atoms, for example, the series ends at $n = 3$, and in kaonic lead the series ends at $n = 7$.

Experimentally, the hadrons are generated by a beam of protons incident on a metallic target. A secondary beam is used to transport the particles to the target in which the atoms are to be made and studied. The arrival of a hadron is signaled by a set of scintillation counters as it is slowed down by passage through a moderator of carbon or beryllium. The thickness of the moderator is adjusted so that a maximum number of hadrons stops in the target under investigation. The x-ray detectors are semiconductors of silicon or germanium. Efficiencies for detecting an x-ray that comes from within the target are around 5×10^{-3}. The energy resolution of the detectors is of paramount importance. The **illustration** is an example of a kaonic x-ray spectrum of chlorine obtained by an ultrapure germanium detector. The lines are labeled according to their hadronic transitions. The intensity of the lines average about 0.3 x-ray per stopped kaon for the principal lines ($\Delta n = -1$). About 5×10^6 kaons were stopped to obtain the spectrum. In special cases where the resolution is of vital importance, as was the case in determination of the pionic mass, a crystal spectrometer is used. *See* PARTICLE DETECTOR.

Pionic atoms. The interpretation of the x-ray spectra of pionic atoms is complicated by the necessity for the pion to react with two nucleons rather than one nucleon alone, as in reaction (1), where *N* stands

$$\pi + N + N \rightarrow N + N + \text{kinetic energy} \qquad (1)$$

for either a proton or a neutron. The two-nucleon final state is required for the reaction to conserve momentum.

Through the use of the line width and energy shift data of all the measurements on many elements throughout the periodic table, a calculation was made to determine the parameters of the low-energy pion-nucleus interaction. An important ingredient of this calculation is the short-range nucleon-nucleon correlation in nuclei, which determines the probability for reaction (1). The experimental data on the pion absorption in the nuclei can be fairly well accounted for in this phenomenological approach.

Kaonic atoms. Kaonic atoms were expected to yield valuable information concerning the surface of nuclei because a kaon reacts very strongly with either a single neutron or a single proton, as in reaction (2).

$$K^- + N \rightarrow \pi + \text{hyperon} \qquad (2)$$

On the basis of theory it had been predicted that more neutrons than protons would be found on the surfaces of nuclei. One of the first interpretations of the behavior of the series of x-ray lines of various elements ranging from $Z = 3$ through 92 suggested that the neutron dominance of nuclear surfaces was verified. However, it was later pointed out that a resonance between K^- and proton, $Y^*(0)1405$, would probably enhance the affinity of kaons for protons on the nuclear outskirts, and this could account for the increased capture rate observed experimentally. The data on the kaonic atoms could not be analyzed uniquely in terms of the difference of the proton and neutron distributions at the nuclear surface. The K^- nucleon resonances at the threshold complicate the parametrization of elementary interaction (2) so much that the present quality of the data is not sufficient to determine the elementary interaction on the one hand and the nuclear matter distribution on the other. *See* NUCLEAR STRUCTURE.

Σ^- hyperonic atoms. When kaons are absorbed by nucleons, about 20% of the hyperons produced are Σ^- particles. In the light elements most of the Σ^- hyperons are ejected from the nucleus in which they are generated. Some of them are captured by target nuclei; Σ^- hyperonic atoms are formed and emit characteristic x-rays, just as do the kaonic atoms. Weak x-ray lines due to the hyperonic atoms are found along with the kaonic x-ray lines (see illus.). The hyperonic lines are of special interest because they are actually doublets due to the magnetic moment of the Σ^-. X-ray lines of Σ^- atoms in some heavy elements have been measured, and the Σ^- magnetic moment was determined to be $\mu = -(1.097 \pm 0.044)$ nuclear magnetons with an additional systematic error of $\pm$ 0.04 nuclear magneton. Perhaps Σ^- will turn out to be a suitable probe of the nucleus.

Antiproton atoms. Antiproton atoms are the latest in the series of hadronic atoms to be observed. Their x-ray lines are doublets due to the magnetic moment of $\bar{p}$. The splitting has been measured and μ found to be -2.795 ± 0.019 nuclear magnetons, which agrees very well with the proton magnetic moment but has a sign opposite to that expected for antiparticles.

The main research effort involving antiproton atoms has been dedicated to the investigation of the x-ray spectra of the antiprotonic hydrogen. The transitions to the ground state depend directly on the elementary antiproton-proton interaction at the threshold. If this interaction turns out to be simple enough, the antiprotonic atoms will be a future tool for measuring the matter distribution of the nuclear surface. Another source of low-energy antiprotons—the Low Energy Antiproton Ring (LEAR), which makes precision measurements on antiprotonic atoms feasible—was put into operation at CERN near Geneva, Switzerland.

There are two additional hadrons with lifetimes that are long enough to be candidates for hadronic atom formation: the negative xi (Ξ^-) and the negative omega (Ω^-), but even at the largest accelerators, these particles are too scarce for their atoms to be detected. Clyde E. Wiegand; Bogdan Povh

Bibliography. C. J. Batty, Exotic atoms, *Sov. J. Part. Nuc.*, 13:71–96, 1982; J. B. Warren (ed.), *Nuclear and Particle Physics at Intermediate Energies*, 1976.

Haemophilus

A genus of gram-negative, pleomorphic bacteria that are facultative anaerobes and are nonmotile and non-spore-forming. They have fastidious growth requirements, including either factor X (provided by hemin or other porphyrins) or factor V [provided by nicotinamide adenine dinucleotide (NAD), nicotinamide adenine dinucleotide phosphate (NADP), or nicotinamide nucleoside] or both, depending on the species. Erythrocytes represent an exogenous source of these factors.

Classification. *Haemophilus influenzae* was the first of the species to be isolated and is considered the type species. It was originally recovered during the influenza pandemic of 1889 and for a time was believed to be the causative agent of influenza; thus it was called the influenza bacillus. However, when this fallacy became apparent, the organism was renamed, still reflecting the historical association with influenza.

Haemophilus species are distinguished by a number of criteria, including the requirement for factor X or factor V or both (see **table**), the pattern of sugar fermentation, the ability to lyse horse erythrocytes, and the presence of catalase. Strains of *H. influenzae* and *H. parainfluenzae* can be subdivided into biotypes according to tests for the production of indole and the presence of urease and ornithine decarboxylase. Strains of *H. influenzae* can be separated into encapsulated and nonencapsulated forms. Encapsulated strains express one of six biochemically and antigenically distinct capsular polysaccharides that are designated serotypes a through f. Nonencapsulated strains are defined by their failure to agglutinate with antisera against the known capsular types and are referred to as nontypable. *See* INFLUENZA; MENINGITIS.

Human infections. *Haemophilus influenzae* is a human-specific pathogen that inhabits the upper respiratory tract and is acquired by exposure to airborne droplets or contact with respiratory secretions. Nontypable strains can be isolated from the nasopharynx of up to 80% of normal children and adults at any given time, usually in association with asymptomatic colonization. In the setting of compromised mucociliary clearance (due to viral respiratory infection, exposure to cigarette smoke, or underlying cystic fibrosis, chronic bronchitis, or allergic disease), nontypable *H. influenzae* can spread contiguously to produce localized disease in the lungs (bronchitis,

***Haemophilus* species associated with human disease**

Species	Growth factor X	Growth factor V	Primary clinical manifestations
H. influenzae nontypable	+	+	Exacerbation of chronic bronchitis, otitis media, sinusitis
serotype b			Bacteremia, meningitis, epiglottitis, pneumonia, septic arthritis
serotype a, c-f			Meningitis, pneumonia, septicemia
biotype IV			Neonatal septicemia, amnionitis, maternal bacteremia, acute salpingitis, tubo-ovarian abscess
biogroup aegyptius			Purulent conjunctivitis, Brazilian purpuric fever
H. parainfluenzae	−	+	Meningitis, endocarditis, brain abscess, genitourinary tract infection
H. aphrophilus	+*	−	Endocarditis, brain abscess
H. paraphrophilus	−	+	Endocarditis, brain abscess
H. haemolyticus	+	+	Pharyngitis
H. parahaemolyticus	−	+	Endocarditis, pharyngitis
H. segnis	−	+	Endocarditis
H. ducreyi	+	−	Chancroid (genital ulcers)

*Although *H. aphrophilus* was originally reported to require factor X for growth, recent studies suggest that it can grow independent of this factor.

pneumonia), middle ears (otitis media), or sinuses (sinusitis). Overall, these organisms are the leading cause of exacerbations of chronic bronchitis, and the second most common etiology of acute otitis media and sinusitis (see table).

On occasion, nontypable *H. influenzae* causes invasive disease such as meningitis, septicemia, endocarditis, epiglottitis, or septic arthritis. Invasive disease occurs most often in neonates and in patients with underlying immunodeficiency, especially when abnormalities in humoral immunity are present. The majority of individuals who develop nontypable *H. influenzae* meningitis have a passage between the respiratory tract and the subarachnoid space. Typical symptoms include fever, headache, and stiff neck. Since the 1970s, nontypable *H. influenzae* has emerged as an important cause of neonatal septicemia, particularly in premature infants. In most infants, symptoms develop within the first few hours of life, with manifestations of respiratory distress dominating the presentation. A large proportion of isolates from these patients are biotype IV, a subgroup of nontypable *H. influenzae* that is uncommon at other sites of infection. In 1984 a fulminant septicemic illness due to *H. influenzae* biogroup aegyptius was recognized in children in São Paolo and was called Brazilian purpuric fever. *Haemophilus influenzae* biogroup aegyptius is a subgroup of *H. influenzae* biotype III that was first described in the late 1800s and has a predilection for causing purulent conjunctivitis. In most patients, Brazilian purpuric fever begins with purulent conjunctivitis, which is followed 3–15 days later by fever, abdominal pain, vomiting, and vascular collapse. Although conjunctivitis due to *H. influenzae* biogroup aegyptius occurs in the United States and other countries, cases of Brazilian purpuric fever have been reported only in Brazil and Australia, apparently because only select clones are endowed with the necessary virulence determinants.

Encapsulated strains of *H. influenzae* are present in the nasopharynx of only 2–5% of children and an even smaller percentage of adults. Historically, *H. influenzae* type b strains were the primary cause of childhood bacterial meningitis and a majority of other bacteremic diseases in children. However, in recent years the incidence of disease due to *H. influenzae* type b has plummeted in the United States and other developed countries, reflecting the routine use of *H. influenzae* conjugate vaccines. The existing *Haemophilus* vaccines contain the type b capsular polysaccharide (a polymer of ribose and ribitol-5-phosphate) conjugated to an immunogenic carrier protein. These vaccines provide effective protection against disease due to *H. influenzae* type b but fail to protect against non-type b strains.

Haemophilus aphrophilus, *H. haemolyticus*, *H. parahaemolyticus*, *H. parainfluenzae* and *H. segnis* are members of the normal flora in the human oral cavity and oropharynx and have low pathogenic potential. *Haemophilus aphrophilus*, *H. haemolyticus*, and *H. segnis* are often present in dental plaque and gingival scrapings. *Haemophilus parainfluenzae* is occasionally isolated from the urethra and vagina. Among these species, *H. parainfluenzae* is the most common pathogen and has been reported in association with a variety of diseases, including meningitis, endocarditis, brain abscess, epidural abscess, epiglottitis, pneumonia, empyema, septicemia, and urinary and genital tract infections. *Haemophilus aphrophilus* and *H. paraphrophilus* are responsible for a similar spectrum of diseases (see table). *Haemophilus haemolyticus* and *H. parahaemolyticus* rarely produce illness in humans.

Haemophilus ducreyi is distantly related taxonomically to other *Haemophilus* species and is the causative agent of chancroid, a sexually transmitted disease characterized by genital ulceration and inguinal adenitis. Transmission occurs during sexual contact when the organism penetrates a break in the epithelium. Following an incubation period of 4–7 days, a papule develops, usually surrounded by a rim of redness. Over the next 2–3 days, the papule evolves to a pustule, which ruptures to form a sharply circumscribed ulcer. Occasionally multiple ulcers coalesce to form giant ulcers. Other diseases associated with a similar presentation include syphilis, lymphogranuloma venereum, donovanosis, and genital herpes. Chancroid is most common in developing countries in Africa, Asia, and Latin America. Although relatively uncommon in the United States, a number of large outbreaks have been identified since 1981, usually revolving around prostitution and illicit drug use. Evidence suggests that chancroid may be an important cofactor in heterosexual spread of human immunodeficiency virus (HIV). *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); SEXUALLY TRANSMITTED DISEASES.

Therapy. Strains of *H. influenzae* are increasingly resistant to a wide variety of antibiotics, including penicillins, erythromycin, trimethoprim-sulfamethoxazole, and occasionally chloramphenicol. In most cases, penicillin resistance is mediated by a plasmid-encoded beta-lactamase, which has been detected in up to 60% of isolates in some geographic areas. Accordingly, an extended-spectrum cephalosporin is generally recommended for empiric treatment of serious disease. For localized respiratory tract infection, options include a combination of amoxicillin and clavulanate (a beta-lactamase inhibitor), an oral second- or third-generation cephalosporin, trimethoprim-sulfamethoxazole, azithromycin, or clarithromycin. Among isolates of *H. ducreyi*, plasmid-mediated resistance to ampicillin, chloramphenicol, tetracyclines, and sulfonamides is well described. Effective agents for the treatment of chancroid include ceftriaxone, azithromycin, erythromycin, ciprofloxacin, and amoxacillin-clavulanate. Traditionally, other *Haemophilus* species have been treated with ampicillin. However, resistance to ampicillin, usually related to betalactamase production, is recognized with increasing frequency. As a result, a second- or third-generation cephalosporin should be used until antibiotic susceptibility results are available. *See* ANTIBIOTIC; DRUG RESISTANCE.

Veterinary diseases. *Haemophilus agni* is the etiologic agent of acute septicemia in lambs. This illness typically presents as a fulminant process with death in less than 12 hours.

Haemophilus paragallinarum commonly colonizes the nasal passages of birds and is the causative agent of fowl coryza, which manifests as purulent nasal discharge.

Haemophilus parasuis is a normal resident in the oropharynx of swine and is the etiologic agent of Glasser's disease. Infection typically occurs on serosal surfaces in a wide range of locations, including the meninges (brain), pleura (lung), pericardium (heart), peritoneum (abdominal cavity), and synovia (joints). This disease occurs in young swine and presents with fever, respiratory distress, convulsions, and paresis (focal weakness).

Haemophilus parahaemolyticus typically infects young swine via the respiratory route, leading to pneumonia, which is sometimes accompanied by meningitis or arthritis.

Haemophilus somnus can be found in the nasopharynx and female genital tract of healthy cattle and causes infectious thromboembolic meningoencephalitis. This disease typically presents with abrupt onset of fever, weakness, and somnolence. Other neurological signs include ataxia, blindness, and paralysis. Occasionally pneumonia and arthritis are present. Alternatively, *H. somnus* can cause isolated arthritis, endometritis, or pneumonia, especially in young cattle. *See* MEDICAL BACTERIOLOGY.

Graham P. Krasan; Joseph W. St. Geme

Bibliography. S. J. Barenkamp, Other *Haemophilus* species, in R. D. Feigen and J. D. Cherry (eds.), *Textbook of Pediatric Infectious Disease*, 4th ed., W. B. Saunders, Philadelphia, 1998; J. G. Holt et al. (eds.), *Bergey's Manual of Determinative Bacteriology*, 9th ed., Williams & Wilkins, Baltimore, 1994; M. Killian, *Haemophilus*, in A. Balows et al. (eds.), *Manual of Clinical Microbiology*, 5th ed., American Society for Microbiology Press, Washington, DC, 1991; E. R. Moxon, *Haemophilus influenzae*, in G. L. Mandell, J. E. Bennett, and R. Dolin (eds.), *Principles and Practice of Infectious Diseases*, 4th ed., Churchill Livingstone, New York, 1995; J. I. Ward and K. M. Zangwill, *Haemophilus influenzae*, in R. D. Feigen and J. D. Cherry (eds.), *Textbook of Pediatric Infectious Disease*, 4th ed., W. B. Saunders, Philadelphia, 1998.

Haemosporina

A relatively small and generally rather compact group of protozoa in the subphylum Sporozoa. Authorities differ as to the group's taxonomic status; that assigned it by the Committee on Taxonomy and Taxonomic Problems of the Society of Protozoologists is followed here: a suborder of the order Eucoccida, subclass Coccidia, class Telosporea, subphylum Sporozoa. The Haemosporina are common protozoan parasites of vertebrates, and some of them are important as causes of illness and death. The best known of the group are the four species of malarial parasites of humans. Not so well known are at least 60 other species of malarial parasites, with a wide host distribution among terrestrial vertebrates, as well as numerous species of *Haemoproteus* and *Leucocytozoon*, and some species of *Hepatocystis*. All three genera are closely related to the genus *Plasmodium*, in which all the true malarial parasites are placed (see **table**). *See* MALARIA.

Transmission of these parasites is probably always effected in nature by the bite of some blood-sucking invertebrate. In the vertebrate host they reproduce asexually; sometimes this occurs in the tissues of certain internal organs, such as the lungs, liver, spleen, and brain; sometimes in the red blood cells (erythrocytes), or even in other types of blood and blood-forming cells; and often in both tissues and blood cells. The immature sex cells, gametocytes, always occur in erythrocytes or leukocytes (white blood cells). Gametocytes mature into gametes after ingestion by an intermediate host (arthropod). Fertilization ensues, with a subsequent period of development culminating in the production of numerous

Some of the better-known species of malarial parasites

Plasmodium species	Host	Vector
Human		
P. falciparum	Humans	*Anopheles* sp.
P. malariae	Humans and anthropoid apes	*Anopheles* sp.
P. ovale	Humans	*Anopheles* sp.
P. vivax	Humans	*Anopheles* sp.
Simian		
P. brasilianum	South American monkeys of various species	?
P. cynomolgi (includes several subspecies)	Monkeys, especially macaques (occasionally humans)	*Anopheles* sp.
P. inui	Monkeys, especially macaques; very widely distributed	*A. hackeri* and *A. leucosphyrus*
P. knowlesi	Monkeys, especially *Macaca irus* (occasionally humans)	*A. hackeri* and probably others
P. reichenowi	Anthropoid apes	?
Rodent		*A. dureni*
P. berghel	Congo tree rat; laboratory rats and mice	
Avian		
P. cathemerium	Numerous avian species (mostly passerines)	*Culex* sp. (possibly some anophelines)
P. circumflexum	Numerous avian species (mostly passerines)	*Culex* sp. and *Theobaldia* sp.
P. elongatum	Numerous species of passerines and others	*Culex* sp.
P. gallinaceum	Natural hosts Asiatic wildfowl; also occurs in domestic fowl	*Aedes* sp.
P. hexamerium	Numerous species of passerines	?
P. juxtanucleare	Domestic fowl and partridges (found originally in Brazil, but probably from Far East)	*C. pallens* and *C. sitiens*
P. nucleophilum	Catbirds and others	?
P. pinottii	Toucan of Brazil (various laboratory hosts, especially pigeon)	?
P. polare	Cliff swallow in North America; in other birds elsewhere	?
P. relictum	Numerous species of passerine and other birds	*Culex* sp., *Aedes* sp., *Theobaldia* sp., and probably even some anophelines
P. rouxi	English sparrows in Near East; elsewhere in other species (occurrence questionable in New World)	*C. pipiens*
P. vaughani	Mainly in robin in North America; in a number of other passerines elsewhere	?
Reptilian		
P. mexicanum	Lizards of a number of species	?

sporozoites. These tiny filamentous forms are infective for the vertebrate host. Since they can develop no further in the arthropod, infection in this host is self-limited, in the sense that no further buildup is possible; the insect is seldom harmed by the parasite. However, sporozoites may remain infective for a long time in the invertebrate host, perhaps as long as the insect lives.

Taxonomy. Protozoologists disagree about the number of families that should comprise the suborder Haemosporina. Some include two families, the Haemoproteidae and Plasmodiidae; others combine these two as a single family. Still others recognize another family, the Babesiidae, as belonging in the suborder. However, the Babesiidae, containing parasites of the red blood cells and usually referred to as babesias or piroplasms, differ in certain important respects, one of them being that transmission is by ticks and another that, although they inhabit red blood cells, they produce no pigment (hemozoin) from the homoglobin.

Subgenera. With the steady increase in the number of species of *Plasmodium* regarded as valid, proposals have been made to place all species infecting birds into four subgenera: *Haemamoeba,* to include parasites that produce round gametocytes; *Huffia,* to include those that produce elongate gametocytes, if all types of blood and blood-forming cells are parasitized; *Giovannolaia*, to include only those that produce relatively large erythrocytic schizonts; and *Novyella*, to include those that produce small erythrocytic schizonts. Similarly, it has been proposed to put mammalian species of *Plasmodium* into three subgenera: *Plasmodium, Laverania*, and *Vinckeia*. Differentiaton would be based on a variety of characters, but chiefly on shape of the gametocytes (crescentic for *Laverania* and spherical for *Plasmodium* and *Vinckeia*), and exoerythrocytic stages (primary and secondary for *Plasmodium*, primary only in the other two). Species infecting primates other than lemurs would fall into the first two of these three genera. Reptilian species of *Plasmodium* are less easily grouped, but three subgenera, *Sauramoeba*, *Carinamoeba*, and *Ophidiella*, have been created. Those species with larger schizonts go into the first, and species with smaller schizonts into the second; and the third genus is reserved for the single species known from snakes.

Plasmodiidae. A family of Haemosporina inhabiting the erythrocytes of the vertebrate host (and, in a few cases, other blood and blood-forming cells), in which both asexual reproduction and the formation of gametocytes occur. Hemoglobin metabolism results in the formation of a by-product, hemozoin, typical of these parasites. The invertebrate host is generally thought always to be a mosquito, but for many species it is still unknown.

The Plasmodiidae are the true malarial parasites. Although there can be many types of host, the majority of known species are found in birds, mammals, and reptiles (especially lizards). Though, according to P.C.C. Garnham, the avian malarial parasites are “probably of the greatest antiquity,” those of reptiles may be even older. Many of the reptilian and avian species of *Plasmodium* parallel one another very closely; this similarity in itself suggests that parasitism in this group of Haemosporina may have occurred in birds and reptiles much earlier than in mammals. Although primates seem to be the preferred hosts among the mammals, malaria also occurs in a number of other mammals, such as rats, squirrels, bats, antelope, and water buffalo. Some mammalian blood parasites formerly regarded as species of *Plasmodium* have been transferred to the genus *Hepatocystis* because they reproduce only in the tissues (in this case, the liver), as does *Haemoproteus*. It is thought that the malarial parasites of mammals may have evolved from *Hepatocystis;* the host distribution of both is much the same, though the latter is more restricted.

Haemoproteidae. Only the gametocytes of this family of Haemosporina occur in blood cells, the asexual stages being confined to the internal organs. If occurring in the erythrocytes, the parasites produce hemozoin as in malaria, except in the case of *Leucocytozoon*. This parasite was long thought to invade only leucocytes (usually assumed to be lymphocytes), in which it caused changes such as to make the type of host cell unrecognizable, but these cells have been shown to be erythrocytes, at least in some species. This finding is additional evidence of the close relationship of the Haemoproteidae to the true malarial parasites. The invertebrate host, where known, is a blood-sucking fly, such as *Lynchia* or *Simulium*, or a midge of the genus *Culicoides*. *Lynchia* may transmit *Haemoproteus columbae* of the pigeon, *Simulium* (blackfly) may transmit *Leucocytozoon* of ducks, and *Culicoides* may transmit both *Hepatocystis* and certain species of *Haemoproteus*.

Included in the family Haemoproteidae are the three genera *Haemoproteus, Leucocytozoon,* and *Hepatocystis*. It is believed (though without much proof) that parasites of the first two genera are highly host-specific. Whether this is also true of *Hepatocystis* is unknown. If anything like strict host specificity exists, there are indeed many species in all three genera. *Haemoproteus* and *Leucocytozoon* occur very widely in birds throughout the world, but for some reason *Hepatocystis* is limited to the Old World, and is found most commonly in rodents, bats, and monkeys. *Haemoproteus* may sometimes occur in reptiles.

Vector-host, parasite relations. The vectors for many species of *Haemoproteus, Leucocytozoon,* and *Hepatocysti* remain undiscovered. It is believed that species with mammalian hosts are usually transmitted by mosquitoes of the genus *Anopheles*, and those of birds by *Aedes* or *Culex*. This aspect of the life cycle of the reptilian malarias is almost wholly unknown, though development to the oocyst stage has been demonstrated in three species of *Culex* which had fed on lizards infected with *Plasmodium floridense*. Since well over 2700 species of mosquitoes are known, with almost equal variety in host preferences and biting habits, there is much room for research. The vector for only one species of *Hepatocystis*, *H. kochi* of African monkeys, has been discovered; it is a midge, *Culicoides adersi*.

It is noteworthy that many individuals are often malaria-resistant, even when bitten by a mosquito species known to be a very effective malaria vector. Some degree of racial immunity also exists. Blacks are less susceptible to vivax malaria than whites, and often more resistant to the consequences of falciparum malaria. In the latter case it is usually because they are heterozygotes for the red blood cell defect known as sickling; homozygotes are likely to be so handicapped that they fail to reach maturity.

A somewhat similar situation exists among mosquitoes, at least in the species *Culex pipiens*. C. Huff showed many years ago that resistance to the avian malarial parasite (*Plasmodium cathemerium*) behaved as a simple mendelian recessive.

Pathogenicity. The Haemosporina vary greatly in pathogenicity, as with other large groups of parasites, and while they are not on the whole very harmful to their hosts, there are some notable exceptions. Human malaria, though on the wane in many parts of the world, was and is a very important disease. Though there are no statistics for the number of cases that occur annually, it is likely that there are still 200,000,000; the majority occur in Africa and southeastern Asia.

With the discovery that several species of simian malaria may be transmitted to humans by mosquitoes (especially of the so-called leucosphyrus group of anophelines, which have a strong preference for primate blood), these malarias have become potentially important. Fortunately, humans are completely susceptible to only 4 of the 23 species and subspecies of simian *Plasmodium* generally recognized as valid; humans are partially susceptible to at least 4 more.

It is also possible that the numerous species of *Plasmodium* occurring in the lower vertebrates may play a significant part in the maintenance of biological balances, particularly since such infection is often acquired early in the life of the host, when susceptibility is likely to be greatest.

Certain species of *Haemoproteus* and *Leucocytozoon* are known to be quite pathogenic to their avian hosts. Notorious examples are *H. lophortyx* in various species of quail, *L. simondi* (*anatis*) in ducks, and *L. smithii* in turkeys. The incidence of infection with these three species is often high, and the latter two have been the causes of disastrous and fulminating epizootics. Young birds are especially susceptible.

Control of haemosporidian infections always depends primarily on control of the vectors. This usually involves thorough screening and the intelligent use of insecticides: DDT is still one of the most valuable, despite the tendency of insects to develop resistance to it. (That it is also highly poisonous to humans, pets, and useful insect species must never be forgotten.)

Life cycles. The life cycles of the various Haemosporina seem to be very similar, although many are still incompletely known, and are exemplified by the malaria parasites of humans. Such life cycles also closely resemble those of the coccidia, a group of sporozoans to which the Haemosporina are closely related, and from which they probably evolved.

Sporozoites (the infective forms) introduced by the vector develop into intracellular asexually reproducing stages (exoerythrocytic forms) in the tissues. In malaria, after one or several generations of this stage, some of the young parasites (merozoites) escape into the bloodstream and infect red blood cells. Reproduction continues in these cells (and the exoerythrocytic cycle may also persist, it is believed) sometimes for years. Some parasites go on to further asexual reproduction, while others under unknown stimuli become gametocytes. *Haemoproteus* and *Leucocytozoon* undergo similar life cycles, except that reproduction is limited to the internal organs. So does *Hepatocystis*, though its behavior in *Culicoides* (a vector) presents interesting variations. The sexual phase of the cycle occurs only in the vectors and involves gametogenesis, fertilization, growth of the oocyst, and eventual development of large numbers of sporozoites. Some of these migrate to the salivary glands, where they await an opportunity to reach another vertebrate host.

One of the interesting aspects of the haemosporidian life cycle is its frequent correlation with the physiological activities of the host. The periodicity of asexual reproduction is known to be conditioned by both the genetic constitution of the parasite and diurnal variations in host physiology, at least in some species. Relapses in *Leucocytozoon* infections, and perhaps in malaria, are particularly common in the spring because of seasonal changes in reproductive activity.

Biological aspects. Much of what is known about the basic biology of malaria is the result of research on the malarias of the lower animals. Earlier investigations were done chiefly on malaria-infected birds, especially canaries, and later on ducks and chickens. *Plasmodium relictum* (*praecox*), *P. cathemerium*, *P. lophurae*, and *P. gallinaceum* were the species most used, and *P. fallax* has been added to the list.

Plasmodium berghei, a species occurring naturally in the African tree rat, has been found to be a very convenient tool for research on malaria in rats and mice. Their use has provided much knowledge about the life cycle of the malaria parasites, the mechanism of relapse, and chemotherapy.

Use of the electron microscope has supplied information about the ultrastructure of the malaria parasites. An especially interesting finding was the demonstration of an organelle interpreted as a cytostome in the erythrocytic stages of three avian and two simian species of *Plasmodium*. Though essentially similar in all five species, the organelle was about twice the size in the avian parasites (140–190 micrometers) as in the simian parasites (50–80 μm). By the use of this organelle, minute bits of host cell cytoplasm are said to be ingested. Food vacuoles form around these particles and, as digestion proceeds, malaria pigment appears.

A similar structure has also been observed in exoerythrocytic merozoites, but in them it does not seem to function in the ingestion of food. Its real use, and the uses of other organelles, remains obscure.

Sporozoites also have a cytostome, or micropyle, as it is called by Garnham and associates. They thought it might function as an avenue through which the contents of the parasite escape into the host liver or other cells to initiate the exoerythrocytic cycle. Not the least interesting of Garnham's findings is the resemblance of sporozoite structure to that of *Toxoplasma*. Garnham's studies included three of the species causing human malaria.

One of the most important unsolved problems in malariology is a simple and reliable method of culturing the parasites. Nevertheless, through the use of existing cultural methods, much has been learned about parasite physiology, which is known to be remarkably like that of the host cell. Glucose is the chief fuel of the erythrocytic stages, and both the Krebs and Embden-Meyerhof cycles are concerned with energy production. Little is known about the physiology of exoerythrocytic parasites, or of those in mosquitoes.

Protein requirements of the erythrocytic forms are largely met by the consumption of hemoglobin, though this must be supplemented by methionine indirectly derived from blood plasma. Also needed are *para*-aminobenzoic acid and the vitamin ascorbic acid. *See* PARA-AMINOBENZOIC ACID; ASCORBIC ACID.

The chief metabolic residue is malarial pigment, or hemozoin. Formerly thought to be the same as hematin, it is now believed to be a porphyrin-denatured protein complex. Since hemoglobins of different species (and even sometimes of different individuals) are not identical, it probably differs somewhat in different parasite and host species. *See* CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; FERTILIZATION (ANIMAL); GAMETOGENESIS; REPRODUCTION (ANIMAL); SPOROZOA; TOXOPLASMEA.

Reginald D. Manwell

Bibliography. L. R. Ash et al., *Atlas of Human Parasitology*, 4th ed., 1997; P. C. Beaver and R. C. Jung (eds.), *Animal Agents and Vectors in Human Disease*, 5th ed., 1985; R. Goldsmith, *Tropical Medicine and Medical Parasitology*, 1988; M. Tiru and W. Hennessen (eds.), *Diagnostics and Vaccines for Parasitic Diseases*, 1986.

Hafnium

A metallic element, symbol Hf, atomic number 72, and atomic weight 178.49. There are five naturally occurring isotopes. It is one of the less abundant elements in the Earth's crust. *See* PERIODIC TABLE.

Hafnium is a lustrous, silvery metal that melts at about 2222°C (4032°F). Reported values of the boiling point vary greatly, from about 2500 to about

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**

5100°C (4530 to 9200°F). There are virtually no uses of the metal other than in control rods for nuclear reactors.

The chemistry of hafnium is almost identical with that of zirconium. The similarity of hafnium to zirconium is a consequence of the lanthanide contraction, which brings the ionic radii to very nearly identical values. Before (and since) the discovery of hafnium, this element was extracted with zirconium from its ores and passed with zirconium into all derivatives. Since the chemical properties are so similar, there has been no incentive to separate the hafnium except for making nuclear studies and components of nuclear reactors. *See* ZIRCONIUM. Warren B. Blumenthal

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. Hala (ed.), *Halides, Oxyhalides and Salts of Halogen Complexes of Titanium, Zirconium, Hafnium, Vanadium, Niobium and Tantalum*, 1989; M. F. Lappert, *Comprehensive Organometallic Chemistry II: Scandium, Yttrium, Lanthanides and Actinides, and Titanium, Zirconium, and Hafnium*, vol. 4, 1995.

Hail

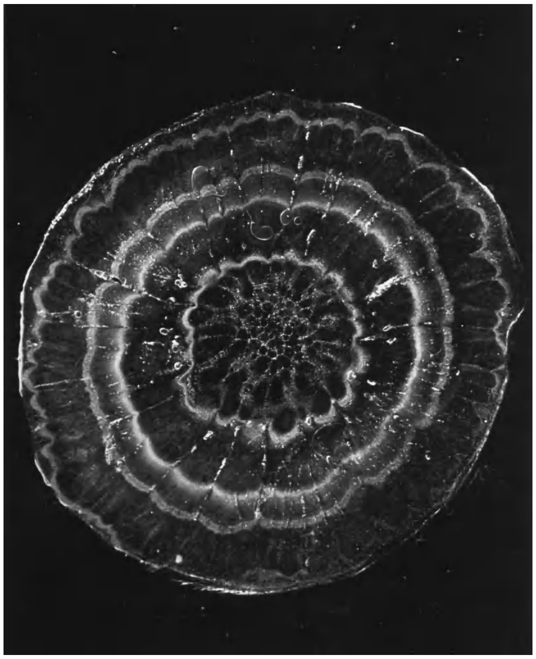
Precipitation composed of lumps of ice formed in strong updrafts in cumulonimbus clouds. Individual lumps are called hailstones. Most hailstones are spherical or oblong, some are conical, and some are bumpy and irregular. By definition, hail must have a diameter of at least 0.2 in. (5 mm), and has been known to reach over 6 in. (150 mm). Thus, the largest stones are grapefruit or softball size, and the smallest are pea size.

Hailstone structure. Very often hailstones are observed to be made of alternating rings of clear and white ice (see **illus.**). These rings indicate the growth processes of the hail. The milky or white portion of the growth occurs when small cloud droplets are collected by the hailstone and freeze almost instantaneously, trapping bubbles of air between the droplets. The clear portion is formed when many droplets are collected so rapidly that a film of water spreads over the stone and freezes gradually, giving time for any trapped air bubbles to escape from the liquid.

Conditions for formation. The most favorable conditions for hail formation occur in the mountainous, high plains regions of the world, such as in the western United States, the Russian Federation, northern India, South Africa, and Argentina. The high mountainous region of Kenya also has a very high incidence of hail.

The growth of hail normally occurs in a two-stage process. First, hailstone embryos form as graupel, or frozen rain particles. Then they grow to hailstone size as the updrafts strengthen. Hailstone embryos can also form in weaker cloud cells feeding into the main storm. Hailstorms generally have relatively high, cool cloud bases and very strong updrafts within the clouds to carry the hail embryos and hailstones into the cooler regions of the cloud, where maximum growth occurs. Both small ice particles (the embryos) and supercooled liquid water (at temperatures below 32°F or 0°C) are needed to form hailstones. A hailstone larger than 0.5–1 in. (1–2.5 cm) will change size very little during its fall to earth; one smaller than 0.4 in. (1 cm) will probably melt completely by the time it reaches the ground. The largest hailstones fall at speeds approaching 88 mi/h (40 $\mathrm{m \cdot s^{-1}}$), while the more numerous, smaller stones fall at 22–44 mi/h (10–20 $\mathrm{m \cdot s^{-1}}$).

Hailstorm modification. Much damage to property and crops is caused by hailstorms. The damage to crops depends on the time of year; much hail occurs in the late summer in the Dakotas, but earlier in states such as Illinois, where less (but still significant) damage is done. Property damage due to hail can be as great as that done to crops. In May 1995 a hailstorm caused $1 billion worth of damage in the Dallas–Fort Worth area in Texas, an amount comparable



Cross section of a large hailstone showing the structure of alternating rings of clear and white ice. (*Alberta Research Council, Edmonton*).

to the annual crop damage by hail in the entire United States. Consequently, considerable effort has been made to modify hailstorms, with limited success. The experiments and operations use silver iodide to form many small ice particles (more than nature provides) which compete for the available supercooled liquid water to become hailstones and to cause more rapid transformation from supercooled liquid water to ice. The more numerous but smaller hailstones of the modified storm melt as they fall to earth and produce rain. Controversy exists as to whether this is the way the cloud seeding works in every situation, and whether it might cause more hail and less rain in certain cases. The tremendous complexity of a hailstorm makes the experiments to test the theories very expensive.

Evidence from the late 1980s is encouraging. Target and control analyses of hail suppression in southwestern North Dakota, and in southwestern France, where a 30-year-long operational project is in place, have indicated 40–45% decreases in hail in the target areas. A randomized crossover-designed hail suppression project in Greece has indicated similar hail reduction effects. Small rain increases have been detected, but in nonrandomized tests. Experiments in the field and with hailstorm models run on supercomputers continue to reveal hailstorm processes and how they might be changed. *See* CLOUD PHYSICS; PRECIPITATION (METEOROLOGY); WEATHER MODIFICATION.

Harold D. Orville

Bibliography. R. D. Farley, Numerical modeling of hailstorms and hailstone growth, pt. III: Simulation of an Alberta hailstorm—natural and seeded cases, *J. Climatol. Appl. Meteorol.*, 26:789–812, 1987; G. Foote and C. Knight (eds.), *Hail: A Review of Hail Science and Hail Suppression*, American Meteorological Society, December 1977; C. A. Knight and P. Squires (eds.), *Hailstorms of the Central High Plains*, vols. 1 and 2, National Hail Research Experiment, National Center for Atmospheric Research, 1982; P. L. Smith et al., An exploratory analysis of crop hail insurance data for evidence of cloud seeding effects in North Dakota, *J. Appl. Meteorol.*, 36:463–473, 1997.

Hair

Nonliving, specialized epidermal derivatives characteristic only of modern mammals. However, it is now thought that hair was present in at least some therapsid reptiles. It consists of keratinized cells, tightly cemented together, which arise from the matrix at the base of a follicle. A follicle is a tubular epidermal downgrowth that penetrates into the dermis and widens into a bulb (the hair root) at its deep end (**Figs. 1** and **2**). The follicle, together with a lateral outgrowth called the sebaceous gland, forms the pilosebaceous system. Rapid cell production in the matrix, and differentiation in the regions immediately above, produces a hair shaft which protrudes from the follicle mouth at the skin surface. *See* SEBACEOUS GLAND.

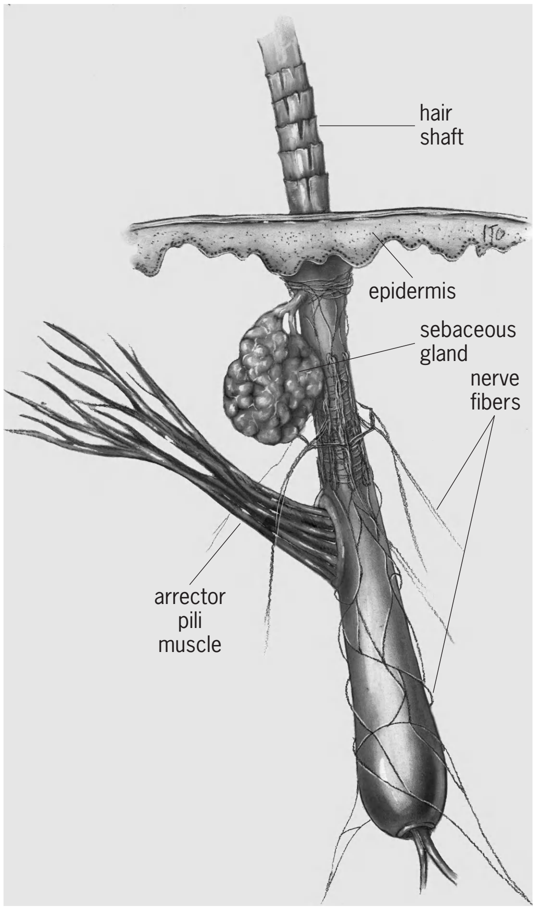


Fig. 1. Human hair follicle and associated structures.

Structure. The shaft consists of an outer cuticle and an innermost medulla, with a cortex in between (**Fig. 3**). The superficial cuticular cells overlap one another, with the free edges directed toward the hair tip. The cortex, which makes up the bulk of each hair, consists of longitudinally arranged spindle-shaped cells which are firmly attached to each other. The color of hair is due to the effect of the pigment in the cortical cells on light reflected from the central medulla. The medulla is less dense than the cortex, and its finer cells are more loosely attached to one another; it may therefore be continuous, discontinuous, or fragmented.

Growth and replacement. Hair follicles do not produce hairs continuously; growing phases alternate with periods of rest. In association with this alternation, the follicle and shaft pass through a complex series of morphological change which is referred to as the hair cycle (**Fig. 4**). During the growth phase, called the anagen phase (Fig. 4*a*), the follicle penetrates deep into the dermis; sections through its deepest portions show complex maturation patterns of cells as the various regions of the shaft are formed. Cytocrine activity of follicle melanocytes deposits pigment in the maturing cells. As the shaft attains its specific length, the follicle shortens (Fig. 4*b*), and eventually, during the quiescent or telogen phase (Fig. 4*c*), the dead "club" hair protrudes from

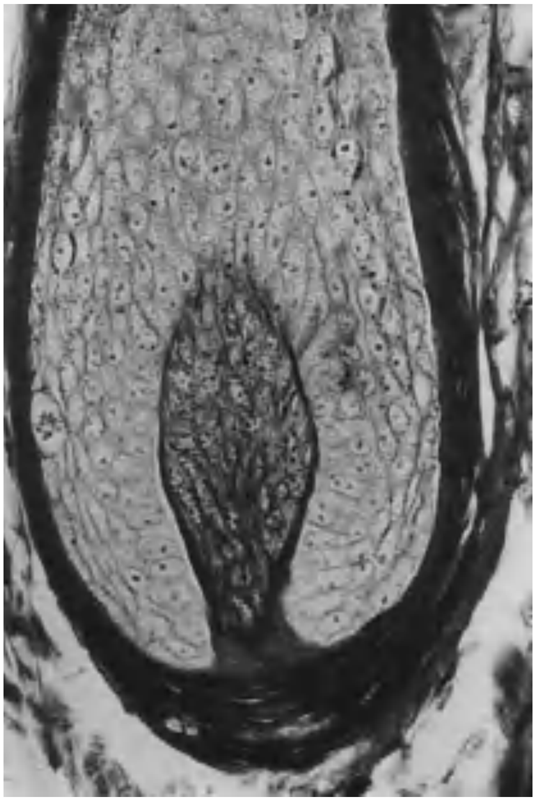


Fig. 2. Photomicrograph of longitudinal section of the lower half of a hair follicle, the root.

a follicle wherein no proliferative or differentiative activity can be seen. The growing period may be as long as 3 years or more, as in the human scalp, or as short as 12 days, as on the belly of a rat. Hairs cannot "turn gray" or "fall out" overnight (traditionally, shock is believed to have such effects), since they are dead structures joined proximally at the follicle base. However, stress may alter both the strength and the color of hairs during the growth phase, probably through a disruption of the endocrine balance during this critical period.

Hairs are not permanent structures but are continually replaced throughout the life of a mammal. In some species, for example, the rat, hamster, mouse, chinchilla, and rabbit, the replacement pattern is undulant, and waves of follicular activity can be traced across the body. In other species, for example, humans, cats, and guinea pigs, each follicle appears to cycle independently of others in the immediate area. Control of a cyclical follicular activity is imperfectly understood, but it certainly involves hormones. However, a systemic stimulant such as a hormone, if acting alone, would affect all follicles over the body, since it is carried by the bloodstream. It would be disastrous for a mammal if all its follicles were to grow in synchrony, since inevitably all functional club hairs would be shed simultaneously, thus depriving the animal of its thermal insulation. It is remarkable that both mammalian hairs and avian feathers, which function as thermal insulators, are developed and shed in various asynchronous patterns, so that the animals are never "naked." Since both hairs and feathers are epidermal appendages controlled and maintained by dermal papillae, it seems probable that part of the secret of the subtle cyclic control mechanisms must lie in these mesenchymal cells. *See* FEATHER.

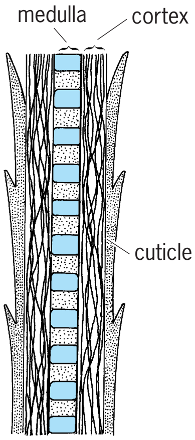


Fig. 3. Longitudinal section of a hair shaft.

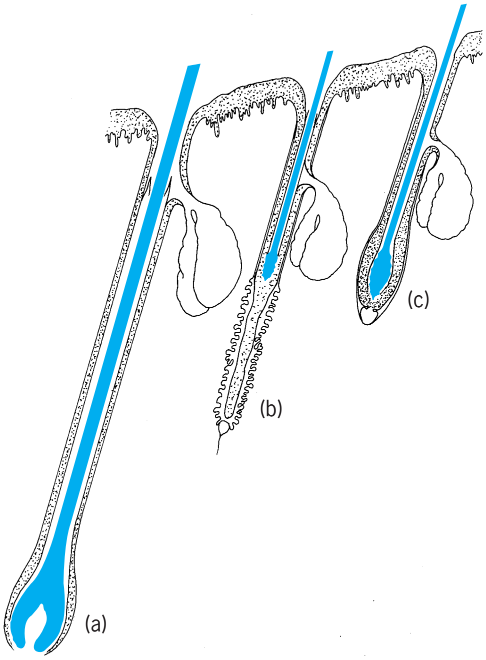


Fig. 4. Stages of the hair growth cycle. (*a*) Growing (anagen). (*b*) Intermediate. (*c*) Quiescent (telogen).

Apart from the sexual or ecological significance of annual pelage changes in many mammals, wild species could not survive if their follicles grew continuously. Merino sheep, whose follicles have a continuous (or, at least, extremely long) growth phase, are prized for the length and texture of their hair shaft, which makes the finest wool. However, strays which are not sheared regularly become entangled in shrubs and bushes and die.

Evolution and adaptive radiation. The major function of hairs in modern mammals is that of thermal insulation. However, individual hairs constitute the pelage, and, strictly, it is the pelage which insulates, since single hairs alone do not have an insulatory function. A major evolutionary problem has been to determine how such complex individual units, constituting such a complex system, could have been produced by natural selection. Mammalian hairs were probably "protoadapted" to form a pelage; that is, they performed another function before they became units in an insulatory system. This ancestral function was initially the subject of selection; and eventually, as a by-product, the units became sufficiently dense to have an insulatory function which was later improved by selection.

It seems most likely that the primary, ancestral function was that of mechanoreception, and this purpose is suggested by the existence of so-called tactile or sinus hairs in most mammals, also known as vibrissae or whiskers. Such structures, which are characterized by a blood-filled sinus in the dermis surrounding the follicle, which acts as an erectile tissue, are associated with a complex network of nerve fibers. Although all hairs seem to have some degree of sensory function, most of the body is covered by coat hairs which lack the complex, innervated sinus. They may occur as coarse, tough guard hairs, as in pig bristles or porcupine quills, especially, or as short, fine underhairs grouped around the guard hairs. Although guard hairs serve predominantly for physical protection, the dense underhairs fulfill the insulatory function.

Humans appear to lack vibrissal-type hairs, but show a diversity of coat hair at different stages of life. Infants have fine, unpigmented lanugo hairs, which are replaced by coarser terminal hairs in the adult. Vellus hairs are intermediate in texture and in time of appearance. The time sequence of replacement of lanugo hair by terminal hairs varies according to the region of the body. *See* WOOL. P. F. A. Maderson

Bibliography. C. E. Orfanos et al. (eds.), *Hair Research: Status and Future Aspects*, 1981; R. I. Spearman and P. A. Riley (eds.), *The Skin of Vertebrates*, 1981.

Half-life

The time required for one-half of a given material to undergo chemical reactions; also, the average time interval required for one-half of any quantity of identical radioactive atoms to undergo radioactive decay.

Chemical reactions. The concept of the time required for all of the material to react is meaningless, because the reaction goes very slowly when only a small amount of the reacting material is left and theoretically an infinite time would be required. The time for half completion of the reaction is a definite and useful way of describing the rate of a reaction.

The specific rate constant k provides another way of describing the rate of a chemical reaction. This is shown in a first-order reaction, Eq. (1), where c_0 is

$$k = \frac{2.303}{t} \log \frac{c_0}{c} \tag{1}$$

the initial concentration and c is the concentration at time t. The relation between specific rate constant and period of half-life, $t_{1/2}$, in a first-order reaction is given by Eq. (2). In a first-order reaction, the period

$$t_{1/2} = \frac{2.303}{k} \log \frac{1}{1/2} = \frac{0.693}{k} \tag{2}$$

of half-life is independent of the initial concentration, but in a second-order reaction it does depend on the initial concentration according to Eq. (3).

$$t_{1/2} = \frac{1}{kc_0} \tag{3}$$

See CHEMICAL DYNAMICS. Farrington Daniels

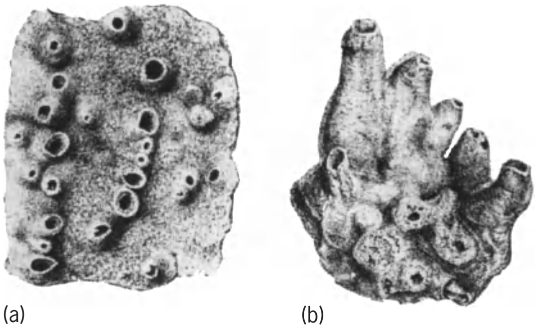
Radioactive decay. The activity of a source of any single radioactive substance decreases to one-half in 1 half-period, because the activity is always proportional to the number of radioactive atoms present. For example, the half-period of ^{60}Co (cobalt-60) is $t_{1/2} = 5.3$ years. Then a ^{60}Co source whose initial activity was 100 curies will decrease to 50 curies in 5.3 years. The activity of any radioactive source decreases exponentially with time t, in proportion to exp $-0.693t/t_{1/2}$. After 1 half-period (when $t = t_{1/2}$) the activity will be reduced by the factor $e^{-0.693} = {}^1\!/_2$. In 1 additional half-period this activity will be further reduced by the factor $^1\!/_2$. Thus, the fraction of the initial activity which remains is $^1\!/_2$ after 1 half-period, $^1\!/_4$ after 2 half-periods, $^1\!/_8$ after 3 half-periods, $^1\!/_6$ after 4 half-periods, and so on.

The half-period is sometimes also called the half-value time or, with less justification, but frequently, the half-life. The half-period is 0.693 times the mean life or average life of a group of identical radioactive atoms. The probability is exactly $^1\!/_2$ that the actual life-span of one individual radioactive atom will exceed its half-period. *See* RADIOACTIVITY.

Robley D. Evans

Halichondrida

A small order of sponges of the class Demospongiae, subclass Ceractinomorpha, with a skeleton of diactinal or monactinal siliceous megascleres or both. Some megascleres may be arranged in loose tracts, but most are distributed irregularly in the flesh. Spongin is present in small amounts; microscleres are absent. A skinlike dermis is present and is often reinforced with tangentially placed spicules.



***Halichondria panicea*, shallow-water sponge. (*a*) Encrusting form. (*b*) Fistular form.**

Halichondrid sponges are encrusting, massive, lobate, or branching in shape. Common shallow-water species, such as *Halichondria panicea*, exhibit extensive intraspecific variations in shape associated with environmental conditions (see **illus.**). Halichondrids inhabit all seas, occurring chiefly in tidal areas and shallow waters of the continental shelf. Some species occur down to depths of at least 4900 ft (1500 m). Fossil species are unknown. *See* DEMOSPONGIAE. Willard D. Hartman

Halide

A compound containing one of the halogens [fluorine (F), chlorine (Cl), bromine (Br), iodine (I)] and another element or organic group. Halides have the general formula M_xX_y, where M is a metal or organic group and X is a halogen. Halides are composed of almost every element in the periodic table, and they are referred to as fluorides, chlorides, bromides, or iodides. *See* HALOGEN ELEMENTS; PERIODIC TABLE.

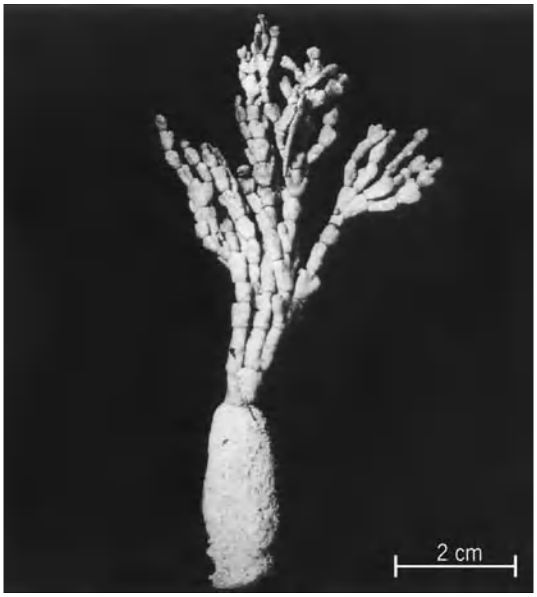
The halides are divided into classes that reflect the nature of bonding between the halogen and metal or organic species. The bonding of halides ranges from purely ionic to essentially covalent. The classes include ionic halides, molecular halides, halides and halogens that behave as ligands in coordination complexes, and organic halides.

Ionic halides such as sodium chloride (NaCl) and potassium chloride (KCl) are prepared from the vigorous reaction of the alkali and alkaline-earth metals with the halogens. These compounds possess high melting and boiling points and are soluble in very polar solvents. Ionic halides are extremely important to the chemical industry, where they are used to produce commodity chemicals such as sodium hydroxide (NaOH), hydrochloric acid (HCl, a hydrogen halide), and potassium nitrate (KNO_3).

The organic halides are divided into the alkyl halides (haloalkanes) and the aryl halides. The alkyl halides have the general formula RX, where R is any alkyl group and X is one of the halogens; for example, 1-chlorobutane ($CH_3CH_2CH_2CH_2Cl$). Halides are good leaving groups in nucleophilic substitution reactions and are good nucleophiles. The aryl halides are compounds where the halogen is attached directly to an aromatic ring and have the general formula ArX, where Ar is an aromatic group. *See* COORDINATION CHEMISTRY; COORDINATION COMPLEXES; ELECTROPHILIC AND NUCLEOPHILIC REAGENTS; HALOGENATED HYDROCARBON; HALOGENATION. Thomas J. Meade

Halimeda

A genus of marine, benthic, green algae (Chlorophyta) belonging to the family Codiaceae. Plants are attached by a holdfast, generally several centimeters high, and consist of calcareous segments separated by flexible, little-calcified nodes. Most species have a distinct, erect habit (see **illus.**), but some deep-water species have a vinelike growth form.



Halimeda from southern Florida.

Internally, *Halimeda* segments are made up of longitudinal filaments that develop lateral branches terminating in the surface layer. Calcification, consisting of the mineral aragonite, is variable within segments; also, older segments in a plant are more thoroughly calcified than younger ones.

Halimeda is an exclusively marine alga restricted to tropical waters, except for one or two species known from subtropical regions. These algae colonize sand and mud substrates, where rhizoids of the plant penetrate the soft bottom to develop holdfasts. Light requirements and other factors permit these plants to range in depth from just below low tide to about 330 ft (100 m). They are most common at shallow depths of a few meters, especially in tropical marine shelf and lagoonal environments.

Halimeda has been a prodigious sediment producer throughout its history, which extends from the Mesozoic. The first occurrence of *Halimeda* is not certain because of taxonomic problems in distinguishing morphologically similar genera; however, Middle Jurassic time is probably the earliest appearance of this modern calcareous green alga. *See* ALGAE; ATOLL. John L. Wray

Bibliography. J. L. Wray, *Calcareous Algae*, 1977.

Halite

One of the group of minerals referred to as evaporites, halite is commonly known as salt. Halite is one of many substances that are essential for human life. Evaporite minerals form when ions are concentrated to their saturation point by the progressive evaporation of seawater or saline lake water. Halite precipitates after calcium sulfate, but before the highly soluble salts of potassium and magnesium.

Halite (chemical formula NaCl) is composed of sodium cations and chlorine anions in equal proportion. It is the most common chloride mineral in natural sequences which proceed beyond the precipitation of sulfates. Even in sequences which contain a high percentage of potassium and magnesium salts, halite is often the most common chloride present.

Crystals of halite are generally cubic or hopper-shaped (skeletal). The mineral is isometric with symmetry $4/m\ \bar{3}\ 2/m$. Although the mineral is colorless generally, impurities can color it gray, red, orange, or brown. Blue or violet halite results from exposure to radioactivity, which produces dislocations and defects in the crystal structure. Halite is characterized by a hardness of 2.5 on Mohs scale and a specific gravity of 2.16.

The U.S. Geological Survey ranks halite as a basic raw material essential to modern industry along with coal, limestone, iron, and sulfur. Chlorine gas and hydrochloric acid are two of the major products for which halite is a raw material. The agricultural industry uses halite for a variety of purposes, for example, in cattle feed stocks, in fertilizers, and as a weed killer. In many areas with severe winter weather, halite is used regularly as a road deicer. Salt is used for seasoning food and as a preservative both in the home and in the food processing industry. *See* SALT (FOOD).

The deformation of bedded halite deposits is of importance to the petroleum industry. Salt rises, in part as a result of density contrasts, to form dome-like structures. Hydrocarbons (oil and gas) are commonly associated with salt domes. Exploration for these structures by geophysical techniques often results in major discoveries by the petroleum industry. *See* GEOPHYSICAL EXPLORATION; PETROLEUM GEOLOGY; SALT DOME.

Salt resources are found throughout the world and are considered virtually unlimited for any future need. The United States is the world's largest producer, with over 50% of the production coming from Texas and Louisiana and most of the rest from New York, Ohio, and California. Other major producers include China, Russia, Germany, Indonesia, Canada, Mexico, France, Italy, Brazil, Spain, and Romania. *See* HALOGEN MINERALS; SALINE EVAPORITES.

Marc L. Helman; B. Charlotte Schreiber

Bibliography. Bureau of Mines, U.S. Department of the Interior, *Minerals Yearbook*, 1988, vol. 1: *Metals and Minerals*, 1990; Bureau of Mines Staff, *Chemical Industry Applications of Industrial Minerals and Metals*, 1993; W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 5: *Non-Silicates*, 1962; C. J. Dixon, *Atlas of Economic Mineral Deposits*, 1979; C. S. Hurlbut, Jr., and C. Klein, *Manual of Mineralogy*, 21st ed., 1998; C. C. Plummer and D. M. McGeary, *Physical Geology*, 8th ed., 1999.

Hall effect

An effect whereby a conductor carrying an electric current perpendicular to an applied magnetic field develops a voltage gradient which is transverse to both the current and the magnetic field. It was discovered by E. H. Hall in 1879. Important information about the nature of the conduction process in semiconductors and metals may be obtained through analysis of this effect.

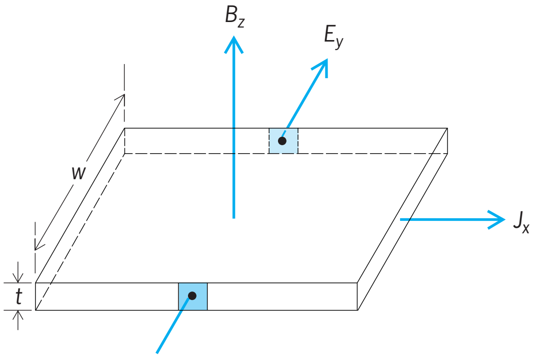


Fig. 1. Configuration of fields and currents in the Hall effect experiment.

Theory. A simple model which accounts for the phenomenon is the following. For a magnetic field of strength B in the z direction (**Fig. 1**), particles flowing with speed v in the x direction suffer a Lorentz force F_L in the y direction given by Eq. (1), where

$$F_L = -qvB \tag{1}$$

q is the charge of the particles. This force deflects the particles so that a charge imbalance develops between opposite sides of the conductor. Deflection continues until the electric field E_y resulting from this charge imbalance produces a force $F_y = qE_y$ which cancels the Lorentz force. In practice, the equilibrium condition $F_L + F_y = 0$ is achieved almost instantaneously, giving a steady-state Hall field as in Eq. (2). The current density is $J_x = nqv$, where n

$$E_y = vB \tag{2}$$

is the carrier density. The Hall resistivity, defined by Eq. (3), is thus given by Eq. (4). For a sample of thick-

$$\rho_{yx} = \frac{E_y}{J_x} \tag{3}$$

$$\rho_{yx} = \frac{B}{nq} \tag{4}$$

ness t and width w, the Hall voltage is $V_H = -E_y w$ and the total current is $I = J_x tw$ so that the Hall resistance, defined by Eq. (5) [which is the experimentally measured quantity], obeys Eq. (6) and is independent of

$$R_H = \frac{V_H}{I} \tag{5}$$

$$R_H = \frac{-B}{nqt} \tag{6}$$

the width but inversely proportional to the sample thickness. The Hall coefficient, defined by Eq. (7),

$$R_0 = \frac{\rho_{yx}}{B} \tag{7}$$

satisfies Eq. (8) and thus R_0 provides a measure of

$$R_0 = \frac{1}{nq} \tag{8}$$

the sign and magnitude of the mobile charge density in a conductor. Within the free-electron theory of simple metals, q is expected to be the electron charge $-e$, and n is taken to be $n = Zn_A$, where Z is the valence of the metal and n_A is the density of the atoms. This yields Eq. (9).

$$R_0 = \frac{-1}{n_A Ze} \tag{9}$$

See FREE-ELECTRON THEORY OF METALS.

Equation (9) is approximately valid in simple monovalent metals but fails drastically for other materials, often even giving the wrong sign. The explanation of the failures of Eq. (9) was one of the great early triumphs of the quantum theory of solids. The theory of band structure shows how collisions with the periodic array of atoms in a crystal can cause the current carriers to be holes which have an effective positive charge which changes the sign of the Hall coefficient. Band structure theory also accounts for the observed dependence of R_0 on the orientation of the current and the magnetic field relative to the crystal axes, an effect which is very useful for studying the topology of the Fermi surface. *See* BAND THEORY OF SOLIDS; FERMI SURFACE; HOLE STATES IN SOLIDS.

Effect in semiconductors. Because of band structure effects, semiconductors generally contain two types of carriers and their number density is highly temperature-dependent. The presence of two types of carriers causes the Hall coefficient to vary with the strength of the magnetic field. This effect can be used to deduce the relaxation times of each carrier species. *See* SEMICONDUCTOR.

Effect in magnetic materials. In a magnetic material the Hall resistivity has the form $\rho_{yx} = R_0B + R_SM$, where M is the magnetization and R_S is the anomalous (or extraordinary) Hall coefficient. The contribution of the anomalous term due to the magnetization may be as large as 100 to 1000 times that of the ordinary term due to the Lorentz force. The origin of this intrinsically quantum-mechanical effect is the asymmetry of the scattering of the electrons from the atoms when there is a net magnetization. The electrons scatter more frequently to the right (say) than the left, producing an extra transverse force distinct from the Lorentz force. For example, a plot of the Hall resistivity of amorphous ferromagnetic cobalt as a function of applied field displays an initial sharp rise, which is due to the increase in the normal component of magnetization of the sample and the resulting anomalous Hall effect. The magnetization reaches its saturation value M_S at an applied field of about 1.8 teslas. A subsequent linear decrease in the Hall resistivity is due to the ordinary Hall effect, which happens to have the opposite sign.

Quantization of Hall resistance. In certain special field-effect transistors, it is possible to create an electron gas which is effectively two-dimensional. The Hall resistance for an idealized system in two dimensions is given by Eq. (10), where n_S is the density of

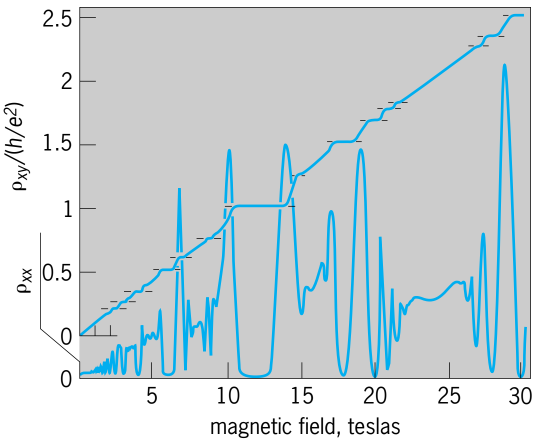


Fig. 2. Hall resistivity (ρ_{xy}, upper trace) and dissipative resistivity (ρ_{xx}, lower trace) as functions of magnetic field in a GaAs-AlGaAs heterostructure at a temperature close to absolute zero (0.085 K). (*After R. Willett et al., Observation of an even-denominator quantum number in the fractional quantum Hall effect, Phys. Rev. Lett., 59:1776–1779, 1987*)

$$\rho_{xy} = -\rho_{yx} = \frac{B}{n_S e} \tag{10}$$

electrons per unit area (rather than volume). However, if the measured value of ρ_{xy} for a high-quality (low-disorder) device is plotted as a function of B (**Fig. 2**), the linear behavior predicted by Eq. (10) is observed only at low fields. At high fields the Hall resistance exhibits plateau regions in which it is a constant independent of B. Furthermore, the values of ρ_{xy} on these plateaus are given quite accurately by the universal relation of Eq. (11), where h is Planck's

$$\rho_{xy} = \frac{h}{e^2\nu} \tag{11}$$

constant and ν is an integer or simple rational fraction (Fig. 2). The precision and reproducibility of measurements of the quantized values of ρ_{xy} have reached a few parts in 10^8. The absolute accuracy with which Eq. (11) has been verified is better than 1 part in 10^6.

This extremely accurate quantization of ρ_{xy} allows the realization of a new standard of resistance based solely on fundamental constants of nature. In addition, the quantum unit of Hall resistance, $h/e^2 \simeq$ 25,812.80 ohms, determines the fine-structure constant. *See* ELECTRICAL UNITS AND STANDARDS; FUNDAMENTAL CONSTANTS.

The explanation of this remarkable phenomenon involves several subtle quantum-mechanical effects. In the quantum regime (small ν), ρ_{xx}, which is the dissipative (longitudinal) resistivity, approaches zero on the Hall plateaus. The quantization of the Hall resistance is intimately connected with this fact. It is speculated that at zero temperature the dissipation is zero and that Eq. (11) is then obeyed exactly. *See* QUANTUM MECHANICS.

The nearly complete lack of dissipation in the quantum Hall regime is reminiscent of superconductivity. In both effects the ability of the current to flow without dissipation has its origin in the existence of a quantum-mechanical excitation gap, that is, a minimum threshold energy needed to disturb the special microscopic order in the system. In an ordinary conductor, flow of charge past the (inevitably present) impurities and random crystalline defects induces turbulence and hence dissipation, much as occurs in a rocky stream. Here, however, quantum mechanics causes the flow to be highly ordered, and production of turbulence costs a small but finite amount of energy which is not available to the system at temperatures close to absolute zero. Hence, just as in a superconductor, the current flow is nearly ideal and carries essentially zero entropy. *See* ENTROPY; SUPERCONDUCTIVITY.

In the integer quantum Hall effect [where ν in Eq. (11) in an integer], this excitation gap is a single-particle effect associated with the quantization by the strong magnetic field of the kinetic energy of the individual electrons into discrete states called Landau levels. In the fractional effect, the gap is associated with the highly collective, many-body ordering of the electrons into a quantum state which minimizes the strong Coulomb repulsion and hence lowers the overall energy. Thus, while the integer and fractional quantum Hall effects look superficially similar on a plot of resistivities versus magnetic field (Fig. 2), their physical origins are actually quite different. *See* DE HAAS-VAN ALPHEN EFFECT; GALVANOMAGNETIC EFFECTS.

Steven M. Girvin

Bibliography. N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, 1976; J. P. Eisenstein and H. L. Störmer, The fractional quantum Hall effect, *Science*, 248:1510–1516, 1990; B. I. Halperin, The quantized Hall effect, *Sci. Amer.*, 254(4):52–60, April 1986; R. E. Prange and S. M. Girvin (eds)., *The Quantum Hall Effect*, 2d ed., 1990; C. T. Van Degrift, M. E. Cage, and S. M. Girvin, Resource letter QHE-1: The integral and fractional quantum Hall effects, *Amer. J. Phys.*, 58:109–123, 1990.

Halley's Comet

The most famous of comets, associated with many important events in history. Records of Halley's Comet appear at least as far back as 240 B.C., and they are found in the Bayeux Tapestry (the apparition of the comet in 1066) and in the *Nuremberg Chronicle* (the apparition of 1456 and probably those of 684 and 1301). The comet's size, activity, and favorably placed orbit, with the perihelion roughly halfway between the Sun and the Earth's orbit, ensure its visibility to the naked eye at each apparition.

Despite major observing campaigns on recent bright comets, specifically comets Hyakutake and Hale-Bopp, Halley's Comet is the basis for many ideas about comets in general. No other comet has had so organized a campaign as that of Halley's in 1985–1986, during which several space missions were launched that made on-site measurements and produced several images of the nucleus. Although detailed knowledge of one comet is invaluable, comets are likely to be highly individualistic, and it should not be assumed that they are all similar to Halley's Comet.

History. This comet was the first to have its return predicted, a feat accomplished by Edmond Halley in 1705. He computed the orbits of several comets with Isaac Newton's new gravitational theory. The orbits of comets observed in 1531, 1607, and 1682 were remarkably similar. Halley assumed that the sightings were of a single comet and predicted its return in 1758–1759. The prediction was verified, and the comet was named in his honor. Halley's Comet was observed in 1835 by F. W. Bessel and by numerous astronomers in 1910 and 1986. Halley's Comet displays the gamut of known cometary phenomena, including a long tail when sufficiently close to the Sun (**Fig. 1**). Because of its predictable orbit, brightness, and extensive activity, it was the prime target of the six spacecraft making up the Halley Armada in March 1986. *See* COMET.

Orbital properties. The comet's orbit is a very elongated ellipse, with an eccentricity of 0.967, which has a perihelion of 0.59 astronomical unit (1 AU = the average Earth–Sun distance = 9.30×10^7 mi = 1.496×10^8 km) and an aphelion of 35 AU, between the orbits of Neptune and Pluto. Halley's Comet was farthest from the Sun in 1948, and has been moving away from the Sun since its perihelion on February 9, 1986. Halley's Comet will return to perihelion in 2061. The average period of revolution is 76 years, and the comet's motion is retrograde, that is, opposite the planets' motion. The present relative positions of the comet's orbit and the Earth's orbit mean that the comet can approach Earth as close as 0.15 AU at the descending node. *See* CELESTIAL MECHANICS.

1986 apparition. Halley's Comet was detected for the first time on its most recent approach to the Sun by D. C. Jewitt and G. E. Danielson of the California Institute of Technology. The comet was recovered on October 16, 1982, by using the 200-in. (5-m) telescope on Palomar Mountain and an advanced electronic detector originally designed for the *Hubble Space Telescope*.

Casual viewers, if they were far enough south and well away from city lights, saw the comet in March and April 1986. However, Halley's Comet was not the spectacular object for public viewing that it was in 1910. The excitement for the 1985–1986 apparition lay primarily in the massive cooperative effort that scientists launched to observe the comet.

In March 1986, six uncrewed spacecraft successfully encountered Halley's Comet on the sunward side and made measurements in its vicinity (see **table**). These missions produced data that have greatly enhanced the understanding of comets. Observations of Halley's Comet were also made by spacecraft in orbit around the Earth and the planet Venus, and by ground-based instruments. The ground-based observations were coordinated by the International Halley Watch, composed of networks of astronomers and institutions worldwide formed to coordinate the total observing effort and to archive results.

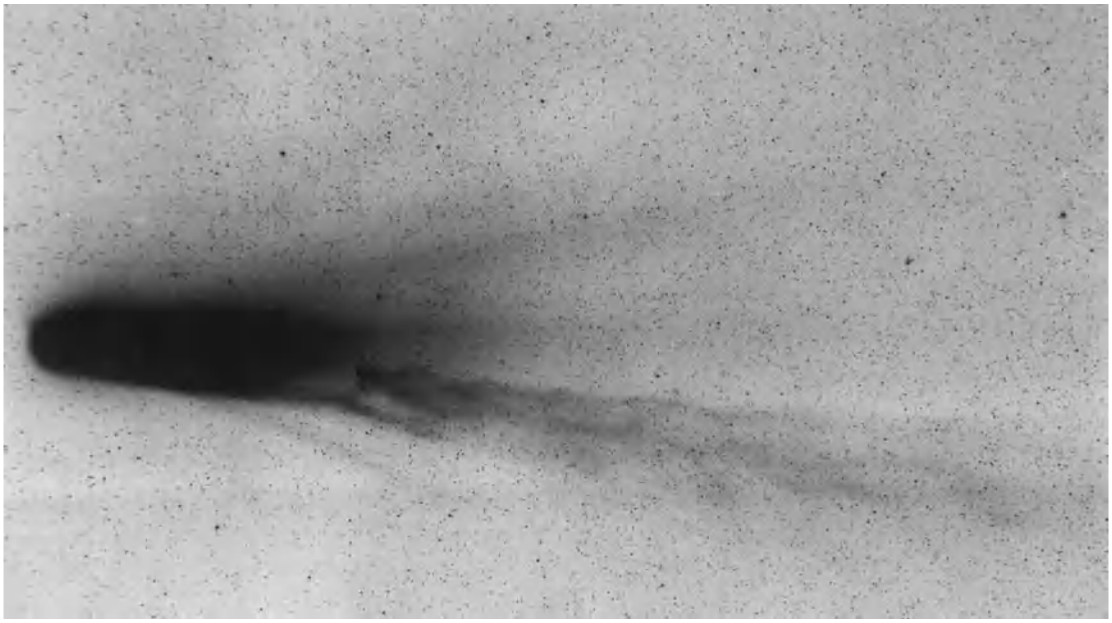


Fig. 1. Halley's Comet as photographed by the United Kingdom Schmidt telescope in Australia on March 9, 1986. Dust-tail structures are visible (above), and the plasma tail (below) also shows a completely detached portion called a disconnection event. (*Copyright © by Photolabs, Royal Observatory, Edinburgh*)

Results from both space and ground-based observations focus on three general areas. The first concerns the interaction of the comet with the solar wind. H. Alfvén's basic picture has been confirmed: the comet's plasma tail is indeed formed by the comet–solar wind interaction. Molecular ions from the comet are trapped onto solar-wind magnetic field lines, causing the magnetic field lines to drape around the comet to form the plasma tail. This process has been confirmed in detail by the spacecraft observations, including the missions to Halley's Comet and the *International Cometary Explorer*, which also encountered Comet Giacobini-Zinner at a distance of 7800 km (4845 mi) on September 11, 1985, and passed directly through the plasma tail. Somewhat surprising is the immense distance over which the interaction takes place, up to approximately 1×10^7 km (6×10^6 mi) or more from the comet, as measured by the spacecraft. Plasma processes in comets can produce spectacular results, such as disconnection events (Fig. 1).

The second area is chemical composition. The composition of the gas as measured in the inner coma is approximately 80% water (H_2O), roughly 10% carbon monoxide (CO) as determined from a rocket observation, approximately 3.5% carbon dioxide (CO_2), a few percent in complex organic compounds such as polymerized formaldehyde $(H_2CO)_n$, and the remainder in trace elements. The deuterium-to-hydrogen ratio (D/H) was initially found to be close to the value for terrestrial ocean water, although the uncertainty in the initial value was high. These results supported the view that comets may have supplied an important fraction of the volatile elements to the terrestrial planets, possibly including prebiotic molecules to Earth. However, a refined analysis gives a D/H value well above the value for ocean water. Thus, if the D/H value for Halley's Comet is typical, comets cannot have been the sole source of ocean water. Note that the composition values refer to a comet that has passed through the inner solar system many times. The values for the deep interior of Halley's Comet or for other comets may be different and would probably have a higher value of carbon dioxide.

The dust composition was found to be as follows. Some particles are composed primarily of the light atoms hydrogen (H), carbon (C), nitrogen (N), and

Space missions to Halley's Comet

Spacecraft	Sponsor	Approx. closest approach	Day in 1986
*Vega 1**	Soviet Union	8890 km (5520 mi)	March 6
Suisei	Japan	151,000 km (94,000 mi)	March 8
*Vega 2**	Soviet Union	8030 km (4990 mi)	March 9
Sakigake	Japan	7,000,000 km (4,350,000 mi)	March 11
*Giotto**	European Space Agency	605 km (375 mi)	March 14
International Cometary Explorer	NASA, United States	28,000,000 km (17,400,000 mi)	March 25

*Produced imaging of comet nucleus.

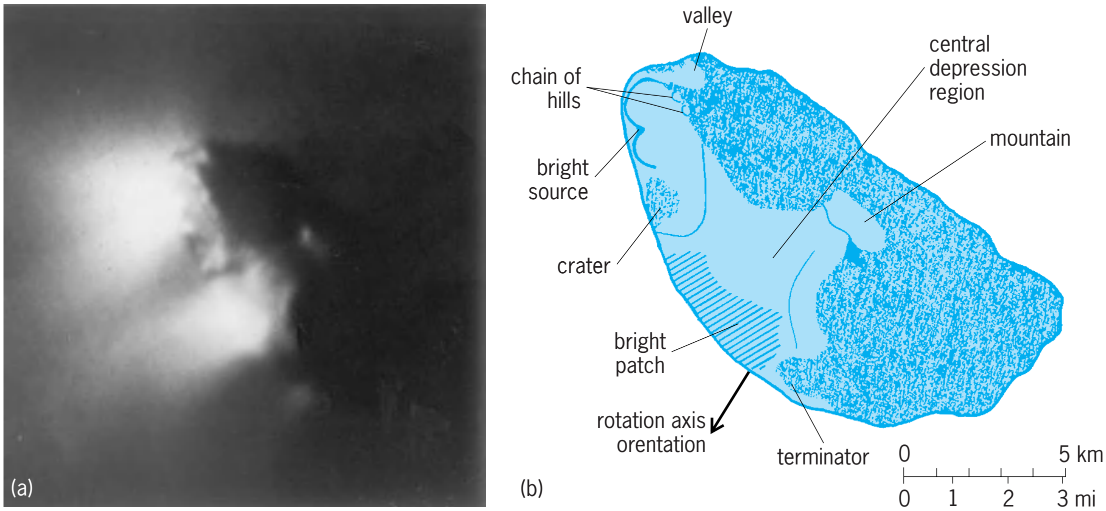


Fig. 2. **Nucleus of Halley's Comet from *Giotto* spacecraft. (*a*) Composite image formed from 60 individual images. The resolution varies from about 800 m (2600 ft) at the lower right to about 80 m (260 ft) at the upper left. The material in the bright jets streams toward the Sun, leftward. (*b*) Matching drawing labeling the features on the nucleus. (*Harold Reitsema, Ball Aerospace*; *copyright © 1986 by Max Planck Institut für Aeronomie*)**

oxygen (O), and are called CHON particles. Another kind has a silicate composition similar to the rocks that make up the crusts of Earth, Moon, and Mars, and of most meteorites. Most dust particles resemble a mixture of these two types, that is, they resemble carbonaceous chrondrites enriched in the light elements (hydrogen, carbon, nitrogen, and oxygen). These should resemble the Brownlee particles collected in the Earth's upper atmosphere.

The third general area on which the results focus is the cometary nucleus. Researchers have long based their understanding of the nucleus on F. Whipple's dirty, water–ice snowball model. According to this model, the Sun's heating of the nuclear surface produces sublimation of the ices, causing gas and dust to be released to form the cometary atmosphere and tails. All available evidence from the Halley observations, including images of the nucleus obtained by *Giotto* and the *Vega*s, confirms Whipple's model. The spacecraft observations also show that the nuclear body, shaped like a potato or peanut, is roughly 15 km (9 mi) long and 8 km (5 mi) across, somewhat larger than expected (**Fig. 2**). The surface of the comet, found to be darker than expected, has a reflectivity comparable to black velvet or coal. The nuclear surface is likely a crust of dust covering the sublimating ices. Solar energy heats the surface, and energy is conducted downward to the ices, where sublimation occurs. Some of the gases produced by the sublimation escape through thin areas or openings in the crust, forming the jets that were observed originating from the surface of Halley's nucleus.

Studies of the rotation of Halley's nucleus have provided important insights into the rotational state and interior structure of nuclei. Evidence for the rotation period of the nucleus has been presented to support values of 2.2 days and 7.4 days. In retrospect, the rotation of the highly asymmetrical nucleus is likely to be complex. Indeed, the final answer to this important question involves a model with five dominant active regions and a nucleus that rotates and spins and also precesses. This model satisfies the constraints placed on the motion by the images from the *Vega* and *Giotto* spacecraft, the observations of jets and shells, and the observed variations in brightness. The model implies an approximately constant density throughout the interior of the nucleus.

Some insight into the active comet's dusty environment can be inferred from the dust's effect on the spacecraft. The three spacecraft that passed closest to Halley—*Giotto* and the two *Vega*s—survived but were seriously damaged because of the density of the dust and the high speed of the spacecraft. The *Giotto* spacecraft was retargeted to intercept Comet Grigg-Skjellerup in July 1992.

1991 outburst. On February 12, 1991, when it was 14.4 AU from the Sun, Halley's Comet displayed a major outburst which lasted for months. The dust cloud was about 300,000 km (185,000 mi) across. Major activity this far from the Sun is unprecedented and nicely illustrates the range of physical processes in cometary science. Crystallization of amorphous ice, widely believed to be the form of water ice in the cometary interior, could be the cause of the outburst. The amorphous to crystalline ice transition releases energy that produces pockets of gas under pressure. These gas pockets could punch through the interior material and flow to the surface. The circumstances of the outburst are consistent with this model.

Associated meteors. Halley's Comet is associated with two meteor showers, the η-Aquarids and the Orionids. The origin of these showers undoubtedly involves meteoroids from the comet's nucleus that were gravitationally perturbed into their current orbits, which intersect the Earth's orbit. *See* METEOR.

John C. Brandt

Bibliography. J. C. Brandt, *Comets*: *A Scientific American Reader*, 1981; J. C. Brandt and R. D. Chapman, *Introduction to Comets*, 2d ed., 2004; J. Crovisier and T. Encrenaz, *Comet Science*, 2000;

M. Grewing, F. Praderie, and R. Reinhard (eds.), *Exploration of Halley's Comet*, 1988; W. F. Huebner (ed.), *Physics and Chemistry of Comets*, 1990.

Halloysite

A clay mineral similar in structure to kaolinite, having a 1:1 structure in which a silica tetrahedral sheet is joined to an alumina octahedral sheet. Unlike kaolinite, however, the structure is disordered in both the *a* and the *b* axis directions in successive layers, and it frequently contains water between the layers. *See* KAOLINITE.

Two principal modifications exist: a less hydrous form with a composition and structure near to that of kaolinite, $Al_2Si_2O_5(OH)_4$; and a hydrous form with the composition $Al_2Si_2O_5(OH)_4 \cdot 2H_2O$. The less hydrous form has a *c*-dimension of about 0.72 nanometer, whereas the hydrous form has a *c*-dimension of about 1.01 nm, the difference between them being roughly the thickness of a single sheet of water molecules. The hydrated form converts spontaneously and irreversibly into the less hydrous form when dried. The terminology for these two forms is confused. The 0.72-nm variety has been called metahalloysite, dehydrated halloysite, and halloysite. The 1.01-nm variety has been called halloysite, hydrated halloysite (or hydrohalloysite), and endellite. The recommended terminology is halloysite (0.7 nm) and halloysite (1.0 nm).

Electron microscopy reveals that the morphology of halloysite is usually tubular. Because the 1:1 layers in halloysite generally are separated from each other by water, halloysite has a larger cation exchange capacity, surface area, and catalytic activity than does kaolinite.

Halloysite is formed in nature from the weathering of feldspar under intense leaching conditions, and may also form in low-temperature hydrothermal systems. It has not been synthesized in the laboratory beyond doubt, although products resembling halloysite have been obtained by the artificial weathering of feldspar, and by the intercalation of kaolinite. Halloysite may precede kaolinite as a weathering product, and the transformation of halloysite into kaolinite may explain why halloysite is not common in sediments. *See* FELDSPAR; WEATHERING PROCESSES.

Halloysite is used as a catalyst and in the manufacture of ceramic products. *See* CLAY MINERALS.

Dennis Eberl

Bibliography. G. W. Brindley and G. Brown (eds.), *Crystal Structures of Clay Minerals and Their X-ray Identification*, 1980.

Halo

Either of two large circles of light surrounding the Sun or Moon that result from the refraction of sunlight by small, hexagonal ice crystals falling slowly through the air. Light passing through the side faces

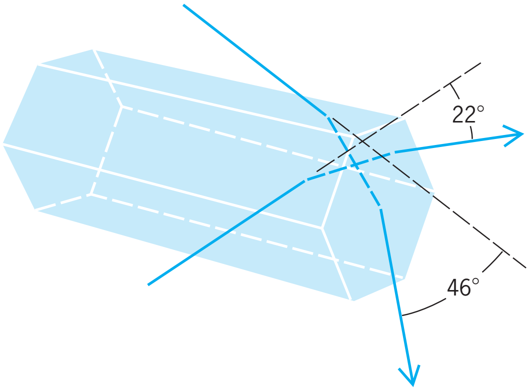


Fig. 1. Light rays passing through a hexagonal crystal to produce the 22° halo and the 46° halo.

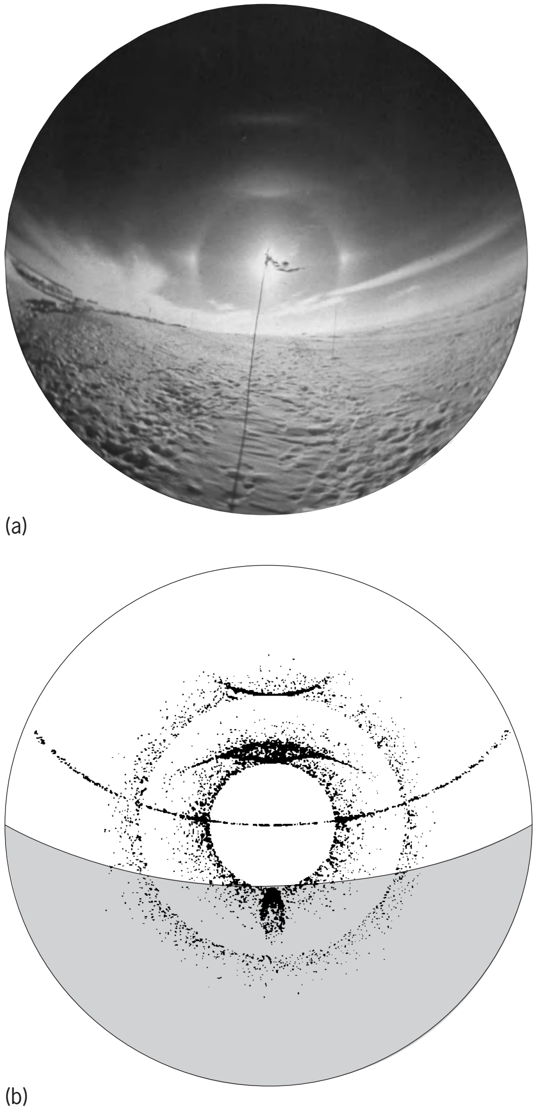


Fig. 2. Halo effects. (*a*) Photograph taken at the South Pole. (*b*) Matching computer simulations including the 22° and 46° halos, a sundog on either side of the Sun at the 22° halo position, a parhelic circle parallel to the horizon passing through the Sun, a combination of the upper tangent arc and the rarer Parry arc at the top of the 22° halo, and a circumzenithal arc at the top of the 46° halo. (*From R. Greenler*, *Rainbows*, *Halos*, *and Glories*, *Cambridge University Press*, *1980*)

of a hexagonal prism is refracted by an amount that depends on the orientation of the crystal; but a collection of many crystals refracts light passing through two side faces by an average angle of about 22° (**Fig. 1**). If such crystals tumble randomly as they fall, they will produce the 22° halo, a circle around the Sun with an angular radius of 22°. Rays that pass through a side face and an end face of the prism similarly produce the larger and fainter 46° halo. The halos sometimes have a red inner edge and otherwise appear nearly white.

Many similar effects result from rays passing through ice crystals that assume special orientations as they fall, and from rays undergoing combination of reflection and refraction in an ice crystal. Usually, all of these effects are referred to as halo effects; the wide-angle photograph in **Fig. 2*a*** shows several of these. The corresponding dot diagram (Fig. 2*b*) shows the simulation of the effects by a computer tracing of the rays through the crystal, made in order to identify the ray paths that cause each effect. *See* METEOROLOGICAL OPTICS; SUN DOG. Robert Greenler

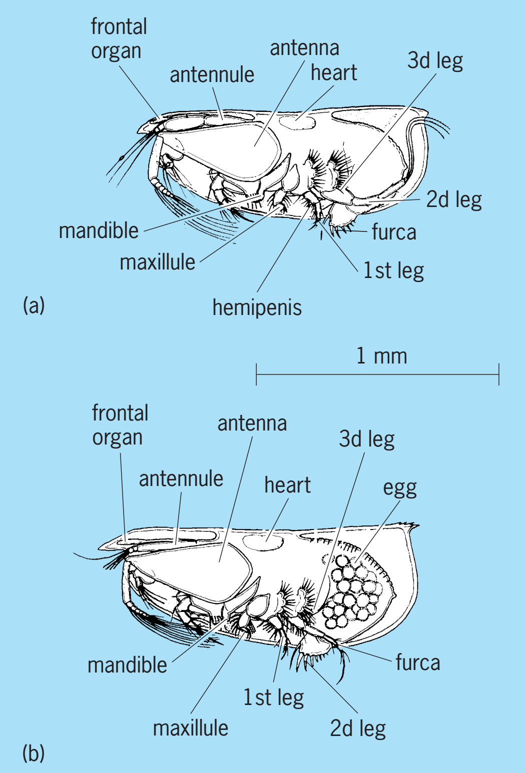
Halocyprida

An order of the subclass Myodocopa, class Ostracoda, characterized by biramous antennae with the endopod and exopod of similar size, reduction or absence of the seventh pair of appendages, an unpaired male copulatory organ, and the absence of a median eye. The taxon is subdivided into two suborders, Halocypridina and Cladocopina; each has both Recent and fossil representatives. Extant halocyprids (see **illus.**) are small ostracodes, measuring less than 0.4 in. (10 mm). The shell, which may be smooth or ornamented, may be strongly calcified or soft and may or may not bear a rostrum at its anterior end. The two orders exhibit major differences. The Halocypridina possess a uniramous maxillule, whereas that of the Cladocopina is distinctly biramous. The seventh pair of appendages is reduced to one or two segments and a pair of bristles in the Halocypridina, but the sixth pair is elongate; in the Cladocopina both the sixth and seventh pairs are absent. A heart may or may not be present in the halocypridinids, but it is totally absent in cladocopinids. Differences are also evident in the mandible and armature of the caudal furca.

With the exception of one brackish-water representative, halocyprids are marine, cosmopolitan in all oceans, and found from surface waters to abyssal depths. Most are planktonic species, although a few are epibenthic or benthic, and they include filter feeders, carnivores, and possibly detritus feeders.

The stem protostracode has been hypothesized as having several characters in common with certain halocyprids, such as the structure of the sixth and seventh appendages and the pair of furca. Several primitive characters are recognized in the family Polycopacea, and the extinct Entomozoacea could be ancestral to that family. Although classified in the suborder Halocypridina, and presumably ancestral to both of its recent major taxa, the extinct Entomoconchacea are also thought to exhibit a distant connection with the Cladocopina. *See* CRUSTACEA.

Patsy A. McLaughlin



Conchoecia elegans **(superorder Halocyprida). (*a*) Male with left valve removed; long frontal organ extends into the rostrum; endopod of the antenna terminates in a hook-shaped clasper; second leg shows dimorphic enlargements. (*b*) Female with left valve removed; antennule lacks the sensory enlargement of one seta seen in the male; both first and second legs bear branchial plates. (*After R. C. Moore, ed., Treatise on Invertebrate Paleontology, pt. Q: Arthropoda 3, Geological Society of America, 1961, and D. L. McGregor and R. V. Kesling, Contrib. Mus. Paleontol. Univ. Mich., 1969*)**

Bibliography. L. G. Abele (ed.), *The Biology of Crustacea*, vol. 1: *Systematics, the Fossil Record, and Biogeography*, 1982; L. S. Kornicker and I. G. Sohn, Phylogeny, ontogeny, and morphology of living and fossil Thaumatocypridacea (Myodocopa: Ostracoda), *Smith. Contrib. Zool.*, vol. 219, 1976; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Halogen elements

The halogen family consists of the elements fluorine, F; chlorine, Cl; bromine, Br; iodine, I; and astatine, At. Estimates of the relative abundances of these elements in seawater in parts per million, and in the lithosphere (Earth's crust to a depth of 10 mi or 16 km) are given in **Table 1**. From direct measurements it is also known that all the halogen elements except astatine exist in the Earth's atmosphere.

The halogens are the best-defined family of elements. They have an almost perfect gradation of

TABLE 1. Relative abundances of the halogens, in ppm

Element	Seawater	Lithosphere
Fluorine	1.4	770
Chlorine	18,980	550
Bromine	65	1.6
Iodine	0.05	0.3

physical properties. The increase in atomic weight from fluorine through iodine is paralleled by increases in density, melting and boiling points, critical temperature and pressure, heats of fusion and vaporization, and even in progressively deeper color (fluorine is pale yellow; chlorine, yellow-green; bromine, dark red; and iodine, deep violet).

Chemically, fluorine is the most powerful oxidizing agent known. The heavier halogens have progressively less oxidizing ability. Each forms an acid with hydrogen, and salts with metals. The properties of these acids and salts show as consistent a relationship as the elements themselves. Organic halogen compounds generally show progressively increased stability in the order iodine, bromine, chlorine, fluorine.

Although all halogens generally undergo the same types of reactions, the extent and ease with which these reactions occur vary markedly. Fluorine in particular has the usual tendency of the lightest member of a family of elements to exhibit reactions not comparable to the other members. Each halogen must be considered individually, both in its preparation and in its reaction. *See* ASTATINE; BROMINE; CHLORINE; FLUORINE; HALIDE; HALOGENATION; IODINE; PERIODIC TABLE.

Despite the chemical similarities of the halogen elements, a rich variety of chemical forms and processes are involved in the injection of halogen-containing material into the atmosphere, the chemical and physical transformations that occur in the air, and the eventual removal of the halogen atoms from the atmosphere. Although their concentrations are known to vary with time and location, the values listed in **Table 2** are typical.

Particulates. In airborne particles whose diameters range from 0.1 micrometer to over 10 micrometers in the marine atmosphere, there is typically 5000 nanograms of chlorine per standard cubic meter (scm) of air, and lesser amounts of F, Br, and I as shown in Table 2. In continental air there is normally less chlorine, bromine, and iodine in airborne particles (aerosols) than near the sea, while the opposite is true for fluorine. In polluted areas there are often much higher levels of halogens in particles; specific examples include fluoride from aluminum and steel mills and phosphate-fertilizer plants and bromide from the burning of gasoline additives. Generally, the halogen form in particles and in precipitation is the halide: fluoride (F^-), chloride (Cl^-), bromide (Br^-), and iodide (I^-), although iodate (IO_3^-) may also occur. An interesting anomaly is that the amount of iodine in the marine aerosol is much larger compared to that for Cl, Br, and Na than that expected from seawater. Part of the explanation for the iodine richness of marine aerosols could be a correspondingly high iodine content of organic surface films on seawater that enter the air on bursting bubbles and sea spray. Albert A. Gunkler

Gases. In gaseous form the halogens exist in the atmosphere in both organic and inorganic molecules. While there is at least one atmospherically significant naturally occurring organohalogen, methyl chloride (CH_3Cl), much of the atmospheric burden is anthropogenic. Table 2 shows that a typical concentration of organic F compounds is 1000 parts per trillion (ppt) by volume. Corresponding values for organochlorine, -bromine, and -iodine compounds are 2500 ppt, 20 ppt, and 3 ppt. The dominant individual species in this class of compounds are anthropogenic. They include the chlorofluorocarbons CF_2Cl_2, $CFCl_3$, $C_2F_4Cl_2$, $C_2F_3Cl_3$; the chlorocarbons CH_3CCl_3 and CCl_4 (which might also have a natural source), CH_2Cl_2, and C_2Cl_4; and perfluoromethane, CF_4. Generally, the environmental stability or inertness of these organohalogens increases with the number of halogen atoms substituted for hydrogen atoms and from the heaviest to the lightest halogen elements. The unusual stability of chlorofluorocarbons in sea level air allows accumulation in the air; their levels are increasing each year.

Dissolved halides. Most of the measurements of halides in precipitation are of the halide ion dissolved in rainwater, but some studies have been performed on melted snow. Table 2 shows typical concentrations measured in micrograms per liter of liquid precipitation. For fluoride the lowest levels, 5 micrograms/ liter, are seen far from areas influenced by continental dust or industrial pollution. For chloride the lowest values (100) are measured over continents, away from the influence of chloride-rich sea-salt aerosol. However, in polluted urban areas hydrochloric acid, HCl, is often a major component in acidic rain. The low values in Table 2 for Br^- and I^- in rainfall represent marine precipitation.

Chlorofluorocarbons. One class of halogen compounds, the anthropogenic chlorofluorocarbons mentioned above, despite being very useful as aerosol-spray-can propellants and as refrigerants,

TABLE 2. Atmospheric halogen concentrations

Source	F	Cl	Br	I
Particles (ng/scm) in marine air	0.2	5000	7	2
Organic gases (ppt)	1000	2500	20	3
Inorganic gases (ppt)	?	1000	1–10	1–5
Precipitation (μg/liter)	5–150	100–10,000	10	5

causes serious side effects. The same property of chemical inertness that makes CF_2Cl_2 and $CFCl_3$ useful also allows their concentrations to grow. Once in the air, there appears to be only one means of destroying these tightly constructed artificial molecules: upward air motions eventually carry these molecules into the upper atmosphere (stratosphere, 8–35 mi or 13–56 km above the surface) where they are attacked and decomposed by harsh ultraviolet (UV) light and energetic oxygen atoms. Chlorine atoms are released as the chlorofluorocarbons decompose, and the chain reaction below ensues:

$$\begin{array}{rcl} Cl + O_3 & \rightarrow & ClO + O_2 \\ ClO + O & \rightarrow & Cl + O_2 \\ \hline \text{net: } O + O_3 & \rightarrow & 2O_2 \end{array}$$

Ozone (O_3) and its precursor O atoms are thus destroyed in the high stratosphere. Stratospheric ozone is very important as a natural shield against biologically damaging ultraviolet rays from the Sun. Also, the absorption of solar energy by stratosphere O_3 is a major source of heat for the upper atmosphere. Natural wind systems are largely generated in this way. In addition to the chain reaction above, the chlorine species participate in many other photochemical reactions in the stratosphere. It is clear that human usage of CF_2Cl_2, $CFCl_3$, and similar chemicals is having a large impact on the chemistry of the high-altitude air, but it is not clear how much the atmospheric O_3 is being depleted by these pollutants. Continued usage of CF_2Cl_2 and $CFCl_3$ is being predicted to be capable of causing a 12% loss of atmospheric O_3 and to cause a large redistribution of the ozone with altitude. The latter distubance could affect global wind systems and climate. The facts that these chlorofluorocarbons are present in air everywhere on the Earth and have reached high altitudes are well established by reliable measurements. A second, equally important side effect of the chlorofluorocarbons is that their increasing concentrations are causing more of the Earth's infrared energy to be trapped. An appreciable global warming through a greenhouse effect is indicated. *See* GREENHOUSE EFFECT; PHOTOCHEMISTRY; STRATOSPHERIC OZONE.

The projected stratospheric consequences of chlorofluorocarbon use are largely due to the fact that each chlorofluorocarbon molecule can reside in the atmosphere for 50–100 years, thus allowing accumulation to occur. A great deal of elemental chlorine, Cl_2, is used in water treatment and in bleaching with no apparent long-lasting or global environmental effects. Chlorinated solvents are also used widely, but most of them are similarly short-lived in the open air. Potential problems due to accidental release of radioactive iodine (the 129 and 131 isotopes) from uranium fission include the possibility of accumulation in the human thyroid. *See* AIR POLLUTION; ATMOSPHERIC CHEMISTRY; FLUOROCARBON.

Ralph J. Cicerone

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; R. T. Morrison and R. N. Boyd, *Organic Chemistry*, 7th ed., 2000; D. F. Shriver and P. W. Atkins, *Inorganic Chemistry*, 3d ed., 1999.

Halogen minerals

Naturally occurring compounds containing a halogen as the sole or principal anionic constituent. There are over 70 such minerals, but only a few are common and can be grouped according to the following methods of formation. *See* HALOGEN ELEMENTS.

1. *Saline deposition by evaporation of seawater or salt lakes.* Halite (rock salt), NaCl, is the most important of this type and is found in beds covering many hundreds of square miles and ranging in thickness from a few feet to over 1000 ft (300 m). Of the other minerals associated with halite, sylvite, KCl, and carnallite, $KMgCl_3 \cdot 6H_2O$, are the most important. *See* HALITE; SALINE EVAPORITES.

2. *Hydrothermal deposition.* Fluorite, CaF_2, is the chief representative of this type and occurs in veins by itself or associated with metallic ores. Cryolite, Na_3AlF_6, may be of primary deposition or may result from the action of fluorine-bearing solutions on preexisting silicates. *See* CRYOLITE; FLUORITE.

3. *Secondary alteration.* Chlorides, iodides, or bromides of silver, copper, lead, or mercury may form as surface alterations of ore bodies carrying these metals. The most common are cerargyrite, AgCl, and atacamite, $Cu_2(OH)_3Cl$. *See* CERARGYRITE.

4. *Deposition by sublimation.* Halides formed as sublimation products about volcanic fumaroles include sal ammoniac, NH_4Cl; malysite, $FeCl_3$; and cotunnite, $PbCl_2$. At Mount Vesuvius, Italy, is the most noted occurrence of such minerals. *See* HALIDE.

5. *Meteorites.* Lawrencite, $FeCl_2$, has been found in iron meteorites. *See* METEORITE.

Cornelius S. Hurlbut, Jr.

Bibliography. L. H. Ahrens, *Origin and Distribution of the Elements*, 1979; J. D. Dana, *System of Mineralogy*, vol. 2: *Halides*, *Nitrates*, *Borates*, *Carbonates*, *Sulfates*, *Phosphates*, *Arsenates*, *Tungstates*, *Molybdites*, 7th ed., 1951.

Halogenated hydrocarbon

An aliphatic or aromatic hydrocarbon in which one or more hydrogen atoms are substituted by halogen. *See* HALOGEN ELEMENTS; HALOGENATION.

Alkyl halides. These are compounds in which one hydrogen of an alkane has been replaced by halogen [fluorine (F), chlorine (Cl), bromine (Br), or iodine (I)], for example, bromoethane (ethyl bromide; CH_3CH_2Br). Many alkyl halides have been prepared; the chlorides and bromides are most useful and most common. Since halogen atoms are much more electronegative than carbon, alkyl halides are polar molecules [dipole moment $(\mu) = 1.94$ for chloromethane, CH_3Cl]. This property and the higher molecular weight lead to much higher boiling points and densities compared to the parent alkanes; for example, ethane has a boiling point of −89°C (−128°F) and a density at this temperature of 0.54 g/cm^3 (0.31 oz/in.^{3}), while ethyl bromide has a boiling point of 38°C (100°F) and a density of 1.46 g/cm^3 (0.84 oz/in.^{3}). *See* DIPOLE; ELECTRONEGATIVITY; HALIDE; POLAR MOLECULE.

Preparation. Alkanes readily undergo chlorination by a radical substitution process [reactions (1), where RH represents an alkane, and R· is the free radical formed from the alkane]. Methyl chloride (CH_3Cl) and ethyl chloride (C_2H_5Cl) can be prepared in this way. However, the reaction is not selective for higher alkanes, and isomer mixtures are usually obtained. Free-radical chlorination is an important process in the industrial production of polychloro compounds. *See* ALKANE; FREE RADICAL.

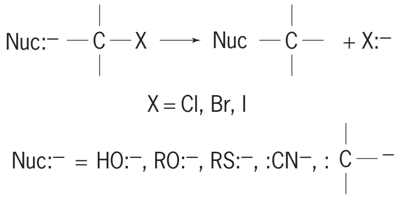
$$\begin{gathered} Cl_2 \longrightarrow Cl\cdot + RH \longrightarrow R\cdot + HCl \\ R\cdot + Cl_2 \longrightarrow RCl + Cl\cdot \end{gathered} \qquad (1)$$

A very general method for obtaining any alkyl chloride, bromide, or iodide is the reaction of the alcohol (ROH) with the corresponding hydrohalide (HX) or phosphorus halide (PX_3) [reactions (2) and (3), where RX represents the alkyl halide].

$$ROH + HX \longrightarrow RX + H_2O \qquad (2)$$

$$3ROH + PBr_3 \longrightarrow 3RX + H_3PO_3 \qquad (3)$$

Reactions. Alkyl halides are important starting materials for the preparation of many other functionally substituted compounds. A general reaction for chlorides, bromides, and iodides is nucleophilic substitution, in which an ion or molecule with an available electron pair (a nucleophile; Nuc:) displaces a halide ion [reaction (4)]. For example, *n*-propyl bromide (**1**) with sodium cyanide (**2**) gives the alkyl cyanide [**3**; reaction (5)]; with a neutral nucleophile such as a phosphine (**4**), a phosphonium salt (**5**) is the product [reaction (6)]. Alkyl iodides are more reactive than the chlorides or bromides, but they are more expensive and are useful only for small-scale laboratory purposes. Alkyl fluorides do not undergo displacement. *See* REACTIVE INTERMEDIATES; SUBSTITUTION REACTION.

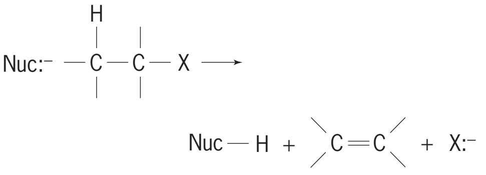


(4)

$$\underset{\textbf{(1)}}{CH_3CH_2CH_2Br} + \underset{\textbf{(2)}}{NaCN} \longrightarrow \underset{\textbf{(3)}}{CH_3CH_2CH_2CN} \qquad (5)$$

$$CH_3CH_2CH_2Br + \underset{\textbf{(4)}}{(C_6H_5)_3P} \longrightarrow \underset{\textbf{(5)}}{CH_3CH_2CH_2P^+(C_6H_5)_3Br^-} \qquad (6)$$

A related reaction, which competes with substitution, is elimination. In this case, the nucleophile acts as a base to remove a proton, and the product is an alkene [reaction (7)].

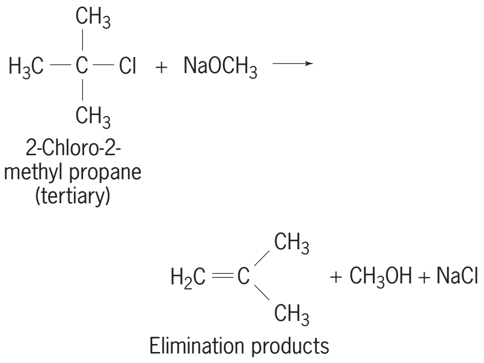


(7)

See ALKENE.

The occurrence of elimination as opposed to substitution depends on several factors. Substitution is most favorable with primary halides [RCH_2X; reaction (8)], and it is generally impractical with tertiary or other sterically encumbered halides, which undergo elimination instead [reaction (9)].

$$\underset{\substack{\text{1-Chlorobutane} \\ \text{(primary)}}}{CH_3CH_2CH_2CH_2Cl} + NaOCH_3 \longrightarrow \underset{\text{Substitution product}}{CH_3CH_2CH_2CH_2OCH_3 + NaCl} \qquad (8)$$



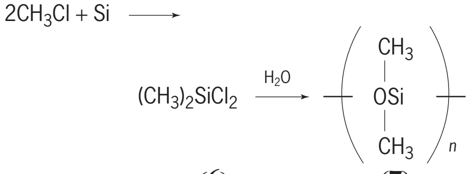
(9)

Strongly basic reagents and higher reaction temperatures promote elimination versus substitution. Alkyl fluorides are particularly prone to give the elimination product.

Another very important role of halides is in the formation of organometallic compounds. The reaction can be considered as insertion of a metal atom into the carbon-halogen bond. The organic group in the resulting organometallic compound acquires nucleophilic character, and these compounds have wide application in organic synthesis [reaction (10)].

$$R\text{—}Br + Mg \xrightarrow{\text{ether}} R\text{—}MgBr \rightleftharpoons R_2Mg + MgBr_2 \qquad (10)$$

The first organometallic compounds were alkylmagnesium halides (for example, R-MgBr, a Grignard reagent) which are readily available from any alkyl chloride, bromide, or iodide and magnesium. A major industrial use of this insertion reaction is the manufacture of dichlorodimethylsilane (**6**), which is the precursor to silicone polymers (**7**). About 600,000 tons (540,000 metric tons) per year of methyl chloride is consumed in this process [reaction (11)].

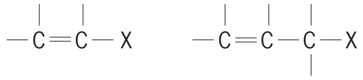


(11)

(6) (7)

See GRIGNARD REACTION; ORGANIC SYNTHESIS; ORGANOMETALLIC COMPOUND; SILICONE RESINS.

Unsaturated halides. When a halogen atom is located on a doubly bonded carbon or a carbon adjacent to a double bond, the halides are termed vinylic and allylic, respectively.



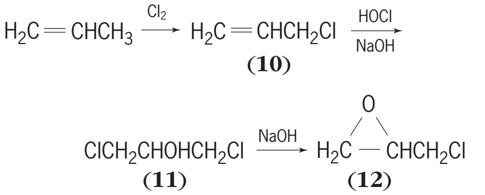
Vinylic Allylic

In vinylic halides, the C-X bond is shorter and stronger and undergoes substitution less readily than in a saturated halide. Conversely, halogen in an allylic position is activated toward nucleophilic substitution, and compounds with this structure are versatile intermediates in a number of synthetic applications. *See* BOND ANGLE AND DISTANCE; CHEMICAL BONDING.

The parent compound vinyl chloride (chloroethene) is one of the most important industrial organic chemicals. In the major manufacturing process, ethylene chloride is obtained by addition of chlorine to ethylene, and then it is converted to vinyl chloride (**8**) by elimination of hydrogen chloride (HCl). The chlorine used in the first step is generated by the oxidation of HCl with air. The overall process is called oxychlorination [reaction (12)]. Most of the output goes into the manufacture of poly(vinyl chloride) [PVC] and related copolymers. Vinyl chloride is a highly toxic and potentially carcinogenic compound, and stringent exposure limits are required in its manufacture and use. *See* COPOLYMER; POLYVINYL RESINS.

$$\begin{array}{l} H_2C{=}CH_2 + Cl_2 \longrightarrow ClCH_2CH_2Cl \longrightarrow H_2C{=}CHCl + HCl \\ 2HCl + O_2 \xrightarrow{CuCl_2} Cl_2 + H_2O \\ \hline H_2C{=}CH_2 + HCl + {}^1/_2O_2 \longrightarrow \underset{(8)}{H_2C{=}CHCl} + H_2O \end{array} \qquad (12)$$

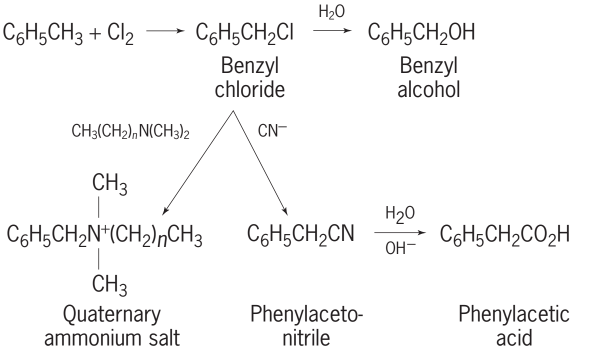
$$H_2C{=}CHCl + HF \longrightarrow CH_3CHClF \longrightarrow \underset{(9)}{H_2C{=}CHF} + HCl \qquad (13)$$



(10)

(14)

(11) (12)



(15)

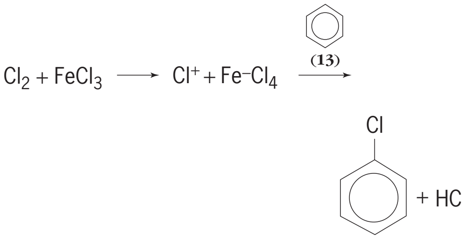
Vinyl fluoride (**9**), also used in polymers, is obtained by addition of hydrogen fluoride (HF) to vinyl chloride followed by elimination of hydrogen chloride [HCl; reaction (13)].

Allyl chloride [**10**; reaction (14)] is another important industrial chemical, prepared by free-radical chlorination of propene. Addition of hypochlorous acid (HOCl) gives the dichloroalcohol (**11**), which is then converted to the oxide, epichlorohydrin (**12**), by reaction with sodium hydroxide (NaOH). The oxide is an intermediate in the manufacture of glycerol and in epoxy resins.

Halogen on a carbon adjacent to a benzene or other aromatic ring is termed benzylic; a compound with this structure has properties like those of an allylic halide. Benzyl chloride or bromide is obtained by free-radical halogenation of toluene under mild conditions. Further chlorination gives the dichloro- and trichloro-compounds.

Substitution reactions of benzyl chloride occur readily; benzyl alcohol is obtained by hydrolysis, and phenylacetic acid is prepared via the nitrile. Substitution with a long-chain tertiary amine leads to quaternary ammonium salts, which are widely used as bacteriostatic cationic detergents [reaction scheme (15)]. *See* QUATERNARY AMMONIUM SALTS.

Aromatic halides. Compounds with halogen bonded directly to a benzene (**13**) or other aromatic ring are called aryl halides. Halogen is introduced by electrophilic substitution, with a Lewis acid catalyst such as $FeCl_3$ or $FeBr_3$ to enhance the positive character of the halogen [as illustrated for chlorobenzene (**14**) in reaction (16)].

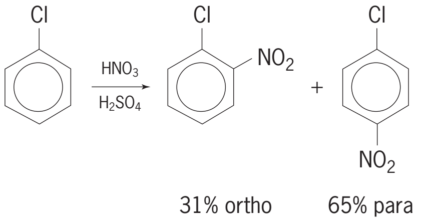


(16)

(14)

See ELECTROPHILIC AND NUCLEOPHILIC REAGENTS.

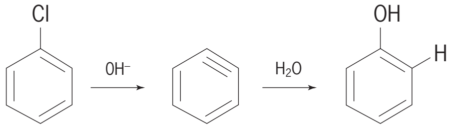
Halogen is an ortho-para directing, slightly deactivating group. Further substitution by electrophiles leads to a mixture of ortho (*o*) and para (*p*) isomers, with the latter predominating. Thus nitration gives *o*- and *p*-chloronitrobenzene in a 1:2 ratio; both of these products are important commercial intermediates [reaction (17)].



(17)

See NITRATION.

Chlorobenzene (**14**) is unreactive under conditions used for nucleophilic substitution in alkyl halides. At high temperature, chlorine is substituted by hydroxide ion to give phenol, but the process actually occurs by initial elimination to give the highly reactive intermediate benzyne (**15**) [reaction (18)].



(18)

(**14**) (**15**) (**16**)

This reaction was for some years the basis of a major industrial process for phenol (**16**). *See* ORGANIC REACTION MECHANISM.

A few aryl halides that were at one time produced on a large scale are now proscribed for use in the United States. The insecticide DDT is the condensation product of chlorobenzene and chloral. Another product, a mixture of polychlorobiphenyls, was at one time widely used as a dielectric fluid in transformers. *See* AROMATIC HYDROCARBON; POLYCHLORINATED BIPHENYLS.

Fluorocarbons. In these compounds, every hydrogen atom is replaced by fluorine. Fluorocarbons can be named simply by using the prefix perfluoro- with the parent name. Because of the small atomic radius and high electronegativity of fluorine, these compounds are chemically inert and have properties quite unlike those of other halogenated organic compounds. *See* FLUORINE; FLUOROCARBON.

Hydrofluorocarbons. These organic compounds contain combinations of fluorine and hydrogen to satisfy the valency requirement of carbon. Hydrofluorocarbons are also known as HFCs. The development of specific molecules for particular applications, previously satisfied by chlorofluorocarbons, has been international in scope. Because they contain hydrogen, hydrofluorocarbons are more likely to be degraded in the lower regions of the atmosphere. Since they do not contain chlorine, these compounds do not contribute to ozone depletion.

Hydrofluorocarbons are prepared by methods similar to that used for chlorofluorocarbons and hydrochlorofluorocarbons. Process control, however, is much more critical. Asymmetrical tetrafluoroethane [HFC-134a (CF_3CH_2F)] can be prepared by addition of HF to the double bond of trichloroethylene ($Cl_2C{=}CClH$), followed by replacement of chlorine by fluorine [reaction (19)].

$$Cl_2C{=}CClH + 4HF \longrightarrow CF_3CH_2F + 3HCl \qquad (19)$$

Another useful method is the replacement of a C-Cl bond by a C-H bond by reaction with hydrogen. This conversion can begin with a suitable chlorofluorocarbon or hydrochlorofluorocarbon, as in reactions (20) and (21).

$$CF_3CCl_2F + 2H_2 \longrightarrow CF_3CH_2F + 2HCl \qquad (20)$$

$$CF_3CHCl_2 + 2H_2 \longrightarrow CF_3CH_3 + 2HCl \qquad (21)$$

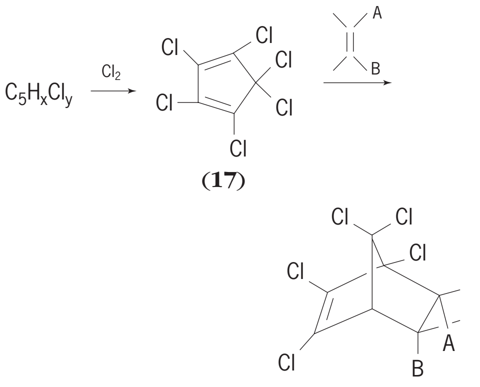
Perfluorocarbons. These saturated compounds contain only carbon and fluorine. They are named by using the prefix perfluoro- along with the name of the equivalent hydrocarbon. Perfluorocarbons are chemically very inert and also have excellent thermal stability. This inertness and the resulting long atmospheric lifetime is reflected in higher global warming potentials compared to hydrofluorocarbons, hydrochlorofluorocarbons, and chlorofluorocarbons. This is due, partly, to the high strength of the C-F bond. The electronegativity of fluorine shields the carbon backbone from chemical attack. Under normal conditions, perfluorocarbons are unaffected by strong acids or bases and by oxidizing or reducing agents. *See* ELECTRONEGATIVITY.

Perfluorocarbons are dense, colorless substances. They possess unique physical and chemical properties when compared with their hydrocarbon analogs. Low boiling points, critical temperatures, and vapor pressures are the norm. Perfluorocarbons wet almost any solid because of their low surface tensions. They have unusual solubility characteristics, and their electrical properties make them outstanding insulators. These compounds do not mix with polar liquids such as water or nonpolar liquids such as hydrocarbons.

There are several methods available for the synthesis of perfluorocarbons. The first member of the series, carbon tetrafluoride (CF_4), can be prepared in a pure state by the direct reaction of fluorine and carbon. Higher homologs can also be made by this method. Another useful procedure is to react a hydrocarbon with a high-valency metal fluoride, such as cobalt trifluoride. Process conditions are adjusted for maximum fluorination. Electrochemical fluorination is also used. Here, a hydrocarbon, dissolved in anhydrous liquid HF, is subjected to an electric current. Fluorinated product and hydrogen is the end result. The replacement of a carbon-chlorine bond with a carbon-fluorine bond is also possible. This is done with HF as the fluorine source.

Polyhaloalkanes. A number of compounds with two or more halogen atoms are of special importance. Methane (CH_4) can be substituted with as many as four halogen atoms to give compounds such as CH_2Cl_2, CHI_3, and CF_3Br. Several polyhalomethane, -ethane, and -ethylene derivatives have major uses, and they are industrial chemicals produced in large quantities.

The accumulation of halogen atoms on a saturated chain reaches a limit at three carbons, except for the important case of fluorine. The completely chlorinated (perchloro) propane C_3Cl_8 is known; but with longer chains, the perchloro compounds are alkenes and dienes such as hexachlorobutadiene. Similarly, exhaustive chlorination of a variety of C-5 alkanes or alkenes leads to further chlorination, elimination, and cyclization; the final product is hexachlorocyclopentadiene (**17**). Diels-Alder adducts of this diene [reaction (22)] have been used in the past as insecticides but they have become greatly restricted in the United States. *See* DIELS-ALDER REACTION.



(**17**)

(22)

A; B = various groups

Chlorinated C_1 and C_2 compounds. Chlorination of methane leads to mixtures of mono-, di-, tri-, and tetrachloro products. The relative amounts can be controlled by adjusting the ratio of starting materials [reaction (23); with equimolar amounts of methane and chlorine].

$$\begin{array}{l} CH_4 + Cl_2 \longrightarrow \\ \qquad \begin{array}{cccc} CH_3Cl + & CH2Cl_2 + & CHCl_3 + & CCl_4 \\ 37\% & 41\% & 19\% & 3\% \end{array} \end{array} \quad (23)$$

Methyl chloride is manufactured by this chlorination process and also by reaction of methanol and HCl. Methylene chloride, CH_2Cl_2, is the major product from methane, and is utilized primarily as a cleaning solvent or as a blowing agent for plastic foam. Methylene chloride is more volatile and much less toxic than $CHCl_3$ or CCl_4.

The trihalomethanes are known as haloforms. Chloroform is produced by methane chlorination. Its major use is in the manufacture of chlorofluorocarbons. Chloroform found some early use as an inhalation anesthetic; while the margin between anesthetic and lethal concentrations was narrow, chloroform possessed the virtue of nonflammability.

Chloroform, bromoform, and iodoform (general formula CHX_3) can all be obtained by the haloform reaction, which was at one time the manufacturing process for chloroform. In this process a methyl ketone (**18**) is treated with halogen in the presence of base [reaction (24)]. The three α-hydrogen atoms are substituted via the enol by Cl, Br, or I, and the trihaloketone (**19**) is then susceptible to cleavage by a base to give the haloform and a carboxylate anion. Before the availability of spectroscopic methods, this reaction was used as a diagnostic test for the —$COCH_3$ group in an unknown compound. Reaction of the substance in question with I_2 and a base gave a yellow precipitate of iodoform if this group was present.

$$\underset{(\mathbf{18})}{R—\overset{\overset{O}{\|}}{C}CH_3} \xrightarrow[OH^-]{3X_2} \underset{(\mathbf{19})}{—\overset{\overset{O}{\|}}{C}CH_3} \xrightarrow{OH^-} R—\overset{\overset{O}{\|}}{C}O^- + CHX_3 \quad (24)$$

The ease of formation of haloforms in aqueous solution has led to the problem that these compounds, referred to collectively as THM (trihalomethane), have become ubiquitous environmental contaminants. A major source is the halogenation process used in municipal water treatment. In the United States a maximum contaminant level (MCL) of 0.1 mg/liter for THM has been set by the Environmental Protection Agency. *See* WATER TREATMENT.

The main process for production of carbon tetrachloride has become the high-temperature chlorinolysis of mixed, partially chlorinated two- and three-carbon alkenes and alkanes that are by-products from other processes. Tetrachlorethylene ($Cl_2C{=}CCl_2$) is also obtained, and the relative amounts of CCl_4 and $Cl_2C{=}CCl_2$ can be varied by controlling the temperature and amount of chlorine.

Three chlorinated C_2 compounds are produced in very large volume, primarily for solvent use. 1,1,1-Trichloroethane (Cl_3CCH_3) is one of the least toxic of the high-chlorine compounds and is used extensively as a solvent in the manufacture of electronic components. Trichloroethylene ($HClC{=}CCl_2$) has been found to contribute significantly to photochemical smog, and its use is being curtailed. Tetrachloroethylene ($Cl_2C{=}CCl_2$) is the principal solvent in textile dry cleaning. *See* BROMINE; CHLORINE; IODINE.

Chlorofluorocarbons. These substances are methane and ethane derivatives with all hydrogen atoms replaced by combinations of chlorine and fluorine. Chlorofluorocarbons are known collectively as CFCs. The first of these compounds, dichlorofluoromethane (CCl_2F_2), was introduced in the 1930s as a nontoxic, nonflammable working fluid in refrigeration equipment to replace ammonia and sulfur dioxide. Other compounds were developed to meet the requirements of specific uses such as air-conditioning equipment in buildings and vehicles, and propellants for aerosols.

The principal method of preparation is replacement of chlorine by fluorine with HF in the presence of a metal fluoride. The reaction can be controlled to give the desired fluorine content [reaction (25)].

$$2CCl_4 + 3HF \longrightarrow CCl_3F + CCl_2F_2 + 3HCl \qquad (25)$$

Hydrochlorofluorocarbons. These simple organic compounds contain combinations of hydrogen, chlorine, fluorine, and carbon to satisfy the valency requirement of carbon. Also known as HCFs, the hydrochlorofluorocarbons have been developed as interim substitutes for chlorofluorocarbons. Since they have at least one hydrogen atom in the molecule, they are more likely to be degraded in the troposphere by reaction with hydroxyl (OH) radicals. Thus, the potential that hydrochlorofluorocarbons have to deplete ozone by migration to the stratosphere is reduced.

The hydrochlorofluorocarbons with lower molecular weights are dense, colorless gases or water-immiscible liquids. Depending on the number of hydrogen atoms present in the molecule compared to fluorine and chlorine, they may be flammable or nonflammable. Nonflammability is required for most commercial applications.

Hydrochlorofluorocarbons are used as working fluids in refrigeration and air-conditioning equipment, as foam-blowing agents, and as solvents. Both rigid and flexible polyurethane, polystyrene, or polyethylene foam are used extensively in insulation and in packaging applications. The foaming agent reduces foam density and increases softness. Previously, trichlorofluoromethane ($CFCl_3$), designated as CFC-11, was used widely for this purpose. Hydrochlorofluorocarbons are also used as solvents to remove materials such as grease in the electronic and aerospace industries.

Simple hydrochlorofluorocarbons are prepared by methods similar to those used for chlorofluorocarbons. Because of the presence of hydrogen and the reduced stability, the process conditions are much more complex.

Ozone depletion potential. The relative potential of chlorofluorocarbons and other compounds to deplete ozone (O_3) in the stratosphere is known as the ozone depletion potential.

The Earth's atmosphere is a complicated and dynamic system. Simplistically, there are several regions of the atmosphere, defined by their temperature and structure. In the lowest region, the troposphere, temperature decreases with increasing altitude, and mixing of the air is rapid. The next region is called the stratosphere. Here, the temperature increases with altitude, and mixing is slower. These two regions are separated by the tropopause. The tropopause is roughly 5–6.8 mi (8–11 km) above the Earth's surface; and the stratosphere, which lies above it, extends to about 30 mi (50 km) above the Earth's surface. The stratosphere contains ozone, which shields the Earth's surface from high-energy solar radiation, ultraviolet-B. *See* ATMOSPHERE; ULTRAVIOLET RADIATION.

The concentration of ozone in the stratosphere is the result of a dynamic balance of a series of reactions that continually produce and destroy ozone, and winds that move and mix the air. Stable chlorofluorocarbons are transported by winds from the troposphere into the stratosphere. Once there, they are subjected to solar ultraviolet radiation of very high energy. Chlorine atoms are produced in this process, and they react with the ozone. A simplified ozone destruction cycle can be written as shown in reactions (26)–(30). This chain reaction continues

$$CCl_3F + \text{ultraviolet} \longrightarrow Cl + CCl_2F \qquad (26)$$

$$Cl + O_3 \longrightarrow ClO + O_2 \qquad (27)$$

$$O_3 + \text{light} \longrightarrow O + O_2 \qquad (28)$$

$$ClO + O \longrightarrow Cl + O_2 \qquad (29)$$

$$O_3 + O_3 \longrightarrow 3O_2 \qquad (30)$$

until stopped by other mechanisms. The net result is shown in reaction (27). *See* CHAIN REACTION (CHEMISTRY).

Because of concern about ozone depletion, the use of chlorofluorocarbons in aerosol products in the United States was banned in 1978. Extensive studies linking seasonal ozone depletion over the Antarctic continent to chlorofluorocarbons and other compounds led to calls for action. An international agreement, known as the Montreal Protocol, required a phase-out of production of chlorofluorocarbons in developed countries by the year 1996. This accelerated the search for suitable alternatives in applications hitherto fulfilled by chlorofluorocarbons.

The ozone depletion potential depends on several factors. These include the number of chlorine atoms in the molecule and the ability of the compound to degrade in the troposphere before significant amounts can reach the stratosphere. In order to reduce the impact on the ozone layer, substitutes for chlorofluorocarbons have been designed to contain hydrogen to decrease their atmospheric stability. *See* AEROSOL; ATMOSPHERIC CHEMISTRY; STRATOSPHERIC OZONE.

Greenhouse effect. The thermal and radiation balance of the Earth is determined by a balance of heating by incoming solar energy and cooling by outgoing infrared radiation from the Earth's surface. Many compounds block the release of infrared radiation from lower regions of the atmosphere back into space. Chlorofluorocarbons, carbon dioxide (CO_2), and methane are a few examples. Gases, such as those mentioned, block release of infrared energy by absorbing energy frequencies that are characteristic of their structure. This retained energy results in a warming of the troposphere and the surface of the Earth known as the greenhouse effect. Carbon-halogen (especially fluorine) bonds are very efficient in absorbing certain frequencies of this infrared energy. The ability of compounds to trap radiation depends on such factors as stability in the atmosphere, extent of halogen substitution, and molecular structure. *See* GREENHOUSE EFFECT; HEAT BALANCE, TERRESTRIAL ATMOSPHERIC.

The global warming potential of halocarbons can be reduced by decreasing their stability in the atmosphere. This can be done by including hydrogen as part of the molecule. James A. Moore; V. N. M. Rao

Bibliography. R. E. Banks, B. E. Smart, and J. C. Tatlow (eds.), *Organofluorine Chemistry, Principles and Commercial Applications*, 1994; D. A. Fisher et al., Model calculations of the relative effects of CFCs and their replacements on stratospheric ozone, *Nature*, 344:508–512, 1990; M. Hudlicky, *Chemistry of Organic Fluorine Compounds*, 2d ed., 1992; L. E. Manzer and V. N. M. Rao, Catalytic synthesis of chlorofluorocarbon (CFC) alternatives, *Adv. Catal.*,

39:329–350, 1993; D. A. O'Sullivan, International gathering plans ways to safeguard atmospheric ozone, *Chem. Eng. News*, 67(26):33–36, 1989; F. Rowland, Stratospheric ozone in the 21st century: The chlorofluorocarbon problem, *Environ. Sci. Technol.*, 25:622–628, 1991; A. Streitwieser, C. Heathcock, and E. M. Kosower, *Introduction to Organic Chemistry*, 4th ed., 1992, revised 1998.

Halogenation

A chemical reaction or process which results in the formation of a chemical bond between a halogen atom and another atom. Reactions resulting in the formation of halogen-carbon bonds are especially important, and the emphasis in this article is primarily on such reactions. The halogenated compounds produced are employed in many ways, for example, as solvents, intermediates for numerous chemicals, plastic and polymer intermediates, insecticides, fumigants, sterilants, refrigerants, additives for gasoline, and materials used in fire extinguishers. *See* HALOGEN ELEMENTS.

Numerous compounds are halogenated by a wide variety of commercial and laboratory processes. Halogenation reactions can be subdivided in several ways, for example, according to the type of halogen (fluorine, chlorine, bromine, or iodine), type of material to be halogenated (paraffin, olefin, aromatic, hydrogen, and so on), and operating conditions and methods of catalyzing or initiating the reaction.

Halogenation reactions with elemental chlorine, bromine, and iodine are of considerable importance. Because of high exothermocities, fluorinations with elemental fluorine tend to have high levels of side reactions. Consequently, elemental fluorine is generally not suitable for direct fluorination. Two types of reactions are possible with these halogen elements, substitution and addition.

Substitution halogenation. This type of halogenation is characterized by the substitution of a halogen atom for another atom (often a hydrogen atom) or group of atoms (or functional group) on paraffinic, olefinic, aromatic, and other hydrocarbons. A chlorination reaction of importance that involves substitution is that between methane and chlorine. The overall reaction between methane and chlorine is shown in reaction (1).

$$CH_4 + Cl_2 \longrightarrow CH_3Cl + HCl \qquad (1)$$

Substitution reactions with paraffins involve several free-radical reaction steps: only the most important steps are shown below. In the initiating step, chlorine free radicals are formed as follows: Cl_2 (elemental chlorine) → 2 Cl· (chlorine free radical). Breaking elemental chlorine into free radicals requires energy, and can be accomplished thermally at high temperatures (about 480°F or 250°C, or higher) or at lower temperatures (including ambient temperatures) by means of radiation (gamma radiation, sunlight, and so on).

Heats of reaction in radical halogenation*

Reaction	F_2	Cl_2	Br_2
$RX + X\cdot \rightarrow R\cdot + HX$	−33	−3	+13
$R\cdot + X_2 \rightarrow RX + X\cdot$	−71	−24	−21

*In kilocalories per mole (1 kcal = 4184 joules).

The chlorine free radical starts a chain mechanism with methane, as in reactions (2) and (3).

$$CH_4 + Cl\cdot \longrightarrow CH_3\cdot + HCl \qquad (2)$$

$$CH_3\cdot + Cl_2 \longrightarrow CH_3Cl + Cl\cdot \qquad (3)$$

Similar reaction steps also occur with other paraffins or hydrogen. The relative rates of halogenation vary according to the type of carbon-hydrogen bond, as follows: tertiary > secondary > primary. Free-radical reactions such as those shown above are sometimes affected to a considerable extent by the surface of the reactor employed. The kinetics of chlorination can vary by factors of at least four or five, depending on the material of construction of the reactor or the past history of the reactor.

The **table** gives approximate heats of reaction for the two steps in the chain mechanism, such as those shown in reactions (2) and (3) for substitution halogenation with fluorine, chlorine, and bromine. Halogenation reactions are exothermic; the level of exothermicity is $F_2 > Cl_2 > Br_2 > I_2$. *See* SUBSTITUTION REACTION.

Addition halogenation. Chlorine, bromine, and iodine react readily with most olefins; the reaction between ethylene and chlorine, to form 1,2-dichloroethane, reaction (4), is one of considerable

$$CH_2{=}CH_2 + Cl_2 \longrightarrow CH_2Cl{-}CH_2Cl \qquad (4)$$

commercial importance, since it is used in the manufacture of vinyl chloride. This reaction occurs readily in the gas phase at temperatures in the range of 175 to 265°F (80 to 130°C) when the inner surface of the reactor is coated with calcium or lead chlorides.

Reaction (4) also occurs readily in the liquid phase by means of ionic reactions at temperatures from 85 to 158°F (30 to 70°C). Ferric chloride, often formed by reaction of chlorine with iron in the walls of the reactor, is an excelent catalyst.

Addition reactions with bromine or iodine are frequently used to measure quantitatively the number of —CH=CH— (or ethylenic-type) bonds on organic compounds. Bromine numbers or iodine values are measures of the degree of unsaturation of the hydrocarbons.

Halogenation of aromatics. Substitution halogenation on the aromatic ring can be made to occur by means of ionic reactions. The chlorination reactions with elemental chlorine occur under conditions similar to those used for addition chlorination of olefins; temperatures of 68–122°F (20–50°C) are suitable, and ferric chloride is an effective catalyst. Polychlorination of benzene results in various isomers of dichloro-, trichloro-, and tetrachlorobenzenes; even higher levels of chlorination sometimes occur.

When liquid benzene is contacted with chlorine at essentially ambient temperature in the presence of radiation (light radiation, gamma radiation, and so on), but in the absence of catalysts, the chlorine adds to the benzene by means of a free-radical chain mechanism to form benzene hexachloride, an effective insecticide. The overall reaction is shown as reaction (5).

$$C_6H_6 + 3Cl \xrightarrow[\text{radiation}]{\substack{86\text{–}104^\circ F \\ \text{or } 30\text{–}40^\circ C}} C_6H_6Cl_6 \qquad (5)$$

Bromine and benzene also react in the presence of sunlight; benzene hexabromide is formed, however, with difficulty.

When alkyl benzenes are halogenated, the halogen can react either on the aromatic nucleus or on the alkyl group. Free-radical (and gas-phase) chlorinations generally result in substitution, primarily on the alkyl groups; such reactions occur readily at 248–266°F (120–130°C). Ionic, low-temperature reactions, however, favor substitution on the nucleus.

At 750–930°F (400–500°C), chlorobenzenes are produced from benzene by free-radical chlorination steps. The resulting products have an isomer distribution quite different from that obtained by means of ionic chlorinations at essentially ambient temperatures.

Halogenations with hydrogen halides. Numerous methods are available for halogenating with hydrogen halides. Oxychlorination reactions can be used with olefinic, paraffinic, and aromatic hydrocarbons. The reaction with ethylene (6) is of major industrial importance for the production of 1,2-dichloroethane in modern vinyl chloride plants.

$$CH_2{=}CH_2 + 2HCl + {}^1/{}_2O_2 \xrightarrow[\substack{180\text{–}750^\circ F \\ \text{or } 250\text{–}400^\circ C}]{CuCl_2} CH_2Cl{—}CH_2Cl + H_2O \qquad (6)$$

Hydrogen halides can also be made to react with many ethylenic and acetylenic compounds. Examples are shown in reactions (7) and (8).

$$CH_2{=}CH_2 + HCl \xrightarrow[\substack{212^\circ F \\ \text{or } 100^\circ C}]{ZnCl_2} CH_3CH_2Cl \qquad (7)$$

$$HC{\equiv}CH + HCl \xrightarrow[\substack{94\text{–}284^\circ F \\ \text{or } 90\text{–}140^\circ C}]{HgCl_2} H_2C{=}CHCl \qquad (8)$$

Before 1965 reaction (8) was the main method used in producing vinyl chloride. Fluorocarbons widely used as refrigerants are produced from carbon tetrachloride; the production of trichlorofluoromethane is shown in reaction (9). Alcohols react with hydrogen halides to form alkyl halides, as shown in reactions (10) and (11).

$$CCl_4 + HF \xrightarrow[\substack{212^\circ F \\ \text{or } 100^\circ C}]{SbF_5} CCl_3F + HCl \qquad (9)$$

$$CH_3OH + HBr \rightarrow CH_3Br + H_2O \qquad (10)$$

$$CH_3CH_2OH + HCl \xrightarrow[\substack{230\text{–}284^\circ F \\ \text{or } 110\text{–}140^\circ C}]{ZnCl_2} C_2H_5Cl + H_2O \qquad (11)$$

Electrochemical techniques are employed commercially for substituting fluorine atoms for hydrogen on the alkyl group of organic acids, amines, nitriles, and alcohols. In an electrolytic cell operated at low voltages hydrogen is also a product.

Miscellaneous halogenating agents. Many compounds have been employed as halogenating agents. Several examples are sodium hypochlorite, phosgene ($COCl_2$), thionyl chloride, sulfuryl chloride, ferric chloride, antimony chlorides, and phosphorus chlorides, which can be used for certain chlorination reactions; hypobromites, *N*-bromosuccinimide, and *N*-bromoacetamide, which have proved to be brominating agents; antimony pentafluoride, silver difluoride, and lead tetrafluoride, metal fluorides which are used as fluorinating agents; and iodine monochloride and alkali hypoiodites, which are iodination agents. *See* HALOGENATED HYDROCARBON. Lyle F. Albright

Bibliography. P. B. De la Mare and R. Bolton, *Electrophilic Additions to Unsaturated Systems*, 2d ed., 1982; R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998; V. Gutmann, *Main Group Elements*: *Groups 6–7*, 1975.

Halophilism (microbiology)

The requirement of high salt (NaCl) concentrations for growth of microorganisms. Microorganisms (mainly bacteria) can be classified by their physiological tolerance to salt. Most normal eubacteria, such as *Escherichia coli* and *Pseudomonas fluorescens*, and most fresh-water microorganisms, are nonhalophiles (best growth at less than 1.2% NaCl). Slight halophiles (1.2–3% NaCl) include many marine microorganisms. Moderate halophiles (3–15% NaCl) include *Vibrio costicola*, *Paracoccus halodenitrificans*, and many others. Borderline extreme halophiles (9–25% NaCl) include the photosynthetic bacterium *Ectothiorhodospira halophila*, the actinomycete *Actinopolyspora halophila*, and the halophylic archaebacteria *Halobacterium volcanii* and *H. mediterranei*. Extreme halophiles (require at least 10% NaCl; optima 15–30% NaCl) are *Halobacterium salinarium* and *Halococcus morrhuae*. *See* METHANOGENESIS (BACTERIA).

The halophilic aerobic archaebacteria give a striking red color to hypersaline waters. They are found in the Dead Sea, the Great Salt Lake, Lake Magadi in Kenya, and other alkaline salt lakes, and in solar salterns where salt is prepared by evaporating seawater. Their red color is due to carotenoid pigments (bacterioruberins), which seem to protect them from strong sunlight in their natural environments. *See* CAROTENOID.

Some genera fit into more than one halophilic classification. Thus, some *Ectothiorhodospira* species can grow in lower salt concentrations than *E. halophila*. Some species of *Dunaliella*, a flagellated alga found in salt lakes, behave as moderate halophiles, and others as borderline extreme halophiles.

The range of salt concentration at which any microorganism grows may be affected by nutritional conditions and by temperature. Generally, at lower temperatures cells grow better at lower salt concentrations, and they can grow at a wider range of salt concentrations in richer media.

Many species of salt-tolerant microorganisms also exist which do not require salt but can grow in high salt concentrations. These include lactic acid bacteria, which can grow in 8% NaCl or higher, and some bacilli and micrococci that can grow in 25% NaCl. Yeasts (for example, *Saccharomyces rouxii*) and filamentous fungi (for example, *Xeromyces* spp.) can grow in very high concentrations of salt or sugars, such as sucrose; usually, they do not require such concentrations for growth. Such microorganisms are often used in the processing of salted foods, such as sauerkraut or soy sauce; but they will spoil jams and jellies.

It is not always possible to distinguish between the direct action of salts and sugars on cells and their indirect action, through reduction of available water in the cell's environment. True halophiles, however, have a specific requirement for NaCl whose ions can be replaced only partially, if at all, by K^+, $NO_3{}^-$, or other ions.

Adaptation to salt; compatible solutes. Microorganisms that live in high concentrations of salt or other solutes do not exclude solutes from the interior of the cell. However, the internal solute composition is quite different from the outside composition. *Dunaliella* species have internal glycerol concentrations corresponding to the external concentration of NaCl. Other salt-tolerant algae and yeasts also have high internal concentrations of glycerol, or other polyols. Solutes which maintain osmotic equilibrium between inside and outside of the cell without interfering with the cell's physiological processes are called compatible solutes. *See* OSMOREGULATORY MECHANISMS.

In halophilic eubacteria (and in some nonhalophiles exposed to moderate salt concentrations), compatible solutes include sugars and amino acids, especially glutamate and proline. However, for most halophilic organisms, the concentrations of known compatible solutes do not add up to the osmotic equivalent of external salts.

Red halophiles. Mechanisms of adaptation to a highly saline environment have been best characterized in aerobic halophilic bacteria whose enzymes are able to function in high salt concentrations; indeed, most of them require such salt concentrations for activity, stability, or both. For a number of enzyme systems, KCl rather than NaCl is required, confirming the role of KCl as a compatible solute in these bacteria.

Other parts of the cells of these bacteria also require high salt concentrations for function of stability. Halobacteria lyse and their cell walls may completely dissolve unless salt concentrations are high. NaCl is specifically required for active transport of ions and nutrients in all halophilic bacteria.

Transport and sensory pigments. When halobacteria are grown in conditions of limiting oxygen supply, up to 50% of their cytoplasmic membrane becomes covered by organized patches of purple material. The color is due to bacteriorhodopsin, a combination of a protein, bacterioopsin, and retinal, the same visual pigment found in animal eyes. When membranes containing bactriorhodopsin are exposed to visible light, protons (H^+ ions) are transferred from inside to outside of the cell. The resulting proton gradient can be used as a source of cellular energy for adenosine triphosphate (ATP) production, for active transport, and for anaerobic growth. The purple-membrane system has probably provided the best direct evidence for the chemiosmotic hypothesis since in this system a proton gradient can be established by light alone. *See* ION TRANSPORT.

Purple membranes also exist in nature in blooms of halobacteria in hypersaline waters. They are thought to provide an extra source of energy to these microorganisms under conditions of low oxygen availability, which might occur often, since oxygen is much less soluble in concentrated salt solutions than in pure water.

Molecular biology and genetics. Halophilic archaebacteria, which are often easier to grow than methanogens or extreme thermophiles, provide much information regarding central metabolism, chromosome structure, transcription, translation, and the control of synthesis of ribosomal proteins and other proteins in archaebacteria. Some mutations of *Halobacteria* species take place at a very high rate because of the presence of insertion sequences. In *Haloferax* species, partly because of the relatively simple nutritional requirements of the latter halophiles, a number of genetic markers have been found. A good part of the genome of *Haloferax volcanii* has been mapped. Virtually nothing is known of the genetics of halophilic eubacteria, and few mutations have been characterized. *See* BACTERIAL GENETICS; MUTATION.

Phages. A number of phages of halophilic archaebacteria (especially *Halobacterium salinarium* and closely related species) have been isolated from salt ponds, salted fish, and apparently pure bacterial cultures. Some phages may exist in either lytic or carrier (lysogenic) states in their hosts, depending on environmental conditions. The phages of halophilic archaebacteria have the same salt requirement for stability as their hosts, and are inactivated within minutes in dilute solutions. In contrast, phages of halophilic eubacteria are relatively stable for days in the absence of salt. *See* ARCHAEA; BACTERIAL PHYSIOLOGY AND METABOLISM.

D. J. Kushner

Bibliography. A. D. Brown, *Microbial Water Stress Physiology: Principles and Perspectives*, 1990; M. Kates, D. J. Kushner, and A. T. Matheson (eds.), *The Biochemistry of Archaea (Archaebacteria)*,

1993; T. Kauri et al., A bacteriophage of a moderately halophilic bacterium, *Arch. Microbiol.*, 156:435–438, 1991; D. J. Kushner (ed.), *Microbial Life in Extreme Environments*, 1978; F. Rodriguez-Valera (ed.), *Halophilic Bacteria*, vols. 1 and 2, 1988; R. H. Vreeland and L. I. Hochstein (eds.), *The Biology of Halophilic Bacteria*, 1992.

Haloragales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Rosidae of the class Magnoliopsida (dicotyledons). The order consists of 2 families, with about 150 species in all. The Haloragales are herbs with perfect or often unisexual, more or less reduced flowers. The perianth is minute or vestigial and was probably originally tetramerous. The ovary is inferior, with only one ovule in each locule, and the seeds have a more or less abundant endosperm. Many of the species are aquatic. Entomophily (pollination by insects) has been largely abandoned in the group, and the pollen is commonly distributed by wind or water. The aquarium plant called parrot's feather (*Myriophyllum*, family Haloragaceae) and the very large-leaved plant *Gunnera* (family Gunneraceae) are well-known members of the Haloragales. *See* MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM; ROSIDAE.

Arthur Cronquist; T. M. Barkley

Hamamelidae

A small subclass of the class Magnoliopsida (dicotyledons) in the division Magnoliophyta (Angiospermae), the flowering plants, consisting of 11 orders (Trochodendrales, Hamamelidales, Daphniphyllales, Didymelales, Eucommiales, Urticales, Leitneriales, Juglandales, Myricales, Fagales, and Casuarinales), 24 families, and about 3400 species. They have strongly reduced, often unisexual flowers with poorly developed or no perianth. In the more advanced types, the flowers of one or both sexes are borne in catkins, and the mature fruit is unilocular and indehiscent, with only a single seed. With the notable exception of some of the Urticales, they are all woody plants. Pollination is usually by wind. Many of the families formerly grouped under the Amentiferae belong to the Hamamelidae. See articles on each order. *See* MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM.

Arthur Cronquist; T. M. Barkley

Hamamelidales

A small order of flowering plants, division Magnoliophyta (Angiospermae), which gives its name to the subclass Hamamelidae in the class Magnoliopsida (dicotyledons). The family Hamamelidaceae contains about 100 species and the Platanaceae about 6 species; the other 3 families have only 2 species each. Within its subclass, the order is more advanced than the Trochodendrales in having vessels in the wood, but less advanced than the other orders in that the gynoecium consists either of separate carpels or of united carpels that open at maturity to release the seeds. Witch hazel (*Hamamelis*; see **illus.**), sweet gum (*Liquidambar*, family Hamamelidaceae), and the plane tree or sycamore (*Platanus*) are familiar members of the Hamamelidales. *See* HAMAMELIDAE; MAGNOLIOPHYTA; PLANT KINGDOM; TROCHODENDRALES.

Arthur Cronquist; T. M. Barkley

American witch hazel (*Hamamelis virginiana*), a characteristic member of the family Hamamelidaceae and order Hamamelidales. The flowers of this species open in the autumn, when the leaves are already beginning to deteriorate. (*John H. Gerard, National Audubon Society*)

Hamilton-Jacobi theory

A theory that uses canonical transformations to construct a general procedure for solving mechanical problems. Either of two methods can be used.

1. If the Hamiltonian is conserved, one can transform to a set of canonical coordinates such that only the momenta appear. That is, the coordinates do not appear in the transformed Hamiltonian. Such coordinates are termed cyclic, or ignorable. This procedure generates equations with trivial solutions.

2. One can find a cononical transformation from the coordinates and momenta (q,p) at a time t to a new set of constant quantities. Usually these quantities are chosen to be the values of the coordinates and momenta (q_0,p_0) at $t = 0$.

With such a transformation, the equations of transformation relating old and new canonical variables are the desired solution to the problem. The new coordinates depend only upon the initial values and time. This latter method is more general and can also be used when the Hamiltonian is time-dependent.

Either of these procedures comprises the Hamiltonian-Jacobi theory. Both require a canonical transformation of the coordinates so that the transformed Hamiltonian is identically zero. The transformation eliminates the original momenta and results in a set of $(f + 1)$ differential equations in the coordinates q_i and t, where f is the number of degrees of freedom of the system. These equations can, in principle, be solved. *See* CANONICAL COORDINATES AND TRANSFORMATIONS; DEGREE OF FREEDOM (MECHANICS); DIFFERENTIAL EQUATION; HAMILTON'S EQUATIONS OF MOTION.

Schrödinger equation. The resulting Hamilton-Jacobi equation bears a close formal resemblance—in fact, much more than formal—to the Schrödinger wave equation. The Hamilton-Jacobi equation determines the canonical transformation from a set of initial coordinates and momenta to those at time t. Likewise, in quantum mechanics the state vector at t is found from the state vector at an earlier time t_0 by a unitary transformation, as in Eq. (1). The

$$\psi(t) = U(t,t_0)\psi(t_0) \tag{1}$$

unitary operator U satisfies the Schrodinger equation (2), which resembles the Hamilton-Jacobi equa-

$$H\left(q, \frac{ih}{2\pi}\frac{\partial}{\partial q}, t\right) U = \frac{ih}{2\pi}\frac{\partial U}{\partial t} \tag{2}$$

tion for the canonical transformation. *See* NONRELATIVISTIC QUANTUM THEORY; SCHRÖDINGER'S WAVE EQUATION.

Action-angle variables. A system is said to be separable and multiply periodic if, for each coordinate pair q_j, p_j, there is a separate equation of motion such that each pair is periodic in the time, as in oscillatory motion, or p_j is a periodic function of q_j, as in rotational motion. In such systems it is useful to introduce the action variables J_k, also termed the action-angle variables, defined by Eq. (3). In this case

$$J_k = \oint p_k dq_k \tag{3}$$

the energy is expressible as a function of the action variables, as in Eq. (4).

$$E = H(J_1, \ldots, J_f) \tag{4}$$

See ACTION; QUANTUM MECHANICS.

Each action-angle variable J_k corresponds to a frequency ν_k at which the corresponding coordinate returns to its previous value. To calculate the frequencies ν_k of a separable, multiply periodic system, it is sufficient to find the energy as a function of the action variables. The derivative of this function with respect to any one action variable is the frequency of the corresponding motion. This relation provides a basis for perturbation theory.

John L. Safko; Philip Stehle

Bibliography. V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2d ed., 1989; M. Born, *Mechanics of the Atom*, 1960; T. L. Chow, *Classical Mechanics*, 1995; H. C. Corben and P. Stehle, *Classical Mechanics*, 2d ed., Dover reprint, 1994; H. Goldstein, C. P. Poole, Jr. and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; I. Percival and D. Richards, *Introduction to Dynamics*, 1982.

Hamilton's equations of motion

A set of first-order ordinary differential equations that may be used to describe the motion of a mechanical system. Because of their remarkably symmetrical form [which appears in Eqs. (4), below], they are often referred to as the canonical equation of motion (where "canonical" is used in the sense of designating a simple general set of standard equations). The Lagrangian formulation of a system of f degrees of freedom generates f differential equations of second order in the time derivatives of the variables. Hamilton's equations, which are equivalent to Lagrange's equations, consist of $2f$ first-order and highly symmetrical equations. These properties make Hamilton's equations very useful for general discussions of the motion of systems. *See* DEGREE OF FREEDOM (MECHANICS); DIFFERENTIAL EQUATION; LAGRANGE'S EQUATIONS.

Definitions. Hamilton's equations follow from Lagrange's equations in a straightforward manner. The generalized coordinates of a system with f degrees of freedom can be labeled q_j $(j = 1, 2, \ldots, f)$, and the dynamical description of the system can be given by the Lagrangian $L(q,\dot{q},t)$, where q denotes all the coordinates and a dot denotes the total time derivative. Lagrange's equations are then given by Eq. (1).

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j} = 0 \tag{1}$$

The momentum p_j canonically conjugate to q_j is defined by Eq. (2).

$$p_j = \frac{\partial L}{\partial \dot{q}_j} \tag{2}$$

It is assumed that Eq. (2) is soluble for the velocities $\dot{q}_j$ in terms of the coordinates and momenta. *See* LAGRANGIAN FUNCTION.

The Hamiltonian $H(q,p,t)$ is defined by Eq. (3). Dif-

$$H(q, p, t) = \sum_{j=1}^{f} p_j\dot{q}_j - L(q, \dot{q}, t) \tag{3}$$

ferentiating this equation and then using Lagrange's equations leads to Eqs. (4), which are Hamilton's

$$\dot{q}_j = \frac{\partial H}{\partial p_j} \qquad p_j = -\frac{\partial H}{\partial q_j} \tag{4}$$

equations. It is essential that H be written as a function of the coordinates and momenta without the velocities appearing.

Phase space. Hamilton's equations are most easily interpreted by introducing the phase space of the system. This is a space of $2f$ dimensions in which the canonical coordinates and momenta serve as the Cartesian coordinates of a point, the system's phase point, that represents the state of motion of the system. Hamilton's equations give the velocity of this phase point as a function of its position in phase space and the time. The important Liouville theorem follows directly from Hamilton's equations. Consider many points in a neighborhood of phase space at a given time. These points represent different possible states of motion of the system, or the motions of many distinct identical systems. As time proceeds, these points move according to Hamilton's equations. If these points are pictured as particles of a fluid flowing through phase space, this flow is that of an incompressible fluid because the divergence of the velocity field in the fluid vanishes. For additional information on Liouville's theorem and phase space *see* STATISTICAL MECHANICS

Applications. As they stand, Hamilton's equations (4) are seldom easier to integrate directly than Lagrange's. Elimination of the *p*'s from Hamilton's equations leads to Lagrange's equations. Hamilton's equations are of great advantage in more general discussions, and they permit the making of canonical transformations that can lead to simplifications. They also lend themselves to numerical integration in some cases where Liouville's theorem is important, as in ion and electron optics. *See* CANONICAL COORDINATES AND TRANSFORMATIONS.

The Hamiltonian function H of classical mechanics is used in forming the quantum-mechanical Hamiltonian operator for systems having a classical analog. *See* NONRELATIVISTIC QUANTUM THEORY.

John L. Safko; Philip Stehle

Bibliography. V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2d ed., 1989; H. C. Corben and P. Stehle, *Classical Mechanics*, 2d ed., 1960, reprint, Dover, New York, 1994; H. Goldstein, C. P. Poole, Jr., and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; I. Percival and D. Richards, *Introduction to Dynamics*, 1982.

Hamilton's principle

Hamilton's principle obtains Lagrange's equations by variations about possible motions of the system between two times (that is, it is an integral principle) rather than changes in the instantaneous state (which characteristic of a differential principle). *See* LAGRANGE'S EQUATIONS.

The meaning that is attached to the phrase "motion of the system between two times" (t_1 and t_2) needs explanation. The problem of determining how a system moves may be formulated in the following way: If the configuration of the system at time t_1 is specified by the generalized coordinates $q_1(t_1), \ldots, q_f(t_1)$ and at the time t_2 by $q_1(t_2), \ldots, q_f(t_2)$ [where f is the number of degrees of freedom of the system], then it is required to find the trajectory along which the system travels from the initial to the final configuration. Hamilton's principle addresses this problem similarly to the way that a geometer addresses the problem of finding the shortest path lying in a curved surface between two given points on the surface. The geometer specifies the distance ds between any two close-lying points in terms of the coordinates q_i of the two points and their differences, the coordinate differentials dq_i. Hamilton defined a characteristic function Φ,

$$\Phi = \int_{t_1}^{t_2} L(q, \dot{q}, t)\, dt$$

using the Lagrangian function $L(q, \dot{q}, t)$, called the Lagrangian of the system, in a way analogous to the geometer's metric. Hamilton's principle states that the system follows the trajectory that makes the integral in the equation above have a minimum value, provided the time interval between times t_1 and t_2 is not too great. This principle produces Lagrange's equations of motion for the system. Either variational principle (Lagrange's or Hamilton's) produces the same Lagrange's equations. *See* DEGREE OF FREEDOM (MECHANICS); DIFFERENTIAL GEOMETRY; LAGRANGIAN FUNCTION.

Hamilton's principle should not be confused with the Hamiltonian or Hamilton's equations of motion, discussed elsewhere. *See* CANONICAL COORDINATES AND TRANSFORMATIONS; HAMILTON-JACOBI THEORY; HAMILTON'S EQUATIONS OF MOTION.

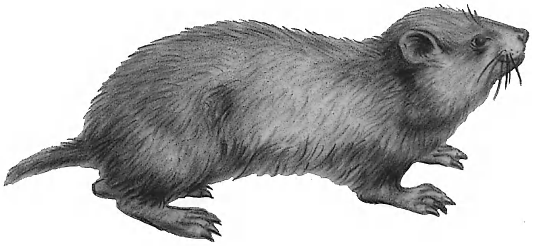
Philip Stehle; John L. Safko

Bibliography. T. L. Chow, *Classical Mechanics*, 1995; H. C. Corben and P. Stehle, *Classical Mechanics*, 2d ed., 1960, reprint, Dover, New York, 1994; H. Goldstein, C. P. Poole, Jr. and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; D. Hestenes, *New Foundations for Classical Mechanics*, 2d ed., Kluwer, Dordrecht, 1999; W. Yourgrau and S. Mandelstam, *Variational Principles in Dynamics and Quantum Theory*, 3d ed., 1968, reprint 1979.

Hamster

The common name for any of 14 species of rodents in the family Cricetidae. The natural range of most of these species is Asia but a few, such as the common hamster (*Cricetus cricetus*), are found in Europe.

The common hamster is a solitary, aggressive animal with interesting burrowing and hoarding habits (see **illus.**). The burrows are not deep, rarely more than 2 ft (0.6 m), and consist of a large central chamber with radiating side chambers for special purposes, such as for hoarding food, for living quarters, and for excretion. Hamsters are clean animals and avoid soiling their living quarters. A so-called summer chamber is used for breeding. A single animal may hoard as much as 200 lb (90 kg) of roots, seeds, nuts, and various tubers, and each type of food is



The hamster, a pugnacious and solitary rodent.

stored in a separate chamber. The hamster goes into its burrow in the autumn, closes off the entrance, and goes to sleep. It does not hibernate deeply and wakes up from time to time to eat. *See* HIBERNATION AND ESTIVATION.

The hamster uses its incisors to carry large food materials, such as carrots, while small foods, such as corn or nuts, are carried in the large cheek pouches. Foraging is always nocturnal. This animal also hunts lizards, birds, mice, insects, and snakes. The hamster has 16 teeth and the dental formula is I 1/1 C 0/0 Pm 0/0 M 3/3.

This rodent has scent glands on each flank, and by rubbing any structure with its flanks, it marks a territory as a warning to other hamsters. When mating, the male invades the territory of the female after first mixing his scent with hers. After mating, the female drives the male away and raises the young herself.

The golden hamster (*Mesocricetus auratus*) is a closely related species, first known from one specimen found in 1839; it was not seen again until 1930, when a family of 13 animals was dug up in Syria. Since that time, it has not been found alive in the wild state. This species makes a good pet and is extensively used as a laboratory animal for experimental purposes, especially in studies of the physiological aspects of hibernation. Prior to hibernation, changes occur in many animals in which the body fats change chemically and form what is termed brown fat. It has been postulated that brown fat is related to the rapid production of heat during arousal. The gestation period is only 15 days and adulthood is attained in about 11 weeks. Other species of hamsters are *Calomyscus bailwardi*, the mouselike hamster found in Asia, and nine species of *Cricetelus*, the Eurasian hamsters found in Europe and western Asia. *See* RODENTIA; SCENT GLAND; TERRITORIALITY.

Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Haplopoda

An order of carnivorous branchiopod crustaceans formerly included in the order Cladocera. Only one species, the fresh-water *Leptodora kindti*, is included.

The body is about 9 mm (0.4 in.) long. *Leptodora* is among the most transparent of multicellular freshwater animals. It swims slowly by means of enormous antennae and seizes its prey with its six pairs of segmented, grasping, thoracic limbs. The mandibles are styliform and of a type unique within the Branchiopoda. There is a single, median sessile eye. The carapace is reduced to a dorsal brood pouch and does not protect the body.

Parthenogenetic eggs and young are carried in the brood pouch in summer. In autumn, fertilized resting eggs are carried there which are shed freely, and the eggs overwinter. They hatch in spring as nauplii, a stage eliminated from the parthenogenetic phase of the life cycle.

Leptodora is widely distributed in the plankton of larger lakes of the Holarctic region. *See* BRANCHIOPODA.

Geoffrey Fryer

Haplosclerida

An order of sponges of the class Demospongiae, including species with a skeleton made up of a single category of siliceous megascleres embedded in spongin fibers or joined together in a network by a spongin cement. The megascleres are usually diactinal, sometimes monactinal. Microscleres (never asters) are present or absent. A modified dermal skeleton

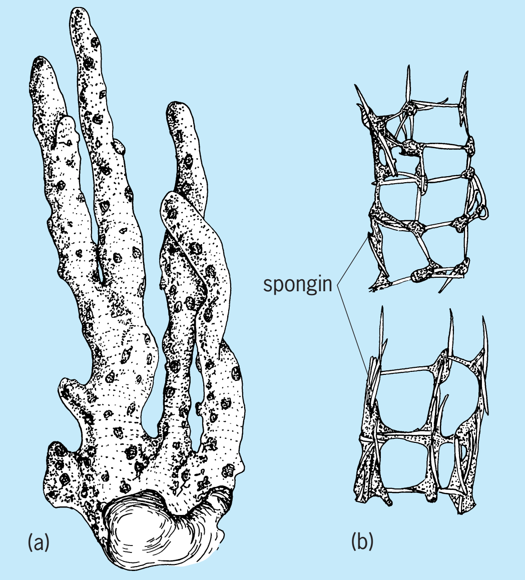


Fig. 1. *Haliclona* morphology. (*a*) Whole sponge. (*b*) Portions of the skeleton of two specimens of *H. oculata* showing varying amounts of spongin joining the spicules.

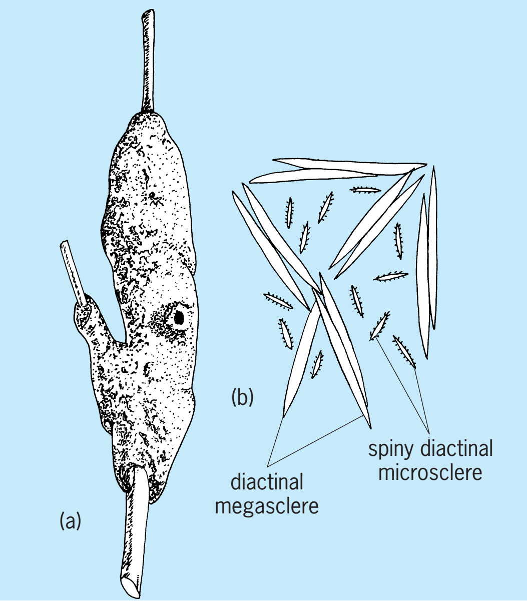


Fig. 2. Spongillid morphology. (*a*) An encrusting sponge growing on a twig. (*b*) Spiculation of *Spongilla lacustris*.

is absent. Several genera lack spicules and have a skeleton of spongin fibers only. Species of this order are encrusting, massive, or lobate, and many species form large upright branching colonies (**Fig. 1**), the branches often being hollow.

Haplosclerid sponges inhabit all seas and are especially abundant in tidal and shallow waters of the continental shelf. Some species occur down to depths of 6600 ft (2000 m).

The family Spongillidae is restricted to fresh water except for a few species which have secondarily invaded brackish water. Spongillids (**Fig. 2**) are encrusting, massive, or branching in form and typically inhabit the edges of streams, ponds, and lakes where light shade is present. They are chiefly gray, brown, or white in color. One species is bright yellow and many are green from the occurrence of zoochlorellae in the cells.

Fossil spongillids are known from the Jurassic Period. Spongillid spicules commonly occur in Pleistocene lake sediments. *See* DEMOSPONGIAE.

Willard D. Hartman

Haplosporea

A class of protozoa, often known as Acnidosporidia, in the subphylum Sporozoa. Haplosporea are distinguished from other similar groups by the production of spores lacking polar filaments. The spores are enclosed in a membrane, and each spore contains a single sporozoite.

Two orders make up the class according to the scheme proposed by the Committee on Taxonomy and Taxonomic Problems of the Society of Protozoologists. These are the Haplosporida and Sarcosporida. Haplosporida are parasites of invertebrates and primitive chordates (Ascidia). Certain species found in fish have also been assigned to this order by some investigators. Sarcosporida are muscle parasites of vertebrates, particularly warm-blooded ones. In many respects Sarcosporida resemble fungi, which some investigators have thought them to be. *See* HAPLOSPORIDA; PROTOZOA; SARCOSPORIDA; SPOROZOA.

Reginald D. Manwell

Haplosporida

A group of Protozoa sometimes regarded as an order within the class Haplosporea. The chief distinguishing characteristic of the Haplosporida, and the one from which their name is derived, is the production of uninucleate spores that lack polar capsules and polar filaments. A spore of *Urosporidium fuliginosum,* a haplosporidan parasitizing a polychaete annelid, is shown in **Fig. 1**. The spores of some species have lids.

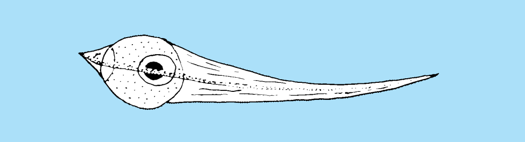


Fig. 1. Spore of *Urosporidium fuliginosum*.

Taxonomy. The number of species within the order is highly uncertain, partly because some species originally assigned to the Haplosporida have had to be reassigned to other groups after more intensive study, and partly because organisms which have no evident importance often receive little study. M. Caullery, who with F. Mesnil proposed the order in 1899, recognized only 5 genera, which include fewer than 20 valid species. However, he also published a table listing 25 other genera which have been assigned to the order by various authors. Some of them may remain there, but the later research mentioned above resulted in the transfer of certain genera to other groups, for example, the fungi. There has also been disagreement among authorities as to whether existing knowledge justifies the recognition of certain genera as Haplosporida; R. R. Kudo includes several that Caullery rejected or regarded as uncertain.

Life cycle. Very little is known of the life cycles of most Haplosporida. The life cycle of *Icthyosporidium giganteum*, a fish parasite, was worked out many years ago (**Fig. 2**); however, this organism is possibly not a true haplosporidan. It is not in Caullery's list of valid genera, but is included among the Haplosporida by Kudo. The cycle begins with a uninucleate spore that lacks polar capsules and polar filaments. These spores are often of a bizarre shape and represent the infective stage. The spore develops into a multinucleate plasmodium in the coelom or tissues of the host. The multinucleate plasmodium gives rise to uninucleate forms known as sporoblasts, which in turn produce the spores.

Parasitic species. Haplosporida are mainly parasites of invertebrates, such as rotifers, annelids, crustacea, insects, and mollusks, but they also occur in

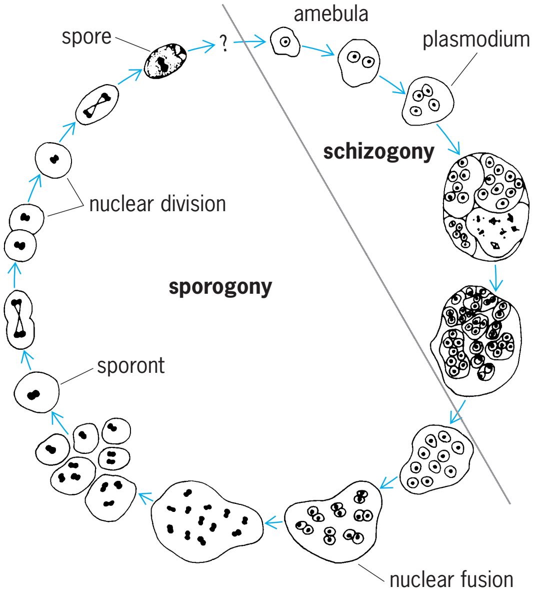


Fig. 2. Stages in the development of *Icthyosporidium giganteum*.

the bodies of Ascidia (tunicates), which are primitive chordates. If the genus *Icthyosporidium* is properly included in the order, Haplosporida also parasitize fish. One species causes "neon" fish disease. A species of *Rhinosporidium* was described from a human, but this genus has since been assigned to the fungi.

It is of considerable biological interest that some species of Haplosporida are hyperparasites, that is, parasitic on other parasites. They have been found in gregarines parasitizing annelids, and also in flukes.

Whether Haplosporida are of any real importance in nature remains unknown. Doubtless, they play some part in the maintenance of biological balances, but evidence suggests that few of them are pathogenic. *See* HAPLOSPOREA; PROTOZOA; SPOROZOA. Reginald D. Manwell

Bibliography. A. R. Anderson, *Comparative Protozoology*, 1988; M. C. Meyer et al., *Essentials of Parasitology*, 4th ed., 1988; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Harbors and ports

A harbor is a geographic location with a body of water of sufficient depth for vessels to enter to seek rest, obtain supplies, and find shelter from storms or other natural phenomena. Harbors typically are naturally occurring but may be constructed. The modern harbor is a place where vessels are built, launched, anchored, and repaired, and usually contains terminals for incoming and outgoing vessels to transport cargo and people. Harbors typically are used for commercial, naval, and fishery vessels, as well as refuge for small craft. Most harbors that accommodate large vessels are situated at the mouth of a river or at some point where it is easy to transfer cargoes inland by river barges, railroads, or trucks.

Harbors may be classified as three types: landlocked with access to the sea usually by an inlet or channel of limited size; unprotected with significant exposure to the hazards of changing tides and weather-induced phenomena (such as waves, fog, and ice); and artificial. The artificial harbors are fashioned by dredging ship channels and by constructing jetties, breakwaters, and other features to protect and accommodate vessels as necessary.

Ports. The term harbor sometimes is confused with the term port. A port is located usually within a harbor and contains facilities dedicated to the movement of cargo and passengers to another destination. A modern port must be deep enough for large vessels to transit and anchor safely, and must provide a direct channel to open water. A modern port also must have enough room for docks, loading and unloading machinery, warehouses or open spaces, and efficient access and distribution. Many ports are situated within cities with large populations. Consequently, the modern harbor must help ensure that the working port does not deny the population access to the waterfront for recreation.

With the exception of some artificial harbors constructed during wartime, few harbors are entirely human-made. The ideal port is deep enough to accommodate the largest cargo freighters, container ships, tankers, and cruise ships in use. It must have enough room for large ships to maneuver around each other. The bed of the harbor should not be so rocky, sandy, or muddy that anchors cannot be safely dropped and secured. In latitudes where sea ice is present, harbors should be situated so that shipping operations are curtailed as little as possible during the ice season.

Harbors are developed according to prevailing natural conditions (**Fig. 1**). A natural coastal harbor is sheltered from the wind and sea by virtue of its location within a natural coastal indentation or in the protective lee of an island, cape, reef, or other natural barrier. When needed, breakwaters are constructed to provide shelter or supplement inadequate shelter already provided by natural sources. In some coastal harbors, locks or other mechanical devices may be used to provide sufficient water to float vessels at all stages of the tide. Facilities may consist of quays or wharves parallel to the banks of the channel, or piers or jetties that extend into the channel. In a natural river harbor, the waters are not restrained by any artificial means. In some river harbors, slips or basins have been excavated in the banks, obliquely or at right angles to the axis of the stream. As with coastal harbors, river harbors may be equipped with locks or other mechanical devices to provide sufficient water to float vessels at all stages of the tide. A canal or lake harbor is located in the interior portion of a canal or lake that is connected with the sea by a navigable waterway. An open roadstead is a port that has no natural or artificial barrier to provide shelter from the wind, sea, and swell.

Port construction. Among the most important structures at any modern port are breakwaters. These barriers are designed to intercept or deflect the force of moving water. Their main function is to protect the harbor area from severe wave action. Thus, breakwaters provide an artificial means of forming suitable harbors. They also are used to deflect currents and to control silting of channels. They must be strong enough to withstand the impact of waves and should be located to avoid detrimental current or wave deflections. Fine examples of ports protected by breakwaters are Dover, England, and Madras, India. *See* NEARSHORE PROCESSES.

Dredging operations are done at many ports, especially those located at the mouths of rivers to keep the channels free of silt and other obstructions that hamper navigation. The viability of some ports cannot be maintained without dredging. For example, the Port of New York and New Jersey, the third largest port in the United States and one of the largest ports in the world, naturally is about 6 m (19 ft) deep. Virtually all of its channels must be dredged regularly to maintain required depths, ranging from

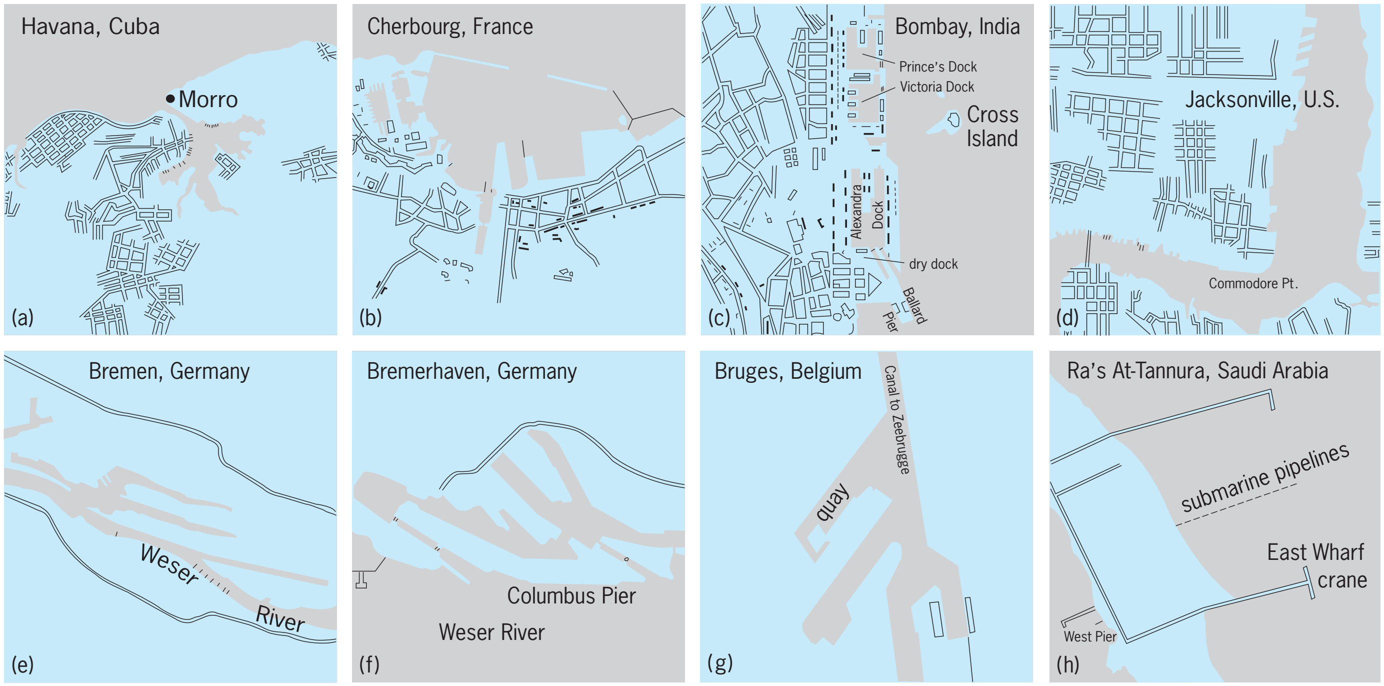


Fig. 1. Examples of different types of harbors. (*a*) Coastal [natural]. (*b*) Coastal [breakwater]. (*c*) Coastal [tide gates]. (*d*) River [natural]. (*e*) River [basins]. (*f*) River [tide gates]. (*g*) Canal or lake. (*h*) Open roadstead. (*U.S. Navy Hydrographic Office*)

10 to 15 m (35 to 50 ft). The large volume of material that is removed from the bottoms of the channels is used for beneficial purposes, ranging from beach nourishment to the upland capping and remediation of brownfield sites (properties where pollutants and contaminants complicate use). *See* RIVER ENGINEERING.

Port efficiency. The amount of cargo moving around the world via ships increases steadily each year. Constructing new ports to accommodate this cargo is neither practical nor inexpensive when such a need must compete with other, more desirable land uses. Therefore, existing ports must improve their efficiency, especially when ports compete with each other for business. The efficiency of a port depends predominantly on the size of the ships that will use it. The largest ports load and unload cargo mainly from container ships and tankers, and must dedicate large areas of land and buildings to store cargo temporarily as it awaits transit to end users. Electric cranes, automatic hoists, fast belt loaders, pneumatic tubes for grains, and rapid pumping facilities for oil, water, and other liquids have been developed to speed loading and unloading. Electronically controlled equipment regulates these operations where previously longshoremen accomplished the same tasks at considerable cost and time. To move cargo out of the port, a port must have both an efficient system of roads and railways. Provision must also be made for customs inspections, banking, and other commercial enterprises that handle the paper work involved in marine insurance, finance, and the preparation of shipping documents. *See* MATERIALS-HANDLING EQUIPMENT.

Cargo security and navigation safety also are prime considerations. The millions of enclosed containers that carry cargo around the world are a major focus of security. Security initiatives undertaken by ports include the hardening of the physical infrastructure, closed-circuit television, early warning systems, security awareness and training, evacuation and emergency planning, stakeholder coordination, cargo and container inspections, and coordination with various levels of government. To help ensure the safe navigation of ships, ports use lighthouses, buoys, beacons, radio direction finders, radar, sonar, geographical positioning systems (GPS), and tugboats with competent harbor pilots to facilitate vessel transit. *See* BUOY; COASTAL ENGINEERING; ELECTRONIC NAVIGATION SYSTEMS; LIGHTHOUSE; PILOTING.

Cargo facilities. General cargo facilities handle nonspecialized cargo such as fruit, coffee, fish, metals, cement, coal, meat, rubber, and wood. Cargo of this nature may be packaged in bags, on pallets, or loosely in large storage areas within the confines of the ship. General cargo is unloaded by a ship's gear (such as derricks, winches, electric cranes, or pneumatic processes) at varying locations along the length of the ship, or with equipment located on the pier. Unloaded cargo is transported from shipside by truck or train to landside temporary storage. While awaiting transport to its final destination, cargo may be stored in open areas or in sheds, depending upon the nature and volume of the cargo as well as the time interval before final transport.

Container terminal facilities. Container terminal facilities move cargo between land and ship in preloaded standard-sized metal boxes [typically, 12 m (40 ft) long by 2.4 m (8 ft) wide by 2.6 m (8.5 ft) high]. Container terminals require open, paved storage areas adjacent to a ship berth to store containers temporarily while they await loading onto a ship or final shipment to their destination inland. Containerization began in the 1950s with the goal of shipping cargo more rapidly and using fewer personnel. Containers are brought to and from upland holding areas (container yards) by truck or train. They are brought close to the ship, onto which they are individually loaded. A special crane lifts a container from the truck or train vertically and then moves it horizontally to the appropriate storage location aboard ship. Containers are placed first in the hold (interior storage area) and then stacked on the deck of a ship. This process is reversed to unload a container from a ship. The first container ships were able to transport approximately 2000 containers. Today, one ship can move as many as 10,000 containers, and container ships continue to grow in size. *See* MERCHANT SHIP.

A container yard must be organized and managed carefully to ensure that all containers (which can number in the thousands or hundreds of thousands at any given time) can be accessed in the minimum amount of time. Containers also move refrigerated cargo.

Equipment for moving containers within the yard includes straddle carriers and side loaders (**Fig. 2**). Straddle carriers are versatile and provide for the efficient movement and repositioning of containers that will be loaded onto the ship, train, or truck by other equipment. Side loaders move containers primarily from the yard on and off trucks or trains, but are being used increasingly to help optimize temporary yard storage. Ports are seeking to improve the efficiency of handling containers within the yard and are assessing the feasibility of automated systems. *See* MARINE CONTAINERS.

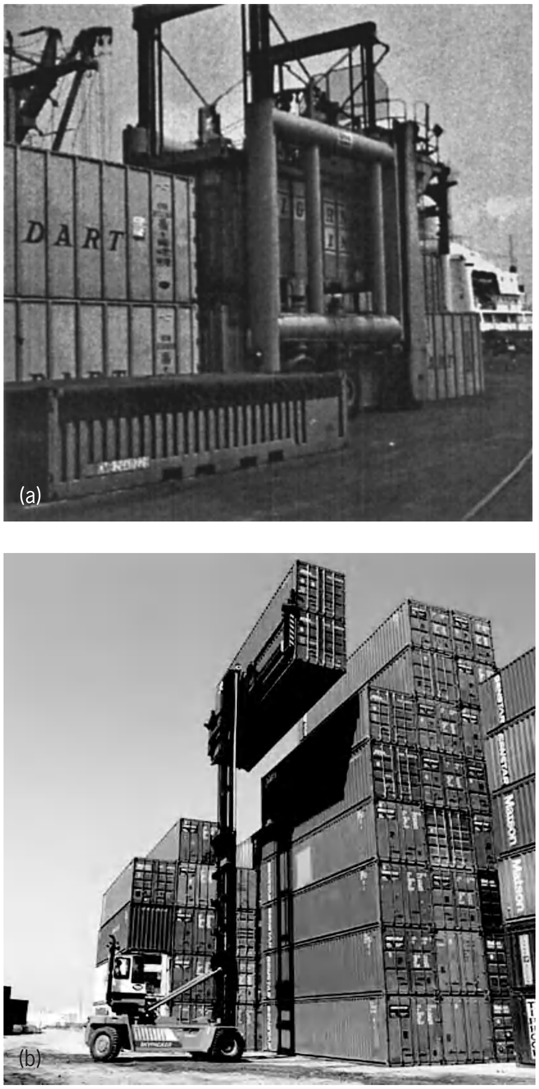


Fig. 2. Examples of container-moving equipment (*a*) Straddle carrier. (*b*) Side loader.

Other types of terminal facilities. General cargo-handling facilities include roll-on/roll-off terminals that accommodate ships that carry cargo in the form of wheeled vehicles. Ferry terminals are designed to provide the efficient transfer of people or vehicles between ferries and high-volume roads, highways, or mass-transit facilities in large urban environments.

Bulk-liquid terminals accommodate vessels transporting liquid cargoes in large volumes, such as crude oil. Bulk liquid is transferred between vessels and storage tanks on land by means of pipelines, with the cargo normally discharged by the ship's pumps. The terminal supplies the energy for pumping the bulk liquid from storage onto the vessel. Since many bulk liquids are flammable or toxic, bulk-liquid terminals are located away from population centers and are equipped with special fire and clean-up installations that are maintained continuously. *See* PIPELINE; PUMP.

Dry bulk terminals are designed to handle cargo such as minerals (for example, iron ore), coal, foodstuffs (for example, grain), and other dry commodities (for example, cement). Depending on the type of commodity, storage is of three basic types: open storage for commodities that are not subject to serious degradation from the elements, sheds for cargo that may be degraded in the short term by the elements, and silos for commodities such as grain or cement where protection from the elements is essential.

Thomas F. Costanzo

Bibliography. H. Agershore, *Planning and Design of Ports and Marine Terminals*, 2004; R. W. Gastil, *Beyond the Edge: New York's New Waterfront*, 2002; J. Gaywaithe, *Design of Marine Facilities for the Berthing, Mooring and Repair of Vessels*, 2004; H. Stevens, *The Institutional Position of Seaports: An International Comparison*, 1999; G. P. Tsinker, *Port Engineering: Planning, Construction, Maintenance and Security*, 2004.

Hardness scales

Arbitrarily defined measures of the resistance of a material to indentation under static or dynamic load, to scratch, abrasion, or wear, or to cutting or drilling. Standardized tests compare similar materials according to the particular aspect of hardness measured by the test. Widely used tests for metals are Brinell, Rockwell, and Scleroscope tests, with modifications depending upon the size or condition of the material. Indentation tests compare species of wood or flooring materials, and abrasion tests serve as an index of performance of stones and paving materials.

Hardness tests are important in research and are widely used for grading, acceptance, and quality control of manufactured articles. The hardness designation or scale is associated with the test method or instrument used.

Scratch hardness. Resistance to scratching is defined by comparison with 10 selected minerals, which are numbered in the order of increasing hardness. This mineralogical scale, called Mohs scale, is 1, talc; 2, gypsum; 3, calcite; 4, fluorite; 5, apatite; 6, orthoclase; 7, quartz; 8, topaz; 9, corundum; and 10, diamond. Minerals lower in the scale are scratched by those with higher numbers. The scale is extended to provide finer distinction of harder materials by additional minerals: 7, vitreous pure silica; 8, quartz; 9, topaz; 10, garnet; 11, fused zirconia; 12, fused alumina; 13, silicon carbide; 14, boron carbide; and 15, diamond.

File-test hardness. Materials are differentiated qualitatively according to resistance to scratching or cutting by files especially selected for the purpose. Whether or not a visible scratch is produced on the material indicates its hardness in comparison with a sample of desired hardness. The method is used for routine inspection of hardened surfaces in production.

Brinell hardness. Resistance to indentation by a hardened steel or tungsten carbide ball under specified load is the basis for Brinell hardness. Standard procedure uses a 10-mm ball with loads of 3000 kg for hard material, 1500 kg for intermediate, and 500 kg or less for soft materials. Various machines apply and control the specified load. The diameter of the impression is measured with a micrometer microscope. Brinell hardness number (Bhn), expressed in kilograms per square millimeter, is obtained by dividing the load by the spherical surface area of the impression. Different-size balls may be used according to size, thickness, and hardness of the specimen, and give the same hardness number provided the loads are proportional to the square of the ball diameter. Carboloy balls are used for very hard material. Time of load application, minimum thickness, and size of specimen are are standardized. A close relation exists between Bhn and ultimate tensile strength.

Vickers hardness. Indentation of a square-based diamond pyramid penetrator with an angle between opposite faces of 136° measures Vickers hardness. Applied load may be varied from 5 to 120 kg in increments of 5 kg according to size of test piece. Vickers hardness number, also called diamond pyramid hardness, is equal to the load divided by the lateral area of the pyramidal impression. The area is computed from measurements of the diagonals of the square impression. Vickers hardness is the most reliable measure for very hard material and is applicable to thin sheets and hardened surfaces.

Rockwell hardness. Depth of indentation of either a steel ball or a 120° conical diamond with rounded point, called a brale, under prescribed load is the basis for Rockwell hardness. The ball is normally $^{1}/_{\sqrt{16}}$ in. in diameter, but $^{1}/_{8}$-, $^{1}/_{4}$-, or $^{1}/_{2}$-in. balls are used for soft materials. A specially designed machine applies loads of 60, 100, or 150 kg. The depth of impression, referred to the position under an initial minor load, is indicated on a dial whose graduations represent the hardness number. Hardness is designated by a number with a standard system of prefix letters to indicate type of penetrator and load used.

Superficial Rockwell hardness is measured by a special machine differing from the standard Rockwell tester in that it applies lighter loads with a more sensitive depth-measuring dial. It produces a shallow impression and is suitable for thin sheet material and where surface hardness is of limited depth.

Monotron hardness. The pressure in kilograms per square millimeter required to embed a 0.75-mm hemispherical diamond penetrator to a depth of 0.0018 in., producing an impression 0.36 mm in diameter, is the measure of Monotron hardness. The depth is controlled by a separate dial graduated to 1 kg/unit area. The method is applicable to the entire range of hardness and is suited to thin sheet and case-hardened surfaces.

Shore Scleroscope hardness. Height of rebound of a diamond-tipped weight or hammer falling within a glass tube from a height of 10 in. and striking the specimen surface measures Shore Scleroscope hardness. The standard hammer is $^{1}/_{4}$ in. in diameter, $^{3}/_{4}$ in. long, and weighs $^{1}/_{\sqrt{12}}$ oz. The hardness number is the height of rebound referred to an arbitrary scale graduated to 140 divisions within the glass tube. The method is a dynamic load test, and the rebound reflects the size of indentation produced, which determines the energy absorbed by deformation and hence that available for rebound. A recording instrument with a dial indicates the rebound hardness directly. Both instruments are portable and permit rapid determinations.

Herbert pendulum hardness. Resistance to cold working is measured as Herbert pendulum hardness. The apparatus consists of a rocking device, called a pendulum, supported on a 1-mm steel ball in contact with the specimen. A curved level bubble measures amplitude of oscillation. The time hardness number is the time in seconds for 10 complete swings of the pendulum through a small arc. The work-hardening capacity is measured by the maximum time hardness after previously repeated single swings of the pendulum. The scale hardness number is the angular oscillation of a half swing after tilting the pendulum through a definite angle before release. Scale hardness is taken as a measure of resistance to flow as in rolling, drawing, and stamping. The method is applicable to studies of machining and forming of metals.

Microhardness. Resistance to indentation over very small areas (as on small parts, the constituents of metal alloys, or for exploration of hardness variations) is called microhardness. One tester employs the Vickers square-based pyramidal diamond penetrator attached to the end of a vertically guided shaft having a weight of 25 g. An arrangement of microscopes permits centering and measurement of the diagonals of the impression. The hardness number is the pressure intensity in kilograms per square millimeter.

Another procedure employs a Tukon tester applying loads of 25–3600 g using a Knoop indenter, which is a diamond ground to produce a diamond-shaped impression with ratio of diagonal lengths of 7:1. The location of the indenter and measurement of the diagonals of the impression are accomplished with microscopes. The hardness number is the ratio of the applied load to the projected area of the impression.

Hardness of wood. The load required to embed a 0.444-in.-diameter steel ball to half its diameter expresses the hardness of wood. Used as a means of comparison, the values vary with species and grain characteristics. Hardness values of poplar and Douglas-fir are approximately 400 and 900 lb, respectively.

Hardness of paving. Wear or abrasion hardness applies primarily to natural stones, paving or flooring materials. It is measured by a specified test providing an index of service performance. Hardness of stone is reflected by weight loss of a cylindrical core rubbed on a sand bed. Deval abrasion test determines weight loss of a charge of broken stone tumbled in a cylinder. Similarly, the Los Angeles rattler and the standard rattler test for paving block tumble a charge including steel balls in a drum to determine percentage loss of weight as an index of wear. Special wear tests are applied to floor surfaces.

William J. Krefeld/Waldo G. Bowman

Bibliography. H. E. Davis, G. E. Troxell, and F. W. Hauck, *The Testing of Engineering Materials,* 4th ed., 1982; A. A. Ivan'ko, *Handbook of Hardness Data*, 1971; V. K. Sarin (ed.), *Science of Hard Materials*, vol. 3, 1989.

Hardy-Weinberg formula

A basic mathematical relation used in population genetics. It gives the proportion of the various genotypes in a randomly mating population in terms of the frequencies of the genes. The formula is useful for genetic analysis of populations, such as human populations or plants and animals in nature where experimental matings are not possible. It was discovered independently in 1908 by G. H. Hardy, a British mathematician, and W. Weinberg, a German physician who made a number of important contributions to the methodology of human genetics. *See* HUMAN GENETICS; POPULATION GENETICS.

In its simplest form the Hardy-Weinberg formula may be stated thus: If p is the proportion of gene A in the population and $q\,(= 1 - p)$ is the proportion of gene a, then after one generation of random mating the three genotypes AA, Aa, and aa will occur in the proportions p^2, $2pq$, and q^2. In other words the genotypes are given by the appropriate terms in the expansion of the binomial $(p + q)^2$. The extension to multiple alleles is direct.

As a numerical example, the recessive gene a for phenylketonuria, which is a metabolic deficiency resulting in mental retardation, has a frequency in the United States of about 0.01. Therefore the proportions of the three genotypes are:

Homozygous normal	AA	$(0.99)^2$	D 0.9801
Heterozygous normal	Aa	$2(0.99)(0.01)$	D 0.0198
Feeble-minded	aa	$(0.01)^2$	D 0.0001

There are about 200 times as many persons who are heterozygous carriers (Aa) of the gene as there are persons homozygous (aa) for it. This is characteristic of rare recessive factors: almost all of the affected children have normal (heterozygous) parents.

The formula holds only for an infinite population and assumes random mating in the absence of significant mutation pressure or gene transfer between populations. However, it is an accurate approximation in many populations. Random mating in this context means that matings occur without regard to the characters determined by the genes in question or the degree of relationship of the mates, and it is possible for a population to be mating at random with respect to some genes and not for others at the same time. For example, it is appropriate to regard the human population as mating at random for blood group genes, and the data actually show excellent agreement with Hardy-Weinberg predictions. However, the formula would not be expected to hold for genes that determine such characters as skin color or intelligence, which strongly influence the choice of mates.

Gene frequency analysis is a technique for testing genetic hypotheses in randomly mating populations. The Hardy-Weinberg formula or an extension of it is used to predict the frequency of certain types in the population or in the progeny of certain parental combinations, and these are compared with the frequencies actually observed. *See* GENETICS.

James F. Crow

Bibliography. L. M. Cook, *Case Studies in Population Biology*, 1988; D. S. Falconer, *Introduction to Quantitative Genetics*, 4th ed., 1996; E. L. Green, *Genetics and Probability*, 1979; J. Roughgarden, *Theory of Population Genetics and Evolutionary Ecology: An Introduction,* rev. ed., 1987.

Harmonic (periodic phenomena)

A sinusoidal quantity having a frequency that is an integral multiple of the frequency of a periodic quantity to which it is related. *See* MODE OF VIBRATION.

A harmonic series of sounds is one in which the basic frequency of each sound is an integral multiple of some fundamental frequency. The name exists for historical reasons, even though according to the usual mathematical definition such frequencies form an arithmetic series. An ideal string (or air column) can vibrate as a whole or in a number of equal parts, and the respective periods of vibration are proportional to the lengths. These increasingly shorter lengths or periods form a harmonic series. The name came from the harmonious relation of such sounds, and the science of musical acoustics was once called harmonics. Nowadays, it is customary to deal with ratios of frequency rather than ratios of length and, because frequency is the reciprocal of period, the definition of harmonic in acoustics becomes that given here. *See* MUSICAL ACOUSTICS.

Robert W. Young

Harmonic motion

A periodic motion that is a sinusoidal function of time. It is often called simple harmonic motion. It is the simplest possible type of vibratory motion. The motion is symmetric about its midpoint, at which the velocity is greatest and the acceleration is zero. At the extreme displacements or turning points, the velocity is zero, and the acceleration is a maximum. The motion is characterized by a unique frequency (without overtones).

Harmonic motion may be present in very simple mechanisms. For example, if a wheel is rotating at constant speed about a fixed axis, the projection on any fixed line of the motion of a point on the wheel is simple harmonic. Harmonic motion may also result from the response of a vibrating system to a periodic—in particular a sinusoidal—force. Harmonic motion is the typical motion of most simple systems that have been displaced from a position of stable equilibrium and then released, provided that the damping is negligible. The motion of a pendulum is approximately simple harmonic for small amplitudes. *See* PENDULUM.

If x represents the displacement measured from the midposition and t the time, then harmonic motion can be described by either of the forms in Eqs. (1). The constants A, B, C, and δ are not all

$$\begin{aligned} x &= A\cos(\omega t) + B\sin(\omega t) \\ x &= C\sin(\omega t - \delta) \end{aligned} \tag{1}$$

independent, but have the relations $A = -C\sin\delta$, $B = C\cos\delta$. The amplitude C represents maximum displacement in one direction from the center (half the total motion between extreme positions); δ is a phase angle whose value depends on the precise instant at which the oscillation was started, or alternatively on the phase of the motion when $t = 0$. These quantities are illustrated in **Fig. 1**.

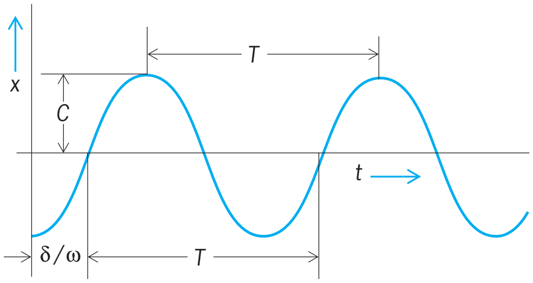


Fig. 1. Representation of simple harmonic motion.

The remaining constant, ω, is known as the angular frequency. Dimensionally, ω is the reciprocal of time. Thus the product ωt is a pure number, to be interpreted as an angle measured in radians. When ωt increases by 2π, the motion repeats. Thus the angular frequency ω is related to ordinary frequency f (number of complete oscillations in unit time) and period T (duration of one complete oscillation) by Eq. (2). The velocity v and acceleration a, obtained

$$\omega = 2\pi f = \frac{2\pi}{T} \tag{2}$$

by differentiating Eq. (1), are given by Eqs. (3) and (4). Because the net force acting on a body is equal to

$$v = \frac{dx}{dt} = \omega C\cos(\omega t - \delta) = \omega\sqrt{C^2 - x^2} \tag{3}$$

$$a = \frac{dv}{dt} = -\omega^2 C\sin(\omega t - \delta) = -\omega^2 x \tag{4}$$

the mass of the body multiplied by its acceleration, Eq. (5) shows that in simple harmonic motion the

$$F = ma = -m\omega^2 x \tag{5}$$

force must be proportional to the displacement.

Conversely, simple harmonic motion occurs whenever, for a body that is displaced from an equilibrium position, there is a net restoring force proportional to the displacement.

Energy considerations. The importance of harmonic motion lies in the simplicity of its time dependence, as given by Eqs. (1), and in the frequent occurrence of linear restoring forces. For sufficiently small displacements of almost any mechanical system from equilibrium, the restoring force or torque is always approximately proportional to the displacement.

This can be explained by some considerations about potential energy. At a point of stable equilibrium, potential energy is necessarily a minimum. If potential energy is a well-behaved function of position, it may be expanded as a series of powers of the distance from any point. (An important exception is electrostatic potential, which varies inversely as the distance from the charge and so is not well-behaved in the immediate neighborhood of the charge.) In the expansion of potential energy about a point of stable equilibrium, the special minimum property guarantees that the linear term must vanish. The next term,

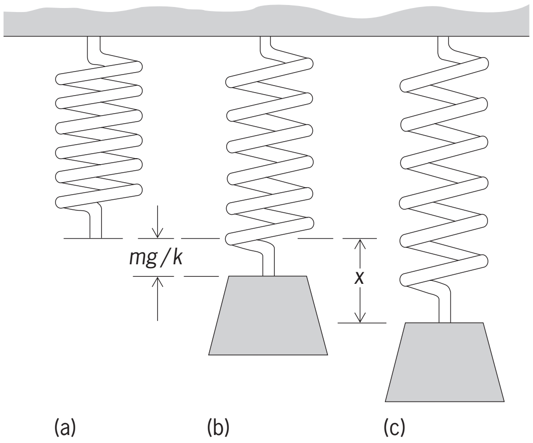


Fig. 2. Weight on an elastic spring. (*a*) Unloaded. (*b*) Statically loaded. (*c*) Weight and spring oscillating.

quadratic in the distance from the equilibrium point, does not ordinarily vanish and provides a good approximation of the potential energy for sufficiently small displacements. A quadratic, or parabolic, variation of potential with distance is equivalent to a force that varies linearly.

In a system oscillating freely about an equilibrium position, the energy changes from potential to kinetic and back again. For simple harmonic motion the potential energy, proportional to the square of the displacement, is greatest at the extremities of the motion. Conversely, the kinetic energy $^1/_2 mv^2$ is zero at the turning points and maximum when the body is going past its equilibrium position. The total energy (kinetic plus potential) is constant. In simple harmonic motion it is proportional to the square of the amplitude of the motion, and the average potential energy just equals the average kinetic energy.

The frequency of a freely oscillating system is determined by the stiffness and inertia of the system. If the motion is harmonic, it is also isochronous, which means that the frequency is independent of the amplitude of the motion.

Weight on elastic spring. If an elastic spring is stretched a distance x beyond its natural length, or compressed a distance x short of its natural length, it exerts a restoring force equal to $-kx$. The stiffness of the spring is measured by the spring constant k. Consider a spring which is suspended vertically and set into vertical oscillations, with a mass m attached to its free end (**Fig. 2**). If m is large compared to the mass of the spring, the latter can be neglected.

The equation of motion for the mass m is Eq. (6),

$$mg - kx = ma = m\frac{d^2x}{dt^2} \qquad (6)$$

where g is the acceleration of gravity. The solution of this equation can be written in the form of Eq. (7).

$$x = \frac{mg}{k} + C\sin(\omega t - \delta) \qquad \omega^2 = \frac{k}{m} \qquad (7)$$

The term mg/k is simply the extension of the spring due to the static weight, and marks the loaded equilibrium position. Displacement from this equilibrium position calls forth a linear restoring force, and the resulting motion is simple harmonic.

Potential energy (PE) for this example is the sum of an elastic energy $^1/_2kx^2$ and a gravitational energy $-mgx$. Then Eq. (8) may be written, showing the

$$\begin{aligned} \text{PE} &= {}^1/_2kx^2 - mgx \\ &= {}^1/_2k\left(x - \frac{mg}{k}\right)^2 - \frac{m^2g^2}{2k} \end{aligned} \qquad (8)$$

quadratic dependence on displacement from equilibrium. The constant term $-m^2g^2/2k$ is of no importance to the motion; it can always be removed by a new choice for the zero of potential energy.

Rotational harmonic motion. Rotational as well as translational simple harmonic motion may occur. In this case the angular displacement is a sinusoidal function of time. A free angular vibration will be simple harmonic if the angular displacement from an equilibrium orientation produces a restoring torque proportional to the displacement. An example is the torsional pendulum (**Fig. 3**). One end of a torsionally flexible elastic rod is held fixed. To the other end is fastened a disk, or another body of large moment of inertia. If the rod is twisted and then released, the rod and disk will undergo angular simple harmonic motion, provided that the torque in the rod is proportional to the angle of twist.

Atomic vibrations. The realization that atoms are continually vibrating in motions that are nearly harmonic is essential for understanding many properties of matter, including molecular spectra, heat capacity, and heat conduction.

In a diatomic molecule, the distance between the atoms is not precisely fixed. There is an equilibrium

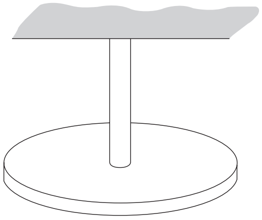


Fig. 3. Torsional pendulum.

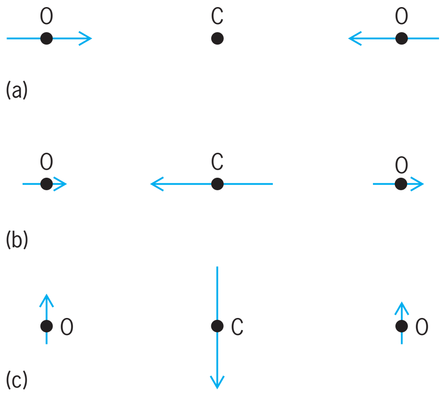


Fig. 4. Three normal modes of the CO_2 molecule. (*a*) Lower-frequency in-line mode. (*b*) Higher-frequency in-line mode. (*c*) Bending mode.

separation corresponding to a minimum in potential energy. The actual magnitude of the interatomic distance oscillates about the equilibrium distance with a motion that is approximately simple harmonic, if the energy of oscillation is small.

In a polyatomic molecule or in a crystalline solid, the atoms also vibrate about equilibrium positions. Because of the large number of degrees of freedom, however, the situation is more complicated. For example, in the carbon dioxide (CO_2) molecule neither oxygen atom by itself moves harmonically, or even periodically. On the other hand, the motion of the CO_2 molecule can be analyzed into a number of independent motions, called normal modes, each of which is by itself simple harmonic. In one such motion, for example, the two oxygen atoms move in phase toward or away from the carbon atom (**Fig.** 4*a*). The actual motion of the atoms in a molecule is a superposition of the various normal modes and a rotation of the molecule as a whole. *See* DAMPING; FORCED OSCILLATION; HARMONIC OSCILLATOR; LATTICE VIBRATIONS; MOLECULAR STRUCTURE AND SPECTRA; PERIODIC MOTION; VIBRATION; WAVE MOTION. Joseph M. Keller

Bibliography. R. Baierlain, *Newtonian Dynamics*, 1983; V. Barger and M. G. Olsson, *Classical Mechanics*, 2d ed., 1995; K. R. Symon, *Mechanics*, 3d ed., 1971.

Harmonic oscillator

Any physical system that is bound to a position of stable equilibrium by a restoring force or torque proportional to the linear or angular displacement from this position. If such a body is disturbed from its equilibrium position and released, and if damping can be neglected, the resulting vibration will be simple harmonic motion, with no overtones. The frequency of vibration is the natural frequency of the oscillator, determined by its inertia (mass) and the stiffness of its restoring force.

The harmonic oscillator is not restricted to a mechanical system, but might, for example, be electric. Typical electronic oscillators, however, are only approximately harmonic.

If a harmonic oscillator, instead of vibrating freely, is driven by a periodic force, it will vibrate harmonically with the period of the force; initially the natural frequency will also be present, but any damping will eventually remove the natural motion. The response of a harmonic oscillator driven by a general force $f(t)$, an arbitrary function of time, is described by a linear differential equation. If $x(t)$ represents the displacement as a function of time t, then the equation of motion is Eq. (1). The left side of Eq. (1) represents the

$$m\frac{d^2x}{dt^2} = -c\frac{dx}{dt} - kx + f(t) \qquad (1)$$

mass m multiplied by the acceleration. The force on the right side of the equation includes, in addition to the restoring force $-kx$ and the driving force, a "viscous damping" force $-c\ dx/dt$ that is proportional to the velocity. (This damping may vanish.) The explicit solution of Eq. (1) is Eq. (2), where $b = c/2m$,

$$x = e^{-bt}\left\{A\sin(\omega t - \delta) + \frac{e^{-bt}}{m\omega}\int_{-\infty}^{t} e^{bT}\sin[\omega(t-T)]f(T)\,dT\right\} \qquad (2)$$

$\omega^2 = (k/m) - (c/2m)^2$, and A and δ are arbitrary constants which can be set equal to zero if the oscillator displacement and velocity were zero before application of the force $f(t)$. The variable of integration T is the time at which the force $f(t)$ is considered to act. *See* DAMPING; FORCED OSCILLATION; HARMONIC MOTION.

In both quantum mechanics and classical mechanics, the harmonic oscillator is an important problem. It is one of the few rigorously soluble problems of quantum mechanics. The quantum mechanical description of electromagnetic, electronic, mesonic, and other fields is usually carried out in terms of a (time) Fourier analysis. The individual Fourier components of noninteracting fields are independent harmonic oscillators.

The hamiltonian for a harmonic oscillator is Eqs. (3) where ω is the angular frequency character-

$$H = \frac{p^2}{2m} + \frac{m\omega^2 q^2}{2} \qquad (3a)$$

$$= \frac{P^2}{2} + \frac{1}{2}\omega^2 Q^2 \qquad (3b)$$

istic of the oscillator. The mass m in Eq. (3*a*) is made to disappear by replacing the coordinate q and the conjugate momentum p by $Q = m^{1/2}q$ and $P = m^{-1/2}p$ in Eq. (3*b*). The corresponding Schrödinger equation for the wave function ψ is then Eq. (4), where E is

$$-\frac{\hbar^2}{2}\frac{d^2\psi}{dQ^2} + \frac{\omega^2 Q^2}{2}\psi = E\psi \qquad (4)$$

the energy, and $\hbar$ is Planck's constant divided by 2π. This equation possesses quadratically integrable solutions only for the characteristic energy values as shown by Eq. (5), where n is any positive integer.

$$E = (n + {}^1/_2)\hbar\omega \qquad (5)$$

For these cases, ψ can be expressed in terms of the Hermite polynomial $H_n(y)$ of degree n, as in Eq. (6) where $y = (\omega/\hbar)^{1/2}Q$ is dimensionless, and Eq. (7) applies.

$$\psi_n = C_n e^{(-y^2/2)} H_n(y) \qquad (6)$$

$$C_n^2 = \left(\frac{\omega}{\pi\hbar}\right)^{1/2}\frac{1}{2^n n!} \qquad (7)$$

See ANHARMONIC OSCILLATOR; LAGRANGE'S EQUATIONS; NONRELATIVISTIC QUANTUM THEORY.

Joseph M. Keller

Bibliography. R. Baierlein, *Newtonian Dynamics,*1983; R. M. Eisberg, *Fundamentals of Modern Physics*, 1961; R. M. Eisberg and R. Resnick, *Quantum Physics*, 2d ed., 1985; H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; C. Kittel, W. D. Knight, and M. A. Ruderman, *Mechanics*, Berkeley Physics Course, vol. 1, 2d ed., 1973; J. W. S. Rayleigh, *The Theory of Sound*, 2d ed., 1894, reprint 1945.

Harmonic speed changer

A mechanical-drive system used to transmit rotary, linear, or angular motion at high ratios and with positive motion. In the rotary version (**Fig. 1**), the drive consists of a rigid circular spline, an input wave generator, and a flexible spline. Any one of these can be fixed, used as the input, or used as the output. Any combination (fixed, driver, or driven) may be used. *See* SPLINES.

In Fig. 1 the fixed member is a rigid circular spline with 132 internal teeth. The driven part is a flexible spline, a ring gear with 130 external teeth of same size as on the rigid spline. The driving member is shown as a two-lobed member generating a traveling circular wave on the flexible spline.

The flexible spline meshes with the rigid circular spline at two diametrically opposite regions of the spline, which is shown flexed as an ellipse with its major axis nearly vertical. The teeth on the splines clear each other along the nearly horizontal minor axis.

As the wave generator rotates, the axes of the flexible spline correspondingly rotate along with the regions of contact and clearance between the teeth on the splines. Because the number of teeth on the flexible spline is less than the number of teeth on the fixed and rigid circular spline, when the wave generator rotates one turn, the flexible spline will rotate $^{2}/_{132}$ parts of a turn in a direction opposite that of the driving wave generator. The speed reduction is therefore 2:132 (or 1:66). By increasing the number of teeth on the two splines, the flexible spline always having two less (with a two-lobed wave generator) than the number on the circular spline, the speed reduction can be increased correspondingly.

In actual construction (**Fig. 2**), the wave generator is composed of a ball bearing with an elliptical inner race. In rotating, the inner race flexes the outer race, engaging its external teeth along its major axis with the internal teeth of the rigid spline.

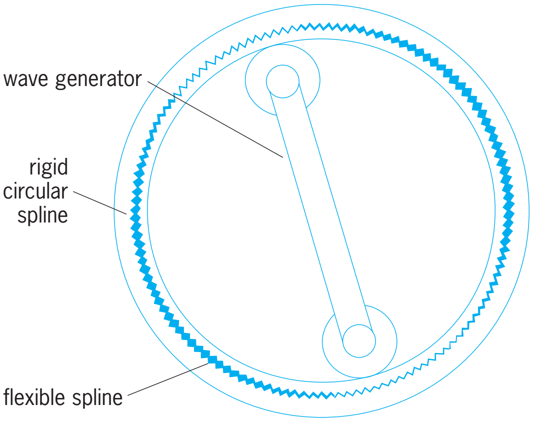


Fig. 1. Rotary-to-rotary harmonic speed changer.

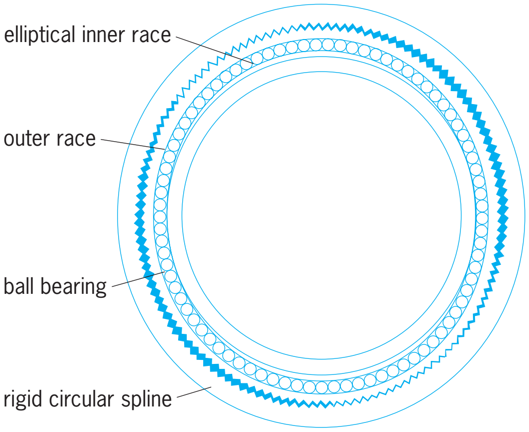


Fig. 2. Diagram of actual construction of a rotary-to-rotary harmonic speed changer.

The advantages of this drive are (1) high-ratio gearing in small space, (2) speed reduction (or increase) in fixed ratio up to 1000:1 in one unit, (3) negligible wear of teeth, (4) balanced bearing loads, (5) negligible backlash, (6) efficiency of approximately 80% in a gear ratio of 400:1, and (7) adaptability to rotary-to-rotary, rotary-to-linear, and linear-to-linear drives.

Paul H. Black

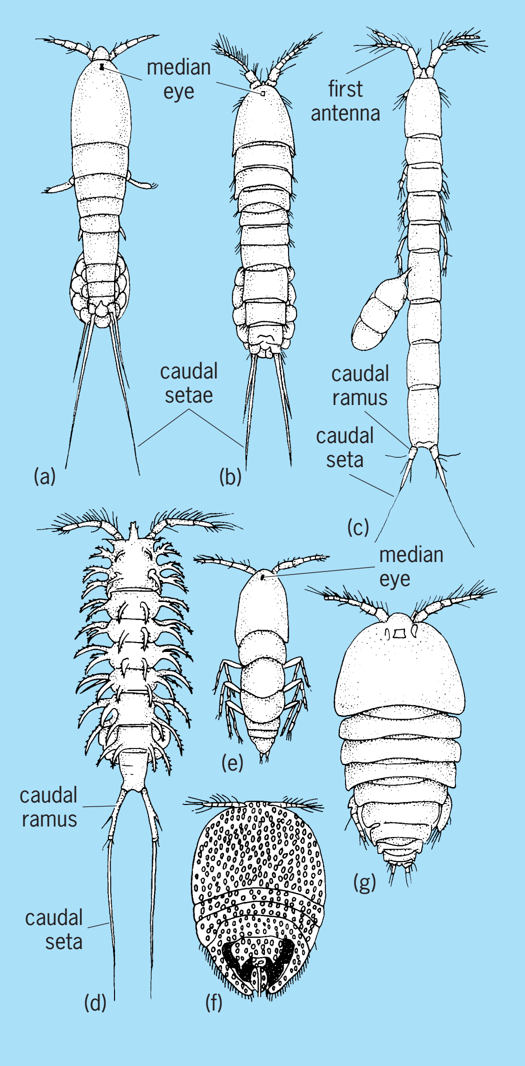
Bibliography. E. A. Dijksman, *Motion Geometry of Mechanism*, 1976; H. H. Mabie and F. W. Ocvirk, *Mechanisms and Dynamics of Machinery*, 4th ed., 1987.

Harpacticoida

An order of the Copepoda derived from the genus *Harpacticus* Milne-Edwards 1840, with about 1700 species known. This group of crustaceans is referred to as Podoplea Copepoda. Their form is variable but is generally linear and more or less cylindrical (see **illus.**). These animals are minute and vary in length from 0.016 to 0.12 in. (0.4 to 3 mm). As a rule, the first thoracic segment is incorporated with the cephalothorax, and the last thoracic segment is included in the abdomen. First antennae in the males are prehensile, while second antennae in both sexes are usually biramous.

The mouthparts of free-living forms are well developed, while in parasitic forms they are reduced or transformed. Maxillipeds as a rule are also prehensile. The first pair of swimming legs generally is modified for grasping. The second to fourth pairs of legs are natatory, with exopodites almost always three-jointed, and endopodites three- to one-jointed or absent. Sexual dimorphism can occur in all legs mentioned. The last pair of legs is reduced and dissimilar in the two sexes. The ovisac may be single or double, but is always ventral. Ova of a few species are laid free. *See* SEXUAL DIMORPHISM.

The nervous system consists of a supra- and a subesophageal ganglion, a circumesophageal ring, and a ventral chain with one pair of ganglia in each



Some typical harpacticoids. (*a*) *Harpacticus chelifer*. (*b*) *Ameira longipes*. (*c*) *Cylindropsyllus laevis*. (*d*) *Anchorabolus mirabilis*. (*e*) *Parategastes spaericus*. (*f*) *Porcellidium ravanae*. (*g*) *Alteutha interrupta*.

segment. The supraesophageal ganglion innervates the eye, the first antennae, and if they are present, the frontal organs. The circumesophageal ring innervates the dorsal musculature of the cephalothorax, the second antennae, and the labrum, while the subesophageal ganglion innervates the rest of the mouthparts. The last-mentioned ganglion is provided with two transverse commissures.

As a rule, a tripartite median eye, the nauplius eye, is present. The parts of the eye, however, may be independent, or the eye may have disappeared. One pair of frontal organs is sometimes found. Antennulae have an olfactory organ, the esthete, on the fourth joint or its equivalent; males often have esthetes on other joints also.

The gonads are single. Their ducts open ventromedially on the first true abdominal segment. Two oviducts open into a seminal receptacle. The vas deferentia are single or paired, and sperm are contained in spermatophores.

The alimentary canal is similar to that in most other copepods. The excretory organ consists of a maxillary gland which is always present, and a rudimentary antennal organ occurs in some forms. Development follows the general plan as it occurs in other copepods. *Elaphoidella bidens* exhibits parthenogenesis, while heterogony is found in *Epactophanes*.

Harpacticoids have a worldwide distribution. They occur in all kinds of aquatic habitats, especially marine, but also among moss and leaves. In the sea they range from the shore to abyssal depths. In general, they are free-living and benthonic; some species are pelagic, parasitic, or commensal. The species may be mono- or polycyclic. A few fresh-water species of Harpacticoida produce resting eggs, or they estivate within cysts. *See* COPEPODA. Karl Lang

Bibliography. K. Lang, *Monographie der Harpacticiden*, 1948; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; G. O. Sars, *Crustacea of Norway*, vol. 5, 1911, vol. 7, 1921.

Hassium

A chemical element, symbol Hs, atomic number 108. It was synthesized and identified in 1984 by using the Universal Linear Accelerator (UNILAC) at Darmstadt, West Germany, by the same team (led by P. Armbruster and G. Müzenberg) which first identified bohrium and meitnerium. The isotope ^{265}Hs was produced in a fusion reaction by bombarding a ^{208}Pb target with a beam of ^{58}Fe projectiles. *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**

The discovery of bohrium and meitnerium was made by detection of isotopes with odd proton and neutron numbers. In this region, odd-odd nuclei show the highest stability against fission. Elements with an even atomic number are intrinsically less stable against spontaneous fission. The isotopes of hassium were expected to decay by spontaneous fission—which explains why meitnerium was synthesized before hassium.

As in the case of bohrium and meitnerium, the isotope was produced by fusion in a one-neutron deexcitation channel; in this case the compound system was ^{266}Hs. The reaction mechanism again was cold fusion. The isotope ^{265}Hs has a half-life of about 2 ms and decays by emission of an alpha particle of 10.36 MeV.

The elements bohrium, hassium, and meitnerium are stabilized by shell effects against spontaneous fission; this special stability may occur because these nuclei may prefer a sausage shape which is predicted to be energetically most favorable for them. *See* BOHRIUM; MEITNERIUM; TRANSURANIUM ELEMENTS. Peter Armbruster

Bibliography. S. Cwiok et al., Fission barriers of transfermium elements, *Nucl. Phys.* A410:254–270, 1983; A. G. Demin et al., On the properties of the element 106 isotopes produced in the reaction Pb + 54Cr, *Z. Phys.*, A315:197–200, 1984; S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; G. Muzenberg et al., The identification of elements 108, *Z. Phys.*, A317:235–236, 1984; G. Muzenberg et al., Observation of the isotopes 264108 and 265108, *Z. Phys.*, A328:49–59, 1987; Y. T. Oganessian et al., On the stability of the nuclei of element 108 with A = 263–265, *Z. Phys.*, A319:215–217, 1984.

Hatchettite

Mountain tallow, a yellow-white to yellow-green hydrocarbon occurring in Belgian coal seams. The material, also called hatchettine, frequently is found in geodes and also in fractures in associated rocks. *See* GEODE.

Hatchettite is translucent but darkens on exposure to air. It is soft, has no odor, is greasy to the touch, and consists of 85.5% carbon and 14.5% hydrogen leading to the empirical formula C_nH_{2n}. Its index of refraction is 1.47–1.50, it melts at 46–47°C (115–117°F), is sparingly soluble in alcohol or ether, decomposes in concentrated sulfuric acid, and has a specific gravity of 0.89–0.98. Irving A. Breger

Hazardous waste

Any solid, liquid, or gaseous waste materials that, if improperly managed or disposed of, may pose substantial hazards to human health and the environment. Hazardous waste can be generated from many types of processes (**Table 1**). It generally has been considered a subset of solid waste and has been distinguished from municipal wastes and nonhazardous industrial wastes. This characterization is not based on science and engineering but on the regulatory history of hazardous wastes. For example, in the United States in the mid-1970s, statutes existed to control air and water pollution, as well as the design of landfills and other facilities to address solid wastes; hazardous waste, however, was not being addressed. As a result, the U.S. Congress enacted the Resource Conservation and Recovery Act (RCRA) in 1976. The law's primary goals are to protect human health and the environment from the potential hazards of waste disposal, to conserve energy and natural resources, to reduce the amount of waste generated, and to ensure that wastes are managed in an environmentally sound manner. The RCRA regulations are found in the Code of Federal Regulations at Title 40, Parts 260 through 280. In fact, RCRA is an amendment to the Solid Waste Disposal Act of 1956. The 1984 amendments to RCRA are known as the Hazardous and Solid Waste Amendments (HSWA). Subtitle C (hazardous waste) and Subtitle D (solid, primarily nonhazardous, waste) provide the structure for comprehensive waste management programs. In addition, RCRA regulates underground storage tanks under Subtitle I and medical waste under Subtitle J. Many other countries have similar programs to control hazardous wastes. Of the 13 billion tons of industrial, agricultural, commercial, and household wastes generated annually in the United States, 2% (more than 279 million tons) are hazardous as defined by RCRA regulations.

Properties. Hazardous wastes may be of any phase—solid, liquid, gas, or mixture. The hazard of the waste can result from its inherent properties, such as its physicochemical properties, including its likelihood to ignite, explode, react, cause irritations, or elicit toxic effects. The inherent properties of hazardous wastes can also be biological, such as that of infectious medical wastes.

Since wastes are transported by fluids (liquids and gases), especially in air and water, the likelihood of

TABLE 1. Typical hazardous wastes generated by selected industries

Waste generator	Types of wastes produced
Chemical manufacturers	Strong acids and bases Reactive wastes Ignitable wastes Discarded commercial chemical products
Vehicle maintenance shops	Paint wastes Ignitable wastes Spent solvents Acids and bases
Printing industry	Photography waste with heavy metals Heavy-metal solutions Waste inks Spent solvents
Paper industry	Ignitable wastes Corrosive wastes Ink wastes, including solvents and metals
Construction industry	Ignitable wastes Paint wastes Spent solvents Strong acids and bases
Cleaning agents and cosmetic manufacturing	Heavy-metal dusts and sludges Ignitable wastes Solvents Strong acids and bases
Furniture and wood manufacturing and refinishing	Ignitable wastes Spent solvents Paint wastes
Metal manufacturing	Paint wastes containing heavy metals Strong acids and bases Cyanide wastes Sludges containing heavy metals

SOURCE: U.S. Environmental Protection Agency, *RCRA: Reducing Risks from Wastes*, Rep. EPA530-K-97-004, September 1997.

TABLE 2. Densities of some important environmental fluids

Fluid	Density (kg m^{-3}) at 20°C unless otherwise noted
Air at standard temperature and pressure (STP) = 0°C and 101.3 N m^{-2}	1.29
Air at 21°C	1.20
Ammonia	602
Diethyl ether	740
Ethanol	790
Acetone	791
Gasoline	700
Kerosene	820
Turpentine	870
Benzene	879
Pure water	1000
Seawater	1025
Carbon disulfide	1274
Chloroform	1489
Tetrachloromethane (carbon tetrachloride)	1595
Lead	11,340
Mercury	13,600

SOURCE: D. A. Vallero, *Environmental Contaminants: Assessment and Control*, Elsevier Academic Press, Burlington, MA, 2004.

contamination strongly depends on physical properties, especially solubility, density, and vapor pressure. If the substances composing the hazardous waste are quite soluble in water, they are hydrophilic. If a substance is not easily dissolved in water, it is hydrophobic. Since many contaminants are organic (consisting of molecules containing carbon-to-carbon bonds and/or carbon-to-hydrogen bonds), the solubility can be further differentiated as to whether the substance is easily dissolved in organic solvents. Such substance are said to be lipophilic. A substance that is not easily dissolved in organic solvents is said to be lipophobic. *See* SOLVENT; VAPOR PRESSURE.

Density is a very important fluid property for environmental situations. For example, in responding to a leak or release of a hazardous substance, the density of the substance must be known. If a substance is burning, whether it is of greater or lesser density than water will be one of the factors on how to extinguish the fire. If the substance is less dense than water, the water will likely settle below the substance layer, making water a poor choice for fighting the fire. So, any flammable substance with a density less than water (**Table 2**), such as benzene or acetone, will require other fire-extinguishing substances. For substances heavier than water, such as carbon disulfide, water may be the proper fire-fighting material. *See* DENSITY.

Ground-water contamination. One of the chief concerns from hazardous wastes is the potential to contaminate ground water, which can be affected by any type of hazardous waste from numerous sources. When a fluid that is dense and miscible (that is, able to be mixed in any concentration without separation of physical phases) seeps underground through the vadose zone, below the water table and into the zone of saturation, the dense contaminants move downward (**Fig. 1**). When these contaminants reach the bottom of the aquifer, the bottom shape dictates their continued movement, along with the slope of the underlying bedrock or other relatively impervious layer. Solution and dispersion near the

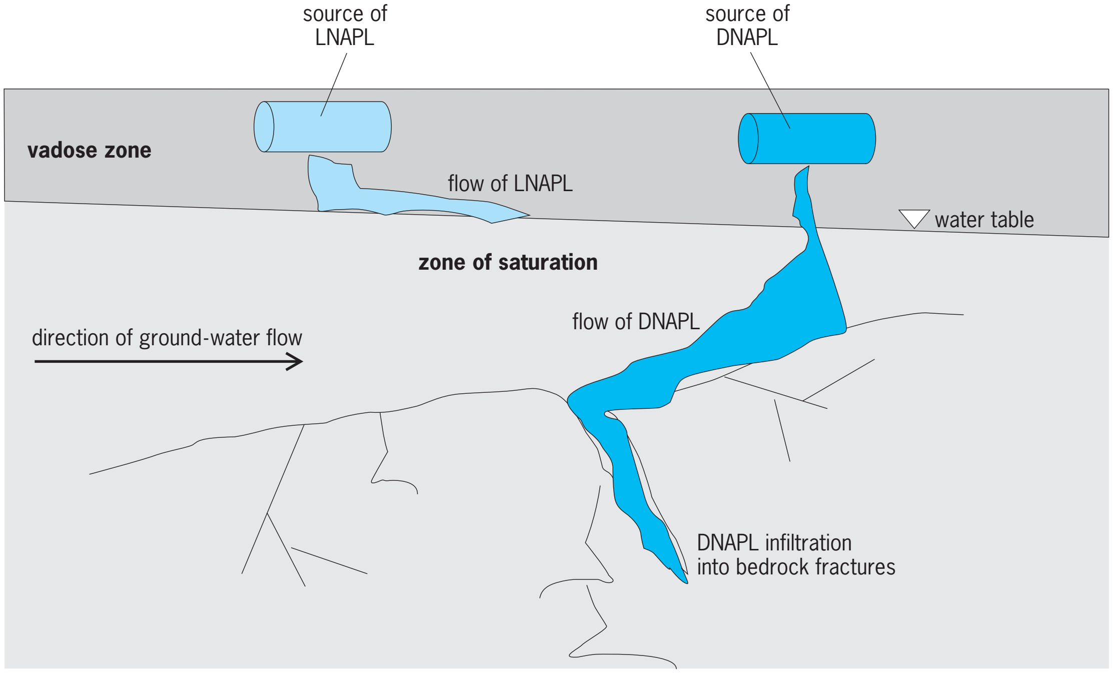


Fig. 1. Importance of density of fluids in ground-water contamination scenarios. The dense nonaqueous phase liquids (DNAPLs) can penetrate more deeply into the aquifer than do the light nonaqueous phase liquids (LNAPLs). The density reference for whether a compound is a DNAPL or an LNAPL is whether it is denser or lighter than water, respectively. The DNAPL may even be against the general flow of the ground water. (*After D. A. Vallero, Environmental Contaminants: Assessment and Control, Elsevier Academic Press, 2004; H. Hemond and E. Fechner-Levy, Chemical Fate and Transport in the Environment, Academic Press, 2000*)

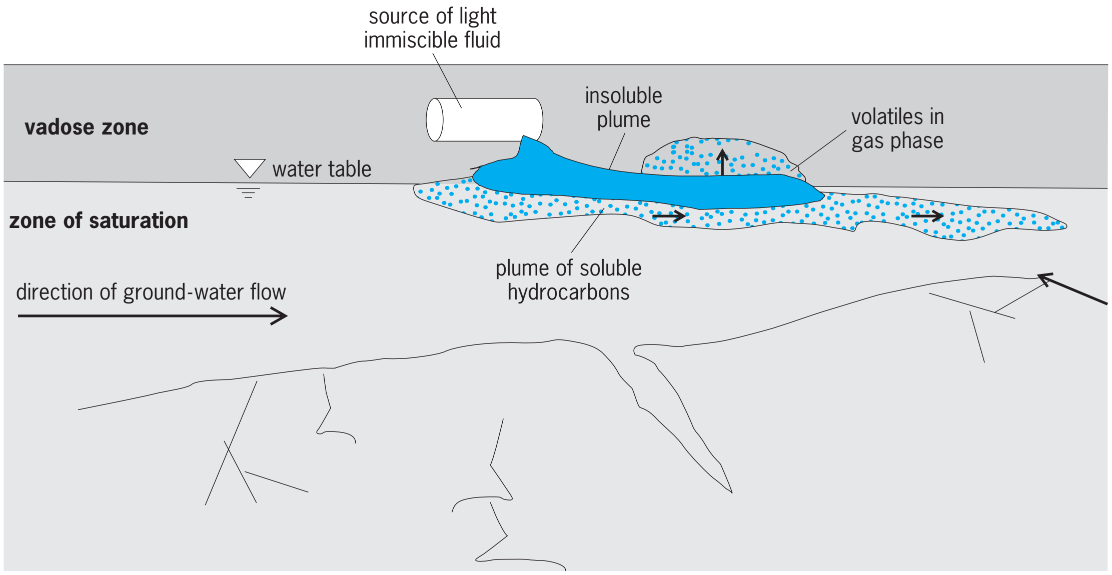


Fig. 2. Hypothetical plume of hydrophobic fluid. (*After D. A. Vallero, Environmental Contaminants: Assessment and Control, Elsevier Academic Press, 2004; and M. N. Sara, Groundwater Monitoring System Design, in Practical Handbook of Ground-Water Monitoring, D. M. ed. by Nielsen, Lewis Publishers, 1991*)

boundaries of the plume will generate a secondary plume that will generally follow the general direction of ground-water flow (**Fig. 2**). Organics more dense than water (DNAPLs) will penetrate more deeply, while the lighter organics (light nonaqueous-phase liquids, or LNAPLs) will float near the top of the zone of saturation.

The physics of this system determines the direction of the contaminant plume's movement. For example, the movement is greatly affected by the solubility and density of the contaminants, but other factors also influence the rate of transport, such as sorption, presence of other solvent besides water, and the amount of organic matter in the soil. In

TABLE 3. Key legal definitions related to hazardous wastes in the United States

Term	Definition
Hazardous waste	A solid waste, or combination of solid wastes, which because of its quantity, concentration, or physical, chemical, or infectious characteristics may (a) cause, or significantly contribute to an increase in mortality or an increase in serious irreversible, or incapacitating reversible, illness; or (b) pose a substantial present or potential hazard to human health or the environment when improperly treated, stored, transported, or disposed of, or otherwise managed.
Medical waste	Any solid waste which is generated in the diagnosis, treatment, or immunization of human beings or animals, in research pertaining thereto, or in the production or testing of biologicals. Such term does not include any hazardous waste identified or listed under subchapter III of this chapter or any household waste as defined in regulations under subchapter III of this chapter. (For "infectious waste" definition, see CFR Part 243 below.)
Mixed waste	Waste that contains both hazardous waste and source, special nuclear, or by-product material subject to the Atomic Energy Act of 1954 (42 U.S.C. 2011 et seq.).
Solid waste	Any garbage, refuse, sludge from a waste treatment plant, water supply treatment plant, or air pollution control facility and other discarded material, including solid, liquid, semisolid, or contained gaseous material resulting from industrial, commercial, mining, and agricultural operations, and from community activities, but does not include solid or dissolved material in domestic sewage, or solid or dissolved materials in irrigation return flows or industrial discharges which are point sources subject to permits under section 1342 of title 33, or source, special nuclear, or by-product material as defined by the Atomic Energy Act of 1954, as amended (68 Stat. 923) (42 U.S.C. 2011 et seq.). (For another definition, see CFR Part 243 & 257 below.)
Transuranic waste	Material contaminated with elements that have an atomic number greater than 92, including neptunium, plutonium, americium, and curium, and that are in concentrations greater than 10 nanocuries per gram, or in such other concentrations as the Nuclear Regulatory Commission may prescribe to protect the public health and safety.
High-level radioactive waste	(A) The highly radioactive material resulting from the reprocessing of spent nuclear fuel, including liquid waste produced directly in reprocessing and any solid material derived from such liquid waste that contains fission products in sufficient concentrations; and (B) other highly radioactive material that the Commission, consistent with existing law, determines by rule requires permanent isolation.
Low-level radioactive waste	Radioactive material that—(A) is not high-level radioactive waste, spent nuclear fuel, transuranic waste, or by-product material as defined in section 2014(e)(2) of this title; and (B) the Commission, consistent with existing law, classifies as low-level radioactive waste.

SOURCE: *U.S. Code, Title 42—The Public Health and Welfare*, Chapter 82, Solid Waste Disposal Subchapter I, General Provisions, §6903; Chapter 23, Development and Control of Atomic Energy, Division A, Atomic Energy Subchapter I, General Provisions, §2014; and Chapter 108, Nuclear Waste Policy, §10101.

Fig. 2, note the importance of vapor pressure. The volatile contaminants (that is, those with relatively high vapor pressure) can move upward from the plume, often reaching the atmosphere. Thus, hazardous wastes can contaminate the air, soil, aquifers, and surface waters. *See* GROUND-WATER HYDROLOGY; WATER POLLUTION.

Hazardous-waste characterization. The concentration and quantity determine whether a waste is hazardous. In most countries, these factors are codified to encourage proper handling of wastes (**Table 3**). In the United States, the amount of hazardous waste generated by a source determines the stringency of the rules. Any source that generates more than 1000 kg (2200 lb) of hazardous wastes per month is considered a large-quantity generator. A large-quantity generator is also one that has produced in any single month or has accumulated at any time 1 kg (2.2 lb) of acutely hazardous waste, or that generates or has accumulated at any time more than 100 kg (220 lb) of spill cleanup material contaminated with acutely hazardous waste. Depending on the hazard of the waste, a source that generates less than 100 kg of waste per month may be conditionally exempted. Other hazardous waste operations that are regulated include those that transport, as well as those that treat, store, and dispose of hazardous wastes.

Hazardous wastes often have been improperly disposed of and managed in the past, creating a need for response and remedial operations to clean up contaminated sites and operations. This has led to extensive environmental and public health threats with associated substantial legal and financial costs to the responsible parties. Cleanup continues to be a challenge to environmental engineers and scientists, but often the most prudent course of action is to eliminate the production of hazardous wastes in the design phase. Thus, pollution prevention, waste minimization, life-cycle analysis, and other "green engineering" programs are increasing. *See* HAZARDOUS-WASTE ENGINEERING. Daniel Vallero

Bibliography. H. M. Freeman (ed.), *Hazardous Waste Minimization*, 1990; H. M. Freeman (ed.), *Standard Handbook of Hazardous Waste Treatment and Disposal*, 2d ed., 1997; H. Hemond and E. Fechner-Levy, *Chemical Fate and Transport in the Environment*, 2000; D. M. Nielsen, *Practical Handbook of Ground-Water Monitoring*, 1991; D. A. Vallero, *Environmental Contaminants: Assessment and Control*, 2004.

Hazardous waste engineering

The control and management of hazardous wastes are truly among the most important challenges of our times. Environmental engineers play crucial roles in reducing the amount of hazardous substances produced, treating hazardous wastes to reduce their toxicity, and applying sound engineering controls to reduce or eliminate exposures to these wastes. Engineers design the facilities that generate the chemicals that, under the wrong circumstances, become hazards. The engineers also design and operate the containment and treatment facilities to deal with the wastes after they have been released, and are frequently called upon to address these wastes once they are in the environment.

A hazard is expressed as the potential of unacceptable outcome (**Table 1**). For chemicals, the most important hazard is the potential for disease or death, which is measured by epidemiologists as morbidity and mortality, respectively. The hazards to human health are referred to collectively in the medical and environmental sciences as toxicity. Toxicology is the study of these health outcomes and their potential causes. *See* EPIDEMIOLOGY; TOXICOLOGY.

In the United States, as in many countries, the term hazardous waste is defined by the federal government. Section 1004(5) of the Resource Conservation and Recovery Act (RCRA) defines a hazardous waste as a solid waste that may pose a substantial present or potential threat to human health and the environment when improperly treated, stored, transported, or otherwise managed.

Toxicity testing. The U.S. Environmental Protection Agency (EPA) has developed standard approaches and has set criteria to determine if substances exhibit hazardous characteristics. Because RCRA defines as hazardous a waste that presents a threat to human health and the environment when it is improperly managed, the government identified a set of assumptions that would allow for a waste to be disposed of if it is not subject to the controls mandated by Subtitle C of RCRA. (*See* HAZARDOUS WASTE.) This mismanagement scenario was designed to simulate a plausible worst case. Under a worst-case scenario, a potentially hazardous waste is assumed to be disposed along with municipal solid waste in a landfill with actively decomposing substances overlying an aquifer. When the government developed the mismanagement scenario, it recognized that not all wastes would be managed in this manner but that a dependable set of assumptions would be needed to ensure that the hazardous waste definition was implemented. The Hazardous and Solid Waste Amendments of 1984 (HSWA) established the Toxicity Characteristic Leaching Procedure (TCLP) to provide replicable results for contaminants commonly found in hazardous waste sites (**Table 2**). The specific medium used in the test is dictated by the alkalinity of the waste. The liquid extracted from the waste is analyzed for the 39 listed toxic constituents, and the concentration of each contaminant is compared to the TCLP standards specific to each contaminant.

Other hazards besides toxicity are important in hazardous waste engineering. The outcome may relate to environmental quality, such as an ecosystem stress, loss of important habitats, and decreases in the size of the population of sensitive species. Outcomes related to public and personal safety are also important. These may include a substance's potential to ignite, its corrosiveness, flammability, or explosiveness. A substance may be a public welfare hazard

TABLE 1. Properties of hazardous wastes

Designation	Description
H1	“Explosive”: subtances and preparations which may explode under the effect of flame or which are more sensitive to shocks or friction than dinitrobenzene.
H2	“Oxidising”: substances and preparations which exhibit highly exothemic reactions when in contact with other substances, particularly flammable substances.
H3A	“Highly flammable” liquid substances and preparations having a flashpoint of below 21°C (including extremely flammable liquids), or substances and preparations which may become hot and finally catch fire in contact with air at ambient temperature without any application of energy, or solid substances and preparations which may readily catch fire after brief contact with a source of ignition and which continue to burn or to be consumed after removal of the source of ignition, or gaseous substances and preparations which are flammable in air at normal pressure, or substances and preparations which, in contact with water or darnp air, evolve highly flammable gases in dangerous quantities.
H3B	“Flammable”: liquid substances and preparations having a flashpoint equal to or greater than 21°C and less than or equal to 55°C
H4	“Irritant”: noncorrosive substances and preparations which, through immediate, prolonged, or repeated contact with the skin or mucous membrane, can cause inflammation.
H5	“Harmful”: substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may involve limited health risks.
H6	“Toxic”: substances and preparations (including very toxic substances and preparations) which, if they are inhaled or ingested or if they penetrate the skin, may involve serious, acute, or chronic health risks and even death.
H7	“Carcinogenic”: substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may induce cancer or increase its incidence.
H8	“Corrosive”: substances and preparations which may destroy living tissue on contact.
H9	“Infectious”: substances containing viable microorganisms or their toxins which are known or reliably believed to cause disease in humans or other living organisms.
H10[2]	“Toxic for reproduction”: substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may produce or increase the incidence of nonheritable adverse effects in the progeny and/or of male or female reproductive functions or capacity.
H11	“Mutagenic”: substances and preparations which, if they are inhaled or ingested or if they penetrate the skin, may induce hereditary genetic defects or increase their incidence.
H12	Substances and preparations which release toxic or very toxic gases, in contact with water, air, or an acid.
H13	Substances and preparations capable by any means, after disposal, of yielding another substance, e.g., a leachate, which possesses any of the characteristics listed above.
H14	“Ecotoxic”: substances and preparations which present or may present immediate or delayed risks for one or more sectors of the environment.

SOURCE: United Kingdom Environment Agency, *Interpretation of the Definition and Classification of Hazardous Waste*, Technical Guidance WM2, Hazardous Waste Directive Annex III, 2003.

that damages property values or physical materials—for example, due to its corrosiveness or acidity. The hazard may be inherent to the substance. But more than likely, the hazard depends upon the situation and conditions where the exposure may occur. The substance is most hazardous when a number of conditions exist simultaneously, such as the hazard to firefighters using water in the presence of oxidizers.

Waste management hierarchy. There is growing acceptance throughout the world to use waste management hierarchies for solving hazardous waste problems. A typical sequence involves source reduction, recycling, treatment, and disposal. The waste generator is encouraged to begin at the top of the hierarchy and proceed down the hierarchy only if necessary. In the United States, the term pollution prevention is often used instead of the term source reduction, and waste minimization is used as an umbrella term for strategies and technologies that use source reduction and recycling techniques (**Fig. 1**).

Source reduction. Source reduction measures include process modifications, feedstock substitutions, improvements in feedstock purity, changes in housekeeping and management practice, increases in the efficiency of equipment, and recycling within a process. In the United States, there is a requirement in the federal environmental regulations that all large hazardous waste generators must have a program at their facility to encourage source reduction and recycling. Other countries, such as Austria, Germany, and Denmark, have initiated direct subsidies to encourage preferable waste management options.

Recycling. This is the use or reuse of hazardous waste as an effective substitute for a commercial product or as an ingredient or feedstock in an industrial process. It includes the reclamation of useful constituent fractions within a waste material or the removal of contaminants from a waste to allow it to be reused. *See* RECYCLING TECHNOLOGY.

Treatment. This refers to any method, technique, or process that changes the physical, chemical, or biological character of any hazardous waste to neutralize it; to recover energy or material resources from the waste; or to render the waste nonhazardous, less hazardous, safer to manage, amenable for recovery, amenable for storage, or reduced in volume.

Disposal. This is the discharge, deposit, injection, dumping, spilling, leaking, or placing of hazardous waste into or on any land or body of water so that the waste or any constituents may enter the air or be discharged into any waters, including ground water.

Treatment technologies. There are various alternative waste treatment technologies, including physical treatment, chemical treatment, biological treatment, incineration, and solidification or stabilization treatment. These processes are used to recycle and reuse waste materials, reduce the volume and toxicity of a waste stream, or produce a final

TABLE 2. Toxicity Characteristic Chemical Constituent Regulatory Levels for 39 Hazardous Chemicals [pursuant to Resource Conservation and Recovery Act of 1976 (RCRA) regulations, 40 CFR 261 and 262.11]

Contaminant	Regulatory level, mg L^{-1}
Arsenic	5.0
Barium	100.0
Cadmium	1.0
Chromium	5.0
Lead	5.0
Mercury	0.2
Selenium	1.0
Silver	5.0
Endrin	0.02
Lindane	0.4
Methoxychlor	10.0
Toxaphene	0.5
2,4-D	10.0
2,4,5 TP (Silvex)	1.0
Benzene	0.5
Carbon tetrachloride	0.5
Chlordane	0.03
Chlorobenzene	100.0
Chloroform	6.0
o-Cresol	200.0
m-Cresol	200.0
p-Cresol	200.0
Cresol	200.0
1,4-Dichlorobenzene	7.5
1,2-Dichloroethane	0.5
1,1-Dichloroethylene	0.7
2,4-Dinitrotoluene	0.13
Heptachlor (and its hydroxide)	0.008
Hexachloroethane	3.0
Hexachlorobutadiene	0.5
Hexachloroethane	3.0
Methyl ethyl ketone	200.0
Nitrobenzene	2.0
Pentachlorophenol	100.0
Pyridine	5.0
Tetrachloroethylene	0.7
Trichloroethylene	0.5
2,4,5-Trichlorophenol	400.0
2,4,6-Trichlorophenol	2.0
Vinyl chloride	0.2

residual material that is suitable for disposal. The selection of the most effective technology depends upon the wastes being treated.

The characteristics of the media in need of treatment determine the performance of any contaminant treatment or control. For example, sediment, sludge, slurries, and soil characteristics that will influence the efficacy of treatment technologies include particle size, solids content, and high contaminant concentration (**Table 3**).

Particle size is an important limiting characteristic for applying treatment technologies to sediments. Most treatment technologies work well on sandy soils and sediments. The presence of fine-grained material reduces the effectiveness of treatment system emission controls because it increases particulate generation during thermal drying, it is more difficult to dewater, and it has greater attraction to the contaminants (especially fine-grained clays). Clayey sediments that are cohesive also present materials-handling problems in most processing systems.

Solids content generally ranges from high (30–60% solids by weight) to low (10–30% solids by weight). Treatment of slurries is better at lower solids contents, but this can be achieved for high solids contents by adding water at the time of processing. It is more difficult to change a lower to a higher solids content, but evaporative and dewatering approaches, such as those used for municipal sludges, may be used. Also, thermal and dehalogenation processes are decreasingly efficient as the solids content is reduced. More water means increased chemical costs and an increased need for wastewater treatment.

Elevated levels of organic compounds or heavy metals in high concentrations must also be considered. Higher total organic carbon (TOC) content favors incineration and oxidation processes. The TOC can be the contaminant of concern or any organic, since they are combustibles with caloric value. Conversely, higher metal concentrations may make a technology less favorable by increasing contaminant mobility of certain metal species following processing.

A number of other factors may affect the selection of a treatment technology other than its effectiveness for treatment (**Table 4**). For example, vitrification and supercritical water oxidation have been used only for relatively small projects and have not been proven for use in full-scale sediment projects. Regulatory compliance and community perception are always a part of decisions regarding an incineration system. Land use considerations, such as the amount of acreage needed, are commonly confronted in solidification and solid-phase bioremediation projects, as well as in sludge farming and land application. Disposing of ash and other residues following treatment must be part of any process. Treating water effluent and air emissions must be part of the decontamination decision-making process.

Physical treatment. This includes processes that separate components of a waste stream or change the physical form of the waste without altering the chemical structure of the constituent materials. Physical treatment techniques are often used to separate the materials within the waste stream so that they can be reused or detoxified by chemical or biological treatment or destroyed by high-temperature incineration. These processes are very useful for separating hazardous materials from an otherwise nonhazardous waste stream so that the materials may be treated in a more concentrated form, separating various hazardous components for different treatment processes, and preparing a waste stream for ultimate destruction in a biological or thermal treatment process.

Physical treatment processes are important to most integrated waste treatment systems regardless of the nature of the waste materials or the ultimate technologies used for treatment or destruction. The physical processes that are commonly used in waste treatment operations include screening, sedimentation, flotation, filtration, centrifugation, dialysis, membrane separations, ultrafiltration, distillation, solvent extraction, evaporation, and adsorption. *See* ACTIVATED CARBON; ADSORPTION; CENTRIFUGATION; DIALYSIS; DISTILLATION; EVAPORATION; FILTRATION; FLOTATION; MEMBRANE SEPARATIONS;

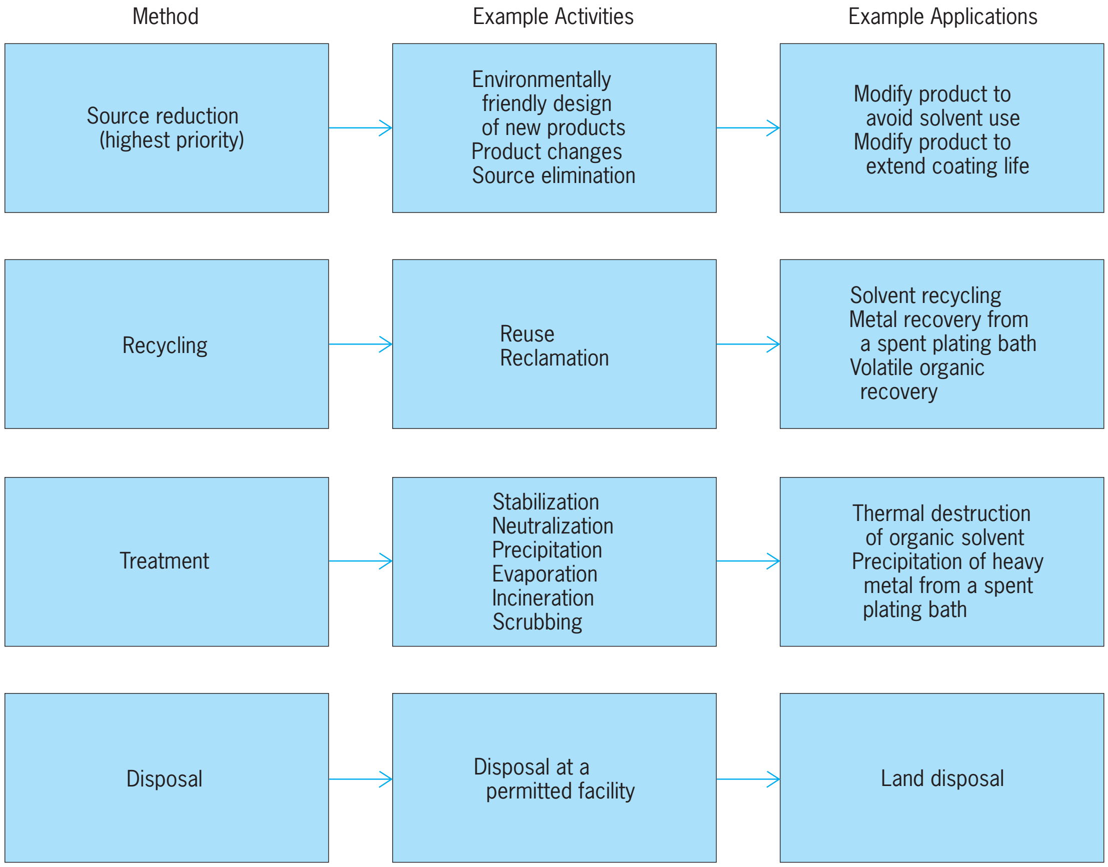

Method
Example Activities
Example Applications
Source reduction (highest priority)
Environmentally friendly design of new products
Product changes
Source elimination
Modify product to avoid solvent use
Modify product to extend coating life
Recycling
Reuse
Reclamation
Solvent recycling
Metal recovery from a spent plating bath
Volatile organic recovery
Treatment
Stabilization
Neutralization
Precipitation
Evaporation
Incineration
Scrubbing
Thermal destruction of organic solvent
Precipitation of heavy metal from a spent plating bath
Disposal
Disposal at a permitted facility
Land disposal

Fig. 1. Flow chart for management of waste: reduction from manufacture, use, treatment, and disposal. (*After U.S Environmental Protection Agency, Facility Pollution Prevention Guide, EPA/600/R-92/088, 1992*)

SCREENING; SEDIMENTATION (INDUSTRY); SOLVENT EXTRACTION; ULTRAFILTRATION.

Chemical treatment. Chemical treatment processes alter the chemical structure of the constituents of the waste to produce either an innocuous or a less hazardous material by taking advantage of chemical reactions. This can be said of many biological processes as well, so chemical treatment processes are understood to be abiotic (that is, chemical reactions anywhere outside of a living organism). Four

TABLE 3. Effect of particle size, solids content, and extent of contamination on decontamination efficiencies*

	Predominant particle size			Solids content		High contaminant concentration	
Treatment technology	Sand	Silt	Clay	High (slurry)	Low (in situ)	Organic compounds	Metals
Conventional incineration	N	X	X	F	X	F	X
Innovative incineration	N	X	X	F	X	F	F
Pyrolysis	N	N	N	F	X	F	F
Vitrification	F	X	X	F	X	F	F
Supercritical water oxidation	X	F	F	X	F	F	X
Wet air oxidation	X	F	F	X	F	F	X
Thermal desorption	F	X	X	F	X	F	N
Immobilization	F	X	X	F	X	X	N
Solvent extraction	F	F	X	F	X	X	N
Soil washing	F	F	X	N	F	N	N
Dechlorination	U	U	U	F	X	X	N
Oxidation	F	X	X	N	F	X	X
Bioslurry process	N	F	N	N	F	X	X
Compositing	F	N	X	F	X	F	X
Contained treatment facility	F	N	X	F	X	X	X

*F = sediment characteristic is favorable to the effectiveness of the process.
N = sediment characteristic has no significant effect on process performance.
U = effect of sediment characteristic on process is unknown.
X = sediment characteristic may impede process performance or increase cost.
SOURCE: U.S. Environmental Protection Agency, *Remediation Guidance Document*, EPA-905-B94-003, Chap. 7, 2003.

TABLE 4. Factors for selecting decontamination and treatment approaches*

Treatment technology	Implementability at full scale	Regulatory compliance	Community acceptance	Land requirements	Residuals disposal	Wastewater treatment	Air emission control
Conventional incineration		●	●				●
Innovative incineration		●	●				●
Pyrolysis		●					●
Vitrification	●	●					●
Supercritical water oxidation	●						
Wet air oxidation							
Thermal desorption					●	●	●
Immobilization				●			●
Solvent extraction					●	●	
Soil washing					●	●	
Dechlorination							●
Oxidation	●						
Bioslurry process	●						●
Composting				●			●
Contained treatment facility				●		●	●

*● indicates that the factor is critical in the evaluation of the technology.
SOURCE: U.S. Environmental Protection Agency, *Remediation Guidance Document*, EPA-905-B94-003, Chap. 7, 2003.

categories of chemical reactions are available to treat hazardous wastes: synthesis or combination, decomposition, single replacement, and double replacement.

Synthesis or combination. In combination reactions (1),

$$A + B \rightarrow AB \qquad (1)$$

two or more substances react to form a single substance. Two types of combination reactions are important in environmental systems: formation and hydration. Formation reactions are those where elements combine to form a compound. An example is the formation reaction (2) of ferric oxide.

$$4Fe\,(s) + 3O_2\,(g) \rightarrow 2Fe_2O_3\,(s) \qquad (2)$$

Hydration reactions involve the addition of water to synthesize a new compound, for example, in reaction (3) when calcium oxide is hydrated to form calcium hydroxide, and reaction (4) when phosphate is hydrated to form phosphoric acid.

$$CaO\,(s) + H_2O\,(l) \rightarrow Ca(OH)_2\,(s) \qquad (3)$$

$$P_2O_5\,(s) + 3H_2O\,(l) \rightarrow 2H_3PO_4\,(aq) \qquad (4)$$

Decomposition. Decomposition is often referred to as degradation when discussing organic compounds in toxicology, environmental sciences, and engineering. In decomposition reactions (5),

$$AB \rightarrow A + B \qquad (5)$$

one substance breaks down into two or more new substances. For example, in reaction (6) calcium car-

$$CaCO_3\,(s) \rightarrow CaO\,(s) + CO_2\,(g) \qquad (6)$$

bonate breaks down into calcium oxide and carbon dioxide.

Single replacement (single displacement). This reaction (7)

$$A + BC \rightarrow AC + B \qquad (7)$$

commonly occurs when one metal ion in a compound is replaced with another metal ion, such as in reaction (8) when trivalent chromium replaces monovalent silver.

$$3AgNO_3\,(aq) + Cr\,(s) \rightarrow Cr(NO_3)_3\,(aq) + 3Ag\,(s) \qquad (8)$$

Double replacement (metathesis or double displacement). In metathesis reactions (9), cations and anions trade

$$AB + CD \rightarrow AD + CB \qquad (9)$$

places. They are commonly encountered in metal precipitation reactions, such as when lead is precipitated, as in reaction (10) of a lead salt with an acid

$$Pb(ClO_3)_2\,(aq) + 2KCl\,(aq) \rightarrow PbCl_2\,(s) + 2KClO_3\,(aq) \qquad (10)$$

such as potassium chloride.

Chemical processes are attractive because they produce minimal air emissions, they can often be carried out on the site of the waste generator, and some processes can be designed and constructed as mobile units. A few commonly applied chemical treatment processes are neutralization, precipitation (for example, waste steams containing heavy metals), ion exchange, dehalogenation (especially dechlorination and debromination), and oxidation-reduction for detoxifying toxic wastes. *See* ACID AND BASE; ION EXCHANGE; OXIDATION-REDUCTION; PRECIPITATION (CHEMISTRY).

Biological treatment. Biological waste-treatment processes are those that use microorganisms to decompose organic wastes into water, carbon dioxide, and simple inorganic substances, or into simpler organic substances such as aldehydes and acids. Contaminants, if completely organic in structure, are in theory completely destructible using microorganisms, with the engineering inputs and outputs summarized in reaction (11).

$$\text{Hydrocarbons} + O_2 + \text{microorganisms } (+\text{energy}) \rightarrow CO_2 + H_2O + \text{microorganisms } (+\text{energy?}) \qquad (11)$$

Contaminant wastes are mixed with oxygen and aerobic microorganisms, sometimes in the presence of an external energy source in the form of added nutrition for the microorganisms. In time, the by-products of gaseous carbon dioxide and water are produced and exit the top of the reaction vessel, while a solid mass of microorganisms is produced to exit the bottom.

If the waste contains other chemical constituents, in particular chlorine and/or heavy metals, and if the microorganisms are able to withstand and flourish in such an environment, the simple input and output relationship is modified to reaction (12).

$$\begin{aligned}&\text{Hydrocarbons} + O_2 + \text{microorganisms } (+\text{energy?}) \\ &\quad + \text{Cl or heavy metal(s)} + H_2O + \text{inorganic salts} \\ &\quad + \text{nitrogen compounds} + \text{sulfur compounds} \\ &\quad + \text{phosphorus compounds} \rightarrow CO_2 + H_2O\ (+\text{energy?}) \\ &\quad + \text{chlorinated hydrocarbons or heavy-metal(s)} \\ &\qquad \text{inorganic salts} + \text{nitrogen compounds} \\ &\quad + \text{sulfur compounds} + \text{phosphorus compounds} \qquad (12)\end{aligned}$$

If the microorganisms do survive in this complicated environment, the potential exists for the transformation to a potentially more toxic molecule that contains chlorinated hydrocarbons, higher heavy-metal concentrations, as well as more mobile or more toxic chemical species of heavy metals.

All bioreactor systems share common attributes. All rely on a population of microorganisms to metabolize organic contaminants, ideally into the harmless by-products of $CO_2 + H_2O$ (+energy?). In all the systems, the microorganisms either must be initially cultured in the laboratory to be able to metabolize a specific organic waste, or must be given sufficient time to evolve to be able to digest the contaminant. The engineer must plan for and undertake extensive and continual monitoring and fine-tuning of each microbiological processing system during its operation. *See* BIOCHEMICAL ENGINEERING.

The advantages of biological treatment systems include (1) the potential for energy recovery; (2) volume reduction of the hazardous waste; (3) detoxification as selected molecules are reformulated; (4) basic scientific principles, engineering designs, and technologies that are well understood from a wide range of other applications, including municipal wastewater treatment at facilities; (5) application to most organic contaminants which as a group compose a large percentage of the total hazardous waste generated; and (6) possibility to scale the technologies to handle a single gallon/pound (liter/kilogram) or millions of gallons/pounds (liters/kilograms) of waste per day.

The disadvantages of the biotreatment systems include (1) operation of the equipment requires very skilled operators and is more costly as input contaminant characteristics change over time and correctional controls become necessary; and (2) ultimate disposal of the waste microorganisms is necessary and particularly troublesome and costly if heavy metals and/or chlorinated compounds are found during the monitoring activities. In general, hazardous waste containing heavy metals should not be bioprocessed without specific additional steps.

Variations on three different types of bioprocessors are available to treat hazardous wastes: trickling filter, activated sludge, and aeration lagoons.

Trickling filter. The classical design of a trickling filter system includes a bed of fist-sized rocks enclosed in a rectangular or cylindrical structure through which is passed the waste (**Fig. 2**). Biofilms are selected from laboratory studies and encouraged to grow on the rocks. As the liquid waste moves downward through the bed, the microorganisms come in contact with the organic contaminant/food source and ideally metabolize the waste into $CO_2 + H_2O +$ microorganisms (+energy?). Oxygen is supplied by blowers from the bottom of the reactor and passed upward through the bed. The treated waste moves downward through the bed and subsequently enters a quiescent tank where the microorganisms that are sloughed off the rocks are settled, collected, and disposed. Trickling filters actually are considered mixed treatment systems because aerobic bacteria grow in the upper, higher-oxygen layers of the

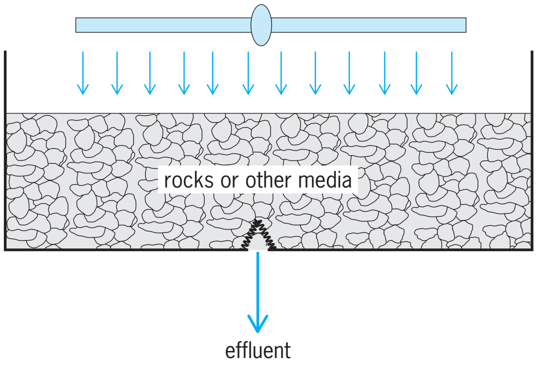


Fig. 2. Trickling filter treatment system. (*After D. Vallero, Engineering the Risks of Hazardous Wastes, Butterworth-Heinemann, 2003*)

media, while anaerobes grow in the regions lower in the system. *See* BIOFILM; INDUSTRIAL WASTEWATER TREATMENT.

Activated sludge. The key to the activated sludge system is that the microorganisms are recycled within the system (**Fig. 3**). This reuse enables the microorganisms to evolve over time and adapt to the changing characteristics of the hazardous waste.

In an activated-sludge system, a tank of liquid hazardous waste is injected with a mass of microorganisms. Oxygen is supplied in the aeration basin as the microorganisms come in contact, sorb, and metabolize the waste, ideally into $CO_2 + H_2O +$ microorganisms (+energy?). The heavy (fed) microorganisms then flow into a quiescent tank, where they settle by gravity and are collected and disposed. Depending on the operating conditions of the facility, some or many of the settled (now hungry) and active microorganisms are returned to the aeration basin, where they feed again. Liquid effluent from the activated-sludge system may require additional microbiological and/or chemical processing prior to release into a receiving stream or sewer system.

Aeration lagoons. These ponds treat liquid and dissolved contaminants for long terms, from months to years (**Fig. 4**). Persistent organic molecules, those not readily degraded in trickling filter or activated sludge systems, are potentially broken down by certain microbes into $CO_2 + H_2O +$ microorganisms (+energy?) if given enough time. The ponds are open to the weather, and ideally oxygen is supplied directly to the microorganisms from the atmosphere.

Since biological treatment systems do not alter or destroy inorganic substances, and high concentrations of such materials can severely inhibit decomposition activity, chemical or physical treatment may be required to extract inorganic materials from a waste stream prior to biological treatment. *See* SEWAGE TREATMENT.

Thermal treatment. If completely organic in structure, hazardous wastes are completely destructible using principles based in thermodynamics with the engineering inputs and outputs as in reactions (13) and (14).

$$\text{Hydrocarbons} + O_2\ (+\text{energy?}) \rightarrow CO_2 + H_2O\ (+\ \text{energy?}) \qquad (13)$$

$$(CH)_x + O_2 \rightarrow CO_2 + H_2O \qquad (14)$$

Complete combustion may also result in the production of molecular nitrogen (N_2) when nitrogen-containing organics are burned, such as in the combustion of methylamine. Incomplete combustion can produce a variety of compounds. Some are more toxic than the original compounds being oxidized, such as polycyclic aromatic hydrocarbons (PAHs), dioxins, furans, and carbon monoxide. *See* COMBUSTION.

Wastes are mixed with oxygen, sometimes in the presence of an external energy source, and within several seconds the by-products of gaseous carbon dioxide and water are produced and exit the top of the reaction vessel, while a solid ash is produced that exits the bottom of the reaction vessel. Energy may also be produced during the reaction, and the heat may be recovered.

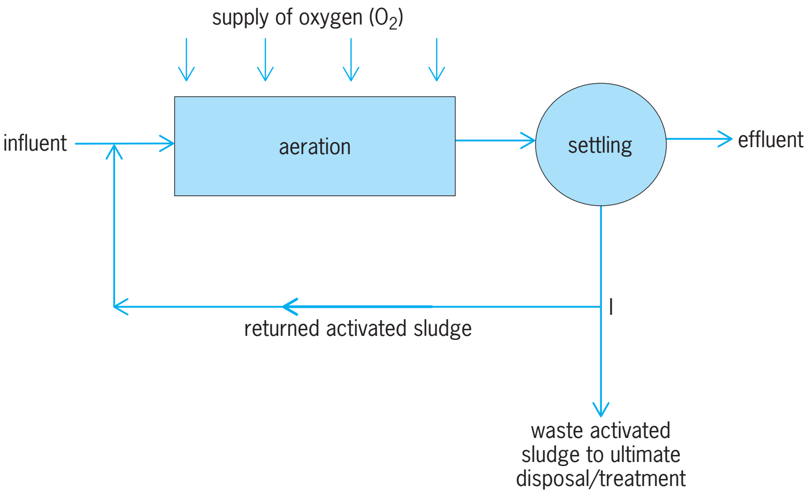

supply of oxygen (O_2)
influent
aeration
settling
effluent
returned activated sludge
waste activated sludge to ultimate disposal/treatment

Fig. 3. Activated sludge treatment system. (*After D. Vallero, Engineering the Risks of Hazardous Wastes, Butterworth-Heinemann, 2003*)

If the wastes contain other chemical constituents, particularly chlorine and/or heavy metals, the situation becomes complex. For such wastes, the potential exists for destroying the initial contaminant as well as exacerbating the problem by generating hazardous off-gases containing chlorinated hydrocarbons or ashes containing heavy metals. For example, the improper incineration of certain chlorinated hydrocarbons can lead to the formation of the highly toxic chlorinated dioxins, furans, and hexachlorobenzene.

The advantages of thermal systems include (1) the potential for energy recovery; (2) volume reduction of the contaminant; (3) detoxification as selected molecules are reformulated; (4) basic scientific principles, engineering designs, and technologies that are well understood from a wide range of other applications, including electric generation and municipal solid-waste incineration; (5) application to most organic contaminants which compose a large percentage of the total contaminants generated worldwide; (6) possibility to scale the technologies to handle a single gallon per pound (liter per kilogram) of waste or millions of gallons per pound (liter per kilogram) of waste: and (7) small land area requirements relative to other hazardous waste management facilities such as landfills.

In general, the same reaction applies to most thermal processes of gasification, pyrolysis, hydrolysis, and combustion. The actual reactions from test burns

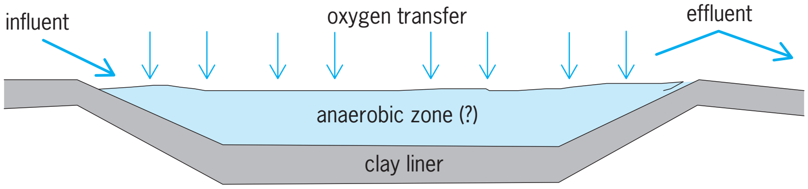


Fig. 4. Aeration pond. (*After D. Vallero, Engineering the Risks of Hazardous Wastes, Butterworth-Heinemann, 2003*)

for commonly incinerated compounds are provided in **Table 5**.

Of all of the thermal processes, incineration is the most common process for destroying organic contaminants in industrial wastes. Incineration simply is heating wastes in the presence of oxygen to oxidize organic compounds. There are a number of stages in the incineration process where new compounds may need to be addressed. Ash and other residues may contain high levels of metals, at least higher than the original feed. The flue gases are likely to include both organic and inorganic compounds resulting from temperature-induced volatilization or newly transformed products of incomplete combustion with higher vapor pressures than the original contaminants.

The disadvantages of hazardous waste incinerators include: (1) the equipment is capital-intensive, particularly the refractory material lining the inside walls of each combustion chamber which must be replaced as cracks form whenever a combustion system is cooled and heated; (2) the equipment requires very skilled operators and is more costly to operate when fuel must be added to the system; (3) the disposal of the ash is necessary and particularly troublesome and costly if heavy metals and/or chlorinated compounds are present; and, (4) the air emissions may be hazardous and thus must be monitored for chemical constituents and must be controlled, including the release of heavy metals and their compounds.

Pyrolysis is the process of chemical decomposition induced in organic materials by heat in the absence of oxygen. It is impossible to achieve a completely oxygen-free atmosphere, so pyrolytic systems run with less than stoichiometric quantities of oxygen. During pyrolysis, organic compounds are converted to gaseous components (such as CO, H_2, CH_4, and other hydrocarbons), along with some liquids, coke (the solid residue of fixed carbon), and ash. Pyrolysis generally takes place well above atmospheric pressure at temperatures exceeding 430°C (800°F). The secondary gases need their own treatment, such as by a secondary combustion chamber. Particulates must be removed by additional air pollution controls, such as fabric filters or cyclones.

The target contaminant groups for pyrolysis include semivolatile organic compounds, including pesticides, polychlorinated biphenyls (PCBs), dioxins, and PAHs. Pyrolysis allows for separating organic contaminants from various wastes, including those from refineries; coal tar; wood-preservative, creosote-contaminated, and hydrocarbon-contaminated soils; mixed radioactive and hazardous wastes; synthetic rubber processing; and paint and coating processes. Pyrolysis systems may be used to treat a variety of organic contaminants that chemically decompose when heated (cracking). Pyrolysis is not effective in either destroying or physically separating inorganic compounds that coexist with the organics in the contaminated medium. Volatile metals may be removed and transformed, but their mass balance will not change. *See* PYROLYSIS.

TABLE 5. Balanced combustion reactions for selected organic compounds

Compound	Combustion reaction
Chlorobenzene	$C_6H_5Cl + 7O_2 \rightarrow 6CO_2 + HCl + 2H_2O$
TCE	$C_2Cl_4 + O_2 + 2H_2O \rightarrow 2CO_2 + HCl$
HCE	$C_2Cl_6 + {}^1/_2O_2 + 3H_2O \rightarrow 2CO_2 + 6HCl$
CPVC	$C_4H_5Cl_3 + 4{}^1/_2O_2 \rightarrow 4CO_2 + 3HCl + H_2O$
Natural gas fuel (methane)	$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$
PTFE Teflon	$C_2F_4 + O_2 \rightarrow CO_2 + 4HF$
Butyl rubber	$C_9H_{16} + 13O_2 \rightarrow 9CO_2 + 8H_2O$
Polyethylene	$C_2H_4 + 3O_2 \rightarrow 2CO_2 + 2H_2O$
Wood*	$C + O_2 \rightarrow CO_2$ $H + 0.25O_2 \rightarrow 0.5H_2O$

*Wood is considered to have the composition $C_{6.9}H_{10.6}O_{3.5}$. Therefore the combustion reactions are simple carbon and hydrogen combustion.

SOURCE: U.S. Environmental Protection Agency.

Emerging thermal technologies. New thermal processes include high-pressure oxidation and vitrification. High-pressure oxidation combines two related technologies—wet air oxidation and supercritical water oxidation—which combine high temperature and pressure to destroy organics. Wet air oxidation has generally been limited to conditioning of municipal wastewater sludges, but can degrade hydrocarbons (including PAHs), certain pesticides, phenolic compounds, cyanides, and other organic compounds.

Vitrification uses electricity to heat and destroy organic compounds and immobilize inert contaminants. The product is a solid, glasslike material that is very resistant to leaching.

Solidification and stabilization. Solidification and stabilization are treatment systems designed to improve handling and the physical characteristic of the waste, decrease the surface area across which transfer or loss of contained pollutants can occur, and limit the solubility of, or detoxify, any hazardous constituents contained in the wastes. In solidification, these results are obtained primarily by producing a monolithic block of treated waste with high structural integrity. Stabilization techniques limit the solubility or detoxify waste contaminants even though the physical characteristics of the waste may not be changed. Stabilization usually involves the addition of materials that ensure that the hazardous constituents are maintained in their least soluble or least toxic form.

Disposal. After treatment is completed, an inorganic valueless residue remains that must be disposed of safely. There are five options for disposing of hazardous waste: (1) Underground injection wells are steel- and concrete-encased shafts placed deep below the surface of the earth into which hazardous wastes are deposited by force and under pressure. Some liquid waste streams are commonly disposed of in underground injection wells. (2) Surface impoundment involves natural or engineered depressions or diked areas that can be used to treat, store, or dispose of hazardous waste. Surface impoundments are often referred to as pits, ponds, lagoons, and basins. (3) Landfills are disposal facilities where hazardous waste is placed in or on land. Properly designed and operated landfills are lined to prevent leakage, and contain systems to collect potentially contaminated surface water runoff. Most landfills isolate wastes in discrete cells or trenches, thereby preventing potential contact of incompatible wastes. (4) Land treatment is a disposal process in which hazardous waste is applied onto or incorporated into the soil surface. Natural microbes in the soil break down or immobilize the hazardous constituents. Land treatment facilities are also known as land-application or land-farming facilities. (5) Waste piles are noncontainerized accumulations of solid, nonflowing hazardous waste. While some are used for final disposal, many waste piles are used for temporary storage until the waste is transferred to its final disposal site.

Of the hazardous waste disposed of on land in the United States, nearly 60% is disposed of in underground injection wells, approximately 35% in surface impoundments, 5% in landfills, and less than 1% in waste piles or by land application.

Comprehensive treatment facilities. In the United States, Canada, and many western European countries, there are large facilities capable of treating and disposing of many types of hazardous wastes. These facilities are able to realize economies of scale by incorporating many treatment processes that might not be economical for individual generators. Comprehensive facilities are also able to exploit the synergistic opportunities made possible by having many different types of waste present at a single site. These include using waste acids and alkalies to neutralize one another, waste oxidants to treat cyanides and organic contaminants in water, salts and acids to salt out organic compounds from wastewater, onsite incinerators to dispose of organic vapors generated by other onsite processes, ash and calcium and magnesium oxides to aid in stabilization processes, and combustible solids and liquids to produce blended liquid fuels. Although these technical advantages are compelling, the siting of these facilities is often resisted by neighbors and is often opposed politically. This opposition is known as NIMBY, “not in my back yard.”

Abandoned disposal sites. There are sites in many countries where hazardous waste has been disposed of improperly in the past and where cleanup operations are needed to restore the sites to their original state. In the United States, about 20,000 of these sites have been identified, and approximately 2000 sites require immediate action. Denmark, the Netherlands, and Sweden have also identified sites requiring immediate attention. Environmental regulations lay out explicit steps for cleaning up contamination (**Fig. 5**). The first step of a contaminant cleanup is a preliminary assessment (PA). During the PA of a site, readily available information about a site and its surrounding area are collected to determine the threat to human health and the environment. Any possible emergency response actions may also be identified. If information beyond the preliminary data in the PA is needed, a site inspection is performed.

Cleaning up abandoned disposal sites involves isolating and containing contaminated material, removing and redepositing contaminated sediments, and in-situ and direct treatment of the hazardous wastes involved.

Ex-situ and in-situ treatment. Contaminated soil and sediment must often be removed and treated

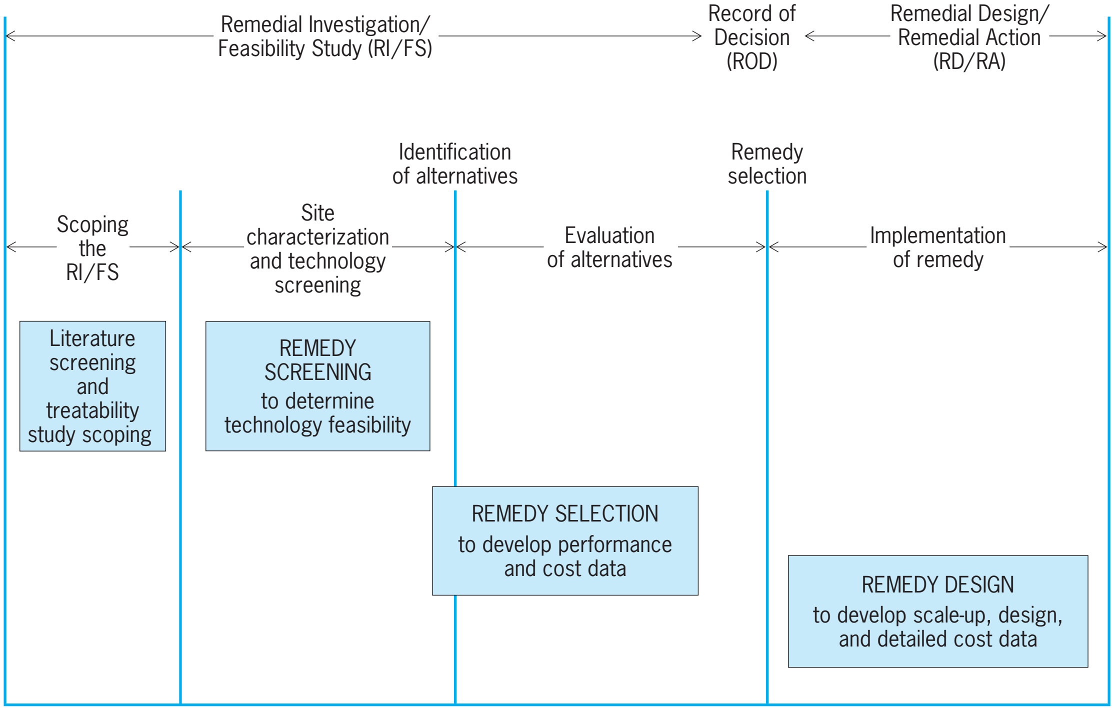


Fig. 5. Steps needed in the cleanup of an abandoned hazardous waste site. (*U.S. Environmental Protection Agency, Guide for Conducting Treatability Studies under CERCLA: Thermal Desorption, EPA/540/R-92/074 B, 1992*)

off-site (ex situ). Contaminated soil may be excavated and transported to kilns or other high-temperature operations, where it is mixed with combustible material. High-sand-content soils may be used as part of an asphalt mix. Contaminated soil may also be distributed onto an impermeable surface, allowing the more volatile compounds to evaporate. Microbial biodegradation can be accelerated and enhanced by adding nutrients and moisture to the soil. A faster process, thermal desorption, entails heating the soil to evaporate the contaminants and capture the compounds, and then burning them in a vapor treatment device.

Ground water generally is treated by drilling recovery wells to pump contaminated water to the surface. Commonly used ground-water treatment approaches include air stripping, filtering with granulated activated carbon (GAC), and air sparging. Air stripping transfers volatile compounds from water to air. Ground water is allowed to drip downward in a tower filled with a permeable material through which a stream of air flows upward. Another method bubbles pressurized air through contaminated water in a tank. Filtering ground water with GAC entails pumping the water through the GAC to trap the contaminants. In air sparging, air is pumped into the ground water to aerate the water. Most often, a soil venting system is combined with an air sparging system for vapor extraction.

Hazardous wastes treated where they are found, without first removing them, is known as in-situ remediation. Bioremediation makes use of living microorganisms to break down toxic chemicals or to render them less hazardous. This is often done by using bacteria, and in some instances algae and fungi, that are already living in the soil, sediment, or water. These microbes are exposed to increasing amounts of the chemical, so that they adapt to using the chemical as an energy (food) source. This process is known as acclimation. The acclimated microbes can then be taken from the laboratory and applied to the waste in the field. The most passive form of bioremediation is natural attenuation, where no direct engineering intervention is used, and the contaminants are allowed to be degraded by resident microbes over time. The only role for the engineer is to monitor the soil and ground water to measure the rate at which the chemicals are degrading. Natural attenuation can work well for compounds that are found in the laboratory to break down under the conditions found at the site. For example, if a compound is degraded under reduced, low-pH conditions in the laboratory, it may also degrade readily in soils with these same conditions (for example, in deeper soil layers where bacteria have adapted to these conditions naturally). Natural attenuation can sometimes be enhanced, for example, when soil moisture levels are increased.

Plant life may also be used to reduce the amount of hazardous wastes. In phytoremediation, contaminated areas are seeded, and plant roots extract the chemicals from the soil. The harvested plants are either treated on-site or transferred to a treatment facility. In reality, both microbial and macrophytic processes occur simultaneously. Fast-growing trees (for example, poplars), can help to treat areas contaminated with agricultural chemicals. Plants, such as grasses and field crops, have even been used to treat the very persistent polychlorinated biphenyls (PCBs), wood preservatives, and petroleum. Plants have also been used to extract heavy metals and radioactive substances from contaminated soil. Bioremediation has been used successfully to treat numerous other organic and inorganic compounds.

Various methods are available for treating substances after they have been released into the environment. However, eliminating the wastes before they are released is the best means of reducing hazards and the associated risks, especially by applying the principles of green chemistry and sustainable design. *See* GREEN CHEMISTRY. Daniel Vallero

Bibliography. Federal Remediation Technologies Roundtable, *Remediation Technologies Screening Matrix and Reference Guide*, 4th ed., 2002; H. M. Freeman (ed.), *Hazardous Waste Minimization*, 1990; H. M. Freeman (ed.), *Standard Handbook of Hazardous Waste Treatment and Disposal*, 2d ed., 1997; A. Gaudy and E. Gaudy, *Elements of Bioenvironmental Engineering*, 1988; C. N. Haas and R. J. Ramos, *Hazardous and Industrial Waste Treatment*, 1995; Metcalf and Eddy, Inc., by G. Tchobanoglous, F. L. Burton, and H. D. Stensel, *Wastewater Engineering*, 4th ed., 2002; J. J. Peirce, R. Weiner, and P. A. Vesilind, *Environmental Pollution and Control*, 4th ed., 1998; M. N. Sara, Groundwater monitoring system design, in D. M. Nielsen (ed.), *Practical Handbook of Ground-Water Monitoring*, Lewis Publishers, Chelsea, MI, 1991; H. Tammenagi, *The Waste Crisis: Landfills, Incinerators, and the Search for a Sustainable Future*, 1999; U.S. Environmental Protection Agency, *Remediation Guidance Document*, EPA-905-B94-003, Chap. 7, 2003; U.S. Environmental Protection Agency, *Test Methods for Evaluating Solid Waste*, vols. I and II (SW-846), 3d ed, 1986; D. A. Vallero, *Engineering the Risks of Hazardous Wastes*, 2003; D. A. Vallero, *Environmental Contaminants: Assessment and Control*, 2004; C. Wentz, *Hazardous Waste Management*, 1989; D. G. Wilson, History of Solid Waste Management, in *Handbook of Solid Waste Management*, 1997.

Head

The region of the body consisting of the skull, its contents, and related structures. Its two principal parts are the cranium, or braincase, and the face.

The skin, hair, and subcutaneous tissues over the top of the skull are collectively known as the scalp. The regions of the cranium take their names from the underlying bones, for example, the temporal, parietal, frontal, and occipital regions.

The intracranial contents include the brain and uppermost portion of the spinal cord with their coverings (meninges), blood vessels, and the important cranial nerves, as well as the cerebrospinal fluid system. Many openings, or foramina, afford means of passage from within the skull for nerves and blood vessels. *See* BRAIN.

The face contains the mouth and jaws, the nose and eyes (the ears are in the cranial region), and many related structures. Muscles acting on the lower jaw and on the skin (facial muscles, or muscles of expression) cover much of the face and part of the cranium. Thomas S. Parsons

Headache

Pain within the head. It is probably the most common complaint for which people seek a physician's help. Headaches can be grouped into three primary categories: vascular, muscle-contraction, and organic.

Vascular headaches. Vascular headaches include classic and common migraine as well as cluster, toxic, and hypertensive headaches. All are caused by dilation of cerebral blood vessels. Constriction of the blood vessels may also occur in any part of the cerebral vasculature and cause the neurologic symptoms associated with some forms of vascular headache.

Migraine. Migraine is a French word derived from the Greek *hemicrania*, which means half a head. An estimated 14 to 16 million people in the United States suffer from migraine.

Migraine affects one side of the head but may be bilateral. This may occur several times each month, although some individuals experience attacks only once or twice yearly. Neurologic symptoms, especially visual disturbances, are common. The associated symptoms of migraine include nausea, vomiting, loss of appetite, and irritability. During the migraine attack, some patients complain of sensitivity to light and sound. In classic migraine, the attacks may be preceded for several minutes to an hour by visual disturbances such as blind spots, flashing lights, and decreased visual field, or other neurologic symptoms such as numbness, speech disorders, or perception of strange odors.

Migraine tends to be a familial disorder. Migraine usually begins in young adulthood but can appear early in childhood; it rarely first occurs after age 40. An estimated 60% of migraine sufferers are women, and often the first migraine coincides with menarche. The onset of an acute migraine attack has been linked to such factors as diet, stress, fatigue, menstruation, hormonal changes, hypoglycemia, and oversleeping. A small number of persons can be treated by the elimination of certain foods, such as chocolate, aged cheese, fermented foods, and alcoholic beverages from the diet. A regular schedule of meal times, adequate sleep, and a consistent time for awakening have also been shown to be beneficial. *See* ALLERGY.

After an attack has begun, the drugs of choice for treatment are the ergotamines. If used early in an attack, they may prevent its progression. For those in whom ergotamine is contraindicated or not tolerated, isometheptene mucate has been effective in the abortive treatment of migraine, as have the nonsteroidal anti-inflammatory agents such as naproxen sodium. *See* ERGOT AND ERGOTISM.

Prophylactic therapy should be considered when attacks occur more often than twice a month. The agent of choice in migraine prophylaxis is the beta blocker, propranolol. Other beta blockers have been used successfully in migraine treatment, but only propranolol has been approved for this therapeutic indication by the U.S. Food and Drug Administration.

Antidepressants and the nonsteroidal anti-inflammatory drugs have been used successfully in treating migraine, and calcium channel blockers are another potentially effective remedy. Serotonin, a substance found in blood platelets, has been implicated in blood vessel changes, and serotonin blockers may offer additional treatment options. *See* SEROTONIN.

Many physicians recommend nonpharmaceutical treatment such as biofeedback, a method by which the patient learns to control a physiological function that was previously considered involuntary. By using self-hypnotic phrases and progressive relaxation exercises, as well as instrumentation, the patient may learn to relax the muscles in the upper part of the body and to increase finger temperature, which have been shown to abort migraine attacks. Biofeedback training has been especially successful with children. *See* HYPNOSIS.

Cluster headache. Cluster headache is the occurrence of migraines in groups or series. Men are more prone to cluster headaches. The cluster headache is characterized by its one-sided, excruciating attack that is usually localized around one eye. Typically, the attack is brief, lasting from several minutes to 1 h, and it often awakens the person from sleep. The associated manifestations include facial flushing, nasal congestion or draining, tearing, and on occasion drooping of the eyelid and constriction of the pupil. The series of headaches lasts from a few weeks to several months and then disappears for several months or years. The series appear most frequently in spring and fall. Although the cause remains unknown, any alcohol consumption during a series of cluster headaches is known to precipitate an acute attack.

The drugs of choice for treating cluster headaches are the ergotamines, but they must be used as early as possible in the attack to be effective. The inhalation of pure oxygen has also proved beneficial. Because cluster headaches are self-limiting, therapy is short-term. Methysergide, a potent serotonin receptor antagonist, is the agent of choice for prophylaxis. Corticosteroids, including prednisone, can prevent the onset of attacks. For those with chronic cluster headaches, lithium carbonate has provided some relief.

Other forms of vascular headache may be caused by systemic infection or fever, which causes dilation of the blood vessels. The ingestion of alcohol, poisons, or some medications used to treat hypertension or cardiac disease may produce adverse effects, including vascular headaches. *See* HYPERTENSION.

Muscle-contraction headaches. The most common form of form of headache is the muscle-contraction or tension headache. It is characterized by dull, constricting pain that can either occur intermittently or continue for days, months, or years. Intermittent muscle-contraction headaches usually occur once or twice a week or less frequently and show no definite pattern. Over-the-counter analgesics, including aspirin, acetaminophen, and the ibuprofen compounds, provide relief in most cases.

Muscle-contraction headaches usually affect both sides of the head and may be described as having a hat-band distribution of pain. The chronic form occurs on a daily or near-daily basis, with a worsening of pain in the morning and late evening. If the headache is associated with a sleep disturbance, it may be symptomatic of depression, and antidepressant therapy or psychological counseling may be beneficial.

A daily headache pattern carries the accompanying risk of habituation to prescription and over-the-counter analgesics. Hospitalization in a specialized unit dedicated to the problem of headaches may yield alternative means of obtaining pain relief.

Organic headaches. Very few headaches have an organic cause, such as brain tumor or aneurysm. Headache is not a prominent symptom of brain tumor: if present, headache will become progressively worse and constant, and it may not appear until late in the course of the tumor development. Other, more prominent neurologic symptoms usually arise before the onset of headache.

The headache associated with an aneurysm is usually mild until the aneurysm is at the point of rupture. If a patient complains of an exceptionally severe headache, organic disease, such as aneurysm, must be ruled out. Rupture of an aneurysm may have morbid consequences and must be treated immediately. *See* ANEURYSM.

Sinus headache is a common misnomer: chronic sinus disease does not cause headache. Acute sinus headache is characterized by nasal congestion and fever. The headache is minimal in the morning and increases in severity through the day. Treatment consists of decongestants as well as antibiotic therapy.

Temporamandibular joint (TMJ) disease involves a faulty bite or misalignment of the teeth and can cause a headache, but many people whose headaches have been incorrectly attributed to TMJ disease have not found relief from pain through dental appliances or surgery.

Eye conditions may also cause headache. The increased intraocular pressure of glaucoma, for example, may cause a headache, and so complaints of a recent onset of headache, particularly in the elderly, should prompt a screening for glaucoma. *See* GLAUCOMA; PAIN.

Seymour Diamond

Bibliography. S. Diamond and D. J. Dalessio, *The Practicing Physician's Approach to Headache*, 6th ed., 1999; H. G. Wolff, *Headache and Other Head Pain*, 5th ed., 1987.

Health physics

A branch of the environmental and occupational safety health sciences and professions concerned with the protection of people and the environment from possibly harmful effects of radiation, while providing for the utilization of radiation for the benefit of society. Health physicists are interdisciplinary radiation protection and safety specialists whose expertise draws from environmental science, mathematics, medicine, radiological health, radiation biology, chemistry, and physics. The subject requires understanding of the generation, measurement, and characteristics of radiation; environmental transport of radionuclides; dosimetry; effects of radiation in biological systems; and the regulations and recommendations governing the use of radiation. The field has expanded to include nonionizing as well as ionizing sources of radiation.

Although radiation and radioactivity were discovered in 1895, radiation protection activities began almost immediately, around 1900. The field received a major stimulus, expansion, and formalization during World War II as a part of the atomic bomb (Manhattan) project. Contemporary health physics includes basic and applied research on fundamental biological and physical radiation interaction mechanisms; on dosimetry, shielding, and instrumentation; on quantification of radiation effects; and on radioecology. Industrial hygiene, epidemiology, environmental engineering, and electronics and instrument development also include fundamental research in health physics.

Health physics specialists are involved in nuclear power radiation monitoring, control, and regulation required for the safe operation of facilities. In medicine, health physicists help design and survey radiation therapy facilities, and develop techniques to minimize patient doses so as to derive the maximum benefit from diagnostic radiology and nuclear medicine. Industrial applications employing health physics include industrial radiography and x-ray diffraction, gages, product and food sterilization by radiation, and radioactive tracer techniques. Health physics is employed in many aspects of nuclear waste management and remedial actions. Environmental health physics covers measuring and possibly reducing exposure to naturally occurring radioactivity, especially radon; monitoring nuclear facilities and waste sites; consumer-product radiation safety; and radiation accident monitoring, management, and mitigation. Health physics includes all aspects of radiation protection, from initial measuring and planning, through monitoring and implementation, to routine control and emergency management. *See* DECONTAMINATION OF RADIOACTIVE MATERIALS; DOSIMETER; ENVIRONMENTAL RADIOACTIVITY; NUCLEAR MEDICINE; RADIATION INJURY (BIOLOGY); RADIATION SHIELDING; RADIOACTIVE TRACER; RADIOACTIVITY; RADIOACTIVITY AND RADIATION APPLICATIONS; RADIOACTIVITY STANDARDS; RADIOECOLOGY; RADIOGRAPHY; RADIOLOGY; RADON.

Marvin Goldman

Bibliography. H. Cember, *Introduction to Health Physics*, 3d ed., 2000; National Council on Radiation Protection and Measurements, *Limitation of Exposure to Ionizing Radiation*, Rep. 116, 1993; B. Shleien (ed.), *The Health Physics and Radiological Health Handbook*, 3d ed., 1997.

Hearing (human)

The general perceptual behavior and the specific responses that are made in relation to sound stimuli. The auditory system consists of the ear and the auditory nervous system. The ear comprises outer, middle, and inner ear. The outer ear, visible on the surface of the body, directs sounds to the middle ear, which converts sounds into vibrations of the fluid that fills the inner ear. The inner ear contains the vestibular and the auditory sensory organs. The vestibular organ's function is the key to balance: this organ converts movements of the head into nerve signals. *See* EAR (VERTEBRATE).

The auditory part of the inner ear, known as the cochlea because of its snaillike shape, analyzes sound in a way that resembles spectral analysis. It contains the sensory cells that convert sounds into nerve signals to be conducted through the auditory portion of the eighth cranial nerve to higher brain centers. The neural code in the auditory nerve is transformed as the information travels through a complex system of nuclei connected by fiber tracts, known as the ascending auditory pathways (**Fig. 1**). They carry auditory information to the auditory cortex, which is the part of the sensory cortex where perception and interpretation of sounds are believed to take place. The auditory cortex is linked to other parts of the brain, but these connections are usually not considered part of the auditory system. Interaction between the neural pathways of the two ears makes it possible for a person to determine the direction of a sound's source. *See* BINAURAL SOUND SYSTEM; BRAIN.

Outer ear. The pinna collects sound, but because it is small in relation to the wavelengths of sound that are important for human hearing, the pinna plays only a minor role. The ear canal acts as a resonator: it increases the sound pressure at the tympanic membrane in the frequency range between 1500 and 5000 Hz; at 3500 Hz the increase is greatest, about 10 decibels, but together with the gain from the diffraction of the head, the sound pressure may be as much as 17 dB higher than outside the head. Because of diffraction by the head, the sound that reaches the ears depends on the angle to a sound source. The difference between the arrival time of a sound at each of the two ears and the difference in the intensity of the sound that reaches each ear are direct functions of the angle between a line through the two ears and the sound source. This difference is used by the auditory nervous system to determine the location of the sound source. That sound-localizing ability is most highly developed in the horizontal plane.

Middle ear. Sound that reaches the tympanic membrane causes the membrane to vibrate, and these vibrations set in motion the three small bones of the middle ear: the malleus, the incus, and the stapes. The footplate of the stapes is located in an opening of the cochlear bone—the oval window. Moving in a pistonlike fashion, the stapes sets the cochlear fluid into motion and thereby converts sound (pressure fluctuations in the air) into motion of the cochlear fluid. Since motion of the fluid in the cochlea begins the neural process known as hearing, the transformation of sound into motion of the cochlear fluid must occur as efficiently as possible. Sound is a weak force compared to the force required to set the cochlear fluid into motion; in other words, the air in which sound is transmitted has a low impedance compared to the impedance of the fluid that the sound must move. Because the most efficient transfer of energy occurs when the impedances of the energy source and the energy receiver are equal, direct transfer of sound to a fluid would be inefficient. In fact, 99.9% of the sound energy would be lost by reflection from the air–fluid interface. To improve efficiency,

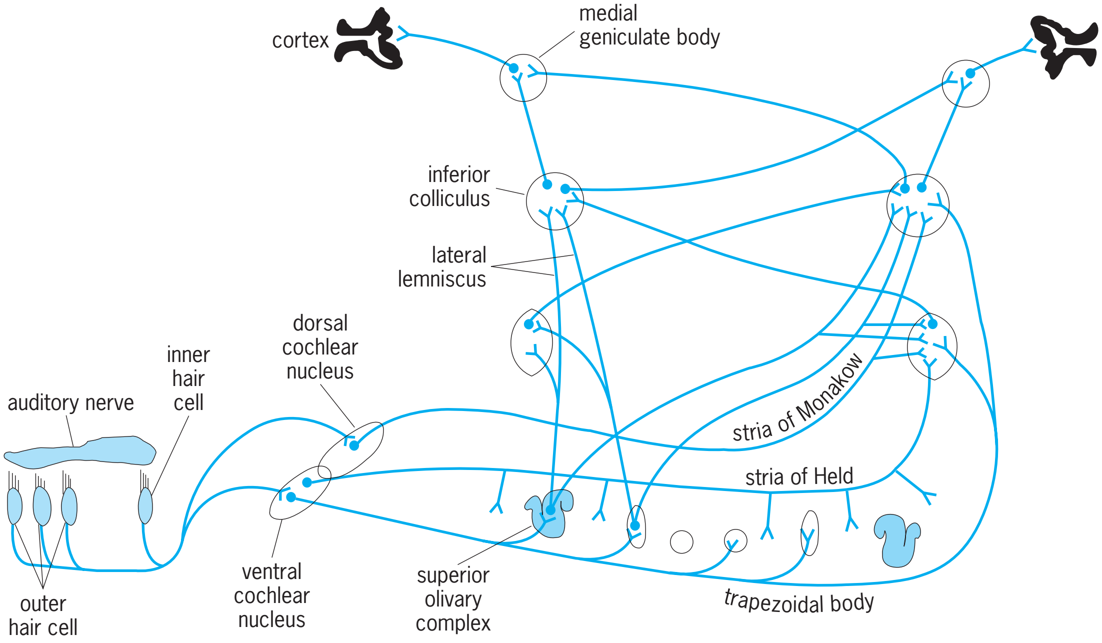


Fig. 1. Ascending auditory pathway. (*After A.R. Møller, Auditory Physiology, Academic Press, 1983*)

the middle ear in humans acts as an impedance transformer: it matches the high impedance of the cochlear fluid to the low impedance of air, thereby facilitating transmission of sound to the cochlear fluid. This transformer action mainly depends upon the ratio between the surface areas of the tympanic membrane and the stapes footplate. Optimum transfer of sound to the inner ear would occur if this ratio was equal to the square root of the ratio between the impedances of air and the cochlear fluid, giving a perfect impedance match. The properties of the human middle ear almost fit that criterion, reducing to a minimum the loss of sound energy from reflection.

The middle-ear bones are enclosed in the air-filled middle-ear cavity; it is important that the air pressure in this cavity be closely maintained at the ambient pressure. The Eustachian tube, which connects the middle-ear cavity to the nasopharynx, maintains that balance. This tube is usually closed, but it opens briefly during swallowing and yawning, thereby equalizing the air pressure in the middle-ear cavity.

There are two small muscles in the middle ear: the tensor tympani and the stapedius muscles. The former pulls the manubrium of the malleus inward, while the latter is attached to the stapes and pulls the stapes in a direction that is perpendicular to its pistonlike motion. The stapedius muscle is the smallest striated muscle in the body, and it contracts in response to an intense sound (of about 85 dB above the threshold of hearing). This is known as the acoustic middle-ear reflex. The muscle's contraction reduces sound transmission through the middle ear and thus acts as a regulator of input to the cochlea. Perhaps a more important function of the stapedius muscle is that it contracts immediately before and during a person's own vocalization, reducing the sensitivity of the speaker's ears to his or her own voice and possibly reducing the masking effect of an individual's own voice. The role of the tensor tympani muscle is less well understood, but it is thought that contraction of the tensor tympani muscle facilitates proper ventilation of the middle-ear cavity.

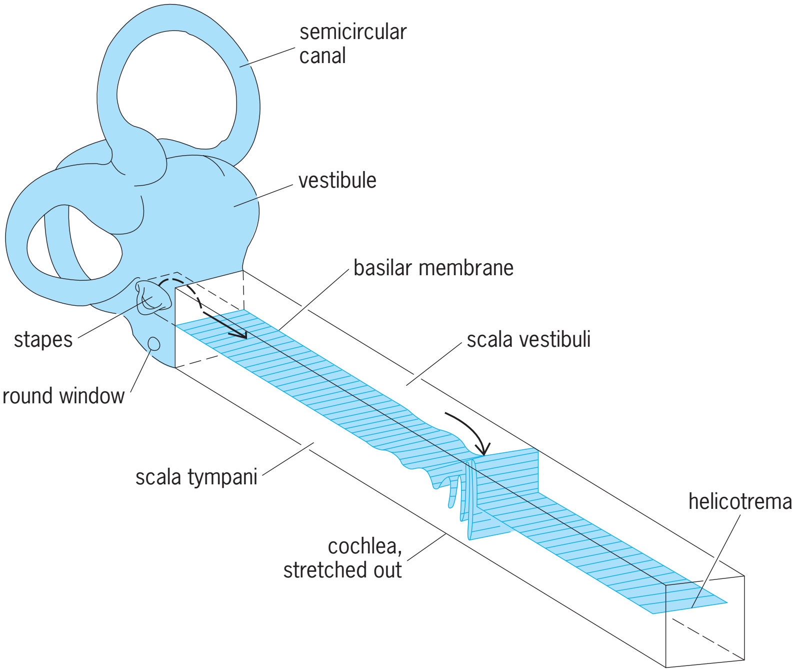


Fig. 2. Traveling wave along the basilar membrane in response to a pure tone that produces a maximum displacement of the membrane in its middle portion. (*After G. Zweig, R. Lipes, and J. R. Pierce, The cochlear compromise, J. Acoust. Soc. Amer., 29:1312–1317, 1957*)

These two muscles are innervated by different cranial nerves: the facial (VIIth) nerve for the stapedius and the trigeminal (Vth) nerve for the tensor tympani. The acoustic stapedius reflex plays an important role in the clinical diagnosis of disorders affecting the middle ear, the cochlea, and the auditory nerve. A contraction of the stapedius muscle can easily be recorded by using a change in the ear's acoustic impedance as an indicator, because a contraction of the stapedius muscle makes the middle ear more stiff.

Cochlea. In the 1930s, G. von Békésy proved that the basilar membrane of the cochlea has frequency-selective properties. Using human cadaver ears, he showed that vibrations in the cochlear fluid set up a traveling wave on the basilar membrane and that this wave always progresses from the base of the cochlea toward the apex. The amplitude of the wave increases until it reaches a certain point, and then rapidly decreases. The location of the point of maximal deflection is a direct function of the frequency of the sound. When tones are used to set the cochlear fluid into vibration, one specific point on the basilar membrane will vibrate with a higher amplitude than any other. Therefore, a frequency scale can be laid out along the basilar membrane, with low frequencies near the apex and high frequencies near the base of the cochlea (**Fig. 2**). The propagation velocity of this traveling wave is much lower than the propagation velocity of the sound wave in the fluid, and most of the energy is transferred to the basilar membrane at its basal part. As the traveling wave approaches the point of maximal deflection, the propagation velocity decreases markedly. These properties of the basilar membrane are largely a result of the change in compliance along the membrane, which is greater than 1:100.

The sensory cells that convert the motion of the basilar membrane into a neural code in individual auditory nerve fibers are located along the basilar membrane (**Fig. 3**). They are also known as hair cells, because they have hairlike structures on their surfaces. These "hairs" are actually stereocilia, and they are deflected when the basilar membrane vibrates. It is this deflection of the hairs that controls the discharges in individual fibers of the auditory nerve.

Sensory transduction in hair cells. The sensory cells in the mammalian cochlea are similar to the cells that are found in, for instance, the airways, where they transport mucus by rhythmic movements (beats) of their hairs (cilia). Cells in the airways have two types of cilia: stereocilia and kinocilia. The hair cells of the mammalian cochlea lose their kinocilia around the time of birth, however, and only stereocilia remain in the hair cells of the adult mammalian cochlea.

The hair cells in the mammalian cochlea function as mechanoreceptors: motion of the basilar membrane causes deflection of the hairs, starting a process that eventually results in a change in the discharge rate of the nerve fiber connected to each hair cell. This process includes the release of a chemical transmitter substance at the base of the hair cells that controls the discharge rate of the nerve fiber (**Fig. 4**). The hair cells also move when subjected to an electrical field, a property that may be the basis for the positive feedback in the cochlea.

Frequency tuning in auditory nerve fibers. Because of the frequency selectivity of the basilar membrane, individual auditory nerve fibers respond in accordance with the frequency of a sound; therefore, the frequency selectivity of the basilar membrane can be assessed by recording the neural impulses from single nerve fibers (**Fig. 5**). Tuning curves can be obtained by using a standard neurophysiological technique far less complex than that required for measuring the vibration of the basilar membrane.

The frequency selectivity of the basilar membrane had long appeared to be much less sharp than the tuning of single auditory nerve fibers, and so it was assumed that the neural transduction process in the cochlea included some unknown mechanism that sharpened the frequency selectivity of the basilar membrane. With the development of more refined methods for quantifying small vibration amplitudes, measurements could be made in living, anesthetized animals. Such measurements revealed that the frequency selectivity of the basilar membrane is in fact as sharp as neural tuning. Such studies have also shown that the frequency analysis that takes place in the cochlea is far more complex than was believed earlier, and some kind of positive feedback probably improves the frequency selectivity of the basilar membrane. This positive feedback in the cochlea may be compared with that used in early radio receivers to increase their frequency selectivity and so improve differentiation between two radio stations. The positive feedback in the cochlea requires biological energy and is therefore active only in the living, intact cochlea, which explains why von Békésy could not observe it in his studies. Because the effect

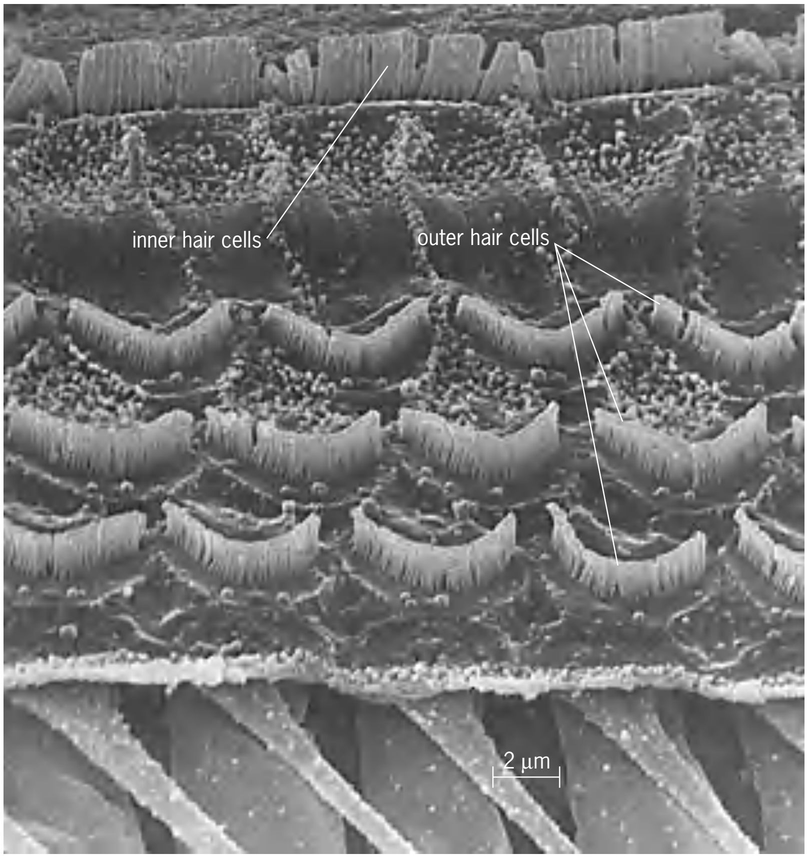


Fig. 3. Electron micrograph of hair cells along the basilar membrane. (*From A. R. Møller, Auditory Physiology, Academic Press, 1983*)

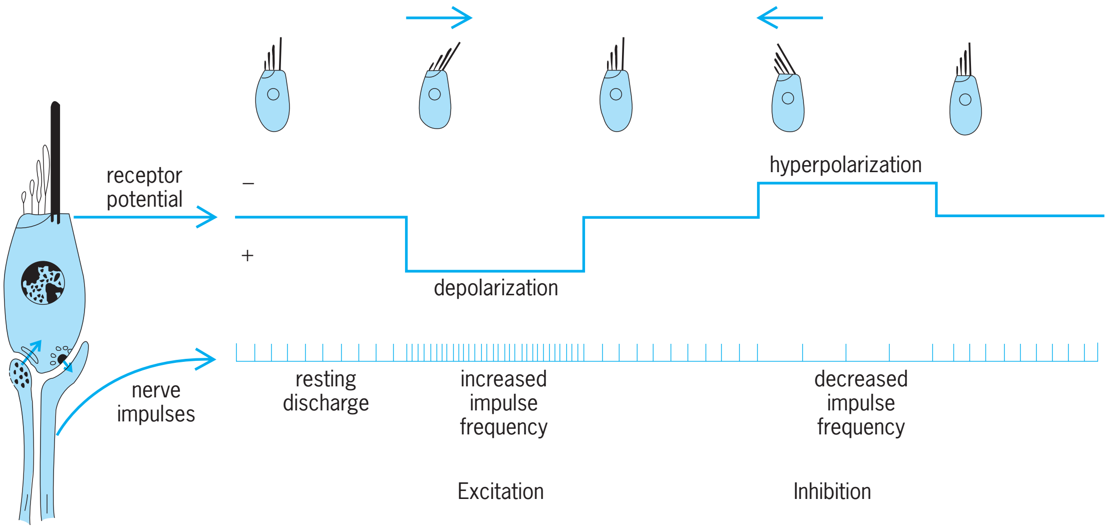


Fig. 4. Schematic illustration of the excitation in hair cells. A deflection of hairs in one direction results in an increase in the neural discharge rate in the nerve fiber associated with the hair cell, whereas a deflection in the opposite direction causes inhibition, or slowing, of this discharge rate. (*After A. Fiock, Transducing mechanisms in lateral line canal organ receptors, Proceedings of the Cold Spring Harbor Symposia on Quantitative Biology, 30:133–146, 1965*)

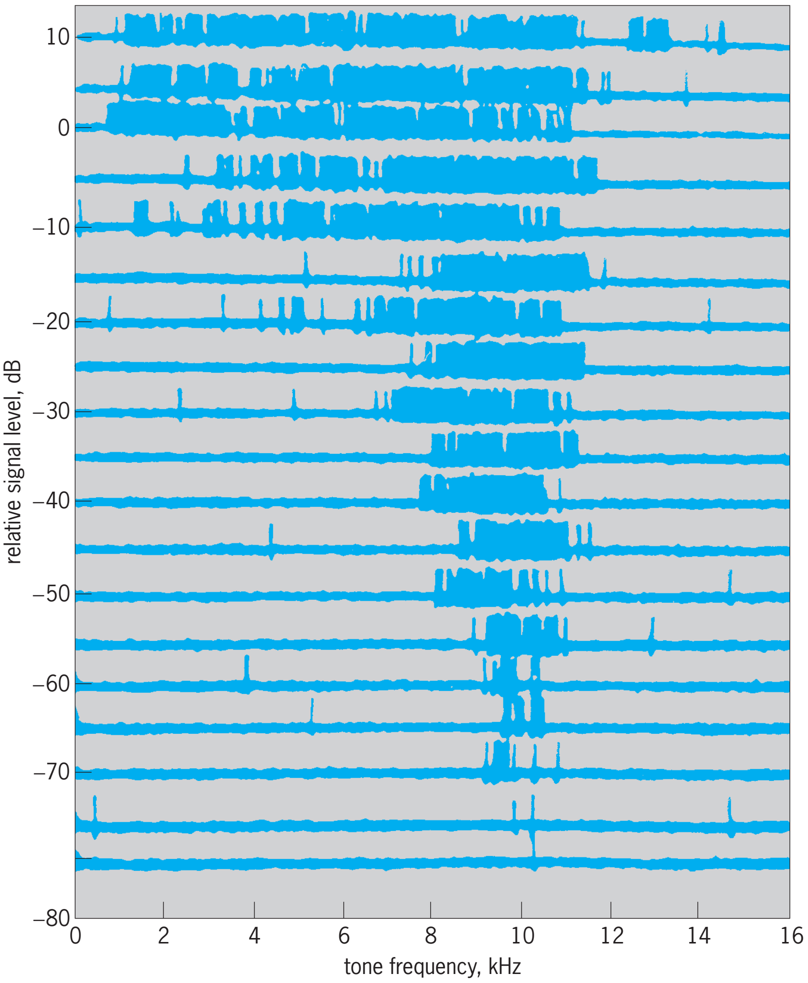


Fig. 5. Neural discharges evoked in a single auditory nerve fiber of an experimental animal exposed to tones of different frequencies. When a continuous tone was presented at different intensities and at variable frequencies, the neural discharges were recorded by using a microelectrode placed close to a single auditory nerve fiber. When the sound intensity is low, the fiber responds only within a narrow range of sound frequencies; as the sound intensity is raised, the nerve fiber responds over a larger and larger range. The contour that surrounds the area of increased neural activity is known as a tuning curve. (*After E. F. Evans*, *The frequency response and other properties of single fibers in the guinea pig cochlear nerve*, *J. Physiol. (London)*, *226:263–287*, *1972*)

of positive feedback probably decreases as sound intensity increases, the cochlea has less frequency selectivity for intense sounds than for weak sounds. The basilar membrane therefore acts as a nonlinear filter—unlike most manufactured frequency filters, such as those used in telecommunications, that are usually linear. The nonlinearity of the basilar membrane compresses the sound intensities that reach the ear, thereby extending the range of sound intensities in which the hearing sense can function adequately.

Neural frequency code. The frequency selectivity of the basilar membrane provides the central nervous system with information about the frequency or spectrum of a sound, because each auditory nerve fiber is "tuned" to a specific frequency. However, the frequency of a sound is also represented in the time pattern of the neural code, at least for frequencies up to 5 kHz, because the probability of the occurrence of a nerve impulse is highest during a certain phase of motion of the basilar membrane. This means that the time at which activity occurs in auditory nerve fibers provides information to the central nervous system about the frequency of a sound. Thus, the frequency or spectrum of a sound can be coded for place and time in the neural activity in the auditory nervous system. Both ways of coding the frequency or spectrum of a sound require neural processing in order to be interpreted, and although the precise mechanism remains unclear, psychoacoustic experiments reveal that the auditory system is capable of such an analysis. *See* AUDIOMETRY; PITCH.

Auditory nervous system. The ascending auditory nervous system consists of a complex chain of clusters of nerve cells (nuclei), connected by nerve fibers (nerve tracts). The chain of nuclei relays and transforms auditory information from the periphery of the auditory system, the ear, to the central structures, or auditory cortex, which is believed to be associated with the ability to interpret different sounds.

The four main nuclei in the ascending auditory pathway are not connected in a simple series. Rather, each nucleus relays information not only to the next higher nucleus but also to other nuclei in the chain. The complexity of the auditory nervous system is evidenced by the first relay station of the ascending auditory nervous system, the cochlear nucleus: each auditory nerve fiber connects with nerve cells in at least three divisions of the cochlear nucleus. Such parallel processing, a prominent feature of the auditory nervous system, is not confined to the right or left side: there are connections between the nuclei of the two sides. A descending auditory nervous system, which connects nuclei above with nuclei at lower levels, is also present.

When pure tones are used as stimuli to study the responses from individual neurons in the various nuclei of the ascending auditory nervous system, it is evident that individual neurons respond selectively to tones with frequencies in a discrete range. Such frequency tuning of nerve cells is similar to the frequency specificity of single auditory nerve fibers that is described above. However, the shapes of the tuning curves of nerve cells in the nuclei of the ascending auditory pathway differ from those of single auditory nerve fibers, with the difference being less marked for more peripherally located nuclei such as the cochlear nucleus.

Neurons in the entire auditory nervous system are, in general, organized anatomically according to the frequency of a tone to which they respond best, which suggests a tonotopical organization in the auditory nervous system and underscores the importance of representations of frequency in that system. However, when more complex sounds were used to study the auditory system, qualities of sounds other than frequency or spectrum were found to be represented differently in different neurons in the ascending auditory pathway, with more complex representation in the more centrally located nuclei. Thus, the response patterns of the cells in each division of the cochlear nucleus are different, which indicates that extensive signal processing is taking place. Although the details of that processing remain to be determined, the cells appear to sort the information and then relay different aspects of it through different channels to more centrally located parts of the ascending auditory pathway. As a result, some neurons seem to respond only if more than one sound is presented at the same time, others respond best if the frequency or intensity of a sound changes rapidly, and so on.

Another important feature of the ascending auditory pathway is the ability of particular neurons to signal the direction of sound origination, which is based on the physical differences in the sound reaching the two ears. Certain centers in the ascending auditory pathway seem to have the ability to compute the direction to the sound source on the basis of such differences in the sounds that reach the ears.

Knowledge of the descending auditory pathway is limited to the fact that the most peripheral portion can control the sensitivity of the hair cells. *See* SIGNAL DETECTION THEORY.

Cortical representation of sounds. The human auditory cortex, unlike that of many mammals, is buried deep in the temporal lobe and is, therefore, not easily accessible for study. As a result, most knowledge about cortical representation of sounds has been gained through experiments that use animals. Early in the history of neurophysiological research, it became evident that the surface of the cortex responds to simple sounds such as tones because it is tonotopically organized. However, with more complex sounds, complex maps better describe the cortical representation of sounds because they delineate the response not only to the frequency of a single tone but to many other qualities that may be related to behavior. Perhaps the best-known of such maps are those generated by the navigational sounds of the flying bats, which use their auditory systems much like radar for navigational purposes. The bat emits a sound and then listens for the echo, using the delay between emission of the sound and detection of the echo to determine the distance to an object. Actually, each animal emits a complex sound that contains components with rapidly changing frequencies and a series of harmonics. Analysis of the harmonic content of the echo helps to identify an animal's unique echo among those of other bats flying nearby. *See* ECHOLOCATION; PSYCHOACOUSTICS.

Neuronal computation of the bat's echoes has been shown to lead to specific cortical maps in which features relevant to navigation are represented on the surface of the cortex. These features are derived from the physical characteristics of a sound, which in this case are the echoes of the sounds emitted by the animal itself. Such features as the distance to obstacles or to a target (prey) and pure navigational data such as velocity and direction to a target were all represented in maps on the animal's auditory cortex. Maps that show how behavioral features appear for sounds that have meaning for other animals, such as humans, have yet to be devised.

Acoustic reflexes. Contraction of the middle-ear muscles in response to sound is involuntary and is called the acoustic middle-ear reflex. It is mediated through connections from the lower levels of the ascending auditory pathway to the motonuclei of the facial nerve (stapedius muscle) and trigeminal motonucleus (tensor tympani muscle). Other, more diffuse acoustic reflexes are the pinna reflex, prominent in some animals, and the startle reflex, which causes contractions of many skeletal muscles. *See* PHYSIOLOGICAL ACOUSTICS.

Auditory evoked potentials. Most knowledge about auditory nervous system function comes from studies of the discharge patterns of single nerve fibers or single nerve cells, which are recorded by placing microelectrodes near a single nerve fiber or nerve cell to record the electrical activity of that particular fiber or cell. Such recorded action potentials are identical waveforms, so that only the time of occurrence of a discharge is significant. Electrical activity of the nervous system can be studied, as well, by recording evoked potentials or compound action potentials, so named because they represent electrical activity in many hundreds or thousands of nerve cells or nerve fibers. By placing an electrode on the surface of those structures, electrical activity generated in specific parts of the nervous system can be recorded. Evoked potentials can also be recorded at a distance from these neural generators, that is, auditory evoked potentials can be recorded from electrodes on the scalp. Although evoked potentials provide

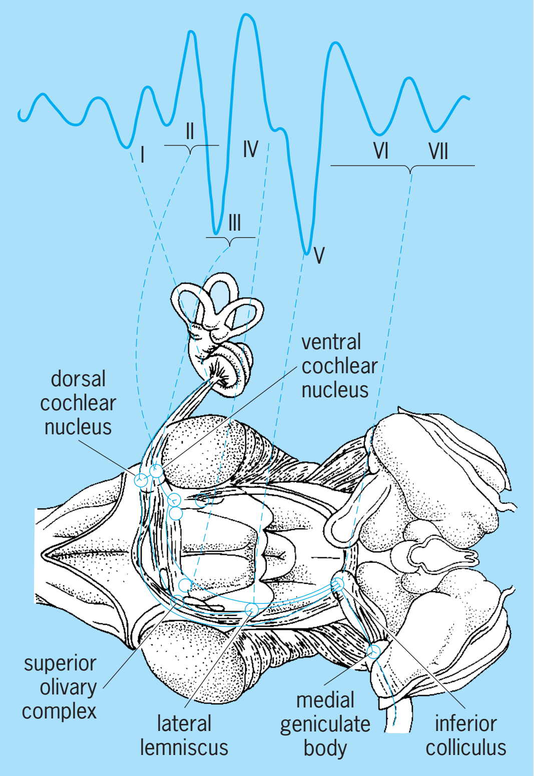


Fig. 6. Main generators of the five earlier peaks in the brainstem auditory evoked potential. (*After A. R. Møller and P. J. Janetta, Neural generators of the auditory brainstem response, in J. T. Jacobson, ed., The Auditory Brainstem Response, College Hill Press, 1984*)

only indirect information about the function of the auditory nervous system, they can be recorded noninvasively. Auditory evoked potentials recorded from the scalp, known as far-field potentials, have very small amplitudes, but repeating the same stimulus many times and adding the responses makes it possible to separate those potentials from background noise. The exact origin of the different components of the far-field potentials from the ascending auditory pathway, known as brainstem auditory evoked potentials or BAEP, is unclear, but sufficient information about their origin has been gathered (**Fig. 6**) to make them important tools in the diagnosis of disorders that affect the ear and the auditory nervous system. *See* HEARING IMPAIRMENT; LOUDNESS; MASKING OF SOUND; PHONORECEPTION; SOUND; SPEECH PERCEPTION. Aage R. Møller

Bibliography. R. V. Harrison, *The Biology of Hearing and Deafness*, 1988; R. D. Luce, *Sound and Hearing*, 1992; B. C. Moore, *Frequency Selectivity in Hearing*, 1986; J. O. Pickels, *An Introduction to the Physiology of Hearing*, 2d ed., 1988; W. A. Yost, *Fundamentals of Hearing: An Introduction*, 4th ed., 2000.

Hearing (vertebrate)

The ability to perceive sound arriving from distant vibrating sources through the environmental medium (such as air, water, or ground). The primary function of hearing is to detect the presence, identity, location, and activity of distant sound sources. Sound detection is accomplished using structures that collect sound from the environment (outer ears), transmit sound efficiently to the inner ears (via middle ears), transform mechanical motion to electrical and chemical processes in the inner ears (hair cells), and then transmit the coded information to various specialized areas within the brain. These processes lead to perception and other behaviors appropriate to sound sources, and probably arose early in vertebrate evolution.

Sound. Sound is gathered from the environment by structures that are variable among species. In many fishes, sound pressure reaching the swim bladder or another gas-filled chamber in the abdomen or head causes fluctuations in volume that reach the inner ears as movements. In addition, the vibration of water particles that normally accompany underwater sound reaches the inner ears to cause direct, inertial stimulation. In land animals, sound causes motion of the tympanic membrane (eardrum). In amphibians, reptiles, and birds, a single bone (the columella) transmits tympanic membrane motion to the inner ears. In mammals, there are three interlinked bones (malleus, incus, and stapes). Mammals that live underground may detect ground-borne sound via bone conduction. In whales and other sea mammals, sound reaches the inner ears via tissue and bone conduction.

Auditory anatomy. The inner ears of all vertebrates contain hair-cell mechanoreceptors that transform motion of their cilia to electrochemical events resulting in action potentials in cells of the eighth cranial nerve. Patterns of action potentials reaching the brain represent sound wave features in all vertebrates.

In fishes the auditory portions of the ears are otolith organs containing a group of hair cells and a solid calcium carbonate otolith which is in contact with the hair-cell cilia. The sacculus is the major auditory organ in most fishes, but other organs (utricle and lagena) may function in hearing as well. In mammals these otolith organs serve the sense of balance.

Most amphibians have two auditory organs: the amphibian and basilar papillae that respond to lower and higher frequencies, respectively. The auditory organ of reptiles and birds is the basilar papilla situated on the basilar membrane.

The mammalian cochlea contains hair cells in the organ of Corti on the basilar membrane. The cochlea varies in length from 10 mm (in some rodents) to 60 mm (elephant). The inner ears of mammals and birds perform a mechanical sorting of frequencies through variations of basilar membrane stiffness from base to tip. In some amphibians, reptiles, and birds, frequency selectivity may also depend on the electrochemical mechanisms of hair-cell membranes. Eighth nerve cells are selective for different sound frequencies. All vertebrates have an analogous set of auditory brain centers. *See* EAR (VERTEBRATE); PHYSIOLOGICAL ACOUSTICS.

Hearing abilities. Experiments show that vertebrates have more commonalities than differences in their sense of hearing. The major difference between species is in the frequency range of hearing, from below 1 Hz to over 100,000 Hz. In other fundamental hearing functions (such as best sensitivity, sound intensity and frequency discrimination acuity, time and frequency analysis, and source localization), vertebrates have much in common. The hearing abilities of over 180 species from all vertebrate classes have been investigated (see **illus.**). All detect sound within a restricted frequency range. Audiograms graphically display the lowest sound intensities that an animal can hear (threshold) for the range of audible frequencies. All species are able to detect sounds in the presence of interfering sounds (noise), discriminate between different sound features, and locate the sources of sound with varying degrees of accuracy.

Fishes. Fishes with greatest sensitivity and frequency range are the "hearing specialists." These species have a mechanical link between the gas bladder and the ear (for example, goldfish and catfish). Fishes without a gas bladder (such as flatfish) detect only acoustic particle motion. The best sensitivity falls between 100 and 1000 Hz. The most sensitive fishes have the best sensitivity in the range between −20 and 0 dB.

Amphibians and reptiles. There are very few audiograms for amphibians and reptiles because it is difficult to train these animals. The two lowest curves in illus. *b* were obtained for two frogs using conditioning methods. The audiogram for a turtle (broken line) shows poor sensitivity. For amphibians, best sensitivity approaches 0 dB.

Birds. The audiograms for 20 bird species are similar in sensitivity and frequency range, with the best sensitivity between −15 and 25 dB and the best frequency between 1000 and 6000 Hz. The barn owl is exceptional with best hearing between 2000 and 9000 Hz.

Mammals. The audiograms for 57 mammals show great variation in frequency range. The low-frequency limit varies from 20 to 10,000 Hz. The high-frequency limit varies from 11,000 to 150,000 Hz. No single species hears within this entire frequency range. Some marine mammals and bats have the best high-frequency hearing. Some rodents, elephants, and primates hear well at lower frequencies. Most mammals are similarly sensitive at their best frequency of hearing.

The sensitivity range is similar among all groups, with some species in all groups having a best sensitivity in the region of −20 to 0 dB. Fishes, amphibians, reptiles, and birds hear best between 100 and 5000 Hz. Only mammals hear at frequencies above 10,000 Hz. Humans and elephants have the poorest high-frequency hearing.

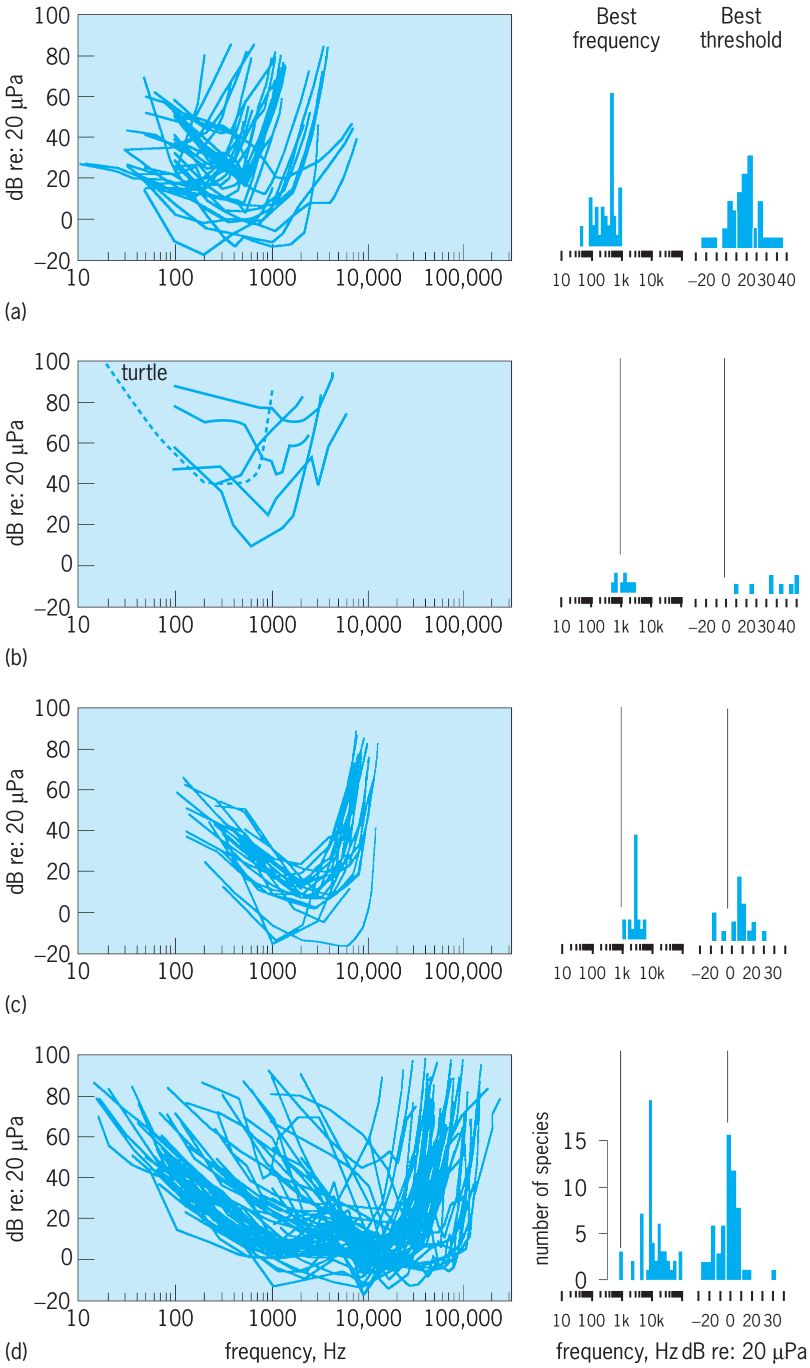
Sound intensity discrimination. All vertebrates have the ability to discriminate between sounds that differ only in sound level or intensity, but the acuity of this discrimination varies among species (1 to 3 dB differences can be discriminated). Level discrimination thresholds are approximately the same at all frequencies and intensities.

Temporal pattern. Most natural sounds vary over time in level or intensity. These patterns are important for the detection, identification, and classification of sources. The threshold for detecting a brief sound increases with sound duration in all vertebrates. At durations greater than 100 to 400 milliseconds, sensitivity reaches a maximum value. Features with durations as brief as 1 to 4 ms can be resolved by birds and mammals. Animals trained to discriminate between sounds differing only in duration can discriminate differences of 5–10%.

Sound source localization. All vertebrates are able to determine the direction of sound sources in the horizontal and vertical directions with some degree of accuracy. In land animals the information used for localization in the horizontal plane comprises differences in the times and intensities at which sound features reach the two ears.

Fishes localize sound sources with an accuracy of 10–20°, but the mechanisms for localization are not well understood. Time and intensity differences at the two ears are too small to be useful. However, the angle of acoustic particle motion is represented in eighth nerve cells due to the directional selectivity of the otolith organs and their hair cells.

Many amphibians, reptiles, and birds have close-set ears, so that time and intensity differences reaching the two ears are reduced. These animals localize sound sources in the horizontal and vertical planes with an accuracy of 10–20°. In some species, there are specialized mechanisms that exaggerate small time and intensity differences reaching the two ears.



Behavioral audiograms for fishes ($n = 49$), amphibians ($n = 8$), a reptile ($n = 1$), birds ($n = 20$), and mammals ($n = 57$). Thresholds are expressed in decibels with respect to 20 micropascals. Thresholds for aquatic animals tested underwater are the sound pressure levels in air corresponding to the sound intensity levels underwater calculated from underwater sound pressure thresholds assuming far-field conditions. On the right are shown frequency distributions of best frequency and best, or lowest, threshold obtained from the audiograms. [*Modified after R. R. Fay (1988), with permission, where the references and tabled data for all audiograms can be found*]

Mammals with small heads also have reduced time and intensity difference cues for low frequencies. These animals rely on intensity difference cues for high-frequency sounds. Mammals with large heads and good sensitivity at low frequencies use time difference cues for horizontal localization. Acuity varies among mammals between 1 and 20°. Humans and porpoises have the most acute localization abilities.

Frequency processing. All vertebrates tend to break down complex sounds into their constituent frequency components. This process enhances sound detection and is important for sound source identification and localization. Acuity in frequency selectivity includes the minimum discriminable difference between pure-tone frequencies, and measures of the frequency range within which one sound may interfere with, or mask, the detection of another sound.

Minimum discriminable frequency differences decline as sound level is raised, but grow larger in proportion to the frequency at which the discrimination is measured. Detecting the presence of a pure tone in the presence of masking noise bands of different center frequency and frequency range indicates that masking interference occurs only in a restricted frequency range surrounding the signal (the critical band).

For all vertebrates, the estimates of frequency resolution tend to be increasing, parallel functions of frequency. The greatest frequency acuity occurs in animals having poor high-frequency hearing and long cochleas (such as the human and elephant). These hearing functions likely have had a long and stable evolutionary history. Thus, all vertebrates have solved common problems of sound perception. *See* AUDIOMETRY; LOUDNESS; MASKING OF SOUND; PHONORECEPTION; SOUND. Richard R. Fay

Bibliography. R. R. Fay, *Hearing in Vertebrates: A Psychophysics Databook*, Hill-Fay Associates, Winnetka, IL, 1988; R. R. Fay, Structure and function in sound discrimination among vertebrates, in D. Webster, R. Fay, and A. Popper (eds.), *The Evolutionary Biology of Hearing*, Springer-Verlag, New York, 1992; R. S. Heffner and H. E. Heffner, Evolution of sound localization in mammals, in D. Webster, R. Fay, and A. Popper (eds.), *The Evolutionary Biology of Hearing*, Springer-Verlag, New York, 1992; W. C. Stebbins and M. A. Berkeley (eds.), *Comparative Perception: Complex Signals*, Wiley, New York, 1990.

Hearing aid

A device that amplifies sound for someone with a hearing loss. A typical aid consists of a microphone, an electronic amplifier, an earphone (called a receiver), and a battery. The sound is coupled to the ear with an earmold. The great majority of earmolds are cast from impressions of the individual patient's ear canal, but over-the-counter and mail-order hearing aids use prefabricated earmolds, commonly offered in several sizes. In the case of in-the-ear hearing aids, all components are small enough to be contained in an earmold shell that fits within the ear canal and concha, or even within the canal alone (**Fig. 1**). *See* AMPLIFIER; EAR (VERTEBRATE); EARPHONES; MICROPHONE.

Development. Pre-electric ear trumpets provided acoustic gain as early as the late eighteenth century. By the early 1920s, hearing aids containing miniature vacuum tubes could provide an amount of amplification limited only by the quality of the seal between the earmold and the ear (to prevent whistling caused by feedback). By 1962 the transistor plus subminiature microphones and earphones had made the components small enough to produce all-in-the-ear hearing aids. *See* TRANSISTOR.

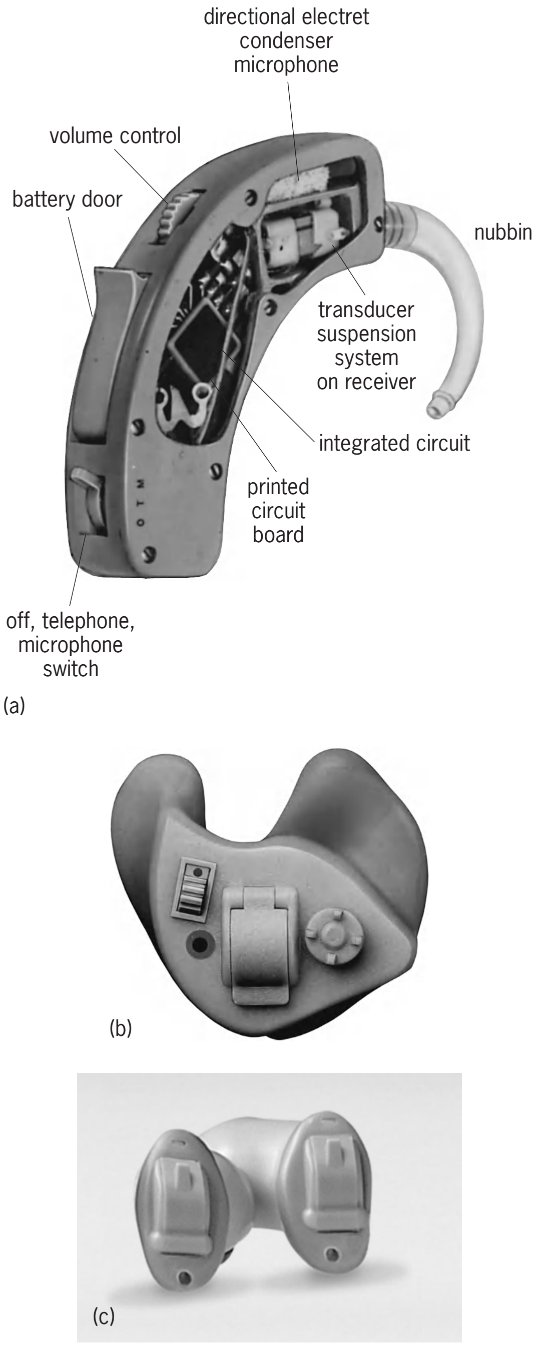

directional electret condenser microphone
volume control
battery door
nubbin
transducer suspension system on receiver
integrated circuit
printed circuit board
off, telephone, microphone switch
(a)
(b)
(c)

Fig. 1. Hearing aids. (*a*) Over-the-ear aid with case open to show components, and (*b*) in-the-ear aid (*Beltone USA, Glenview, IL*). (c) In-the-canal aid (*Siemens Hearing Instruments, Piscataway, NJ*).

By 1989, hearing-aid designers had come to understand that they needed to take the transmission characteristics of the ear into account to create normal frequency response at the eardrum, and that

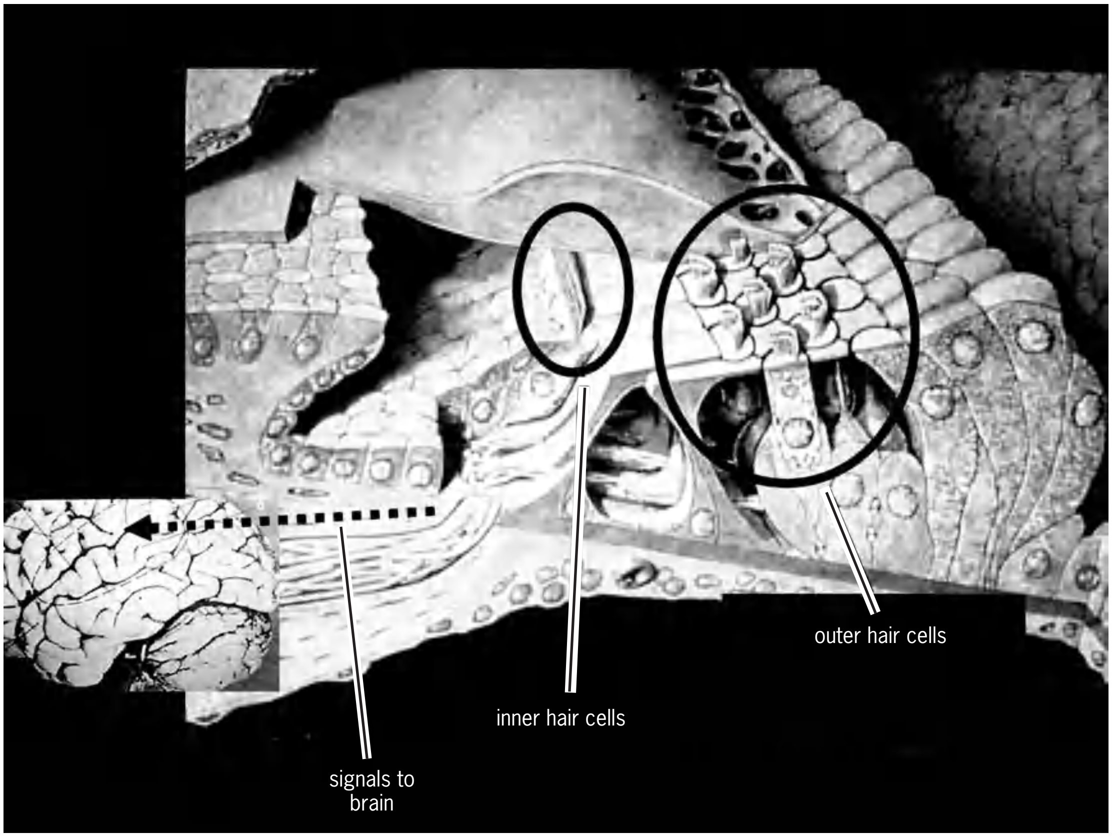


Fig. 2. Inner and outer hair cells found on the basilar membrane of the inner ear. Tiny hairs, or stereocilia, are visible on the top of each cell. Inner hair cells provide all the signals to the brain. Outer hair cells amplify quiet sounds.

hearing loss is typically accompanied by hearing distortions that need to be compensated by electronic processing of the signal. This understanding, combined with integrated circuits, high-quality miniature microphones and receivers, zinc-air batteries, and high-efficiency class D amplifiers, made it practical to produce hearing aids with a fidelity (exclusive of the processing) equal to the best home stereo systems. *See* BATTERY; INTEGRATED CIRCUITS.

Types of hearing loss. Hearing loss caused by abnormalities in the mechanical parts of the ear is called conductive; that created by loss of hair cells of the cochlea is called cochlear or sensorineural; and that created by pathology of the eighth nerve (much less common) is called neural. Conductive hearing loss is effectively overcome with surgery or hearing aids, while the latter two types of loss present a greater challenge for hearing aids.

In the normal inner ear, approximately 15,000 outer hair cells provide some 30–50 dB of mechanical amplification for quiet sounds, while approximately 5000 inner hair cells convert sound-induced mechanical energy into nerve impulses that go to the brain (**Fig. 2**). A major cause of damage to these hair cells is excessive exposure to loud noise and music. A loud ringing in the ear after such an exposure is an indication of likely death of some of the cells. The loss of hair cells causes two symptoms: a loss of hearing sensitivity to quiet sounds, and a loss of ability to understand speech in noise. The two symptoms can occur in various ratios. When loss of sensitivity is the dominant symptom, hearing aids can restore nearly normal hearing by amplifying these quiet sounds without overamplifying louder sounds, and by providing extra gain for those high-frequency sounds made inaudible by the typical high-frequency hearing loss. A signal-processing system called wide-dynamic-range compression (WDRC) does all of that. Although this system was widely criticized when it was introduced, it is now used in almost all analog and digital hearing aids. *See* SIGNAL PROCESSING.

Unfortunately, a loss of speech cues available to the brain often accompanies a loss of inner (as opposed to outer) hair cells, and the loss of redundant cues makes it difficult for the listener to understand speech in noise. This difficulty can be helped by increasing the signal-to-noise ratio in several ways: using directional-microphone hearing aids, leaning closer to the talker, using a microphone held close to the talker, or using a system of microphones worn by each of several talkers (a development that can improve the signal-to-noise ratio by a factor of 10). *See* HEARING (HUMAN); HEARING IMPAIRMENT.

User satisfaction. Surveys show that a nearly constant 60% of hearing-aid users are satisfied overall, with no increase of satisfaction from analog to digital aids. The dissatisfied 40% include those whose hearing aids are not properly fitted or are inappropriate to the loss; those whose earmold fit does not prevent feedback or prevent their own voices from sounding hollow (the latter effect is called occlusion); those— a small percentage—for whom wearing a hearing aid is disturbing under any circumstances; and most important, those who are unable, even with the hearing aids, to understand speech in the presence of noise or competing speech. For the last group, both analog and digital aids may fail equally by providing audibility but not the ability to separate speech from interference. The surveys show no difference in overall satisfaction between inexpensive analog hearing aids and expensive digital hearing aids, although both analog and digital aids with directional microphones produce higher satisfaction. (As noted above, an even greater improvement in signal-to-noise ratio can be obtained using a microphone held close to the talker or using a system of microphones worn by each of several talkers.)

While many of the claims of the superiority of digital over analog hearing aids have been exaggerated, digital circuits do have potential advantages deriving from their flexibility, and the better ones provide sophisticated feedback reduction that does not simultaneously reduce audibility or intelligibility. From the standpoint of convenience, a digital circuit can analyze a listening environment and adjust the hearing aid to the processing the manufacturer has chosen for that environment (directional versus omni mode, for example), so that the user does not have to readjust the controls. Although many digital hearing aids have not achieved the sound quality of the best analog aids, digital aids have neither an inherent advantage or disadvantage in sound quality compared to analog aids.

Binaural hearing aids. Some years ago the use of binaural hearing aids (an aid in each ear) was looked on with some suspicion, as a way to get people to buy a second aid that offered no advantage. The Federal Trade Commission strictly limited advertising claims for the advantages of binaural hearing aids. It is now recognized that an aid in each ear offers three real advantages: (1) Binaural release from masking: because of the directional cues created by binaural listening, the ability to separate speech from interference is improved by several decibels. (2) Binaural loudness summation: a sound heard by two ears is louder than the same sound heard by one ear. This means that less amplification is required of each hearing aid of a binaural pair, and for the same loudness the physical level of sound going into each ear is less. The danger of feedback, and the danger of distortion created by overload of the mechanical parts of the ear, are thereby reduced. (3) The loss of intelligibility in the unaided ear that often occurs over time from the auditory deprivation is avoided. *See* MASKING OF SOUND.

Mead C. Killion; Edgar Villchur

Bibliography. K. W. Berger, *The Hearing Aid: Its Operation and Development*, National Hearing Aid Society, 1984; M. C. Killion, Guest editorial: Hearing aids, past, present, future—moving toward normal conversations in noise, *Brit. J. Audiol.*, 31:141–148, 1997; M. C. Killion, High fidelity and hearing aids, *Audio*, 75(1):42–44, 1990; S. Kochkin, On the issue of value: Hearing aid benefit, price, satisfaction, and brand repurchase rate, *Hear. Rev.*, 10(2):12–26, February 2003; S. Kochkin, 10-year customer satisfaction trends in the U.S. hearing instrument market, *Hear. Rev.*, 9(10):14–25, 46, October 2002; S. Silman, S. A. Gelfand, and C. A. Silverman, Late-onset auditory deprivation: Effects of monaural versus binaural hearing aids, *J. Acous. Soc. Amer.*, 76:1357–1362, 1984; E. Villchur, A different approach to the noise problem of the hearing impaired, *Amer. J. Audiol.*, 2(2):47–51, July 1993.

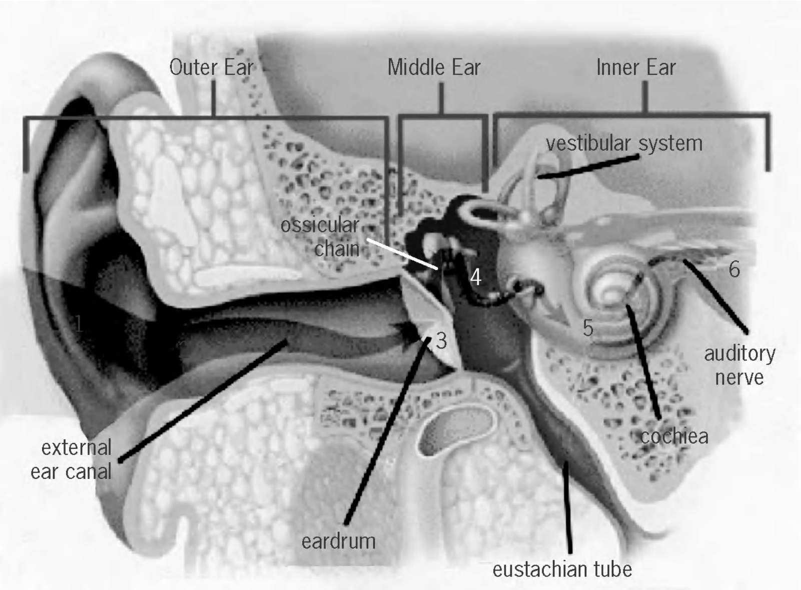
Hearing impairment

Any alteration of hearing capacity. Hearing impairment can be of various degrees, including mild, moderate, severe, profound, or total. The degree of impairment typically is categorized by the loss of hearing sensitivity, that is, how loud sounds must be for a listener to hear them. The degree of impairment can refer either to the loss of hearing sensitivity for individual pitches of sounds for each ear separately or to an overall loss of hearing sensitivity for both ears. Hearing impairment is further categorized as unilateral if present in only one ear and as bilateral if present in both ears.

Hearing impairment may be present at birth or acquired later in life. Because hearing serves many functions, from the simple detection of warning sounds to the more complex functions necessary for the perception of speech, the effects of a hearing impairment are quite varied. The most prominent effects depend on a variety of factors, including the degree of the impairment, the pitch range of the impairment, and the stage of language and speech development when the impairment occurred. Congenital hearing loss greatly interferes with normal language and speech development if it is bilateral and of severe or greater magnitude over the pitch range that covers speech sounds. Acquired hearing loss can occur gradually or suddenly at any time of life and therefore can also be defined in relation to the development of language and speech. Hearing impairment is often termed prelingual, perilingual, or postlingual if the hearing loss occurred prior to, during, or after the development of language and speech, respectively.

The term deafness has two meanings. If it refers only to a total lack of hearing function, the term is lowercase. If it refers to an individual with bilateral hearing loss who does not use oral language and speech, the term has an initial capital. Thus, an individual may be deaf in one ear and not Deaf; another individual may be Deaf, yet have some residual hearing ability.

The auditory system (see **illustration**) consists of the outer ear (1), the external ear canal (2), the eardrum (3), the system of bones in the middle ear called the ossicular chain (4) that conducts sound from the eardrum to the inner ear, which consists of the hearing organ called the cochlea (5), and the vestibular system, which controls balance. The cochlea senses and analyzes sounds and sends corresponding impulses along the auditory nerve (6) to the brain. Hearing loss can be classified by the part of the auditory system that is defective. Conductive hearing loss results from abnormalities or diseases of the outer or middle ear; sensorineural hearing loss results from abnormalities or diseases of the inner ear or auditory nerve; and central hearing impairment results from abnormalities or diseases of the auditory portions of the central nervous system. Combined conductive and sensorineural hearing loss is referred to as mixed hearing loss. *See* AUDIOMETRY; EAR (VERTEBRATE).



Major portions of the human auditory system.

Causes. Hearing loss results from a broad range of causes, some of which are reversible, and so hearing loss may be either temporary or permanent.

Conductive abnormalities. Many conditions exist whereby the transmission of sound through the conductive mechanism is impeded. These conductive abnormalities result in a reduction of sound reaching the inner ear, so that sounds must be made louder to be heard. The degree of reduction in sound transmission caused by the abnormality determines the degree of the hearing loss.

External canal abnormalities may be due to ear wax (cerumen), which can become hardened or impacted in the external ear canal. Bacteria, fungi, or other microorganisms may infect the skin of the external ear canal and cause sufficient swelling of the walls to block sound waves. Congenital malformations may also occlude or block the canal.

Middle-ear abnormalities may be associated with a sequence of conditions that often results in otitis media, an infection in the middle-ear cavity that is a common cause of conductive hearing loss, especially in children. This sequence may begin with a common cold, allergy, or upper respiratory tract infection. Nasal secretions can pass to the middle ear through the eustachian tube; if the tube becomes involved and cannot open properly, the surrounding tissues absorb the air in the middle-ear cavity. The higher air pressure in the external ear forces the drum membrane inward, restricting movement of the ossicular chain and inhibiting sound conduction. Watery secretions then form in the middle-ear space and may be thickened by the infection, resulting in suppurative (pus-forming/discharging) otitis media. If the infection does not resolve, the secretions may expand enough to perforate the eardrum and further impede the transmission of sound.

Air travel can be hazardous to the ears for some people. Aero-otitis media begins during flight when the difference in barometric pressure on the two sides of the eardrum is not corrected by swallowing. High-speed movement in elevators and deep-water diving may produce a similar effect.

Another middle-ear abnormality is otosclerosis, a hereditary disease. Portions of the bony capsule surrounding the inner ear, normally the hardest bone in the body, become decalcified and are replaced by a soft spongy bone that enlarges and hardens, often fixing the ossicular chain at the point where the middle-ear conductive mechanism joins the inner ear. The volume of the sound reaching the inner ear becomes greatly reduced. Otosclerosis is a progressive disease that usually begins in youth. The hearing loss is first apparent in adolescence or during the early twenties. It is more common among whites and East Indians than other racial groups, with a slight predominance in females.

Sensorineural abnormalities. Many conditions affect the generation of electrical potentials in the inner ear and their transmission by the auditory nerve. These abnormalities result in two different effects on hearing. When hearing sensitivity becomes impaired, sounds must be amplified to be heard. In addition, because the sensory and neural structures are involved (the structures that analyze sound into its components), the sound that is heard becomes distorted.

Inner-ear abnormalities may be due to congenital malformations that may arise independently or in association with other syndromes such as Waardenburg syndrome (which causes hearing loss and pigmentary anomalies) or Usher's syndrome (which causes hearing loss and progressive vision loss). Many drugs and chemicals can be damaging to the inner ear (that is, ototoxic). The effect may be temporary or permanent and depends on the dosage and how they are applied. The types of ototoxic drugs (with examples) include salicylates (aspirin), nonsteroidal anti-inflammatory drugs (naproxen), antibiotics (gentamycin), diuretics (ethacrynic acid), chemotherapeutic agents (cisplatin), and quinine. Inner-ear damage can be caused by bacterial infections such as meningitis and viral infections such as measles and mumps. Loud noise is an insidious environmental cause of inner-ear damage; it may be in the form of either a sudden explosive sound such as a blast, or a prolonged exposure. Initially, noise exposure results in a temporary loss of hearing for the high pitches, known as temporary threshold shift. The condition may become permanent with habitual exposure. Damage is cumulative and usually occurs without pain. *See* ACOUSTIC NOISE; LOUDNESS.

Aging is one of the most common causes of inner-ear abnormalities, but it also affects other parts of the auditory system. Hearing loss associated with the aging process is called presbycusis, and frequently leads to complaints by the older person that other people do not speak distinctly. This condition results from the loss of sensory cells of the inner ear, particularly those associated with the hearing for high pitches that are characteristic of certain speech sounds. Auditory nerve abnormalities are occasionally caused by a tumor within the skull which can press against the auditory nerve and disturb its function.

Central nervous system abnormalities. Even though the hearing mechanism is sending the correct signals to the brain, the auditory pathways in the central nervous system can experience abnormalities that result in hearing impairments. Some causes of central nervous system deficits that may affect hearing include degenerative diseases such as multiple sclerosis, cerebral vascular accidents from hemorrhage or clotting, birth deformities, and brain trauma. *See* NERVOUS SYSTEM (VERTEBRATE).

Nature of hearing impairment. Conductive hearing loss impairs sensitivity to sound; if the sound is amplified, the impairment can usually be overcome. In noisy places where speaking more loudly is an automatic response, a person with mild conductive hearing loss, who hears less background noise and has learned to listen carefully for quiet speech, may actually have an advantage over a person with normal hearing. Most of these conditions can be reversed with medical or surgical intervention.

In sensorineural hearing loss, the abnormality can affect different portions of the inner ear so that often hearing sensitivity may be normal or nearly so for low-pitched tones but falls off sharply for higher tones. In addition, what is heard may be distorted. Persons with mild-to-moderate sensorineural hearing loss may hear the louder low-pitched portions of speech such as vowels but may not hear, or hear with distortion, quieter, high-pitched speech such as the voiceless consonants (that is, consonants produced without sound from the vocal cords). The person thus has the disquieting sensation of hearing speech but not understanding it. When hearing loss occurs gradually, the adult learns to accommodate with supplemental speech reading (lip reading) and filling in from the context of the conversation. The speech of the hearing-impaired person may deteriorate over time because the high-pitched speech sounds cannot be heard. A child born with sensorineural hearing loss that is undiagnosed or untreated may fail to acquire intelligible speech and may wrongly be thought inattentive, developmentally delayed, or ill behaved. *See* SPEECH DISORDERS; SPEECH PERCEPTION.

In most cases of sensorineural or mixed hearing loss, both sensitivity and clarity are impaired, so the perceived sound is weak and distorted. A hearing aid that amplifies sound fails to correct the distortion, and so the difficulty of understanding speech is only partially remedied. With severe or greater loss, the listener may hear only the cadence and the intensity variations of speech without being able to discriminate all the speech sounds. Speech deteriorates for adults who suffer this type of hearing loss after speech has developed normally. For children born with severe or profound bilateral sensorineural hearing loss, special education is required for the acquisition of speech and language. *See* HEARING AID; PSYCHOACOUSTICS.

Other symptoms characterize sensorineural hearing loss. Tinnitus head noise or ringing in the ears that is heard with no related acoustic stimulus may occur continuously or intermittently, and is described as a rushing or roaring noise. Vertigo and nausea may accompany the hearing deficit if the abnormality affects the vestibular system, as in Ménière's disease (a disorder of the inner ear causing fluctuating hearing loss, vertigo, tinnitus, and a sensation of pressure in the affected ear). In double hearing, or diplacusis, a single tone is heard at a different pitch in each ear, or simple tones may sound fuzzy or rough. Loudness recruitment, an abnormal increase in perceived loudness as a sound is intensified, may be associated with a low tolerance for loud noises.

Treatment and management. Prevention is the best approach to possible loss of hearing. Universal neonatal hearing screening programs are often mandated and have proven very successful. Many newborns have their hearing screened before being discharged from the hospital nursery. Though such programs do not prevent hearing loss, they do allow for early intervention that then can lead to normal language and speech development. Desirable preventive measures include immunization for viral and bacterial diseases, early medical intervention for upper respiratory infections or earaches, keeping both nostrils open when blowing the nose, control of allergies, avoidance of ototoxic drugs, and reduced exposure to loud sounds. Routine hearing tests throughout life are also advisable.

Antibiotic control of ear infections is a common treatment for conductive hearing loss, and surgical procedures are also available. Myringotomy is a simple surgical perforation of the tympanic membrane to drain fluids from the middle ear. Tympanoplasty, a more extensive surgical procedure, can eradicate middle-ear infection. A stapedectomy is the reconstruction of the ossicular chain that has been fixed in place in the oval window of the inner ear by otosclerosis. Under local anesthetic, the footplate of the third bone in the ossicular chain, the stapes, is removed and replaced with a graft of connective tissue and a pistonlike prosthesis that connects the flexible oval-window graft to the remainder of the ossicular chain. Many of these procedures for conductive hearing loss allow hearing to return to normal or near-normal levels.

Sensorineural hearing impairment is generally permanent and not amenable to medical intervention. For sensorineural hearing losses that have not yet become permanent, or that fluctuate or that are the result of processes that may be reversible, the measures to correct the hearing loss are varied. Tranquilizing and antivertiginous drugs, vasodilative agents, and low-salt diets are among the treatments prescribed for the symptomatic management of Ménière's disease, but they have met with only moderate success. Various medications are sometimes used with certain sudden hearing loss, again with only moderate success. The lack of widely accepted and effective pharmaceutical approaches to sensorineural hearing loss has placed an emphasis on the use of a variety of devices to correct for hearing impairment.

Cochlear implants. For permanent severe or profound bilateral sensorineural hearing loss, the cochlear implant seeks to restore hearing by direct electrical stimulation of the auditory nerve, thereby producing an auditory sensation that represents environmental sounds and speech sounds. The cochlear implant consists of a very small electrode array that is inserted into the cochlea adjacent to the auditory nerve fibers. This electrode array is in turn connected to a small module located under the skin behind the ear. An external processor is then added that picks up acoustic signals in the environment and transmits them across the skin where they are picked up by the internal module and converted to electrical signals on the electrode that then activate any remaining auditory nerve fibers. Over 60,000 patients already have been implanted, and this number is growing at a very rapid rate. For the adult who experiences a permanent, severe or profound sudden bilateral sensorineural hearing loss, the cochlear implant often restores sufficient auditory function to allow normal or near-normal communication ability, including face-to-face conversation and telephone use. The device also helps substantially in the educational process for teaching oral language and speech to prelingual hearing-impaired children. Universal neonatal hearing screening programs are now a part of routine neonatal health care in many countries throughout the world. Many deaf children are detected at birth by such programs, then receive a cochlear implant at 1 year of age, attend special and intensive education programs centered on language and speech development, and by age 5 or 6 are able to enter regular schools with their hearing peers and with normal language and normal or near-normal speech.

Hearing aids. A conventional hearing aid is a battery-operated device that amplifies sound. Hearing aids come in a variety of configurations from small modules that fit behind the outer ear to tiny modules that fit completely in the ear canal. Over 90% of all hearing aids are now digital, which allows many additional features. These include adaptive directional microphones to help users screen out background or extraneous sounds; automatic reduction of the extraneous whistling associated with older hearing aids; more settings to fine-tune performance in different environments, such as crowds; automatic and coordinated adjustment of settings in both ears to account for changing conditions in the listening environment; controls linked to other devices, such as the user's wristwatch, which allow the owner to change settings more easily and discreetly; and improved integration with cell phones and other personal listening devices. Current hearing aids that are adjusted properly and worn for at least a short period often provide substantial benefit to many hearing-impaired individuals. For others, additional improvement can be obtained with additional aural rehabilitation which includes special auditory training to use this amplification and instruction in speech reading (lip reading).

Because the spontaneous development of language and speech is highly dependent on an infant's ability to hear speech, the child with severe to profound congenital hearing loss requires either a hearing aid or a cochlear implant (or both) and intensive special education in order to talk. Hearing-impaired individuals with this type of training are known as oral deaf. Hearing-impaired individuals who do not develop oral language and speech but instead are taught sign language belong to the Deaf community, comprising individuals who have sign language as their primary language. Gerald R. Popelka

Bibliography. H. Cooper and L. Craddock, *Cochlear Implants: A Practical Guide*, 2d ed., 2006; J. Katz (ed.), *Handbook of Clinical Audiology*, 5th ed., 2001; H. A. Newby and G. R. Popelka, *Audiology*, 6th ed., 1992; J. L. Northern and M. P. Downs, *Hearing in Children*, 5th ed., 2001; M. Valente (ed.), *Strategies for Selecting and Verifying Hearing Aid Fittings*, 2002.

Heart (invertebrate)

Hearts of invertebrates can be categorized according to the source of the electrical rhythmicity that underlies their beat. Rhythmic electrical activity can arise in the muscle itself (myogenic hearts) or in neurons that drive the heart muscle (neurogenic hearts). Most mollusks and some insects appear to have purely myogenic hearts; these hearts beat normally when isolated from neural inputs. Conversely, the hearts of the higher crustaceans and the xiphosuran *Limulus* are usually considered to be purely neurogenic: motor neurons impose their rhythmic electrical activity on heart muscle fibers by means of direct excitatory synapses. Without neural input, the heart ceases to beat. Other invertebrates, including gnathobdellid leeches and some insects, have hearts that can produce a myogenic beat but

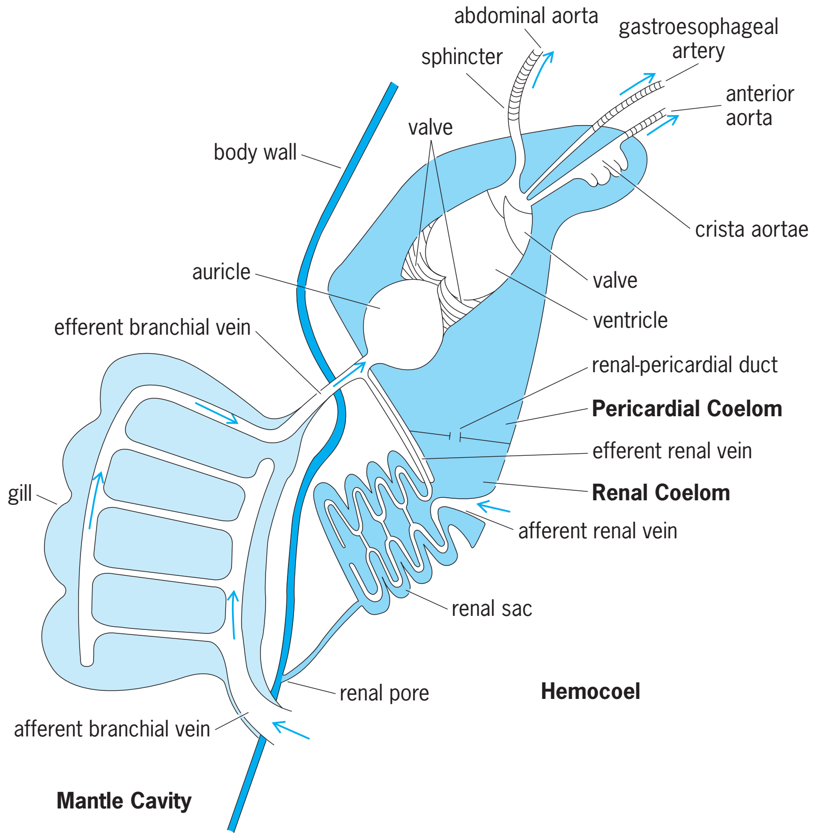


Fig. 1. *Aplysia* **circulatory system. Arrows indicate the normal direction of blood flow. (After E. Mayeri et al., *Neural control of circulation in Aplysia, I. Motoneurons, J. Neurophysiol., 37:458–475, 1974*)**

require rhythmic neural input to coordinate that beat and maintain the proper rate.

Myogenic hearts. In the marine snail *Aplysia*, a muscular heart consisting of an auricle and a ventricle is located in a dorsal pericardial cavity (**Fig. 1**). The rhythmic contractions of the auricle fill the ventricle with hemolymph, which is then pumped through the open circulatory system by the rhythmic contractions of the ventricle. The normal heartbeat period lasts about 3 s.

A pair of semilunar valves prevents backflow of hemolymph into the auricle during ventricular contraction. Three arteries issue from the ventricle toward the anterior, and a single semilunar valve prevents backflow from them during ventricular expansion. The arteries carry the hemolymph to the various body organs, where they end in tissue spaces. The hemolymph then collects in the hemocoel and returns to the heart by two parallel veins, one through the kidney and one through the gill.

Although the heart is innervated, its normal beat persists after denervation, and so the heart is myogenic. There is no discernible cardiac pacemaker muscle, but contraction of the ventricle usually begins at its base near the valves that separate it from the auricle.

The cardiovascular motor cells which innervate the heart are located in the abdominal ganglion. Two heart inhibitors which slow the heartbeat and a heart excitor which speeds the heartbeat have been identified. In addition, an excitatory motor neuron of unknown function innervates the heart. Three excitatory cardiovascular motor neurons which innervate certain main arteries have also been identified. The activity of these neurons causes the arteries to constrict and thus allows them to regulate blood pressure.

The cardiovascular motor cells are controlled by an elaborate network of heart interneurons within the abdominal ganglion. The interneuron network coordinates the activity of the motor cells so that appropriate cardiovascular adjustments can be made. For example, under anoxic conditions the snail displays a behavior called respiratory pumping, which is designed to maximize aeration of the gill. During this response, cardiac output decreases in an orderly way. The heart interneurons excite the heart inhibitors while inhibiting both the heart excitor and the vasoconstrictor motor neurons. This

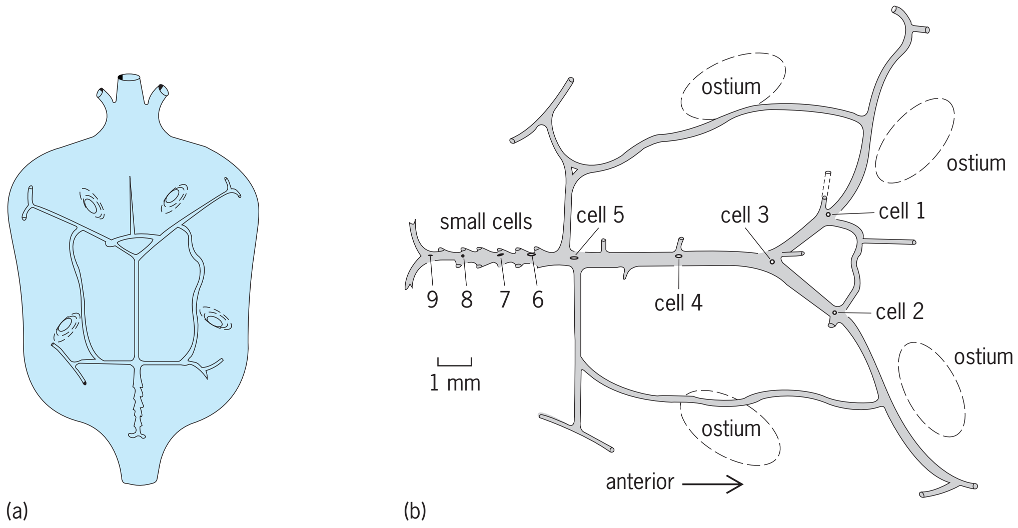


Fig. 2. Cardiac ganglion of *Homarus*. (*a*) Nerves on the inner dorsal surface of the heart. (*b*) Cell body position of all nine intrinsic neurons. (After D. K. Hartline, *Integrative neurophysiology of the lobster cardiac ganglion, Amer. Zool., 19:53–65, 1979*)

interneuronal action slows the heartbeat, lowers blood pressure, and prevents conflicting cardiovascular responses.

The heartbeat rhythm of *Aplysia* derives directly from the myogenic properties of the cardiac muscle fibers. However, a central neural network of cardiovascular motor cells and heart interneurons regulates the heartbeat rate and blood pressure. This network ensures that the cardiovascular system will respond appropriately to the environmental demands confronting the animal. The precise neural control of the heart and the existence of the functionally undefined heart motor neuron call into question the purely myogenic nature of the heart.

Neurogenic hearts. A muscular heart pumps hemolymph through the open circulatory system in lobsters (**Fig. 2***a*). This heart is located dorsally along the thoracic midline and is suspended within a pericardial cavity by ligaments. The heartbeat period lasts about 2 s. Large anterior- and posterior-going arteries, which branch extensively to supply various body organs, issue from the heart. Semilunar valves, located at the juncture of each artery with the heart, prevent backflow of blood into the heart when it relaxes. Hemolymph enters the heart from the pericardial sinus through six ostia, which have valves to ensure unidirectional flow.

The rhythmic discharge of motor neurons innervating the heart by way of excitatory chemical synapses produces the heartbeat. These motor neurons are located in the cardiac ganglion on the inner dorsal surface of the heart (Fig. 2*a* and *b*). The cardiac ganglion contains only nine neurons, which generate a simple two-phased rhythm of electrical activity. The four posterior small cells (cells 6–9) are interneurons, and the five anterior large cells (cells 1–5) are the motor neurons.

The neurons of the cardiac ganglion receive direct synaptic input from the central nervous system through one pair of inhibitory axons and two pairs of excitatory axons. Cardiac ganglion cells are excited by heart stretch, presumably by the direct activation of neuron collaterals which enter the heart muscle. A neurohormone released by the pericardial organ also directly excites cardiac ganglion cells.

All nine neurons of the ganglion produce nearly concurrent rhythmic impulse bursts of about 0.5 s in duration, separated by quiescent periods of 1.5 s. Impulse burst production begins with the interneurons, followed closely by the motor neurons which discharge at 50–100 Hz. These high-frequency impulse bursts in the motor neurons drive the heartbeat. In the absence of this rhythmic input, the heart ceases to beat; thus the heart is truly neurogenic.

The cardiac ganglion neurons generate normal rhythmic impulse bursts when totally isolated from the heart, pericardium, and central nervous system. Thus, the heartbeat rhythm is generated solely by the intrinsic neurons of the cardiac ganglion. Sensory feedback from heart stretch, input from the central nervous system, and neurohormones modulate only the rate and intensity of the heartbeat rhythm generated by the ganglion. This modulation adjusts the heartbeat rate to the lobster's internal state and to its environmental needs. For example, heartbeat rate is affected by activity, illumination, and feeding.

The interneurons of the cardiac ganglion form a positive feedback network by means of mutually excitatory chemical synapses. The interneurons also make excitatory chemical synapses on the motor neurons. All the neurons of the ganglion appear to be capable of generating an endogenous electrical rhythm which underlies impulse burst production. The motor neurons, however, have their activity cycle imposed upon them by the interneurons through strong excitatory synapses. Primary control of the overall impulse burst rhythm of the ganglion is exerted by a pacemaker interneuron, cell 9, which fires first in the ganglionic impulse burst. The excitatory synaptic interactions among the interneurons initiate and coordinate the production of impulse bursts in the other interneurons, once the pacemaker has begun to fire. Thus the interneurons produce a coordinated impulse burst, which is synaptically imposed on the motor neurons. The motor neurons then produce a synchronous impulse burst, which is synaptically imposed upon the heart muscle and causes it to contract.

The lobster cardiac ganglion can be thought of as a semiautonomous nervous system comprising two tiers of neurons (interneurons and motor neurons). This nervous system produces rhythmic impulse bursts in the motor neurons that drive the heartbeat. Sensory feedback, input from the central nervous system, and neurohormones modulate the frequency and intensity of the impulse burst rhythm produced by the ganglion cells.

In *Limulus*, which has a neurogenic heart with an organization similar to that of the lobster, the neuropeptide proctolin induces myogenic properties in the heart muscle. Although similar action of proctolin has not been reported in lobsters, whose nervous system contains the neuropeptide, this observation calls into question whether the lobster heart is purely neurogenic.

Hearts with combined properties. The medicinal leech (*Hirudo medicinalis*) is a segmented worm, and its segmentation is reflected in the anatomy and neural control of its circulatory system. The closed circulatory system comprises four longitudinal vessels which run the length of the animal (**Fig. 3**). These vessels are in fact coelomic sinuses. The sinuses communicate with one another at each end of the animal and in each segment by a series of transverse vessels. Smaller vessels and capillaries branch off the segmental transverse vessels and invade the body tissues. Spiral muscle cells located in the walls of the two lateral sinuses, or heart tubes, contract rhythmically to produce the rhythmic constrictions which move blood through the circulatory system. These constrictions of the lateral sinus constitute the heartbeat.

Normally the constriction cycle of each segmental heart tube section is coordinated with its neighbors

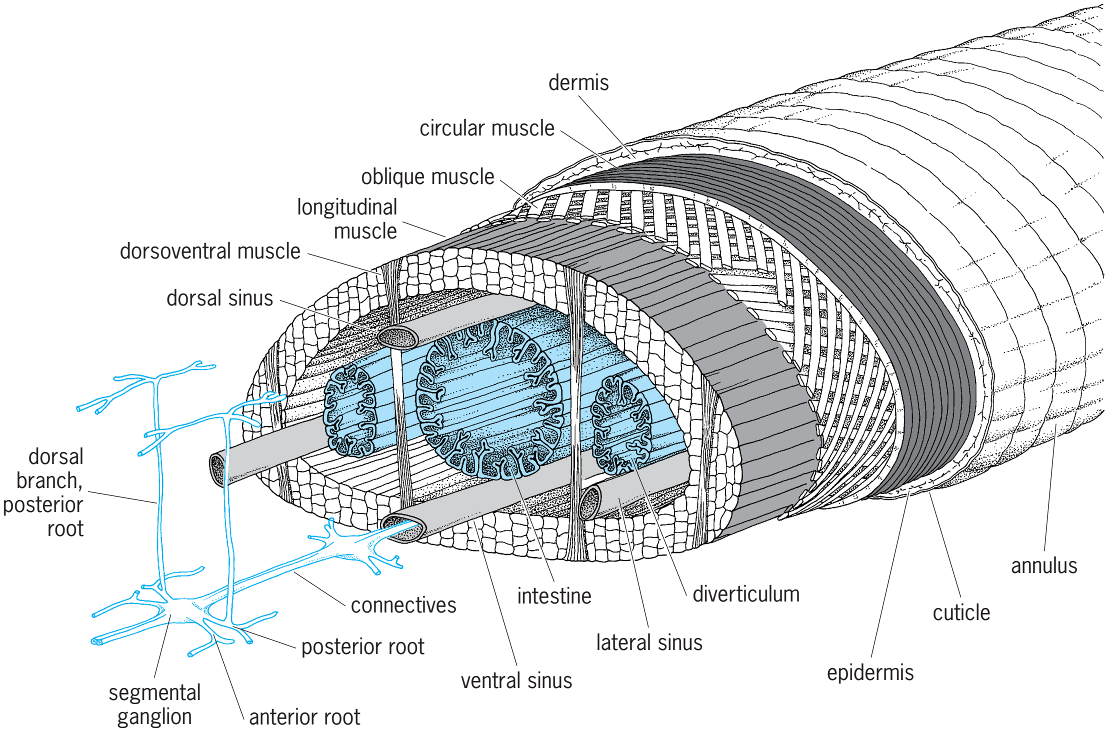


Fig. 3. **Cross section of the leech *Hirudo medicinalis*, showing its major organs and muscles, including its ventral nerve cord and the four longitudinal coelomic sinuses which compose the circulatory system. The nerve cord is enclosed in the ventral sinus. The lateral sinuses constrict rhythmically and are called heart tubes. (*After J. G. Nicholls and D. van Essen*, *The nervous system of the leech, Sci. Amer.*, *230*(*1*)*:38–48*, *1974*)**

so that on one side they constrict in a rear-to-front progression (peristalsis), while on the other they constrict nearly in synchrony. Reciprocal, right-left transitions between peristaltic and synchronous coordination states occur spontaneously every 10–20 heartbeat cycles. The heartbeat period varies from 10–30 s depending upon temperature and other undefined influences. The muscle of each segmental heart tube section, from the third through eighteenth segment, is innervated from each corresponding segmental ganglion of the ventral nerve cord by a rhythmically active segmental HE motor neuron. Each heart is also innervated intersegmentally by two cells, the HA modulatory neurons, that reside in the fifth and sixth segmental ganglia of the nerve cord.

Denervation of the hearts does not cause cessation of their rhythmic beat. However, the segmental heart tube sections lose all trace of intersegmental coordination, and their beat cycle period is prolonged. Thus, rhythmic neural activity is necessary to establish the complex coordination of the segmental heart tube sections and to maintain a normal heartbeat period. In contrast, the rhythmic activity of the HE cells in completely isolated nerve cords faithfully mimics the heartbeat pattern with all its subtle nuances. This observation shows that the complex heartbeat pattern, which normally paces the heart, is generated within the central nervous system.

The myogenic activity of the heart muscle is entrained by the rhythmic activity of the HE motor neurons through excitatory cholinergic synapses and is regulated by neuropeptides contained in these neurons and the HA neurons. The HA neurons contain the peptide FMRFamide, while the HE neurons contain that peptide or a closely related one. The HA modulatory neurons appear to regulate mainly beat strength through peptide release, while HE motor neurons appear to ensure that the heart will be rhythmically active (myogenic) through peptide release.

A set of segmental interneurons, HN cells, controls the activity of the HE cells by means of direct inhibitory synapses. Bilateral pairs of HN cells have been identified in the first seven segmental ganglia. Not all HN cells connect with HE cells, but those that do connect with all ipsilateral HE cells in more posterior ganglia.

HN cells produce rhythmic impulse bursts through combination of endogenous membrane properties and reciprocal inhibitory synapses. Their inhibitory synaptic interactions also coordinate the activity of the entire ensemble into a precise pattern. The HE cells are tonically active in the absence of inhibitory input from the HN cells. Normally they are driven into rhythmic impulse bursts by periodic inhibition from the HN cells. The intersegmental nature of HN cell to HE cell synaptic contacts ensures the proper intersegmental coordination of the HE cells. The rhythmic activity of the HE cells is then imposed upon the hearts themselves by the excitatory cholinergic synapses that they make on heart muscle. The establishment of the two different heartbeat coordination states and the reciprocal changes between them are mediated by the HN cell pair in the fifth segmental ganglion.

Although the hearts of the medicinal leech are capable of disjointed myogenic contractions and their myogenic properties are regulated by modulatory neural inputs, an elaborate central neural network of heart interneurons and motor neurons also exists. This central network imposes its activity upon the hearts by means of direct excitatory synapses and causes them to beat with the proper period and intersegmental coordination. The central network also mediates the reciprocal changes in coordination state between the hearts. *See* NERVOUS SYSTEM (INVERTEBRATE).

Ronald L. Calabrese; Christine S. Cozzens

Bibliography. R. B. Hill and F. Lang (eds.), Symposium on the Comparative Physiology of Invertebrate Hearts, *Amer. Zool.*, 19:3–175, 1979; A. I. Selverston (ed.), *Model Neural Networks and Behavior*, 1985; G. S. Stent, W. J. Thompson, and R. L. Calabrese, Neural control of heartbeat in the leech and some other invertebrates, *Physiol. Rev.*, 59:101–136, 1979.

Heart (vertebrate)

The muscular pumping organ of the cardiovascular system.

Anatomy. The heart typically lies ventrally, near the anterior end of the trunk; it is ventral and medial to the gills in fish and at the base of the neck or in the chest region of tetrapods. In humans it is located behind the breastbone and ribs between the third and fifth costal cartilages. Its anterior portion or base is directed to the right and dorsally and is the area where the great vessels enter and leave the heart. The lower muscular portion ends in a blunt apex which lies behind the fifth costal cartilage on the left.

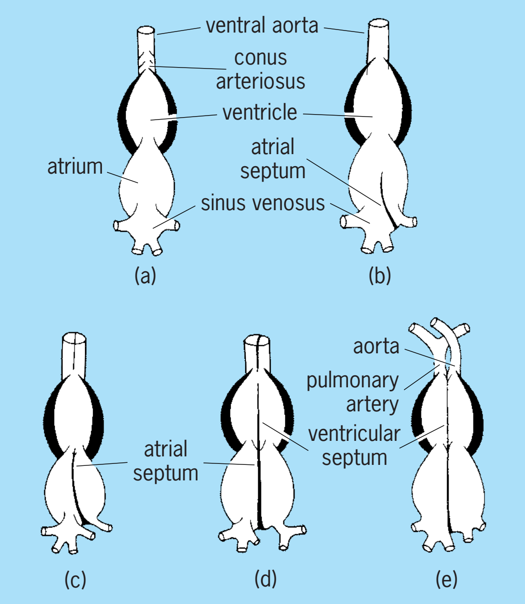


Fig. 1. Vertebrate hearts in ventral view, drawn in the primitive "stretched out" position to show the different chambers. (*a*) Dogfish. (*b*) Crossopterygian. (*c*) Amphibian. (*d*) Reptile. (*e*) Mammal.

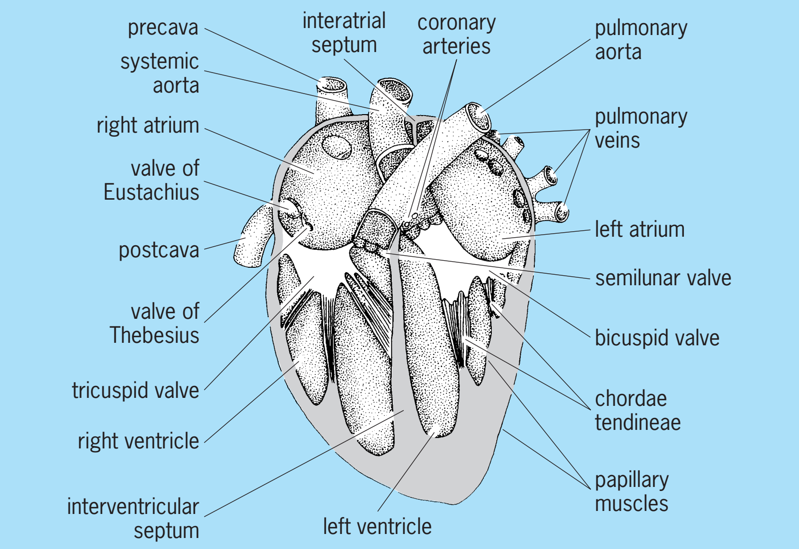


Fig. 2. Internal structure of four-chambered mammalian heart, ventral view. (*After C. K. Weichert*, *Anatomy of the Chordates*, *2d ed.*, *McGraw-Hill*, *1958*)

The muscular wall of the heart, the myocardium, is lined by an inner endocardium and is covered externally by the membranous visceral pericardium. There are coronary arteries and veins to and from the heart, which has a specialized neuromuscular conducting system and autonomic nerve supply.

Water breathers. In fishes the heart is basically a simple tube (**Fig. 1**) which becomes subdivided into four successive chambers, the sinus venosus, atrium, ventricle, and conus arteriosus. Blood from the body enters the sinus and leaves the conus to go to the gills to be oxygenated. The ventricle supplies the main pumping force.

In fishes and all other animals with backbones the heart becomes bent to fit into the available space so that the atrium, originally posterior to the ventricle, comes to lie dorsal and finally anterior to it; in Fig. 1 the heart is shown as if it were still a simple straight tube.

Air breathers. When lungs are introduced into the system in lungfish and tetrapods, the mixing of oxygenated and nonoxygenated blood becomes a problem. In brief, the sinus venosus and conus arteriosus disappear, becoming incorporated into the other chambers or the bases of the great vessels. At the same time the atrium and later the ventricle become divided into right and left chambers by a median septum. Various stages in this development are seen in different amphibians and reptiles.

In birds and mammals including humans (**Fig. 2**) the medial fibromuscular septum divides the heart into two lateral halves, each consisting of a thin-walled receiving chamber or atrium (auricle, often used as a synonym of atrium, properly refers to a small, earlike projection of the atrium) and a thicker, muscular pumping chamber or ventricle. Blood enters the right atrium from the superior and inferior venae cavae which drain most of the body. It passes through the tricuspid valve to the right ventricle and is pumped to the lungs during systole, or contraction of the heart. Blood returns from the lungs by way of the pulmonary veins to the left atrium, passes into the left ventricle through the mitral valve, and during contraction is pumped out into the aorta. *See* CARDIOVASCULAR SYSTEM. Thomas S. Parsons

Heart disorders

Pathologies of the heart and its blood vessels. Almost all cardiovascular disorders eventually progress to serious debilitating stages characterized by heart failure (reduced pumping function), dysrhythmias (abnormal electrical rhythms), and sudden death. Coronary atherosclerosis, a disease causing obstruction of the arteries that supply nutrient blood to heart muscle, is the leading cause of cardiovascular mortality. Hypertension (high blood pressure) is another important cause. Valvular heart disease, cardiomyopathies (disease of heart muscle), and congenital heart disease are less common.

Coronary disease. The coronary arteries carry oxygen and nutrients to the heart muscle. Obstructive disease of these arteries is almost always caused by atherosclerosis, a pathologic process that begins in late adolescence and early adulthood. In its initial stages, the disease produces a fatty deposition and thickening of the inner surface of the artery. This thickening gradually progresses to the atherosclerotic narrowing, eventually restricting the flow of blood through the artery. The major clinical manifestations of this disease are not usually seen until the sixth decade of life; men are affected earlier than women. *See* ARTERIOSCLEROSIS.

Because the size or cross-sectional area of the arterial lumen is progressively reduced by growth of the fatty plaque, the capacity to deliver blood to the heart muscle is increasingly limited. During exercise or other hemodynamic stress, the increased oxygen demand of the heart muscle cannot be met by an increase in coronary blood flow. Therefore, heart muscle function declines, and the individual experiences characteristic discomfort in the chest called angina pectoris. Angina pectoris is relieved by rest or medications that improve the balance between myocardial oxygen demands and coronary blood (and oxygen) supply. Electrocardiographic changes, which usually accompany these events, provide a basis for the exercise-stress test that can be used to diagnose coronary artery disease.

Myocardial infarction. Coronary artery disease can be complicated by the abrupt formation of a blood clot (thrombus) at the site of an atherosclerotic plaque. This formation usually occurs as a result of cracks or fissures on or adjacent to the plaque. Such rupture or disruption of plaque results in an irregularity along the inner surface of the artery that promotes formation of a thrombus with a reduction in blood flow through the artery. When the artery is incompletely obstructed, unstable angina or angina at rest may develop. This is a dynamic condition that is complicated by spasms of the involved artery. When the artery becomes completely occluded, a myocardial infarction or heart attack results, causing irreversible injury to the heart muscle (myocardium) supplied by the occluded artery. Complications and deaths are directly related to the size of the infarction. *See* INFARCTION; THROMBOSIS.

The diagnosis of a myocardial infarction is based on a clinical history of prolonged chest pain, serial electrocardiographic changes, and abnormal blood levels of enzymes that are released from damaged heart muscle. Treatment consists of prompt administration of a clot-dissolving drug so as to limit the extent of myocardial injury. The agents are administered with heparin and aspirin, which provide anticoagulant functions. On occasion, mechanical procedures, such as coronary angioplasty or bypass surgery, are used to reestablish flow in the occluded vessel.

During the hours and days after a myocardial infarction, the heart can lose its pumping ability, and consequently heart failure may develop. Other serious complications include destruction of a muscle that supports the mitral valve within the heart or rupture of the weakened wall of the heart. The injured myocardium also becomes susceptible to disorders of electrical rhythm, the most serious disorder being ventricular fibrillation. If this condition is not treated promptly, the dysrhythmia can be fatal. Most other disorders of cardiac rhythm, including atrial fibrillation, are much less dangerous.

Risk factors. Several characteristics or risk factors are associated with a high likelihood that coronary disease will develop. They include cigarette smoking, elevated levels of blood cholesterol, hypertension, and a family history of the disease. Other risk factors include diabetes, a lack of exercise, and obesity. By modifying or eliminating these factors, the risk of developing coronary disease can be reduced.

Cholesterol. Total blood cholesterol exceeding 240 mg per deciliter is associated with an increased risk of developing coronary disease. Values below 200 mg are desirable. This value represents the sum of all cholesterol carried by a number of distinct lipoproteins. The lipoproteins are combinations of cholesterol, other fats, and proteins. High-density lipoproteins (HDL) appear to provide some protection against coronary disease, while the low-density lipoproteins (LDL) are strongly associated with the development of coronary disease and its complications. However, this low-density fraction of the total cholesterol can be reduced by limiting the dietary intake of saturated fats and cholesterol. Drugs that are effective in reducing low-density lipoproteins include bile sequestrants, nicotine acid, and the inhibitors of cholesterol synthesis. When an elevated blood level of cholesterol is treated, the risk of the development of coronary disease is reduced. *See* CHOLESTEROL; LIPID METABOLISM.

Hypertension. Hypertension or high blood pressure is defined as a pressure exceeding 160/100 mmHg;

values below 140/90 mmHg are normal. A variety of vascular and neuroendocrine abnormalities contribute to the complications of hypertension, including stroke and renal failure. After a prolonged asymptomatic period, heart failure, coronary events, and stroke lead to disability or death. High blood pressure can be reduced by limiting dietary sodium and by weight loss. Several drugs, including diuretics, beta blockers, calcium channel blockers, and angiotensin converting enzyme inhibitors, are effective antihypertension agents. Lowering the blood pressure of hypertensive individuals reduces the incidence of heart failure and stroke and the probability of premature death. *See* HYPERTENSION.

Diagnosis. The diagnosis of coronary disease is often based on stress tests that use electrocardiographic monitoring during exercise. In addition, some diagnostic laboratories use an intravenous injection of radioactive tracers to study the distribution of myocardial blood flow during exercise. Other laboratories use ultrasonography (an echocardiogram) to detect changes in myocardial function during the hemodynamic stress of exercise. Unfortunately, the accuracy of such tests is imperfect, and further evaluation is necessary in some individuals, especially those with atypical symptoms. The most reliable procedure for detecting coronary disease and evaluating its severity is coronary arteriography. This procedure involves injection of a small amount of radiopaque dye into the coronary arteries (the dye is delivered through a small catheter that is inserted into an artery in the groin). A radiographic motion picture is made as the dye circulates through the artery, and atherosclerotic narrowings are identified and their severity is evaluated. *See* ELECTRODIAGNOSIS.

Treatment. The treatment of coronary disease can be considered in three stages. First, the well-known risk factors should be eliminated or modified: smoking must be discontinued, elevated blood cholesterol levels should be reduced, and high blood pressure should be treated. Second, attempts should be made to improve the balance between myocardial oxygen demand and supply with medications; nitroglycerin, beta blockers, and calcium channel blockers are effective therapeutic agents. Third, invasive procedures, such as angioplasty or surgery, may be necessary in individuals whose disease is too serious for medical therapy.

Coronary angioplasty is a procedure that reestablishes flow through a diseased segment of a coronary artery. In a manner similar to that used in coronary angiography, a balloon-tipped catheter is inserted at the groin and is directed to the diseased coronary artery. Under fluoroscopic visualization, the balloon is inflated, the atherosclerotic obstruction is stretched, and the lumen of the vessel is opened (**Fig. 1**). In most cases this procedure is effective in relieving symptoms, but on occasion the procedure damages the artery and urgent bypass surgery becomes necessary. Symptoms recur in as many as one-third of the individuals, and many of them require repeat angioplasty.

Coronary artery bypass surgery involves the placement of a vein (or artery) graft to bypass the diseased segment of a coronary artery (**Fig. 2**). This is a major thoracic surgical procedure that involves the use of a heart-lung machine. The risk of death or serious complications with this procedure is about 2–3%. Most individuals experience complete relief of symptoms; moreover, the procedure can prolong life in

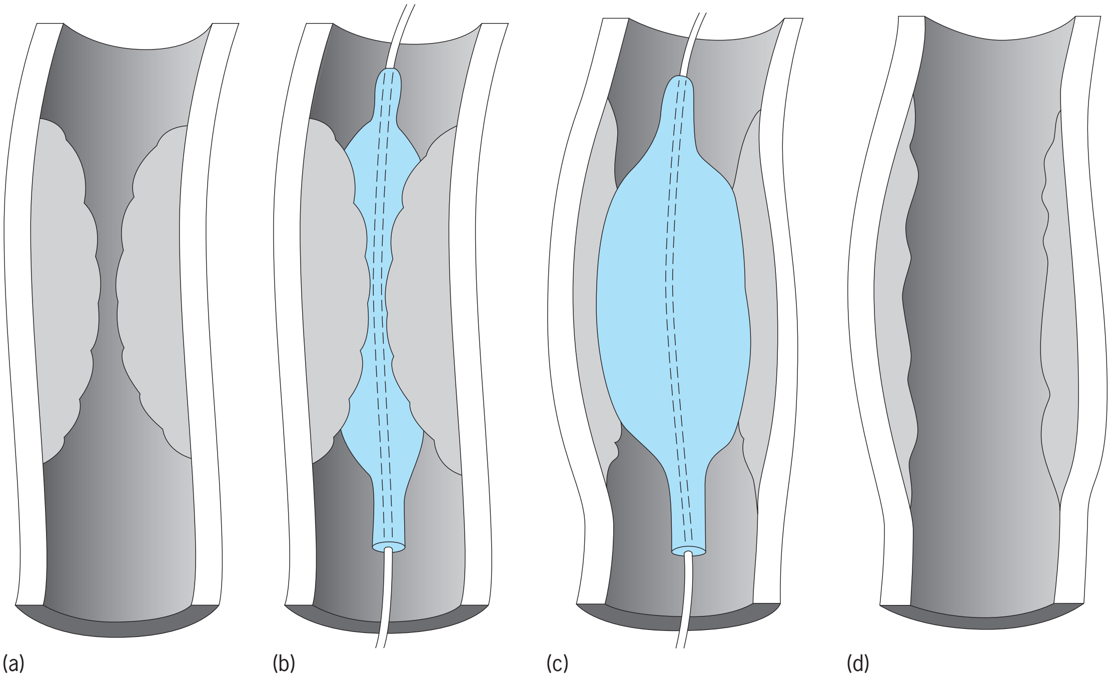


Fig. 1. Percutaneous transluminal coronary angioplasty. (*a*) Cholesterol deposits in the wall of the artery cause a partial obstruction. (*b*) A balloon-tipped catheter is positioned in the narrowed section of the diseased artery. (*c*) Inflation of the balloon compresses the atherosclerotic obstruction. (*d*) After the lumen is restored, the balloon is deflated and removed.

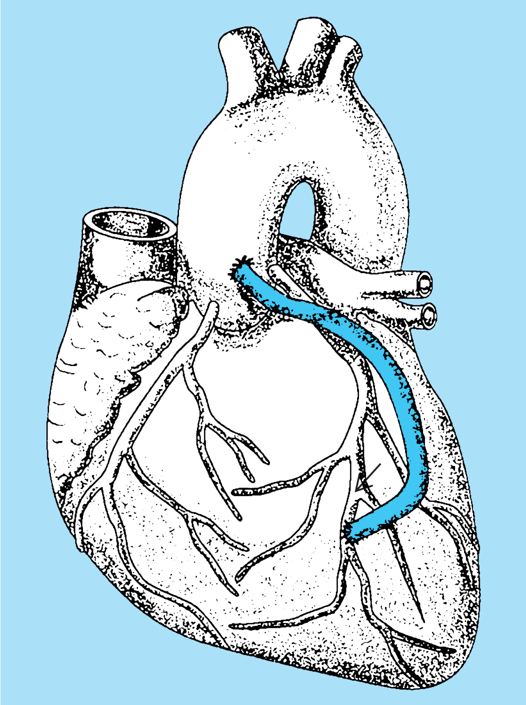


Fig. 2. Coronary artery bypass graft (shaded vessel). This graft provides blood flow to a region that is underperfused by a diseased or obstructed artery (arrow).

selected cases with severe disease. Unfortunately, the atherosclerotic process is not arrested, and therefore risk-factor modification remains important after operation. If symptoms recur, medical treatment and invasive procedures may be repeated.

Hypertensive heart disease. High blood pressure results in an excessive cardiac work load. The heart responds to the high pressure through a growth process called hypertrophy. As a result, the mass of the heart muscle is increased, and the work requirement of each unit of heart muscle returns toward normal. Unfortunately, this compensatory increase in cardiac mass eventually results in heart failure and other morbid cardiovascular events. Antihypertensive drugs are effective in the treatment of hypertension and its consequences.

Valvular heart disease. Heart valves ensure an efficient forward blood flow through the heart. Disease of these valves can be caused by rheumatic fever or degenerative noninflammatory processes. Mitral valve prolapse is a common cause of mitral regurgitation. In this condition, a portion of the blood pumped by the left ventricle leaks backward into the left atrium and causes increased pressure in the atrium and pulmonary veins and shortness of breath. Rheumatic mitral stenosis produces similar symptoms, but the cause is a narrowed valve that obstructs blood flow. A balloon valvuloplasty procedure is used to open a stenotic mitral valve, but surgical repair or replacement with an artificial valve is necessary in mitral regurgitation.

When the aortic valve is incompetent, a portion of the ejected blood leaks backward from the aorta to the left ventricle and causes enlargement of the chamber. By contrast, in aortic stenosis a narrowed valve obstructs blood flow through the valve. Thus, the ventricle is exposed to pressure overload, and it hypertrophies in a fashion similar to that in hypertensive heart disease. Surgical replacement of the valve not only relieves symptoms but can be life-saving. Unfortunately, artificial valves can malfunction or form blood clots, and on occasion additional surgery is necessary.

Congenital heart disease. About 1% of all newborns have a structural abnormality of the heart or adjacent blood vessels. Such congenital heart disease can be caused by maternal rubella, or it may be inherited, as in Down syndrome. However, the cause of most congenital heart disease is unknown.

Defects in the walls between the atria and the ventricles are the most common congenital malformations. In an atrial septal defect, a hole is present in the wall between the right and left atria. In a ventricular septal defect, a hole connects the right and left ventricles. Both conditions cause a shunting of blood from the left to the right heart chambers; thus, oxygenated blood is recirculated through the right heart and lungs. A blood shunt between the pulmonary artery and the aorta also causes a recirculation of oxygenated blood through the lungs; this defect occurs when there is an abnormal persistence of the fetal connection between the two vessels. These and a variety of other more complex abnormalities that occur alone or in combination can be cured or reduced by modern surgical techniques. *See* CONGENITAL ANOMALIES; DOWN SYNDROME.

Cardiomyopathies. Heart muscle disease that is caused by excessive intake of alcohol, some drugs, or infections may cause depression of myocardial function and cardiac enlargement. The resulting dilated cardiomyopathy causes shortness of breath and fatigue. Less commonly, inappropriate growth of the myocardium has occurred in the absence of enlargement of the chamber. This hypertrophic cardiomyopathy causes anginalike chest pain, shortness of breath, and fainting spells. If medical treatment is unsuccessful, surgical removal of a small portion of the myocardium can benefit selected individuals.

Heart failure. Congestive heart failure is a clinical syndrome that consists of shortness of breath, fatigue, and retention of fluid. It is caused by failure of the heart as a pump; thus, heart failure can be caused by almost any form of heart disease. Medical treatment includes digitalis to strengthen the heart muscle, diuretics to reduce fluid retention, and vasodilators to reduce the work load of the heart.

Cardiac transplantation can be performed in individuals whose heart failure does not respond to medical therapy. A normal donor heart (obtained from a victim of irreversible brain injury) is used to replace the diseased heart. After the operation, medications are given to prevent rejection of the transplanted heart. Unfortunately, a shortage of donor organs remains a major problem, and the use of mechanical hearts is limited by the formation of blood clots, infection, and a variety of engineering problems. *See* TRANSPLANTATION BIOLOGY.

Sudden cardiac death. Most cases of sudden death result from inadequate pumping and low cardiac output during a rapid cardiac dysrhythmia, such as ventricular tachycardia or ventricular fibrillation. Fortunately, antiarrhythmic drugs, implantable cardiac defibrillators, or surgical procedures can control the abnormal rhythms in selected individuals. *See* HEART (VERTEBRATE).

William H. Gaasch; Ferdinand J. Venditti, Jr.

Bibliography. E. Braunwald (ed.), *Heart Disease: A Textbook of Cardiovascular Medicine*, 5th ed., 1997; R. C. Schlant and R. W. Alexander (eds.), *Hurst's The Heart*, 9th ed., 1998.

Heartwater disease

A rickettsial disease, also known as cowdriosis, which is caused by the microorganism *Cowdria ruminantium* and is transmitted by ticks of the genus *Amblyomma*. The disease occurs in wild and domestic ruminants, primarily cattle, sheep, and goats, in sub-Saharan Africa and some Caribbean islands (for example, Guadeloupe, Antigua, and Marie-Galante); it is a major obstacle to improvement of livestock production in Africa.

Heartwater disease is characterized by fluid in the pericardium of the heart, high fever, lung edema, and nervous symptoms that range from mild incoordination and exaggerated reflexes to convulsions seen in acute infections. The course of acute heartwater disease is 2–6 days, and recovery is rare. However, young animals have a high rate of natural resistance.

Macrophages and the endothelial cells that line blood vessels, predominantly those in the brain, become infected with the rickettsia. The organisms develop within a membrane-bound vacuole, forming large inclusions. Within the inclusions, *C. ruminantium* divides by binary fission and subsequently develops into denser elementary bodies that are infective for other cells. In the tick vector, transmission of the rickettsia occurs while nymphal and adult ticks feed. The developmental cycle in ticks is similar to the one in endothelial cells, and involves replication of *C. ruminantium* within inclusions in gut and salivary gland cells.

Diagnosis of heartwater is by clinical symptoms and by demonstration of inclusions of *C. ruminantium* in brain tissue. Treatment is difficult because the course of the disease is rapid and must be given on the basis of clinical signs without waiting for laboratory confirmation. The organism is susceptible to tetracycline antibiotics. However, once marked nervous symptoms have developed, recovery usually does not occur. *See* ANTIBIOTIC.

Control and prevention of heartwater is achieved by tick control or immunization. Ticks are controlled by using acaracide dips or pour-ons to provide tick-free stock. An alternative program is to allow young animals to be exposed naturally to *C. ruminantium* by ticks. However, this approach requires enzootic stability of heartwater in which sufficient numbers of naturally infected ticks and animals are present in the field for continual exposure. A heartwater vaccine (available only in the Republic of South Africa) involves controlled exposure and infection of calves with frozen sheep blood infected with *C. ruminantium*. If a clinical reaction or temperature occurs, vaccinated cattle are treated with tetracyclines.

Katherine M. Kocan

Bibliography. J. L. Howard, *Current Veterinary Therapy: Food Animal Practice*, 4th ed., 1999; Z. Woldehiwet and M. Ristic (eds.), *Rickettsial and Chlamydial Diseases of Domestic Animals*, 1993.

Heartworms

Heartworm (*Dirofilaria immitus*) is a nematode parasite that resides within the host's large pulmonary arteries and right heart chambers. It primarily infests dogs but may also infest foxes, wolves, coyotes, ferrets, sea lions, horses, and cats.

A dog can be infested with one to several hundred adult heartworms, which can grow to 12 in. (30 cm). During their 3–5-year life-span, heartworms can cause serious and often life-threatening damage to the heart and lungs.

Initially a problem in southeastern coastal areas, heartworm infestation has spread rapidly throughout the United States in the past two decades. Endemic areas require a reservoir of infected animals (usually dogs) and the presence of mosquitoes, the intermediate host, which transmit the larval stages to a new host.

Life cycle. Heartworms are transmitted from dog to dog by several species of mosquitoes. Adult female heartworms release their microscopic offspring called microfilariae into the bloodstream. A mosquito becomes infested with these circulating microfilariae while taking a blood meal from the dog. The microfilariae develop into mature larvae within the mosquito during the next 10–14 days. As the mosquito feeds again, the mature larvae are injected into the new host. Once in the dog, it takes approximately 6 months for these larvae to complete the cycle by migrating to the large arteries of the lung and right chambers of the heart.

Pathology. Adult heartworms stimulate a progressive proliferation of the artery lining (endarteritis) that gradually restricts the blood flow to the lungs. The resulting increase in the pulmonary artery blood pressure (pulmonary hypertension) causes the right ventricle to pump harder, eventually leading to right heart failure. In advanced cases the lung fibrosis and heart changes may be permanent.

When adult worms die either naturally or with treatment, their fragments become lodged distally in the smaller pulmonary arteries causing an exaggerated proliferation of the vessel lining, the formation of blood clots, and an intense local inflammatory reaction. Blood flow is severely restricted or totally blocked, resulting in severe coughing, coughing up blood, and difficulty in breathing (dyspnea).

Symptoms of infestation. The clinical symptoms of heartworm disease (dirofilariasis) are proportional to the number of infecting heartworms, the duration of the infection, and the host's response. Dogs with only a few adult worms or early in the course of the disease usually show no outward symptoms. Initial signs of disease include coughing, exercise intolerance, and weight loss. As the disease progresses, these symptoms become more pronounced. With advanced disease, dogs begin to exhibit progressive signs of pulmonary disease and associated heart failure, including fainting spells, collapse, difficulty in breathing, coughing up blood, and fluid accumulation around the lungs (hydrothorax) or within the abdominal cavity (ascites).

Rarely, a rapidly fatal condition called vena caval syndrome may be observed in young dogs with massive heartworm infestation. A large number of heartworms block the right heart valves and occlude the large vein returning to the heart (vena cava), severely compromising cardiac, liver, and kidney function. A dog with vena caval syndrome will suddenly stop eating, become listless, very weak, and pale, and may become jaundiced (icteric). An enlarged jugular vein is common. The urine is dark brown (hemoglobinuria). Without prompt surgical removal of the worm burden, most of these dogs will die within 24–72 h.

Diagnosis. Heartworm infection can usually be determined by examining a blood sample for the presence of circulating microfilariae. Unfortunately, circulating microfilariae are found in only 30–60% of infected dogs and in less than 10% of infected cats. Those infestations in which the adults produce no circulating microfilariae are termed occult infections. An occult heartworm infestation can be accurately diagnosed by identifying specific circulating immunologic substances (uterine antigens) released into the blood by adult females.

Treatment. Managed properly, all but the most advanced cases of heartworm disease can usually be treated successfully. There are six steps in heartworm treatment.

1. The severity of the infestation is assessed, and the degree of pathologic disease present is staged. A thorough medical history, a thorough physical examination, laboratory evaluation of the liver and kidneys, urine analysis, chest radiographs, and an electrocardiogram complete the evaluation. Dogs with advanced chronic heartworm disease, especially those with heart failure, are at greatest risk of complications with the adulticide therapy. Stabilizing the lung and cardiac function is strongly advised prior to initiating therapy to eliminate the adult worms. In these cases, the prospect of complete recovery is poor despite successfully killing all the adults.

2. The adult worms are killed with an adulticide drug. The treatment involves a series of injections with an arsenical compound. Serious side effects are possible; therefore the dog should be hospitalized and closely monitored for potential adverse reactions.

3. Exercise is restricted for 4–6 weeks following administration of the adulticide to minimize the potential for all the dead worms becoming lodged in the smaller pulmonary arteries at one time.

4. Medication is administered to eliminate the microfilariae, if present, once the threat of embolic pneumonia has passed. The microfilariae can cause damage to the dog's kidney (for example, glomerular nephritis). They are sources of infestation for other dogs in the area as well.

5. Once the microfilariae have been eliminated, daily or monthly preventive medication is initiated.

6. The antigen-based (occult) adult heartworm test is repeated to determine the success of the adulticide therapy.

Prevention. Daily or monthly heartworm preventive medication is strongly recommended, especially during the mosquito season, in infested areas. The objective is killing the infectious larval stages before they develop into adults. Only dogs that test negative for heartworms should be placed on preventive medication because of the risk of serious reactions.

William D. Fortney

Bibliography. P. A. Aiello and S. E. Aiello (eds.), *The Merck Veterinary Manual*, Merck & Company, 1998; American Heartworm Society, *Heartworm Disease in Dogs*, 1995; S. J. Ettinger and E. C. Feldman (eds.), *Textbook of Internal Medicine: Disease of the Dog and Cat*, W. B. Saunders, 1995.

Heat

For the purposes of thermodynamics, it is convenient to define all energy while in transit, but unassociated with matter, as either heat or work. Heat is that form of energy in transit due to a temperature difference between the source from which the energy is coming and the sink toward which the energy is going. The energy is not called heat before it starts to flow or after it has ceased to flow. A hot object does contain energy, but calling this energy heat as it resides in the hot object can lead to widespread confusion. *See* ENERGY; INTERNAL ENERGY.

Heat flow is a result of a potential difference between the source and sink which is called temperature. Work is energy in transit as a result of a difference in any other potential such as height. Work may be thought of as that which can be completely used for lifting weights. Heat differs from work, the other type of energy in transit, in that its conversion to work is limited by the fundamental second law of thermodynamics, or Carnot efficiency. This natural law is that the fraction of the heat Q convertible to work is determined by the relation $dW = Q(dT/T)$ for processes where the source and sink are differentially different in temperature, or by the relation $dW = dQ(T_1 - T_2)/T_1$ where the source (at T_1) and the sink (at T_2) differ by a finite temperature interval. *See* WORK.

For the above relations to be valid, temperature must be expressed on a thermodynamic temperature scale. Conversely, any temperature scale for which the above relations are valid, irrespective of the substance or material under investigation, is a thermodynamic temperature scale. The perfect gas law defines a scale in which the temperature is proportional to the thermodynamic temperature. In order to make the two scales be identical, the triple point of water (temperature and pressure at which ice, water, and vapor are in equilibrium) is defined to be at 273.16 kelvins on both the ideal-gas and the thermodynamic scales. *See* TEMPERATURE; THERMODYNAMIC PRINCIPLES.

Harold C. Weber; William A. Steele

Bibliography. C. O. Bennett and J. E. Myers, *Momentum, Heat, and Mass Transfer*, 3d ed., 1982; K. S. Pitzer,*Thermodynamics*, 3d ed., 1995.

Heat balance

An application of the first law of thermodynamics to a process in which any work terms are negligible.

Closed systems. For a closed system, one that always consists of the same material, the first law is $Q + W = \Delta E$, where Q is the heat supplied to the system, W is the work done on the system, and ΔE is the increase in energy of the material forming the system. It is convenient to treat ΔE as the sum of changes in mechanical energy, such as kinetic energy and potential energy in a gravitational field, and of internal energy ΔU that depends on changes in the thermodynamic state of the material. (Scientific disciplines do not all use the same sign convention for work, so care is needed when consulting text books.)

The net heat input Q may have positive and negative components. For example, a satellite in space has no work interaction with its surroundings. It receives heat by radiation from the Sun on some surfaces (positive Q) and loses heat by radiation to space from others (negative Q). If the net heat input is positive, the internal energy of the satellite increases and its temperature rises; if the heat inputs and outputs are in balance, the internal energy of the satellite does not change and its temperature remains constant. Because the rates at which any changes occur are usually of interest, heat balances are often written in terms of heat flow rates (heat per unit time), sometimes denoted by a dot over the symbol, $\dot{Q}$, so that for a process with negligible work, kinetic energy and potential energy terms $\dot{Q} = \dot{Q}_{\mathrm{IN}} - \dot{Q}_{\mathrm{OUT}} = dU/dt$, the rate of change of internal energy with time.

The thermodynamic distinctions in terminology between heat, work, and energy are not always applied rigorously. For example, in a heat balance on an electric light bulb, the electrical power is strictly a rate of doing work on the system but might be treated as an electrical heat input, to be balanced against the rate of heat loss by radiation and convective cooling of the bulb by the surrounding air.

Open systems. Often it is more convenient to apply the first law or a heat balance to an open system, a fixed region or control volume across the boundaries of which materials may travel and inside which they may accumulate, such as a building, an aircraft engine, or a section of a chemical process plant. Then the first law is expressed by the equation below,

$$\dot{Q} + \dot{W}_S = \sum \dot{m}\,(h + c^2/2 + gz)_{\mathrm{OUT}} - \sum \dot{m}\,(h + c^2/2 + gz)_{\mathrm{IN}} + dE/dt$$

where $\dot{W}_S$ is the rate of doing shaft work on the system; $\dot{m}$ is the mass flow rate of any stream entering or leaving the control volume; h is the enthalpy per unit mass; c is the velocity; gz is the gravitational potential for each stream at the point of crossing the boundary of the control volume; and E is the energy of all material inside the control volume. When conditions inside the control volume do not change with time, although they need not be spatially uniform, $dE/dt = 0$, and the balance equation is known as the steady-flow energy equation. There is also a mass balance equation: the total incoming and outgoing mass fluxes must be equal if material cannot accumulate inside the control volume.

Enthalpy is a thermodynamic property defined by $h = u + pv$, where u is the specific internal energy (enthalpy per unit mass), p the pressure, and v the specific volume. It is used, along with shaft work $\dot{W}_S$, because the derivation of the first-law equation for a control volume from the more fundamental equation for a closed system involves work terms pv that are not available for use outside the control volume. Changes in enthalpy occur because of changes in temperature, pressure, physical state (for example, from liquid to vapor), and changes in chemical state.

For example, for a control volume around a domestic heating boiler, the entering material streams are the fuel and the air needed for combustion and the incoming flow of cold water; the outgoing streams are the combustion products and the hot water or steam. The only heat flow would be a small term for heat losses from the hot casing of the boiler to the surroundings. There might be a small electrical work input to drive a circulating pump for the water. The main terms in this heat balance would be the changes in enthalpy. The enthalpy change from fuel to combustion products is negative and is related to, but not quite equal to, the calorific value of the fuel. *See* ENTHALPY; THERMODYNAMIC PRINCIPLES; THERMODYNAMIC PROCESSES.

D. B. R. Kenning

Bibliography. Y. A. Cengel and M. A. Boles, *Thermodynamics: An Engineering Approach*, 3d ed., McGraw-Hill, 1998; M. C. Potter and C. W. Somerton, *Engineering Thermodynamics*, McGraw-Hill, 1995; K. Wark, Jr., *Advanced Thermodynamics for Engineers*, McGraw-Hill, 1994; P. B. Whalley, *Basic Engineering Thermodynamics*, Oxford University Press, 1992.

Heat balance, terrestrial atmospheric

The balance of various types of energy in the atmosphere and at the Earth's surface. Essentially all the energy that the Earth–atmosphere system receives

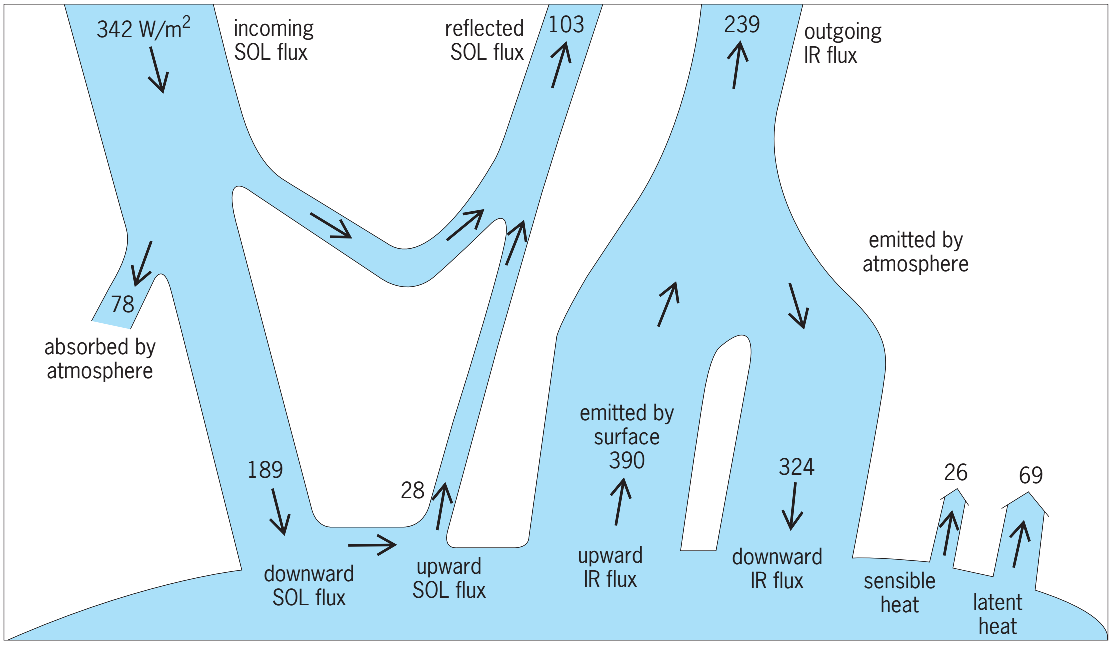


Fig. 1. Heat balance of the Earth and the atmosphere system. The solar (SOL) constant used in the calculations is 1366 W/m^2 so that the effective incoming solar flux is 342 W/m^2, while the global albedo used is 30%. The atmosphere given in the figure contains molecules, aerosols, and clouds. The atmospheric thermal infrared (IR) flux is emitted both upward and downward. The surface albedo and surface temperature used are 15% and 288 K, respectively. The width of the shaded area with an arrow is approximately proportional to the flux value. (*Data taken from K. N. Liou, 2002*)

comes from the Sun. This energy is conventionally referred to as the solar constant, and is defined as the flux of solar energy (energy per time) available on 1 square meter facing the Sun at the top of the atmosphere when the Earth is at its mean distance from the Sun. On the basis of recent satellite observations, a value of about 1336 watts per square meter (W/m^2) has been suggested. Because the area of the spherical Earth is four times that of its cross section facing the parallel solar beam, the top of the Earth's atmosphere receives an average of about 342 W/m^2. Based on analysis of the observed data from satellite radiation budget experiments in the last 40 years, about 30% of this is reflected back to space and is referred to as the global albedo. The reflecting power of the Earth–atmosphere system includes the scattering of molecules, aerosols, and clouds, as well as reflection of different types of surfaces. As a consequence of this global albedo, only about 70% of the incoming solar flux (that is, about 239 W/m^2) is available on average to warm the Earth–atmosphere system. For this system to be in thermodynamic equilibrium or balance so that an equilibrium temperature can be defined, it must radiate the same amount of energy (239 W/m^2) back to space. The emitted (or outgoing) terrestrial radiation from the Earth and the atmosphere having an equilibrium temperature of about 254 K (−2.5°F) is in the infrared portion of the electromagnetic spectrum and is called the thermal infrared radiation or longwave radiation. This is differentiated from the solar radiation or shortwave radiation from the Sun, which has an effective temperature of about 5800 K (10,000°F). *See* ALBEDO; ATMOSPHERE; HEAT BALANCE; SOLAR CONSTANT; SOLAR ENERGY; TERRESTRIAL RADIATION.

On global average and over a climatological period of say 40 years, the incoming solar flux of 342 W/m^2 must be balanced by the sum of the reflected solar flux of 103 W/m^2 and the emitted infrared flux of 239 W/m^2 at the top of the atmosphere, as displayed in **Fig. 1**. This is referred to as the heat balance of the Earth–atmosphere system.

Determination of the radiative heat-balance components in the atmosphere and at the surface requires theoretical calculation and analysis. Climatological profiles of temperature, cloud, aerosol, water vapor, ozone, molecular make up, and pertinent absorbing gases are needed in radiative transfer calculations. Also required are the globally averaged surface albedo, estimated to be about 15%, and the climatological surface temperature of 288 K (59°F). Using these parameters, the solar flux reaching the surface is found to be about 189 W/m^2, while the reflected component is 28 W/m^2. Thus, the absorbed solar flux at the surface is 161 W/m^2, which amounts to about 47% of the incoming solar flux. As a consequence, the atmosphere, including clouds, aerosols, ozone, air molecules, and other minor gases, absorbs 78 W/m^2, or about 23% of the incoming solar flux.

On the basis of theoretical calculations, the emitted downward infrared flux from the atmosphere reaching the surface is 324 W/m^2. The climatological surface temperature of 288 K emits 390 W/m^2 of infrared flux, leading to a net loss of the thermal infrared flux by the Earth's surface of 66 W/m^2. Consequently, the radiative heat balance at the Earth's surface is the sum of the gain due to the absorption

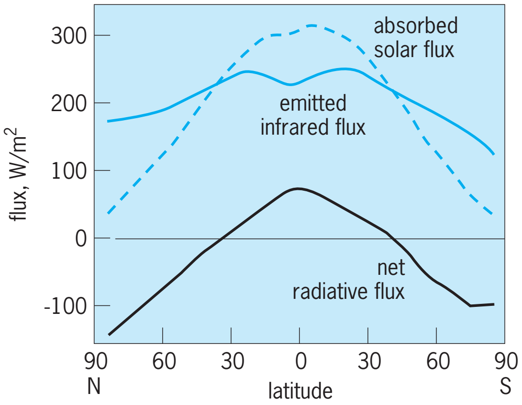


Fig. 2. Components of the annual mean absorbed solar flux, emitted infrared flux, and net radiative flux as functions of latitude at the top of the atmosphere derived from satellite broadband radiation measurements. (*Data taken from K. N. Liou, 2002*)

of solar flux and the net emission loss associated with thermal infrared flux, resulting in a positive gain of 95 W/m^2. This gain is countered by the transport of fluxes of sensible heat (related to potential temperature) and latent heat (related to water substances) out of the surface in order to maintain an overall heat balance. These fluxes are 26 W/m^2 and 69 W/m^2, respectively, based on the so-called Bowen ratio, which has a global value of 0.27. The surface heat balance is illustrated in Fig. 1.

On global average, the incoming solar flux absorbed by the Earth-atmosphere system is balanced by the emitted infrared flux, as presented above. However, substantial local variability occurs. **Figure 2** illustrates the latitudinal distribution of solar, infrared, and net radiative flux at the top of the atmosphere. The net radiative flux pattern depicts a maximum gain of radiative heat at the Equator associated with a minimum of emitted infrared flux due to the colder towering cumulus clouds frequently occurring in this region and the large amount of solar flux absorbed there. In the polar regions, we see large negative net fluxes associated with the high albedo of snow and ice as well as the sharp decrease in solar insolation during the winter season. This configuration clearly demonstrates the gain and loss of radiative energy in the tropical and subtropical areas and the polar regions, respectively, a pattern that establishes a temperature gradient from the Equator to the Poles. Consequently, thermally driven circulations are produced either directly or indirectly by horizontally heating and cooling gradients in the atmosphere. Thus, there is a substantial transport of sensible and latent heat by large-scale atmospheric circulations in addition to their upward transport from the surface to the atmosphere by convective activities, as shown in Fig. 1. *See* ATMOSPHERIC GENERAL CIRCULATION; INSOLATION.

The atmospheric latent heat component has maxima in the subtropics between about 20° and 30° latitude in both hemispheres, where evaporation rates are highest. Minimum patterns are evident between about 10°N and 10°S because of excess precipitation in the tropical regions, and between about 40° and 60° in both hemispheres because of surplus precipitation associated with storm activity. The atmospheric sensible heat component shows a large maximum in the tropics associated with higher temperatures and two small maxima located at about 40°N and 40°S related to the transport of heat by circulations. On an annual basis and over a climatological time period, about 20 W/m^2 of sensible and latent heat is lost between 40°S and 40°N, while about 50 to 70 W/m^2 is gained poleward of 60°. In addition to the transport of sensible and latent heat fluxes by atmospheric circulations, the ocean currents can transport heat. Best estimates show that about 40 W/m^2 is moved out of the tropics and about 25 W/m^2 is transported into latitudes poleward of 40° by ocean currents, but this transport is terminated at about 70°S where there are no oceans. The sum of the heat flux transport by the atmosphere and the ocean must be balanced by the net radiative flux at the top of the atmosphere on an annual basis. K. N. Liou

Bibliography. K. N. Liou, *An Introduction to Atmospheric Radiation*, 2d ed., Academic Press, 2002.

Heat capacity

The quantity of heat required to raise a unit mass of homogeneous material one unit in temperature along a specified path, provided that during the process no phase or chemical changes occur is known as the heat capacity of the material in question. The unit mass may be 1 g, 1 lb, or 1 gram-molecular weight (1 mole). Moreover, the path is so restricted that the only work effects are those necessarily done on the surroundings to cause the change to conform to the specified path. The path, except as noted later, is at either constant pressure or constant volume. This definition conforms to an average heat capacity for the chosen unit change in temperature.

Instantaneous heat capacity at a particular temperature is defined as the rate of heat addition relative to the temperature change at the temperature in question; that is, on a plot of heat addition Q as a function of temperature T, instantaneous heat capacity is given by the slope of the curve at the temperature in question. Units of heat capacity are energy units per unit mass of material per unit change in temperature.

In accordance with the first law of thermodynamics, heat capacity at constant pressure C_p is equal to the rate of change of enthalpy with temperature at constant pressure $(\partial H/\partial T)_p$. Heat capacity at constant volume C_v is the rate of change of internal energy with temperature at constant volume $(\partial U/\partial T)_v$. Moreover, for any material, the first law yields the relation in Eq. (1).

$$C_p - C_v = \left[P + \left(\frac{\partial U}{\partial V}\right)\right]_T \left(\frac{\partial U}{\partial T}\right)_P \qquad (1)$$

See ENTHALPY; INTERNAL ENERGY.

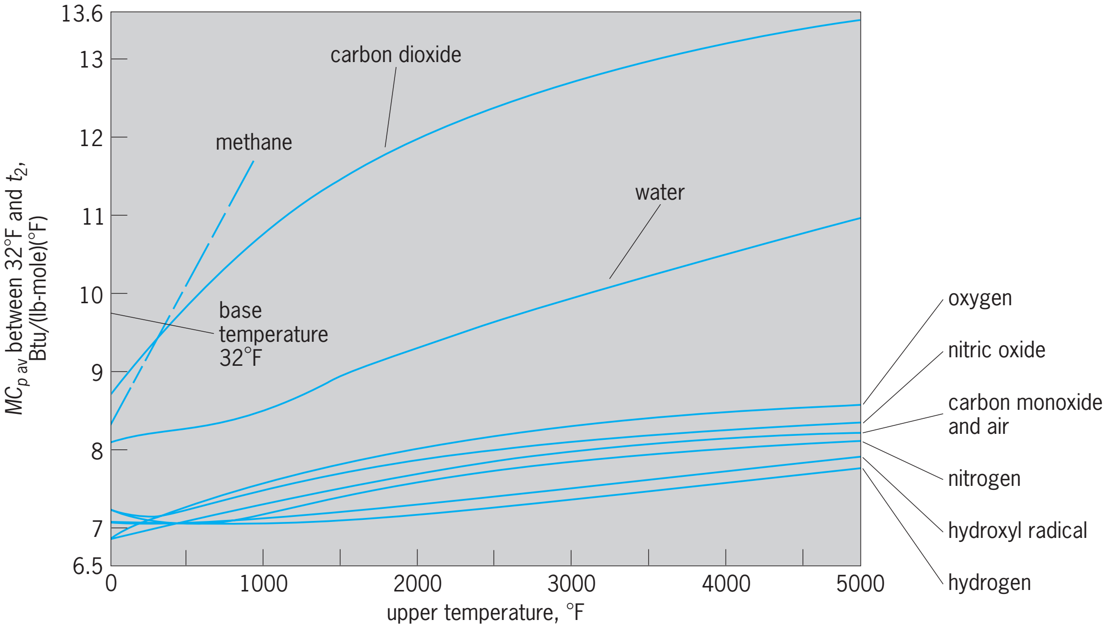


Fig. 1. Variation of average molar heat capacity with temperature for several common gases. t^2 = a given upper temperature. °C = (°F − 32)/1.8.

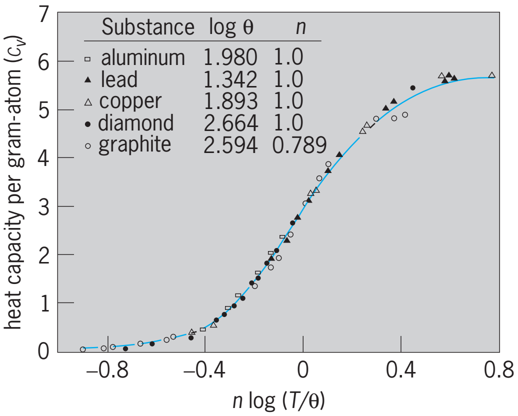


Fig. 2. Atomic heat capacity as a function of temperature (K).

Gases. For one mole of a perfect gas, the preceding relation becomes $MC_p - MC_v = R$, where M is the molecular weight of the gas under discussion and R is the perfect gas law constant. For gases the ratio C_p/C_v is usually designated by the symbol K.

For monatomic gases at moderate pressures, MC_v is about 1.67, and the heat capacity changes but little with temperature. For diatomic gases, MC_v is approximately 5 at 68°F (20°C) and moderate pressures. Change of heat capacity with temperature is usually small. The value of K is between 1.40 and 1.42. For triatomic gases at moderate pressures, MC_v varies from 6 to 7 and changes rapidly with temperature. The value of K varies but is always smaller than that for the less complex molecules at the same conditions of pressure and temperature.

For gases with more than three atoms per molecule, no generalizations are reliable. However, as molecular complexity increases, heat capacity increases, the influence of temperature on heat capacity increases, and K decreases. **Figure 1** shows average MC_p for several common gases. Up to pressures of a few atmospheres, the effect of pressure on heat capacity of gases is small and is usually neglected.

Solids. For solids, the atomic heat capacity (heat capacity when the unit mass under discussion is 1 at. wt) may be closely approximated by equation of type (2), where $n = 1$ for elements of simple crystalline

$$C_v = J\left(\frac{T}{\theta}\right)^n \qquad (2)$$

form, but has a smaller value for those of more complex structures; θ is characteristic of each element; J is a function that is the same for all substances; T is absolute temperature. **Figure 2** compares measured with calculated values.

For all solid elements at room temperature, C_v is about 6.4 calories per gram atom per degree Celsius. This approximation may be used when no experimental data are available, but errors may be considerable, particularly for elements with atomic weights less than 39. Kopp's law states that for solids the molal heat capacity of a compound at room temperature and pressure approximately equals the sum of heat

Heat capacities of some elements

Element	Heat capacity
All heavy elements	6.4
Boron	2.7
Carbon	1.8
Fluorine	5.0
Hydrogen	2.3
Oxygen	4.0
Phosphorus and sulfur	5.4
Silicon	3.5

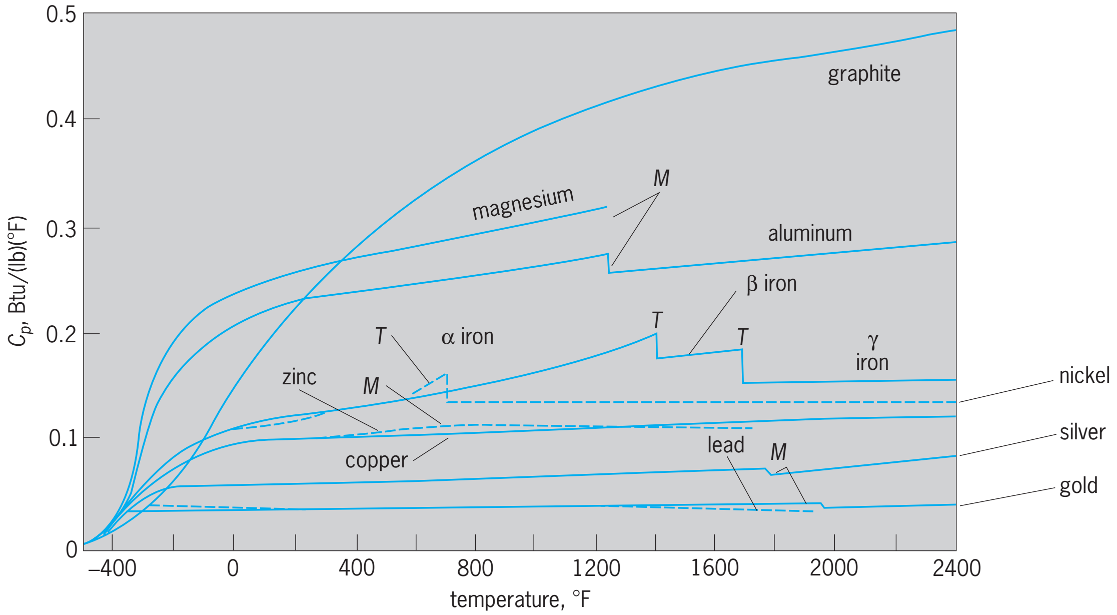


Fig. 3. Change in heat capacity of some industrially important solids with temperature. *M* = melting point; *T* = transition temperature. °C = (°F − 32)/1.8.

capacities of the elements in the compound. Errors are considerable but may be reduced by judicious choice of atomic heat capacities for the lighter elements. Recommended values for some of these are given in the **table** of constants for Kopp's law. Use of Kopp's law is justified only when no experimental data are available.

Figures 3 and **4** give instantaneous heat capacities for some industrially important solids.

Liquids. For liquids and solutions no useful generally applicable approximations are available. For aqueous solutions of inorganic salts the approximate heat capacity of the solution may be estimated by assuming the dissolved salt to have negligible heat capacity. Thus, in a 20% by weight solution of any salt in water 0.8 would be the estimated heat capacity.

Effect of pressure on heat capacities at any temperature may be calculated by the relations in Eqs. (3*a*) and (3*b*).

$$\left(\frac{\partial MC_p}{\partial P}\right)_T = -T\left(\frac{\partial^2 V}{\partial T^2}\right)_P \qquad (3a)$$

$$\left(\frac{\partial MC_v}{\partial V}\right)_T = T\left(\frac{\partial^2 P}{\partial T^2}\right)_V \qquad (3b)$$

Constant temperature. Not so familiar as C_p and C_v are the heat necessary to cause unit change

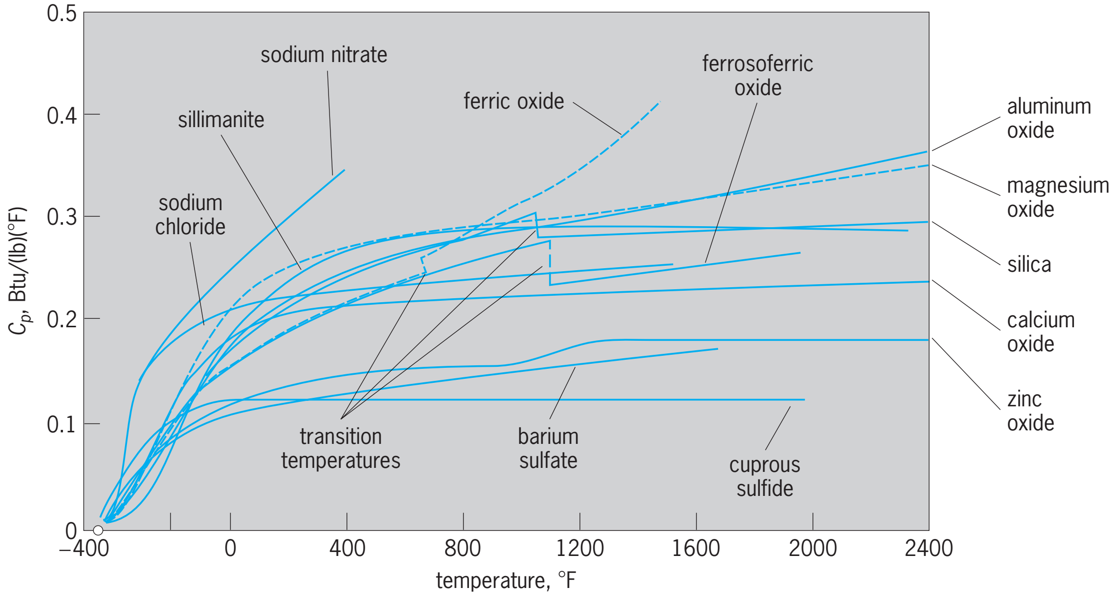


Fig. 4. Change in heat capacity of compounds with temperature. °C = (°F − 32)/1.8.

in pressure in a unit mass of material at constant temperature and the heat required to cause unit change in volume at constant temperature. These are designated $\partial Q_P/\partial P$ and $\partial Q_T/\partial V$. Similarly, $\partial Q_V/\partial P$ and $\partial Q_p/\partial V$ may be called heat capacities. *See* SPECIFIC HEAT; THERMODYNAMIC PRINCIPLES.

Harold C. Weber

Bibliography. I. M. Klotz and R. M. Rosenberg, *Classical Thermodynamics*, 5th ed., 1994; K. C. Rolle, *Thermodynamics and Heat Power*, 5th ed., 1998.

Heat exchanger

A device used to transfer heat from a fluid flowing on one side of a barrier to another fluid (or fluids) flowing on the other side of the barrier.

When used to accomplish simultaneous heat transfer and mass transfer, heat exchangers become special equipment types, often known by other names. When fired directly by a combustion process, they become furnaces, boilers, heaters, tube-still heaters, and engines. If there is a change in phase in one of the flowing fluids—condensation of steam to water, for example—the equipment may be called a chiller, evaporator, sublimator, distillation-column reboiler, still, condenser, or cooler-condenser.

Heat exchangers may be so designed that chemical reactions or energy-generation processes can be done within them. The exchanger then becomes an integral part of the reaction system and may be known as a nuclear reactor or catalytic reactor.

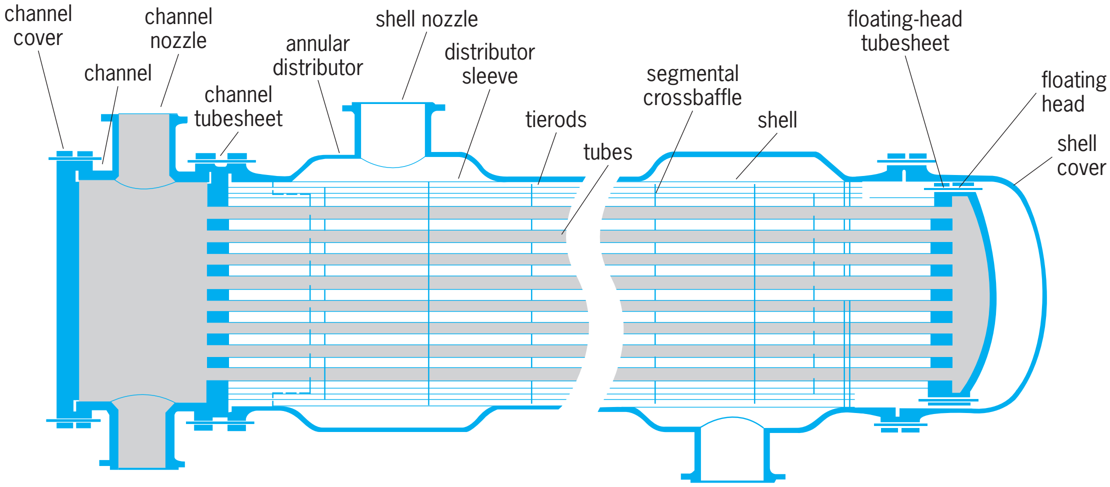
Heat exchangers are normally used only for the transfer and useful elimination or recovery of heat without an accompanying phase change. The fluids on either side of the barrier are usually liquids, but they may also be gases such as steam, air, or hydrocarbon vapors; or they may be liquid metals such as sodium or mercury. Fused salts are also used as heat-exchanger fluids in some applications.

With the development and commercial adoption of large, air-cooled heat exchangers, the simplest example of a heat exchanger would now be a tube within which a hot fluid flows and outside of which air is made to flow for cooling. By similar reasoning, any container of a fluid immersed in any fluid could serve as a heat exchanger if the flow paths were properly connected, or any container of a fluid exposed to air becomes a heat exchanger when a temperature differential exists.

Most often, the barrier between the fluids is a metal wall such as that of a tube or pipe. However, it can be fabricated from flat metal plate or from graphite, plastic, or other corrosion-resistant materials of construction. If the barrier wall is that of a seamless or welded tube, several tubes may be tied together into a tube bundle (see **illus.**) through which one of the fluids flows distributed within the tubes. The other fluid (or fluids) is directed in its flow in the space outside the tubes through various arrangements of passes. This fluid is contained by the heat-exchanger shell. Discharge from the tube bundle is to the head (heads) and channel of the exchanger. Separation of tube-side and shell-side fluids is accomplished by using a tube sheet (tube sheets).

Applications. Heat exchangers find wide application in the chemical process industries, including petroleum refining and petrochemical processing; in the food industry, for example, for pasteurization of milk and canning of processed foods; in the generation of steam for production of power and electricity; in nuclear reaction systems; in aircraft and space vehicles; and in the field of cryogenics for the low-temperature separation of gases. Heat exchangers are the workhorses of the entire field of heating, ventilating, air-conditioning, and refrigeration.

Classifications. The exchanger type described in general terms above and illustrated by the diagram is the well-known shell-and-tube heat exchanger. Shell-and-tube exchangers are the most numerous, but constitute only one of many types. Exchangers in use range from the simple pipe within a pipe—with a few square feet of heat-transfer surface—up to the



Schematic diagram of heat exchanger.

complex-surface exchangers that provide thousands of square feet of heat-transfer area.

In between these extremes is a broad field of shell-and-tube exchangers often specifically named by distinguishing design features; for example, U tube, fin tube, fixed tube sheet, floating head, lantern-ring packed floating head, socket-and-gland packed floating head, split-ring internal floating head, pull-through floating head, nonremovable bundle with floating head or U-tube construction, and bayonet type.

Also, varying pass arrangements and baffle-and-shell alignments add to the multiplicity of available designs. Either the shell-side or tube-side fluids, or both, may be designed to pass through the exchanger several times in concurrent, countercurrent, or cross flow to the other fluids.

The concentric pipe within a pipe (double pipe) serves as a simple but efficient heat exchanger. One fluid flows inside the smaller-diameter pipe, and the other flows, either concurrently or countercurrently, in the annular space between the two pipes, with the wall of the larger-diameter pipe serving as the shell of the exchanger.

To solve new processing problems and to find more economical ways of solving old ones, new types of heat exchangers are being developed. There has been much emphasis on cramming more heat-transfer surface into less and less volume. Extended-surface exchangers, such as those built with fin tubes, are finding wide application.

Water shortage has added a new dimension to heat-exchanger design and has led to use of air-cooled exchangers.

Plate-type heat exchangers, long used in the milk industry for pasteurization and skimming, have moved into the chemical and petroleum industries. Coiled tubular exchangers and coiled-plate heat exchangers are winning new assignments. Spiral exchangers offer short cylindrical shells with flat heads, carrying inlets and outlets leading to internal spiral passages. These passages may be made with spiral plates or with spiral banks of tubes. Exchangers with mechanically scraped surfaces are finding favor for use with very viscous and pastelike materials.

A somewhat unusual type of plate heat exchanger is one in which sheets of 16-gage metal, seam- and spot-welded together, are embossed to form transverse internal channels which carry the heat-transfer medium. This type is often used for immersion heating in electroplating and pickling.

Materials of construction. Every metal seems to be a possible candidate as a material of construction in fabrication of heat exchangers. Most often, carbon steels and alloy steels are used because of the strength they offer, especially when the exchanger is to be operated as a pressure vessel. Because of excellent heat conductance, brass and copper find wide use in exchanger manufacture.

Corrosion plays a key role in the selection of exchanger construction materials. Often, a high-priced material will be selected to contain a corrosive tube-side fluid, with a cheaper material being used on the less corrosive shell side.

For special corrosion problems, exchangers are built from graphite, ceramics, glass, bimetallic tubes, tantalum, aluminum, nickel, bronze, silver, and gold. *See* CORROSION.

Problems of use. Each of the fluids and the barrier walls between them offers a resistance to heat transfer. However, another major resistance that must be considered in design is the formation of dirt and scale deposited on either side of the barrier wall. It may become so great that the exchanger will have to be removed from service periodically for cleaning.

Chemical and mechanical methods may be used to remove the dirt and scale. For mechanical cleaning, the exchanger is removed from service and opened up. Perhaps the entire tube bundle is pulled from the exchanger shell if the plant layout has provided space for this to be done. If the deposit is on the inside of straight tubes, cleaning may be accomplished merely by forcing a long worm or wire brush through each tube.

More labor is required to remove deposits on the shell side. After removal of the tube bundle, special cleaning methods such as sandblasting may be necessary.

Much engineering effort has gone into the design of heat exchangers to allow for fouling. However, it has been suggested that methods are available to design heat exchangers that, by accommodating a certain amount of dirt in a thermal design, will allow heat exchangers to run forever without shutdown for cleaning. Commercial units designed in this fashion are in operation.

Another operating problem is allowance for differential thermal expansion of metallic parts. Most operating difficulties arise during the startup or shutdown of equipment. Therefore, the following general rules have been suggested: (1) Startup. Always introduce the cooler fluid first. Add the hotter fluid slowly until the unit is up to operating conditions. Be sure the entire unit is filled with fluid and there are no pockets or trapped inert gases. Use a bleed valve to remove trapped gases. (2) Shutdown. Shut off the hot fluid first, but do not allow the unit to cool too rapidly. Drain any materials which might freeze or solidify as the exchanger cools. (3) Steam condensate. Always drain any steam condensate from heat exchangers when starting up or shutting down. This reduces the possibility of water hammer caused by steam forcing the trapped water through the lines at high velocities. *See* CONDUCTION (HEAT); CONVECTION (HEAT); COOLING TOWER; DISTILLATION; EVAPORATOR; FURNACE CONSTRUCTION; HEAT RADIATION; HEAT TRANSFER; VAPOR CONDENSER. Raymond F. Fremed

Bibliography. N. Afgan et al. (eds.), *New Developments in Heat Exchangers*, 1996; S. Kakaç and H. Liu, *Heat Exchangers: Selection, Rating and Thermal Design*, 2d ed., 2002; T. Kuppan, *Heat Exchanger Design Handbook*, 2000; R. K. Shah and D. P. Sekulic, *Fundamentals of Heat Exchanger Design*, 2002.

Heat insulation

Materials whose principal purpose is to retard the flow of heat. Thermal- or heat-insulation materials may be divided into two classes, bulk insulations and reflective insulations. The class and the material within a class to be used for a given application depend upon such factors as temperature of operation, ambient conditions, mechanical strength requirements, and economics.

Examples of bulk insulation include mineral wool, vegetable fibers and organic papers, foamed plastics, calcium silicates with asbestos, expanded vermiculite, expanded perlite, cellular glass, silica aerogel, and diatomite and insulating firebrick. They retard the flow of heat, breaking up the heat-flow path by the interposition of many air spaces and in most cases by their opacity to radiant heat.

Reflective insulations are usually aluminum foil or sheets, although occasionally a coated steel sheet, an aluminized paper, or even gold or silver surfaces are used. Refractory metals, such as tantalum, may be used at higher temperatures. Their effectiveness is due to their low emissivity (high reflectivity) of heat radiation. *See* EMISSIVITY.

Thermal insulations are regularly used at temperatures ranging from a few degrees above absolute zero, as in the storage of liquid hydrogen and helium, to above 3000°F (1650°C) in high-temperature furnaces. Temperatures of 4000–5000°F (2200–2760°C) are encountered in the hotter portions of missiles, rockets, and aerospace vehicles. To withstand these temperatures during exposures lasting seconds or minutes, insulation systems are designed that employ radiative, ablative, or absorptive methods of heat dissipation.

Heat flow. The distinguishing property of bulk thermal insulation is low thermal conductivity. Under conditions of steady-state heat flow the empirical equation that describes the heat flow through a material is Eq. (1), where q = time rate of heat flow,

$$\frac{q}{A} = -k\frac{\theta_2 - \theta_1}{l} \tag{1}$$

A = area, θ_1 = temperature of warmer side, θ_2 = temperature of colder side, l = thickness or lenght of heat-flow path, and k = thermal conductivity, representative values being listed in the **table**. For a given thickness of material exposed to a given temperature difference, the rate of heat flow per unit area is directly proportional to the thermal conductivity of the material. *See* CONDUCTION (HEAT).

In the unsteady state, or transient heat flow, the density and specific heat of a material have a strong influence upon the rate of heat flow. In such cases, thermal diffusivity $\alpha = k/\rho\ C_p$ is the important property. Here ρ = density and C_p = specific heat at constant pressure. In the simple case of one-dimensional heat flow through a homogeneous material, the governing equation is Eq. (2), where t =

$$\frac{d\theta}{dt} = \alpha \frac{d^2\theta}{dx^2}\bigg|_0^l \tag{2}$$

time and x is measured along the heat-flow path from 0 to l.

Thermal conductivity. In general, thermal conductivity is not a constant for the material but varies with temperature. Generally, for metals and other crystalline materials, conductivity decreases with increasing temperature; for glasses and other amorphous materials, conductivity increases with temperature. Bulk insulation materials in general behave like amorphous materials and have a positive temperature coefficient of conductivity.

Thermal conductivity of bulk insulation depends upon the nature of the gas in the pores. The conductivities of two insulations, identical except for the gases filling the pore spaces, will differ by an amount approximately proportional to the difference in the conductivities of the two gases.

Thermal conductivities of selected solids[*]

Material	Density, lb/ft^3 (kg/m^3)	Temperatures, °F (°C)	Conductivity (k),[†] Btu/(h)(ft^2)(°F/in.) [W/(m^2)(°C/m)]
Asbestos cement board	120 (1920)	75 (24)	4 [0.6]
Cotton fiber	0.8–2.0 (15–32)	75 (24)	0.26 [0.037]
Mineral wool, fibrous rock, slag, or glass	1.5–4.0 (24–64)	75 (24)	0.27 [0.039]
Insulating board, wood, or cane fiber	15 (240)	75 (24)	0.35 [0.050]
Foamed plastics	1.6 (26)	75 (24)	0.29 [0.042]
Glass			3.6–7.32 [0.52–1.056]
Hardwoods, typical	45 (720)	75 (24)	1.10 [0.159]
Softwoods, typical	32 (510)	75 (24)	0.80 [0.115]
Cellular glass	9 (150)	75 (24)	0.40 [0.058]
Fine sand (4% moisture content)	100 (1600)	40 (4)	4.5 [0.65]
Silty clay loam (20% moisture content)	100 (1600)	40 (4)	9.5 [1.37]
Gypsum or plaster board	50 (800)	75 (24)	1.1 [0.16]

[*] From American Society of Heating, Refrigerating, and Air Conditioning Engineers, *Heating, Ventilating and Air Conditioning Guide*, 1959.

[†] Typical: suitable for engineering calculations.

Increasing the pressure of the gas in the pores of a bulk insulation has little effect on the conductivity even with pressures of several atmospheres. Decreasing the pressure has little effect until the mean free path of the gas is in the order of magnitude of the dimensions of the pores. Below this pressure the conductivity decreases rapidly until it reaches a value determined by radiation and solid conduction. A few materials have such fine pores that at atmospheric pressure their dimensions are smaller than the mean free path of air. Such insulations may have conductivities less than that of still air. *See* HEAT RADIATION; HEAT TRANSFER. Harry F. Remde

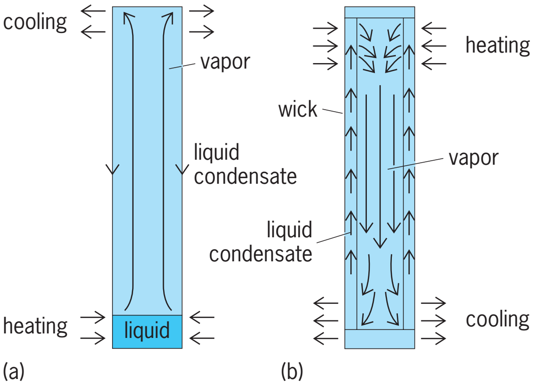
Bibliography. D. Brandreth (ed.), *Improved Thermal Insulation*, 1991; D. E. Croy and D. A. Dougherty, *Handbook of Thermal Insulation Applications*, 1984; R. M. Diamant, *Thermal and Acoustic Insulation*, 1986; F. J. Powell and S. L. Matthews (eds.), *Thermal Insulation: Materials and Systems*, 1987.

Heat pipe

A device for transferring heat efficiently between two locations by using the evaporation and condensation of a fluid contained therein. Originally developed for refrigeration and spacecraft thermionic generator applications, heat pipes (and their close relative, the thermosiphon) have become routine components in systems ranging from missiles to domestic tuner-amplifiers and furnaces. They have many applications in areas where reliable performance and low cost are of prime importance—for example, in electronics and heat exchangers.

Operating principle. The heat pipe, the idea of which was first suggested in 1942, is similar in many respects to the thermosiphon. A large proportion of applications for heat pipes do not use heat pipes as strictly defined below, but employ thermosiphons (**illus.** *a*), sometimes known as gravity-assisted heat pipes. A small quantity of liquid is placed in a tube from which the air is then evacuated, and the tube sealed. The lower end of the tube is heated, causing liquid to vaporize and the vapor to move to the cooler end of the tube, where it condenses. The condensate is returned to the evaporator section by gravity. Since the latent heat of evaporation is generally high, considerable quantities of heat can be transported with a very small temperature difference between the two ends. Thus the structure has a high effective thermal conductance. The thermosiphon, also known as the Perkins tube, has been used for many years. A wide variety of working fluids have been employed, ranging from helium to liquid metals.

One limitation of the basic thermosiphon is that in order for the condensate to be returned by gravitational force to the evaporator region, the latter must be situated at the lowest point. The heat pipe is similar in construction to the thermosiphon, but in this case provision is made for returning the condensate against a gravity head. Commonly a wick, for example a few layers of fine gauze, is used. This is fixed to the inside surface of the tube, and capillary forces return the condensate to the evaporator (illus. *b*). Since the evaporator position is not restricted, the heat pipe may be used in any orientation. If, of course, the heat pipe evaporator happens to be in the lowest position, gravitational forces will assist the capillary force. Alternative techniques, including centripetal forces, osmosis, and others, may be used for returning the condensate to the evaporator (see **table**).



Heat transfer devices. (a) Thermosiphon, (b) Heat pipe. The heat pipe can be in any position, not just vertical as shown.

Capillary forces are by far the most common form of condensate return employed, but a number of rotating heat pipes are used for cooling of electric motors and other rotating machinery. In some applications a mechanical pump is used to return condensate in two-phase run-around coil heat recovery systems. While this may be regarded as a retrograde step, it is a much more effective method for condensate return than reliance on capillary forces.

Applications. The applications for which heat pipes are developed are related to five principal functions of the heat pipe: separation of heat source and sink, temperature flattening, heat flux transformation, temperature control, and acting as a thermal diode or switch. The two major applications, cooling of electronic components and heat exchangers, can involve all of these features. In the case of electronics cooling and temperature control, all features can be important. In heat exchangers employing heat pipes,

Methods of condensate return in heat pipes

Method	Device
Gravity	Standard thermosiphon
Capillary forces	Standard heat pipe
Centripetal forces	Rotating heat pipe Rotary heat pipe
Electrostatic volume forces	Electrohydrodynamic heat pipe
Magnetic volume forces	Magnetohydrodynamic heat pipe
Osmotic forces	Osmotic heat pipe
Vapor bubble pump	Inverse thermosiphon
Mechanical pump	Two-phase run-around coil

the separation of heat source and sink, and the action as a thermal diode or switch, are most significant.

Cooling of electronic components. The use of heat pipes for cooling or temperature control of electronic components is the largest field of application, and a wide variety of heat pipe geometries may also be employed. The basic tubular heat pipe remains the most common type, but flat heat pipes, direct-contact systems, and flexible heat pipes all play an increasingly important role. The most widely applied heat pipes in electronics cooling are units of conventional form that were developed for cooling of semiconductor devices and are manufactured by using a low-cost production method for spirally grooved wicks. By using water as the working fluid in a copper container, heat transport capabilities in excess of 1 kW are achieved. The heat pipes, which normally operate horizontally, typically have a diameter of 0.6 in. (16 mm), and give a 30% improvement in heat dissipation and a 50% weight reduction when compared with a conventional heat sink.

A wide range of audio amplifiers employ heat-pipe heat sink systems of this type. As well as improving heat dissipation, the use of heat pipes also yields electrical benefits, including improvements in the distortion factor. The heat pipes are used to transfer heat to the periphery of the amplifier, where a finned heat sink can dissipate it to atmosphere. This prevents local overheating within the electronics enclosure.

A second major application is in power amplifiers. The heat pipe radiator is used as a heat sink for high-power transistors and thyristors. It contributes to low package sizes and general costbenefits.

Heat exchangers. The use of bundles of multiple heat pipes as heat exchangers is second only to electronics thermal control in terms of number of units in operation. While heat-pipe heat exchangers in the form of air-air or air-liquid heat recovery systems are used in domestic and commercial buildings for conserving energy in heating, ventilating, and air conditioning, it is in the industrial field that the heat-pipe heat exchanger has the most applications. Gas-gas heat-pipe heat exchangers are used in a wide range of industrial processes, normally to recover heat from hot or humid exhaust streams to preheat make-up air. Heat may also be recovered for space heating.

The conventional heat-pipe heat exchanger is available in sizes covering heat transport duties ranging from a few tens of watts to several megawatts. The ability to separate source and sink conveniently has led to many derivatives which show the versatility of heat pipe and thermosiphon heat exchangers. These include systems for recovering heat from wastewater, condensers, and exchangers for improving the performance of domestic heat pumps; heat recovery in flue gas desulfurization; and heat transfer in spent nuclear fuel stores. Heat-pipe waste heat boilers (which use hot gases to generate steam for processes or space heating) have been used in situations where reactive fluids have to be separated, and the ability of heat-pipe heat exchangers (in common with the run-around coil, another heat recovery system) to accommodate different external surfaces on the evaporators and condensers benefits many energy-intensive processes. *See* HEAT EXCHANGER; HEAT TRANSFER.

David A. Reay

Bibliography. P. D. Dunn and D. A. Reay, *Heat Pipes*, 4th ed., 1994; D. A. Reay, *Advances in Heat Pipe Technology*, 1982; M. Terpstra and J. G. Van Veen, *Heat Pipes: Construction and Application*, 1987; L. L. Vasiliev, Low temperature heat pipes, *J. Heat Recovery Sys.*, 5:203–216, 1985.

Heat pump

The thermodynamic counterpart of the heat engine. A heat pump raises the temperature level of heat by means of work input. In its usual form a compressor takes refrigerant vapor from a low-pressure, low-temperature evaporator and delivers it at high pressure and temperature to a condenser (**Fig. 1**). The pump cycle is identical with the customary vapor-compression refrigeration system. *See* REFRIGERATION CYCLE.

Application to comfort control. For air-conditioning in the comfort heating and cooling of space, a heat pump uses the same equipment to cool the conditioned space in summer and to heat it in winter, maintaining a comfortable temperature at all times (**Fig. 2**). *See* AIR CONDITIONING; COMFORT HEATING.

This dual purpose is accomplished, in effect, by placing the low-temperature evaporator in the conditioned space during the summer and the high-temperature condenser in the same space during the winter (**Fig. 3**). Thus, if 70°F (21°C) is to be maintained in the conditioned space regardless of the season, this would be the theoretical temperature of the evaporating coil in summer and of the condensing coil in winter. The actual temperatures on the refrigerant side of these coils would need to be below 70°F in summer and above 70°F in winter to permit the necessary transfer of heat through the coil surfaces.

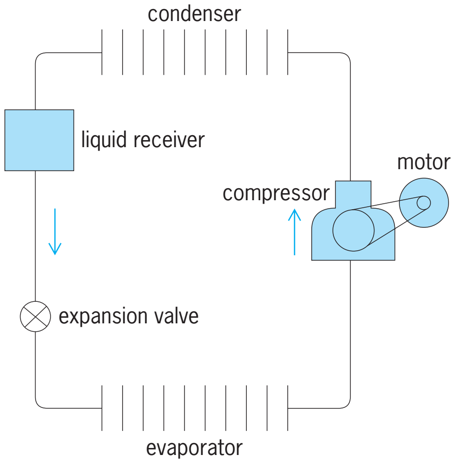


Fig. 1. Basic flow diagram of heat pump with motor-driven compressor. For summer cooling, condenser is outdoors and evaporator indoors; for winter heating, condenser is indoors and evaporator outdoors.

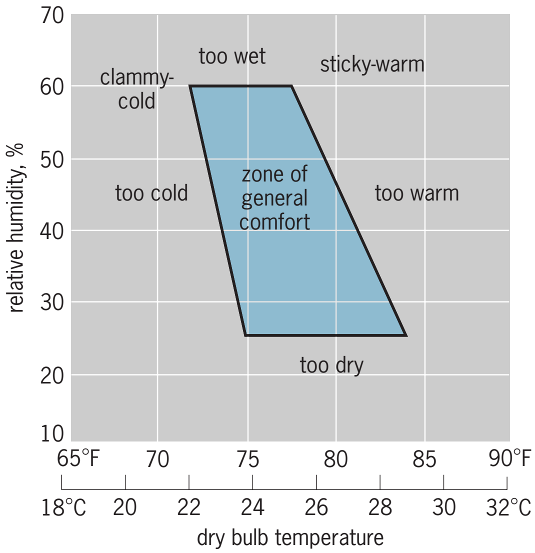


Fig. 2. Indoor climatic conditions acceptable to most people when doing desk work; continuous air motion with 5–8 air changes per hour.

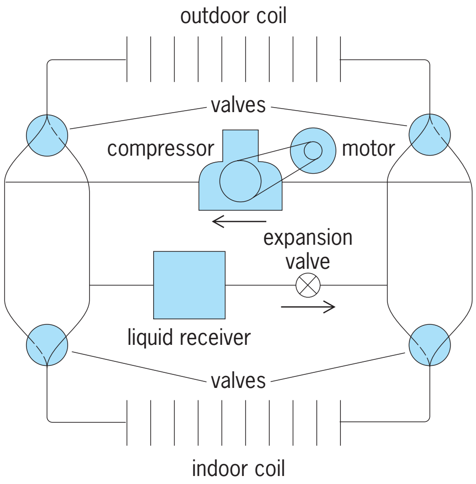


Fig. 3. Air-to-air heat pump installation; fixed air circuit with valves in the summer positions (the broken lines show the winter positions).

If the average outside temperatures are 100°F (38°C) in summer and 40°F (4°C) in winter, the heat pump serves to raise or lower the temperature 30°F (17°C) and to deliver the heat or cold as required. The ultimate ideal cycle for estimating performance is the same Carnot cycle as that for heat engines. The coefficient of performance COP_c as cooling machine is given in Eq. (1), and the coefficient COP_w

$$\mathrm{COP}_c = \frac{\text{refrigeration}}{\text{work}} = \frac{T_c}{T_h - T_c} \qquad (1)$$

as a warming machine is given in Eq. (2), where T is

$$\mathrm{COP}_w = \frac{\text{heat delivered}}{\text{work}} = \frac{T_h}{T_h - T_c} \qquad (2)$$

temperature in degrees absolute (Rankine) and the subscripts c and h refer to the cold and hot temperatures, respectively.

For the data cited, the theoretical coefficients of performance are as in Eqs. (3). The significance of

$$\begin{aligned} \mathrm{COP}_c &= \frac{460+70}{(460+100)-(460+70)} = 17.7 \\ \mathrm{COP}_w &= \frac{460+70}{(460+70)-(460+40)} = 17.7 \end{aligned} \qquad (3)$$

these coefficients is that ideally for 1 kilowatt-hour (kWh) of electric energy input to the compressor there will be delivered 3413 × 17.7 = 60,000 Btu/h as refrigeration or heating effect as required. This is a great improvement over the alternative use of resistance heating, typically, where 1 kWh of electric energy would deliver only 3413 Btu. The heat pump uses the second law of thermodynamics to give a much more substantial return for each kilowatt-hour of electric energy input, since the electric energy serves to move heat which is already present to a desired location. *See* CARNOT CYCLE; THERMODYNAMIC PRINCIPLES.

Effect of seasonal loads. When summer comfort cooling of space is desired, it is entirely possible and practical to use the same compressor equipment and coils for winter heating and for summer cooling and to dispense with the need for direct-fired apparatus using oil, gas, coal, or wood fuel.

For an economical installation, equipment must be of correct size for both the summer cooling and winter heating loads. Climatic conditions have a significant influence and can lead to imbalance on sizing. If the heating and cooling loads are equal, the equipment can be selected with minimum investment. However, generally the loads are not balanced; in the temperate zone the heating load is usually greater than the cooling load. This necessitates (1) a large, high-horsepower compressor fitted to the heating demand, (2) a supplementary heating system

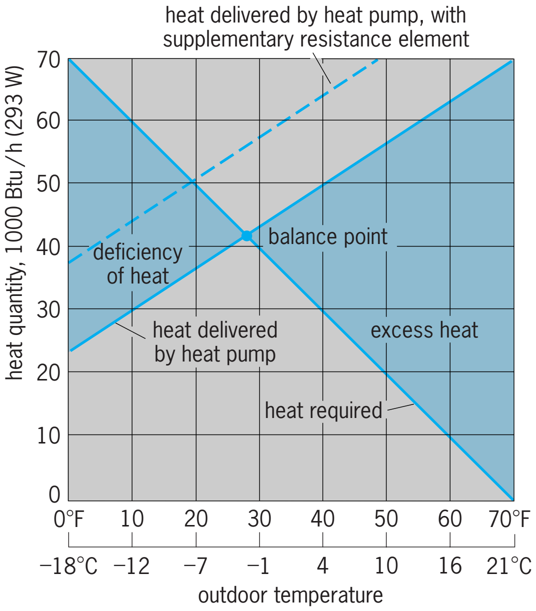


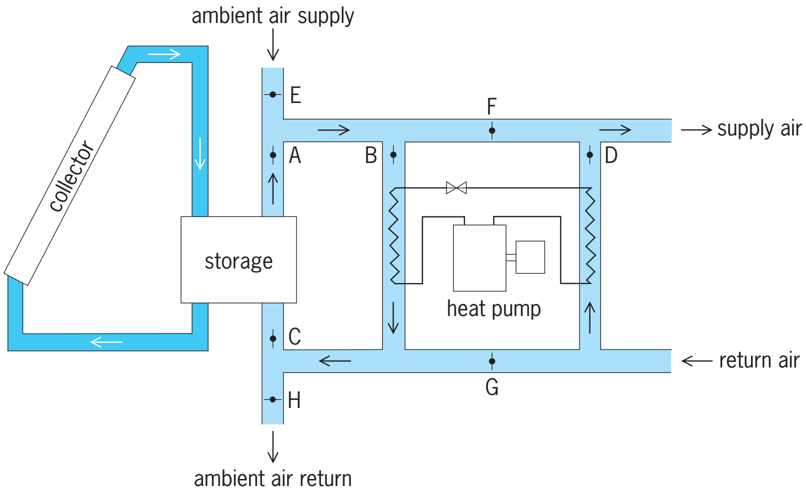
Fig. 4. Performance of air-to-air, self-contained, domestic heat pump on the heating cycle.

(electrical resistance, fuel, or solar energy), or (3) a heat-storage system. *See* COMPRESSOR; RESISTANCE HEATING; SOLAR ENERGY.

If well water or the ground serves as the heat source, the imbalance is less severe than when atmospheric air is the source. However, the uncertain heat transfer rates with ground coil, the impurities, quality, quantity, and disposal of water, and the corrosion problems mitigate the use of these sources.

Atmospheric air as a heat source is preferable, particularly with smaller domestic units. A self-contained, packaged unit of this type offers maximum dependability and minimal total investment. The performance of such a unit for the heating cycle is illustrated in **Fig. 4**. Curves of heat required and heat available show the limitations on capacity. The heat delivered by the pump is less than the heat required at low temperatures, so that there is a deficiency of heat when the outside temperature goes below, in this case, about 28°F (2°C). The intersection of the two curves is the balance point. There is an area of excess heat to the right and of deficiency of heat to the left of the balance point. Many devices and methods are offered to correct this situation, such as storage systems, supplementary heaters, and compressors operating alternatively in series or in parallel.

In temperate regions heat-pump installations achieve coefficients of performance on the order of 3 on heating loads when all requirements for power, including auxiliary pumps, fans, resistance heaters, defrosters, and controls are taken into account. Automatic defrosting systems, when air is the heat source, are essential for best performance, with the defrost cycle occurring twice a day.



	open	close	mode
1	A,B,C,D	E,F,G,H	temperature of stored energy too low for direct use but higher than ambient air
2	A,C,F,G	E,B,D,H	stored solar energy temperature sufficiently high; direct utilization of solar energy (bypass heat pump)
3	B,D,E,H	A,C,F,G	stored solar energy temperature lower than ambient air temperature; heat pump source is ambient air

Fig. 5. Schematic arrangement of solar-assisted heat-pump system.

Heat pumps are uneconomical if used for the sole purpose of comfort heating. The direct firing of fuels is generally more attractive from an overall financial viewpoint. The investment in heat-pump equipment is higher than that for the conventional heating system. Unless the price of electric energy is sufficiently low or the price of fuels very high, the heat pump cannot be justified solely as a heating device. But if there is need for comfort cooling of the same space in summer, the heat pump, to do both the cooling and heating, becomes attractive.

The heat pump is also used for a wide assortment of industrial and process applications such as low-temperature heating, evaporation, concentration, and distillation. Theodore Baumeister

Solar energy-assisted system. For the heat pump to be more attractive in colder regions, where ambient temperatures remain about (or below) 0°F (−18°C) for long periods, the source temperature for the heat pump should be increased. One possible method for achieving this is solar energy.

Solar collectors. For space heating utilizing solar energy, flat-plate collectors are generally used. A typical flat-plate collector contains a metallic plate painted black, with one or two glass covers and the sides and bottom of the collector well insulated. Solar energy is transmitted through the glass, and a significant part of that reaching the metallic plate—about 90%—is absorbed by the plate, increasing its temperature. The energy absorbed by the plate is, in turn, transferred to a working fluid—usually air or water. If air is the working fluid, it passes over the metallic plate; if water, it flows through tubes attached to the metallic plate.

Because of the heat transfer from the collector to the surroundings, not all the radiant energy reaching the collector is transferred to the working fluid. The efficiency of a collector, defined as the ratio of the heat transfer to the working fluid to the radiant energy reaching the collector surface, decreases with an increase in the absorber temperature. The higher the temperature at which the working fluid is operated, the higher is the temperature of the absorber surface and the lower the efficiency.

Working fluid and size. A heat pump system utilizing solar energy for space heating is schematically illustrated in **Fig. 5**. Air is the working fluid for the solar collectors, and the system uses an air-to-air heat pump. Energy storage is necessary as, during some periods, solar energy may not be available at all, such as during the night, or may be available in very limited quantities, such as on cloudy days. The system can operate in three different modes (Fig. 5). In mode 1, the solar energy stored is not at a sufficiently high temperature for direct use but at a temperature higher than ambient temperature—conditions likely to be met on cold, sunny days. The heat pump uses the stored solar energy in this mode. In mode 2, the stored solar energy is at a sufficiently high temperature for direct utilization—probable conditions during sunny days in the fall and spring. The heat pump is then bypassed. In mode 3—during extended periods of cloudy days or very cold weather—the stored solar energy is depleted and outside air is the source for heat pumps. *See* SOLAR HEATING AND COOLING.

N. V. Suryanaryana

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; L. Miles, *Heat Pumps: Theory and Service*, 1993; M. Taylor and J. Roberts, *Heat Pump Systems*, 1992.

Heat radiation

The energy radiated by solids, liquids, and gases as a result of their temperature. Such radiant energy is in the form of electromagnetic waves and covers the entire electromagnetic spectrum, extending from the radio-wave portion of the spectrum through the infrared, visible, ultraviolet, x-ray, and γ-ray portions. From most hot bodies on Earth this radiant energy lies largely in the infrared region. *See* ELECTROMAGNETIC RADIATION; INFRARED RADIATION.

Radiation is one of the three basic methods of heat transfer, the other two methods being conduction and convection. *See* CONDUCTION (HEAT); CONVECTION (HEAT); HEAT TRANSFER.

A hot plate at 400 K (260°F) may show no visible glow; but a hand which is held over it senses the warming rays emitted by the plate. A temperature of more than 1000 K (1300°F) is required to produce a perceptible amount of visible light. At this temperature a hot plate glows red and the sensation of warmth increases considerably, demonstrating that the higher the temperature of the hot plate the greater the amount of radiated energy. Part of this energy is visible radiation, and the amount of this visible radiation increases with increasing temperature. A steel furnace at 1800 K (2800°F) shows a strong yellow glow. If a tungsten wire (used as the filament in incandescent lamps) is raised by resistance heating to a temperature of 2800 K (4600°F), it emits a bright white light. As the temperature of a substance increases, additional colors of the visible portion of the spectrum appear, the sequence being first red, then yellow, green, blue, and finally violet. The violet radiation is of shorter wavelength than the red radiation, and it is also of higher quantum energy.

In order to produce strong violet radiation, a temperature of almost 3000 K (5000°F) is required. Ultraviolet radiation necessitates even higher temperatures, and there is no solid on Earth which can withstand such temperatures without melting. The Sun emits considerable ultraviolet radiation, as evidenced by the sunburn it produces. The spectral distribution of the Sun's radiation has been measured, and the temperature of the Sun's surface has been determined from Wien's displacement law and corresponds to about 6000 K or 10,000°F (Wien's law is discussed later). Such temperatures have been produced on Earth in gases ionized by electrical discharges. The vapor lamps used on highways, the fluorescent lamp used in offices, and the xenon compact-arc lamp used in searchlights are good examples of such gas discharges. They emit large amounts of ultraviolet radiation. Temperatures up to 20,000 K (35,000°F), however, are still much too low to produce x-rays or γ-radiation. Approaches to the utilization of nuclear energy have made use of the fusion of deuterons in magnetically constricted arcs at extremely high currents and with multiple, intense laser beams bombarding small spheres. By these means, temperatures above 1×10^6 K (2×10^6 °F) have been obtained for small fractions of a second. These devices require enormous amounts of energy to produce such high temperatures. Matter maintained at such temperatures emits x-rays and γ-rays. *See* NUCLEAR FUSION; SUN; ULTRAVIOLET RADIATION.

Heat radiators emit energy over a wide range of angles and wavelengths at a single temperature. Lasers, in a sense, are completely opposite to heat radiators, emitting a very narrow range of wavelengths within a narrow beam. In addition, a laser will not operate if the laser medium has only a single temperature. *See* LASER.

Theory. The emission of radiation is explained in terms of excited atoms and nuclei. For example, electrons in an atom can be ejected from their normal orbits around the atom into those farther from the nucleus. When this happens the atom is said to be in an excited state. This occurs when energy, supplied from outside a substance, is converted into thermal motion and finally into excitation. A short time after excitation, the electrons return to their normal orbits and give off their excess energy ΔE in the form of radiation of a particular frequency ν. This wavelength may be determined by the relation $\Delta E = h\nu$, where h is Planck's constant. *See* ATOMIC STRUCTURE AND SPECTRA; PLANCK'S CONSTANT.

In a gas, the thermal motion consists of substantially unhindered movement of the individual particles with different velocities. In a solid, on the other hand, the thermal motion is an oscillating movement of the particles, with varying displacements, about their fixed positions. The extent of the thermal motion depends upon the temperature. The hotter the substance, the greater the thermal motion and the higher the intensity and energy of the radiation. An energy distribution of the radiation intensity results, for example, from the distribution of velocities of the particles in a gas or from the distribution of displacements of the particles about their positions in a solid.

Further, the maximum available energy (excitation energy) depends upon temperature, and this explains why the energy of emitted radiation shifts to shorter wavelength (that is, higher energy) as the temperature is increased. For instance, a temperature of 1000 K (1300°F) produces just enough excitation energy for the dark red glow which contains the longest wavelengths within the visible portion of the spectrum. As explained before, higher temperatures or greater excitation energies are necessary to excite measurable quantities of the shorter wavelength regions. It is obvious that with decreasing temperatures, less excitation energy is available, and the amount of heat radiation decreases until finally at absolute zero of temperature (0 K or −459.67°F) substances radiate no energy because all atomic motion has ceased. However, for a definition of zero point energy *see* QUANTUM MECHANICS

The radiated energy per second is commonly expressed in terms of joules per second, or watts. Other units often used are ergs per second or calories per second. These are related to each other as follows: 1 watt = 1 joule/s = 10^7 ergs/s = 0.239 cal/s. For instance, the Sun radiates onto 1 cm^2 of the Earth's surface 2 cal/min or $^1/_{30}$ cal/s or about $^1/_7$ W. The total energy radiated from 1 cm^2 (0.155 $in.^2$) of a tungsten wire in an incandescent lamp at 2800 K (4600°F) is 112 W. The same wire at room temperature emits only 0.0015 W. *See* SOLAR CONSTANT.

Energy distribution curves. To evaluate the usefulness of a heat radiator, energy distribution curves are used. These are graphs of relative or absolute radiated energy versus the wavelength of radiation (expressed in micrometers, nanometers, or angstroms) or the frequency (velocity of light/wavelength) expressed in hertz (cycles per second).

Such graphs show how the energy radiated from a substance at a certain temperature is distributed over the various portions of the spectrum (**Fig. 1**). The usefulness of these graphs lies in the fact that they provide information, for example, on the effectiveness of a radiator as a light source or as a heating element. Furthermore, the area under the energy distribution curve is equivalent to the total radiated energy.

The energy distribution of various substances differs because of their internal properties and their surface condition. As a common rule, which holds well above 3 micrometers, substances with good electrical conductivity, especially metals, are poor emitters of radiation and are good reflectors (for example, silver or aluminum). Insulators radiate strongly in the infrared region of the spectrum and have gaps of low radiation intensity near the visible portion of the spectrum, as shown in Fig. 1*c*. These gaps are due to the electronic band structure of insulators. Roughening the surface of all radiators increases the emitted energy. This is true because tiny holes in the surface act as cavity radiators, radiating almost blackbody energy. (Cavity radiators and blackbody radiation are discussed later.) *See* ELECTRIC INSULATOR.

Energy distribution curves are obtained by passing white (heterochromatic) light through a monochromator (quartz prism, grating, and the like) and measuring the spectral intensity of radiation with a phototube or a thermopile. The measured intensities at the various wavelengths are then plotted either as percent of the maximum intensity (relative energy) or as absolute intensity (absolute energy) versus the wavelength. *See* RADIOMETRY.

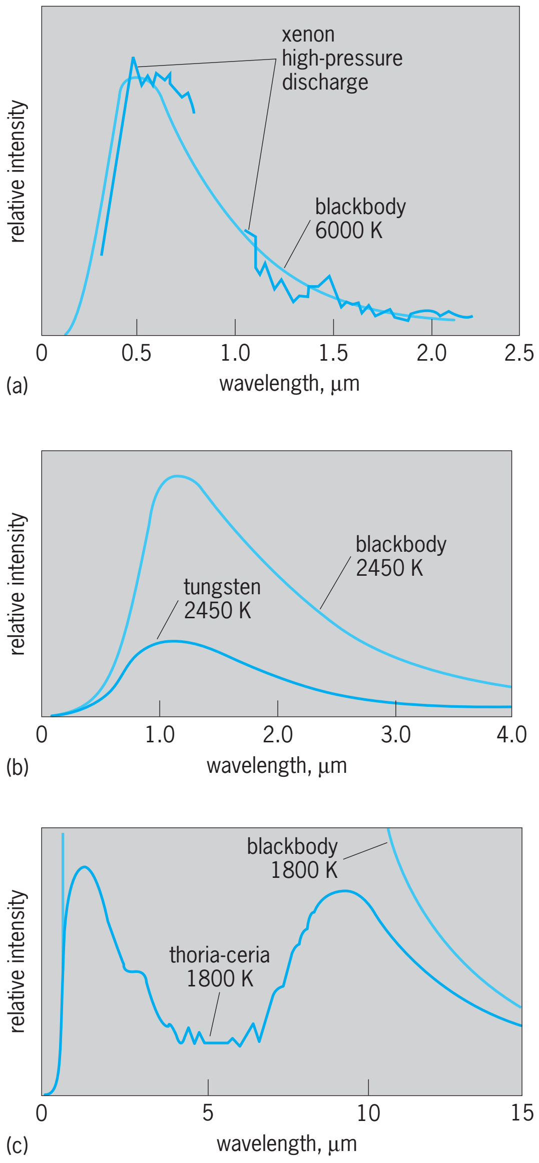


Fig. 1. Energy distribution curves for (*a*) xenon high-pressure electrical gas discharge (*after W. Meyer, ed., Technischwissen-schaftliche Abhandlungen aus dem Osramkonzern, vol. 6, Springer, 1953*); (*b*) tungsten (*after American Institute of Mining and Metallurgical Engineers, Pyrometry: The Papers and Discussions of a Symposium on Pyrometry, 1920*); (*c*) thoria plus 1% ceria, a typical ceramic (*after R. W. Pohl, Einführung in die Optik, Springer, 1948*). °F = (K × 1.8) − 460.

A radiator used in heating rooms should produce much infrared radiation (heat) and no light, whereas much visible light and little heat is desired from a light source. Unfortunately, an energy distribution curve gives a true picture of the radiator for one particular temperature only. If more information is needed, a set of such curves would have to be provided for the temperature range of interest.

Blackbody radiation. Because of the tedious experimental work involved in determining such curves, a different and more fruitful approach is generally taken. Two quantities characterize a heat radiator completely: the total emissivity and the spectral emissivity, which are designated by ϵ and ϵ_λ, respectively, where the subscript λ designates wavelength. Both emissivities, in conjunction with the radiation properties of a blackbody, describe fully the behavior of a real heat radiator. The radiation properties of a blackbody are completely stated by Planck's radiation law.

Planck's law and the concept of blackbody radiation are of utmost importance for the understanding of heat radiation. The blackbody signifies in the domain of heat radiation what any other standard, such as the standard meter, signifies in its own domain. A blackbody is defined as a body which emits the maximum amount of heat radiation. Although there exists no perfect blackbody radiator in nature, it is possible to construct one on the principle of cavity radiation.

A cavity radiator is usually understood to be a heated enclosure with a small opening which allows some radiation to escape or enter. The escaping radiation from such a cavity has the same characteristics as blackbody radiation. Radiation energy which enters the cavity is almost completely absorbed because of the multiple reflections it encounters (**Fig. 2**). This follows because at each reflecting point some of the energy is absorbed by the walls. The absorptivity of the cavity hole is essentially unity, independent of the wall material. As a consequence of Kirchhoff's law, the emissivity of the cavity is also unity, and this fulfills the definition of a blackbody radiator. In practice, the cavity is approximated by a small hole or even a wedge cut into a surface. *See* BLACKBODY.

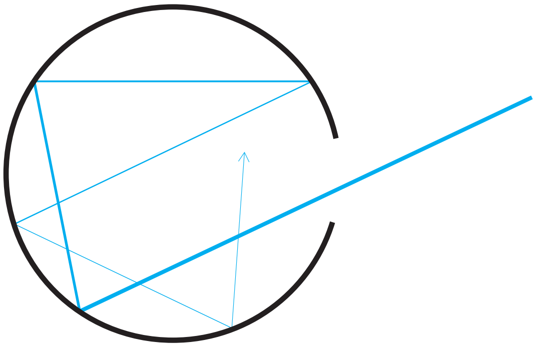


Fig. 2. Diagram of a cavity radiator.

Kirchhoff's law. This law correlates mathematically the heat radiation properties of materials at thermal equilibrium. It is often called the second law of thermodynamics for radiating systems.

Kirchhoff's law can be expressed as follows: The ratio of the emissivity of a heat radiator to the absorptivity of the same radiator is a function of frequency and temperature alone. This function is the same for all bodies, and it is equal to the emissivity of a blackbody. When ϵ is the emissivity of a real radiator, α its absorptivity, and $E = 1$ the emissivity of a blackbody, Kirchhoff's law takes the form of Eq. (1).

$$\frac{\epsilon}{\alpha} = E = 1 \qquad (1)$$

A substance, when brought without contact into an evacuated enclosure the walls of which are at a constant but higher temperature than the body, will assume the wall temperature after some time. However, it will not exceed it. Under these conditions, the exchange of energy can take place only by radiation. As the test body receives radiation from the walls it will absorb some of it, transforming it into motion of its elementary particles, and thereby raising its own temperature. Thermal equilibrium is obtained when the temperature of the walls and the test body is the same; in this case the test body must emit as much energy as it receives. If it absorbs all the impinging radiation, it is a blackbody. If it absorbs only a fraction of the impinging radiation, the other part must be reflected in order to maintain the equilibrium. These statements require that the absorptivity be equal to the emissivity. This is the form in which Kirchhoff's law is often stated. For opaque bodies, absorptivity plus reflectivity must be equal to unity, and therefore the emissivity and the absorptivity respectively must be unity minus the reflectivity. A consequence of Kirchhoff's law is the postulate that a blackbody has an emissivity which is greater than that of any other body.

Planck's radiation law. This celebrated law represents mathematically the energy distribution of the heat radiation from 1 cm^2 of surface area of a blackbody at any temperature. It is the only heat radiation law which is accurate throughout the entire spectrum. The basis of Planck's radiation law was experimental data obtained from measurements on cavity radiators.

Planck's radiation law has great importance. Formulated by Max Planck early in the twentieth century, it laid the foundation for the advance of modern physics and the advent of quantum theory. In determining the heat radiation of hot bodies Planck's radiation law is a basic tool in research and development, both in science and industry. The radiation law can be used to predict light output of incandescent lamps, the cooling time of molten steel, heat dissipation of nuclear reactors, the energy radiated from the Sun, the temperature of the stars, and many other important applications.

Although Planck's law can be derived on theoretical grounds alone, it was deduced from experiment. Prior attempts to calculate the heat radiation of a blackbody had described the radiation as consisting of electromagnetic waves whose energy content could vary continuously. Those attempts did not match the experimental results. Planck replaced the concept of continuous energy with the idea that the energy existed in bundles; that is, the energy was quantized. *See* QUANTUM MECHANICS.

This concept was a drastic innovation at that time. However, upon calculating the radiation, Planck found that the expression in Eq. (2) described the

$$R_\lambda = \frac{c_1}{\lambda^5[e^{c_2/(\lambda T)} - 1]} \qquad (2)$$

experimental results completely. This is the mathematical expression of Planck's radiation law, where

R_λ is the total energy radiated from the body measured in watts per unit area per unit wavelength, at the wavelength λ. The quantity T is the temperature in kelvins, and e is the base of the natural logarithms.

Planck found that $c_1 = 2\pi hc^2$ and $c_2 = hc/k$ where h, c, and k are Planck's constant, the velocity of light, and Boltzmann's constant, respectively. In SI units, $c_1 = 3.7418 \times 10^{-16}$ watt meter^2, $c_2 = 1.4388 \times 10^{-2}$ meter kelvin, and R_λ is in watts per square meter per meter of wavelength. More practical units use area in square centimeters and wavelength in micrometers. Then $c_1 = 37{,}418$ and $c_2 = 14{,}388$, while R_λ is in watts per square centimeter per micrometer of wavelength. Planck's law can be graphed for various temperatures (**Fig. 3**). However, while various substances attain these temperatures, these substances will not radiate as predicted by Planck's law since they are not blackbodies themselves.

The radiation increases at every point of the energy spectrum as the temperature is increased (Fig. 3). At all temperatures, the energy radiated at the extremes of the energy spectrum approaches zero and has a maximum at some place in between. The total area under any of the energy distribution curves (Fig. 1) measures the total energy radiated by the body at the temperature represented by the curve.

Three important aspects of Planck's radiation law can be examined. First, the behavior of the law at the extremes of the energy spectrum leads to a discussion of Wien's radiation law, the Rayleigh-Jeans radiation law, and the so-called ultraviolet catastrophe. Second, the shift of the wavelength at which the maximum energy is radiated can be studied as the temperature is changed. This leads to Wien's displacement law. Finally, the total amount of energy radiated at any temperature can be investigated. This leads to the Stefan-Boltzmann law. The four laws mentioned were well known prior to the formulation of Planck's law. It is the fact that the Planck law so neatly sums up the four earlier laws and introduces the implication of energy quantization which made it of such importance in the development of modern physics during the twentieth century.

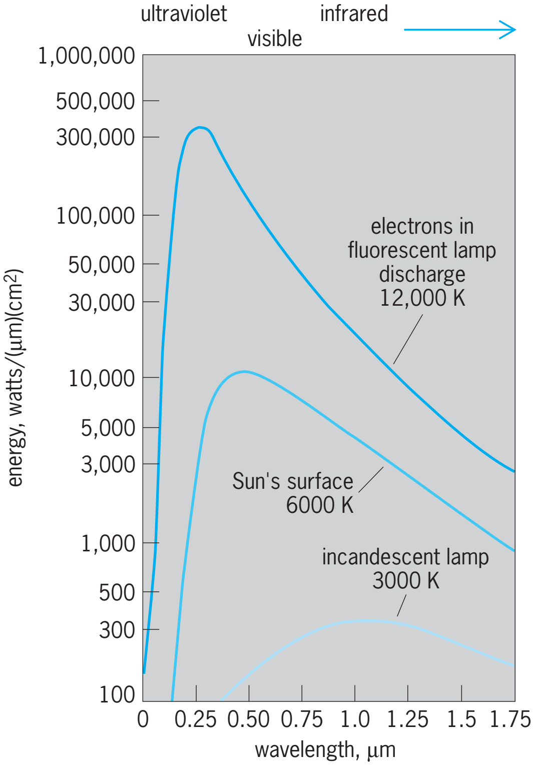


Fig. 3. Curves of Planck's law for various temperatures. Substances which attain these temperatures are also shown, although these substances will not radiate as predicted by Planck's law. °F = (K × 1.8) − 460.

Rayleigh-Jeans law. The heat radiation from a blackbody at long wavelengths is adequately described by the Rayleigh-Jeans radiation law. For larger values of λT, Planck's law simplifies to the Rayleigh-Jeans law, as shown in Eq. (3), where the numerical form

$$R_\lambda = \frac{c_1 T}{c_2 \lambda^4} = 2.6007 \frac{T}{\lambda^4} \tag{3}$$

of the equation is based on the practical units specified above. This law states that the energy radiated at any temperature increases without limit as the wavelength decreases. This law can be accurate only for wavelengths much larger than that at which the maximum occurs (Figs. 1 and 3). For wavelengths shorter than this maximum, the energy radiated from a blackbody actually decreases again. If a blackbody acted as predicted by the Rayleigh-Jeans law, then the energy radiated at very short wavelengths, in the ultraviolet region, would become extremely large and the total energy radiated would be infinite. This is known as the ultraviolet catastrophe, and would be valid at any temperature, no matter how low.

Wien's radiation law. This law is valid at short wavelengths and is obtained from Planck's law by taking λT as very small. Planck's formula then becomes Eq. (4). This law is accurate in the visible region of

$$R_\lambda = \frac{c_1}{\lambda^5} e^{-c_2/(\lambda T)} = \frac{37,418}{\lambda^5} e^{-14,388/(\lambda T)} \tag{4}$$

the spectrum below 3000 K (5000°F).

Wien's displacement law. This law is obtained from Planck's law by the process of differentiation. It describes the shift with temperature of the wavelength at which the maximum amount of radiation occurs by Eq. (5). Thus, the product of the temperature of

$$\lambda_{\max} T = 2898 \qquad \text{(micrometer-kelvins)} \tag{5}$$

a blackbody and the wavelength at which the maximum amount of radiation occurs is a constant. Wien's law has wider significance than this, however. Dividing Planck's law by T^5 results in Eq. (6). On the right-

$$\begin{aligned}\frac{R_\lambda}{T^5} &= \frac{c_1}{(\lambda T)^5[e^{c_2/(\lambda T)} - 1]} \\ &= \frac{37,418}{(\lambda T)^5[e^{14,388/(\lambda T)} - 1]}\end{aligned} \tag{6}$$

hand side of Eq. (6) the wavelength and temperature always appear multiplying each other. This means

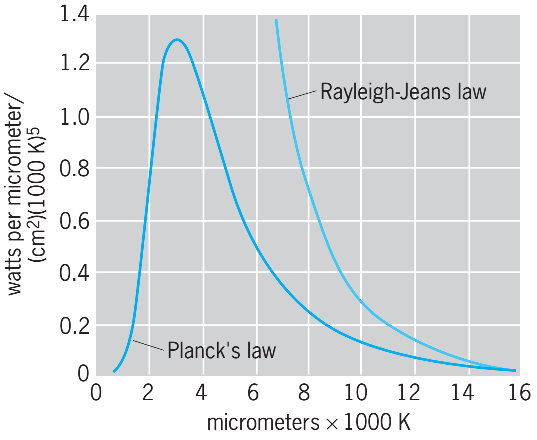


Fig. 4. Planck's law expressed as Wien's displacement law. Rayleigh-Jeans law is shown for comparison.

that only one curve is needed to express Planck's law, for all temperatures, if a graph is used in which the radiation energy divided by the fifth power of the temperature is plotted versus the product of the wavelength and temperature (**Fig. 4**). It is helpful here to measure temperature in thousands of degrees. For comparison, the Rayleigh-Jeans law can also be illustrated.

Wien's displacement law is helpful in determining the temperature of hot bodies. If $\lambda_{\max}$ can be measured, the temperature is immediately obtained from the displacement law. This is how an astronomer measures the temperature of a star. As an example, the radiation from the Sun's surface has a maximum in the green region of the energy spectrum in the vicinity of $\lambda = 0.5\ \mu$m (Fig. 3). From Wien's law the surface temperature of the Sun must then be about 6000 K (10,000°F).

Stefan-Boltzmann law. This law states that the total energy radiated from a hot body increases with the fourth power of the temperature of the body. This law can be derived from Planck's law by the process of integration and is expressed mathematically as Eq. (7), where R_T is the total amount of energy

$$R_T = 5.670 \times 10^{-12} T^4 \tag{7}$$

radiated from the blackbody in watts per square centimeter. When R_T is multiplied by the total emissivity, the total energy radiated from a real heat radiator is obtained.

The rapid increase in heat radiation with temperature is quite evident from the Stefan-Boltzmann law. If the absolute temperature is doubled, say from 273 to 546 K (32 to 524°F), then the energy radiated increases 16-fold. Thus, the attainment of very high temperatures requires large amounts of energy to overcome the loss of energy by heat radiation. Temperatures greater than 3×10^7 K (5×10^7 °F) are encountered in a hydrogen bomb explosion. Such temperatures are 100,000 times higher than room temperature. Therefore, the energy radiated by 1 cm^2 (0.155 in.^{2}) of a substance at this high temperature will be 1×10^{20} times as much as that radiated at room temperature by the same substance. This energy, if radiated by a blackbody, would boil 2×10^7 tons (2×10^{10} kg) of ice water in 1 s.

Temperature determination. The apparent temperature of a real heat radiator can be determined by comparison with a blackbody whose temperature is known. This is done in any one of three customary ways, based upon the radiation laws described. (1) The radiation temperature of a surface is the temperature of the blackbody which radiates the same total energy per unit area in 1 s as does the surface. This temperature is based upon the Stefan-Boltzmann law. (2) The brightness temperature of a surface is the temperature of the blackbody which has the same brightness at a certain wavelength as has the surface. The wavelength is normally taken as 0.655 μm, and the brightness temperature thus measured can be used in conjunction with the spectral emissivity $\epsilon_{0.655}$ and Wien's radiation law to calculate the true temperature of the body. (3) The color temperature of a surface is the temperature of a blackbody whose radiation has the same or approximately the same energy distribution as the surface. This is illustrated in Fig. 1*a*, where it is seen that the xenon high-pressure discharge has a color temperature approximating 6000 K (10,000°F). For a graybody, that is, a body whose emissivity does not vary throughout the spectrum, the color temperature and true temperature are the same. *See* INCANDESCENCE.

Heinz G. Sell; Peter J. Walsh

Bibliography. M. Born, *Atomic Physics*, 8th ed., 1969, reprint 1989; R. P. Feynman, *QED: The Strange Theory of Light and Matter*, 1985; S. Fluegge (ed.), *Handbuch der Physik*, vol. 26, 1958; M. F. Modest, *Radiative Heat Transfer*, 1993; R. Siegel and J. R. Howell, *Thermal Radiation Heat Transfer*, 3d ed., 1992; M. R. Wehr et al., *Physics of the Atom*, 4th ed., 1984.

Heat storage systems

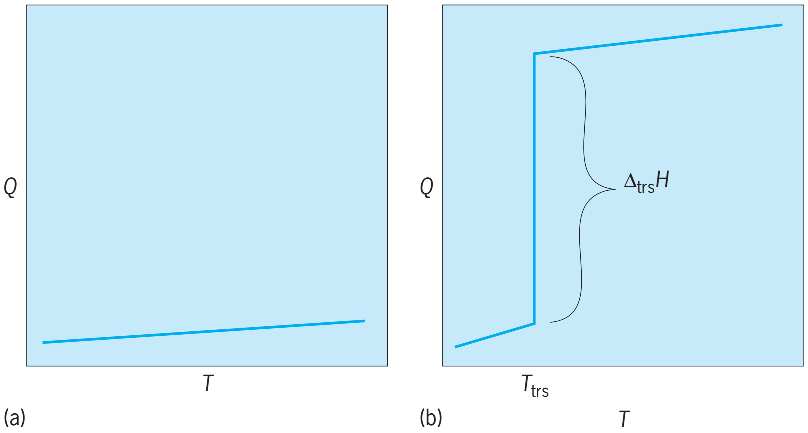
Systems that buffer temperature changes and reclaim energy that would otherwise be lost to the surroundings. Reclaiming energy increases the efficiency of systems in both industrial applications and smaller-scale applications such as household use of solar energy. In addition, heat storage materials and systems can be used in applications where the temperature needs to be kept within a certain range, such as prevention of damage to biological materials or electronic devices. The most important temperature range for heat storage materials for both industrial and consumer use is 0–100°C (32-212°F).

Heat storage materials. Heat storage systems rely on the materials within them to store thermal energy. These materials can be classified as sensible heat storage materials or phase-change materials (PCM).

Sensible heat storage. Every material stores energy within it as it is heated, and in this way is a "sensible heat storage material." The energy stored can be quantified in terms of the heat capacity C, the temperature change ΔT (= final temperature − initial temperature), and the amount of additional heat stored ΔQ, such that $\Delta Q = C\Delta T$. The units for Q, C, and T are joules (J), J K^{-1}, and K, respectively (or Btu, Btu/°F, and °F). Clearly, other factors being equal, the higher the heat capacity (C) of a material, the greater will be the energy stored (ΔQ) for a given temperature rise (ΔT). Typical materials in a sensible heat storage device include rock, sand, wood, and water. An example of sensible heat storage is a night storage heater, which stores energy in bricks that are heated using nighttime electrical energy (available at reduced rates), and radiates this heat to the surroundings during the following day. *See* HEAT; HEAT CAPACITY; SPECIFIC HEAT.

The energy density (defined as $\Delta Q/V$, where V is the volume of the system) of a sensible heat storage system depends on both the heat capacity and the density. For example, balsa wood has a specific heat capacity (heat capacity per unit mass) of 2.9 J K^{-1} g^{-1}, which is rather high compared with that of marble (0.88 J K^{-1} g^{-1}). However, balsa has a rather low density (0.16 g cm^{-3}) compared with marble (2.6 g cm^{-3}), which means that 1 cm^3 of marble can store 2.3 J for a 1°C (or, equivalently, 1 K) temperature rise, whereas 1 cm^3 of balsa can store only 0.46 J over the same temperature increment. Ideally, it would be best to have as small a system as possible to store heat energy. One way to improve the energy density of the heat storage system is to use phase-change materials.

Phase-change materials. Although all materials increase their heat content Q as the temperature is increased, a very large increase in Q occurs when materials change phase. For example, the heat content of water increases considerably as it goes through the phase change from ice to liquid; this is the familiar melting process. The step in Q at the phase change is the latent heat associated with the transition, usually represented as $\Delta_{\text{trs}} H$ (see **illus.**). The step in Q at the transition is in addition to the sensible heat storage capacity of the material. *See* MELTING POINT.



Heat content *Q* as a function of temperature *T*. (*a*) *Q* increases with increasing temperature, even if there is no phase transition, as in a sensible heat storage material. (*b*) When the material undergoes a phase transition at temperature T_{trs}, a dramatic increase in *Q* occurs; its jump corresponds to the value of the latent heat of the transition $\Delta_{\text{trs}}H$ as indicated on the diagram. This large increase in *Q* can be used to advantage in phase-change materials for heat storage.

A phase change can lead to a much larger quantity of energy stored, compared with sensible storage alone. The comparison for water is quite useful. Pure water has a heat capacity of 4.2 J K^{-1} g^{-1}, so for a 1°C temperature rise, 1 g of water can store 4.2 J. However, the latent heat associated with melting of ice is 330 J g^{-1}. So taking 1 g of ice from just below its melting point to just above (with a total temperature difference of 1°C) absorbs 334 J (latent heat plus 4.2 J from sensible heat storage), about 80 times as much as the sensible heat storage capacity alone.

Although phase-change materials can be used to provide more energy storage, they have their limitations. The most critical limitation is that, in order to be most useful, they must have their phase transition in the temperature range at which the system will operate. For example, at ordinary pressures water has only two phase transitions, one associated with melting at 0°C (32°F) and one with boiling at 100°C (212°F). If a heat storage system were required to operate in the temperature range of 20–40°C (68–104°F), water could operate only as a sensible heat storage material, not as a phase-change material.

Glauber salts are commonly used phase-change materials. They are inorganic salt hydrates that undergo melting-phase transitions that absorb heat. While these materials are highly efficient at storing heat, their maximum storage capacity, again, is over a limited range (transition temperature from 30 to 48°C or 86 to 118°F). In addition, there are problems associated with phase separation on melting, and there is a need for supercooling in order to recover all the stored heat. While these materials are inexpensive, their corrosive nature can lead to high storage container costs. However, the energy storage capacity of Glauber salts and related materials is high, and they are used in commercial products.

Paraffin waxes are also used as phase-change materials. When a paraffin wax melts, it absorbs heat and stores energy. However, there are drawbacks associated with using paraffin waxes as heat storage materials such as oxidation, volatility, flammability, and the large volume change on melting (about 20% increase). Once again, containment costs can be high. Paraffins are often microencapsulated in commercial phase-change materials applications. *See* PARAFFIN.

Pentaerythritol [$C(CH_2OH)_4$] and its derivatives absorb energy on going through phase transformations, but its phases (before and after the transition) are both solid. The attractive feature of using a solid-solid phase transition is that the mechanical properties of the material are much the same before and after the transition. In particular, the flow problems associated with melting transitions are overcome. The major problem with pentaerythritol is its relatively high transition temperature (188°C or 370°F), which severely limits its commercial use. The transition temperature can be lowered through chemical modification, but this also lowers the quantity of heat absorbed in the transition. Nevertheless, development of new solid-solid phase transition phase-change materials offer considerable promise for heat storage systems.

Applications. Broadly speaking, applications of heat storage systems fall into two categories: scavenging of heat, and thermal control. Both rely on the principle that as energy is put into the system its temperature rises. Later, the energy is released as the temperature falls. In both cases, the goal is to achieve the maximum energy stored (and later released) for a given temperature change.

Scavenging of heat. These uses range from district heating or cooling to solar energy applications.

In district heating, heat storage systems are used to scavenge energy from what might otherwise be wasted sources, such as smokestacks associated with heavy industry. By appropriate heat exchange between the waste energy source and a fluid, this thermal energy can be moved to another location such as a steam heating system for housing and office buildings.

District cooling makes use of heat storage systems in a somewhat different way. For example, in Minato Mirai 21, a major commercial and residential development in Yokohama, Japan, an underground storage system makes use of off-peak nighttime electricity to cool phase-change materials to their low-temperature state. In the daytime, heat-exchange fluids carry the "cold" to the buildings in this district to provide air conditioning. When the fluids return to the phase-change materials system, their heat warms the material. The high energy density of the phase-change materials allows a large cooling capacity, while leveling out the diurnal electricity consumption.

In solar energy applications, the radiant energy of the Sun is used to increase the energy content of the heat storage system in the daytime. This energy is released from the system as the temperature decreases during the night, either through direct radiation or through exchange of energy with circulating fluids. As with the other heat-scavenging applications, the energy density is greatest if phase-change materials, optimized to the temperature cycle of the system, are employed. *See* SOLAR ENERGY; SOLAR HEATING AND COOLING.

Another example of a scavenging heat storage system is found in some automobiles. The heat storage system uses the engine's waste heat to store thermal energy that is later used to rapidly heat the air for the passenger compartment after a cold start.

Thermal control. Because a heat storage system aims to have a large energy storage for a small temperature change, these systems also have applications in thermal control.

There are now commercially available ski boots that make use of heat storage systems. Phase- change materials in the insole prevent major temperature drops. Any loss of energy because of the cold conditions is tempered by the presence of the material, which requires a large loss of energy (latent heat) before it cools into its low-temperature phase and then cools further. In a similar way, phase-change materials in insulated storage bags are used to keep pizza hot during delivery.

Preventing temperature extremes can be very important in applications involving biological materials or electronic devices. Here again, heat storage systems can be used to "buffer" against major temperature changes in the environment. If necessary, two phase-change materials can be used, with one providing a lower-temperature buffer and the other an upper-temperature buffer. If the temperature of the surroundings is too extreme, or out of range for too long, the storage capacity of the system can be overcome and the temperature can go outside the desired range.

Outlook. The convenience or actual savings associated with a heat storage system must outweigh its capital cost. One way to improve this situation is by the use of more efficient systems. Efficiency has two components: the energy recovered (which should be as high as possible) and the quality of the energy transferred (the fraction of the energy that is available for conversion to useful work, also called the exergy). New materials, such as high-efficiency phase-change materials with a wide or tunable range of phase-transition temperatures, would be very useful. New designs of the heat storage systems can further optimize efficiency. With further realization of the limitations of energy supplies, heat storage materials are likely to play a larger role in consumer and industrial energy distribution systems in the near future. *See* ENERGY STORAGE. Mary Anne White

Bibliography. S. M. Hasnain, Review on sustainable thermal energy storage techniques, Part II: Cool thermal storage, *Energy Convers. Manag.*, 39:1139–1153, 1998; F. R. Kalhammer, Energy-storage systems, *Sci. Amer.*, 241(6):56–65, 1979; C. J. Weinberg and R. H. Williams, Energy from the Sun, *Sci. Amer.*, 263(3):146–155, 1990; M. A. White, *Properties of Materials*, Oxford University Press, New York, 1999.

Heat transfer

Heat, a form of kinetic energy, is transferred in three ways: conduction, convection, and radiation. Heat can be transferred only if a temperature difference exists, and then only in the direction of decreasing temperature. Beyond this, the mechanisms and laws governing each of these ways are quite different. This article gives introductory information on the three types of heat transfer (also called thermal transfer) and on important industrial devices called heat exchangers.

Conduction. Heat conduction involves the transfer of heat from one molecule to an adjacent one as an inelastic impact in the case of fluids, as oscillations in solid nonconductors of electricity, and as motions of electrons in conducting solids such as metals. Heat flows by conduction from the soldering iron to the work, through the brick wall of a furnace, through the wall of a house, or through the wall of a cooking utensil. Conduction is the only mechanism for the transfer of heat through an opaque solid. Some heat may be transferred through transparent solids, such as glass, quartz, and certain plastics, by radiation. In fluids, the conduction is supplemented by convection, and if the fluid is transparent, by radiation.

The conductivities of materials vary widely, being greatest for metals, less for nonmetals, still less for liquids, and least for gases. Any material which has a low conductivity may be considered to be an insulator. Solids which have a large conductivity may be used as insulators if they are distributed in the form of granules or powder, as fibers, or as a foam. This increases the length of path for heat flow and at the same time reduces the effective cross-sectional area, both of which decrease the heat flow. Mineral wool, glass fiber, diatomaceous earth, glass foam, Styrofoam, corkboard, Celotex, and magnesia are all examples of such materials. *See* CONDUCTION (HEAT).

Convection. Heat convection involves the transfer of heat by the mixing of molecules of a fluid with the body of the fluid after they have either gained or lost heat by intimate contact with a hot or cold surface. The transfer of heat at the hot or cold surface is by conduction. For this reason, heat transfer by convection cannot occur without conduction. The motion of the fluid to bring about mixing may be entirely due to differences in density resulting from temperature differences, as in natural convection, or it may be brought about by mechanical means, as in forced convection.

Most of the heat supplied to a room from a steam or hot-water radiator is transferred by convection. In fact, the heat from the fire in the furnace heating the hot water or steam is transferred to the boiler wall by convection, and the hot water or steam transfers heat from the boiler to the radiator by convection. Iced tea is cooled and soup heated by convection. *See* CONVECTION (HEAT).

Radiation. Solid material, regardless of temperature, emits radiations in all directions. These radiations may be, to varying degrees, absorbed, reflected, or transmitted. The net energy that is transferred by radiation is equal to the difference between the radiations emitted and those absorbed.

The radiations from solids form a continuous spectrum of considerable width, increasing in intensity from a minimum at a short wavelength through a maximum and then decreasing to a minimum at a long wavelength. As the temperature of the object is increased, the entire emitted spectrum decreases in wavelength. As the temperature of an iron bar, for example, is raised to about 1000°F (800 K), the radiations become visible as a dark red glow. As the temperature is increased further, the intensity of the radiation increases and the color becomes more blue. This process is quite apparent in the filament of a light bulb. When the bulb is operated at less than normal voltage, the light appears quite red. As the voltage is increased, the filament temperature increases and the light progressively appears more blue.

Liquids and gases only partially absorb or emit these radiations, and do so in a selective fashion. Many liquids, especially organic liquids, have selective absorption bands in the infrared and ultraviolet regions. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION.

Transfer of energy by radiation is unique in that no conducting substance is necessary, as with conduction and convection. It is this unique property that makes possible the transfer of large amounts of energy from the Sun to the Earth, or the transfer of heat from a radiant heater in the home. It is the ready transfer of heat by radiation from a California orange grove to outer space on a clear night that sometimes results in a frost. The presence of a shield of clouds will tend to prevent this loss of heat and often prevent the frost. By means of heat lamps and gold-plated reflectors, heat may be transferred deep into the layer of enamel on a car body, with resultant hardening of the enamel from the inside out. It is also the transfer over great distance of quantities of radiant energy that makes the atomic bomb so destructive. *See* HEAT RADIATION.

Design considerations. By utilizing a knoweledge of the principles governing the three methods of heat transfer and by a proper selection and fabrication of materials, the designer attempts to obtain the heat flow required for his purposes. This may involve the flow of large amounts of heat to some point in a process or the reduction in flow in others. It is possible to employ all three methods of heat transfer in one process. In fact, all three methods operate in processes that are commonplace. In summer, the roof on a house becomes quite hot because of radiation from the Sun, even though the wind is carrying some of the heat away by convection. Conduction carries the heat through the roof where it is distributed to the attic by convection. The prudent householder attempts to reduce the heat that enters the rooms beneath by reducing the heat that is absorbed in the roof by painting the roof white. He may apply insulation to the underside of the roof to reduce the flow of heat through the roof. Further, heated air in the attic may be vented through louvers in the roof.

Heat transferred by convection may be transferred as heat of the convecting fluid or, if a phase change is involved, as latent heat of vaporization, solidification, sublimation, or crystallization. The human body can be cooled to less than ambient temperature by evaporation of sweat from the skin. Dry ice absorbs heat by sublimating the carbon dioxide. Heat extracted from the products of combustion in the boiler flows through the gas film and the metal tube wall and converts the water inside the tube to steam, all without greatly changing the temperature of the water.

Heat exchangers. In industry it is generally desired to extract heat from one fluid stream and add it to another. Devices used for this purpose have passages for each of the two streams separated by a heat-exchange surface in the form of plates or tubes and are known as heat exchangers. Needless to say, the automobile radiator, the hot-water heater, the steam or hot-water radiator in a house, the steam boiler, the condenser and evaporator on either the household refrigerator or air conditioner, and even the ordinary cooking utensils in everyday use are all heat exchangers. In power plants, oil refineries, and chemical plants, two commonly used heat exchangers are the tube-and-shell and the double-pipe exchangers. The first consists of a bundle of tubes inside a cylindrical shell. One fluid flows inside the tubes and the other between the tubes and the shell. The double-pipe type consists of one tube inside another, one fluid flowing inside the inner tube and the other flowing in the annular space between tubes. In both cases, the tube walls serve as the heat-exchange surface. Heat exchangers consisting of spaced flat plates with the hot and cold fluids flowing between alternate plates are also in use. Each of these exchangers essentially depends upon convection heat flow through a film on each side of the heat-exchange surface and conduction through the surface. Countless special modifications, often also utilizing radiation for heat transfer, are in use in industry.

In these exchangers, the fluid streams may flow parallel concurrently or in mixed flow. In most cases, the temperatures of the various streams remain essentially constant at a given point, and the process is said to be a steady-state process. As the streams move through the exchangers, unless there is a phase change, the fluids are continuously changing in temperature, and the temperature gradient from one stream to the other may be continuously varying. To determine the amount of surface needed for a given process, the designer must evaluate the effective temperature gradient for the particular condition and exchanger.

With extremely high temperatures, or with gas streams carrying suspended solids, the use of conventional heat exchangers becomes impractical. Under these conditions, the transfer of heat from one stream to another becomes more economical by the alternate heating and cooling of refractory solids or by checkerwork as in the blast-furnace hot stove, in the glass-furnace regenerator, or in the Royster stove. At lower temperatures, metal packing is frequently employed, as in the Ljungstrom preheaters or in regenerators for liquid-air production. In petroleum refining and in the metallurgical industry, exchangers are being employed in which one or more of the streams are fluidized beds of solids, the large area of the solids tending to produce very high rates of heat exchange. In some of these devices and also in nuclear power reactors, large quantities of heat are being generated in the exchangers. Here one of the principal problems involves the rapid removal of this heat before the temperature rises to the point where the equipment is damaged or destroyed.

Often the heating or cooling of a body is desired. In this case, the body representing the second stream does not remain at constant temperature, the heat being transferred representing a change in the heat content of the body. Such a process is known as an unsteady-state process. The heating or cooling of food and canned products in utensils, refrigerators, and sterilizers; the heating of steel billets in metallurgical furnaces; the burning of brick in a kiln; and the calcination of gypsum are examples of this type of process. *See* HEAT; HEAT EXCHANGER.

Ralph H. Luebbers

Bibliography. Y. Bayazitoglu and M. N. Ozisik, *Elements of Heat Transfer,* 1988; M. Becker, *Heat Transfer: A Modern Approach,* 1986; A. Bejan and J. S. Jones, *Modern Heat Transfer*, 1993; A. J. Chapman, *Fundamentals of Heat Transfer,* 1987; J. P. Holman, *Heat Transfer*, 8th ed., 1997; F. Kreith and M. Bohn, *Principles of Heat Transfer*, 6th ed., 2000.

Heat treatment (metallurgy)

A procedure of heating and cooling a material without melting. The heating and cooling sequence may involve temperatures above, below, and at the ambient. Controlled heating and cooling rates, and a variety of furnace atmospheres and heating media may be used. Plastic deformation may be included in the sequence of heating and cooling steps, thus defining a thermomechanical treatment. Typical objectives of heat treatments are hardening, strengthening, softening, improved formability, improved machinability, stress relief, and improved dimensional stability. Heat treatments are often categorized with special names, such as annealing, normalizing, stress relief anneals, process anneals, hardening, tempering, austempering, martempering, intercritical annealing, carburizing, nitriding, solution anneal, aging, precipitation hardening, and thermomechanical treatment.

Steels and Other Ferrous Alloys

All metals and alloys in common use are heat-treated at some stage during processing. Iron alloys, however, respond to heat treatments in a unique way because of the multitude of phase changes which can be induced, and it is thus convenient to discuss heat treatments for ferrous and nonferrous metals separately. *See* IRON; STEEL.

General principles. At room temperature the equilibrium crystal structure of pure iron is body-centered cubic α-iron (**Fig. 1***a*), also known as

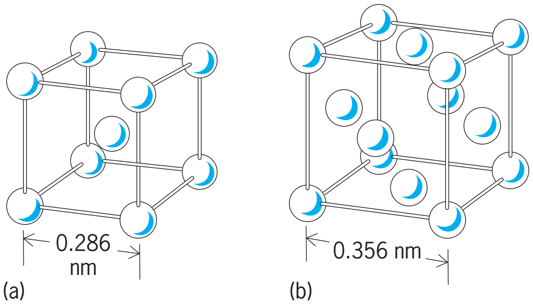


Fig. 1. Crystal structures of iron. (*a*) Body-centered cubic α-iron. (*b*) Face-centered cubic γ-iron.

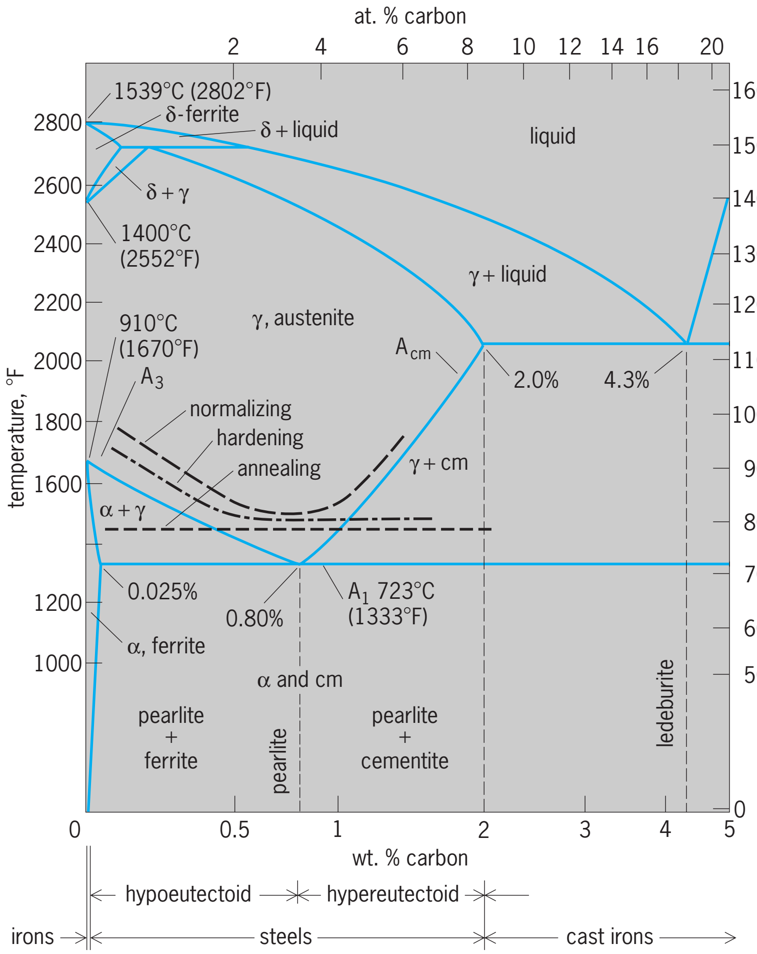


Fig. 2. Iron-carbon phase diagram.

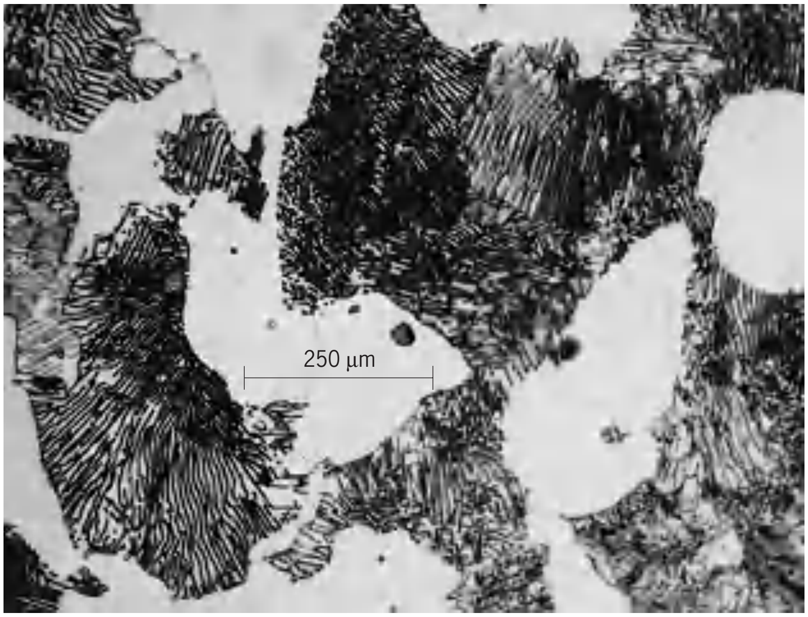


Fig. 3. Microstructure of annealed 0.40% carbon steel, after polishing and etching. White regions are ferrite; lamellar regions, pearlite.

ferrite. On heating above 1670°F (910°C), α-iron is transformed to face-centered cubic γ-iron (Fig. 1*b*), also called austenite. At 2552°F (1400°C) γ-iron transforms to δ-ferrite, which is also body-centered cubic and structurally similar to α-iron, but δ-ferrite is seldom involved in heat-treating procedures. The addition of carbon to iron influences the transformation from one form of iron to another, and the resultant structures are summarized in the iron-carbon phase diagram shown in **Fig. 2**. Practical steels and irons contain other elements such as manganese, silicon, sulfur, phosphorus, aluminum, silicon, chromium, molybdenum, and nickel. These alloy elements influence the shape of the iron-carbon diagram, but if the total alloy content is less than 2%, the phase diagram is not affected substantially.

The general principles of the heat treatment of plain-carbon and low-alloy steels may be understood from the basic iron-carbon diagram. The diagram indicates the microconstituents, or phases, which are observed for a given carbon content at each temperature. In addition to ferrite and austenite, the other principal phase is the intermetallic compound, Fe_3C, or cementite. Cementite is, however, not the most stable form of carbon in iron, and at true equilibrium, graphite is formed after prolonged heating. Graphite is a constituent of the common grades of cast iron, but these irons usually contain 2–4% silicon to promote graphitization. From a practical standpoint, cementite is, however, stable in most alloys with less than 2% carbon. Approximate ranges of the carbon contents for irons, steels, and cast iron are shown in Fig. 2.

Steels which contain 0.80% carbon are called eutectoid steels, those with less carbon are hypoeutectoid, and those with more are hypereutectoid. The changes which occur on slowly heating a typical hypoeutectoid steel containing 0.40% carbon are the following. At room temperature such a steel contains ferrite and cementite. On heating above the A_1 critical temperature, 1333°F (723°C), the cementite dissolves, austenite is formed, and the structure consists of austenite and ferrite. Above the A_3 temperature, which is 1472°F (800°C), for the alloy, the structure is entirely austenitic. Melting is observed at 2640°F (1450°C), and in practice, heat-treating and hot-forming temperatures are below 2200°F (1200°C). On cooling, the sequence of phase changes is reversed, but the shapes and distribution of the phases depend on the cooling rate. If the cooling is very slow, as in furnace cooling, the carbides will tend to have a spheroidal shape. If the cooling rate is more rapid, as in air cooling, the cementite will form a duplex lamellar structure with the ferrite, called pearlite, and the microstructure will consist of 50% ferrite grains and 50% pearlite colonies as shown in **Fig. 3**. A eutectoid steel, containing 0.80% carbon, will consist entirely of pearlite on air-cooling from above A_1. Hypereutectoid steels will consist of pearlite colonies and excess cementite. It should be emphasized that pearlite is not a phase, and it is shown on the phase diagram only for convenience.

Annealing heat treatments. Annealing heat treatments are used to soften the steel, to improve the machinability, to relieve internal stresses, to impart dimensional stability, and to refine the grain size. Several typical annealing treatments are discussed below.

Anneal or grain-refining anneal. Hypo- and hypereutectoid steels are heated just above A_1, and cooled moderately slowly to room temperature. Recrystallization of the ferrite grains will occur during the heating, and the material will be softened by the stress relief and the removal of crystallographic imperfections at temperature and during the slow cooling. Anneals of this type are used to improve the machinability of hypereutectoid steels since the blocky pearlite which forms (Fig. 3) aids in the breakup of the chips during machining.

Spheroidizing anneal. In this treatment, steels are heated just above A_1 and cooled very slowly through the critical temperature, using a programmed reduction of the temperature over a period of 10–15 h to 1200°F (650°C), and then cooling more rapidly to room temperature. In other procedures, the temperature is cycled slowly from just above to just below the critical, or the steel may be cooled to just below the critical and held for a long period. The latter are called isothermal anneals. The objective of all of these treatments is to produce spheroidized carbides. Such structures provide superior cold formability in hypoeutectoid steels and good machinability in hypereutectoid steels.

Stress-relief anneal. This treatment consists of heating below A_1 and cooling slowly. The objective is to relieve residual stresses, such as those introduced by cold forming and machining. Phase transformations are avoided in order to minimize distortion, and this treatment is used frequently to improve the dimensional stability of a complex part.

Normalizing. This treatment involves heating above A_3 or A_{cm} (cm indicates cementite) and air-cooling to room temperature. It is used for hypereutectoid steels to remove carbide networks which may have formed in previous processing (such as carburizing), and for hypoeutectoid steels to produce a material of intermediate strength with high ductility and low residual stresses. Typically such a treatment is specified for large shafts, made of hypoeutectoid steels, after forging. Normalizing is used to improve the machinability of hypoeutectoid steels, because the resultant microstructure contains pearlite, although the pearlite colonies will not be as well developed as those in a steel which has been given a full anneal.

In-process anneal. This is an annealing treatment introduced at the steel mill during the processing of steel bar, wire, sheet, tube, or other product. These may be spheroidizing, softening, normalizing, or stress relief anneals, depending on the ultimate use.

Hardening treatments. Steels are hardened by heating to a temperature at which austenite is formed and then cooling with sufficient rapidity to make the transformation to pearlite and/or ferrite unfavorable. The phase diagram is useful in determining suitable austenitizing, or hardening, temperatures, but the isothermal transformation curve, also called a TTT (time, temperature, transformation) curve, is more helpful in assessing the hardenability of a steel. A typical isothermal transformation diagram for a eutectoid steel austenitized at 1500°F (815°C) is shown in **Fig. 4**. This diagram is obtained experimentally for each combination of steel and austenitizing temperature by austenitizing small pieces, cooling rapidly to a holding bath, typically a lead or salt bath, and quenching into water after various times at the holding temperature. The constituents which form isothermally are determined by metallographic methods and are indicated on the diagram.

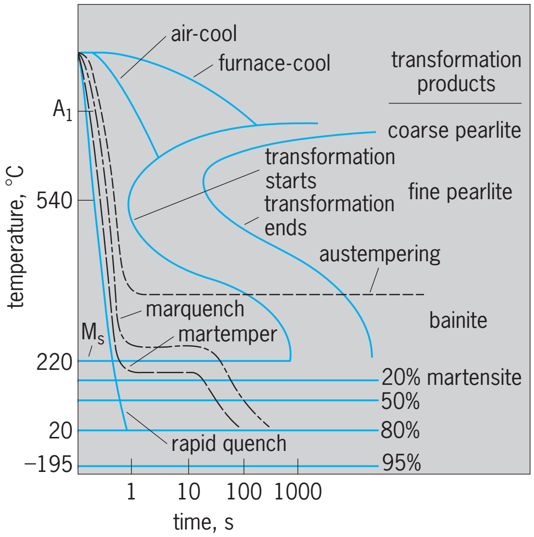
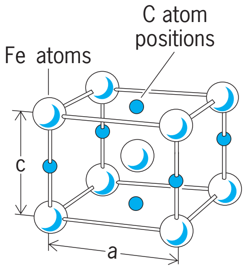


Fig. 4. Schematic isothermal transformation diagram for a eutectoid steel. °F = (°C × 1.8) + 32.

The superposition of typical cooling rates used in heat treatments onto the isothermal diagrams will indicate the constituents which will form initially on cooling. A continuous cooling transformation diagram is used to record all of the phases which form at various cooling rates, but such diagrams are difficult to determine and the isothermal curves are frequently used for this purpose. For example, furnace and air-cooling curves intersect the transformation curve above the nose of the C curve (Fig. 4), and the resultant microstructure will be pearlite. A rapid quench, in water or in some oils, will miss the pearlite nose, and martensite will form.

Martensite has a body-centered tetragonal crystal structure (**Fig. 5**) with the *c* and *a* dimensions dependent on the carbon content. Martensite can be formed only from austenite and contains all of the carbon and alloy elements which are dissolved in the austenite. The hardening of steel occurs by the formation of martensite, and the hardness of the martensite increases with the carbon content, reaching a plateau at 0.80% C.

Martensite begins to form at the M_s temperature, which is 520°F (270°C) in the example shown in Fig. 4. The percentage of martensite is shown as a series of horizontal lines on the transformation diagram



weight percent C	c (nm)	a (nm)
0	0.286	0.286
0.20	0.288	0.2858
0.40	0.291	0.2856
0.80	0.295	0.285

Fig. 5. Crystal structure of martensite.

since very little martensite formation occurs isothermally. At room temperature, such a steel will contain about 80% martensite by volume, with the remainder retained austenite. Some of the retained austenite may be transformed by cooling below room temperature, but the austenite cannot be eliminated solely by refrigeration, even to liquid helium temperatures. The M_s temperature is lowered, and the retained austenite contents increased, by increasing alloy and carbon contents of the austenite.

The position of the nose of the C-curve is dependent on the alloy and carbon contents, and this greatly influences the cooling rates required to harden a steel. In plain-carbon steels the nose is far to the left, and water or rapid oil quenches must be used. For medium-alloy steels (for example AISI 4340, which contains 0.40 C, 1.8 Ni, 0.8 Cr, 0.25 Mo) moderately thick sections may be hardened by oil quenching. For high-alloy steels (for example, type A-2 tool steel, which contains 1.0 C, 5.0 Cr, 1.0 Mo, 0.25 V) air cooling may be sufficient. The hardenability of steel is greater the farther to the right the position of the nose of the C-curve. Hardness should be distinguished from hardenability. Hardness refers to the resistance to indentation. Hardenability refers to the severity of cooling which must be used to produce martensite in a steel. If the steel can be hardened by using a slow cooling rate, such as air cooling, it has high hardenability. If a very rapid cooling rate, such as water quenching, must be used, it has low hardenability.

Freshly quenched martensite containing more than 0.10% C cannot be deformed easily, and a subsequent heating, or tempering, is used to increase the ductility. Microstructural changes occur on tempering, and the sequence of structures in low- and medium-alloy steels follows three stages. In the first, ϵ-carbide, which is hexagonal with the approximate composition Fe_2C, precipitates at temperatures up to 450°F (220°C). In the second stage, 410–590°F (210–310°C), the retained austenite decomposes to bainite, and in the third stage, 590–1000°F (210–540°C), ferrite and cementite are formed. Typically, cutting tools and bearings made of low-alloy steels are tempered through the first and second stages, and high-strength structural steels, such as AISI 4340, are tempered into the third stage in order to obtain sufficient ductility. The tempering sequence is somewhat different from high-alloy tool steels, such as type A-2 or the high-speed tool steels. In these steels, high-alloy carbides precipitate on tempering in the range 900–1000°F (490–550°C), lowering the M_s temperature of the retained austenite. The retained austenite transforms to martensite on cooling, and these steels thus undergo a secondary hardening on tempering in this range.

Austempering. Several variations in the quenching procedure are used to achieve special properties. Austempering involves quenching rapidly to a temperature below the pearlite nose, but above the M_s, as shown in Fig. 4. The objective is the formation of bainite, which is similar in form and has about the hardness of martensite tempered at the same temperature. Bainite, however, has better impact toughness and lower residual stresses than the corresponding martensite.

Marquenching and martempering. Marquenching involves quenching to a temperature just above the M_s and then air-cooling to room temperature. In martempering, the procedure is similar, but the interrupting temperature is just below M_s. The steels are then tempered in the normal fashion. These treatments are used to minimize the distortion which occurs in the direct quench to room temperature.

HSLA and dual-phase steels. High-strength low-alloy (HSLA) steels usually have a low carbon content (typically 0.02–0.10%) and sufficient content of alloy elements such as manganese, silicon, molybdenum, titanium, columbium, chromium, and vanadium to provide enough hardenability for the formation of low-carbon martensite in thin sheets. Such steels may harden on air-cooling from the austenitizing temperatures, although similar results may be obtained by water-quenching thin sheets of plain-carbon steel. Dual-phase (or duplex) steels are HSLA steels for which a particularly high degree of formability has been achieved by a special heat treatment. These steels are heat-treated by austenitizing in the $\alpha + \gamma$ phase region (this is often called an intercritical anneal) and cooling rapidly to room temperature. The temperature is chosen to result in a microstructure with approximately 80% ferrite and 20% martensite-plus-retained-austenite and this provides a combination of good formability and high strength in the finished part.

Surface treatments. Some heat treatments are used to alter the chemistry at the surface of a steel, usually to achieve preferential hardening of a surface layer. *See* SURFACE HARDENING OF STEEL.

Carburizing. Most carburizing is carried out in an atmosphere of partially combusted natural gas which is further enriched in carbon by additions of natural gas, or some other carbonaceous gas. The starting material is a low-carbon steel, and carbon diffuses into the outer layers, thus producing a duplex structure with a hypereutectoid steel in the case and a hypoeutectoid steel in the core. A hardening treatment may then be used to produce a hard case and a soft core.

Nitriding. Steels, usually containing aluminum, molybdenum, titanium, or chromium, are heated at 1000°F (540°C) in an atmosphere containing undissociated ammonia. Nitrogen diffuses into the surface, forming nitrides and causing a substantial increase in hardness of the surface layer. Additional hardening treatments are not required after the nitriding.

Chromizing. Chromium may be added to the surface by diffusion from a chromium-rich material packed around the steel or dissolved in molten lead. This process is quite slow and must be carried out at an elevated temperature. It is used to improve the corrosion resistance of the surface.

Nonferrous Metals and Alloys

Many nonferrous metals do not exhibit phase transformations, and it is not possible to harden them by means of simple heating and quenching treatments as in steel. Unlike steels, it is impossible to achieve grain refinement by heat treatment alone, but it is possible to reduce the grain size by a combination of cold-working and annealing treatments. Some nonferrous alloys can be hardened, but the mechanism is one by which a fine precipitate is formed, and the reaction is fundamentally different from the martensitic hardening reaction in steel. There are also certain ferrous alloys that can be precipitation hardened. However this hardening technique is used much more widely in nonferrous than in ferrous alloys. In titanium alloys, the β phase can transform in a martensitic reaction on rapid cooling, and the hardening of these alloys is achieved by methods which are similar to those used for steels.

Annealing treatments. This process involves heating the metal or alloy to an elevated temperature, particularly after cold-working operations such as wire drawing, deep drawing, cold forming, cold extrusion, or heavy machining, and cooling slowly to room temperature. The range of annealing temperatures for copper-base alloys is 660–1560°F (350–850°C), for aluminum alloys 560–800°F (290–430°C), and for nickel-base alloys 1470–2190°F (820–1200°C). The metals are softened by a recrystallization mechanism, which involves the nucleation of a new set of strain-free grains, and the subsequent growth of these strain-free grains at the expense of the cold-worked grains. The temperature at which recrystallization starts is not precisely defined and depends primarily on the amount of cold work, that is, deformation, the material experienced. Recovery from cold work can also be achieved without the formation of new grains by prolonged heating below the recrystallization temperature. Recovery treatments are used in instances where stress relieving is required for dimensional stability but full softening is undesirable because of the accompanying reduction in strength.

Precipitation and age hardening. Certain nonferrous alloys can be hardened by a precipitation-hardening treatment, sometimes called age hardening. In order to be amenable to such a treatment, the solubility of a phase or a compound in the alloy must be such as to be completely dissolved at a high temperature and supersaturated at a lower temperature. For example, one of the well-known precipitation-hardening alloys is an aluminum alloy containing 4% copper, 0.5% manganese, and 0.5% magnesium. It is solution-treated by heating at about 950°F (510°C) and quenching into water. At the solution temperature all of the copper is in solid solution, and the solid solution is retained on quenching to room temperature. The alloy is soft after this treatment (or solution anneal) and may be readily machined or deformed. If the alloy is subsequently aged at room temperature, the strength and hardness gradually increase, reaching a maximum in 5 or 6 days. The aging may be accomplished more rapidly by heating for several hours at 400°F (175°C).

During the age-hardening period, copper-rich regions form in the aluminum lattice and a precipitate which is coherent with the aluminum matrix is gradually developed. The coherency refers to a matching of the lattice spacings in the copper-rich cluster with those in the aluminum, and the resultant local strains produce the hardening of the alloy. If the alloy is overaged, by heating above the precipitation-hardening temperature, the equilibrium precipitate, $CuAl_2$, which is not coherent, will be formed and the hardness will be reduced. Overaging is utilized where dimensional stability is favored over achieving the maximum strength.

Some precipitation-hardening alloys achieve maximum strength only after a cold-working operation is introduced following the solution treatment and prior to the precipitation treatment. These treatments are usually carried out at the mill, and the material is purchased in the hardened condition.

A variety of heat treatments and thermomechanical treatments are used for aluminum alloys, and these are designated by letters and digits following the alloy designation. Some common designations are the following:

F = fabricated
O = annealed
H = strain-hardened
W = solution heat-treated
T1 = naturally aged only
T2 = annealed (case products only)
T3 = solution heat-treated, cold-worked, and naturally aged
T4 = solution heat-treated and naturally aged
T5 = artificially aged only}
T6 = solution heat-treated and artificially aged
T7 = solution heat-treated and stabilized (over aged)
T8 = solution heat-treated, cold-worked, and artificially aged
T9 = solution heat-treated, artificially aged, and cold-worked
T10 = artificially aged and cold-worked

See ALLOY; METAL, MECHANICAL PROPERTIES OF.

B. L. Averbach

Bibliography. American Society for Metals, *Heat Treating Source Book*, 1986; American Society for Metals, *Metals Handbook*, 9th ed., vol. 4: *Heat Treating*, 1981; American Society for Metals, *Steels: Heat Treatment and Processing Principles*, 1989; O. F. Devereux, *Topics in Metallurgical Thermodynam-*

ics, 1989; G. Krauss, *Principles of Heat Treatment of Steel*, 1980, reprint 2000; P. H. Morton (ed.), *Surface Engineering and Heat Treatment*, 1992.

Heating system

An apparatus consisting of an energy source, a method of converting that energy to heat, and a transport system to convey the energy and heat to the point of use. Most heating systems include some manual or automatic method of controlling the heat output and delivery.

Energy sources. There are many sources of energy for use in heating. The earliest source, and still most common in underdeveloped countries, comprises wood and wood products, such as paper, wood chips, and sawdust; peat is used in some cultures. Solar use for heating and electrical generation has become widespread, although economics discourages more general use. The generation of electrical energy requires the use of fossil fuels, water power, geothermal energy, or nuclear energy. Fossil fuels are used directly in furnaces and boilers. Waste heat from various processes is often used to serve other parts of a common facility. For example, the methane gas generated by sewage disposal plants is sometimes used in special boilers to speed up the sewage reaction process. The heat pump is an adaptation of the refrigeration cycle in which the rejected heat is put to beneficial use. *See* ELECTRIC POWER SYSTEMS; ENERGY SOURCES; HEAT PUMP; SOLAR ENERGY.

Energy conversion. The energy source is converted into heat by various means. Wood and fossil fuels are converted by burning, that is, the combustion process. All combustion processes generate substances that are regarded as pollutants, and are more or less inefficient. Thus, there are frequent attempts to use waste heat and maximize combustion efficiencies. Modern standards and codes require that pollutants be removed to the greatest extent possible. This requirement adds to costs and sometimes has an adverse effect on efficiency. Electrical energy can be converted directly into heat by means of resistance heaters. Heat pump systems typically use electrical energy to drive the refrigeration compressor and fans; part of the useful heat is waste heat from the refrigeration process, but most of the heat comes from the air or water source. Solar energy requires collectors, which convert it into heat or electricity.

Transport systems. Heating systems use many methods to deliver heat to the point of use. Radiation systems take several forms. Cast-iron column radiators, using steam or hot water, have largely been superseded by convector radiators using steam, hot water, or electricity. Panel-type radiators are also used in ceilings or in floors. All require a piping or electrical distribution system. Forced-air warm-air heating, using electric motor–driven circulating fans, is very common. Residential warm-air heating systems use warm air with heat supplied by a central furnace burning fossil fuels or by a heat pump. Commercial, institutional, and industrial systems usually use central boilers for heat generation, with distribution through piping systems to heat exchangers at forced-air handling units, which deliver heated air to the point of use. In most cases the air systems also provide cooling for year-round air conditioning. In some cities, district heating systems supply heat through piping systems from a central plant to many users. *See* COMFORT HEATING; DISTRICT HEATING; HOT-WATER HEATING SYSTEM; PANEL HEATING AND COOLING; RADIANT HEATING; SOLAR HEATING AND COOLING; STEAM HEATING; WARM-AIR HEATING SYSTEM. Roger W. Haines

Bibliography. American Society of Heating, Refrigerating and Air Conditioning Engineers, *ASHRAE Handbooks: HVAC Systems and Equipment*, 1992, *Fundamentals*, 1993.

Heavy minerals

Minerals with a density greater than 2.9 g/cm^3. The term is most commonly used to denote high-density components of siliciclastic sediments. Most heavy mineral studies are undertaken to determine sediment provenance, because heavy mineral suites provide important information on the mineralogical composition of source areas (see **table**). Since heavy minerals rarely constitute more than 1% of sandstones, their study normally requires them to be concentrated. This is achieved by disaggregation of the sandstone, followed by mineral separation using dense liquids such as bromoform, tetrabromoethane, or the more recently developed nontoxic polytungstate liquids. *See* DENSITY; PROVENANCE (GEOLOGY); SANDSTONE.

Geographic and stratigraphic variations in heavy mineral suites within a sedimentary basin can be used to infer differences in sediment provenance. Such differences result either from the interplay between a number of sediment transport systems draining different source regions, or from erosional unroofing within a single source area. Heavy mineral data therefore play an important role in the understanding of depositional history and paleogeography. In some cases, sophisticated mathematical and statistical treatment of heavy mineral data may be required to elucidate the interplay between multiple sediment transport systems. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; PALEOGEOGRAPHY; SEDIMENTOLOGY; STRATIGRAPHY.

The recognition of changes in provenance within a sedimentary sequence provides a basis for correlation of strata that is independent of more traditional biostratigraphic methods. Heavy mineral assemblages have proven to be especially useful for correlating sandstones that lack age-diagnostic fossils, especially nonmarine sequences. *See* SEQUENCE STRATIGRAPHY.

The identification of provenance on the basis of heavy mineral data requires careful consideration of all factors influencing the composition of the assemblages. A number of processes may alter the original

Provenance of the most commonly recorded nonopaque detrital heavy mineral species*

Mineral	Density, g/cm^3	Hardness (Mohs scale)	Stability in acidic weathering	Stability in burial diagenesis	Most common provenance
Andalusite	3.13–3.16	$6\frac{1}{2}$–$7\frac{1}{2}$	High	Low	Metapelites
Amphibole	3.02–3.50	5–6	Low	Low	Various igneous and metamorphic rocks
Apatite	3.10–3.35	5	Low	High	Various igneous and metamorphic rocks
Chloritoid	3.51–3.80	$6\frac{1}{2}$	Moderate	Moderate	Metapelites
Chrome spinel	4.43–5.09	$7\frac{1}{2}$–8	High	High	Ultramafic igneous rocks
Clinopyroxene	2.96–3.52	5–$6\frac{1}{2}$	Low	Low	Basic igneous and metamorphic rocks
Epidote group	3.12–3.52	6–$6\frac{1}{2}$	Low	Low	Low-grade metamorphic rocks
Garnet	3.59–4.32	6–$7\frac{1}{2}$	Moderate	Moderate-high	Metasediments
Kyanite	3.53–3.65	$5\frac{1}{2}$–7	High	Low-moderate	Metapelites
Monazite	5.00–5.30	5	High	High	Metamorphic rocks; granites
Rutile	4.23–5.50	6–$6\frac{1}{2}$	High	High	High-grade metamorphic rocks
Sillimanite	3.23–3.27	$6\frac{1}{2}$–$7\frac{1}{2}$	High	Low	Metapelites
Staurolite	3.74–3.83	$7\frac{1}{2}$	High	Moderate	Metapelites
Titanite	3.45–3.55	5	Moderate	Low-moderate	Various igneous and metamorphic rocks
Tourmaline	3.03–3.10	7	High	High	Metasediments: granites
Zircon	4.60–4.70	$7\frac{1}{2}$	High	High	Granites and other acidic igneous rocks

*High-stability minerals (such as zircon, rutile, and tourmaline) are commonly recycled.

provenance signal at various points in the sedimentation cycle. These include (1) weathering in the source region, which may affect the mineralogy prior to incorporation in the transport system; (2) abrasion during transport, which may cause destruction of mechanically unstable minerals; (3) weathering during periods of alluvial storage on the floodplain, which may cause depletion of chemically unstable components; (4) hydrodynamics during transport and at the final depositional site, which may affect the relativ e abundance of minerals with different hydraulic behavior; and (5) diagenesis, which may cause depletion of minerals that are unstable in the subsurface. The effects of diagenesis are particularly pervasive, with several case studies (for example, North Sea, United States Gulf Coast, Bay of Bengal) showing that mineral suites become progressively less diverse as burial depth increases. This is the result of dissolution of unstable minerals by circulating pore waters, a process known as intrastratal solution. *See* DIAGENESIS; WEATHERING PROCESSES.

Single-grain heavy mineral geochemical analysis (for example, by electron microprobe) gives additional sophistication to provenance determination. For instance, tourmaline compositions can be used to distinguish granitic and metasedimentary sources, and constraints on plate-tectonic settings of sedimentary basins have been inferred from geochemical analysis of detrital augite. Garnet compositions are particularly useful in provenance studies in view of their response to differences in metamorphic grade and protolith composition. Some heavy minerals, notably zircon and monazite, can be dated radiometrically, and the combined mineralogical and isotopic characterization of sediment source areas is a very powerful tool for provenance studies. *See* AUGITE; ELECTRON MICROSCOPE; GARNET; GRANITE; TOURMALINE.

Heavy minerals have important economic applications. Their use in paleogeographic reconstructions, especially in elucidating sediment transport pathways, is of particular value in hydrocarbon exploration, and their use in correlation has important applications in hydrocarbon reservoir evaluation and production. Recent advances have made it possible to utilize the technique on a real-time basis at the well site, where it is used to help steer high-angle wells within the most productive reservoir hori zo ns. Heavy minerals may become concentrated naturally by hydrodynamic sorting, usually in shallow marine or fluvial depositional settings. Naturally occurring concentrates of economically valuable minerals are known as placers, and such deposits have considerable commercial significance. Cassiterite, gold, diamonds, chromite, monazite, and rutile are among the minerals that are widely exploited from placer deposits. *See* DATING METHODS; MONAZITE; PLACER MINING; WELL; ZIRCON.

Andrew Morton

Bibliography. M. A. Mange and H. F. W. Maurer, *Heavy Minerals in Colour*, Chapman and Hall, London, 1992; A. C. Morton and C. R. Hallsworth, Processes controlling the composition of heavy mineral assemblages in sandstones, *Sed. Geol.*, 124:3–29, 1999.

Heavy water

A form of water in which the hydrogen atoms of mass 1 (^{1}H) ordinarily present in water are replaced by deuterium (symbol D or ^{2}H), the heavy stable isotope of hydrogen of mass 2. The molecular formula of heavy water is D_2O (or ^2H_2O).

Properties. Because the mass difference between ^{1}H and ^{2}H is the largest for any pair of stable (nonradioactive) isotopes in the periodic table, many of

TABLE 1. Physical properties of ordinary and heavy water

Property	^1H_2O	^2H_2O (D_2O)
Molecular weight, ^{12}C scale	18.015	20.028
Melting point, °C	0.00	3.81
Normal boiling point, °C	100.00	101.42
Temperature of maximum density, °C	3.98	11.23
Density at 25°C, g/cm^3	0.99701	1.1044
Critical constants		
Temperature, °C	374.1	371.1
Pressure, mPa	22.12	21.88
Volume, cm^3/mol	55.3	55.0
Viscosity at 55°C, mPa · s	0.8903	1.107
Refractive index, $n^{20}{}_D$	1.3330	1.3283

the physical and chemical properties of the pure isotopic species and their respective compounds differ to a significant extent. Selected physical properties of ^1H_2O and ^2H_2O are compared in **Table 1**.

Heavy water, judging from its higher melting and boiling points, its higher viscosity, and its surprisingly high temperature of maximum density, is a distinctly more structured liquid than is ordinary water. Heavy water is more extensively hydrogen-bonded, and the hydrogen bonds formed by ^2H are somewhat stronger than are those of ^1H. Nonpolar solutes induce structure in ^2H_2O to a greater extent than in ^1H_2O, and structure-breaking salts are more disruptive in ^2H_2O, largely because there is more structure to break. With increasing temperature, structure is broken down more rapidly in ^2H_2O than in ^1H_2O, and as a result ^2H_2O may be more structured or less structured than is ^1H_2O, depending on the temperature. Many ionic solutes are distinctly less soluble in D_2O than in H_2O.

The nuclear properties of ^1H and ^2H are very different. The capture cross section (Σ_a) of deuterium for low energy of neutrons is much smaller than that of ordinary hydrogen. For ^1H_2O, Σ_a is 2.2×10^{-2} cm^{-1} as compared with the value of 8.5×10^{-6} cm^{-1} for ^2H_2O. The nuclear spin of ^1H is $^1/_2$, but that of ^2H is 1. Thus, both isotopes can be used in nuclear magnetic resonance spectroscopy. The resonance frequency for ^2H is much lower than that of ^1H at a given magnetic field, and the relative sensitivity for detection is much higher for ^1H than for ^2H.

The principal difference in chemical behavior between ^1H and ^2H derives from the generally greater stability of chemical bonds formed by ^2H. The most important factor contributing to the difference in bond energy is the lower zero-point vibrational energy (of the order of 5.021–5.275 kilojoules/mol) for chemical bonds formed by ^2H. Both kinetic and static isotope effects result. The ion product constant of ^2H_2O is 1.11×10^{-15} at 25°C (77°F) compared to 10^{-14} for ^1H_2O. The difference by a factor of 10 in ionization constant causes values of pH and pD for solutions of identical composition in the two media to differ significantly. The nonidentity of pH and pD for solutions of a given composition can be critical for phenomena sensitive to hydrogen ion activity, and can make for serious problems in the interpretation of experiments in many biological systems. In the pH range 2–9, it has been established that pD = pH + 0.41 (molar scale; 0.45 molal scale) as measured by an ordinary glass electrode.

Replacement of more than one-third of the ^1H by ^2H in the body fluids of animals by administration of heavy water, or two-thirds of the hydrogen in higher plants, is lethal. However, numerous green and blue-green algae have been successfully cultured in >99.5% ^2H_2O on carbon dioxide and inorganic nutrients. Fully deuterated organisms can, in turn, be used to start a food chain for the growth of nutritionally more demanding bacteria, molds, and even protozoa in fully deuterated form. These organisms contain >99.5% D, and cultures of these organisms of unnatural isotopic composition have been successfully maintained for years in a fully deuterated form.

Preparation. Deuterium is present in ordinary water to the extent of 0.0145%. Water is the only practical source of deuterium, and very large amounts of water must be processed to produce relatively small amounts of heavy water. Deuterium can be extracted from water (as highly enriched ^2H_2O) by electrolysis, by distillation, or by chemical exchange. Electrolysis of ordinary water is very costly in energy, and has been used only for small-scale production or to enrich partially concentrated material. Distillation of water has been successfully used to separate highly enriched D_2O from ordinary water, but again the energy costs are high per unit of separative work performed. More efficient is the distillation of liquid hydrogen at cryogenic temperatures. Excellent fractionation factors can be achieved in the distillation of liquid hydrogen. The power requirements are modest, but the requirements for very large hydrogen feeds of very high purity (traces of oxygen, nitrogen, carbon monoxide, and so on, are solids at liquid hydrogen temperatures and may clog the apparatus) have mitigated against large-scale use. As the production of hydrogen for synthetic ammonia or coal liquefaction expands, however, by-product extraction of deuterium from the hydrogen stream by distillation of liquid hydrogen may well become important.

Various chemical exchange reactions have been considered for concentrating deuterium from natural water (**Table 2**). Isotope exchange reactions between hydrogen gas and water or ammonia, or between hydrogen sulfide and water, provide the best point of departure for the large-scale manufacture of heavy water. The H_2O/H_2S process is the chemical exchange process of principal commercial interest. Because the equilibrium constant for the distribution of deuterium between H_2S gas and liquid H_2O is temperature-dependent, the process can be carried out in the form of a dual-temperature exchange process. In the dual-temperature H_2O/H_2S process, exchange of ^2H between H_2S (gas) and H_2O (liquid) is carried out at elevated pressure (∼2 megapascals or 20 atm) at 120–140°C (250–280°F); ^2H displaces ^1H in the H_2S and becomes concentrated in the gas. At 30°C (86°F) the equilibrium is shifted and reversed,

TABLE 2. Chemical exchange reactions for concentrating deuterium from water

	Equilibrium constant	
Reaction	Low temp. (°C)	High temp. (°C)
$H_2O(l) + HDS(g) \rightleftharpoons NDO(l) + H_2S(g)$	2.18 (20°)	1.83 (130°)
$NH_3(l) + HD(g) \overset{\text{catalyst}}{\rightleftharpoons} NH_2D(l) + H_2(g)$	6.60 (–50°)	4.42 (0°)
$H_2O(g) + HD(g) \overset{\text{catalyst}}{\rightleftharpoons} HDO(g) + H_2(g)$	3.62 (25°)	2.43 (125°)

and ^{2}H is stripped from the gas into the liquid water. The dual-temperature exchange process is carried out in a pair of gas-liquid contacting towers; the cold tower operates at 30°C (186°F), the hot tower at 120–140°C (250–280°F). Water fed to the system flows downward through the cold tower and then through the hot tower countercurrent to a stream of hydrogen sulfide. The water is progressively enriched in deuterium as it passes through the cold tower and progressively depleted in transit through the hot tower. Deuterium-enriched water is drawn from the bottom of the cold tower, and H_2S enriched in deuterium is removed from the top of the hot tower. Further enrichment then occurs in a second stage. To produce 1 metric ton of D_2O, the plant must process 41,000 tons (37,000 metric tons) of water and must cycle 135,000 tons (120,000 metric tons) of H_2S. The product from the exchange plant is enriched to around 15% deuterium, and enrichment to 99.5% is accomplished by vacuum distillation. The dual-temperature GS H_2S/H_2O exchange process is used by all heavy-water plants producing more than 20 metric tons of D_2O per year.

Analysis. The principal methods for determining the deuterium content of heavy water are density determinations, infrared spectroscopy in the ~3-micrometer region, and absorption spectrophotometry in the near-infrared region of the spectrum between 1 and 2 μm. Mass spectroscopy, interferometry, falling-drop methods, and nuclear magnetic resonance spectroscopy have also been used for determining the $^1H/^2H$ ratio in heavy water. Absorption spectrophotometry in the near infrared is a particularly useful procedure because it can be carried out in conventional recording spectrophotometers.

Uses. The only large-scale use of heavy water in industry is as a moderator in nuclear reactors. The Canadian heavy-water-moderated natural-uranium-fueled CANDU reactor uses 0.94 ton (0.85 metric ton) of heavy water per electrical megawatt of installed capacity. The heavy water is not consumed by reactor operations, and the demand for heavy water is determined by the rate at which new nuclear power stations are built. *See* NUCLEAR REACTOR.

Deuterium may become important in energy production by nuclear fusion by the reaction $^2_1H \rightarrow {}^3_1H + {}^1_1H + 4.0$ MeV. It has been estimated that the entire annual energy requirements for the United States in the year 2020 could be supplied by the D-D nuclear fusion reaction by the amount of deuterium contained in 5500 tons (5000 metric tons) of D_2O. *See* NUCLEAR FUSION.

Small amounts of heavy water are used to grow fully deuterated organisms, which serve as a source of fully deuterated compounds of biological importance. These are used in research techniques such as small-angle neutron scattering, in high-resolution nuclear magnetic resonance spectroscopy of immobilized samples, and in the study of isotope effects. *See* DEUTERIUM; ISOTOPE SEPARATION; LIQUID; TRITIUM; WATER.

Joseph J. Katz

Bibliography. J. J. Katz and H. L. Crespi, in *Isotope Effects in Chemical Reactions*, ACS Monogr. 167, 1971; I. Kirshenbaum, *Physical Properties and Analysis of Heavy Water*, 1951; G. M. Murphy (ed.), *Production of Heavy Water*, 1955; H. K. Rae (ed.), *Separation of Hydrogen Isotopes*, ACS Symp. Ser. 68, 1978.

Hederellida

A marine bryozoan order in the class Stenolaemata. It is recorded from scattered localities in the Ordovician through Permian systems of the Paleozoic Era but is found prominently in Devonian strata. The colonies comprise tubular zooecia that commonly encrust brachiopod shells, echinoderm plates and columnals, and the outer surfaces of corals. Although usually encrusting, the colonies sometimes may be erect structures. The initial zooecium is a bulbous ancestrula, which gives rise to a tubular zooecium. Succeeding tubular zooecia bud, one at a time, from the lateral wall of the preceding zooecium. The tubular zooecia have perforated walls, and the distal opening of the zooecium is commonly sealed by a plate (possibly perforate). The zooecial wall apparently consisted of two layers, but it lacks the distinctive laminate structures of the order Trepostomata (Stenolaemata) and shows no additional thickening in the outer parts of the colony.

The simple, perforate tubular zooecia, in which the commonly oval zooecial opening is the same diameter as the zooecial tube, and the lack of accessory tubes or structures separate the hederellids from the Paleozoic cyclostomes. There are six recognized genera, including *Hederella, Reptaria, Hederopsis*, and *Clonopora*.

June R. P. Ross

Bibliography. R. S. Bassler, The Hederelloidea, a suborder of Paleozoic cyclostomatous Bryozoa, *Proc. U.S. Nat. Mus.*, 87:25–91, 1939; R. A. Robison (ed.) *Treatise on Invertebrate Paleontology*, revised, 1983.

Hedgehog

Members of the family Erinaceidae (order Insectivora), which includes 7 genera and 19 species of medium- to large-size animals generally

Hedgehog (*Erinaceus europaeus*).

characterized by spines on their back and sides. This family, however, is subdivided into two subfamilies: the spiny hedgehog group which includes the European hedgehog (*Erinaceus europaeus*); and the hairy hedgehogs or gymnuras, long-tailed species which lack spines. The typical spiny hedgehog has well-developed eyes and a rudimentary to moderately long tail. In most species, there are five toes on each foot, although in some species these are reduced to four on the hindfeet. The spines can be erected by strong muscles, and serve as a protection for the naked or hairy belly, head, and limbs when the animal rolls itself into a ball.

The European hedgehog (see **illus.**) grows to a length of about 9 in. (23 cm). It is a nocturnal animal and rests in thickets or undergrowth, or in crevices during the day. It has 36 sharp teeth suited to its omnivorous diet, which includes worms, insects, reptiles, small rodents, and mollusks. The dental formula is I 3/2 C 1/1 Pm 3/2 M 3/3. The adult is relatively immune to the poison of venomous snakes; thus hedgehogs are valuable in controlling the viper population. The gestation period is 30–50 days and there may be two litters each year of from two to seven young, which are born blind and with soft spines. The young are weaned after a month, and reach sexual maturity at 10 months. *See* DENTITION.

The hairy hedgehogs such as the moon rat (*Echinosorex gymnurus*) are found in southern Asia, Borneo, and Sumatra. This is one of the four species of this little-known group of ratlike animals. The body is over a foot long, which makes it one of the largest living species of insectivore. These animals have a characteristic odor due to the presence of scent glands. It is thought that the diet consists of insects and fruits. *See* INSECTIVORA; SCENT GLAND. Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Helicity (quantum mechanics)

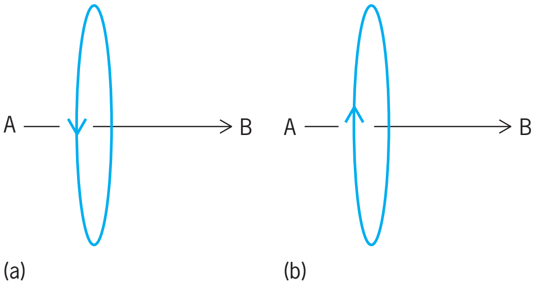
A quantum-mechanical variable that specifies the component of spin-angular momentum of a particle along its direction of motion. The helicity of a particle with rest mass depends on the reference frame, because such a particle has velocity less than the velocity of light c in vacuum. Hence one can make a Lorentz transformation along the direction of motion to a frame in which the particle moves in the opposite direction. For a massless particle, reversal in direction of motion by Lorentz transformation is impossible, and helicity is a Lorentz-invariant quantity. For example, a beam of circularly polarized light consists of photons (zero-mass quanta of the electromagnetic field) with helicity $h = +1$ for right circular polarization, or $h = -1$ for left circular polarization (see **illustration**). *See* LORENTZ TRANSFORMATIONS; PHOTON; POLARIZATION OF WAVES; POLARIZED LIGHT; RELATIVITY; SPIN (QUANTUM MECHANICS).

According to quantum mechanics, the spin angular momentum of a spin-$^1/_2$ particle (for example, the electron or muon) can have only two possible values relative to any given axis: parallel or antiparallel. Thus the helicity of a spin-$^1/_2$ particle can always be thought of as a superposition of two components: a "left-handed" or "left-chirality" portion (spin opposed to direction of motion) and a right-chirality portion (spin parallel to direction of motion). [the word chirality means handedness]. *See* NONRELATIVISTIC QUANTUM MECHANICS.

Neutrinos (ν) and antineutrinos ($\bar{\nu}$) are spin-$^1/_2$ particles of very small rest mass that are produced and absorbed only in weak interactions, and then only in their left-chirality and right-chirality states, respectively. If ν and $\bar{\nu}$ were zero-mass particles (and most physicists believed this to be possible until 1998), their helicities and chiralities would be identical, and we would have $h(\nu) = -1$, $h(\bar{\nu}) = +1$. However, since 1998 neutrino-oscillation experiments have given strong evidence that at least two of three neutrino and antineutrino species have distinct rest masses greater than zero. Thus, although the weak coupling of a neutrino (antineutrino) occurs only in its left (or right) chirality state, the particle, once generated, propagates in a superposition of left- and right-chirality states, yielding a helicity that depends on the reference frame. For example, it is useful to consider the weak interaction (1) in

$$n \rightarrow p + e^- + \bar{\nu}_e \qquad (1)$$

which a neutron decays to a proton, an electron, and



Schematic diagrams of circularly polarized light. (*a*) Right circular polarization. The arrow AB indicates the direction of light propagation. The arrow on the circle indicates the "rotation" associated with the angular momentum of the photon. This angular momentum is parallel to AB. In the description given by classical optics, the arrow on the circle indicates the sense of rotation of the electric and magnetic vectors in the electromagnetic field. (*b*) Left circular polarization. The symbols have the same meaning, but the angular momentum of the photon is in the opposite direction.

an electron-antineutrino. According to the standard theory of weak interactions, modified to account for finite neutrino mass, the average value of helicity for an ensemble of antineutrinos with velocity $v(\bar{\nu})$ is given by Eq. (2), where <...> means average. Since

$$\langle h(\bar{\nu})\rangle = +\frac{v(\bar{\nu})}{c} \qquad (2)$$

in the neutron rest frame virtually all emitted neutrinos in neutron beta decay have kinetic energy much greater than their rest energy, v is very nearly equal to c for these antineutrinos, and thus $\langle h(\bar{\nu})\rangle \cong +1$. In other words, in the neutron rest frame the helicity is nearly, but not exactly, the same as it was in general when a mass $m(\bar{\nu}) = 0$ was assumed. *See* NEUTRINO; WEAK NUCLEAR INTERACTIONS.

Chirality can be an important concept even in circumstances where helicity is irrelevant. For example, in chemistry one encounters examples of two distinct molecules (enantiomorphs) which have the same chemical formula but are mirror images of one another. The spatial structure of one is left-handed (has left chirality), while the other has right chirality. *See* AMINO ACIDS; STEREOCHEMISTRY.

Eugene D. Commins

Bibliography. A. Bertin, R. A. Ricci, and A. Vitale, *Fifty Years of Weak Interaction Physics*, Societa Italiana di Fisica, Bologna, 1984; E. D. Commins and P. H. Bucksbaum, *Weak Interactions of Leptons and Quarks*, Cambridge University Press, 1983; F. Halzen and A. D. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, Wiley, New York, 1984; J. J. Sakurai, *Advanced Quantum Mechanics*, Addison-Wesley, Reading, MA, 1978.

Helicobacter

A genus of gram-negative bacilli whose members are spiral shaped, showing corkscrewlike motility generated by multiple, usually polar flagella. *Helicobacter* require low concentrations of oxygen for maximum growth and produce the enzymes oxidase, catalase, and urease.

Different species of *Helicobacter* are found in the stomachs of different animals: *H. felis* in cats and dogs, *H. mustelae* in the domestic ferret, *H. nemestrinae* in the pigtailed macaque, and *H. acinonyx* in captive cheetahs with gastritis. The species of *Helicobacter* found in the human stomach, *H. pylori*, is extremely common. In the United States and similarly developed countries, its prevalence increases at about 1% per year of age so that the majority of adults above age 50 are infected. In less developed countries, infection rates are dramatically higher, with up to 80% of children infected.

Helicobacter pylori is present in virtually all cases of chronic gastritis, which progresses slowly (years or decades) from asymptomatic to atrophic gastritis with impaired acid secretion. Virtually all individuals with duodenal ulcers are infected with *H. pylori*, which colonize sites in the duodenum. The termination of treatment for duodenal ulcers leads to a high rate of recurrence of the ulcers, but ulcer treatment plus eradication of *H. pylori* from the stomach usually leads to a permanent cure. A significant proportion of individuals with *H. pylori*–associated atrophic gastritis develop intestinal-cell metaplasia in the stomach, a condition which is known to represent a precancerous state. Epidemiological studies have confirmed that 75–80% of individuals with gastric cancer have *H. pylori* gastritis. However, the low incidence of duodenal ulcer and gastric or duodenal cancer in *H. pylori*–infected individuals indicates that *H. pylori* infection, although important, is only one of many risk factors leading to these conditions; other risk factors may include diet and variations in the efficiency of the immune system. *See* CANCER (MEDICINE); ULCER.

Helicobacter pylori virulence factors include its shape and ability to rapidly move into and through the gastric mucous coating, which protects the organism from stomach acid, its surface-associated urease enzyme which neutralizes stomach acid near the organism, a cytotoxin, and a fibrillar adhesin which binds the organism to the surface of gastric epithelial cells. *Helicobacter pylori* also has the ability to survive in large numbers in spite of antibodies which are secreted into the stomach and the host immune cell response (inflammatory response) which is characteristic of gastritis.

Detection of *H. pylori* infection is not difficult. Procedures include detection of circulating anti–*H. pylori* antibody in the blood; histological examination of stomach biopsies in which *H. pylori* may be seen residing in gastric mucus or attached to gastric epithelial cells; cultivation of *H. pylori* from a stomach biopsy; and detection of the potent urease enzyme of *H. pylori* by direct testing of biopsy or indirectly by breath test.

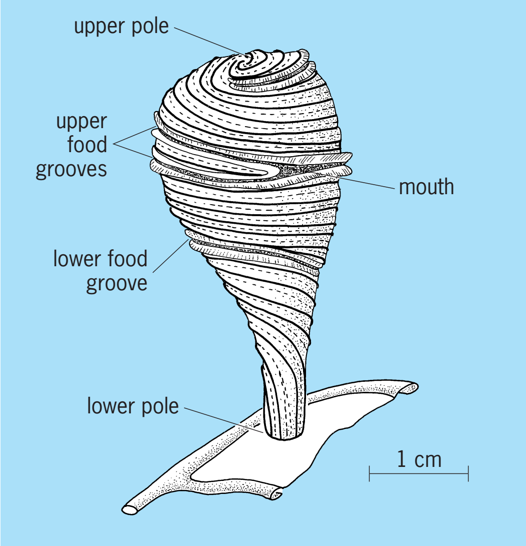
Eradication of *H. pylori* from the stomach is difficult. The most successful therapies are combinations of bismuth compounds plus two or three different antibiotics. *See* MEDICAL BACTERIOLOGY; TOXIN.

Doyle J. Evans, Jr.

Bibliography. S. Baron, *Medical Microbiology*, 4th ed., 1996.

Helicoplacoidea

A small class of spindle-shaped, spirally pleated, primitive echinoderms in the subphylum Echinozoa, from the Early Cambrian (*Nevadella* Zone) in eastern California, western Nevada, and eastern British Columbia. Since their discovery in the early 1960s, three genera and six species have been described based on nearly 600 complete specimens, making helicoplacoids the most diverse and abundant echinoderm class known from the Early Cambrian. Most helicoplacoids have a spindle-shaped theca or body with diagonally spiraled pleats, each made up of three rows of flexibly sutured plates forming an interambulacral ridge. Smaller-plated ambulacra spiral around the theca every 7–13 pleats; the ambulacral



***Helicoplacus gilberti*, side view, reconstructed in living position. The mouth is inferred to be where the ambulacral branches join, and the cover plates over the food grooves are in an open (feeding) position. The lower pointed pole of the helicoplacoid is attached to a trilobite fragment lying on the sea floor. (*After K. Derstler*)**

plates cover a central food groove that has pores for tube feet on each side. The original authors inferred that the mouth was at the more rounded pole of the theca and that a single long ambulacrum branched once about one-third of the way down the theca. However, it has been argued that the mouth was located where the ambulacra split (see **illus.**), so that three ambulacral branches lead away from it, one spiraling partway down the theca and two spiraling up the theca a few pleats apart. No anal opening or pyramid has been found on any of the known specimens.

Helicoplacoids have been reconstructed either as lying recumbent on the sea floor or as standing erect and attached to objects on the sea floor by using the more pointed pole of the theca (see illus.). They were apparently low- to medium-level suspension feeders using their ambulacral tube feet to capture small food particles drifting past the theca. The covered ambulacra bearing pores for tube feet are similar to those of other echinozoans such as edrioasteroids and camptostromatoids. The unusual plating and spiral symmetry of helicoplacoids separate them from all other classes of echinoderms. *See* ECHINODERMATA; ECHINOZOA.

James Sprinkle

Bibliography. K. Derstler, *Studies on the Morphological Evolution of Echinoderms*, Ph.D. dissertation, University of California, Davis, 1985; J. W. Durham and K. E. Caster, *Helicoplacoidea*: A new class of echinoderms, *Science*, 140:820–822, 1963.

Helicopter

An aircraft characterized by its large-diameter, powered, rotating blades, attached to a substantially vertical axis. The helicopter can lift itself vertically by the reactive force generated as the rotating blades accelerate air downward. It can both lift and propel itself by accelerating air downward at an angle to the vertical. The helicopter is the most successful vertical takeoff and landing (VTOL) aircraft developed, by virtue of its relatively high efficiency in performing hovering and low-speed flight missions.

It was not until the 1930s that helicopters began to demonstrate practical capabilities. In 1937 Heinrich Focke built in Germany a twin-rotor helicopter, which ultimately demonstrated meaningful performance capabilities. In 1939 Igor Sikorsky built a relatively simple and controllable single-rotor helicopter, which evolved into the modern standard configuration. In 1946 a single-rotor machine developed by Lawrence Bell received the world's first commercial helicopter license.

Dynamics. The key to understanding the operation and control of a helicopter lies in a knowledge of the forces and resultant motion of each rotor blade as momentum is imparted to the air. Unlike a fixed-wing aircraft, which derives its lift from the translational motion of the fuselage and airfoil-shaped wing relative to the air, the helicopter rotates its wings (or rotor blades) about a vertical shaft and thus is able to generate lift while the fuselage remains stationary.

The rotational motion of the rotor blades creates additional forces that act on the blades as they revolve about the shaft. The principal additional force is the centrifugal force that arises from the circular motion of the blade mass. The centrifugal force creates an effective stiffening of the rotor blade so that a relatively limber blade structure can carry the aerodynamic forces necessary to lift the weight of the helicopter, similar to the phenomenon of supporting a rock attached to the end of a string by whirling it in the horizontal plane.

The major forces on an element of a rotor blade are the lift, drag, weight, and centrifugal force (**Fig. 1**). The forces of greatest interest are the lift, drag, and centrifugal forces, since the blade weight is small relative to the other forces. In order for the rotor to lift the helicopter in hovering flight, the average distributed lift from all blades must equal the weight of the helicopter. With these large lift forces distributed along a blade, it will cone upward until there is a balance of the moments created by the lift and centrifugal forces about the blade attachment point. In hovering flight, both the lift and drag forces are steady, and the blade cones to a constant, equilibrium position. In forward flight, these forces vary as the blade rotates; consequently, the blade's angular position relative to the hub, called the flapping angle (Fig. 1), is a function of these oscillatory aerodynamic forces. In essence, the rotor-blade flapping motion may be considered as a dynamics problem analogous to the forced response of a simple spring-mass-damper system, wherein the centrifugal force is equivalent to the restoring spring and the aerodynamic force provides both the damping and the forcing function for the blade mass.

A unique characteristic of the blade dynamic system just described is that the flapping natural frequency of the blade is approximately equal to the rotational frequency of the rotor. Since the aerodynamic forcing function is also proportional to the rotational frequency, the rotor blade represents one of the few dynamic systems that is forced at its natural frequency or in resonance. This can be a catastrophic occurrence for mechanical systems without damping due to the large amplitudes of motion that may be generated. In the case of a rotor blade, however, the aerodynamic damping generated by the blade flapping motion is so high that it limits the forced response to acceptable amplitudes.

A major consequence of operating at flapping resonance is the resulting relationship between the maximum force on the blade and its maximum displacement. When a dynamic system is forced at its natural frequency, there is a 90° phase lag between the phase angle of the maximum force and the phase angle at which the maximum amplitude of the response occurs. Thus, when a rotor blade experiences a change in the aerodynamic loading, the blade responds correspondingly one-quarter of a revolution (that is, 90°) later. This behavior is important for controlling the flight motion of the helicopter.

For example, to achieve forward flight, the rotor must be tilted in the direction of the desired flight path. This is achieved by decreasing the lift on a blade rotating toward the nose (advancing blade) and increasing the lift on a blade rotating toward the tail of the helicopter (retreating blade). This difference in lift on the advancing and retreating blades, which is maximum when the blades are arrayed laterally, will cause the advancing blade to flap to a lower amplitude when it arrives over the nose while the retreating blade with more lift flaps to a higher amplitude when it arrives over the tail, thus creating the forward tilt of the rotor.

The method for achieving the desired control of the rotor tilt is to vary the angle of the blade (pitch) as the rotor rotates. This variation is achieved by a control device called a swash plate, which, when tilted, will cycle the blade pitch angle in a sinusoidal manner so that blade lift is alternately increased and decreased once every rotor revolution. As discussed below, this cyclic control is also required to accommodate the lift variations encountered when the rotor is in forward flight. *See* AERODYNAMIC FORCE.

Aerodynamics. The basic physics of rotor aerodynamics, particularly the aerodynamic relationships for a rotor in hover or vertical climb, closely parallels the physics of a propeller. In a climb, for example, the air flows through the rotor perpendicular to the rotor disk plane in the same manner that air flows into a propeller disk as it translates horizontally. Thus, the rotor or propeller blades experience the same aerodynamic environment as they rotate. *See* PROPELLER (AIRCRAFT).

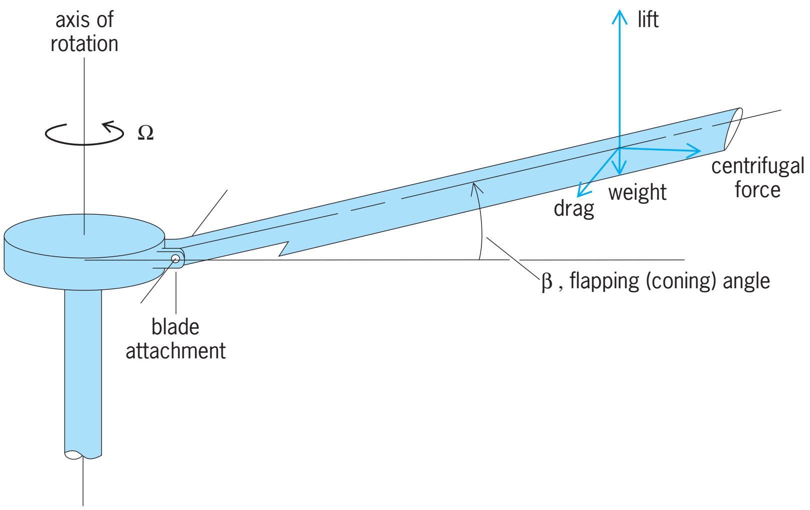


Fig. 1. Forces acting on a helicopter rotor blade.

As the helicopter translates into forward flight, significant differences occur. Because the rotor blades are mounted on a vertical axis and rotate in a horizontal plane, the translational velocity of the helicopter significantly alters the relative airspeed of a rotor blade element as the rotor blade revolves (**Fig. 2**). The blade moving in the direction of flight (advancing blade) encounters a relative airflow velocity which is equal to the vector sum of the helicopter's flight velocity *V* and the blade's rotational velocity. The rotational velocity is proportional to the radial distance along the blade, and at the tip is equal to the tip speed Ω*R*, where Ω is the angular velocity of the rotor and *R* is the radial distance to the blade tip. On the retreating side of the disk, the blade experiences a relative airflow velocity equal to the difference between the two velocities. When the blades are positioned fore and aft, the flight velocity does not contribute substantially to the relative airflow velocity across the blade.

This changing airflow velocity as the blades rotate produces a sinusoidal lift variation if the blade pitch angle is held constant. Since lift is proportional to the velocity squared, an advancing blade with its higher velocity will generate more lift than a retreating blade. Recalling the dynamic response characteristics of the blades discussed above, the blade with

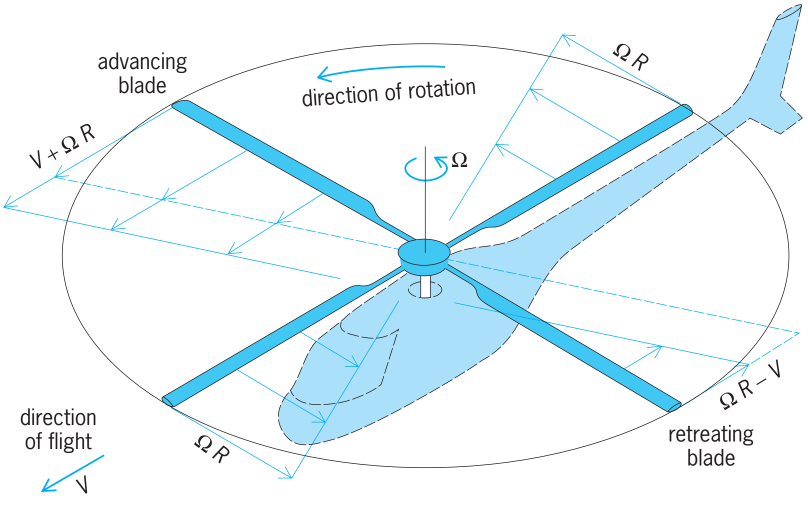


Fig. 2. Variations of relative airflow velocity of rotor blades in forward flight.

greater lift flaps to a larger angle than the blade with less lift, causing the rotor to tilt in a direction opposite to the direction of flight. To counter the difference in lift between the advancing and retreating blades, cyclic pitch control is again introduced to lower the pitch, and consequently the lift of the advancing blades, while increasing the pitch and lift of the retreating blades, thus maintaining the desired tilt of the rotor.

This difference in velocity between advancing and retreating blades is the principal reason that helicopters have a lower speed capability than a fixed-wing aircraft. Since helicopters generally operate with a tip speed ΩR in the range of 600–800 ft/s (180–240 m/s), it does not require much forward speed before the advancing blade is moving relative to the air at a speed approaching that of sound (1117 ft/s or 340 m/s). At these speeds, the airfoil sections of the blade experience compressible flow effects which drastically increase the blade drag and hence the power required. On the retreating blade, the situation is just the opposite. As the helicopter goes faster, the retreating blade experiences a relative airflow that is diminished by the forward velocity. Under these conditions, the blade tip would have no movement relative to the air if the forward velocity were equal to the tip speed of the rotor. Even at forward velocities much less than the tip speed of the rotor, there are regions where the forward velocity equals or exceeds the local velocity due to rotation. The diminishing velocity on the retreating blades is compensated for by increasing the blade pitch and thus the angle of attack of the blades' airfoil. However, an airfoil has a limiting angle that, if exceeded, will result in flow separation (stall). When stall occurs, there is an increase in drag and decrease in lift capability of the airfoil, which again causes an increase in the power required. These conditions make the design of a helicopter a compromise between compressibility effects on the advancing blades and retreating blade stall. *See* AERODYNAMICS; AIRFOIL.

Performance. Helicopter performance is determined by the power required to achieve the desired flight condition. There are four primary functions associated with helicopter flight that require power: the power required to turn the rotor (profile power), the power required to lift the helicopter (induced power), the power required to propel the helicopter in forward flight (parasite power), and the power required to climb or descend (climb power). Each of these elements contributes in a different manner to the total power required as the flight speed varies.

Profile power. The profile power has a moderate value in hover that is a function of the area of the blades being turned, the drag characteristics of the airfoil sections of the blades, and the cube of the effective speed of the rotor. In hovering flight, the effective speed of the rotor is only a function of the tip speed at which the rotor operates. As the helicopter moves into forward flight, the effective speed of the rotor blades increases with a corresponding increase in profile power. For the example illustrated, the drag characteristics of the airfoil section are assumed to be constant. If the rotor encounters stall or compressibility effects, the drag characteristics will increase and, correspondingly, the profile power required will increase more dramatically with airspeed.

Induced power. The induced power has just the opposite trend, as shown by the decrease in induced power as airspeed is increased. Since induced power is associated with the generation of lift, increasing or decreasing the weight of a given helicopter will cause a corresponding change in induced power.

The size of the rotor also plays a major role in determining the induced power. A rotor generates lift by accelerating air downward or, in effect, increasing the momentum of the downward flow. The larger the rotor diameter, the greater the air mass it will act on, and the lower the velocity change required to produce a given lift. Induced power is directly proportional to this induced velocity, which, in turn, is related to the disk loading of the rotor, that is, the ratio of the lift generated to the area of the planar disk described by the blade radius. Helicopters, with their large-diameter rotors, are low disk-loading aircraft, which accounts for their relatively good efficiency in hovering flight compared to other means of achieving powered vertical lift. Even so, the power required is still sizable, which explains the difficulties encountered by the pioneers in providing an engine light enough and powerful enough to achieve vertical flight.

In forward flight, the mass flow of air through the rotor is increased; consequently, the blades do not have to induce as much velocity change to this higher mass flow in order to produce a given lift. This reduction in induced velocity decreases the induced power required as speed is increased.

Parasite power. The power required to propel the helicopter in forward flight (parasite power) is predominantly a function of the drag characteristics of the fuselage and rotor hub and the cube of the flight velocity. Therefore, the parasite power is zero in hovering flight but rises rapidly as airspeed is increased. At high speeds, the parasite power is the predominant cause of power expenditure, illustrating the need for fuselages and rotor hubs that possess very low drag characteristics.

Climb power. The climb power is proportional to the weight of the helicopter and its velocity of climb or descent. When a helicopter is climbing, more power is required; however, when it is descending, the helicopter requires less power than it does in level flight. If the rate of descent is high enough, the helicopter can achieve a condition called autorotation wherein no engine power is required. Autorotation for a helicopter is similar to the glide capability of fixed-wing aircraft. The helicopter, however, has a distinct advantage over an airplane in this regard since a helicopter can autorotate at much lower airspeeds than the glide speed of an airplane. By using the energy derived from the descent, the helicopter can produce the energy needed to generate the necessary lift and to maintain the desired rotational speed of the rotor. As the helicopter nears the ground, the pilot executes a maneuver, called a flare, which arrests any forward velocity, and the aircraft can be landed in a very small landing site.

Rotor configurations. Many different rotor arrangements have been used (**Fig. 3**), and most of the early attempts at vertical flight were made with machines having multiple or coaxial counterrotating rotors. Most modern helicopters employ the single rotor or the tandem rotor configurations.

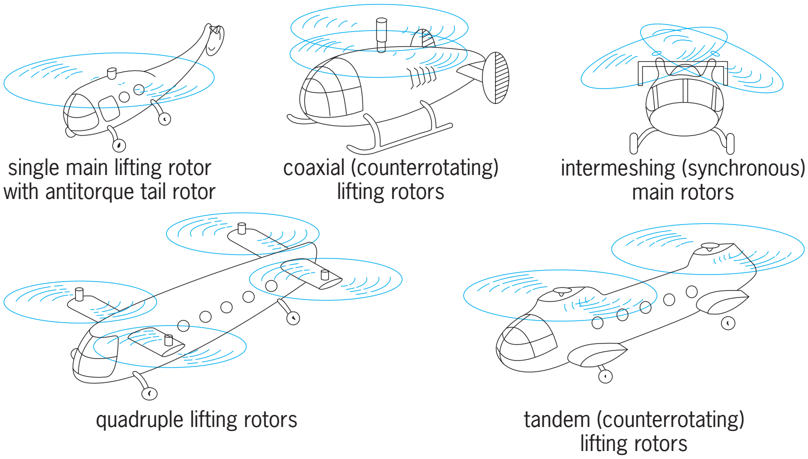

single main lifting rotor with antitorque tail rotor
coaxial (counterrotating) lifting rotors
intermeshing (synchronous) main rotors
quadruple lifting rotors
tandem (counterrotating) lifting rotors

Fig. 3. Principal helicopter (rotor) configurations.

In addition to the selection of the number and location of the lifting rotors, designers have developed varied methods for attaching the blades to the rotor hub. Very early experiments conducted with the blades rigidly attached to the hub were unsatisfactory because of the excessive moments applied to the rotor mast. The first satisfactory solution to this problem was applied by Juan de la Cierva in his development of the autogiro. Cierva incorporated a hinged blade attachment to allow the blades to flap freely and to relieve the undesired moments.

Teetering rotors. Based on the success achieved by the introduction of hinged attachments for the rotor blades, several configurations have been successfully manufactured (**Fig. 4**). The teetering rotor used on two-bladed configurations has one central hinge that allows the blades to move in unison (one up, one down) like a seesaw (Fig. 4*a*). Each blade also has pitch-change bearings to allow the blade pitch angle to be varied as required. The gimbaled rotor (Fig. 4*b*) is essentially equivalent to the teetering rotor and has been used on rotors with three or more blades. It allows the rotor disk to flap as a unit. In both the teetering and gimbaled arrangements, moments are reacted between the blades and are not transferred into the shaft.

Articulated rotor. The articulated rotor (Fig. 4*c*) has each blade attached to the hub by its own flapping hinge. In addition, a hinge is introduced to allow in-plane or lead-lag motion of the blade in order to relieve in-plane bending moments at the root end of the blade. The articulated rotor differs somewhat from the two preceding configurations in that the flapping hinge for each blade is offset from the center of rotation. This offset allows moments to be applied to the shaft and, by selection of a specific hinge offset, the magnitude of the moment is controlled, and is an aid in the control of the helicopter.

Hingeless rotor. The bearingless or hingeless rotor (Fig. 4*d*) is receiving renewed attention in research and development efforts. These efforts represent a closing of the loop in rotor development in that the earliest rotors were basically hingeless designs. There is, however, a major difference in the newer designs which makes them feasible. The early rotors were designed with relatively rigid blade attachments and, given the construction materials of the time, the only way to successfully deal with the large moments was to introduce hinges. With the availability of improved materials, current designs tailor the structural stiffness to allow flexing to take place without having a hinge. This concept provides a lower-weight rotor system and retains the ability to develop hub moments for control.

Flight control. In order to utilize fully the capabilities of a helicopter in hover, vertical, sideward, rearward, and forward flight, the flight controls are of necessity complex (**Fig. 5**).

Cyclic pitch control. The primary control, called cyclic pitch control, is introduced by a control stick located between the pilot's knees. Through a series of control linkages the motion of the stick is transferred to tilt the swash plate either longitudinally or laterally so that blade pitch is cyclically varied in a

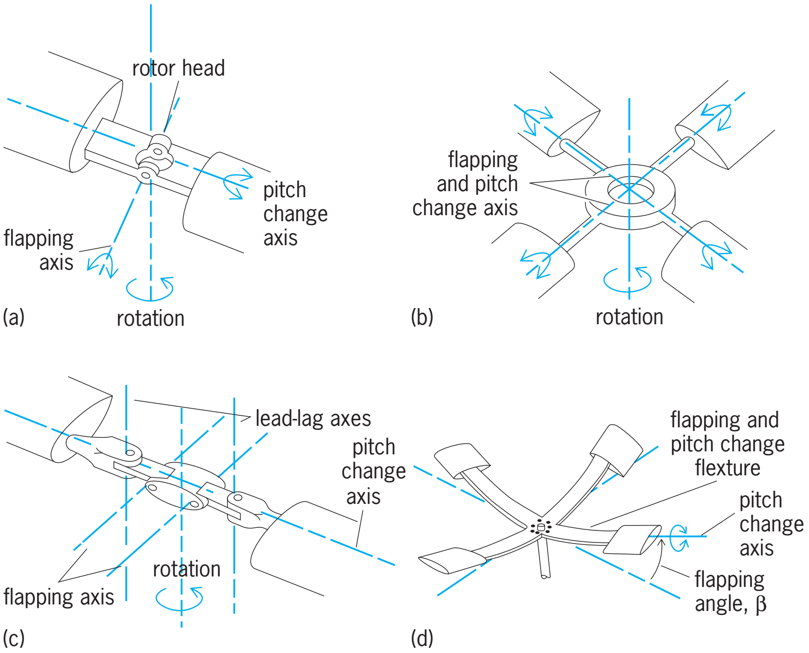


Fig. 4. Principal types of rotor hubs. (*a*) Teetering (semirigid) rotor head. (*b*) Gimbaled rigid rotor. (*c*) Fully articulated rotor head. (*d*) Hingeless rotor.

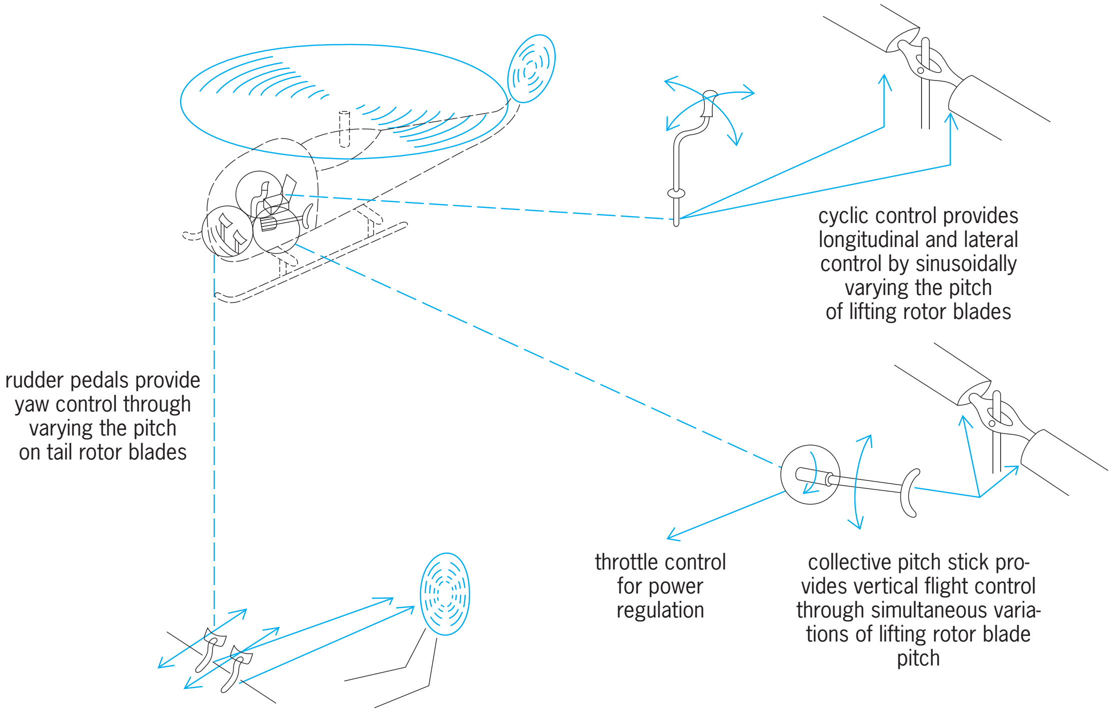

cyclic control provides longitudinal and lateral control by sinusoidally varying the pitch of lifting rotor blades
rudder pedals provide yaw control through varying the pitch on tail rotor blades
throttle control for power regulation
collective pitch stick provides vertical flight control through simultaneous variations of lifting rotor blade pitch

Fig. 5. Functions of helicopter flight controls.

prescribed manner. The cyclic pitch control directs the helicopter in the same direction in which it is moved: if the stick is moved forward, the helicopter moves forward; if it is moved to the right, the helicopter moves to the right, and so forth.

Collective pitch control. The vertical direction of movement is controlled by the collective pitch, which moves the aircraft up if the collective lever is raised and down if the stick is lowered. The collective lever, like the cyclic control, changes the pitch angle of the blades in order to increase or decrease the blade lift. However, a collective input to the blades will change their pitch by an equal amount regardless of their azimuthal position in the plane of rotation. The collective lever is located by the pilot's left hand and, although it may be locked temporarily with a friction lock, most of the time it must be held in the pilot's hand.

Yaw control. The yaw or directional control of the fuselage is provided by pedals similar to those employed by fixed-wing aircraft. On a single-rotor helicopter, the pedals are used to control the pitch of a small tail rotor. The thrust provided by the tail rotor generates the antitorque required to counteract the main-rotor driving torque or to change the directional heading of the helicopter.

Engine fuel control. Because of the varying power requirements of a helicopter, the engine throttle is constantly monitored. Normally, it is controlled by a motorcycle-type twist grip mounted on the collective pitch lever. On early reciprocating engine helicopters, the pilot had to coordinate the throttle position with the power demands. However, on modern turbine-powered helicopters, the engine fuel control is coupled to the engine's power turbine speed by a governor so that rotor rotational speed is maintained automatically. *See* FLIGHT CONTROLS.

Julian L. Jenkins

Higher harmonic control. This is a demonstrated active-control-system concept for helicopters that promises many improvements in such important areas as helicopter vibrations, noise, and even performance. Up to a 90% reduction in helicopter vibration levels has been achieved in flight tests, showing that it is feasible to reduce vibrations to levels comparable to those of a fixed-wing aircraft. Low vibration levels are important for crew and passenger comfort and also to reduce structural fatigue of the airframe, rotor system, drive train, and dynamic components. Efforts in acoustics have been largely directed at reducing or eliminating the annoying blade-slap noise generated by the interaction of a rotor blade with the trailing tip vortex shed by the preceding blade.

Higher harmonic control is an active control concept, in contrast to the conventional passive means of vibration control, such as vibration absorbers, vibration isolators, bifilar or Frahm absorbers, or nodal beam suspension. A passive vibration control device treats the vibratory loads after they have been generated, whereas an active vibration control concept alters or reduces the vibratory excitation at its source. In the case of higher harmonic control, the source of the vibrations is the vibratory aerodynamic forces which act on the rotor blades in flight. The rotor system transmits these unwanted vibrations to the airframe. Higher harmonic control senses these vibrations and applies high-frequency (20–35 Hz) pitch motion to the rotor blades at very small angles (less than 1°) to suppress this aerodynamic excitation. The primary elements of the active vibration suppression system are acceleration transducers that sense the vibratory response of the fuselage; a higher-harmonic blade-pitch actuator system; a flight-worthy microcomputer, which incorporates the algorithm or mathematical model for reducing vibrations; and a signal-conditioning system (electronic control unit), which interfaces between the sensors, the microcomputer, and the higher-harmonic-control actuators. *See* ACTIVE SOUND CONTROL.

E. Roberts Wood

Applications. The growth of the helicopter industry in the United States is founded on the uses made by the armed forces. From the small two-place helicopter built by Sikorsky Aircraft during World War II to the modern air cavalry concept, the military has been a major proponent of the helicopter. The unique capabilities of the helicopter to operate in confined areas and to take off and land in unprepared sites have made it an indispensable tool in the military environment. Large-scale military orders for helicopters during the Korean campaign provided the impetus to the fledgling industry. The growth in the number of helicopters was paralleled by an equally impressive growth in applications. During the Vietnam conflict of the 1960s, the helicopter experienced a greatly expanded role in the military's tactics. From the outset, the Vietnam conflict was dominated by the helicopter. The absence of usable roads, coupled with the need for extreme mobility, required the flexibility of helicopter transportation to deploy troops and equipment rapidly. Procurement and use of helicopters soared in all the services for such varied tasks as observation, troop carriers, gun ships, rescue, and cargo resupply (**Fig. 6**). *See* MILITARY AIRCRAFT.

The technology that evolved to meet the needs of the military provided the base for an impressive growth in commercial applications. Initially, commercial operators were dependent entirely on derivatives of military aircraft for the equipment used. However, as commercial operations expanded into more areas, a market was established which would support the development of distinctly commercial aircraft.

With such diverse operations as crop spraying, logging, construction, police and ambulance service, and passenger and corporate transportation, the industry has responded with a variety of commercial helicopters (**Fig. 7**). One of the more dramatic areas of growth has been in energy exploration, particularly in the offshore oil operations in regions such as the North Sea and the Gulf Coast of the United States. Transportation of workers and equipment to and from the oil rigs is a role easily filled by the helicopter. However, as the oil rigs are moved farther from shore, the longer mission ranges present

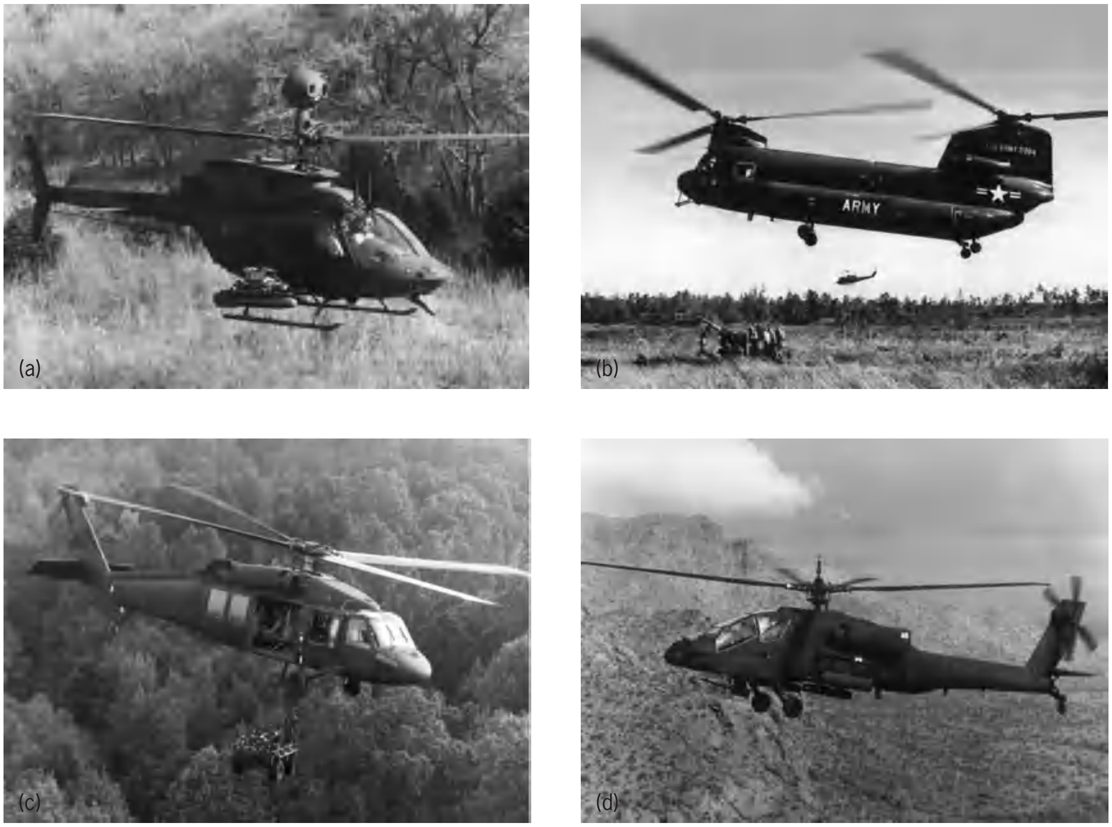


Fig. 6. Military helicopter applications. (*a*) Observation (*Bell Helicopter Textron*). (*b*) Cargo resupply (*U.S. Army*). (c) Tactical transportation (*Sikorsky*). (*d*) Attack (*U.S. Army*).

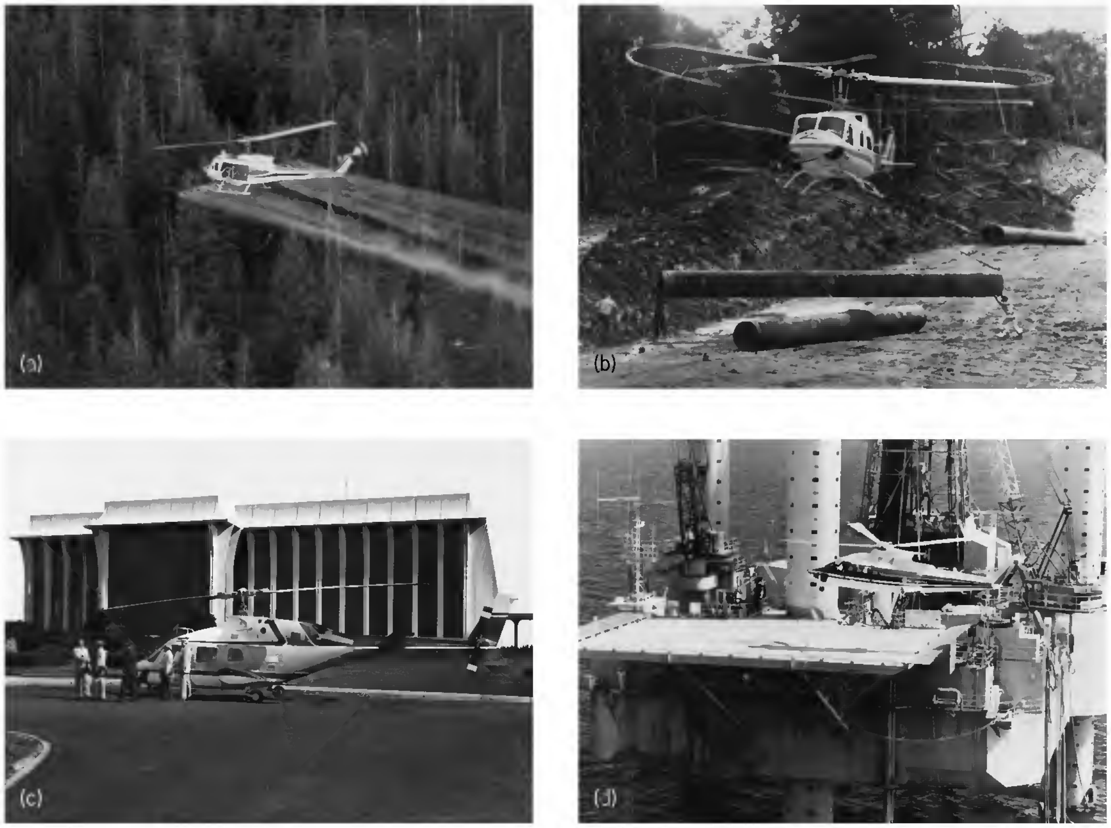


Fig. 7. Commercial helicopter applications: (*a*) Spraying; (*b*) construction; (c) corporate transportation (*Bell Helicopter Textron*). (*d*) Oil exploration (*Sikorsky*).

a significant technical challenge to the industry. *See* OIL AND GAS, OFFSHORE.

Hybrid aircraft. The expanding applications of rotary winged aircraft have created a demand for improved capabilities, including higher speeds, longer range, and greater efficiency. In order to maintain the hovering efficiency of the helicopter while achieving the higher speeds of fixed-wing aircraft, hybrid or convertible aircraft have evolved that maintain the large-diameter rotor efficiency of the helicopter and then convert or change the configuration to achieve high-speed performance. Two configurations that use this approach are the tilt-rotor and the X-wing or stopped-rotor configurations.

Tilt-rotor configuration. The tilt-rotor aircraft (**Fig. 8**) uses large-diameter rotors located on pylons at each end of the wing. For hover and low-speed flight, the tilt-rotor flies in much the same manner as a conventional helicopter. That is, the aircraft is controlled by altering the magnitude and direction of the lift vectors on each of the two side-by-side rotors. For wingborne cruise flight, the pylons are tilted forward so that the rotors function as large propellers. This conversion accomplishes two significant things. First, by rotating the rotor and hub into an axial alignment with the free-stream flow, the parasite drag is considerably reduced, as is the power required for high-speed flight. Second, with the rotor operating in axial flow as a propeller, the sinusoidal flow conditions that exist in the helicopter mode are eliminated, and the rotor can produce propulsive force at much higher speeds.

Full conversion can be accomplished in seconds and it is also possible for the tilt-rotor to fly at intermediate conversion angles over a wide range of airspeeds. Full conversion to the airplane mode can be accomplished only at airspeeds above the stall speed of the wing.

When in the airplane mode, a tilt-rotor is controlled with conventional ailerons, elevators, and rudders in the same manner as any fixed-wing aircraft. In the helicopter mode, the rotors provide the flight control by using collective and cyclic pitch and pedals. Control is accomplished somewhat differently than the single-rotor control described above.

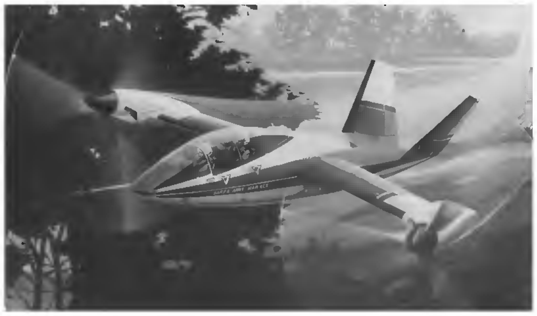


Fig. 8. Concept of advanced tilt-rotor aircraft. (*Bell Helicopter Textron*)

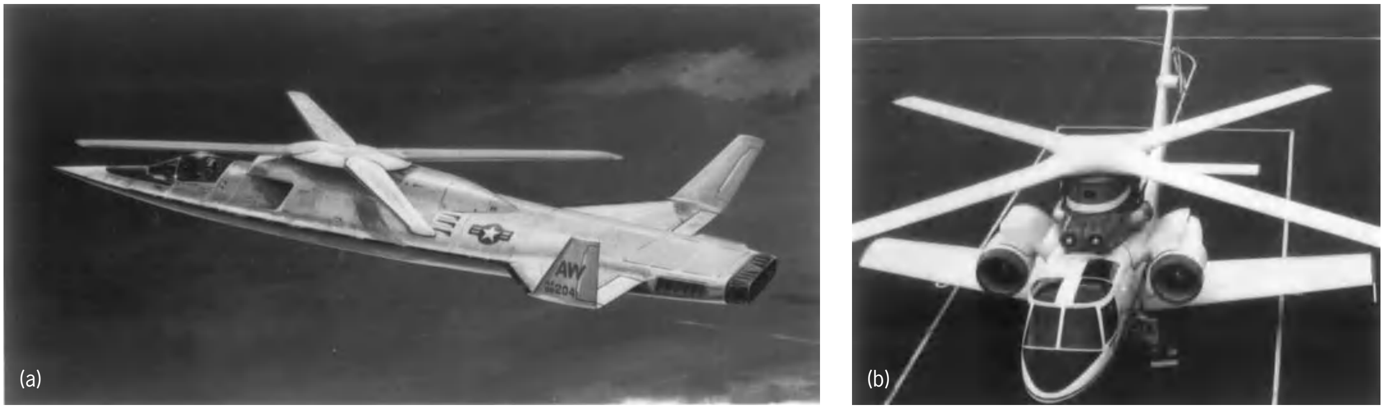


Fig. 9. X-wing aircraft. (*a*) Concept. (*b*) Experimental X-wing, developed by Defense Advanced Research Projects Agency (DARPA), installed on NASA/U.S. Army Rotor Systems Research Aircraft. (*Sikorsky Aircraft Division, United Technologies Corp.*)

Since there are two rotors, roll and yaw control are implemented by introducing differential forces between the rotors. Increasing the collective pitch on one rotor while decreasing it on the other introduces a rolling moment. While differential collective pitch is used for roll control, differential longitudinal cyclic pitch is used for yaw control. Since cyclic pitch tilts the rotor disk in the desired direction of flight, the aircraft executes a yaw turn when one rotor tilts forward and the other back.

The feasibility of the tilt-rotor concept was demonstrated by experimental research aircraft, first flown in 1977. Its success paved the way for the development of the V-22 Osprey tilt-rotor aircraft, flown in March 1989. *See* VERTICAL TAKEOFF AND LANDING (VTOL).

X-wing configuration. In the experimental X-wing or stopped-rotor approach to achieving high-speed flight, rather than conversion of the low-disk-loading rotor into a propeller, the rotor is slowed and stopped in an X position relative to the fuselage (**Fig. 9**). In order to accomplish this conversion, several unique design features must be provided. First, the rotor blades must be extremely rigid since the centrifugal stiffening of a rotating blade is lost when the rotor stops. Second, the ability to generate lift and to control the aircraft while the rotor slows to a stop creates an unusual requirement for both the airfoil section of the rotor/wing and the method of controlling the system. Finally, the power plant must also convert from a system that provides shaft power to the rotor for helicopter flight into a system that provides propulsive force when the rotor is stopped.

To achieve the desired rigidity of the rotor/wing, the X-wing aircraft uses an extremely rigid cantilevered rotor blade. In order to achieve the desired stiffness with a reasonable weight, graphite/epoxy composite materials are used to fabricate an I-beam type of structural spar. The beam is tailored to provide high bending stiffness in order to carry the lifting loads when the rotor is stopped. It is also tailored in torsional bending to allow the beam to be moderately twisted for changes in collective pitch. *See* COMPOSITE MATERIAL.

Unlike a conventional helicopter rotor, however, the X-wing does not require physical movement of the rotor blade as the primary method of achieving collective and cyclic pitch control. Instead, it uses a circulation control airfoil, which relies on the Coanda principle to alter the sectional lift generated. When a thin, low-pressure, high-speed stream of air is blown tangentially to a curved surface, the flow stream tends to follow the surface and thereby alter the velocity field of the airfoil. This change in the velocity field alters the lifting pressure on the airfoil. By modulating the internal air supplied to the slotted airfoil, the lift can be increased or decreased just as if the airfoil were changing its angle of attack. A pneumodynamic control system forces compressed air through ducts in the leading and trailing edges of the rotor/wing blades to provide circulation control lift in rotary and fixed-wing flight and during conversion between the two modes.

The leading edge of a rotor blade in helicopter flight becomes the trailing edge on the left wing when the rotor is stopped for the cruise condition. This unique conversion requires an elliptical airfoil with provisions for blowing air over both the leading and trailing edges of the airfoils during certain azimuth positions in order to maintain the desired aircraft lift and rolling moment trim as the conversion takes place. A convertible engine and declutching arrangement is required to provide an alternative propulsive system once the rotor system is stopped. The convertible engine has two operating modes. In one mode it provides shaft power through a power turbine to turn the main rotor blades when the X-wing operates as a helicopter. In the second mode, the power turbine output shaft is declutched from the rotor and the engine then operates as a turbojet to provide thrust for forward propulsion.

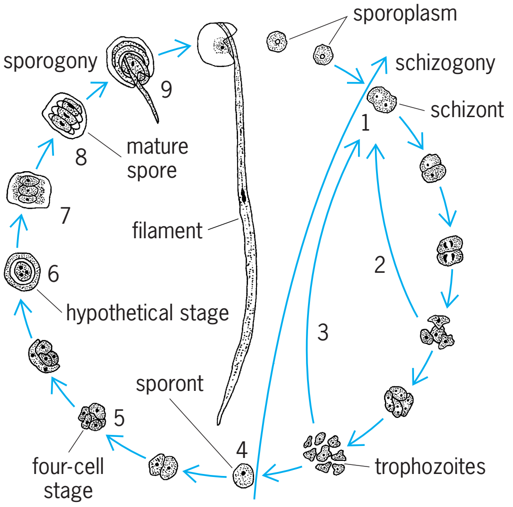
Julian L. Jenkins

Bibliography. J.-P. Harrison, *Principles of Helicopter Flight*, 1993; W. Johnson, *Helicopter Theory,* 1980, reprint 1994; R. W. Prouty, *Helicopter*

Performance, Stability, and Control, 1986, revised 1995; J. Seddon, *Basic Helicopter Aerodynamics*, 1990; W. Z. Stepniewski and C. N. Keys, *Rotary-Wing Aerodynamics*, 1984.

Helicosporida

An order of protozoa in the class Myxosporidea (subphylum Cnidospora). It is characterized by production of spores with a relatively thick, single intrasporal filament and three uninucleate sporoplasms. The spore membrane occurs as a single piece. *Helicosporidium parasiticum* is the only species in the order. The parasite infests the body cavity, fat bodies, and ganglia of mites (*Hericia hericia*) and the body cavity of fly larvae (*Dasyhelea obscura* and *Mycetobia pallipes*) found in the sap of horse chestnut and elm trees. *See* INSECT PATHOLOGY.



Probable development of a helicosporidan, diagrammatic: 1–3, schizogony, with formation of the schizont; 4–9, sporogony, with development of mature spore; 4–7, sporont arising from trophozoites and development into immature spore; 7, mature spore; 8, opening of mature spore; 9, escape of sporoplasms

Although the complete life cycle of helicosporidans is not known, stages in the development and division of the trophozoite and in the development of the spore have been described (see **illus.**). The sporont or sporoblast gives rise to four cells: One forms the spore membrane and three remain as sporoplasms. The origin of the filament is not known. *See* CNIDOSPORA; DIPTERA; MYXOSPORIDEA; PROTOZOA.

Ross F. Nigrelli

Helimagnetism

A property possessed by some metals, alloys, and salts of transition elements in which the atomic magnetic moments, at sufficiently low temperatures, are arranged in a spiral or helix. It may be seen from the **illustration** that simple antiferromagnets and ferromagnets can be considered as nonconical helimagnets with helical angles ϕ of 180 and 0°, respectively. In the same way, nonconical helimagnets may be considered as conical helimagnets with cone angle θ of 0°. Some typical helimagnets are listed in the **table**. The magnetic structures have been detected by neutron diffraction.

Helimagnetism arises when the exchange coupling parameters J_{ij} between the spins $\mathbf{S}_i$ and $\mathbf{S}_j$ of the magnetic atoms lie within a certain range of values. The exchange energy is given by Eq. (1), summed

$$E_{ij} = -2J_{ij}\mathbf{S}_i \cdot \mathbf{S}_j \qquad (1)$$

over all pairs of atoms (i, j). *See* FERROMAGNETISM.

Consider an axial structure with strong ferromagnetic coupling (positive J_{ij}) between atoms which are in the same plane, and weaker ferromagnetic or antiferromagnetic coupling between atoms which are in the same plane, and weaker ferromagnetic or antiferromagnetic coupling between atoms which are in neighbor planes. Let this latter coupling between nearest neighbor planes be J_1, between next neighbor planes be J_2, and so on. Let the angle between the ferromagnetically coupled spins of the nth plane and those of the first plane be ϕ_n. The exchange energy is then given by expression (2), where J is the intraplanar exchange.

$$E \propto -\Sigma_n[J + 2J_1 \cos\left(\phi_{n+1} - \phi_n\right) + 2J_2 \cos\left(\phi_{n+2} - \phi_n\right) + \cdots] \qquad (2)$$

If one assumes a helical array, then Eq. (3) holds,

$$\phi_n = nka + \text{constant} \qquad (3)$$

where ka, which is the ϕ of array (*c*) of the figure, is the phase angle between neighbor planes. With a equal to the spacing between planes, k becomes the wave number describing the pitch of the helix.

On substituting Eq. (3) into Eq. (2), one obtains Eq. (4) where J_k is the Fourier transform of the ex-

$$E \propto -[J + 2J_1 \cos ka + 2J_2 \cos 2ka + \cdots] = -J_k \qquad (4)$$

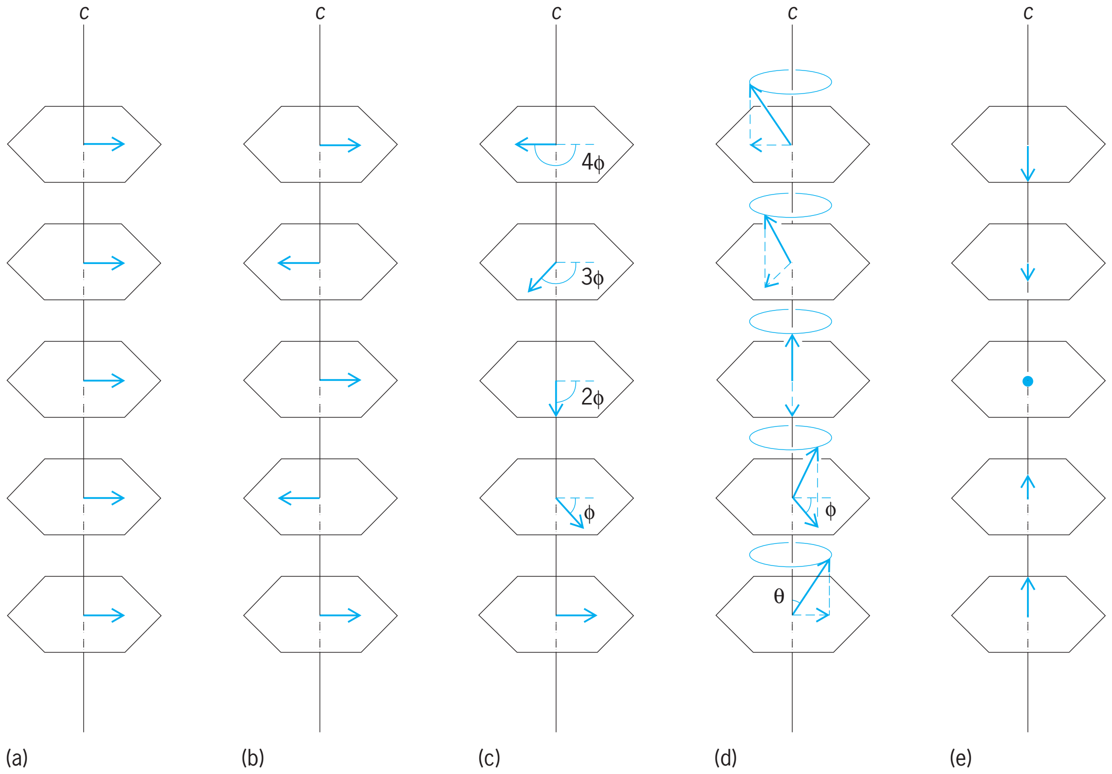
change parameter J_{ij}, when the latter is considered a function of distance between planes only.

It is seen that the exchange energy is minimized by that value of k which maximizes J_k. In this way the pitch of the helix is determined.

Ferromagnets have maximum J_k at $k = 0$. This cannot be achieved if the interplanar J_n are mostly negative, that is, antiferromagnetic. In the latter case what happens physically is that the planes try to order antiferromagnetically with respect to one another. It is impossible for them all to do this, and a helical compromise results.

Conical helimagnets have maximum J_k at $k \neq 0$, and also anisotropic coupling which favors spin alignment along the crystal axis.

Helimagnetism is found in most of the rare-earth metals and their alloys and in a few salts of the iron group. Some helimagnets have quite complicated patterns, such as manganese sulfate which, below



Some magnetically ordered arrays. Shown are directions taken by magnetic moments of consecutive planes along the *c* axes of hexagonal structures. (*a*) Simple ferromagnet; (*b*) simple antiferromagnet; (*c*) nonconical helimagnet of helical angle ϕ; (*d*) ferromagnetic conical helix of cone angle θ and helical angle ϕ; (*e*) sinusoidal antiferromagnet. In *a*, *b*, and *c* the moments all lie in the planes. In *d* there is a net ferromagnetic component along the *c* axis and in *e* there is an oscillating moment along the *c* axis.

Some representative helimagnets

Substance	Magnetic structure	Temperature, K (°F)
MnO	Nonconical helix	$0 < T < 84$ $(-460 < t_F < -308)$
$MnAu_2$	Nonconical helix	$0 < T < 363$ $(-460 < t_F < 194)$
Dy	Nonconical helix	$85 < T < 179$ $(-307 < t_F < -137)$
	Ferromagnet	$0 < T < 85$ $(-460 < t_F < -307)$
$MnCr_2O_4$	Simple ferrimagnet	$18 < T < 43$ $(-427 < t_F < -382)$
	Complex conical helix	$0 < T < 18$ $(-460 < t_F < -427)$
Er	Conical helix	$0 < T < 20$ $(-460 < t_F < -424)$
	Complex oscillation	$20 < T < 53$ $(-424 < t_F < -364)$
	Sinusoidal antiferromagnet	$53 < T < 85$ $(-364 < t_F < -307)$

10 K (−442°F), orders into two antiferromagnetically coupled conical helices. Frederic Keffer

Bibliography. B. Barbara, D. Gignoux, and C. Vettier, *Lectures on Modern Magnetism*, 1988; D. Craik, *Magnetism*, 1995; D. C. Jiles, *Introduction to Magnetism and Magnetic Materials*, 2d ed., 1998; C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996.

Helioporacea

An order of the cnidarian subclass Alcyonaria (Octocorallia). The Coenothecalia have no spicules but form colonies with a massive skeleton composed of fibrocrystalline argonite fused into lamellae. The skeleton is perforated by both numerous wide cylindrical cavities occupied by the polyps, and narrow ones containing the solenial systems. In the calyx of the polyp, septalike structures of stony coral, or pseudosepta, are formed. The order includes a few genera, of which *Heliopora*, or the blue coral, is often found on coral reefs. *See* OCTOCORALLIA (ALCYONARIA); REEF. Kenji Atoda

Helioseismology

A technique for probing the interior of the Sun, using methods akin to terrestrial seismology. The Sun, although the nearest star by far, is a typical star, so what can be learned of its interior through helioseismology is of broad importance to the stars in general.

Solar waves. Like terrestrial seismology, helioseismology entails the analysis of many "seismic" wave modes to determine the structure of the interior. However, although terrestrial seismic waves are initiated by a singular event such as an earthquake, waves within the Sun are continuously excited, probably by the turbulent convective motions in its outer layers. Thus the solar waves are always present at all points within the Sun and on its surface. The Sun is "ringing" like a bell, but not like one struck by a clapper; it vibrates more like a bell suspended in a sandstorm, continuously struck by tiny grains of sand. *See* SEISMOLOGY.

The solar waves are seen at the surface as up-and-down motions of the gases with a speed of about 0.3 mi/s (0.5 km/s) and a vertical displacement of about 30 mi (50 km). These waves are detected through the Doppler shift of the wavelength of absorption lines in the solar spectrum. They have periods clustering near 5 min (that is, with a frequency of one cycle in 5 min or about 0.003 cycle per second). As a result, the solar surface undulates up and down in a so-called five-minute oscillation (**Fig. 1**). The oscillation is actually the superposition of as many as 10^7 individual modes of oscillation of the Sun as a whole, where each mode has its own characteristic frequency (near, but not exactly at, 0.003 cycle per second) and spatial pattern on the solar surface.

Precise observations of the solar oscillations are difficult. The individual oscillation modes have velocities at the solar surface of only 12 in./s (30 cm/s) or less, and only extremely sensitive and stable spectrographs can measure such small motions. In addition, a nearly continuous stream of data extending over days is needed to separate the many individual modes with nearly identical oscillation frequencies. Ground-based observations are hampered because the day-night cycle allows observations of only 10 h or so at most sites. This restriction has been overcome by making observations from near the South Pole during the austral summer, and through networks of similar telescopes spaced at several longitudes around the globe (**Fig. 2**). Another very successful method is to obtain observations from a spacecraft located in an orbit experiencing continuous sunlight, such as the *Solar and Heliospheric Observatory* (*SOHO*) spacecraft, which orbits about the stable L1 Lagrange point, lying between the Earth and the Sun about 9 × 10^5 mi (1.5 × 10^6 km) from the Earth.

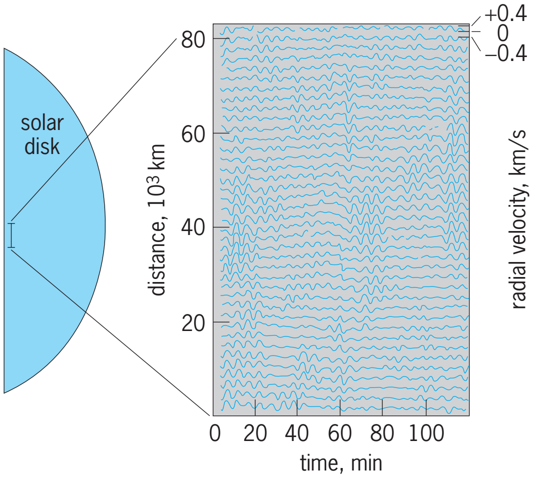


Fig. 1. Variation of radial velocity with time along an 80,000-km (50,000-mi) line on the solar surface. The 5-min oscillations, visible as wave packets localized in space and time, are actually the superposition of millions of discrete oscillation modes. 1 km = 0.6 mi. (*Adapted from S. Musman and D. Rust, Vertical velocities and horizontal wave propagation in the solar photosphere, Solar Phys., 13:261–286, 1970*)

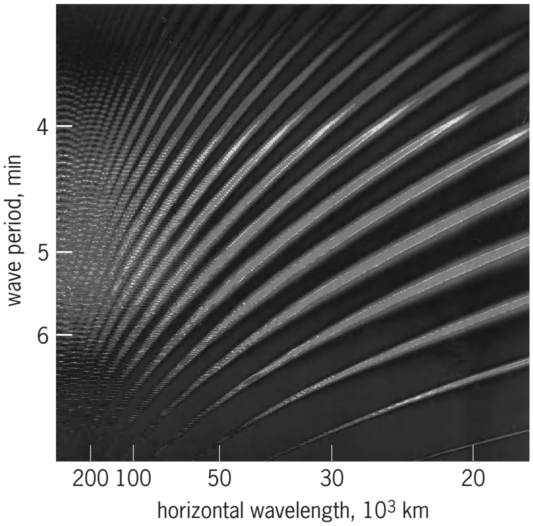


Fig. 2. Space-time spectrum of solar oscillations, from continuous observations obtained by the Global Oscillations Network Group (GONG) network of six solar telescopes spaced in longitude around the Earth. Ridges of power show how oscillation periods vary with horizontal wavelength; from lower right to upper left, each ridge is a higher overtone. 1 km = 0.6 mi. (*National Optical Astronomy Observatories*)

Each solar oscillation mode is produced within a specified region of the interior. The top of this resonant cavity is near the solar surface, which reflects the waves downward when they strike the surface from below. The bottom occurs at the depth where upward refraction reverses the propagation direction of descending waves. Waves with long horizontal wavelength, which travel nearly straight downward into the Sun, penetrate deeply, whereas waves with short horizontal wavelength remain near the surface. Within each cavity there can be both a lowest mode of oscillation and many higher-frequency overtones, or harmonics. Thus, a two-dimensional spectrum, displaying the oscillation power as a function of frequency and horizontal wavelength, shows a number of "ridges" of power, each corresponding to a separate overtone number (Fig. 2). *See* HARMONIC (PERIODIC PHENOMENA); MODE OF VIBRATION; VIBRATION.

Structure of solar interior. It is possible to predict the frequencies of oscillation modes in the Sun, given a model of the interior, that is, a numerical tabulation of temperature, density, pressure, and composition as a function of distance from the surface. When the oscillation frequencies predicted by the standard model in general use before the advent of helioseismology were compared with the observed frequencies, small but significant discrepancies were found.

The discrepancies were lessened when models allowed for a higher initial abundance of helium in the core than first thought, as well as a decrease over the Sun's lifetime in the helium abundance near the surface due to gravitational sinking of helium relative to the lighter hydrogen atoms. The models pinpointed the bottom of the convection zone, at a depth of about 125,000 mi (200,000 km), or about 30% of the solar radius below the surface. They also suggested that the opacity of the solar material just below the bottom of the convection zone was larger than predicted; this was subsequently verified by revised opacity calculations. New calculations of the equation of state relating pressure to temperature and density throughout the interior led to still better agreement between models and helioseismology data. The most significant remaining difficulty is that the models predict a flux of solar neutrinos from the core which is at least twice that observed. It is impossible to adjust the structure so as to yield the observed neutrino flux without producing serious disagreement with helioseismology data, using standard physical descriptions of neutrinos. A possible resolution of this problem is that models of neutrinos as massless particles may be incorrect, and that instead they have a slight amount of mass, thereby causing neutrinos produced in the core to be changed into nondetectable form before reaching detectors at the Earth. *See* NEUTRINO; SOLAR NEUTRINOS.

Rotation in solar interior. Helioseismology offers insight into the structure of the solar interior and also into its rotation. Waves propagating with or against the direction of rotation are carried by it, and their effective propagation speed and frequency are increased or decreased. The frequency shift for any mode depends on the average rotation rate within the resonant cavity for that mode, and comparison of the shift for many modes with different cavities makes it possible to determine how the rotation varies with depth.

The surface of the Sun has long been known to rotate differentially with latitude; that is, at the Equator the surface rotation period is about 25 days while near the Poles it is about 34 days. Helioseismology provides a description of how the rotation varies with depth, down to and below the bottom of the convection zone which constitutes the outer 30% of the solar radius (**Fig. 3**). Roughly speaking, the increase of rotation period from Equator to Pole persists throughout the convection zone. However, at all latitudes the rotation period decreases slightly over the outer 10% of the solar radius, and then increases again to approximately its surface value at the bottom of the convection zone. At the bottom of the convection zone there is an abrupt transition to a deeper interior, which seems to rotate nearly uniformly and at the same speed as surface latitudes of about 35°. That is, the rotation period increases rather sharply across this transition zone near the Equator, and decreases sharply near the Poles. This zone of radial shear, called the tachocline, may be the location of the solar magnetic dynamo, which creates the 22-year cycle of sunspot activity. *See* STELLAR ROTATION; SUN.

Robert W. Noyes

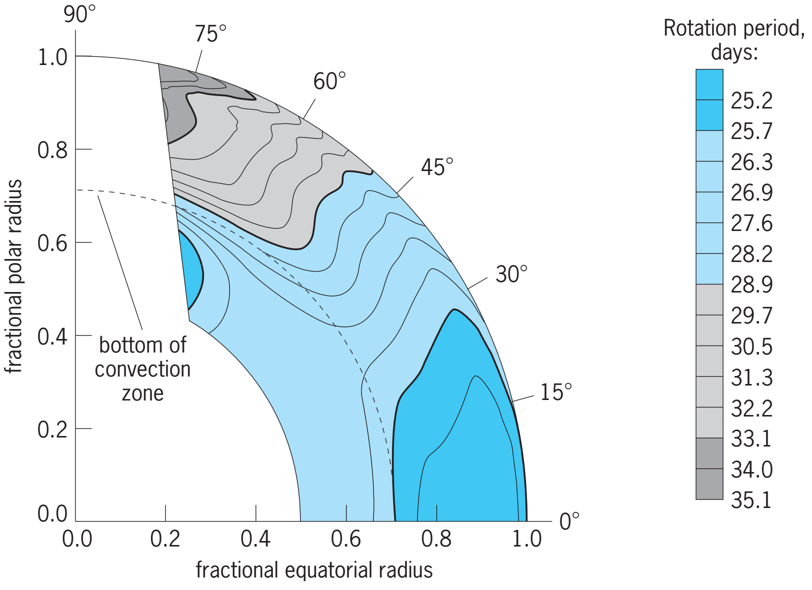
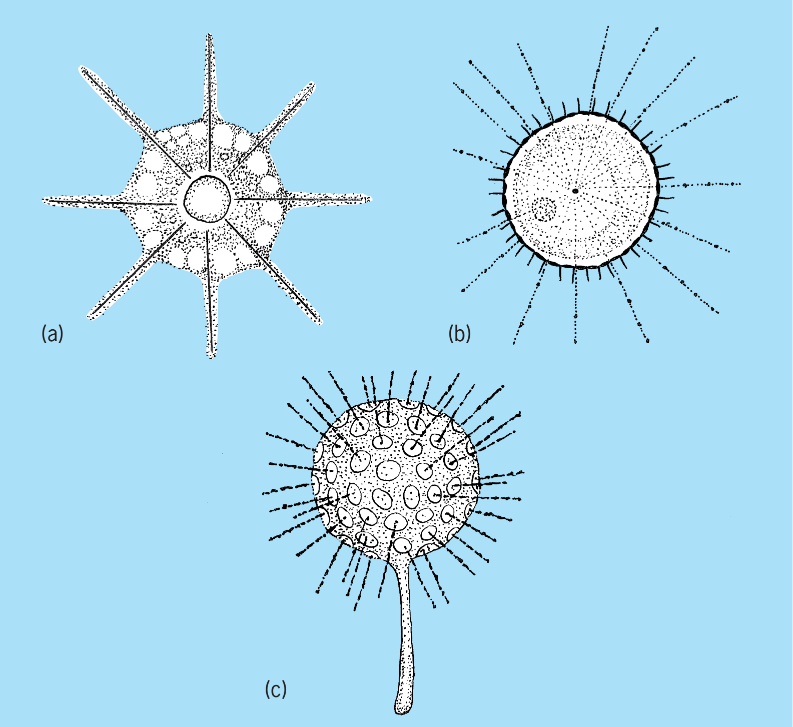


Fig. 3. Contour plot showing lines of constant rotation period within the outer 50% of the Sun, based on data from the Solar Oscillations Investigation (SOI) on the *Solar and Heliospheric Observatory (SOHO)* spacecraft. Tints refer to periods in days given by the key. Light color below the convection zone (broken line) indicates almost uniform rotation of the interior. Solar latitude is indicated from the Equator (0°, horizontal axis) to the Pole (90°, vertical axis). White areas represent regions inaccessible to current helioseismology measurements. (*S. Korzennik, SOI team; NASA*)

Bibliography. J. W. Leibacher et al., Helioseismology, *Sci. Amer.*, 253(3):48–57, September 1985; Special section on helioseismology, *Science*, 272:1281–1309, 1996.

Heliozoia

A subclass of the Actinopodea. Unlike Radiolaria, these protozoans have no central capsule. Most species live in fresh water. Pseudopodia may be either slender with an axial filament surrounded by cytoplasm (axopodia) or filamentous (filopodia). Axopodial filaments can be extended or retracted rapidly by mechanisms not yet explained. Certain floating species can roll along on the tips of their axopodia and also swim by moving their axopodia. Some species are naked; others have skeletal elements ranging from siliceous scales or spicules embedded in a gelatinous capsule to a reticulate chitinous skeleton often impregnated with silica. A centroplast may or may not be present. The subclass has three orders. Actinophryida lack skeletons and centroplasts and include both uninucleate and multinucleate genera (**illus.** *a*). In the Centrohelida a centroplast is present or else is assumed to be because the nucleus is eccentric (illus. *b*). The organisms may be covered with a gelatinous sheath or with a skeleton composed of discrete siliceous elements embedded in such a layer. The Desmothoracida have a perforate continuous skeleton which is composed of chitin, and sometimes impregnated with minerals



Organization of Heliozoa. (*a*) Actinophryida type. (*b*) Centrohelida type. (*c*) Desmothoracida type. (*After R. P. Hall*, *Protozoology*, *Prentice-Hall*, *1953*)

(illus. *c*). *See* ACTINOPHRYIDA; ACTINOPODEA; CENTROHELIDA; DESMOTHORACIDA; PROTOZOA; SARCODINA; SARCOMASTIGOPHORA. Richard P. Hall

Helium

A gaseous chemical element, He, atomic number 2 and atomic weight 4.0026. Helium is one of the noble gases in group 18 of the periodic table. It is the second lightest element. The world's chief source of helium is a group of natural gas fields in the United States. *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**

Helium is a colorless, odorless, and tasteless gas. It has the lowest solubility in water of any known gas. It is the least reactive element and forms essentially no chemical compounds. The density and the viscosity of helium vapor is very low. Thermal conductivity and heat content are exceptionally high. Helium can be liquefied, but its condensation temperature is the lowest of any known substance. The properties of helium are given in the **table**. *See* LIQUID HELIUM.

Helium was first used as a lifting gas in balloons and dirigibles. This use continues for high-altitude research and for weather balloons. The principal use of helium is in inert gas–shielded arc welding. The greatest potential for helium use continues to emerge from extreme-low-temperature applications. Helium is the only refrigerant capable of reaching temperatures below 14 K (−434°F). The chief value of ultralow temperature is the development of the state of superconductivity, in which there is virtually zero resistance to the flow of electricity. Other helium applications include use as a pressurizing gas in liquid-fueled rockets, in helium-oxygen breathing mixtures for divers, as a working fluid in gas-cooled nuclear reactors, and as a carrier gas for chemical analysis by gas chromatography.

Terrestrial helium is believed to be formed in natural radioactive decay of heavy elements. Most of this helium migrates to the surface and enters the atmosphere. The atmospheric concentration of helium (5.25 parts per million at sea level) could be expected to be higher. However, its low molecular weight permits helium to escape into space from the upper atmosphere at a rate roughly equal to its formation. Natural gases contain helium at concentrations higher than in the atmosphere. Arthur W. Francis

Helium is an element with a closed electronic shell, a large ionization potential, and a low polarizability, which makes it a very unlikely candidate to form chemical bonds. However, solid helium compounds have been found to form at high pressure, one with nitrogen [$He(N_2)_{11}$] and one with neon [$Ne(He)_2$]. These compounds belong to a class known as van der Waals compounds. *See* INTERMOLECULAR FORCES.

Other helium compounds have also been observed in a clathrate hydrate, $He(H_2O)_{6+\delta}$, and helium has been detected inside the carbon molecule buckminsterfullerene (C_{60}), forming HeC_{60}. Mixtures of helium and other components prevail under conditions of high pressure in the outer planets of the solar system and their satellites. Therefore, it is believed that helium compounds are important in the modeling of the interiors of such celestial bodies. The formation of helium compounds at high pressures illustrates that under such conditions different chemical behavior occurs compared to that observed under ambient conditions. *See* CHEMICAL BONDING; CLATHRATE COMPOUNDS; FULLERENE. Willem L. Vos

Helium-3 is a rare stable isotope of helium was discovered by L. W. Alvarez and R. Cornog in 1939. Its concentration in nature is so low, approximately one part per hundred million in well helium, that it was 1951 before sufficient quantities of pure gas became available for experimentation. The gas was then, and continues to be, obtained as a by-product from the decay of tritium, the heavy isotope of hydrogen. Tritium is produced in a nuclear reactor from the reaction between lithium and a neutron.

Properties of helium

Property	Value
Atomic number	2
Atomic weight	4.0026
Melting point* at 25.2 atm pressure	−272.1°C (1.1 K)
Triple point (solid, helium I, helium II)	−271.37°C (1.78 K)
Triple point = λ-point (helium gas, helium I, helium II)	−270.96°C (2.19 K)
Boiling point at 1 atm pressure	−268.94°C (4.22 K)
Gas density at 0°C and 1 atm pressure, g/liter	0.17847
Liquid density at its boiling point, g/ml	0.1249
Solubility in water at 20°C, ml helium (STP)/1000 g water at 1 atm partial pressure of helium	8.61

*The melting point varies with the pressure.

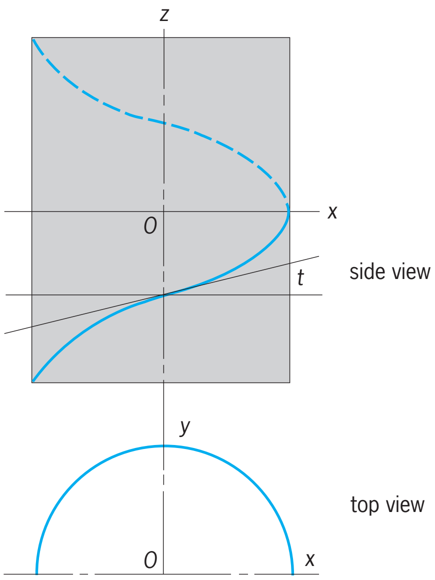
The ^{3}He nucleus is composed of two protons and one neutron, one fewer than for ^{4}He; as a consequence, ^{3}He is a fermion whereas ^{4}He is a boson. The two isotopes are the exemplars of Fermi-Dirac and Bose-Einstein systems, respectively. It is principally for this reason that helium, an apparently featureless chemical element, has been studied intensively. *See* BOSE-EINSTEIN STATISTICS; FERMI-DIRAC STATISTICS; TRITIUM. Bernard M. Abraham

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., 2004; M. Ozima and F. A. Podosek, *Noble Gas Geochemistry*, 2001; J. Wilks and D. S. Betts, *An Introduction to Liquid Helium*, 2d ed., 1987.

Helix

Any nonplanar curve all of whose tangents make the same angle with a fixed line. Other characteristic properties are that all principal normals are parallel to a plane and that the ratio of torsion to curvature is constant. If a helix has constant curvature (and hence constant torsion), it is a circular helix; it lies on a circular cylinder whose elements it cuts at a constant angle (see **illus.**). Parametric equations for a circular helix (the curve of an untapered screw) are $x = a \cos t$, $y = a \sin t$, $z = kt$, $a > 0$, $k \neq 0$, and $-\infty < t < \infty$. *See* ANALYTIC GEOMETRY; DIFFERENTIAL GEOMETRY; PARAMETRIC EQUATION.

Leonard M. Blumenthal



Diagrams of a circular helix.

Helmholtz coils

A pair of flat circular coils of small cross section, with equal numbers of turns and equal diameters, arranged with a common axis, and connected in series to have a common current (**Fig. 1**). The purpose of the arrangement is to obtain a magnetic field H_{AB} that is more nearly uniform than that of a single coil (**Fig. 2**) without the use of a long solenoid. *See* SOLENOID (ELECTRICITY).

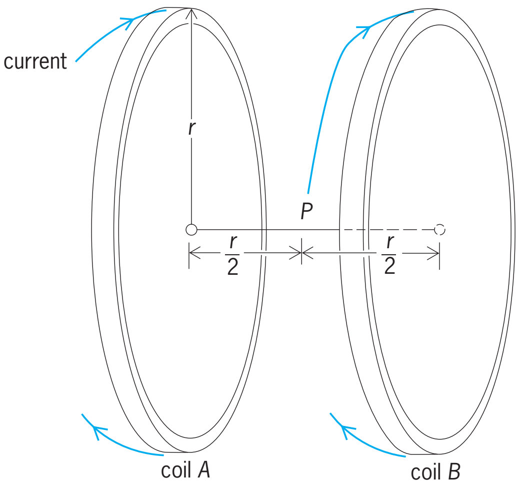


Fig. 1. Helmholtz coils, arranged with common axis and connected in series. (*After L. B. Loeb, Fundamentals of Electricity and Magnetism, 3d ed., Wiley, 1947*)

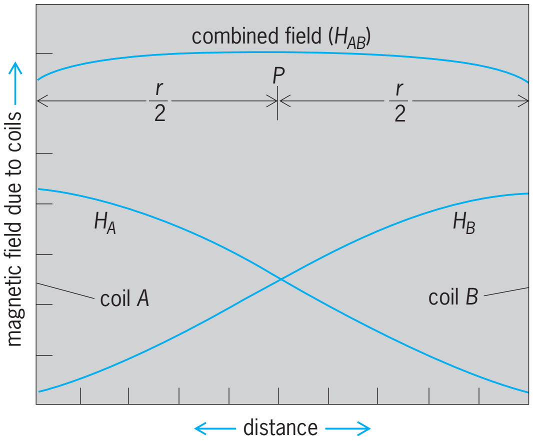
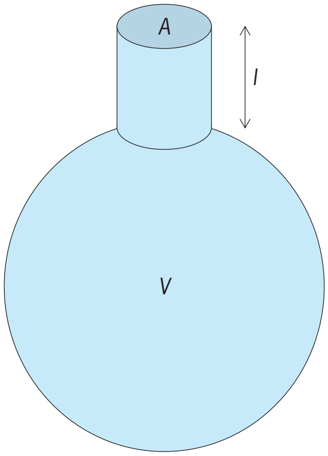


Fig. 2. Separate and combined fields of Helmholtz coils. Note the region of constant resultant field at point *P*. (*After L. B. Loeb, Fundamentals of Electricity and Magnetism, 3d ed., Wiley, 1947*)

The optimum arrangement is that in which the distance between the two coils is equal to the radius of one of the coils (Fig. 1). For this arrangement, the variation of the field strength near the center of the apparatus is a minimum, and the field is nearly uniform near the center. The field is the sum of the fields produced there by the individual coils. More complicated systems having more coils of finite cross section can be computed to give equal or better field uniformity and greater efficiency. Kenneth V. Manning

Bibliography. F. J. Bueche, *Introduction to Physics for Scientists and Engineers*, 4th ed., 1986; D. Halliday, R. Resnick, and K. Krane, *Physics*, 5th ed., 2002; W. H. Hayt and J. A. Buck, *Engineering Electromagnetics*, 7th ed., 2005; W. T. Scott, *The Physics of Electricity and Magnetism*, 2d ed., 1966, reprint 1977.



Helmholtz resonator. Symbols are explained in text.

Helmholtz resonator

An acoustic device consisting of a rigid cavity with one or more small openings. When exposed to sound or vibration, a Helmholtz resonator responds most strongly near a characteristic resonance frequency governed by the dimensions of its cavity and openings. A common example is a bottle or jug that produces sound near its resonance frequency when air is blown over the neck's opening. The Helmholtz resonator is named after the physicist and physiologist Hermann von Helmholtz, who constructed spherical glass resonators with small openings at opposite ends to analyze the frequency content of sounds, including musical tones.

Characteristics. The classical description of the Helmholtz resonator, as derived by Lord Rayleigh, assumes that all dimensions of the resonator are small compared to one-quarter wavelength of its resonance frequency, and that all sound amplitudes are small enough to allow linear acoustic analysis. In this case, the resonator's state can be well described by a uniform cavity pressure and a volume flow rate of air through the cavity openings.

At resonance, the system's oscillating potential energy, associated with compression and expansion of air in the cavity, is exactly balanced by the oscillating kinetic energy of air flowing through the openings. The resulting oscillations can be large compared to the externally applied sound or vibration, and can continue for some time after the external force is removed. This phenomenon occurs at the resonance frequency, also called the natural frequency, which is given for a Helmholtz resonator with a single opening (see **illustration**) by Eq. (1), where c is the speed

$$f_0 = \frac{c}{2\pi}\sqrt{\frac{A}{l'V}} \qquad (1)$$

of sound for air (or other fluid employed), V is the cavity volume, A is the surface area of the opening, and $l' = l + \delta l$ is an effective neck length that includes both the true neck length l and an "end correction" describing an additional volume of fluid constrained to flow into or out of the openings. For an opening small compared to the resonator size, the end correction is approximated by $\delta l = 0.96\sqrt{A}$ for an outer opening within a large plane, and $\delta l = 0.48\sqrt{A}$ for an opening at the end of a long neck in free space.

Energy loss in Helmholtz resonators is associated both with friction of the fluid flowing through the resonator opening and with sound radiation into the surrounding air. These energy loss mechanisms affect the quality factor Q of the resonator, which is defined by Eq. (2), where f_1 and f_2 are the fre-

$$Q = \frac{f_0}{|f_1 - f_2|} \qquad (2)$$

quencies where the resonator response (as measured by its radiation or absorption of sound pressure) is one-half that at the resonance frequency f_0. Low damping results in a high Q value, corresponding to high-amplitude resonator response within a narrow frequency range around resonance. *See* Q (ELECTRICITY).

The behavior of a Helmholtz resonator is analogous to a lumped-element harmonic oscillator with a mass element $M = \rho Al'$, a spring element $K = \rho c^2 A^2/V$, and a damping or dashpot element R, where ρ is the mass density of the air or other fluid. The vibration response of this analog system is greatest near its resonance frequency, given by Eq. (3). Similarly,

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{K}{M}} \qquad (3)$$

the Helmholtz resonator is analogous to a series inductor-resistor-capacitor (LRC) circuit with the natural frequency given by Eq. (4). Modified lumped-

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{1}{LC}} \qquad (4)$$

element models can incorporate deviations from the standard Helmholtz resonator configuration, such as flexible cavity walls in the case of intracranial aneurysms. *See* HARMONIC MOTION; HARMONIC OSCILLATOR; RESONANCE (ACOUSTICS AND MECHANICS); RESONANCE (ALTERNATING-CURRENT CIRCUITS).

If the resonator dimensions are not small compared to one-quarter of the wavelength $\lambda_0 = c/f_0$, the resonator can no longer be accurately regarded as a lumped-element system. In such cases, more detailed analysis, incorporating the spatial distribution of fluid pressure and velocity, is required to fully describe the resonator behavior. For example, a pipe open on one end, such as an organ flue pipe, has additional resonances at frequencies $f_0 = nc/(4l)$, where l is the effective pipe length and $n = 1, 3, 5, \ldots$, is any positive odd integer. *See* SOUND; VIBRATION.

Applications. Helmholtz resonance describes the lowest-frequency oscillation mode of many acoustic systems, including vehicles or dwellings with open windows, low-frequency loudspeakers, and the bodies of stringed musical instruments. Resonance may be induced by an external sound field, mechanical vibration, or an external flow field interacting with the cavity opening. Depending on the application, this resonance may be either a desired feature to be exploited or a nuisance to be controlled.

Vases acting as Helmholtz resonators were used by ancient Greek and Roman architects to modify the acoustic response of amphitheaters, as mentioned in Vitruvius' *De Architectura* (ca. 27 B.C.), and are also found in medieval European churches. In modern architecture, as well as acoustic filter and muffler design, Helmholtz resonators limit the transmission or reverberation of low-frequency sound that can be difficult to control by other means. To absorb sound over a wide frequency band, a collection of resonators with appropriately chosen resonance frequencies and quality factors can be employed. *See* MUFFLER.

T. Douglas Mast

Bibliography. H. L. F. Helmholtz, *On the Sensations of Tone*, 2d English ed., transl. by A. J. Ellis, Dover, New York, 1954; U. Ingard, On the theory and design of acoustic resonators, *J. Acous. Soc. Amer.*, 25:1037–1061, 1953; M. Long, *Architectural Acoustics*, Elsevier Academic, Burlington, MA, 2006; J. W. Strutt, Lord Rayleigh, *The Theory of Sound*, Dover, New York, 1945.

Hematologic disorders

Those disorders marked by aberrations in structure or function of the blood cells or the blood-clotting mechanism. Although many other diseases may be reflected by the blood and its constituents, the abnormalities of red cells, white cells, platelets, and clotting factors are considered to be primary hematologic disorders.

Erythrocytic abnormalities. Red-cell abnormalities are principally represented by the anemias and polycythemias. The anemias are marked by a decrease in the hemoglobin concentration, and may be due to blood loss or decreased production or excessive destruction of red cells. They are considered separately. *See* ANEMIA.

Polycythemias are disorders characterized by an increase in the numbers of circulating red cells and usually by a concomitant increase in hemoglobin. Relative polycythemia is really the result of hemoconcentration in which there is fluid loss from the blood, with corresponding increases in proportions of cellular elements. It is seen in excessive vomiting, diarrhea, dehydration, and similar conditions. *See* HEMOGLOBIN.

Secondary polycythemias result from a compensatory increase in the formation of red cells following hypoxia of the bone marrow. This condition is most often associated with chronic conditions, such as heart or lung diseases and prolonged exposure to high altitudes. The blood volume may be considerably expanded, and the individual may present a typically dusky red appearance. There are rare familial types due to synthesis of an abnormal hemoglobin with high oxygen affinity; this leads to decreased release of oxygen to the tissues, which in turn stimulates an increased production of red cells.

Primary polycythemia or polycythemia vera is a chronic and ultimately fatal disease of middle and old age in which there is a gradual increase in the number of red cells and usually an increase in the number of platelets and leukocytes. The viscosity of the blood is increased, which tends to slow the rate of blood flow; this factor, together with the high platelet count, predisposes to the thromboses commonly seen in the condition. Paradoxically there is also a hemorrhagic tendency, and the subject may bleed from the nose, stomach, and elsewhere. The cause of the disease is unknown, and although more benign, it has been likened to leukemia. The mainstay of treatment is repeated bloodlettings, supplemented when necessary by bone-marrow-suppressing agents such as radioactive phosphorus or hydroxyurea. *See* LEUKEMIA.

Leukocytic abnormalities. In a wide variety of conditions the many forms of white cell present in the circulation, bone marrow, and lymphoid tissues of the body may be altered in form or number.

The suffixes -philia and -penia denote increases and decreases, respectively, for the cells named. Leukopenia, neutrophilia, eosinophilia, and pancytopenia are examples of the wide range of possibilities. The absolute or relative increases or decreases of one or more types of leukocytes are often characteristic of certain disease states; for example, in infections neutrophils are increased both in absolute numbers and in relation to the other white cells. There may also be changes in the proportions of the different cells, and immature or atypical forms may be present.

The leukemias represent a special kind of malignancy in which there is usually an uncontrolled proliferation of one or more types of leukocytes, often reflected by great increases in the white cell count of the peripheral blood. The leukemias are generally classified, on clinical grounds, as acute, subacute, or chronic. They can also be subclassified further on the basis of cell-marker, karyotypes, or other studies. In addition, they are named to indicate the particular cell which shows neoplastic characteristics. The most common forms are lymphocytic and myelogenous leukemia. In the former, large numbers of abnormal lymphocytes may be present in the blood; in the latter, the cells of the granulocyte series are involved. Other varieties exist, such as monocytic, eosinophilic, and plasma-cell leukemia, but they are much less frequently encountered.

Many similarities exist in the clinical course and pattern of lesions seen in the leukemias. In the acute forms, the onset is abrupt, with the appearance of fever, malaise, and weakness. The course in the absence of treatment is usually rapid, with susceptibility to infections and a tendency toward hemorrhage accounting for many complications or leading to death. The outlook for acute lymphocytic leukemia, which is the most common type seen in children, has improved dramatically: long-term remissions and probable cures following treatment are now seen in more than 50% of children and nearly 20% of adults.

Chronic leukemias may be nearly acute to almost completely benign. In general, the granulocytic forms tend to be more severe than the lymphocytic forms. The first evidence of chronic lymphatic leukemia may be discovered more or less by accident, by periodic physical examination, or by routine blood counts. Although more severe forms are seen, the course often extends over many years or even decades, and symptoms attributable to the leukemia may be very late in developing. On the other hand, chronic myelogenous leukemia (chronic granulocytic leukemia) is rarely discovered accidentally and, when so discovered, signs and symptoms attributable to the leukemia soon develop. It may progress into an acute form. Symptoms of the chronic leukemias include fatigue, lack of exercise tolerance, pallor, and weight loss.

The treatment of the leukemias is based on the type, the stage at which it is diagnosed, and the age of the patient. The most common type of acute lymphocytic leukemia in children responds to complex and novel treatment methods that include the use of a combination of two or more chemotherapeutic agents together with steroids. More than half the children will be free of symptoms 10 years after therapy ends. The same treatment has been used for the other types of childhood acute lymphocytic leukemia and all the adult forms, but relatively few respond well. High-dose chemotherapy followed by marrow transplantation is recommended for patients under 50 who have an HLA-compatible sibling. Although this form of therapy is not without great risk, cures have resulted. The prognosis for chronic lymphocytic leukemia that is discovered before signs and symptoms are present is excellent, and the mean survival time is approximately 7 years; survival 20 years or longer is not at all unusual. When symptoms or complications are present, the disease is usually treated with prednisone and an alkylating agent such as chlorambucil. Most patients with chronic myelogenous leukemia respond well initially to therapy with an alkylating agent such as busulfan or hydroxyurea, but relapses are typical and the mean survival time of approximately 3 years is only slightly increased by treatment. Most forms of the acute nonlymphocyctic leukemias are treated with chemotherapeutic agents, and while some show a good initial response, the prognosis is very poor. Bone marrow transplantation as described above for the treatment of acute lymphocytic leukemia has been recommended for the treatment of nonlymphocytic leukemia and also for chronic myelogenous leukemia.

Hemorrhagic disorders. The hemorrhagic disorders result from a large number of known and unknown causes or contributing factors, often of a diverse nature. The initial or primary events that halt bleeding from a very small wound are the formation of a platelet plug which seals the hole in the vessel wall and arteriolar vasoconstriction. The secondary event is the subsequent fortification of the plug by fibrin. The stimulus that causes the platelets to stick together and form the plug is believed to be exposure to substances released from the damaged blood vessels which act in concert with certain plasmatic factors, such as adenosine diphosphate (ADP) released from damaged red cells and the von Willebrand factor. A decrease or abnormality of the von Willebrand factor, a reduction in the number of platelets, a qualitative defect in platelets, or a defect in the vascular wall can all result in failure of the primary hemostatic mechanism, with spontaneous bleeding; this is referred to as purpura. *See* HEMORRHAGE.

Blood vessel damage may occur as a result of direct or indirect damage by microorganisms during infections, as the result of vitamin C deficiency (scurvy), and following hereditary defects in blood vessel development. A number of comparatively rare disorders, usually with a hypersensitivity component, also fall into this category. There are several forms of thrombocytopenia, all of which are characterized by a decrease in thrombocytes or platelets in the circulation. Qualitative changes in the platelets can result in impairment of their function, and may be inherited or acquired, as is seen following aspirin ingestion. Since the platelets are involved in primary plug formation, such deficiencies or abnormalities lead to purpura. Treatment is dependent on the cause and usually is restricted to those occasions when bleeding is severe or life-threatening or when surgery is about to be performed. Thus if the platelet count is very low or the platelets are defective, transfusion of platelets may be indicated. Most forms of von Willebrand's disease are now treated with desmopressin; however, in severe cases cryoprecipitate may be indicated.

Fibrin or clot formation is the result of the sequential or stepwise interaction of platelets and several blood-clotting factors which are proteins; tissue factor is also involved. A defect in fibrin formation gives rise to the type of bleeding seen in hemophilia, in which the primary hemostatic plug is formed normally but breaks down several hours or days later owing to lack of adequate fortification by fibrin. *See* HEMOPHILIA.

Defects in the clotting process that result in excessive bleeding may be due to an acquired or hereditary deficiency or abnormality of a clotting factor, as in hemophilia A or B. The bleeding that may be associated with hereditary conditions is usually controlled by the intravenous administration of concentrates of the deficient factor, which are obtained from plasma. For example, factor VIII concentrates are used in hemophilia A, factor IX concentrates in hemophilia B, and cryoprecipitate in fibrinogen deficiencies. Fresh-frozen plasma is usually used to treat those conditions for which concentrates are not available. Acquired deficiencies of clotting factors are seen in many conditions, including liver disease, vitamin K deficiency, and disseminated intravascular clotting, but they may be caused by the spontaneous development of antibodies directed against one of the clotting factors. The bleeding associated with liver disease may be treated with fresh-frozen plasma and platelet transfusions; the treatment of disseminated intravascular clotting is directed at the primary cause of the condition. The bleeding that results from an antibody is treated with infusions of the clotting factor against which the antibody is targeted, immunosuppressives, and steroids. Sometimes attempts are made to bypass the action of the antibody by intravenous administration of an activated clotting factor. *See* BLOOD. Cecil Hougie

Bibliography. D. L. Bernard, *Clinical Hematology*, 1989; B. Brown, *Hematology*: *Principles and Practice*, 1993; C. Kjeldsberg et al., *Practical Diagnosis of Hematologic Disorders*, 3d ed., 2000; G. R. Lee et al., *Wintrobe's Clinical Hematology*, 10th ed., 1999; W. J. Williams et al., *Hematology*, 4th ed., 1990.

Hematopoiesis

The process by which the cellular elements of the blood are formed. Blood contains many free-floating cells which are moved throughout the body within the blood vessels. The three main types of cells are the red cells (erythrocytes), which serve to carry oxygen, the white cells (leukocytes), which function in the prevention of and recovery from disease, and the thrombocytes, which function in blood clotting. In humans there is only one white cell in the blood for every 700 red cells.

The formation of these cells is one of the most active and important processes in the body. Most of the circulating cells live only for a short time and must be replaced in order to maintain life. For instance, in the human adult a red blood cell has a life of 120 days; 250 billion new red cells have to be produced daily to replace those that are destroyed. Serious consequences result if the balance of replacement and destruction of blood cells is impaired by disease, by genetic defects (as in hemolytic anemia), or by irradiation with x-rays or other high-energy radiations. The lymphocytes, a class of leukocytes, are more sensitive to ionizing radiation than any other cell type in the body, and changes in the number of lymphocytes may be used as a good index of radiation damage. *See* HEMATOLOGIC DISORDERS; RADIATION INJURY (BIOLOGY).

Blood Cell Production in Adults

Blood cells originate in the reticuloendothelial tissue, which is a loose, fibrous, highly vascularized mesh of fibers, endothelial cells, and macrophages. Within the spaces of the tissue are found the precursor (blast) cells of the definitive adult types. For the sake of convenience, the reticuloendothelial tissue is divided into two general but imprecise types: lymphoid and myeloid tissue. This classification certainly does not apply to the situation present in the embryo and in certain blood disorders.

Lymphoid tissue. This tissue is primarily localized in the lymph nodes of the lymphatic system and is also in the spleen, thymus, and bone marrow. Several classes of white cells are produced, including the lymphocytes, macrophages, and monocytes. These cells function in producing antibodies and in removing detritus formed during infectious diseases; organisms in which the lymphoid tissue is reduced or destroyed (such as after sustaining irradiation damage) usually succumb quickly to disease. There is evidence that many of the early blast cells of the lymphoid tissue arise in the embryonic thymus; they then migrate to their final location late in embryonic development. In juveniles and young adults the thymus may also produce a humoral factor which is important for the production of some differentiated cell types and antibodies by the lymphoid tissues. *See* CELLULAR IMMUNOLOGY; LYMPHATIC SYSTEM; THYMUS GLAND.

Myeloid tissue. This tissue is normally limited in humans to the red bone marrow of the ribs, sternum, vertebrae, and proximal ends of the long bones of the body. It is concerned with the production of the erythrocytes and certain types of leukocytes. The latter are the granular leukocytes (called eosinophiles, basophiles, and neutrophiles on the basis of the affinity of granules in their cytoplasm for certain dyes) and megakaryocytes. Fragments of megakaryocytes form the blood platelets (thrombocytes), which are necessary for blood clotting.

Blast differentiation. There are several theories concerning the origin of the blood cells, and they differ mainly in the number of types of blast cells thought to be involved in the process. For instance, the currently favored view holds that there is only one type of blast cell and that it is capable of giving rise to all other blood cells. This pluripotential blast cell is morphologically indistinguishable from a circulating lymphocyte, and may be called a hemocytoblast. It is found in both lymphoid and myeloid tissues. The hemocytoblast may give rise to lymphocytes, macrophages, megakaryocytes, monocytes, granular leukocytes, and erythrocytes, that is, to all the definitive circulating blood cell types. The hemocytoblasts probably are ultimately derived from the mesenchyme of the early embryo. Well-defined stages of the differentiation of the various cell types can be recognized by the hematologist using various stains.

Though the identification and characterization of the progenitor cells that give rise to the red and white cells of the blood are difficult, incontrovertible proof of the existence of pluripotential blast cells has been furnished. If a suspension of bone-marrow cells is injected into the circulation of a lethally irradiated host mouse, some of the injected cells colonize the host's spleen and give rise to localized nodules of either erythrocytes, granulocytes, megakaryocytes, or lymphocytes. These nodules are macroscopically visible and appear before the host dies. It has been shown that the cells that compose a nodule are a clone, derived from a single cell. It has become possible to culture cells from the marrow in very elaborate tissue culture media and to demonstrate cells that have many of the properties expected of precursor blast cells.

Erythropoiesis. More is known about the transformation of a blast cell into an erythrocyte than about any of the other types of blood cells. A hormone, erythropoietin, regulates the production of differentiated cells from the blast cell. Erythropoietin is thought to be produced primarily in the kidney and is essential, at least in higher vertebrates, for the formation of red cells. In physiological or pathological states where a surfeit of red cells exists, less erythropoietin is produced. When a great demand for red cells exists (as in certain anemias), the level of erythropoietin may be high. Colonies of red cells do not form in the spleens of lethally irradiated recipients transfused with bone-marrow cells from untreated donors when the hormone is lacking. The hormone is thought to regulate the production of erythroblasts, the earliest recognizable cells of the erythrocytic series. *See* SPLEEN.

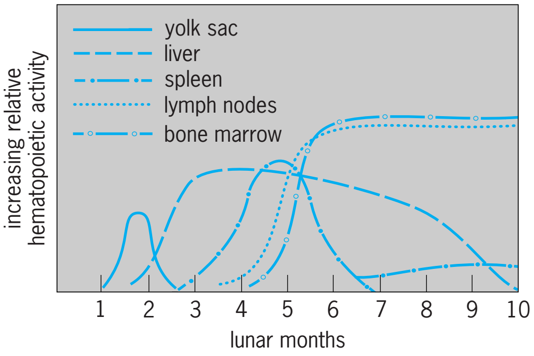
The stages of the transformation of a blast into an erythrocyte are morphologically and functionally distinct. The erythroblasts are spherical cells with a basophilic cytoplasm. They continue to divide and give rise to various types of polychromatic erythroblasts; it is in these cells that hemoglobin is actively synthesized, giving a reddish tinctorial quality to stained cells. Furthermore, the basophilia of the cytoplasm continually decreases because cytoplasmic ribosomes are lost. Toward the end of the polychromatic phase, the nucleus condenses and no further synthesis of nuclear ribonucleic or deoxyribonucleic acid can be detected. In mammals the nucleus is subsequently extruded, but in all other vertebrates the nucleus is retained, although it may be functionally inactive. After the nucleus is extruded some synthesis of hemoglobin still goes on, and the cells are called reticulocytes because of the reticulum of basophilic material present in the hemoglobin-filled cytoplasm. Some reticulocytes may be released to the circulation. Finally all synthesis of hemoglobin in the bone marrow ceases, and the cells are released to the circulation as the definitive erythrocytes. In some anemias the differentiative process may occur very much faster, or some stages may be skipped, and nucleated (immature) cells may be released to the circulation. *See* HEMOGLOBIN.

Embryonic Origins of Blood Cells

Circulating blood cells can be detected very early in the development of vertebrates; they are present as soon as there is a circulatory system to transport them. In the embryo there is a succession of changing sites of primary production. The details may vary among the different vertebrates, but generally one can detect three sites of origin. The yolk sac produces the first blood cells; subsequently certain other organs (spleen, liver, venous sinusoids) become productive. Finally, the normal adult sites (primarily, the bone marrow and lymph nodes) assume their important role near the time of birth. The sequence of different hematopoietic centers in humans is shown in the **illustration**; this sequence comprises most of the features found in all vertebrates.

Yolk sac. Definite areas of the embryo are allocated very early to the production of the first blood cells. These areas, called blood islands, are located in the mesoderm of the yolk sac. Groups of the primordial precursors of blood cells coalesce to form stellate, interconnected groups of highly basophilic cells (hemocytoblasts). The central portions of the blood islands undergo erythropoiesis (although it may differ in detail from the sequence previously described in adult bone marrow), while the cells on the periphery of the interconnected islands flatten and become the endothelial lining of the blood vessels of the yolk sac. Only erythrocytes form during this first wave of hematopoiesis, and the hemoglobins contained in these red cells differ considerably in their amino acid sequence from the hemoglobins of the adult. *See* YOLK SAC.

Formation of definitive blood cells. The yolk sac usually forms red cells for only a relatively short time. As its production wanes, hematopoiesis becomes detectable, and then intense, in a variety of extramedullary sites. In humans the liver next assumes important hematopoietic functions, and the process here is similar to that in the bone marrow of the adult. Nonnucleated red cells are produced, which may (in mice) or may not (in humans) contain hemoglobin identical to the adult. Granular leukocytes and macrophages also appear. Eventually hematopoietic activity in the liver ceases, and other sites, such as the spleen, become active. Later



Hematopoietic activity in humans during the normal gestation period of 10 lunar months of 28 days each.

in development the lymph nodes form and become important centers of origin of lymphocytes. Near the time of birth, the bone marrow becomes active in producing red cells, granular leukocytes, and megakaryocytes.

There are many differences in the hematopoietic activity of the yolk sac compared with subsequent sites. For instance, in the human yolk sac, red cells are nucleated, large, and basophilic and contain an embryonic hemoglobin composed of two α-polypeptide chains (or special variants designated zeta) and two ϵ-polypeptide chains. The liver produces red cells that seem indistinguishable from those of the adult, but they contain a fetal hemoglobin composed of two α- and two γ-polypeptide chains. Eventually the bone marrow produces red cells containing the usual human adult hemoglobin (two α- and two β-polypeptide chains).

The regulation of the changing sites of hematopoiesis and their relation to the nature of the blood cells produced are still mysteries. *See* BLOOD; LIVER. Fred Wilt

Bibliography. D. W. Golde (ed.), *Hematopoiesis*, 1984.

Hemiascomycetes

A class of the phylum Ascomycota that includes the yeasts and yeastlike fungi. These are morphologically simple fungi; no ascoma is formed, and the asci are produced free on the host or substrate. Asexual reproduction occurs by the formation of blastospores (budding) or, less frequently, by fission arthrospores. Two main orders are recognized, the Saccharomycetales and the Taphrinales. *See* YEAST.

The vegetative body (thallus) of the Saccharomycetales may be either unicellular (true yeasts) or mycelial. In unicellular species, asci form when two vegetative cells fuse, and then the fused cell undergoes meiosis to form ascospores. In mycelial species, the hyphae are not very extensive. Sexual reproduction occurs when adjacent cells extend short lateral branches that fuse to form the asci. Variations on these modes of ascus formation, however, are common among the yeasts.

The Saccharomycetales are common on substrates high in sugars, such as plant exudates, ripe fruits, and flower parts. Because they are microscopic, they are recognized mainly from cultures that have a homogeneous appearance and a characteristic odor. The most important genus is *Saccharomyces; S. cerevisiae* is the common bakery and brewery yeast, and *S. ellipsoideus* is used in winemaking. An important mycelial species is *Nematospora coryli*, which causes yeast spot disease of various crops.

The order Taphrinales includes the leaf curl disease fungi. The most widely recognized species are *Taphrina deformans*, cause of leaf curl of peach and almond trees, and *T. caerulescens*, cause of leaf blister of oaks. These fungi produce a well-developed mycelium in the host tissue but, when grown in culture, form only a yeastlike colony of single cells—a phenomenon known as dimorphism. Asci are produced when special binucleate hyphal cells beneath the host cuticle undergo nuclear fusion and the resulting diploid cell elongates to form an ascus on the leaf surface. The nucleus undergoes meiosis, and ascospores are formed. *See* ASCOMYCOTA; EUMYCOTA; FUNGI.

On the basis of molecular data, some workers now propose separating the Taphrinales and the fission yeast, *Schizosaccharomyces*, into a new class, Archiascomycetes. These fungi are considered to be more primitive and phylogenetically basal to the rest of the ascomycetes. Richard T. Hanlin

Bibliography. C. J. Alexopoulos, C. W. Mims, and M. Blackwell, *Introductory Mycology*, 4th ed., John Wiley, New York, 1996; J. A. Barnett, R. W. Payne, and D. Yarrow, *Yeasts: Characteristics and Identification*, 1990; D. R. Berry, *Biology of Yeast*, 1982; N. J. W. Kreger-van Rij (ed.), *The Yeasts: A Taxonomic Study*, 3d ed., 1984; H. Nishida and J. Sigiyama, Archiascomycetes: Detection of a major new lineage within the Ascomycota, *Mycoscience*, 35:361–366, 1994.

Hemicellulose

The term adopted by E. Schulze in 1891 to designate the plant cell components which are made soluble by dilute alkali or which go into solution quite readily in hot dilute mineral acids with the formation of simple sugars. Hemicelluloses constitute about one-fourth of perennial plants and about one-third of annual plants. The term is usually applied to those polysaccharides in the cell wall of land plants which are extractable by dilute alkaline solutions. The term has also been used to include all the polysaccharide components of the cell wall other than cellulose. *See* CELL WALLS (PLANT).

How the hemicelluloses, which are often water soluble after extraction from plant tissue, are anchored in the cell wall so as to be unextractable by water is still a controversial question. As an explanation, hemicellulose-lignin or hemicellulose-cellulose covalent bonds have been postulated. Hemicelluloses extracted from different plant sources, although they often have many common characteristics, are rarely identical. In fact, many different hemicelluloses usually occur intermixed with each molecular type representing different degrees of polymerization. Because of this heterogeneity, few hemicelluloses have been isolated in a homogeneous state. Therefore, relatively little is known of the structure of these compounds that compose almost one-third of the carbohydrates in woody tissue.

Preparation. Hemicelluloses are generally prepared by alkaline extraction of dried, defatted plant tissue or a special delignified plant tissue which is called holocellulose. The latter is usually prepared by treatment of the defatted plant with chlorous acid solution, which solubilizes the lignin but leaves the polysaccharide behind in the same morphological relation as in the original plant tissue. Hemicelluloses are more readily extractable from holocellulose than from dried, defatted plant tissue, and can be divided into two solubility classes, hemicellulose A and hemicellulose B.

Hemicellulose A is insoluble in slightly acidic solution, and therefore precipitates from the alkaline extract upon acidification. This class is composed mainly of high-molecular-weight linear glycans that contain small amounts of uronic acid. The degree of polymerization (DP) of these polysaccharides usually is 70–200. In comparison, the DP of cellulose is 2500–3500. For most plants, the A fraction is the largest because within this fraction occurs the abundant and widely distributed polysaccharide arabinoglucoxylan, which hydrolyzes to L-arabinose, D-glucuronic acid, and D-xylose. Sometimes the D-glucuronic acid unit has a methyl ether group at position C4 of the pyranose ring.

Hemicellulose B, which is precipitated from the neutralized alkaline extract by the addition of excess methanol or ethanol, consists of branched-chain molecules of a lower degree of polymerization than hemicellulose A. Its uronic acid content is higher than that of hemicellulose A. These hemicellulose molecules may also have D-galactopyranose and D-glucopyranose units attached at positions yet undefined. Further fractionation procedures, such as the formation of a complex between the hemicellulose molecules and copper salts and fractional precipitations with organic solvents, must be used to isolate a homogeneous polymer.

D-Xylose is the dominant building unit of the hemicelluloses of most woods and annual plants. The D-xylose content of hemicellulose from hardwood is generally higher (20–25%) than that from softwood (7–12%). D-Mannose is also very abundant in hemicelluloses; the mannose content of softwoods is usually higher than that of hardwood. Often it occurs as a polymer, mannan, or in combination with D-glucose or D-galactose as a glucomannan, galactomannan, or galactoglucomannan.

Uses. Hemicelluloses are important to the paper industry. In chemical wood pulps, hemicellulose is needed for satisfactory pulp quality. Its presence aids the swelling of the pulp, the bonding of the fibers, the bursting strength, tensile strength, tear resistance, folding endurance, opacity, and specific surface of the pulp sheet. The optimum hemicellulose content of a pulp will vary, depending on the type of wood and the conditions of the pulping operations. Excessive amounts of hemicellulose will produce a brittle or glassine paper. *See* PAPER.

On the other hand, if pulps are to be used for purposes other than papermaking, the presence of hemicellulose may be quite disadvantageous. For example, in the production of pulps which are to be used in the manufacture of rayon, cellophane, and cellulose esters and ethers, it is often essential to obtain an α-cellulose very low in hemicellulose. α-Cellulose is insoluble in 17.5% sodium hydroxide solution. In cellulose acetate manufacturing, small amounts of mannan or xylan cause hazy acetate solutions and clogged filter presses.

Hemicelluloses also serve as nutrients for yeasts, and they can be used for raw material in the production of furfural and ethyl alcohol. *See* CELLULOSE.

Roy L. Whistler

Bibliography. M. R. Brown, Jr. (ed.), *Cellulose and Other Natural Polymer Systems*: *Biogenesis*, *Structure*, *and Degradation*, 1982; M. F. Chaplin and J. F. Kennedy (eds.), *Carbohydrate Analysis,* 2d ed., 1994; M. P. Couglan and G. P. Hazelwood (eds.), *Hemicellulose and Hemicellulases*, 1993; R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 5th ed., 1994; F. Loewus and W. Tanner (eds.), *Plant Carbohydrates I*: *Intracellular Carbohydrates*, 1982.

Hemichordata

A group of deuterostome animals that includes the classes Enteropneusta, Pterobranchia, and Planctosphaeroidea. The last, a monospecific class, is represented by *Planctosphaera pelagica*, a planktonic larva that occupies low depths and resembles the larval tornaria of Enteropneusta. *See* ENTEROPNEUSTA; PTEROBRANCHIA.

The hemichordates have a slender tubular diverticulum that projects forward into the protosome from the roof of the buccal cavity. That organ, also called a stomochord or buccal diverticulum, is a supporting axial rod of the protosome and resembles the notochord of Chordata, but it does not have the same position, structure, origin, or function, and so they are not homologous. *See* CHORDATA.

The hemichordates are not plentiful animals. All are marine species that live in a wide range of habitats and depths, from intertidal to abyssal, and show a worldwide distribution. They vary in size from a fraction of an inch, or a few millimeters, such as *Rhabdopleura*, to 7 ft (2 m) or more, as with *Balanoglossus gigas*.

The members of the phylum differ widely. The enteropneusts are vermiform and solitary, whereas the pterobranchs are sacciform and colonial. The adult hemichordate is characterized by a division of both the body and the coelom into three parts: the anterior region, containing the protocoel (protosome); the middle region, containing two mesocoels (mesosome); and the posterior region, containing two metacoels (metasome). Those divisions differ in length and structure. The protosome contains a contractile heartlike vesicle, a blood sinus, and a glomerulus that rest on the buccal diverticulum. The mesosome contains a dorsal thickening of the nervous system, called a neurochord, which is isolated from the epidermis and which shows giant nerve cells and fibers that facilitate fast responses. The mesosome may also bear one or more pairs of dorsal tentaculated arms that resemble a lophophore. The mouth opens anteroventrally in the mesosome. The metasome is the longest region; anteriorly it possesses numerous paired gill slits that extend from the dorsal pharynx wall to the body surface, with one pair or none being found in pterobranchs. The gill slits are considered to be homologous with the chordate gill slits. The gill apparatus itself may function to separate food from water, which is expelled through the gill slits. The gonads are also in the metasome and have a retroperitoneal origin. They are numerous in enteropneusts but limited to one or two in pterobranchs, and they open by separate dorsal pores. Sexes are separate, and mating organs are lacking.

The nervous system is primitive, consisting of an intraepidermal plexus. Two thickenings, middorsal and midventral, form the dorsal and ventral nerve cords. The neurochord perhaps represents a nervous center, but it lacks any special concentration of nervous elements. Scattered monociliated sensory cells respond to tactile and chemical stimuli. The circulatory system is open, with two main longitudinal contractile vessels. The colorless blood moves forward in the dorsal vessel, passes into a blood sinus on the buccal diverticulum, and is then pumped into the glomerulus by a pulsating heartlike vesicle before running backward to the ventral vessel. The glomerulus is formed from an infolding of the hind wall of the protocoel and possibly has excretory and osmoregulatory functions.

The gut is straight in enteropneusts, with a terminal anus at the end of the metasome; in pterobranchs, it is U-shaped and the anus is near the mouth. All hemichordates are filter feeders, taking in microscopic organisms by means of ciliary streams. The enteropneusts trap the food with the protosome, the pterobranchs use their tentacles, and *Planctosphaera* manipulates a ciliary band.

Spawning is synchronized, fertilization is external, and cleavage is holoblastic, radial, and nearly equal. Some species have large eggs that develop directly, without a larval stage; others have a typical larva, the tornaria, which resembles the bipinnaria or auricularia of echinoderms; the pterobranchs have a planulalike larva. In this latter group, asexual reproduction by budding occurs.

The hemichordates are a primitive group, having a tripartite body and coelom; their embryonic development resembles the echinoderms, with which they also share a primitive nervous system. The Hemichordata, therefore, may be a group at a low level of evolution, between echinoderms and chordates. Jesús Benito

Hemicidaroida

An extinct paraphyletic order of regular sea urchins (Echinoidea), identified by having a plain, unsculptured test, tubercles that are perforate and crenulate, and no suranal plate intercalated into the apical disc. They almost certainly include ancestors of later groups that developed imperforate and noncrenulate tuberculation and are thus paraphyletic.

Two families are generally placed within the order, Hemicidaridae and Pseudodiadematidae. Hemicidarids have large primary interambulacral tubercles and much smaller ambulacral tubercles, the ambulacra typically being narrow and sinuous and composed of simple plates adapically. Pseudodiadematids have equal-sized ambulacral and interambulacral tubercles at the ambitus, wider ambulacra, and trigeminate or polygeminate plate compounding throughout.

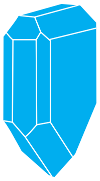
The oldest hemicidaroid is Late Triassic (Upper Norian) and the youngest comes from the Upper Cretaceous (Campanian). They appear to have been epifaunal grazers. *See* ECHINOIDEA. Andrew Smith

Hemidiscosa

An order of the subclass Amphidiscophora in the class Hexactinellida. These sponges are distinguished from the order Amphidiscosa in that the birotulates are hemidiscs with asymmetrical ends. *Hemidiscella* from the Cretaceous is an example. *See* AMPHIDISCOSA; HEXACTINELLIDA. Willard D. Hartman

Hemimorphite

A mineral sorosilicate having the composition $Zn_4Si_2O_7(OH)_2 \cdot H_2O$; an ore of zinc. It crystallizes in the orthorhombic system, pyramidal class, and thus the prismatic crystals have different forms at top and bottom (see **illus.**). Also, as a result of the symmetry, the vertical axis is polar and crystals display the properties of pyroelectricity and piezoelectricity. There is perfect prismatic cleavage. Botryoidal and staclactic aggregates, showing a crystalline surface and frequently impure, are common. Crystals are usually colorless and the aggregates white, but in some cases there are faint shades of green, yellow, and blue. The mineral has a vitreous luster, a hardness of $4^1/_2$ to 5 on Mohs scale, and a specific gravity of 3.45. Hemimorphite frequently resembles the zinc carbonate, or smithsonite, but can be distinguished by the reaction with hydrochloric acid. Hemimorphite is slowly soluble but smithsonite effervesces in cold acid. *See* PIEZOELECTRICITY; PYROELECTRICITY; SMITHSONITE.



Hemimorphite crystal habit. (*After C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*)

Hemimorphite is a secondary mineral found in the oxidized portion of zinc deposits associated with smithsonite, sphalerite, cerussite, anglesite, galena, and more rarely the oxidized minerals of copper. Widely distributed, it has been mined in Belgium, Germany, Romania, England, Algeria, and Mexico. In the United States it is found at Sterling Hill, New Jersey; Friedensville, Pennsylvania; and Elkhorn Mountains, Montana. *See* SILICATE MINERALS; ZINC. Cornelius S. Hurlbut, Jr.

Hemiptera

An order of the class Insecta sometimes referred to as the Heteroptera. Both names refer to the forewings, which are differentiated into a thickened basal area and a membranous apical region. These are the true bugs. The related order, Homoptera, has forewings of uniform texture. Included in Hemiptera are such common insects as the bedbugs, stink bugs, plant bugs, lace bugs, and backswimmers. *See* HOMOPTERA.

The true bugs number about 25,000 species. They are known from all continents except Antarctica and occur on most islands. One group, the marine water striders (*Halobates*), inhabits the open ocean. Hemiptera range in size from small aquatic and ground-inhabiting forms approximately 0.04–0.08 in. (1–2 mm) in length to the giant water bugs that are 4 in. (100 mm) or more.

In habits, the true bugs range from strictly phytophagous types attached to a single host plant to general predators on other insects and even to specialized ectoparasites of bats. Many species are of economic importance as plant pests or vectors of disease. They occur in vegetation, on the ground, in and on the water, and in the nests of termites. One family, the Aradidae, lives on fungi on the bark of dead trees, and certain Reduviidae and Cimicidae live in caves. Among aquatic forms, members of the genus *Rhagovelia* swim on riffles of swift-flowing streams, and *Aphelocheirus* remains permanently submerged in rivers. Most water bugs depend on surface air held to the body by air spaces and hairs on the abdomen. As oxygen is depleted in the air bubble, it is replaced from the surrounding medium by diffusion.

A few Hemiptera, such as *Dacerla*, *Coquillettia*, and *Pilophorus*, mimic ants. Others, including cotton stainers (*Dysdercus*) and certain reduviid predators (*Phonoctonus*), resemble distasteful cantharids and other insects of orange or yellow and black colors, which is a protective Müllerian type of mimicry. Still other bugs resemble their surroundings and hence are concealed from birds, lizards, and other enemies. One of the most striking of these is the flat, barklike *Phloea* of South America. Stridulation, or sound production, is less conspicuous in Hemiptera than in the cicadas of Homoptera and katydids of Orthoptera but is developed in several groups. Stink bugs, flat bugs, the predacious reduviid bugs, and several aquatic groups are examples. The sounds are produced either by rubbing filelike portions of the legs against knife edges on the body wall, by rubbing parts of the wings against the body, or by rubbing the beak along a prosternal cross-striated stridulatory furrow, as occurs in the Reduviidae (**Fig. 1**). *See* PROTECTIVE COLORATION.

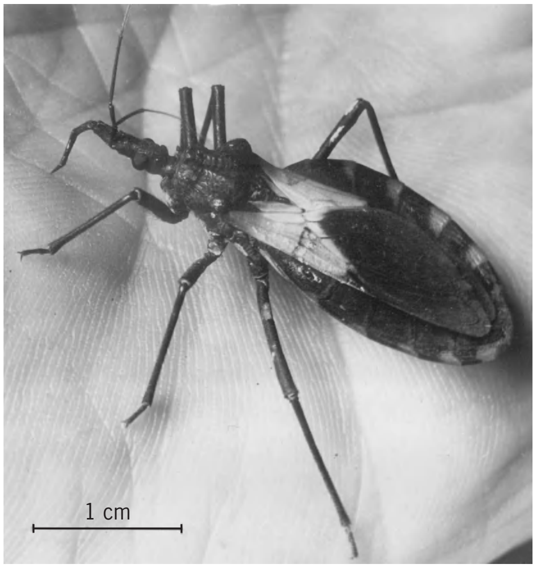


Fig. 1. *Triatoma pallidipennis*, of the Reduviidae. (*Photograph by E. S. Ross*)

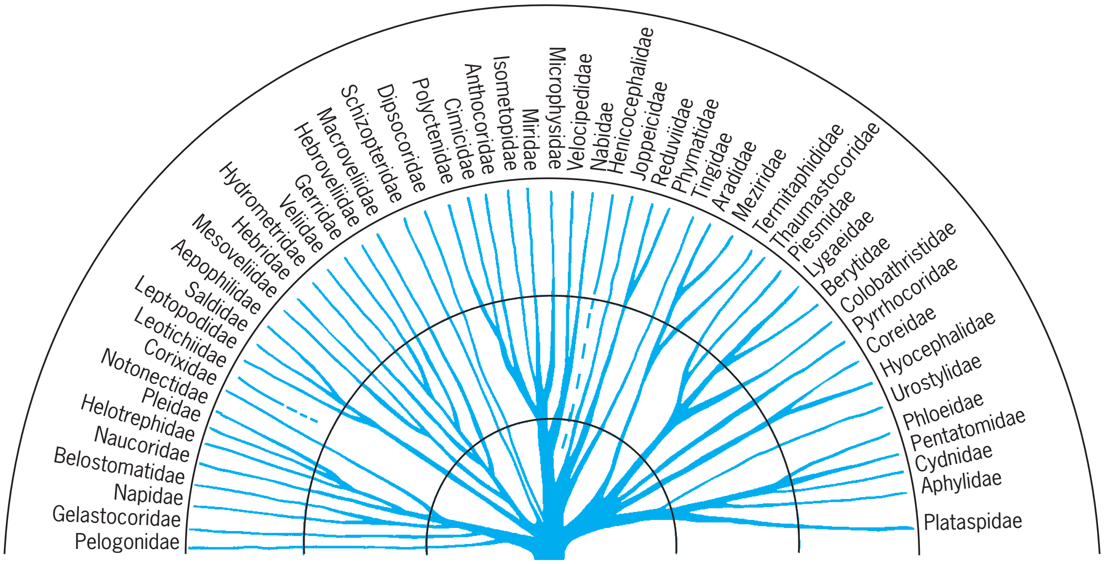


Fig. 2. Relationships of families of Hemiptera.

TABLE 1. Families of Hemiptera

Family	Common name	Distribution	No. of species
Subdivision Hydrocorisae			
Corixidae	Water boatmen	General	300
Nepidae	Water scorpions	General	170
Belostomatidae	Giant water bugs	General	140
Notonectidae	Backswimmers	General	170
Pleidae	None	General	20
Helotrephidae	None	Tropical	20
Naucoridae	Creeping water bugs	General	200
Gelastocoridae	Toad bugs	Tropical and subtropical	80
Ochteridae	Velvety shore bugs	Tropical and subtropical	20
Subdivision Amphibicorisae			
Gerridae	Water striders	General	300
Veliidae	Smaller water striders	General	200
Hydrometridae	Marsh treaders	General	50
Mesoveliidae	Water treaders	General	20
Hebridae	Velvety water bugs	General	40
Subdivision Geocorisae			
Superfamily Leptopodoidea			
Saldidae	Shore bugs	General	200
Leptopodidae	None	Tropical and subtropical	20
Leotichidae	None	Oriental	2
Superfamily Dipsocoroidea			
Dipsocoridae	None	General	
Schizopteridae	Jumping ground bugs	Tropical and subtropical	
Superfamily Cimicimorpha			
Cimicidae	Bat, bed, bird bugs	General	80
Anthocoridae	Flower bugs	General	300
Polyctenidae	Bat bugs	Tropical and subtropical	20
Miridae	Plant bugs	General	5000
Microphysidae	None	Palearctic	30
Plokiophilidae	None	Tropical	20
Nabidae	Damsel bugs	General	250
Tingidae	Lace bugs	General	700
Vianaidae	None	Neotropical	2
Thaumastocoridae	Palm bugs	Tropical	11
Superfamily Enicocephaloidea			
Enicocephalidae	Gnat bugs	General	300
Superfamily Reduvioidea			
Reduviidae	Assassin bugs	General	3500
Superfamily Aradoidea			
Aradidae	Flat bugs	General	800
Termitaphididae	Termite bugs	Tropical	10
Superfamily Pentatomorpha			
Idiostolidae	None	Chilean	2
Lygaeidae	Lygaeid bugs	General	2000
Thaumastellidae	None	Ethiopian	1
Colobathristidae	None	Tropical	70
Berytidae	Stilt bugs	General	100
Malcidae	None	Ethiopian, Oriental	30
Piesmatidae	Ash-gray leaf bugs	General but discontinuous	20
Pyrrhocoridae	Pyrrhocorid bugs	General	300
Largidae	None	General	100
Coreidae	Coreid bugs	General	2000
Rhopalidae	None	General	300
Stenocephalidae	None	Old World, Neotropical	20
Hyocephalidae	None	Australia	1
Pentatomidae	Stink bugs	General	2500
Phloeidae	Bark bugs	Neotropical	5
Plataspidae	None	Old World	400
Lestoniidae	None	Australia	1
Cydnidae	Ground or burrower bugs	General	600
Urostylidae	None	Oriental and Australian	50
Aphylidae (not placed)	None	Australian	2
Joppeicidae	None	Mediterranean	1

Most Hemiptera are bisexual and oviparous, but parthenogenesis is known and a pseudoplacental organ results in a special type of viviparity in Polyctenidae. Mating usually takes place on vegetation or on the ground, the pairing being end-to-end in stink bugs, squash bugs, chinch bugs, and similar species, and with the male above the female in most others. Bedbugs (Cimicidae) have a special organ of Ribaga (or organ of Berlese) through which spermatozoa are introduced to the hemocoel of the female where they penetrate the ovarioles and fertilize the ova. *See* INSECT PHYSIOLOGY; INVERTEBRATE EMBRYOLOGY.

Classification. The Hemiptera were first divided by P. Latreille, in 1825, into Hydrocorisae (water bugs) and Geocorisae (land bugs). L. Dufour, in 1833, proposed a third group, Amphibicorisae (surface water bugs). F. Fieber, in 1851, substituted the Cryptocerata (hidden antennae) for Hydrocorisae and Gymnocerata (exposed antennae) for Geocorisae. Both systems are in use today, but the Latreille-Dufour system conforms best to recent studies. In 1954, the Geocorisae were divided into Cimicomorpha and Pentatomomorpha by D. Leston, J. Pendergrast, and T. Southwood. R. H. Cobben, in 1968, recognized nine more or less equivalent major subdivisions in the Hemiptera. **Table 1** represents a balance between extremes and includes each distinctive group, down to the family level. The cladogram shown in **Fig. 2** summarizes the last published survey (1956) of the possible phylogeny of the families of the Hemiptera.

Habitats. The Hemiptera occupy various terrestrial and aquatic habitats (**Table 2**). Some species are intimately associated with plants and animals.

Morphology and Embryology

In Hemiptera and Homoptera the mouthparts are elongate and slender, forming a sucking mechanism with sheathlike labium and needlelike mandibulary and maxillary stylets (**Fig. 3**). In Homoptera, as in leafhoppers and aphids, the beak arises from the posterior part of the head, whereas in Hemiptera the position is anterior and the head is commonly directed forward or downward rather than backward.

The stylets are grooved and slide up and down during the act of feeding, the mandibles being toothed to grip the plant or other tissue that is being penetrated. Two tubes are formed by the stylets, one for sucking up the fluid food and the other for injecting salivary or anticoagulant fluids. The labium is jointed with three or four segments.

Hemiptera are further characterized (**Fig. 4**) by antennae, usually of four or five segments, a pair of compound eyes, and often two ocelli. The thorax consists of a prominent pronotum, a triangular mesothoracic scutellum, and a broad metathorax which is partly fused with the first abdominal segment. The mesothoracic wings, or hemelytra, overlap at their membranous apices when at rest, leaving the scutellum exposed. They consist of a leathery corium at the base, an inner clavus along the scutellum, and an apical membrane. A marginal fracture separates the apex of the corium into a cuneus in some groups. The venation is seldom clear but has been interpreted from the tracheae of nymphal wing pads as consisting of the subcosta at the anterior edge of the wing, the radius and media more or less fused behind, the cubitus as the posterior vein of the corium, and one or more anal veins in the clavus. The hindwings are hitched to the forewings in flight by grooves and pegs. In metathoracic wings, the subcosta, radius, and media are variously fused but usually form an elongate cell with or without a backward-directed spur vein or hamus. The anal and jugal areas are more or less developed with ill-defined veins and folds.

Wings are sometimes reduced to short pads and may be lacking in certain groups or even in members of a single species. Wing polymorphism, especially in water striders, appears to be determined by environmental factors as well as by genetic factors.

The front legs are frequently enlarged and sometimes chelate in predacious forms, and the middle and hindlegs are adapted for swimming in some groups. The foretibiae have a small plate or comb at the inner apex and sometimes a spongy pad for clinging to smooth surfaces or holding prey. Tarsi usually are of three segments, but some groups have only two segments and front tarsi are lacking in a few specialized types. Two claws are commonly present with or without leaflike or bristlelike arolia between.

The abdomen is of 10–11 segments, but only seven are commonly seen, the first two being fused and the

TABLE 2. Habitats of Hemiptera

Habitat	Families
Terrestrial	
On surface beneath stones and wood	Nabidae, Reduviidae, Lygaeidae, Enicocephalidae, Leptopodidae
In rotting plant material	Schizopteridae, Lygaeidae
Beneath surface of ground	Vianaidae
Aquatic	
True aquatics	Corixidae, Nepidae, Belostomatidae, Notonectidae, Pleidae, Helotrephidae, Naucoridae
Surface bugs	Gerridae, Veliidae, Hydrometridae, Mesoveliidae
Shore bugs	Gelastocoridae, Ochteridae, Hebridae, Mesoveliidae, Saldidae, Leptopodidae, Dipsocoridae
Marine intertidal	Saldidae
Other associations	
In bird nests	Anthocoridae, Cimicidae, Reduviidae, Lygaeidae
Ectoparasites	Polyctenidae
In termite colonies	Termitaphididae, Aradidae, Reduviidae
In spider webs	Plokiophilidae, Nabidae, Reduviidae
In fungi	Aradidae, Plataspidae
On seeds	Lygaeidae
On stems, foliage, flowers	Anthocoridae, Miridae, Nabidae, Reduviidae, Tingidae, Lygaeidae, Bertyidae, Piesmidae, Pyrrhocoridae, Coreidae, Pentatomidae, Cydnidae, Urostylidae, Phloeidae, Plataspidae

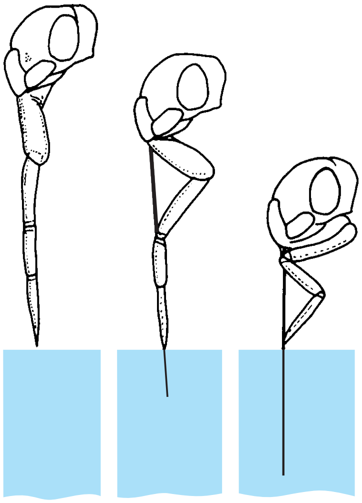


Fig. 3. Mouthparts and method of feeding of a plant bug.

last being partly concealed by the genitalia or withdrawn into the abdomen. Reproductive organs are shown for both sexes in copulation (**Fig. 5**). Males have the ninth segment variously developed as a genital capsule. A pair of parameres is commonly associated with the genital capsule, and the aedeagus, or penis, arises from its floor. The aedeagus is hinged to basal plates. An ejaculatory duct traverses the entire structure, terminating in the gonopore. Clasping organs may be present on either side of the thickened phallosoma, and various lobes and processes are developed on the often membranous endosoma. The latter is evaginated during mating, with the vesica variously inflated and inserted in the female genitalia. The female abdomen is variously modified but commonly has an ovipositor, with bent valves, arising from the eighth and ninth segments. A spermatheca is present in many groups and is distinctive in shape and form in some. Spiracles occur laterally, either in a ventral position or dorsally and most often in or near lateral connexival plates. Long sensory hairs, or trichobothria, occur on the ventrites of stink bugs, squash bugs, chinch bugs, and some others.

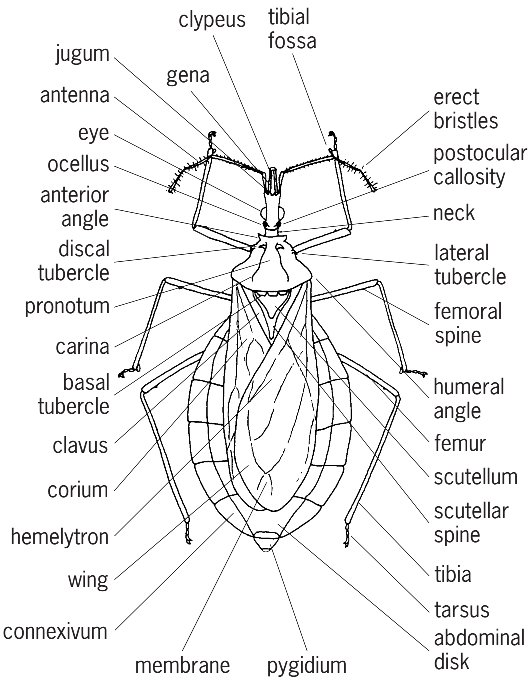


Fig. 4. External anatomy of a bug.

Internal anatomy. Internally the Hemiptera resemble other orders of insects, but several special features are noteworthy. Parts of the midgut have fluted ceca in stink bugs, squash bugs, and some others. These contain symbiotic bacteria which are specific and are transmitted from one generation to the next, in or on the ova. In stink bugs the eggs are smeared at the time of laying, and the first-stage nymphs, when hatched, cluster around the egg shells and suck up bacterial material. Cimicidae have no gastric ceca but possess a pair of mycetomes in about the third abdominal segment. The function of symbiotic bacteria is not clear, but in the Reduviidae it has been shown that vitamins of the B group are produced.

Salivary glands are of two parts, the bilobed principal glands and the tubular accessory glands. They empty into a common duct and thence to a pump which forces the fluid through the stylet tube. The Malpighian tubules are usually four in number.

Respiration is accomplished by spiracles, of which there are 10 pairs or less. In terrestrial forms a closing apparatus is developed to conserve moisture, whereas in water bugs the spiracles are open. Breathing tubes are present in Nepidae, enabling these water scorpions to use surface air while submerged. *Buenoa* and related backswimmers have hemoglobin in clusters of cells adjacent to the spiracles.

Endocrine glands include the corpus cardiacum and corpus allatum located behind the brain. These control molting and metamorphosis. The presence of hormone from the corpus allatum (juvenile, or inhibitory, hormone) prevents metamorphosis in young nymphs. Later, after four or five nymphal instars, the inhibitory hormone is no longer present and the growth and differentiation hormone acts to produce metamorphosis. *See* ENDOCRINE SYSTEM (INVERTEBRATE).

Scent glands are a characteristic feature of bugs, causing the well-known "buggy odor." Most adult bugs have a pair of metathoracic glands with openings on the metapleura, as in Pentatomidae and others, or on the hind coxae, as in Reduviidae.

A few have additional glands in the first abdominal segment. Nymphs of all but a few, such as some Reduviidae and the Hydrometridae, have dorsal abdominal scent glands. The arrangement varies from a full complement of three, opening at hind margins of third, fourth, and fifth tergites; two, on the third and fourth or the fourth and fifth tergites; one, on the fourth tergite; to none. The openings may be single (Miridae) or double and widely separated (Naucoridae). In adults of the water striders (Gerridae), there is a single opening (omphalium) at the middle of the metasternum.

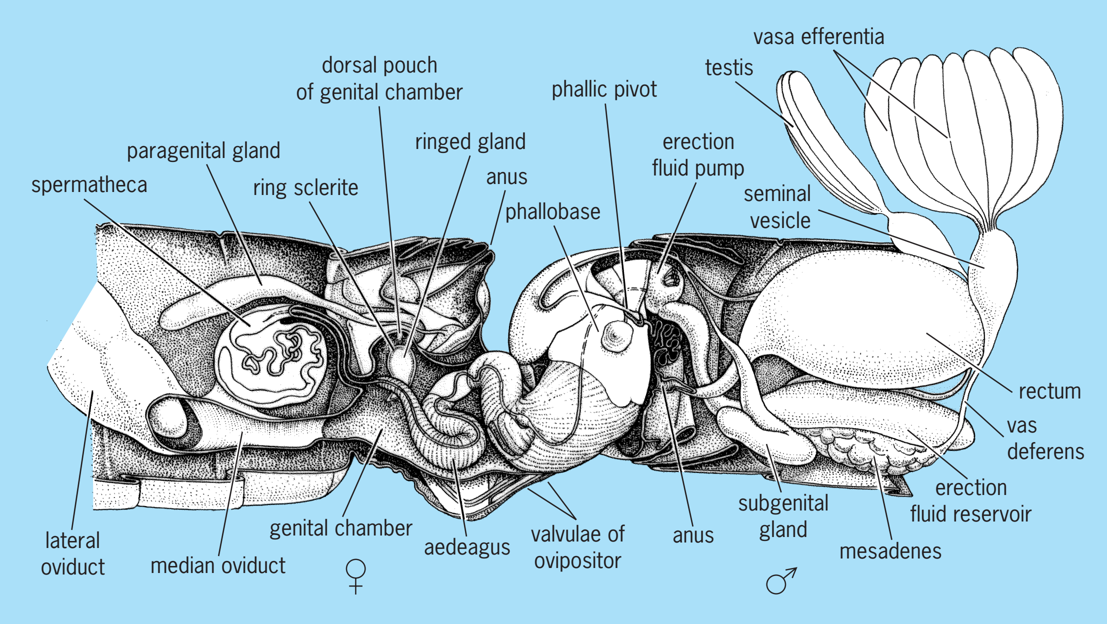


Fig. 5. Reproductive organs of female and male *Oncopeltus fasciatus* in copulation.

Internal reproductive organs include, in females, a pair of acrotrophic ovaries with ovarioles numbering 2–8, with 7 as the commonest number. A spermatheca is well developed and of characteristic shape for each of the main groups of stink bugs, squash bugs, and others, but is absent in plant bugs (Miridae) and their allies. In males, the paired testes comprise 1–7 follicles. The seminal vesicles are usually large and are formed by a glandular reservoir. Accessory glands known as mesadenia empty into the reservoir, and ectadenia join the ejaculatory bulb in some groups.

Chromosomes are best seen in late instar nymphs or during early adult life in males. The 2*n* (diploid) number may be as low as 6, in *Rhytidolomia*, or as high as 43, in *Ranatra*. Commonly, there are 12 autosomes and 2 X chromosomes or an X and a Y. In *Cimex* and some others there are supernumerary X chromosomes. *See* GENETICS.

Development and metamorphosis. Hemiptera undergo simple or incomplete metamorphosis including egg, nymph, and adult stages (**Fig. 6**). Eggs are characterized by a chorion of several layers, usually with hexagonal reticulations on the surface. A lid or cap is present in the plant bugs, bedbugs, and allies, but the anterior end of the egg is marked only by micropylar processes in other groups. An egg burster is well developed and T-shaped in Pentatomidae and less developed but visible in some other groups. Eggs are glued to surfaces by *Lethocerus*, inserted in plant tissues by Miridae, or laid free by some *Triatoma*. In *Belostoma* and *Abedus*, the eggs are laid on the backs of males, presumably for protection, and a corixid (*Ramphocorixa*) regularly lays its eggs on crayfish in pasture ponds. The eggs may be stalked, as in Corixidae; spindle-shaped, as in Hydrometridae; or oval- or barrel-shaped, as in Pentatomidae, and may be laid in clusters, as by *Zelus*, or singly. The female "broods" or "guards" the eggs in *Nerthra* and some Scutelleridae.

Embryonic development is endoblastic and is not very different from that of other insects. Glandular organs, the pleuropodia, occur on the first abdominal segment and produce the embryonic cuticle. In Polyctenidae the pleuropodia function as a pseudoplacental organ. The incubation period varies from a few days to several months, and development is arrested in species that overwinter in the egg stage, as do some Miridae. As development proceeds, red eyespots become visible through the chorion. The eggs of many Miridae are embedded in woody tissue and swell because of growth and absorption of moisture, forcing out a yolk plug (*Notostira*).

Eclosion, or hatching, occurs either by splitting or tearing the chorion or by raising the lid. A postembryonic molt occurs during eclosion, and commonly the exuviae are attached to the egg shell.

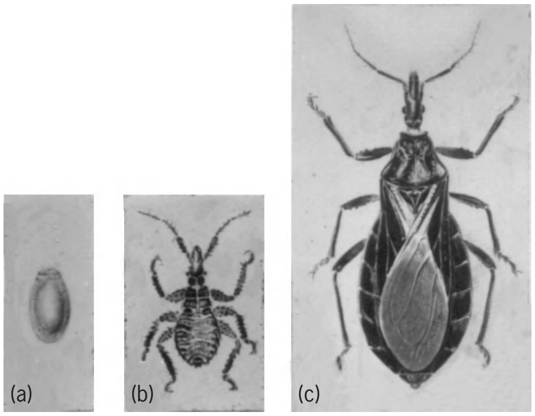


Fig. 6. *Triatoma protracta*. (*a*) Egg. (*b*) Nymphal instar. (*c*) Adult.

First instar nymphs are active or, in Pentatomidae, sedentary and gregarious. They take little or no food, the yolk remaining from the egg stage being sufficient, in most cases, to carry them through to the second instar. Bacteria are taken up from crypts amid the egg shells or from the egg surface in Pentatomidae, thus infecting the gastric ceca.

Wing pads begin to show on the hind margins of meso- and metanota at the third stage, the third nymphal instar. Nymphs, also called larvae or neanides, commonly undergo five instars, but rare cases have been reported with four in some Veliidae and Cimicidae. One case, *Dindymus sanguineus*, of the Pyrrhocoridae, has been reported with nine instars. The final molt and metamorphosis to the adult stage results in an increase in tarsal segmentation, from one to two or from two to three, an increase in number of antennal segments in some cases (four to five in Pentatomidae), and development of the wings and external genitalia. Also, the ocelli are fully developed at this time, and numerous internal changes occur.

Subdivisions

Important families of the Hemiptera and their distinguishing characteristics are treated under the respective subdivisions. Many of the economically important species are given in **Table 3**.

Hydrocorisae. This subdivision contains nine families of water bugs with concealed antennae and without a bulbus ejaculatorius in the male. Many species are predaceous.

Corixoidea. This is a superfamily which contains the single family Corixidae, or water boatmen. Corixidae lack ocelli, have three dorsal abdominal scent-gland openings as nymphs, and have a unique type of mouthpart. The rostrum is broad, with a pseudosegmentation, and the food consists of algae, as well as small animal forms gathered by the comblike front tarsi, or palae. Respiration is through an air bubble obtained at intervals by touching the surface with the pronotum. Corixids swim with the dorsal side uppermost, using the oarlike middle and hind legs. Eggs are stalked and there are five nymphal instars. Adults sometimes fly to lights in great numbers. In Mexico, the eggs are harvested from lakes to make a bread (ahuatle) and the adults (mosco) are dried and sold in the United States as food for pet birds and turtles.

Nepoidea. The Nepidae, or water scorpions, have a long breathing tube at the tip of the abdomen, through which they obtain air directly from the surface. The front legs are chelate, and the beak is short and stout to suck the juices of other insects on which they prey. Two common types of water scorpion are the long, slender *Ranatra*, of worldwide distribution, and the broad *Nepa*, of the eastern United States and Europe. The eggs are embedded in plant tissue and have two flaplike appendages.

Belostomatidae, or electric-light bugs, are related to the Nepidae but have short, straplike respiratory appendages at the tip of the abdomen. Giant water bugs of the genus *Lethocerus*, reaching 4 in. (10 cm) in length, fly to lights at night. They are pests in fish ponds where they attack fry. They can inflict a painful bite when handled carelessly by humans. In China, the adults, called kwai fa shim, are boiled in oil and sold in markets as food for humans. Eggs are laid on cattails and other plants at the water's edge. Control in fish ponds is by draining or by rearing fry and fingerlings in screened pools. Smaller, 1–2-in. (2.5–5-cm) belostomatids (*Belostoma*, *Abedus*) lay eggs on the backs of the males, possibly for protection.

Notonectoidea. The Notonectidae or backswimmers, as the name implies, swim ventral side uppermost, the hindlegs serving as oars. Breathing is facilitated by an air bubble, obtained by touching the tip of the abdomen to the surface. *Notonecta* is worldwide in distribution. It is stout in body form and swims with jerky movements or clings to underwater supports. *Buenoa* in the Western Hemisphere and *Anisops* in the Old World have hemoglobin around the spiracles and maintain hydrostatic balance in the water, thus being truly limnetic forms.

Pleoidea. These minute bugs of the families Pleidae and Helotrephidae are suboval, with legs not fitted for rowing. Unlike Notonectidae, the nymphs have a single scent gland opening on the abdominal tergum.

Naucoroidea. The creeping water bugs are sometimes separated into the Naucoridae and Aphelocheiridae. They are suboval in body form, with chelate front legs and with two nymphal scent gland openings widely separated at the hind margin of the third abdominal tergite. Respiration is either by an air bubble or by means of a plastron. In the Old World genus *Aphelocheirus*, the plastron consists of an ultramicroscopic hair pile (2,000,000 hairs per square millimeter) which acts as a physical gill; oxygen diffuses through it from the water to the spiracles of the bug. *Aphelocheirus* is thus able to remain submerged permanently, inhabiting the bottoms of rivers with swift-flowing, well-oxygenated water.

Gelastocoroidea and Ochteroidea. The Gelastocoridae, or toad bugs, and the Ochteridae (or Pelogonidae) are shore-line or mud inhabitants. Both have ocelli. The former are cryptically colored, resembling the sand or mud background. Their eyes are large and their antennae concealed. Ochterids are black with a silky sheen and the antennae, unlike those of other Cryptocerata, are visible from above, though short and partly concealed beneath the eyes. Eggs are laid in sand and mud.

Amphibicorisae. This subdivision contains surface water bugs with antennae exposed and without a bulbus ejactulatorius in the male. Spermatheca have a fecundation canal.

Only the single, diverse superfamily Gerroidea has been proposed for the surface water bugs. All have conspicuous antennae, and the body is clothed with hydrofuge hairs. Three pairs of sensory hairs, the trichobothria, are inserted between the eyes. All Gerroidea are predacious.

Gerridae are the large water striders with long middle and hind legs and a median scent gland opening, the omphalium, on the metasternum. The claws are

TABLE 3. Economically important Hemiptera

Name	Damage
Cimicidae	
Human bedbug (*Cimex lectularius*)	Sucks blood in human dwellings; eggs with small caps, glued in cracks
Tropical bedbug (*Cimex hemipterus*)	Sucks blood in human dwellings; lives in cracks and sleeping mats
Swallow bug (*Oeciacus vicarius*)	In mud nests of swallows; reported on humans in houses with swallows
Mexican chicken bug (*Haematosiphon inodora*)	Sucks blood in chicken roosts; eggs laid in cracks
Reduviidae	
Bloodsucking conenose bugs (*Triatoma* sp., *Rhodnius* sp., *Panstrongylus* sp.)	Several species live in nests of wood rats (*Neotoma*) and suck blood; they transmit Chagas' disease in Latin America
Miridae	
Tarnished plant bug (*Lygus lineolaris*)	Eggs inserted in plant tissue; overwinters as adult; feeding deforms leaves and stems; punctures cause cat-facing on fruits
Lygus bugs (*Lygus hesperus*)	Several species stunt alfalfa, also attack beans, cotton, and other plants; they overwinter in adult stage and lay eggs in plant tissue
Apple red bug (*Lygidea mendax*)	Feeding punctures cause pitting of fruits; winters in egg stage, with eggs inserted in bark
Ash plant bug (*Neoborus amoenus*)	Feeding stunts new growth in spring; eggs laid in stems and bark scars in summer–overwinter in egg stage
Suckfly (*Cyrtopeltis notatus*)	Feeding and egg laying in ring around stems of tomato and tobacco; causes tomato fruits to drop
Cotton flea hopper (*Psallus seriatus*)	Sucks cotton squares causing shedding and whiplike growth of plant; overwinters in egg stage
Garden flea hopper (*Halticus bracteatus*)	Feeding punctures spot leaves of many garden plants; overwinters as adult
Tingidae	
Azalea lace bug (*Stephanitis pyrioides*)	Introduced from Japan; overwinters in egg stage; three broods a year in New Jersey
Rhododendron lace bug (*Stephanitis rhododendri*)	Overwinters in egg stage; two broods in New Jersey; feeding punctures and fecal spots on undersides of leaves
Basswood lace bug (*Gargaphia tiliae*)	Overwinters in adult stage; eggs laid in spring, partially inserted on undersides of leaves
Eggplant lace bug (*Gargaphia solani*)	Hibernates in adult stage; leaves spotted and yellowed; up to six generations a year in Virginia
Chrysanthemum lace bug (*Corythucha marmorata*)	Overwinters as adult; eggs inserted deep in veins under leaves; feeding punctures and fecal spots
Sycamore lace bug (*Corythucha ciliata*)	Adults hibernate under bark; feeding punctures and fecal spots on under surface of leaves
Lantana lace bug (*Teleonemia scrupulosa*)	Feeds on leaves of lantana, causing leafdrop; introduced into Hawaii from Mexico to control lantana
Ash lace bug (*Leptotypha minor*)	Overwinters in adult stage; eggs, feeding punctures, fecal spots, and cast skins on undersides of leaves
Lygaeidae	
Chinch bug (*Blissus leucopterus*)	Attacks corn and small grains in the Middle Western states, causing wilting and drying of plants; overwinters in adult stage with 2±3 generations a year
Hairy chinch bug (*Blissus hirtus*)	Attacks lawns in eastern U.S., killing grass in spots; life cycle similar to *B. leucopterus*
False chinch bug (*Nysius ericae*)	Several species of the genus *Nysius* breed on weeds and migrate in large numbers to cultivated crops, causing wilting
Coreidae	
Boxelder bug (*Leptocoris trivitatus*)	Overwinters as adult; attacks seeds of boxelder; two generations a year
Squash bug (*Anasa tristis*)	Attacks most cucurbits, wilting leaves or killing small plants; overwinters as adult; the small brown eggs are laid in groups on the plants in the spring and early summer
Pyrrhocoridae	
Cotton stainers (*Dysdercus suturellus and Dysdercus minulus*)	In Florida and Arizona, puncture bolls and seeds and stain the lint
Pentatomidae	
Brown stink bug (*Euschistus servus*)	Breeds on various weeds and attacks most cultivated crops; feeding stains cotton, pits or causes cat-facing of pears, peaches, tomatoes, and other plants; overwinters in the adult stage
Dusky stink bug (*Euschistus tristigmus*)	
One-spot stink bug (*Euschistus variolarius*)	
Conchuela (*Chlorochroa ligata*)	
Say stink bug (*Chlorochroa sayi*)	
Green stink bug (*Acrosternum hilare*)	
Southern green stink bug (*Nezara viridula*)	
Harlequin bug (*Murgantia histrionica*)	Feeds on foliage, causing wilt and death of the plants; attacks cabbage and many other garden plant; breeding is more or less continuous, with adults and nymphs hibernating in cold weather

inserted before the tips of the tarsi. Most species of *Gerris* are pond or slow-moving-stream inhabitants, but *Metrobates* lives on large rivers and *Halobates* inhabits lagoons in tropical seas and even the open ocean. The marine forms are always wingless but all others are polymorphic, with fully winged forms occurring together with short-winged or apterous types. *Halobates* feeds on small anemones and other floating organisms and lays its eggs on feathers, pumice, and presumably anything else that floats. One species, *H. micans*, occurs in all tropical seas but the remaining 40 or more species are Indo-Pacific.

Veliidae are small water striders which have shorter legs and a longitudinal groove between the eyes. The claws are preapical. Like the Gerridae, these are pond inhabitants (*Microvelia*), stream-riffle bugs (*Rhagovelia*), and marine types (*Halovelia*), but the last are found only near shore in tropical reefs.

Hydrometridae are long, slender marsh treaders in which the head is longer than the thorax. The claws are apical. These have been reported as predacious on *Anopheles* mosquito larvae, and attempts have been made to transport them for purposes of biological control.

Mesoveliidae and Hebridae are two small families which differ from others in having the well-developed ocelli and the single dorsal abdominal scent gland openings of the nymphs. *Mesovelia mulsanti* is a swift water strider, but some other species are shoreline inhabitants, as are the Hebridae or velvet water bugs.

Geocorisae. This subdivision contains the land bugs with conspicuous antennae and an ejaculatory bulb in the male reproductive system. This subdivision can be divided into seven groups, each of which may be equivalent in rank to the Hydrocorisae and Amphibicorisae. The two largest groups are the Cimicimorpha and the Pentatomorpha, with the included families having certain common characteristics. The Cimicimorpha includes the Cimicoidea and Tingoidea, which have distinctly operculate eggs and are without a median spermatheca; the accessory salivary glands are vesicular, the abdomen is without trichobothria, and claws are without discrete lateral arolia. The Thaumastocoroidea also belong here. The Pentatomorpha contains the Pentatomoidea, Coreoidea, Pyrrhocoroidea, and Lygaeoidea. In these superfamilies, the eggs are nonoperculate, a median spermatheca is present, accessory salivary glands are tubular, the abdomen has trichobothria, and the claws have well-developed lateral arolia. Superfamilies that do not fit the above groupings are the Reduvioidea, formerly included in the Cimicimorpha; the Saldoidea, with three pairs of head trichobothria like Amphibicorisae; the Arodoidea, which are like the Pentatomorpha but without trichobothria; and the Enicocephaloidea and Dipsocoroidea, of isolated position in the system.

Saldoidea. The superfamily Saldoidea, of the group Leptopodoidea, comprise the shore bugs which have three pairs of trichobothria on the vertex and a single nymphal scent gland opening. Saldids lay their eggs amid moss or stones at the edges of streams and ponds. One genus, *Chiloxanthus*, lives on the tundra and overwinters for 9 months with hundreds of feet of permafrost beneath the winter ice above. An intertidal genus, *Aepophilus*, occurs along the coasts of Britain and France. Leptopodidae differ in the spiny body and appendages. They live under drier conditions, sometimes remote from the water. Leotichidae is a rare group of doubtful affinities; specimens have been found in caves.

Dipsocoroidea. This is a group of minute ground inhabitants, of which the Schizopteridae live in leaf mold and the Dipsocoridae are predators on small insects under bark or in rotten wood, such as *Ceratocombus*, or amid stones at the edge of streams, such as *Cryptostemma*.

Cimicimorpha. This is the largest group of the Geocorisae and is divided into three superfamilies, the Cimicoidea, Tingoidea, and Thaumastocoroidea.

1. Cimicoidea. Anthocoridae, or the flower bugs, are small predators on thrips, mites, and similar species in vegetation (*Orius*), in stored products (*Lyctocoris*), and in the nests of birds. The ocelli are distinct, and the front wings have a marginal fracture, the cuneus. Fertilization of the female is hemocoelic through special organs. The male right genital clasper is large and bladelike; the left is lacking.

Cimicidae contains the bedbugs, which probably were derived from the Anthocoridae but differ in habits, in the absence of ocelli, and in having the short wings reduced to pads. Fertilization is usually through an organ of Ribaga on the second, third, or fourth ventrites or the third, fourth, or fifth tergites. The food consists of blood of birds and mammals. The common bedbugs of humans are *Cimex lectularius*, in temperate regions and large cities, and *Cimex hemipterus*, in the tropics. *Oeciacus* lives on swallows in Europe and North America; *Paracimex*, on cave swiftlets in eastern Asia; *Ornithocoris* and *Haematosiphon*, on chickens in the Western Hemisphere; and *Cimex columbarius*, on pigeons in Europe. Most Old World species, including those of the genus *Cimex*, live on bats, and a few bat parasites are found in the New World. Bedbugs breed continuously in houses and thus are able to complete several generations in a year. Bedbugs and chicken bugs are controlled by residual sprays of 5% DDT or by fumigation. Resistance has been reported to standard dosages of DDT.

The Polyctenidae are bat ectoparasites which resemble the Cimicidae but lack eyes and have ctenidia and strong claws. Reproduction is by hemocoelic fertilization, and the embryo is nourished by a pseudoplacental organ, resulting in true viviparity.

The Miridae family contains a majority of the species of Hemiptera. Included are plant bugs of both herbivorous and predacious type. Ocelli are lacking, a cuneus is present, the antennae are four-segmented, and there is only one dorsal abdominal scent gland opening in nymphs. The eggs are embedded in plant tissue. Of the predators, *Cyrtorhinus mundulus* was introduced into the Hawaiian

Islands and is credited with controlling the sugarcane leafhopper by sucking its eggs. Plant pests include the tarnished plant bug, *Lygus lineolaris*, and other Lygus bugs on alfalfa, cotton, and beans. Control may be obtained with 5% DDT dust applied not later than 30 days before harvest. Lygus bugs overwinter in the adult stage and may undergo two or more generations in a year. Most plant bugs, however, overwinter in the egg stage and are univoltine.

Plant bugs and also some stink bugs and others leave their preferred weed hosts as the growing season advances to suck developing fruit such as apples and pears, thus causing deformity or "cat-facing." Control is by clean cultivation or by DDT dust or spray.

Two small families, the Plokiophilidae and Microphysidae, are close relatives of Miridae. Both are predacious. The Plokiophilids live in the webs of spiders and embiids.

Nabidae are the long "damsel" bugs which are slender predators on other insects. The ocelli are well developed, and the rostrum is four-segmented. The eggs are embedded in plant tissue. Nymphs have three abdominal scent gland openings. *Nabis ferus* is a cosmopolitan predator in grasses and other vegetation. The tropical Velocipedidae are related but differ in the broader form and darker coloration.

2. Tingoidea. This superfamily of the group Cimicimorpha contains the lace bugs of the family Tingidae which have wings with many lacelike areolae. They lack ocelli, have four segments in the antennae and beak, and commonly have one or more bulbous or hoodlike elevations on the thorax. Pests are known on ornamental plants such as Christmas berry (*Corythucha*), ash (*Leptoypha*), and other ornamental plants, and some species attack agricultural crops, such as *Gargaphia solani* on eggplant. The latter is controlled with Malathion as a 5% dust. Lace bug injury is characteristic. The leaves of infested plants have white areas due to feeding, and the undersurfaces of leaves have black fecal spots and usually cast skins adhering to the leaf hairs. Lace bugs commonly overwinter as adults and pass through more than one generation each year. *Teleonemia scrupulosa* was introduced into the Hawaiian Islands from Mexico in an effort to control the lantana plant. The rare South American Vianaidae are also included in this group.

3. Thaumastocoroidea. The only family, the Thaumastocoridae, occurs in Australia and the New World tropics; it includes the royal palm bug (*Xylastodorus*) of Florida.

Enicocephaloidea. This is a unique group in which the head is bilobed, the pronotum trilobed, and the wings completely membranous. There is no prosternal stridulatory groove, the eggs are without an operculum, and the males swarm like gnats. The life cycle is not fully known, but they live under stones, beneath bark, and in leaf mold and are predators. Some species are wingless; others can cast off their wings.

Reduvioidea. This group includes only the assassin bugs or conenose bugs of the family Reduviidae. Nearly all have a stridulatory furrow on the prosternum, which is scraped by the rostrum to produce a squeaking sound. Ocelli are generally present, the beak is three-segmented, and lateral scent gland openings usually lie beneath the hind coxae rather than on the metapleura. The corsair (*Rasahus*) is black with a spot on the membrane of the wings. It lives beneath stones and feeds on insects but flies to light, where it encounters and often bites man. Kissing bugs, the Triatominae, have a longer, more slender rostrum and lack scent glands on the nymphal abdominal tergites. They are exclusively bloodsuckers, living mostly in the nests of wood rats of the genus *Neotoma* in the United States. Several species regularly feed on humans in the American tropics: *Panstrongylus megistus*, *Rhodnius prolixus*, *Triatoma infestans*, and others. Up to 30% of the bugs are commonly infected with *Trypanosoma cruzi* in some places. The trypanosome which causes the disease American trypanosomiasis, or Chagas' disease, is picked up by the bug while sucking blood but is transmitted through the feces. Fecal contamination may be at the place of the bite or in the conjunctiva of the eye—the bites are often about the face (hence the name kissing bug) and the feces are rubbed into the eye. The trypanosomes affect muscles of the heart and other organs. Kissing bugs may be controlled with residual sprays, such as DDT, by eliminating nesting rodents, or by screening bedrooms. *Rhodnius prolixus* is maintained as a laboratory animal for studies of insect physiology. Other reduviids include the thread-legged Emesinae that live in spider webs and elsewhere, the large wheel bug (*Arilus*) of the southern United States, and many other predators of injurious insects. Phymatines with sometimes concealed antennae lie in wait in flowers and attack bees and other insects.

Aradoidea. Flat bugs of the family Aradidae and their specialized relatives, Termitaphididae, have the mandibular and maxillary setae coiled in the clypeus. They lack ocelli and have four-segmented antennae. Most, if not all, are fungus feeders. The Termitaphididae have no wings and the margins of the body have numerous small, setigerous lobes. They are known from termite nests in the Old and New World tropics. Aradidae are nearly cosmopolitan and fully winged, brachypterous, stenopterous, or apterous. Some species are polymorphic, the males having fully developed or stenopterous wings and females being short-winged.

Pentatomorpha. This group includes the superfamilies Lygaeoidea, Pyrrhocoroidea, Coreoidea, and Pentatomoidea.

1. Lygaeoidea. This is the first of the superfamilies with ventral trichobothria on the abdomen. These erect bristles are in groups of three on either side near the middle of the third and fourth segments and lateral on the fifth, sixth, and seventh segments. The antennae are four-segmented, as is the beak. Ocelli are present. There are only four simple veins in the membrane. The Lygaeidae include the chinch bug, *Blissus leucopterus*, the false chinch bugs, *Nysius* spp., and the milkweed bugs,

Oncopeltus. The last is reared as a laboratory animal for experimental purposes, using milkweed seeds as food. The chinch bug, which is a pest of corn and wheat in the Middle West, has two generations per year. Adults overwinter in grass and trash. Control is by spraying with Dieldrin or Toxaphene, applying to bases of plants where the bugs congregate. Barrier strips of Dieldrin may also be used in strips 4 rods wide.

1. Small families related to the lygaeids are the thread-legged Neididae (Berytidae), the small Piesmidae, formerly included in Tingidae, and the tropical Colobathristidae. Some of the last have a unique method of stridulation with a filelike arch on the sides of the head.

2. Pyrrhocoroidea. This group includes the cotton stainers (*Dysdercus*), which attack cotton bolls in the southwestern United States and over most of the tropics, and stout, dark bugs with a reddish border (*Largus*). Pyrrhocorids resemble lygaeids and coreids but lack ocelli. *Dysdercus* may be controlled by a 5% DDT spray.

3. Coreoidea. The squash bugs and their relatives include the Coreidae, *Anasa tristis*; Rhopalidae, *Rhopalus* and *Leptocoris*; Alydidae, *Leptocorisa*; and the Hyocephalidae. They have four-segmented antennae, a beak, distinct ocelli, and many veins in the membrane. Coreids resemble Lygaeids in arrangement of the trichobothria. The squash bug overwinters as an adult. Eggs are laid on squashes and other cucurbits. There are five nymphal instars and usually only a single drawn-out generation per year. Control of these pests is either by DDT 5% dust applied to bugs on soil or under trap boards or by nicotine sulfate in 40% solution.

4. Pentatomoidea. This large group has marginal trichobothria. The antennae are usually five-segmented and the beak is four-segmented.

Scutelleridae, or shield bugs, are not injurious in the United States, but *Eurygaster* is a pest of grains in Russia, and some metallic green species are plant pests in the tropics.

Cydnidae include the ground-burrowing bugs *Amnestus*), which attack strawberries in sandy soil, and the negro bugs (*Corimelaena*).

Smaller families of little or no economic importance in the United States are the large tropical Tessaratomidae, the Acanthosomatidae, the Dinidoridae, the Aphylidae, the Urostylidae, the barklike Phloeidae, the shining, oval Plataspidae, and the Lestoniidae.

The true stink bugs, Pentatomidae, include the black-and-red harlequin cabbage bug, *Murgantia histrionica* (Hahn), and several green stink bugs, *Thyanta*, *Euschistus*, and *Chlorochroa*, on cotton and other crops. Control is by DDT as a dust or spray on the foliage or, on cotton, BHC or Dieldrin dust. Stink bugs overwinter as adults; they may produce two generations in a season. Feeding punctures on pears, which are attacked as the cover crop dries, cause cat-facing. One group, the Asopinae, includes predators (*Podisus*) on caterpillars. *See* ENTOMOLOGY, ECONOMIC; INSECTA. Pedro W. Wygodzinsky

Hemispheric laterality

The human brain is a bilaterally symmetrical structure which is for the most part richly interconnected by two main bridges of neurons called the corpus callosum and anterior commissure. These structures can be surgically sectioned in humans in an effort to control the spread of epileptic seizures. Although there is no apparent change in everyday behavior of these patients, dramatic differences in cognitive function can be demonstrated under specialized testing conditions. Because of these studies it now can be said that in normal humans these cerebral commissures are largely responsible for behavioral unity; the neural mechanism keeps the left side of the body up to date with the activities of the right side, and vice versa.

Results of tests involving speech. Changes in behavioral responses of persons whose cerebral commissures have been sectioned are almost undetectable. The person walks, talks, and behaves in a normal fashion. Dramatic effects are observed only under testing conditions which utilize stimuli that are lateralized exclusively to one hemisphere or the other. When visual stimuli are flashed to the left of fixation of the eye, all information is exclusively projected to the right hemisphere, and vice versa. Likewise, when the left hand manipulates objects that are held out of sight, all relevant sensory information projects to the opposite, right hemisphere. Similarly, information from objects held in the right hand is relayed exclusively to the left hemisphere. This is the normal state and permits the separate input-output testing of each half-brain.

Using these sensory avenues, remarkable "deconnection" effects are seen in split-brain conditions. For example, if a picture of an apple is flashed in the right visual field, the person describes the object normally. However, if the same picture is flashed in the left visual field, in the early days of postoperative testing the person denies that the stimulus was presented at all. Gradually, after many test sessions the person may have the impression that something was flashed, but is unable to say what. In brief, this disparity of recognition in the two sides of the visual field occurs because the information is projected to the right hemisphere, which is incapable of speech. Because the right hemisphere is now disconnected from the left, information arriving in the right hemisphere cannot be communicated by means of speech.

Results of tests not involving speech. When tests are used which do not require a spoken response, numerous mental abilities are observed to be present in the "disconnected" right hemisphere. If, for example, instead of being asked for a verbal response, the person is required with the left hand to retrieve from a series of objects the one most closely related to the flashed visual stimulus, good performance is seen. Even though the person is unable to describe a picture of an orange flashed to the left field, when the left hand searches through a field of objects placed out of view, it correctly retrieves the orange. If asked what the object is, the person would be unable to identify it. Here again the left hemisphere controls speech but cannot solve the problem. The right hemisphere solves the problem but cannot elicit speech.

Right hemisphere. Through the use of this kind of test procedure, a variety of other mental abilities have been observed to go on in the separated right hemisphere. It can read simple nouns but not verbs. It can recognize some adjectives. It can recognize negative constructions but cannot differentiate constructions in the active or passive voice. It cannot pluralize nouns. In brief, it can manage some aspects of language only.

Left hemisphere. Despite its linguistic superiority, the left hemisphere does not excel the right in all tasks. Tests have demonstrated that in some specialized functions, such as arranging blocks to match a pictured design or drawing a cube in three dimensions, the right hemisphere is decidedly superior to the left.

Emotional reactions. In the area of emotional reactions there appears to be equal reactivity in the two hemispheres. Although a brain-bisected person cannot describe an emotional stimulus presented to the right hemisphere, an emotional response such as laughter is nonetheless elicited. In one instance, when a picture of a nude was flashed to the right hemisphere, the person said she did not see anything and then chuckled. When she was asked what the laughing was about, she said that it was “a funny machine.” Here the right hemisphere clearly had enough ability to elicit laughter but not speech. The left hemisphere simply observed the changed emotional state and said that something must be funny, but did not know what.

Localization of the speech processor. Experimental psychological research developed a procedure commonly called dichotic listening. It is a process where two auditory messages are played at the same time—one going to the left ear and one going to the right ear. Previous research showed that under these test conditions normal subjects tend to repeat only the message presented to the right ear and to suppress the message presented to the left. It turns out this phenomenon is exceedingly dramatic in the split-brain person. In one study, when two phonemes (one to each ear) were presented to the split-brain persons, they were totally unable to name phonemes presented to the left ear. This was true despite the fact that the left ear sends projections to the left hemisphere that are not affected by split-brain surgery. Indeed, when a phoneme or any auditory signal is presented to the left ear alone, the split-brain person experiences no difficulty whatsoever in describine auditory event in detail. The suppression comes about only when both ears are simultaneously stimulated, and reflects some kind of dynamic process involving the two cerebral hemispheres that is not well understood.

A more surprising aspect of this study, however, was that the right hemisphere using its own manual response system could not respond at all to phonemes, the basic unit of speech perception. This finding is unexpected in light of the above described language capacities of the right hemisphere. Can the right hemisphere understand simple spoken and written words without being able to understand one of the basic building-block units that make it up? Perhaps the manner in which the right hemisphere comprehends spoken language is drastically different than the way the left processes its information.

Specificity of cortical circuitry. Tests have been conducted in which not all of the subjects have had their entire cerebral commissure sectioned. This is because it is now believed total commissure section is not necessary to stop the interhemispheric spread of some kinds of seizure activity. In neuropsychological testing, these partially sectioned persons showed dramatic breakdown in interhemispheric transfer. When the posterior part of the callosum is sectioned, visual aspects of the syndrome appear. When it is spared and more interior regions are cut, however, tactile and auditory communications are blocked, but not visual ones. It also appears that no fundamental reorganization of the interhemispheric transfer system takes place, since years after surgery these same deficits are present and are not compensated for in any way.

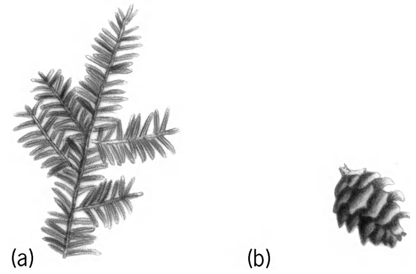
Individual variation and age of surgery. Another fact is that there would appear to be a large variation in the lateralized talents of each half-brain. Thus, while the right hemisphere frequently appears to have some language talent, it is clear that not all split-brain persons have language skills in the right hemisphere. Similarly, visual spatial skills, which are usually present exclusively in the right hemisphere, are frequently bilaterally represented and sometimes represented only in the left speech hemisphere. There is even some evidence that the commissure system itself varies in what is transferred where. In some cases an intact anterior commissure in those persons with a full callosal section will allow for the complete interhemispheric transfer of visual information. Yet, in other cases visual transfer is blocked with a section of only the posterior callosal area. One possible important factor is the age of the surgical section and the extent of early brain damage which was experienced by the person. *See* BRAIN; PSYCHOLOGY.

Michael S. Gazzaniga

Bibliography. M. S. Gazzaniga, *Nature's Mind*, 1994; J. B. Hellige, *Hemispheric Asymmetry: What's Right and What's Left*, 1993; M. Kosslyn, *The Wet Mind*, 1992, reprint 1995.

Hemlock

The genus *Tsuga* of the pine family, characterized by flattened needles with two white lines beneath the needlelike leaves, which have distinct short stalks. The cones are small and pendent. Eastern hemlock (*T. canadensis*) grows to a height of about 90 ft (27 m) and occurs in eastern Canada, the Great Lakes states, and the Appalachians. Minutely toothed leaves, some of the smaller ones growing upside



Eastern hemlock (*Tsuga canadensis*). (*a*) Branch. (*b*) Cone.

down, are characteristic of this species (see **illus.**). The wood is hard and strong, and is used for construction, boxes, crates, and paper pulp. The bark is one of the principal domestic sources of tannin. The eastern hemlock is a common ornamental tree. *See* PINALES.

Carolina hemlock (*T. caroliniana*), a species found in the southern Appalachians, has entire needles and is sometimes grown as an ornamental.

The western hemlock (*T. heterophylla*), which attains a height of 200 ft (60 m) or more, grows in the extreme Northwest and in Alaska. Its needles resemble those of the eastern hemlock, but the white lines beneath them are not so distinct. The greater part of the annual lumber cut comes from Washington. Much of this lumber cut goes into pulpwood for which it is a desired species. It is an important lumber tree, with uses similar to those of the eastern species. Another western species, *T. mertensiana*, is of lesser importance commercially. *See* FOREST AND FORESTRY; TREE. Arthur H. Graves; Kenneth P. Davis

Hemoglobin

The oxygen-carrying molecule of the red blood cells of vertebrates. This protein represents more than 95% of the solid constituents of the red cell. It is responsible for the transport of oxygen from the lungs to the other tissues of the body and participates in the transport of carbon dioxide in the reverse direction. More is known about the chemical structure and function of hemoglobin than any other protein.

Structure. Each molecule of hemoglobin comprises four smaller subunits, called polypeptide chains. These are the protein or globin parts of hemoglobin. A heme group, which is an iron-protoporphyrin complex, is associated with each polypeptide subunit and is responsible for the reversible binding of one molecule of oxygen. There are two different kinds of polypeptide subunits in the hemoglobin of the normal adult human (abbreviated Hb A). One is called the alpha chain and contains 141 amino acid residues. The other is called the beta chain and contains 146 amino acid residues. Two of each kind of polypeptide chain are arranged in the form of a truncated tetrahedron which has the overall shape of an ellipsoid of dimensions 5.5 × 5.5 × 7.0 nanometers (**Fig. 1**). The molecular formula of Hb A is $\alpha_2\beta_2$, and its molecular weight is 66,000. *See* AMINO ACIDS; PEPTIDE; PROTEIN.

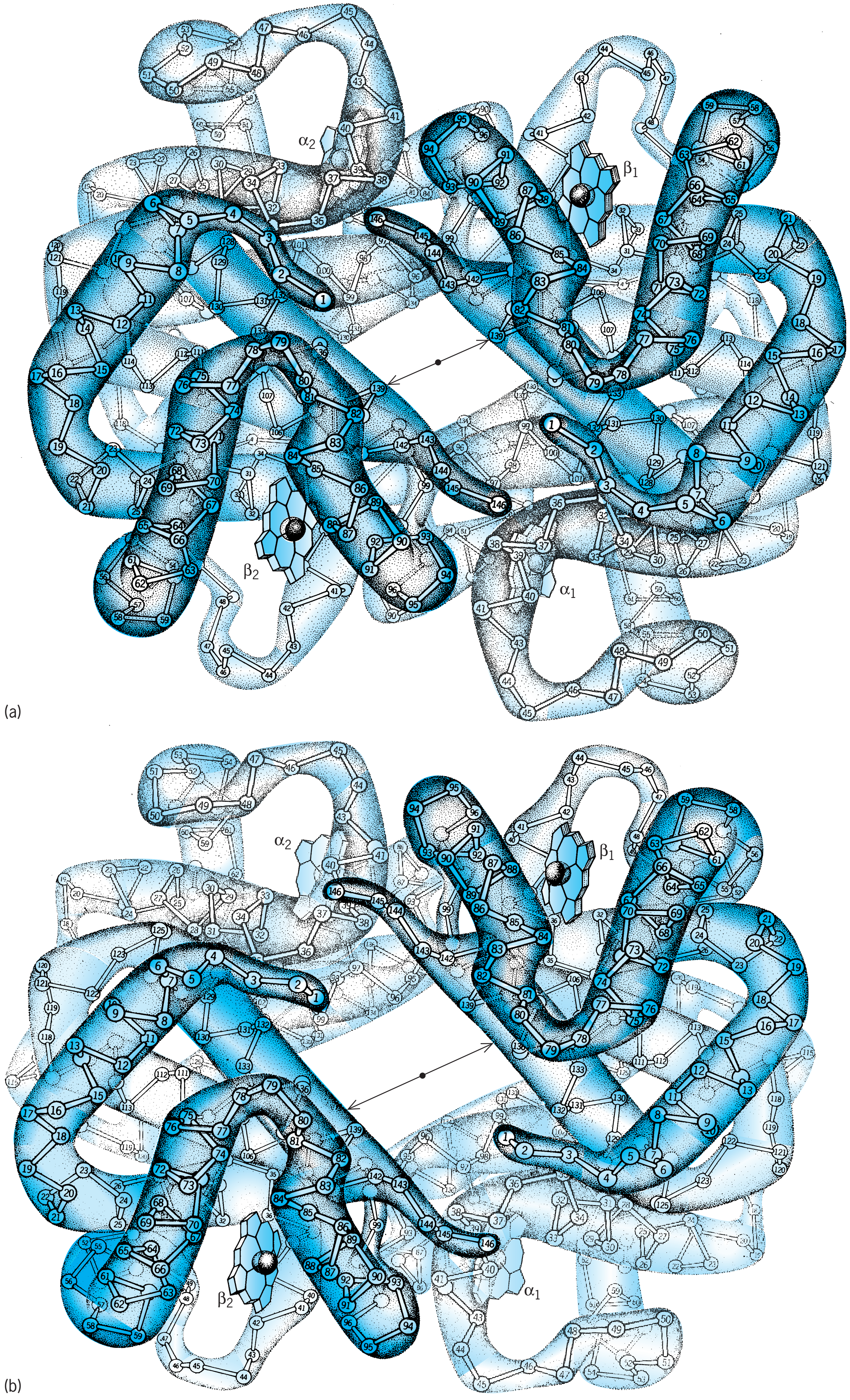
Synthesis and metabolism. The polypeptide chains and the heme are synthesized and combine together in nucleated red cells of the bone marrow. As those cells mature, the nuclei fragment and the cells, now called reticulocytes, begin to circulate in the blood. After sufficient hemoglobin has been formed in the reticulocyte, all nuclear material disappears and the cell is then called an erythrocyte, or red blood cell. Each hemoglobin molecule lasts as long as the red cell, which has an average life of 120 days.

When the old red cell is destroyed, the heme is metabolized to iron, which is reused, and to porphyrin, which is degraded to bile pigments and excreted by the liver. The polypeptide chains are broken down to amino acids, which may then be reused in other metabolic processes, including the synthesis of new proteins. *See* BILIRUBIN; HEMATOPOIESIS; PORPHYRIN.

Heterogeneity and genetic control. Several different kinds of normal hemoglobins occur in humans. Embryonic hemoglobins are present in the first 2 months of intrauterine life. Normal fetal hemoglobin (Hb F) predominates during the rest of fetal life; it has a molecular formula of $\alpha_2\gamma_2$. The alpha chains are identical to those in Hb A, and the gamma chains are similar but not identical to the beta chains. Two types of gamma chains are normally produced. They differ from one another by only a glycine or alanine in position 136 of the gamma polypeptide chain. The gamma chains are not normally synthesized after birth, and Hb F disappears from circulation within the first 3–4 months of life. Another normal hemoglobin, Hb A_2, which has a molecular formula of $\alpha_2\delta_2$, makes up about 2–3% of the total hemoglobin in normal humans. Each type of polypeptide chain is determined by a separate genetic locus (**Fig. 2**). These genes determine the exact sequence of amino acid groups in the polypeptide chains. In addition to these genetically different hemoglobins, other minor components of hemoglobin are formed as the result of chemical modifications that occur after protein synthesis. One of the most interesting is Hb A_{Ic}, which results from the glycosylation of Hb A by glucose from the blood. The relative amount of Hb A_{Ic} is influenced by the concentration of blood sugar and the age of the red cell. *See* GENETIC CODE.

Mutations of these genes result in changes in the amino acid structure of the polypeptide chains. The abnormal gene responsible for sickle-cell anemia causes the normal glutamyl amino acid residue at position six of the beta chain of Hb A to be substituted by a valyl group. This single change out of the total 146 amino acid residues of the beta chain causes a severe, lifelong anemia in individuals who have two mutant genes. Almost 400 other abnormal human hemoglobins are known, but only about one-third are associated with anemia or other diseases. *See* HEMATOLOGIC DISORDERS; MUTATION.

The hemoglobins of different species are similar in molecular size and chemical properties. The corresponding polypeptide chains of closely related animals are very similar in their sequences of amino


α2
β1
β2
α1
(a)
α2
β1
β2
α1
(b)

acid residues. Greater differences occur between hemoglobin of less related animals. Evolutionary differences between animals can be estimated by a comparison of the structure of their hemoglobins. *See* GENETICS.

Normal values and function. Normal adult males and females have about 16 and 14 g, respectively, of hemoglobin per 100 ml of blood; each red cell contains about 29×10^{-12} g of hemoglobin. Red cells normally comprise 40–45% of the volume of whole blood. Determination of these quantities is useful in diagnosis of hematologic disorders. Significant deviations from the values occur under altered conditions of red-cell production and destruction. *See* CLINICAL PATHOLOGY.

The reversible combination of hemoglobin and oxygen can be presented by the reaction shown below. The equilibrium constants for each step are

$$\mathrm{Hb} + 4\mathrm{O}_2 \overset{K_1}{\rightleftharpoons} \mathrm{HbO}_2 + 3\mathrm{O}_2 \overset{K_2}{\rightleftharpoons} \mathrm{Hb(O_2)_2} + 2\mathrm{O}_2 \overset{K_3}{\rightleftharpoons} \mathrm{Hb(O_2)_3} + \mathrm{O}_2 \overset{K_4}{\rightleftharpoons} \mathrm{Hb(O_2)_4}$$

not the same because an oxygen molecule on one heme group changes the affinity of the other hemes for additional oxygen molecules. This alteration in binding affinity during oxygenation is called heme-heme interaction and is due to small changes in the three-dimensional structure of the molecule. Some of the differences between the structure of oxygenated and deoxygenated hemoglobin are shown in Fig. 1. The principal change illustrated is a shift of the $\alpha_1\beta_2$ subunits relative to the $\alpha_2\beta_1$ subunits, resulting in an increase in the space between the beta chains in deoxyhemoglobin. These two structural conformations have different affinities for oxygen. The relationship between the pressure of oxygen and the oxygen content of a solution of hemoglobin can be graphed. The sigmoid curve obtained is a result of heme-heme interaction. Similar measurements of myoglobin, the oxygen storage molecule of muscle, reveal a rectangular hyperbola which indicates no heme-heme interaction (**Fig. 3**). Myoglobin is one-quarter the size and weight of hemoglobin and has only one heme per molecule.

Oxidation states. Hemoglobin combines reversibly with carbon monoxide about 210 times more strongly than with oxygen. This strong affinity for carbon monoxide accounts for the poisoning effects of this gas. The iron of each heme is normally in the ferrous state. It can be oxidized to the ferri state, in which case the hemoglobin, called ferrihemoglobin or methemoglobin, will no longer combine with either oxygen or carbon monoxide. Ferrihemoglobin will combine strongly with ions such as fluoride, cyanide, and hydroxide. The two oxidation states of hemoglobin and various chemical combinations can be determined by spectrophotometric measurements. The dark-bluish color of venous blood is due to the predominance of deoxygenated

Fig. 1. Three-dimensional structure of (*a*) oxyhemoglobin and (*b*) deoxyhemoglobin. The four polypeptide subunits are associated as two identical half-molecules, each containing one alpha and one beta chain. A single iron-containing heme group is associated with each globin chain subunit. The positions of the two alpha-beta half-molecules change relative to each other when oxygen combines with deoxyhemoglobin to form oxyhemoglobin. *(From R. E. Dickerson and I. Geis, Hemoglobin: Structure, Function, Evolution, and Pathology, Benjamin/Cummings, 1983; copyright by I. Geis)*

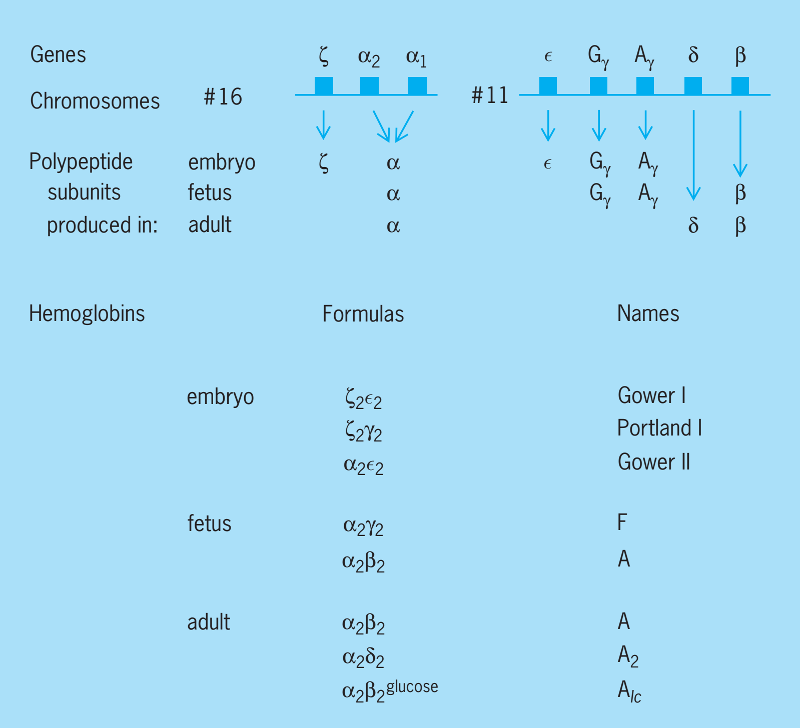


Fig. 2. The chemistry and genetics of hemoglobin. The ζ and two different α-globin genes are located on chromosome 16 (in humans). The other non-α-globin chain genes are located on chromosome 11. The two α genes produce the same α-globin polypeptide chain. The two γ genes make two slightly different polypeptide chains. The β^{glucose} chain results from the glycosylation of the beta chains in Hb A, and occurs continuously at a slow rate throughout the life-span of the red blood cell.

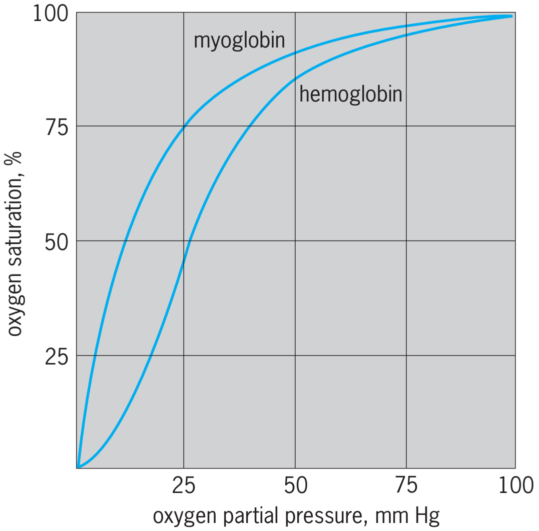


Fig. 3. Myoglobin and hemoglobin oxygen saturation curves. The sigmoid-shaped curve of oxygenation of hemoglobin results from several small changes in the structure of the hemoglobin molecule that occur as O_2 reacts with each heme group. 1 mmHg = 133.3 Pa.

hemoglobin. The bright-red color of arterial blood is due to oxyhemoglobin. A cherry-red color results from the presence of carbonmonoxyhemoglobin.

Blood chemistry. Hemoglobin binds carbon dioxide by means of free amino groups of the protein but not by the heme group. The reversible combination with carbon dioxide provides part of the normal blood transport of this gas. Hemoglobin serves also as a buffer by reversible reactions with hydrogen ions. The acidic property of oxyhemoglobin is greater than deoxygenated hemoglobin. The extra binding of hydrogen ion by deoxyhemoglobin promotes the conversion of tissue carbon dioxide into bicarbonate ion and increases the amount of total carbon dioxide that can be transported by blood. *See* BLOOD; RESPIRATION. Richard T. Jones

Bibliography. H. F. Bunn and B. G. Forget, *Hemoglobin: Molecular, Genetic and Clinical Aspects*, 1986; A. D. Stephens (ed.), *Haemoglobin: Research and Applications*, 1987.

Hemophilia

A rare, hereditary blood disorder marked by a tendency toward excessive bleeding. It is almost entirely restricted to males, and is transmitted as a sex-linked mendelian recessive trait passing from an affected male through an unaffected or very mildly affected daughter to appear again in a grandson. Queen Victoria was a carrier, and several of her male descendants were affected. *See* HUMAN GENETICS; SEX-LINKED INHERITANCE.

Classical hemophilia (hemophilia A) is due to a deficiency of the antihemophilic factor or factor VIII, a clotting factor which is normally present in the blood in trace amounts and is essential for normal fibrin formation. Hemophilia B is a similar sex-linked bleeding disorder affecting males but characterized by a deficiency of another blood clotting factor, factor IX. It can only be distinguished from hemophilia A, which it closely resembles, by laboratory tests. Hereditary deficiencies of other clotting factors may give rise to bleeding disorders similar to hemophilia, but they are inherited as autosomal dominant or recessive characteristics. Most of the hereditary hemorrhagic disorders, including both types of hemophilia, have been observed in animals. Most of the clinical manifestations follow trauma. Even a slight blow may produce severe hemorrhage into the affected body tissues. Bleeding into the joints is seen in the most severe cases, and may result in crippling deformities. Bleeding into muscles and subcutaneous tissue is common, while bleeding from the mucous membranes occurs from time to time. There are large numbers of subtypes of each disorder, which vary markedly in clinical severity. Thus the clinical spectrum ranges from individuals who bleed excessively only following major hemostatic challenges and may live a normal life to those who have crippling deformities. Potent concentrates of factor VIII or IX prepared from human plasma are available, and are very effective in the management of hemophilia A and B, respectively. They must be given intravenously. When given in adequate amounts, they can restore the hemostatic potential of the blood to normal for many hours. *See* BLOOD; HEMORRHAGE. Cecil Hougie

Bibliography. S. Chein et al. (eds.), *Clinical Hemorheology*, 1987; W. J. Williams et al., *Hematology*, 3d ed., 1983.

Hemorrhage

The escape of blood from within the vascular system. Hemorrhage may result from either trauma or disease of the vessel wall.

Causes. The escape of blood following rupture of a vessel wall as a result of trauma is obvious and needs no further explanation. The causes other than trauma can be divided into three main groups.

The first group consists of these conditions in which there is a chronic disease process affecting the vessel wall, such as atherosclerosis or aneurysm formation. Either of these conditions, in association with an elevated blood pressure, can result in a break in the wall and subsequent hemorrhage. An infarct, or tissue death from any cause, may also result in hemorrhage. *See* ANEURYSM; ARTERIOSCLEROSIS; INFARCTION.

The second group consists of those causes in which there is an acute process affecting the vessel wall, such as septicemia, bacterial toxins, or anoxia. *See* HYPOXIA.

The third group consists of those hemorrhagic conditions which result from some defect in the blood itself. Under this heading are leukemia, thrombocytopenia, and the clotting disorders. *See* LEUKEMIA.

Petechiae are hemorrhages no larger than the head of a pin. Hemorrhages of greater size are termed ecchymoses. A localized mass of blood in tissue is a hematoma. Spontaneous hemorrhaging into the skin and mucosal surfaces is termed purpura. This usually denotes a disease of the vascular system or of the blood itself, such as a deficiency of blood platelets. Platelets are thought to be necessary to maintain capillary integrity.

Cerebrovascular accident. Cerebrovascular accident, or stroke, is an acute vascular lesion of the brain. This may be the result of hemorrhage from, thrombosis in, or embolism to a cerebral vessel. Cerebral hemorrhage can result from a rupture of the vessel wall, as from a defect in the wall such as a congenital aneurysm, or it may be secondary to death of cerebral tissue as occurs with an infarct. Intracerebral hemorrhages may be small petechial or perivascular hemorrhages, or they may be massive. Small hemorrhages can result from minimal trauma, poisoning such as from carbon monoxide, or from blood diseases such as leukemia. Massive cerebral hemorrhage is frequently the result of trauma or vascular disease. With cerebral hemorrhage or death of cerebral tissue there is loss of function of the regions involved. One such complication is paralysis of the muscles of the extremities on the side opposite the lesion. Since the brain is enclosed in a nondistensible box, the skull, the increase in size due to the addition of the blood raises intracranial pressure. *See* EMBOLISM; THROMBOSIS.

Compensatory mechanisms. Following the sudden loss of a large quantity of blood, the body reacts to compensate for this loss and the blood clotting mechanism is activated. As the blood volume is decreased, the carotid sinus reflex comes into play. This results in a generalized vasoconstriction which affects the arteries and precapillary and capillary vessels. With a lowering of capillary pressure, there is a shift of fluid from the tissue spaces into the plasma. This tends to restore the blood volume but lowers the hematocrit (the relative percentage of erythrocytes, or red cells, in the blood). Following a loss of 500 ml of blood, a healthy person can restore blood volume in 24 h. The protein content can be restored in a few hours, but the replacement of the cellular elements takes days or even weeks. Failure of these mechanisms to compensate adequately can result in shock with dire consequences. *See* CARDIOVASCULAR SYSTEM.

Morphologic changes of hemorrhages. Small hemorrhages can be completely absorbed leaving no tissue alterations; larger volumes of blood act as a foreign substance which is destroyed and removed by the phagocytes and may result in scarring. The red cells are broken down, liberating hemoglobin. This released hemoglobin is divided into two moieties, the iron-free hematoidin and the iron-containing hemosiderin.

In animals dying of severe or fatal hemorrhage a state of severe ischemia of all the organs is found. The mucous and serous surfaces are pale; the organs are drier than normal and reduced in weight. *See* CIRCULATION DISORDERS; HEMATOLOGIC DISORDERS.

Romeo A. Vidone; Irwin Nash

Bibliography. J. M. Kissane (ed.), *Anderson's Pathology*, 9th ed., 1989; R. S. Cotran et al., *Robbins' Pathologic Basis of Disease*, 6th ed., 1999; W. J. Williams et al., *Hematology*, 3d ed., 1983.

Hemp

The fiber and the plant *Cannabis sativa*. It should not be confused with Manila hemp, which is not related to true hemp. *See* ABACA.

Hemp fiber, which for many years was the major raw material used in the manufacture of rope, now is used mostly in the production of small twines, linenlike fabrics and canvases, and, to some extent, in making special types of paper. *See* NATURAL FIBER.

The plant is about 6–8 ft (2–2.5 m) tall at maturity and has a stem which at midpoint is roughly the diameter of a lead pencil.

Hemp is an annual crop, most of which is produced in Eastern Europe and mainland China, with some production in South Korea, Turkey, and Italy. World production has declined over the years. In the United States, hemp fiber production gradually declined to none by the mid-1950s. Hemp seed production in the United States was centered in Kentucky and fiber production in the upper Midwest.

Hemp is planted about the same time as corn and requires about 4 months of frost-free weather to produce a crop of fiber. Hemp needs a fertile soil and good drainage. The fiber is separated by crushing the retted (partially rotted) stems and beating out the nonfibrous material.

Hemp contains the drug marijuana, and a permit from the Drug Enforcement Administration of the Department of Justice is required to grow or possess any part of this plant in the United States. *See* MARIJUANA.

Elton G. Nelson

Henequen

A fiber obtained from the leaves of *Agave fourcroydes*. It is produced only in Mexico, Cuba, and El Salvador. Henequen is sometimes incorrectly called sisal, which is a closely related plant grown in Brazil and Africa. *See* SISAL.

A well-developed henequen plant has a short stem about 12–18 in. (30–45 cm) in diameter and 3–5 ft (1–1.5 m) in height at maturity. Leaves are 3–6 ft (1–2 m) in length and 4–6 in. (10–15 cm) wide near the base. They are gray-green, thick, succulent, and smooth, with many small, curved spines on the edges and a sharp terminal spine 0.4–1.2 in. (1–3 cm) long. When the plant is 15 to 25 years old, a flower stalk grows up through the top to a height of 13–26 ft (4– 8 m) and develops terminal branches which bear flowers near their tips. The flowers are followed by seed or by bulbils or sometimes by both.

Cultivation. The henequen plant requires a dry, tropical climate and a well-drained soil, preferably of limestone formation. The average annual rainfall in Yucatan, the henequen-growing area of Mexico, is about 30 in. (750 mm) and the temperature is seldom below 59°F (15°C). Many of the plantations there are rocky, and land preparation consists of removal of trees, shrubs, and some of the native vegetation before setting out the plants. Where conditions permit, land for planting should be plowed and cultivated.

Henequen plants are propagated from bulbils or more generally from suckers. Bulbils must be cultivated in a nursery for 1 or 2 years before they are set out in the field. Suckers are dug when 16–24 in. (40–60 cm) in height and, after some of the top is cut back, are set out directly in the field. Plantations are usually in double rows about 3 ft (1 m) apart and spaced about 4 ft (1.25 m) apart in the row, with about 11 to 13 ft (3.5–4 m) between each pair of rows. Some planting is in single rows.

Diseases. Henequen is affected by fewer diseases than most plants. Disease symptoms are frequently caused by nutrient deficiencies. The fungus *Colletotrichum agaves* may attack the leaves and cause leaf spots, especially if they have been punctured by insects. Infestation by the sisal weevil, *Scyphophorus interstitialis,* which bores into the bud, can be reduced by field sanitation and by spraying with aldrin or dieldrin.

Harvesting and processing. The first cutting of henequen leaves for fiber is in the sixth or seventh year. Successive harvests are at 6- to 12-month intervals for periods of 10 to 20 years. About 15 to 20 leaves at an angle of 40° or more from the vertical are cut each time. The terminal and marginal spines are trimmed off, and the leaves are tied in bundles and taken to a defibering machine where they are beaten and scraped. The fiber obtained is usually dried in the sun, but sometimes dryers are used.

Uses. The greatest quantity of henequen fiber goes into farm twine, followed by industrial tying twine and then light-duty rope. Padding for innerspring mattresses is made from the lowest grades of fiber and from flume tow, the short, tangled fiber that can be recovered from the cleaning operation. Henequen is exported as manufactured twine, rope, or padding, not as raw fiber. *See* NATURAL FIBER. Elton G. Nelson

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, vol. 10, 3d ed., 1980.

Heparin

A highly sulfated mucopolysaccharide isolated from mammalian (chiefly beef) tissues, with blood anticoagulant activity. Heparin was first found in abundance in the liver, hence the name, but it is present in substantial amounts in the spleen, muscle, and lung as well. Chiefly composed of D-glucosamine and D-glucuronic acid residues, it usually has a molecular weight from 6000 to 20,000. It normally occurs as the sodium or potassium salts and undergoes some degree of desulfation when the free acid is liberated. Full desulfation of heparin yields heparamine, which on treatment with carboxylic acid anhydrides furnishes heparides, that is, amides of the polysaccharide, free of anticoagulant properties but still valuable for treating hyperlipidemia. In the blood of most mammals, heparin is an antagonist to thrombin, prothrombin, and thromboplastin. It lessens the tendency of platelets to agglutinate. When it is injected, its action is almost instantaneous, but it begins to decline rapidly in 2–3 h. Heparin is only effective when injected, and the site of the injection often becomes swollen or inflamed and causes pain. It is used in the treatment of venous thrombosis, embolism, myocardial infarction, and certain types of cerebral thrombosis. Heparin is the single most important anticoagulant used in medicine. *See* POLYSACCHARIDE. Frank Wagner

Bibliography. R. A. Bradshaw and S. Wessler (eds.), *Heparin-Structure: The Function and Clinical Implications*, 1975; W. D. Comper, *Heparin (and Related Polysaccharides): Structural and Functional Properties*, 1981; N. M. McDuffie (ed.), *Heparin: Structure, Cellular Functions, and Clinical Applications*, 1979; I. Witt (ed.), *Heparin: New Biochemical and Medical Aspects*, Proceedings of the Symposium of the Deustche Gesellschaft fur Klinishe Chemie, 1983.

Hepaticopsida

A class of lower green plants called liverworts that belong to the division Bryophyta. The class is divided into approximately 225 genera and 8500 species. Although there is a great diversity of external form, most of the gametophytes (gamete-producing plants) are dorsoventrally differentiated. These plants are considered among the most primitive of the existing land plants. Fossil remains of liverworts have been found in both the Devonian and Carboniferous. Since the fossils found do not differ significantly from modern liverworts, they are of little value in ascertaining phylogenetic relationships. Liverworts are widely distributed over the world, but have their greatest diversity in the tropics of the Americas and East Indies.

Except when the plants occur in masses, they are quite inconspicuous and are usually confused with mosses, which they resemble somewhat in their external appearance. In the presence of adequate moisture, they grow on soil, rocks, and tree trunks. Usually the plant body is a thin, prostrate thallus, sometimes having a short central axis with leaflike appendages. On the lower surface are rhizoids (rootlike structures) which function in anchorage and absorption. *See* GEOLOGY.

Reproduction. The sex organs, antheridium (male) and archegonium (female), may be produced on the same plant or on different plants. The sperms, produced in the antheridia, are flagellate and motile. The sporophyte (spore-producing generation), which develops as a result of fertilization, remains on the female gametophyte, and for a time is parasitic. The

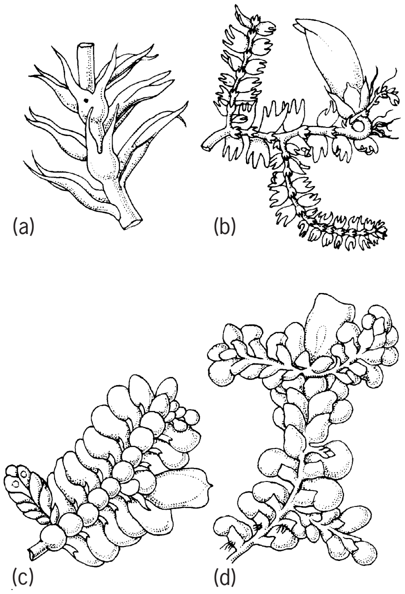


Fig. 1. Jungermanniidae, the leafy liverworts. (*a*) *Herberta*, showing three ranks of equal bifid leaves. (*b*) *Lepidozia*, ventral aspect, showing reduced ventral leaves. (*c*) *Leucolejeunea*, ventral aspect, showing reduced ventral leaves and complicate dorsal leaves. (*d*) *Radula*, ventral aspect, showing absence of ventral leaves but complicate dorsal leaves. (*After E. W. Sinnott and K. S. Wilson*, *Botany: Principles and Problems*, *5th ed.*, *McGraw-Hill*, *1955*)

spores are developed within a sporangium, or capsule, which lacks a sterile region or columella internal to the spore-producing tissue as occurs in the mosses. After discharge from the sporangium the spores germinate, forming short protonemas from which arise new liverwort plants (gametophytes). *See* BRYOPSIDA.

Classification. The Hepaticopsida are divided into two subclasses, the Jungermanniidae and the Marchantiidae.

The Jungermanniidae, with about 190 genera and 8000 species, are called the leafy liverworts because of their chlorophyll-containing ribbonlike or leaflike bodies (**Fig. 1**). The thallus is never differentiated, and it lacks the air spaces and external pores characteristic of the Marchantiidae. Only unicellular, smooth-walled rhizoids are present. The sporophyte often produces a long stalk (sporangiophore) bearing the sporangium, or capsule, on the upper end. When mature the sporangium dehisces (opens) by four valves. The Jungermanniidae are composed of four orders: Takakiales, Calobryales, Jungermanniales, and Metzgeriales.

The Marchantiidae with about 35 genera and 450 species, are called thallose liverworts, the best-known example being *Marchantia polymorpha*. The flat plant body is composed of several distinct layers of tissue, having the chlorophyll-containing cells of the uppermost layer separated by numerous air chambers which are exposed to the external atmosphere through pores. Rhizoids are of two kinds, smooth-walled and tuberculate-walled. The male and female sex organs, each borne on separate plants, are generally found grouped in the tops of special stalked structures known respectively as antheridiophores and archegoniophores. The Marchantiidae are composed of three orders: Sphaerocarpales, Monocleales, and Marchantiales.

Alternation of generations. The life cycle of liverworts is well illustrated by *Marchantia* (**Fig. 2**). The gametophyte, or gamete-producing generation, is the dominant phase. In *Marchantia*, the gametophyte is a flat, dichotomously branched (forked) thalloid structure produced by apical growth of the thallus. As development proceeds, the thallus produces numerous rhizoids from the ventral (under) surface. The internal anatomy shows considerable differentiation of tissues (**Fig. 3**). Starting with the upper surface, there is a single epidermal layer, subtended by an area of air chambers which are connected to the external atmosphere through regularly spaced pores or openings. The pores are surrounded by four rows of four cells each, the upper and lower circles being somewhat smaller, giving the whole structure a barrel-shaped appearance. Below these openings, the vertical pillars of cells and the multibranched filaments at the base of the chambers have numerous oval chloroplasts which give the rich, dark-green color so characteristic of the plant. The remainder of the thallus is made up of densely arranged parenchymatous cells. The lower parenchyma tissue has few chloroplasts, but it contains numerous large mucilage cells which probably function as storage tissue. *See* EPIDERMIS (PLANT); PARENCHYMA.

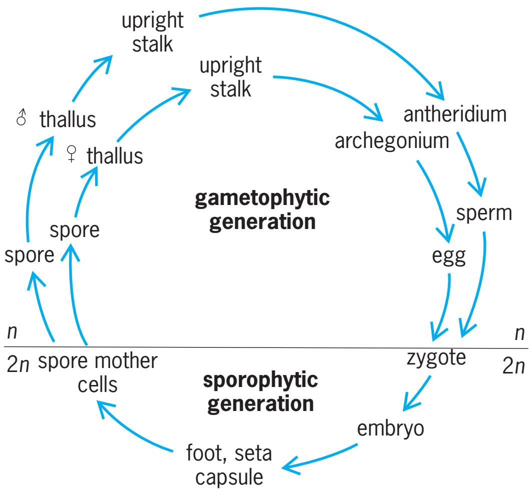


Fig. 2. Diagram of life cycle of the liverwort *Marchantia*, indicating stages at which chromosome changes occur. (*After H. J. Fuller, The Plant World, rev. ed., Holt, 1951*)

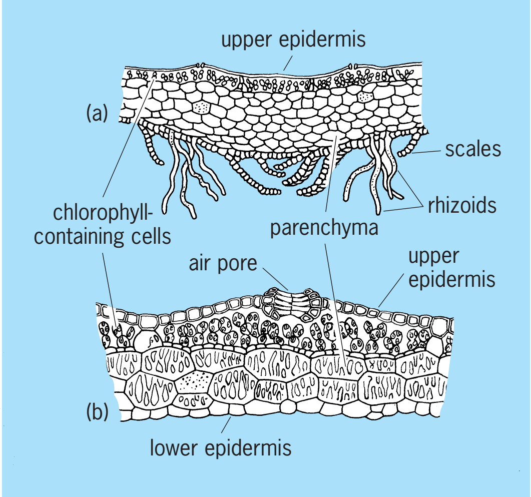


Fig. 3. Transverse sections through the thallus of *Marchantia*. (*a*) Middle portion with scales and rhizoids on the underside. (*b*) Margin of the thallus more highly magnified. (*After H. Kraemer, Applied and Economic Botany, 2d ed., Wiley, 1916*)

The plants are dioecious; that is, they have male and female sex organs on different plants. The antheridia, or male sex organs, are produced on stalked, disk-headed branches called antheridiophores which grow upward from the apical notch of the thallus or gametophyte (**Fig. 4**). The antheridia are embedded in cavities of the disk, each cavity having only one antheridium which, when mature, discharges the numerous biflagellated motile sperms to the exterior through a surface pore. Maturation of antheridia proceeds from the center of the antheridial disk outward and is often indicated by the cell walls becoming deep purple in color.

The archegonia are found on similar stalks called archegoniophores (**Fig. 5**). At the apex of each

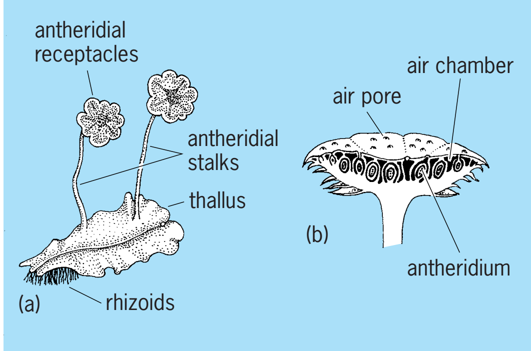


Fig. 4. ***Marchantia*****. (*****a*****) Example of male gametophyte (antheridial plant) (*****after H. J. Fuller and O. Tippo, College Botany, Holt, 1949*****). (*****b*****) Section through an antheridial receptacle (*****after W. J. Robbins and H. W. Rickett, Botany, 3d ed., Van Nostrand, 1939*****).**

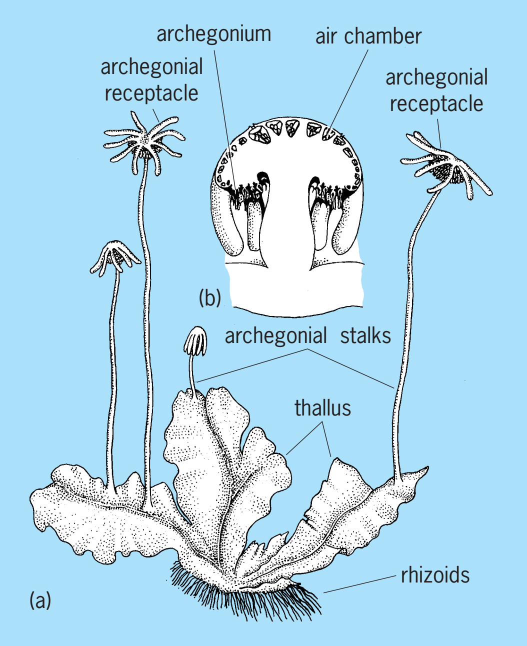


Fig. 5. ***Marchantia*****. (*****a*****) Example of female gametophyte (archegonial plant) (*****after H. J. Fuller and O. Tippo, College Botany, Holt, 1949*****). (*****b*****) Section through an archegonial receptacle (*****after W. J. Robbins and H. W. Rickett, Botany, 3d ed., Van Nostrand, 1939*****).**

archegoniophore about nine finger-shaped structures project outward like the ribs of an umbrella. The archegonia are formed in groups near the bases of these arching projections. Each projection may produce 12–16 archegonia in succession from its tip inward.

There has long been some question as to how the sperms reach the egg cells in the archegonia. It has been observed that raindrops may splash the sperm as much as 2 ft (60 cm). Mites have been observed on the archegoniophore with live sperm on their bodies. Following a rain, or with heavy dew, the water film is probably adequate to allow the sperms to swim from the antheridia to the archegonia.

As a result of fertilization a zygote is formed which develops into a sporophyte (**Fig. 6**). This is accomplished by a series of divisions of the zygote (fertilized egg) which finally results in a structure having a foot embeded in the gametophyte; an elongated seta, or stalk, which shows its most rapid development at the time of spore maturation; and a sporangium, or capsule, within which the spores are developed following a meiotic division of the spore mother cells.

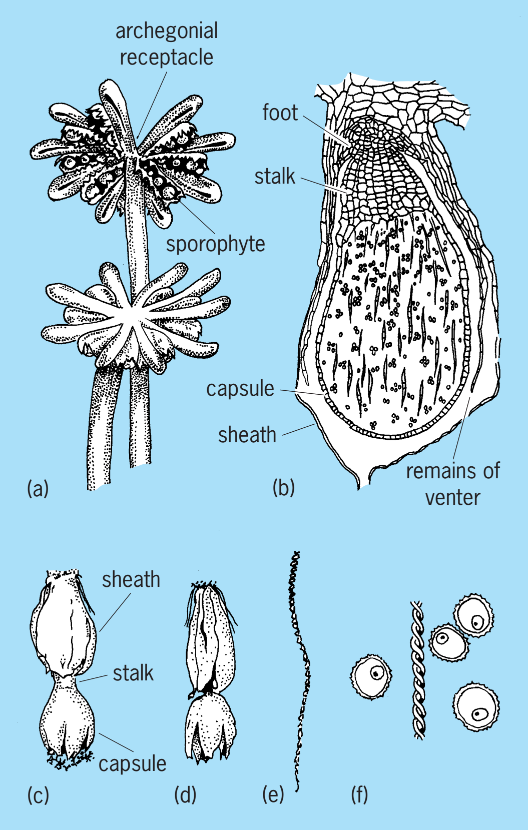


Fig. 6. ***Marchantia*** **sporophyte morphology. (*****a*****) Mature archegonial receptacles of sporophytes (*****after W. J. Robbins and H. W. Rickett, Botany, 3d ed., Van Nostrand, 1939*****). (*****b*****)** ***Marchantia polymorpha*****, longitudinal section showing the internal structure of a nearly mature sporophyte; (*****c, d*****) surface views of sporophytes during and after shedding of spores and elaters; (*****e*****) portion of an elater; (*****f*****) spores and portion of an elater (*****after G. M. Smith, Cryptogamic Botany, vol. 2, 2d ed., McGraw-Hill, 1955*****).**

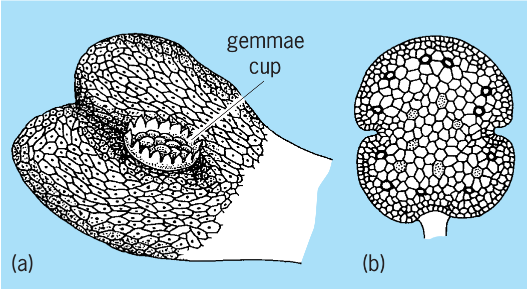


Fig. 7. Structures involved in vegetative reproduction of ***Marchantia*****. (*****a*****) Portion of surface of the thallus. Note gemma cup with gemmae (*****after H. J. Fuller and O. Tippo, College Botany, Holt, 1949*****). (*****b*****) Notched, lens-shaped gemma of** ***M. polymorpha*** **(*****after G. M. Smith, Cryptogamic Botany, vol. 2, 2d ed., McGraw-Hill, 1955*****).**

Special pointed elongated cells, known as elaters, aid in the dispersal of spores by being sensitive to internal water tensions which cause them to coil and uncoil. Spores falling upon a suitable substrate germinate immediately and produce new gametophytes. *See* MEIOSIS.

In *Marchantia* there are also two methods of vegetative reproduction. The first results from the growth habit of the plant. As it grows and branches, the older portions of the plant decay, thus separating the branches into individual plants. The second method is in the formation of gemmae cups on the dorsal surface (**Fig. 7**). Within each cup several small, notched, lens-shaped bodies are produced which, when detached, grow from both ends simultaneously and develop into new plants. This provides for rapid dispersal during the growing season. *See* BRYOPHYTA; CALOBRYALES; JUNGERMANNIALES; JUNGERMANNIIDAE; MARCHANTIALES; MARCHANTIIDAE; METZGERIALES; MONOCLEALES; SPHAEROCARPALES. Paul A. Vestal

Hepatitis

An inflammation of the liver caused by a number of etiologic agents, including viruses, bacteria, fungi, parasites, drugs, and chemicals. All types of hepatitis are characterized by distortion of the normal hepatic lobular architecture due to varying degrees of necrosis of individual liver cells or groups of liver cells, acute and chronic inflammation, and Kupffer cell enlargement and proliferation. There is usually some degree of disruption of normal bile flow, which causes jaundice. The severity of the disease is highly variable and often unpredictable. *See* LIVER.

Viral hepatitis. The most common infectious hepatitis is of viral etiology. Four types of viral hepatitis are recognized: type A (infectious) hepatitis; type B (serum) hepatitis; non-A, non-B (NANB) hepatitis; and delta hepatitis. The clinical course of viral hepatitis is highly variable, ranging from nearly asymptomatic disease to a severe prolonged illness capable of causing death. Signs and symptoms of viral hepatitis are nonspecific and consist of headache, fatigue, malaise, loss of appetite, nausea, jaundice, dark urine, and pain in the upper abdomen. A fever as high as 104°F (40°C) can occur in many with the disease.

Type A viral hepatitis is caused by a small ribonucleic acid (RNA) virus belonging to the picornavirus class. The disease is usually spread by the oral-fecal route, is associated with overcrowding and poor hygiene, and has an incubation period of 15–45 days. It occurs primarily in children and young adults, and is a usually mild disease leading to no residual effects. The virus is usually not detectable in the blood, but an antibody against it is usually detectable by serologic methods and can confirm the diagnosis.

Type B viral hepatitis is caused by a deoxyribonucleic acid (DNA) virus that belongs to a class referred to as hepadna viruses. In 1963 B. Blumberg discovered in an Australian aborigine an antigen which was initially called Australia antigen and, later, hepatitis-associated antigen, and which was subsequently shown to be part of the virus which causes type B hepatitis. The intact virion is called the Dane particle, and has been identified with the electron microscope. Hepatitis B is usually parenterally transmitted, has an incubation period of 50 to 160 days, occurs at any age, and is common in drug addicts. The infectious particle of hepatitis B (the Dane particle) is found in a variety of body fluids of those infected, including saliva, tears, seminal fluid, cerebrospinal fluid, ascites, breast milk, gastric juice, urine, and feces. Hepatitis B virus can be transmitted sexually and through the placenta during the birth process. Type B viral hepatitis has a mortality rate of up to 15%. In Asia and Africa, hepatitis B appears to be related to the development of liver cancer. It may proceed to chronic hepatitis in 10% of the cases. In addition, viral immune complexes of hepatitis B antigens and antibodies have been implicated as a causative agent in several relatively rare conditions—polyarteritis nodosa, membranous glomerulonephritis, and mixed cryoglobulinemia. There is an effective hepatitis B virus vaccine that is in widespread use in medical personnel and others who are at an increased risk. *See* ANTIBODY; ANTIGEN; IMMUNE COMPLEX DISEASE.

Non-A, non-B hepatitis. Liver inflammation referred to as non-A, non-B hepatitis is most common among people who have received transfused blood and whose serologic tests show no evidence of hepatitis A, hepatitis B, or other types of virus such as Epstein-Barr. The causative virus can also be found in clotting factor concentrates and is detected with relative frequency in hemophiliacs who receive large amounts of blood-clotting concentrates. *See* ANIMAL VIRUS; EPSTEIN-BARR VIRUS; HEMOPHILIA.

Delta hepatitis. Delta hepatitis, the most recently recognized hepatitis agent, is caused by the delta agent hepatitis D virus (HDV), a defective ribonucleic acid (RNA) virus that requires the helper function of hepatitis B virus for its replication and expression. Delta hepatitis has a worldwide distribution, but it occurs primarily in northern Africa, southern Europe, and the Middle East.

Alcoholic hepatitis. Another frequently occurring form of hepatitis is caused by excessive ethyl alcohol intake and is referred to as alcoholic hepatitis. It usually occurs in chronic alcoholics and is characterized by fever, high white blood cell count, and jaundice. Microscopically, the hepatic lobular architecture is disrupted by fatty change, acute inflammation with necrosis, and peculiar intracytoplasmic aggregates of keratin proteins called alcoholic hyaline. This form of hepatitis is reversible if the etiologic agent, ethyl alcohol, is withdrawn. Persons who develop alcoholic hepatitis are often misdiagnosed, with obstructed bile ducts being a common assumption, and some undergo exploratory surgery for that condition. *See* ALCOHOLISM.

Drug-induced hepatitis. Some drugs are capable of damaging the liver and can occasionally cause enough damage to produce clinical signs and symptoms. The injury can be so severe that acute liver failure results. Among those drugs known to damage the liver are tetracycline, methotrexate, anabolic and contraceptive steroids, phenacetin, halothane, chlorpromazine, and phenylbutazone. Usually the damage caused by those drugs is relatively mild, and complete reversal follows cessation of medication. Many patients present with jaundice, but only after a careful history is taken will a drug-induced cause be found or considered. *See* ALLERGY; ANALGESIC; ANESTHESIA; ANTIBIOTIC; CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; STEROID.

Chronic hepatitis. Chronic hepatitis is a condition defined clinically by evidence of liver disease for at least 6 consecutive months. It may be caused by a prolonged bout of hepatitis B or non-A, non-B hepatitis. Persons with chronic hepatitis frequently develop nonspecific signs and symptoms such as mild fatigue and abnormal liver function tests. The disease is categorized pathologically into chronic aggressive and chronic persistent forms. The distinction is important, because the chronic aggressive form may progress to cirrhosis and require treatment with corticosteroids. *See* CIRRHOSIS.

Signs and treatment. Clinical features of hepatitis include malaise, fever, jaundice, and serum chemical tests revealing evidence of abnormal liver function. In most mild cases of hepatitis, treatment consists of bedrest and analgesic drugs. In those individuals who develop a great deal of liver cell necrosis and subsequently progress into a condition known as hepatic encephalopathy, exchange blood transfusions are often used. This is done with the hope of removing or diluting the toxic chemicals thought to be the cause of this condition. *See* LIVER DISORDERS.

Samuel P. Hammar

Bibliography. R. J. Gerety, *Hepatitis B*, 1985; G. Gitnick (ed.), *Modern Concepts of Acute and Chronic Hepatitis*, 1988; I. D. Gust and S. M. Feinstone (eds.), *Hepatitis A*, 1988; Y. F. Liaw (ed.), *Chronic Hepatitis*, 1986; R. E. Sampliner, *Preventing Viral Hepatitis*, 1988; T. Shikata, R. H. Purcell, and T. Uchida, *Viral Hepatitis C, D, and E*, 1991; A. J. Zuckerman, *Viral Hepatitis and Liver Disease*, 1988.

Herbarium

A collection of pressed and dried plant specimens, and a description of when, where, and by whom they were collected, arranged in a systematic manner, and serving as a permanent physical record of the occurrence of an individual plant at a specific place and time. Herbaria may contain specimens from the full range of organisms that have classically been considered plants: fungi, lichens, algae, bryophytes, ferns and their allies, gymnosperms, and angiosperms. Many herbaria also accumulate and manage special collections such as liquid-preserved parts for anatomical studies, wood, seeds, or specially preserved material suitable for extraction of deoxyribonucleic acid (DNA) or other chemical constituents. Many groups of plants, especially those with succulent or fleshy parts, are not suitable for preservation as dry, flat specimens because they lose many of their important features in the drying process. Consequently, these plants are often preserved in liquid.

Purpose and uses. A herbarium is basically a library of physical specimens, readily accessible for reference or detailed study. The information embodied in each specimen and its label forms an important resource which can be used to answer many kinds of questions. Much of what is known about the majority of the world's species of plants is based on the specimens in herbaria. A large herbarium contains the results of the accumulated effort and information from thousands of people over hundreds of years. Herbarium specimens represent a sampling of the individual variation exhibited by plants and provide the primary data source for developing concepts of species, naming them, and communicating about them. Examining a large number of individuals allows for a more accurate assessment of the nature of a species and its ecological and geographical distribution. Specimens are used in taxonomic and ecological research, such as morphological studies, and for comparative identification and verification of unknown specimens derived from multidisciplinary scientific projects, governmental or private entities, and interested individuals. Herbaria are also major educational resource tools for training new scientists and professionals who deal with plant materials, and for providing basic information on plants (for example, morphology or phytogeography) for many different education levels.

Herbarium collections may include specimens of both wild and cultivated origin, and can give insights into the evolution of domesticated species or point to wild species of potential interest for new genetic traits. They can also provide information on the historical distribution of rare, threatened, or endangered species. *See* PLANT GEOGRAPHY; PLANT TAXONOMY.

Specimen preparation. A high-quality herbarium specimen should be carefully selected and be representative of the individual from which it is made. For many herbaceous plants one or more entire individuals can be included in a specimen, but for larger individuals the parts to be preserved must be chosen by the collector and should exhibit the full range of morphological features of the plant. A specimen is prepared to fit within the dimensions of a standard mounting sheet (11.5 × 16.5 in. or 29 × 42 cm) and must be arranged before drying so that all pertinent parts will be visible. Two or more mounting sheets may be needed to accommodate large specimens.

Each specimen is placed in a single folded sheet of newspaper and pressed between blotters and corrugated cardboard, forming a sandwich. Alternating units of specimen, blotter, and corrugate are then stacked to form a press with a wooden frame at each end and are bound with two web straps or ropes. The press is then placed over a source of low heat with the corrugations running vertically. The warmed air rises through the corrugations and draws moisture from the plants, while the pressure of the straps ensures that the plants dry flat.

The dried specimen is mounted on a sheet of stiff paper by using glue or gummed tape or by sewing with heavy thread. A standard mounting sheet should be of archival quality and made from 100% rag fiber. The mounting sheet provides support for the specimen and makes it relatively easy to manipulate without damage. Each mounting sheet also has a label glued in the lower right corner with the scientific name of the plant and information on where and when it was collected and by whom. A good label provides detailed information on features of the plant that are not apparent from a direct examination of the dried specimen, such as flower color, height or growth habit, and ecological factors such as habitat, soil type, and associated species.

Structures that are too large to be pressed, such as cones, large fruits, or palm inflorescences, are dried and stored in boxes or bags and are cross-referenced to the conventional, flat specimen to which they belong.

Organization. A herbarium may be arranged in any convenient way as long as the specimens and their information can be easily retrieved. Herbaria are usually arranged by family, genus, species, and infraspecific name, often combined with some geographical scheme. In many collections, specimens are arranged alphabetically by genus and species within a family, or the arrangement may follow an accepted phylogenetic scheme. Specimens are filed in manila folders, often called genus covers, with one or more individual sheets per folder, depending on the thickness of the specimens. In some collections, added protection is provided for each specimen by placing it in a fold of thin, high-quality paper before filing in the genus cover.

Storage units. In the earliest herbaria, specimens were glued to pages, that were bound into volumes. As the number of specimens in collections increased, this method was abandoned, primarily because it was impossible to add to or change the arrangement of the specimens in the bound volumes. Most herbaria today store specimens in steel or wooden cases fitted with two columns of compartments measuring 17 in. (43 cm) deep, 13 in. (33 cm) wide, and 6 in. (15 cm) high (see **illus.**). Because standard cases require permanent aisles wide enough to open the doors, many collections have converted their specimen storage space to mobile modules (compactors), where the units, driven by manual assists or electric motors, move along rails and require no permanent aisles. This is the most compact and efficient means for storing herbarium specimens in limited space.

Protection. Once a plant specimen is prepared and dried, it will essentially last indefinitely if it is properly protected from environmental factors, such as insect pests and high relative humidity. There are specimens from the sixteenth and seventeenth centuries that are as well preserved and useful today as when they were first gathered. As with other items containing cellulose, such as paper, frequent changes in temperature and relative humidity can have an adverse effect and hasten deterioration. The most significant problem, however, for any herbarium is damage from the larvae of various stored-product pests. The most widespread and destructive is the tobacco beetle (*Lasioderma serricorne*). There are no known treatment methods for specimens which will permanently protect them from insect attack. Therefore, a collection must be regularly monitored and any infestations eliminated to prevent damage. An integrated pest management program that combines monitoring, collection isolation and freezing, and judicious use of repellents or pesticides is effective, taking into account the constraints imposed by the physical facility.

Standard herbarium case. (*Harvard University Herbaria; photo by Stephen Jennings*)

Distribution. The *Index Herbariorum* provides information on the majority of the world's herbaria (2639 herbaria in 147 countries holding nearly 273 million specimens). Within the United States there are 628 listed herbaria, with more than 60 million specimens. The oldest herbaria date from the middle of the sixteenth century, but the majority were formed during the twentieth century. Globally, the average herbarium contains about 100,000 specimens, but nearly 65% have less than 25,000 specimens.

James C. Solomon

Bibliography. A. Cronquist, *An Integrated System of Classification of Flowering Plants*, 1992; L. Forman and D. Bridson (eds.), *The Herbarium Handbook*, 3d ed., 1998; P. K. Holmgren, N. H. Holmgren, and L. C. Barnett, *Index Herbariorum*, 8th ed., 1990.

Herbicide

Any chemical used to destroy or inhibit plant growth, especially of weeds or other undesirable vegetation. The concepts of modern herbicide technology began to develop about 1900 and accelerated rapidly with the discovery of dichlorophenoxyacetic acid (2,4-D) as a growth-regulator-type herbicide in 1944–1945. A few other notable events should be mentioned. During 1896–1908, metal salts and mineral acids were introduced as selective sprays for controlling broad-leafed weeds in cereals; during 1915–1925 acid arsenical spray, sodium chlorate, and other chemicals were recognized as herbicides; and in 1933–1934, sodium dinitrocresylate became the first organic selective herbicide to be used in culture of cereals, flax, and peas. Since the introduction of 2,4-D, a wide variety of organic herbicides have been developed and have received wide usage in agriculture, forestry, and other industries. Today, the development of highly specific herbicides that are intended to control specific weed types continues. Modern usage often combines two or more herbicides to provide the desired weed control. Worldwide usage of herbicides continues to increase, making their manufacture and sale a major industry.

The control of weeds by means of herbicides has provided many benefits. Freeing agricultural crops from weed competition results in higher food production, reduced harvesting costs, improved food quality, and lowered processing costs, contributing to an abundant supply of low-cost, high-quality food. Not only are billions of dollars saved through increased production and improved quality, but costs of labor and machinery energy for weed control are reduced, livestock is saved from the effects of poisonous weeds, irrigation costs are reduced, and insect and disease control costs are decreased through the removal of host weeds for the undesirable organisms.

Additional benefits due to appropriate herbicide use result as millions of people are relieved of the suffering caused by allergies to pollens and exposure to poisonous plants. Recreational areas, roadsides, forests, and parks have been freed of noxious weeds, and home lawns have been beautified. Herbicides have reduced storage and labor costs and fire hazards for industrial storage yards and warehouse areas. Modern herbicides even benefit the construction industry, where chemicals applied under asphalt prolong pavement life by preventing weed penetration of the surface.

Classification. There are well over a hundred chemicals in common usage as herbicides. Many of these are available in several formulations or under several trade names. The many materials are conveniently classified according to the properties of the active ingredient as either selective or nonselective. Further subclassification is by method of application, such as preemergence (soil-applied before plant emergence) or postemergence (applied to plant foliage). Additional terminology sometimes applied to describe the mobility of postemergence herbicides in the treated plant is contact (nonmobile) or translocated (mobile—that is, killing plants by systemic action). Thus, glyphosate is a nonselective, postemergence, translocated herbicide.

Selective herbicides are those that kill some members of a plant population with little or no injury to others. An example is alachlor, which can be used to kill grassy and some broad-leafed weeds in corn, soybeans, and other crops.

Nonselective herbicides are those that kill all vegetation to which they are applied. Examples are bromacil, paraquat, or glyphosate, which can be used to keep roadsides, ditch banks, and rights-of-way open and weed-free. An important use for such chemicals is the destruction of vegetation before seeding in the practice of reduced tillage or no tillage. Some are also used to kill annual grasses in preparation for seeding perennial grasses in pastures. Additional uses are in fire prevention, elimination of highway hazards, destruction of plants that are hosts for insects and plant diseases, and killing of poisonous or allergen-bearing plants.

Preemergence or postemergence application methods derive naturally from the properties of the herbicidal chemical. Some, such as trifluralin, are effective only when applied to the soil and absorbed into the germinating seedling, and therefore are used as preemergence herbicides. Others, such as diquat, exert their herbicidal effect only on contact with plant foliage and are strongly inactivated when placed on contact with soil. These can be applied only as a postemergence herbicide. However, the distinction between pre- and postemergence is not always clear-cut. For example, atrazine can exert its herbicidal action either following root absorption from a pre- emergence application or after leaf absorption from a postemergence treatment, and thus it can be used with either application method. This may be an advantage in high-rainfall areas where a postemergence treatment can be washed off the leaf onto the soil and nevertheless can provide effective weed control.

Herbicidal action. Many factors influence herbicide performance. A few are discussed below.

Soil type and organic matter content. Soils vary widely in composition and in chemical and physical characteristics. The capacity of a soil to fix or adsorb a preemergence herbicide determines how much will be available to seedling plants. For example, a sandy soil normally is not strongly adsorptive. Lower quantities of most herbicides are needed on a highly adsorptive clay soil. A similar response occurs with the organic portion of soil; higher organic matter content usually indicates that more herbicide is bound and a higher treatment rate necessary. For example, cyanazine is used for weed control on corn at the rate of 1.5 lb/acre (1.7 kg/ha) on sandy loam soil of less than 1% organic matter, but 5 lb/acre (5.6 kg/ha) is required on a clay soil of 4% organic matter.

Leaching. This refers to the downward movement of herbicides into the soil. Some preemergence herbicides must be leached into the soil into the immediate vicinity of weed seeds to exert toxic action. Excessive rainfall may leach these chemicals too deeply into the soil, thereby allowing weeds to germinate and grow close to the surface.

Volatilization. Several herbicides in use today will volatilize from the soil surface. To be effective, these herbicides must be mixed with the soil to a depth of 2 to 4 in. (5 to 10 cm). This process is termed soil incorporation. Once the herbicide is in contact with soil particles, volatilization loss is minimized. This procedure is commonly employed with thiocarbamate herbicides such as EPTC and dinitroaniline herbicides such as trifluralin.

Leaf properties. Leaf surfaces are highly variable. Some are much more waxy than others; many are corrugated or ridged; and some are covered with small hairs. These variations cause differences in the retention of postemergence spray droplets and thus influence herbicidal effect. Most grass leaves stand in a relatively vertical position, whereas the broad-leafed plants usually have their leaves arranged in a more horizontal position. This causes broad-leaf plants to intercept a larger quantity of a herbicide spray than grass plants.

Location of growing points. Growing points and buds of most cereal plants are located in a crown, at or below the soil surface. Furthermore, they are wrapped within the mature bases of the older leaves. Hence they may be protected from herbicides applied as sprays. Buds of many broad-leafed weeds are located at the tips of the shoots and in axils of leaves, and are therefore more exposed to herbicide sprays.

Growth habits. Some perennial crops, such as alfalfa, vines, and trees, have a dormant period in winter. At that time a general-contact weed killer may be safely used to get rid of weeds that later would compete with the crop for water and plant nutrients.

Application methods. By arranging spray nozzles to spray low-growing weeds but not the leaves of a taller crop plant, it is possible to provide weed control with a herbicide normally phytotoxic to the crop. This directed spray technique is used to kill young grass in cotton with MSMA or broad-leafed weeds in soybeans with chloroxuron in an oil emulsion. Another example of selective application technology is the use of wiper applicators. This apparatus commonly employs absorbent ropes which draw herbicide solution by capillary action from a reservoir. The equipment is arranged so that the ropes can pass above crop plants and transfer the herbicide solution to weeds taller than the crop by a wiping action. Use of a nonselective translocated herbicide in this recirculating sprayer system will then selectively remove the weeds from the crop. Glyphosate is being used commercially in such systems to remove johnson grass from cotton and soybeans.

Protoplasmic selectivity. Just as some people are immune to the effects of certain diseases while others succumb, so some weed species resist the toxic effects of herbicides whereas others are injured or killed. This results from inherent properties of the protoplasm of the respective species. One example is the use of 2,4-DB or MCPB (the butyric acid analogs of 2,4-D and MCPA) on weeds having β-oxidizing enzymes that are growing in crops (certain legumes) which lack such enzymes. The weeds are killed because the butyric acid compounds are broken down to 2,4-D or MCPA. Another example is the control of a wide variety of weeds in corn by atrazine. Corn contains a compound that conjugates with the atrazine atom, rendering it nontoxic; most weeds lack this compound. A third important example of protoplasmic selectivity is shown by trifluralin, planavin, and a number of other herbicides applied through the soil which inhibit secondary root growth. Used in large seeded crops having vigorous taproots, they kill shallow-rooted weed seedlings; the roots of the crops extend below the shallow layer of topsoil containing the herbicides and the seedlings survive and grow to produce a crop.

Properties. Several factors of the commercial herbicide influence selectivity to crops.

Molecular configuration. Subtle changes in the chemical structure can cause dramatic shifts in herbicide performance. For example, trifluralin and benefin are very similar dinitroaniline herbicides differing only in the location of one methylene group. However, this small difference allows benefin to be used for grass control in lettuce, whereas trifluralin will severely injure lettuce at the rates that are required for weed control.

Herbicide concentration. The action of herbicides on plants is rate-responsive. That is, small quantities of a herbicide applied to a plant may cause no toxicity, or even a slight growth stimulation, whereas larger amounts may result in the death of the plants. It has long been known that 2,4-D applied at low rates causes an increase in respiration rate and cell division, resulting in an apparent growth stimulation. At high application rates, 2,4-D causes more severe changes and the eventual death of the plant.

Formulation. The active herbicidal chemical itself is seldom applied directly to the soil or plants. Because of the nature of the chemicals, it is usually necessary that the commercial product be formulated to facilitate handling and dilution to the appropriate concentration. Two common formulations are emulsifiable concentrates and wettable powders. Emulsifiable concentrates are solutions of the chemical in an organic solvent with emulsifiers added which permit mixing and spraying with water. Wettable powders are a mixture of a finely divided powder, active chemical, and emulsifiers which allow the powder to suspend in water and to be sprayed. An additional ingredient called an adjuvant is sometimes added to the formulation or to the spray tank when the spray solution is mixed. These adjuvants are normally surface-active agents (surfactants) which improve the uniformity of spray coverage on plants and the plant penetration of the herbicide. An example is the addition of surfactant to paraquat spray solutions to improve its nonselective postemergence action on weeds. Another material sometimes added to spray solutions is a nonphytotoxic oil. These oils may be used to improve postemergence action of herbicides such as diuron which normally have limited foliar absorption. *See* PESTICIDE; SURFACTANT.

Rodney O. Radke

Bibliography. F. M. Ashton and T. J. Monaco, *Weed Science: Principles and Practices*, 3d ed., 1991; J. R. Crister, Jr., *Herbicides*, 1978; *Herbicide Handbook of the Weed Science Society of America*, 7th ed., 1994; P. C. Kearney and D. Kaufman (eds.), *Herbicides*: *Chemistry, Degradation, and Mode of Action*, 1975–1976; R. D. King, *Farmers Weed Control Handbook*, 1985; H. A. Roberts (ed.), *Weed Control Handbook: Principles*, 8th ed., 1990; W. T. Thomas, *Agricultural Chemicals: Herbicides*, 1993, Book II, 1993.

Herbig-Haro objects

Compact nebulae with peculiar spectra that were first studied in the 1950s by George Herbig and, independently, by Guillermo Haro. Herbig-Haro (HH) objects are typically located within or adjacent to star-forming regions and appear as small patches of nebulosity with highly variable luminosity. They exhibit spectroscopic emission lines that are characteristic of highly ionized gas that has been shocked upon interaction with the surrounding interstellar medium. Doppler shifts of the spectral lines reveal that the gas is moving at supersonic velocities of 100 km/s (60 mi/s) to more than 1000 km/s (600 mi/s). Typical masses of an individual clump of gas are 1–20 times that of Earth with temperatures of 8000–12000 K. They are composed almost entirely of hydrogen and helium with less than 1% consisting of heavier elements. Over 400 individual Herbig-Haro objects are known. Many of the closest ones are observed to emanate from a very young stellar source with proper motions that are discernible in *Hubble Space Telescope* images taken over periods of several years. *See* ASTRONOMICAL SPECTROSCOPY; DOPPLER SHIFT; HUBBLE SPACE TELESCOPE.

The discovery of bipolar outflows from young stars in the late 1970s and early 1980s revealed Herbig-Haro objects to be associated with jets of ionized gas that emanate from young "exciting" stars surrounded by circumstellar accretion disks (**Fig. 1**). When identified, the source stars are pre-main-sequence T Tauri stars or young protostars still hidden at optical wavelengths by a collapsing envelope of dust and gas. The ionized jets are typically highly collimated and oppositely directed along the rotational axis of a disk (**Fig. 2**). Millimeter-wave observations of carbon dioxide emission often reveal less collimated molecular outflows moving in the same general direction but at lower velocities. These are usually found to contain Herbig-Haro objects when studied in sufficient detail. Both ionized jets and molecular outflows may extend for several parsecs from their exciting sources, where they are distributed over a much greater range of angles than nearer to the source. *See* PROTOSTAR; T TAURI STAR.

It is now understood that bipolar jets and outflows responsible for Herbig-Haro objects probably play an important role in triggering continued star formation. A large number of irradiated Herbig-Haro objects are found in clustered star forming regions such as the Orion Nebula, where they deliver enough energy to substantially perturb the medium and could have been responsible for initiating gravitational collapse in the densest clumps of gas and dust. In addition, Herbig-Haro objects reveal centuries-old histories of jet emission from star-disk systems. Consequently, they are important laboratories for

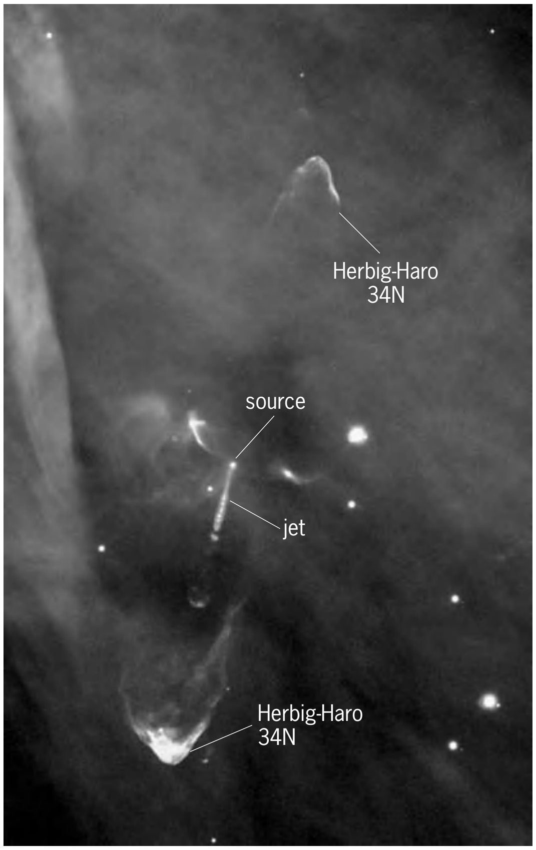


Fig. 1. Jet and bow-shock emission from Herbig-Haro 34. Two opposite jets run into the surrounding interstellar medium, where they terminate in Herbig-Haro objects with a characteristic bow-shock structure. The jets are composed of "bullets" of dense gas ejected at high velocities. (*European Southern Observatory*)

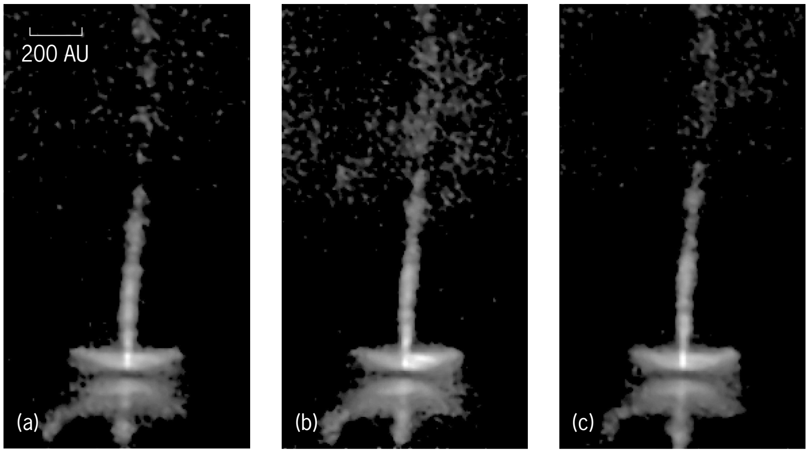


Fig. 2. Disk and jets associated with Herbig-Haro 30, showing changes between images taken in (*a*) 1995, (*b*) 1998, and (*c*) 2000 with the Wide Field and Planetary Camera 2 aboard the *Hubble Space Telescope*. A slightly flared disk of dust and gas appears edge-on at the bottom of the images. The jet and associated Herbig-Haro objects are believed to arise from complex interactions between the magnetic fields associated with accreting disk material and a rapidly rotating young star hidden from view in the surrounding dust and gas. (*NASA; Alan Watson, Universidad Nacional Autonoma de Mexico; Karl Stapelfeldt, Jet Propulsion Laboratory; John Krist, Space Telescope Science Institute; Chris Burrows, European Space Agency/Space Telescope Science Institute*)

star-formation studies, including the conditions prior to collapse and the evolution of outflow systems. *See* ORION NEBULA; STELLAR EVOLUTION. David Koerner

Bibliography. T. P. Ray, Fountains of Youth: Early days in the life of a star, *Sci. Amer.*, 283(4):42–47, August 2000; B. Reipurth and J. Bally, Herbig-Haro flows: Probes of early stellar evolution, *Annu. Rev. Astron. Astrophys.*, 39:403–455, 2001; B. Reipurth and S. Heathcote, Herbig-Haro Objects and the birth of stars, *Sky Telesc.*, 90(4):38–40, October 1995.

Herbivory

The consumption of living plant tissue by animals. Plants and the animals that consume them constitute roughly one-half of the scientifically described species. Herbivorous species occur in most of the major taxonomic groups of animals, including the vertebrates (such as grazing fish, tortoises, geese, and hoofed mammals), echinoderms (sea urchins), mollusks (snails and slugs), nematodes (roundworms), and arthropods (including crabs, lobsters, copepods, amphipods, and isopods; mites; and especially insects, such as beetles, caterpillars of moths and butterflies, larvae of many flies, sawflies, grasshoppers, aphids and their relatives, thrips, and stick insects). Herbivorous insects alone may account for one-quarter of all species. The fraction of all biomass produced by plants that is eaten by herbivores varies widely among plants and ecosystems, ranging from less than 1% to nearly 90%. Thus, in terms of both the number of species involved and the role that herbivory plays in the flow of energy and nutrients in ecosystems, herbivory is a key ecological interaction between species. *See* MARINE ECOLOGY.

Plant adaptations. Either because the herbivore is much smaller than the plant, as when caterpillars feed upon trees, or because part of the plant is inaccessible to the herbivore, herbivory usually does not kill the plant outright, although there are striking exceptions (such as bark beetle outbreaks that decimate conifer trees over thousands of square kilometers). Nevertheless, chronic attack by herbivores can have dramatic cumulative effects on the size, longevity, or reproductive output of individual plants. As a consequence, plants have evolved several means to reduce the level of damage from herbivores and to ameliorate the impact of damage.

Many plants possess physical defenses that interfere mechanically with herbivore feeding on or attachment to the plant. Spines and thorns deter feeding by large herbivores, smaller hairs on the leaf surface can jam insect mouthparts (and in some cases, sharp hooks or sticky resin on these hairs can entrap insects), and the waxy surface layer of plant leaves makes it difficult for insect herbivores to move around the plant without falling off. Mature oak leaves cannot easily be penetrated by the mandibles of newly hatched winter moth caterpillars, which in part explains the small size and low survival of caterpillars fed mature leaves. In addition to these physical defenses, plant tissues may contain chemical compounds that render them less digestible or even toxic to herbivores. For example, tannins and quinones bind to proteins in the food, reducing the nutritional quality of the consumed plant tissue. Many plant compounds (such as heart poisons in milkweed leaves and delphinine in larkspur foliage) can cause death if consumed by unadapted herbivores. While natural selection imposed by herbivores was the likely force driving the elaboration of these plant chemicals, humans have subsequently found many uses for the chemicals as active components of spices (capsacin, which makes cayenne pepper hot), stimulants (caffeine), relaxants (nicotine), hallucinogens (mescaline), poisons (strychnine), and drugs (ephedrine).

An exciting recent finding is that some plants possess induced resistance, elevated levels of physical or chemical defenses that are brought on by herbivore damage and confer enhanced resistance to further damage. Acacia trees grow branches with longer and denser thorns after they are browsed by giraffes. Damaged leaves of wild parsnip (*Pastinaca sativa*) contain higher concentrations of antiherbivore compounds, and caterpillars of the cabbage looper (*Trichoplusia ni*) grow far less rapidly when fed previously damaged leaves. Two reasons why plants may produce certain defensive chemicals only after damage (rather than always producing them) are that those chemicals may be partially toxic to the plant and that their production may involve a physiological cost. For example, wild parsnip leaves after damage have higher rates of respiration that can be accounted for entirely by the production of energy required to synthesize damage-induced antiherbivore compounds. Beetle-damaged soybean plants are initially less palatable to beetles than are undamaged plants, but after about 20 days the damaged plants actually become more palatable, a possible delayed cost of induced resistance. If damage is unpredictable, plants may save energy or scarce nutrients for other uses by defending heavily only those tissues that are highly likely to be attacked by herbivores, and producing defensive chemicals in other tissues only after they are damaged. Wild parsnip fruits, which are consistently attacked by herbivores, always contain high levels of defensive chemicals, whereas root tissues, which receive herbivore damage only rarely, produce high concentrations of chemical defenses only after being damaged.

Induced resistance may eventually be exploited for agricultural purposes. For example, populations of spider mites on grape vines that had received prior damage by another species of mite were substantially lower than populations on previously undamaged plants. Thus, rather than using pesticides, farmers may be able to "vaccinate" their crops with relatively benign herbivores to protect them from later-feeding herbivores that cause more damage. Induced resistance may also be a factor contributing to periodic outbreaks of herbivore populations, through the following sequence: A high herbivore population induces resistance and consequently declines, leading to reduced damage, a decline in the induced resistance as the plants regrow, and finally the return of rapid herbivore population growth to restart the cycle.

When defenses fail to prevent herbivore damage, plants can utilize several mechanisms to compensate for the effects of damage. Fireweed plants (*Epilobium latifolium*) whose growing shoot tips are destroyed by moth larvae grow longer side branches to replace the leaf tissue that they were prevented from producing on the damaged main stem. In one location, scarlet gilia plants whose growing tips were eaten by grazing mammals actually grew larger and produced more seeds than undamaged plants, but such beneficial effects of herbivory are unlikely to occur commonly, even for scarlet gilia across its entire range. Tissues that plants regrow after herbivore damage commonly have higher photosynthetic rates than undamaged tissues, which may hasten the process of compensating for the lost tissue. In at least two species of nonagricultural plants, families of related individuals have been shown to vary in their ability to maintain high seed production after receiving herbivore damage. Tolerance of herbivory in plants thus appears to have a genetic basis, and can evolve in response to herbivore pressure.

Herbivore adaptations. Herbivores can either avoid or counteract plant defenses. Many herbivores avoid consuming the plant tissues that contain the highest concentrations of toxic or antinutritive chemicals. Insects that feed on plants whose leaves contain secretory canals through which toxic substances are transported to sites of herbivore damage often sever the canals before feeding. Herbivores have also evolved an elaborate array of enzymes to detoxify otherwise lethal plant chemicals. Because few herbivores have the ability to detoxify the chemical compounds produced by all the plant species they encounter, many herbivores have restricted diets; the larvae of more than half of all species of butterflies and moths include only a single genus of plants in their diets. Some insect species that have evolved the means to tolerate toxic plant chemicals have also evolved ways to use them in their own defense. Larvae of willow beetles store plant compounds in glands along their back. When the larvae are disturbed, the glands exude droplets of the foul-smelling compounds, which deter many potential predators.

If a plant evolved the ability to produce a novel chemical compound that its herbivores could not detoxify, the plant and its descendants would be freed for a time from the negative effects of herbivory. A herbivore that then evolved the means to detoxify the new compound would enjoy an abundance of food and would increase until the level of herbivory on the plant was once again high, favoring plants that acquire yet another novel antiherbivore compound. These repeated rounds of evolution of plant defenses and herbivore countermeasures have been termed coevolution. Coevolution over long periods of time helps to explain similar patterns of evolutionary relatedness between groups of plant species and the herbivorous insect species that feed on them.

Ecological context. Plants and their herbivores seldom occur in isolation, and other species can influence the interaction between plants and herbivores. Plant defenses that slow the growth of insect larvae increase the insects' vulnerability to their predators. Many parasites use characteristic plant odors to help them find herbivorous host insects with restricted plant diets. Certain fungi that grow within the leaves of grasses are rather harmless to the plants but produce chemicals that deter feeding by insects and mammals, causing infected plants to outperform uninfected plants in the presence of herbivores. Mammalian herbivores often rely upon gut microorganisms to digest cellulose in the plant material they consume. Thus, herbivory occurs against a backdrop of multiple interactions involving the plants, the herbivores, and other species in the ecological community. *See* GRASS CROPS; POPULATION ECOLOGY.

William F. Morris

Bibliography. R. S. Fritz and E. L. Simms, *Plant Resistance to Herbivores and Pathogens: Ecology, Evolution, and Genetics*, 1992; R. Karban and I. T. Baldwin, *Induced Responses to Herbivory*, 1997; G. A. Rosenthal and M. R. Berenbaum, *Herbivores: Their Interactions with Secondary Plant Metabolites*, 1992; L. M. Schoonhoven, T. Jermy, and J. J. A. van Loon, *Insect-Plant Biology*, 1998.

Hermaphroditism

A condition in which components of both testes and ovaries are present in the same individual. Although true hermaphroditism is common among lower forms of animals such as annelids and mollusks, it is rare in humans. A more common condition in humans is pseudohermaphroditism, which simulates hermaphroditism. In female pseudohermaphroditism, or gynandry, the external sexual characteristics are in part or wholly of the male aspect, but internal female genitalia are present. In male pseudohermaphroditism, or androgyny, the individual has external sexual characteristics of female aspect, but has testes (usually undescended).

The four general forms of hermaphroditism are bilateral, the presence of an ovary and testis on each side; lateral, the presence of an ovary on one side and a testis on the other; ovatesticular, the presence of an ovatestis on one or both sides; and unilateral, the combination of an ovatestis on one side with an ovary or testis on the other.

During the fourth to sixth week of embryonic life the genital ridge differentiates into cortical and medullary components. If the embryo is a genotypic male (XY), its destiny is to be a phenotypic male; the cortical component atrophies while the medullary component develops into the seminiferous tubules. If the embryo is a genotypic female (XX), destined to be a phenotypic female, the medullary component atrophies and the cortical component persists to become the ovary. Any failure of the tissue components to develop or function, or any aberration of the chromosomes during embryogenesis, results in some degree of abnormal sexual characteristics; sexual alterations may also occur at puberty. Thus, variable phenotypic patterns of masculinization and feminization occur in accordance with genotypic determinants, and with the elaboration of hormones. *See* OVARY; REPRODUCTIVE SYSTEM DISORDERS; TESTIS.

Sybil P. Parker

Hernia

The protrusion of an organ, part of an organ, or other structure through the wall of the body cavity normally containing it. Various organs may be involved, including the bladder, brain, esophagus, intestine, ovary, and rectum. The most common location for a hernial bulge to appear is the abdominal wall, particularly the groin.

Among the most infrequent but life-threatening hernias is a cerebral hernia in which part of the brain protrudes through an opening in the skull.

A diaphragmatic hernia, which occurs when a defect is present in the muscular diaphragm separating chest from abdomen, may be present at birth or result from an injury later in life. Abdominal organs, such as the liver, spleen, stomach, and intestine, can pass through the diaphragmatic defect and lodge in the chest cavity, so that the lungs become compressed and breathing is impaired. Hiatal or esophageal hernia results when a portion of the stomach slides into the chest cavity through the normal diaphragmatic opening for the esophagus. Hiatal hernias occur principally in the elderly and primarily produce symptoms of indigestion or heartburn. Both diaphragmatic and hiatal hernias can be cured surgically.

There are various abdominal-wall hernias. Among the most common is an incisional or ventral hernia which occurs when an abdominal incision does not heal correctly or separates because of abnormal strain. The defect allows a protrusion of the intestine, necessitating surgical repair. Umbilical or navel hernias are usually congenital, representing a defect left by the passage of umbilical cord structures during fetal life. Unlike all other hernias, childhood umbilical hernias tend to close spontaneously. In adults, umbilical hernias may be caused by a number of predisposing factors, including pregnancy and abdominal distension due to an abnormal collection of fluid or obesity. Epigastric hernias are located strictly in the midline of the abdomen and above the navel. Both adult umbilical and epigastric hernias require little more than a simple surgical procedure for permanent cure.

A Spigelian hernia is located on the lateral abdominal wall, while a lumbar or dorsal hernia occurs through the posterior abdominal wall. Both types are quite rare and difficult to diagnose. They are associated with vague complaints of discomfort and an associated ill-defined mass. Operative repair is accomplished by the simple reapproximation of separated muscle and connective tissue.

Groin hernias consist of two major types, inguinal and femoral. Inguinal hernias account for 75% of all hernias of the body, and are divided into two anatomic variants, indirect and direct. In large inguinal hernias, the groin protrusion may stretch into the scrotum. Indirect inguinal hernias are caused by a weakness in the abdominal wall that corresponds to an area where the testis descended into the scrotum during embryological development. With direct inguinal hernias, the defect results mainly from strain on the abdominal muscles which have been weakened by age—hence the term “older man’s” hernia. Inguinal hernias are 10 times as common in men as in women. Femoral hernias are more common in women, but are infrequent. The weakness in a femoral hernia originates in the area where the major veins, arteries, and nerves pass from the abdomen into the lower extremities. A femoral hernia bulge is always located in the upper inner part of the thigh, just below the groin crease.

Clinical manifestations. A groin hernia develops when the outer layers of the lower abdominal wall weaken or actually separate. Through this defect, the inside lining or peritoneal membrane of the abdominal cavity gradually protrudes and forms a sac. The hernia sac contains displaced intestine or other structures. When combined with the outer layers of stretched-out muscle, connective tissue, and skin, a hernial bulge forms.

Development of a groin hernia usually results from chronic strain on the lower abdominal muscles which in turn causes general loss of tissue tone. Age and heredity also play important roles. Other risk factors include a chronic cough, constipation, generalized obesity, and pregnancy. Activities such as lifting a heavy object or sudden twists or pulls have long been regarded as instant causes of groin hernias, but do not reflect the complicated pathophysiology of groin herniation.

Most inguinal hernias cause little acute pain. They are usually associated with a variety of milder symptoms, including weakness or pressure in the groin, an aching discomfort in the region, a burning feeling in the area of the bulge, or the presence of intestinal gurgling.

A groin hernia is termed reducible if the protruding sac and its intestinal contents can be pushed back into place inside the abdominal cavity. However, upon standing or especially with exercise the bulge quickly reappears. If the sac cannot be manually reduced, the condition is termed an incarcerated hernia. In rare instances, the blood supply to the incarcerated intestine becomes pinched off. This complication is termed a strangulated hernia and is a surgical emergency. Rapidly worsening pain, occasional vomiting, and a firm tender bulge are signals that the herniated tissue has become strangulated and is in the process of turning gangrenous and dying. Within a few hours, this condition becomes life-threatening and requires immediate surgical attention.

Diagnosis and treatment. Most groin hernias are self-diagnosed. However, the presence of a visible bulge in the groin or scrotum is enough to warrant the attention of a surgeon. It is important that other disease entities, including accumulation of fluid in the testis (hydrocele), tumor of the lymph gland (lymphoma), and swelling of spermatic cord (varicocele), be ruled out.

With rare exceptions, all hernias should be corrected surgically to prevent the possibilities of incarceration, intestinal obstruction, and strangulation. Groin herniorrhaphy is typically an elective operation, performed on an ambulatory basis and over 98% successful. Surgical repair involves strengthening the area of weakness through one of many surgical techniques.

Groin hernia repair methods utilize tension-free techniques in which a small incision is made at the site of the hernia and a piece of plastic mesh is inserted to cover the area of the abdominal-wall defect. There is no tension because the surrounding tissues are not sewn together. For example, a conical mesh, termed a hernia plug, can be applied so that it adheres to the confines of the hernia defect, much like placing a cork in a bottle. The mesh is safe and well tolerated by the body’s natural tissues. Such modern herniorrhaphies take approximately 20 min to perform, recovery is swift, and the likelihood of the hernia recurring is less than 2%. Also, since these methods of groin hernia repair cause little postoperative discomfort, individuals are able to resume most normal activities in 2–3 days. Ira M. Rutkow

Bibliography. I. M. Rutkow, *Hernia Surgery*, 1993; G. E. Wantz, *Atlas of Hernia Surgery*, 1991.

Herpes

Any virus of the herpesvirus group, which comprises a family of 70 species, 5 of which are pathogenic to humans; the term also refers to any infection caused by these viruses. Since these pathogens are ubiquitous in nature, most individuals of all populations are exposed to and thus immunized to these viruses. The five pathogenic groups include herpes simplex I and II, varicella-zoster, cytomegalovirus, and the Epstein-Barr virus.

In nonimmunized hosts, the vast majority of all herpes infections are subclinical; that is, they do not cause specific symptoms or signs attributable to herpes infections, but present symptoms of nonspecific viral illnesses which resolve spontaneously. However, the infections that cause clinical disease in fact may cause serious morbidity and mortality in afflicted individuals. A whole range of problems accompanies infections, including the mental anguish of the victims of the venereal disease caused by herpes simplex II (genitalis), and the great concern of public health officials over the epidemic proportions of this disorder. Reactivation of herpes infection, characteristic of the immunocompromised host, is an important cause of mortality in the treatment of patients with advanced cancer, and is a dreaded potential complication of an otherwise possibly curable systemic disease. The outlook for the treatment of patients afflicted with serious herpesvirus infections becomes brighter as potential treatments for several herpesvirus disorders are identified.

Characteristics of the virus. All known viruses are composed of a core of nucleic acids and a protein barrier, the capsid, which protects the gene structure. Herpesviruses have a deoxyribonucleic acid (DNA) core and are 150 to 200 nanometers in size with icosahedral symmetry, and are coated by an envelope derived from the infected host cells. The surface of the virions in general contains protein-carbohydrate structures which allow cellular attachment and thus cellular penetration. All viruses require living cells for their replication, since they carry no metabolic machinery themselves, but rely upon the host-cell metabolic processes. The virus may replicate and destroy the cell, or replicate and allow cell survival, or incorporate its viral gene structure into the host gene structure.

This incorporation phenomenon is designated as latency, and, for example, herpes simplex virus exhibits the phenomenon of latency within nerve cells in the area of previous infection. The Epstein-Barr virus characteristically causes latent infection in lymphocytes (white blood cells in the circulating blood), and the cytomegalic virus also causes latent infection within lymphocytes and possibly within nerve cells. Once the viral genome is incorporated into the host cell, antiviral drugs are of no use, since therapeutic agents cannot selectively destroy or inhibit the viral genome. Viral biologic avenues of infection are still not completely understood. Factors which are possibly involved in the reactivation of latent virus generally revolve around some depression of the host immune response system, and this may be mediated in part through damage by irradiation, chemical agents, mechanical injuries, hormones, emotions, or superimposed infections. Viral genome incorporation into host cells is of great interest as several herpesvirus types are implicated in the development of cancer (herpes simplex II with uterine cervical cancer, Epstein-Barr virus with Burkitt's lymphoma and nasopharyngeal carcinoma). *See* VIRUS INFECTION, LATENT, PERSISTENT, SLOW.

The fact that viruses require host-cell machinery for replication implies that metabolites can be used to compete with or inhibit the assembly of viral DNA. This is the foundation of therapeutic intervention for all herpesviruses, which involves a series of chemicals with structures similar to the base pairs which compose the DNA structure. The base analogs compete with or inhibit viral enzymes necessary for the assembly of viral DNA. The agents are useful in selected clinical situations for given age groups and virus types, and have been only for investigational use in the prophylaxis of certain herpes-type infections. *See* VIRUS.

Herpes simplex I and II. These infections are spread by intimate contact of mucocutaneous surfaces during the period of virus shedding from active lesions. Clinical infections involving herpes type II usually affect the genitalia, but may affect the oral mucosa, causing exquisitely painful ulcerations which crust and heal. Upon healing, the virus resides in latent form within local nerve cells. Factors related to viral reactivation are poorly understood, but may relate in part to the host immune system. The type II virus has been linked to the development of uterine cervical carcinoma, since women with primary genital herpes display roughly a 12-fold greater risk than women without primary genital herpes. Also, women with uterine cervical cancer have greater blood antibody levels against herpesvirus type II than controls. Uterine cervical cancer often is preceded by an abnormal growth of cervical cells known as dysplasia, and a similar correlation between incidence of dysplasia and the presence of herpes antigens in these precancerous cells is established. Other factors which predispose to the development of cervical cancer include the beginning of sexual intercourse at an early age and multiple sexual partners. Thus, in part, cervical cancer is a venereal disease, and this lends support to the notion that a transmissible agent is involved. However, the precise role of herpes type II remains a question, as the virus may act as a cocarcinogen along with some other agent in the cervix, or in fact may be an incidental phenomenon which tends to occur with greater frequency in incipient cancers of the cervix. *See* SEXUALLY TRANSMITTED DISEASES.

Active herpes lesions in prepartum women are an indication for cesarean section, since neonatal herpes may be acquired through the birth canal and may result in serious neonatal morbidity and mortality.

Herpes simplex virus I (cold sores, fever blisters) afflicts 20–40% of the population in the United States and usually affects the oropharynx, causing pharyngitis, tonsillitis, gingivostomatitis, or keratitis (eye inflammation) as primary infections. Inflammation of the mouth, eye, or brain may occur as a secondary infection.

Acyclovir is recommended for the treatment of primary genital herpes, but no therapy is recommended for recurrent herpes genitalis. Also, no therapy is recommended for primary labial herpes, but Acyclovir may be used in high-risk groups, and topical triflurothymidine may be used for keratitis.

Herpes varicella-zoster. Primary infection (airborne) due to varicella-zoster usually affects preschool children, causing chickenpox, a systemic cutaneous vesicular skin eruption. Rare complications include encephalitis, hepatitis, and pneumonia, with the complications usually affecting the immunocompromised host. Secondary infection usually afflicts the elderly when latent viral reactivation occurs, presumably due to an immune imbalance in the host, and involves the spread of virus along the skin in the anatomic distribution of nerve (this disorder is known as shingles). Infection in the first trimester of pregnancy may cause fetal anomalies, and infection of the mother near time of delivery may cause neonatal herpes, which can be treated with hyperimmune serum containing antibodies to herpes varicella-zoster. Vidarabine is recommended as therapy in the compromised host or high-risk patient. *See* CHICKENPOX AND SHINGLES.

Cytomegalovirus. Like all herpes, cytomegalovirus is ubiquitous, with the majority of infections remaining subclinical. Adult syndromes which are well recognized include a mononucleosislike syndrome and hepatitis, both of which are self-limited diseases in the normal host. However, reactivation of latent infection is a major source of morbidity and especially mortality in the compromised host, for example, the patient being treated with chemoradiotherapy for advanced malignant disease. Disseminated cytomegalic disease in these patients is often fatal. *See* HEPATITIS.

Since rubella infections in pregnancy have been controlled, cytomegalovirus has become one of the most important serious pregnancy infections as serious fetal sequela are common, for example, prematurity, malformations, mental retardation, or death. There is no recommendation for drug therapy, although investigational drugs are in use. *See* CYTOMEGALOVIRUS INFECTION.

Epstein-Barr virus. The characteristic clinical syndrome caused by Epstein-Barr virus infection includes generalized lymphadenopathy, hepatosplenomegaly, pharyngitis, tonsillitis, and general fatigue and fever. This disorder affects individuals of all ages, but predominantly adolescents. The majority of children are subclinically infected. This mononucleosis syndrome is usually a self-limited disorder, and investigational drugs in use for prophylaxis of high-risk individuals include interferons and acyclovir. Epstein-Barr virus is suspected to be of etiologic importance in Burkitt's African lymphoma, in which the genomes of Epstein-Barr virus are routinely displayed in tumor cells of lymphocytes. However, the Epstein-Barr genome is not found in the non-African cases of Burkitt's lymphoma. Virological and immunological data support an etiological link of Epstein-Barr virus and a cancer of the nasopharynx. *See* ANIMAL VIRUS; EPSTEIN-BARR VIRUS; VIRUS CHEMOPROPHYLAXIS.

David J. Dabbs

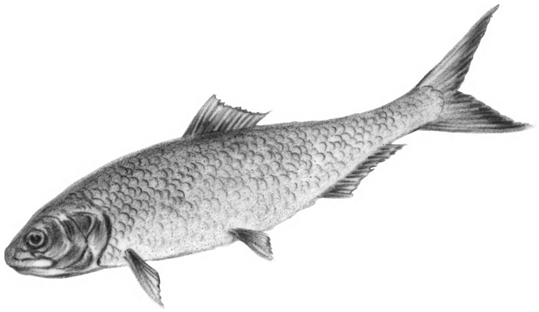
Bibliography. D. L. Ablashi (ed.), *Epstein-Barr Virus and Human Diseases*, 1987; D. A. Baker, *Acyclovir Therapy for Herpes Virus Infection*, 1990; E. Lycke and E. Norrby (eds.), *Textbook of Medical Virology*, 1983; S. Robbins and R. Cotran, *Pathologic Basis of Disease*, 6th ed., 1999; W. C. Russell and J. W. Almond (eds.), *Molecular Basis of Virus Disease*, 1987; D. O. White and F. Fenner, *Medical Virology*, 4th ed., 1994.

Herrings

The common name for a family (Clupeidae) of about 70 genera of fishes in the order Clupeiformes. Used extensively as food all over the world, they occur in all seas except the Arctic and Antarctic. Many live in vast shoals and are caught in great numbers.

These fishes are the most primitive of the higher bony fishes. The fins have no supporting spines and are soft-rayed. There are usually four gill clefts, with the pectoral fins behind the gill openings. Scales are present on the body but absent on the head, and the swim bladder and lateral line may be missing. *See* SWIM BLADDER.

Some species live permanently in fresh water, while many species are anadromous, spending most of their life in the sea and spawning in fresh water. When spawning, the female extrudes the eggs in estuaries while rubbing against seaweeds, rocks, or other objects to which the eggs adhere. The male swims closely behind the female and fertilizes the eggs. The herrings along the North American shores breed once a year, while those in the seas along the Scandinavian coast breed twice yearly, in the spring and autumn. Herrings mature at the age of 4 years, but are not suitable for spawning until about 5–8 years. After spawning, mature herrings disperse to their feeding grounds and congregate in large schools. Mature fish return to their specific spawning grounds during the breeding season. Some herrings live 20 years.



A shad, typical representative of the herring family.

The herring *Clupea harengus* has a circumpolar distribution. About eight other species of this genus are recognized, including the sprat or brisling (*C. sprattus*), which occurs in the Mediterranean and seas of western Europe, and the gizzard shad (*Dorosoma cepedianum*), which is a common species in the Potomac River. In Europe the herring is either salted, pickled, or smoked and cured as kippers. In Canada and the United States young herring are canned as "sardines."

Sardine. This herring, *Sardina pilchardus*, is known commonly as the pilchard and is found along the European coasts in the Atlantic. *Sardina sagax* occurs along the coasts of Japan, Chile, California, and South Africa. Pilchards occur in large schools, feed on pelagic organisms, and spawn in the summer. The entire fish may be processed and preserved in oil, since the bones are soft and all parts are edible.

Shad. These herrings, members of the genus *Alosa*, occur in northern waters on both sides of the Atlantic (see **illus.**). They spawn in fresh water but spend most of their life in the sea. They breed in rivers from spring to early summer, and the fry remain in fresh water for a year or so. The allis shad (*A. alosa*) occurs along the coasts of Europe and may reach a length of 30 in. (75 cm) and weigh 8 lb (3.6 kg). The American shad (*A. sapidissima*) is valued for food.

Anchovies. A family of herringlike fish, the Engraudidae together with the Clupeidae belong to the suborder Clupoidea. They are found in the Mediterranean and range along the European coast as far north as Norway. Spawning occurs in the summer; the female lays eggs which are sausage-shaped and transparent, and float near the surface. Within 3 days the minute, transparent larvae hatch. *See* CLUPEIFORMES.

Charles B. Curtin

Hertzsprung-Russell diagram

A two-dimensional diagram used extensively in astronomy, developed independently by Ejnar Hertzsprung in 1911 and Henry Norris Russell in 1913. In its original form, the Hertzsprung-Russell (H-R) diagram was a plot of absolute visual magnitude versus spectral type (O, B, A, and so on). Variants are now commonly used, avoiding requirements of and uncertainties due to spectral classification. The vertical axis of the diagram is some suitable measure of the power output of the star, while the horizontal axis indicates the temperature (or color) of the star's visible surface, or the corresponding spectral type. Each point in the plot represents a nearby star of known distance. In any of its forms, the diagram reveals the most fundamental correlation among observed stellar properties discovered to date. *See* MAGNITUDE (ASTRONOMY); SPECTRAL TYPE.

Observational and theoretical forms. Different quantities and units may be used for the two axes of the Hertzsprung-Russell diagram depending on whether it relates observational or theoretical stellar properties. Observationally, a star's power output is estimated from measurements made in a specific photometric (wavelength) band, with resulting absolute magnitude determined from the measured apparent magnitude using a knowledge of the star's distance. In its observational form, also referred to as a color-magnitude diagram, absolute magnitude is used as ordinate, although apparent magnitude may be used for a collection of stars at a common distance. Brighter stars (that is, those with higher luminosities, and smaller numerical values of the magnitude) appear at the top of the diagram. The color scale is usually a color index constructed as the difference between the magnitudes measured in two chosen spectral bands. For historical reasons, the bluest color index (corresponding to the highest temperatures) appears at the left. *See* COLOR INDEX.

In its theoretical form, the Hertzsprung-Russell diagram relates more physically fundamental stellar quantities, which can be predicted or modeled from theories of stellar structure and evolution. In this case the ordinate gives the total bolometric magnitude, or the bolometric luminosity, which measure the star's total power output, usually on a logarithmic scale normalized to the luminosity of the Sun. An effective temperature is used as the theoretical counterpart to the color index. Closely related to the temperature of the stellar surface, the effective temperature is tied to the luminosity and stellar radius through the Stefan-Boltzmann law. Consequently, lines of constant stellar radius run diagonally across the Hertzsprung-Russell diagram, from top left to bottom right, with stellar radius increasing from bottom left to top right. *See* HEAT RADIATION.

Theoretical modeling describes stellar structure and evolution in terms of fundamental properties such as mass, luminosity, surface gravity, and effective temperature. These properties may in principle be transformed into observable quantities such as magnitudes and color indices. It is the comparison between observed and theoretical quantities which allows stellar structure and evolutionary theories to be tested and refined. As such, the Hertzsprung-Russell diagram provides a very powerful framework for the verification and development of theories of star formation and evolution, ultimately leading to a rigorous foundation for exploring and explaining the evolutionary history of the Milky Way Galaxy as a whole.

There are numerous complications in transforming between theoretical and observed quantities. A major problem in deriving the power output is that a star's apparent magnitude measured at a telescope must be converted into an absolute magnitude, requiring an accurate knowledge of the star's distance. Direct distance determinations in astronomy are based on trigonometric parallax

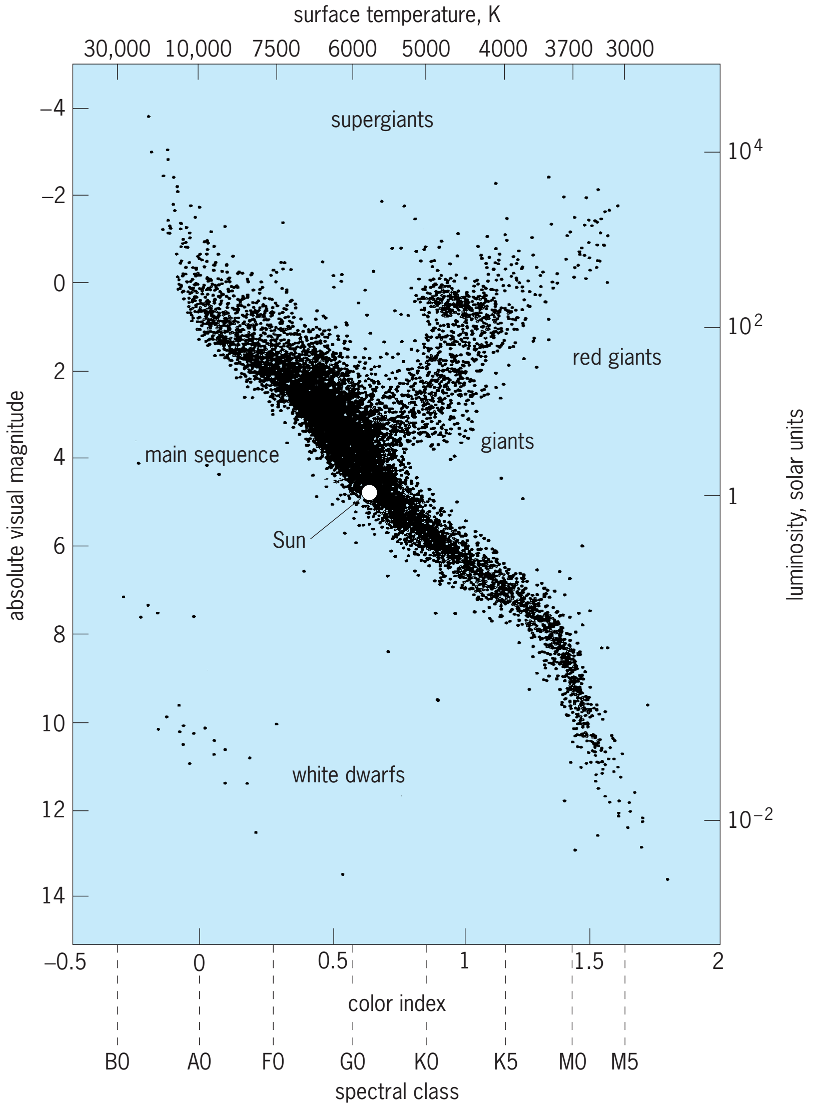


Fig. 1. Hertzsprung-Russell diagram for about 15,000 stars within a sphere of radius 100 parsecs, taken from the *Hipparcos Catalogue*. The color index and absolute visual magnitude scales are directly measured. The spectral class, surface temperature, and luminosity (in terms of solar luminosity) are approximate relationships appropriate for the main sequence.

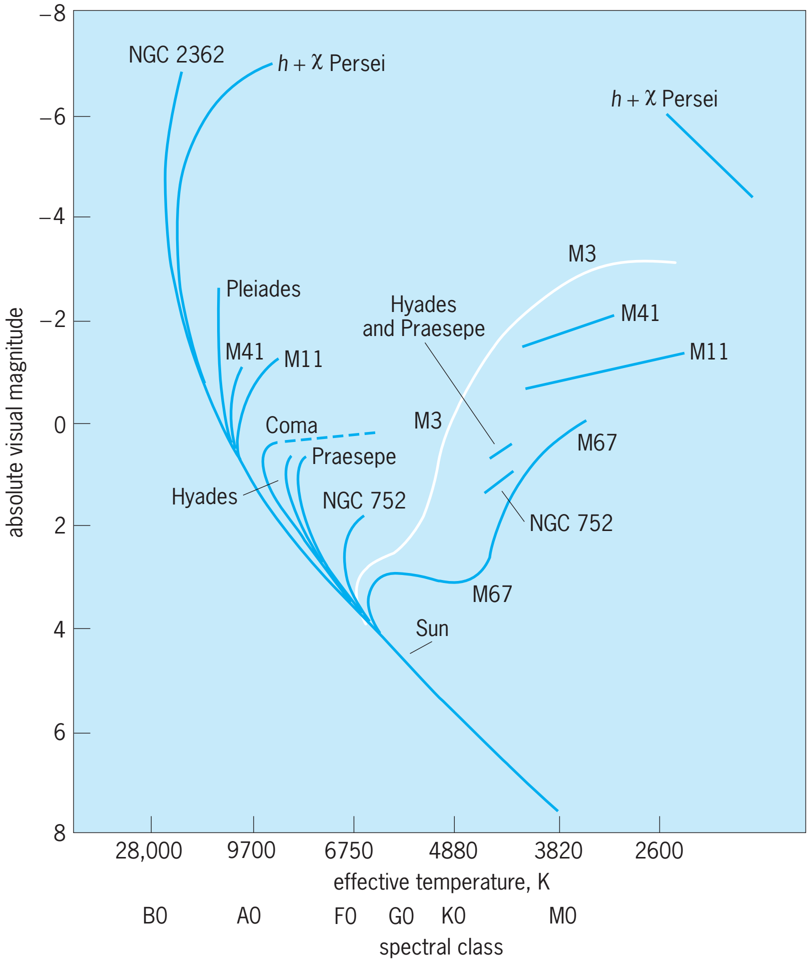


Fig. 2. Composite Hertzsprung-Russell diagram for several representative galactic clusters. NGC 2362 is the youngest and M67 the oldest cluster shown in the diagram, with ages of about 2×10^6 and 5×10^9 years, respectively. (*After A. Sandage, Observational approach to evolution, I. Luminosity functions, Astrophys. J., 125:435–444, 1957*)

measurements, which employ the motion of the Earth around the Sun to provide an estimate of the star's distance by triangulation. Ground-based parallax determinations are complicated by atmospheric motions and other disturbances, and such distance estimates have provided accuracies better than about 20% for only a few hundred nearby stars. The *Hipparcos* satellite, which was devoted to stellar distance measurements from space, has allowed accurate determination of individual stellar distances for many tens of thousands of stars out to distances of a few hundred light-years from the Sun, providing a significant clarification in the structure of the observational Hertzsprung-Russell diagram. *See* PARALLAX (ASTRONOMY).

General features. **Figure 1** shows the Hertzsprung-Russell diagram for about 15,000 single stars from the compilation of nearly 120,000 stellar distances measured by the *Hipparcos* satellite. The sample has been limited to those with relative distance errors smaller than 10%, and lying within a sphere of radius 100 parsecs (326 light-years, 3.1×10^{15} km, or 1.9×10^{15} mi), which is around 1% of the distance from the Sun to the center of the Galaxy. The figure illustrates the basic features of the classical Hertzsprung-Russell diagram. The absolute visual magnitude scale runs from −5 to 15, corresponding to a range of 10^8 in star luminosity. The color index scale corresponds to effective temperatures ranging from around 100,000 K (180,000°F) at the left to about 2500 K (4000°F).

From the upper left (blue, high-luminosity stars) to the lower right (red, low-luminosity stars) a prominent concentration of objects defines the main sequence. Stars located on the main sequence are also called dwarfs. They include stars such as Sirius, and are assigned luminosity class V in the MK stellar classification system. (In this system, two parameters, spectral type and luminosity class, categorize each star.) Along the main sequence, the luminosity of a star and its surface temperature are tightly correlated. Stellar structure theory successfully models this relationship. The main-sequence stars are at the early phases of their lives, and are powered by the fusing of hydrogen to helium in their centers. Masses of the main-sequence stars increase going toward the upper left of the diagram (reaching almost 100 times the Sun's mass) and decrease going to the lower right (to about one-tenth of the Sun's mass). Due to their higher central temperatures and pressures, the more massive stars are burning hydrogen more rapidly and are therefore brighter. The main sequence has a certain width, which arises from numerous contributions: remaining uncertainties in the distance and magnitude measurements, stellar variability, the presence of undetected binary systems, evolutionary changes, and variations in the chemical abundances of the stars. *See* BINARY STAR; DWARF STAR; VARIABLE STAR.

Extending from the main sequence in the direction toward lower temperatures, and at roughly constant luminosity, are the luminosity class III giant stars (such as Vega) and the clump of more luminous red giants. Even more luminous supergiants, of luminosity class I (such as Arcturus and Procyon), are sparsely represented but occupy a broad range of color index at the very highest luminosities. They reach absolute magnitudes of less than −5, corresponding to luminosities some 10^4 times brighter than the Sun, and with radii around 1000 times that of the Sun. The lower left part of the diagram is not entirely empty and contains the white dwarfs: hotter than the Sun, but much less luminous (typically 10^4 times fainter) and of much smaller radius (about 1% of the Sun's radius). *See* GIANT STAR; SUPERGIANT STAR; WHITE DWARF STAR.

The distribution of stars in the Hertzsprung-Russell diagram based on observations such as those in Fig. 1 gives a somewhat incomplete impression of the occurrence of stars of different spectral type and luminosity class for two reasons, related to the luminosity and to the space density of different types of stars. First, such measurements generally extend only down to some observational limit in apparent magnitude, around a visual magnitude of 12 in the case of the *Hipparcos* satellite observations. In any given volume of space, say one extending to a radius of 100 parsecs, the intrinsically low-luminosity stars, such as the white dwarfs and the fainter red main-sequence stars, are actually too faint to be observable farther out, even though their space density may be very high. Second, when the Hertzsprung-Russell diagram is constructed using well-defined limits on the accuracy of the distance measurements, the very luminous supergiants, which are relatively rare but can nevertheless be seen in considerable numbers out to very large distances, are poorly represented. For the same reason, other relatively uncommon stars, such as the hot but subluminous nuclei of planetary nebulae, which would appear between the main sequence and the white dwarf sequence, are not represented in Fig. 1. To determine distances (and hence luminosities) for rarer types of stars farther out in the Milky Way Galaxy, a range of alternative, indirect methods must be adopted. Their corresponding locations in the Hertzsprung-Russell diagram, the interpretation of their physical properties and, ultimately, a complete and unified picture of stellar and galactic evolution rest heavily on the confidence that can be placed in these distance estimates.

Stellar evolution. The Hertzsprung-Russell diagram says nothing, at least directly, about the mass, chemical composition, or age and state of evolution of a star. However, comparisons between observations (such as Fig. 1) and the predictions of stellar evolution theory allow stringent constraints to be placed on models of the structure, chemical composition, and evolution of stars. In the pre-main-sequence evolutionary phase, protostars form from collapsing material contained in interstellar clouds, liberating gravitational potential energy, and increasing in mass and temperature to the point where thermonuclear fusion of hydrogen into helium becomes an important energy source. At this point, the star settles onto the main sequence, its precise position depending primarily on its mass but also on its chemical composition. The subsequent evolution, in which hydrogen is slowly consumed in the stellar core, has such a long duration that most visible stars are found in this phase. The Sun, for example, is roughly halfway through its estimated main-sequence lifetime of about 10^{10} years. When the supply of hydrogen in the central region is depleted, structural adjustments occur, making the star expand and decreasing the temperature of its outer layers. The star becomes a red giant, and the point representing it on the Hertzsprung-Russell diagram moves up and to the right of the main sequence. Other changes occur at successive evolutionary stages and depend upon the mass, chemical composition, and other detailed properties. In particular, high-mass stars evolve at a much faster rate than low-mass stars and exhaust their nuclear fuel over a much shorter interval of time. *See* PROTOSTAR.

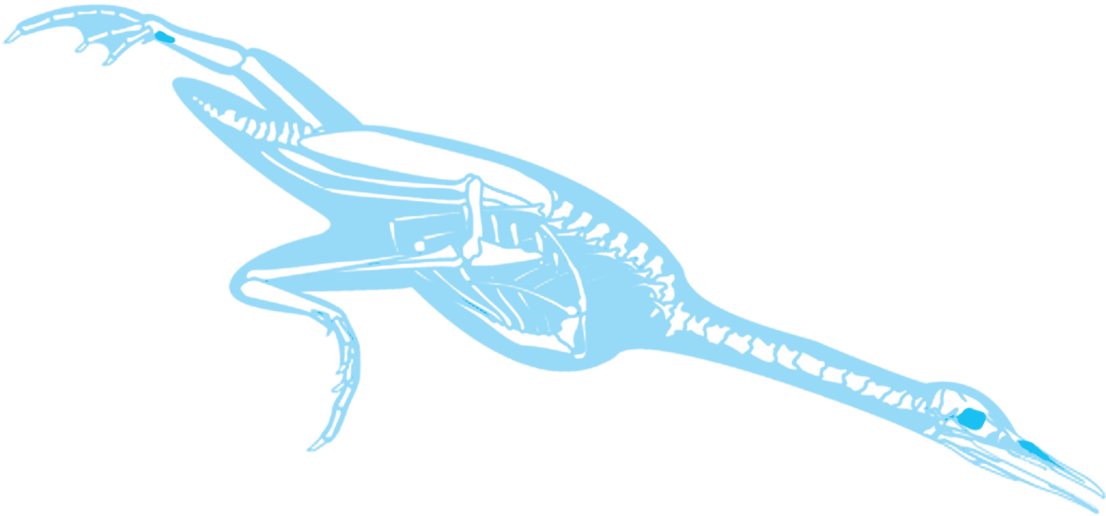
The effects of evolution are particularly evident in open clusters, physical aggregations of typically several hundred stars, in which all objects are essentially at the same distance from the Sun. Due to the common distance of the objects, the detailed features of the diagram are preserved when working in terms of apparent magnitude, even though the zero point of the absolute magnitude scale may be undetermined by the observations. Clusters are important laboratories for testing theories of star evolution, since all stars in a cluster are believed to have been formed at the same time. In a cluster of a given age, all stars brighter than a given limit will have evolved off the main sequence (**Fig. 2**). Thus, while the youngest clusters, such as NGC 2362, still contain hot O stars, progressively older clusters contain no stars bluer than, or to the left of, a certain turn-off point. As the cluster ages, the turn-off point moves down the main sequence to stars of progressively lower mass. In this way, the age of a star cluster can be inferred from its appearance of its Hertzsprung-Russell diagram. *See* STAR CLUSTERS.

Interpretation is further complicated by the effects of the star's initial chemical composition. Stars with the same mass but of different chemical composition do not lie at the same location in the Hertzsprung-Russell diagram. Stars categorized by astronomers as population II (for example, globular cluster stars) are old objects, having estimated ages nearly equal to that of the universe itself. These objects formed early in the history of the Milky Way Galaxy, with low abundances of elements heavier than hydrogen and helium, and a chemical composition nearly reflecting that of the primordial material of the universe. In contrast, open clusters are population I objects, having chemical abundances similar to that of the Sun. Their constituent stars were formed from material already enriched by the products of earlier phases of star formation, and have a wide range of ages, dating anywhere from the initial formation of the Milky Way Galaxy up to the present time. *See* MILKY WAY GALAXY; STAR; STELLAR EVOLUTION; STELLAR POPULATION. Michael A. C. Perryman

Bibliography. S. P. Maran, *The Astronomy and Astrophysics Encyclopedia*, John Wiley, 1997; S. Mitton, *The Cambridge Encyclopedia of Astronomy*, Cambridge University Press, 1984; F. H. Shu, *The Physical Universe: An Introduction to Astronomy*, University Science Books, 1982.

Hesperornithiformes

An extinct group of Cretaceous toothed birds. All members were foot-propelled, seagoing birds that could reach the size of a modern Emperor penguin. The largest known hesperornithiform reached a maximum length of over 5 ft (1.5 m). All of the known fossils are from marine rocks that are Cretaceous in age (about 100–65 million years old); they are most abundant from North America, especially Kansas. Although these birds are thought to be closely related to living birds (the group or clade Neornithes)—along with Ichthyornithiformes—the evolutionary interrelationships among Hesperornithiformes are not well understood by paleontologists. *See* AVES; CRETACEOUS; ICHTHYORNITHIFORMES; NEORNITHES.



***Hesperornis*, a marine hesperornithiform.**

Fossil record. Hesperornithiforms are known from all over the Northern Hemisphere—North America, England, Sweden, Russia, Kazakhstan, and Mongolia—and are preserved in rocks that record shallow, epicontinental seas. As these birds had only rudimentary wings, it is known that they were flightless; the miniscule forelimbs may have been used for steering when swimming. Soft-tissue preservation has revealed that the feathers of Hesperornithiformes were filamentlike, much like many flightless birds today.

Description. Hesperornithiforms had highly reduced, nonfunctional wings. They probably spent almost all their time in the water, except presumably in the breeding and egg-laying season. Like modern-day cormorants, hesperornithiforms appear to have led an aquatic lifestyle, probably diving to catch fish.

One of the best-known hesperornithiforms is *Hesperornis* (see **illustration**). When this bird was first described from the Late Cretaceous rocks of Kansas in the 1870s, it caused a great deal of excitement around the world. *Hesperornis* was touted as a modern bird, living at sea, and some 50 million years younger than its counterpart (and the only other fossil bird known at the time), *Archaeopteryx* ("ancient-wing") from the Jurassic of Germany (about 150 million years old). *Hesperornis* caused a stir because, unlike *Archaeopteryx*, it had lost its ability to fly, but like *Archaeopteryx*, still possessed teeth. The lower jaw (dentary) and the back portion of the upper jaw (maxilla) of *Hesperonis* had many sharp teeth. It is likely that they swam and fed much like modern penguins, preying on small fish. *See* ARCHAEORNITHES.

Extinction. Hesperornithiformes are one of the several lineages of nonmodern birds (along with, for example, the Ichthyornithiformes) that disappear from the fossil record at the end of the Cretaceous. No records of these birds are from the succeeding geological period, the Tertiary, although this was the time of the massive modern avian evolutionary radiation. *See* EXTINCTION (BIOLOGY).

Gareth Dyke

Bibliography. S. Chatterjee, *The Rise of Birds*, Johns Hopkins Press, Baltimore, 1997; L. M. Chiappe and L. M. Witmer, *Mesozoic Birds: Above the Heads of Dinosaurs*, University of California Press, Berkeley, 2002; A. Feduccia, *The Origin and Evolution of Birds*, 2d ed., Yale University Press, New Haven, 1999; S. L. Olson, *The Fossil Record of Birds*, in *Avian Biology*, vol. 8, Academic Press, New York, 1985.

Heterochrony

An evolutionary phenomenon that involves changes in the rate and timing of development. As animals and plants grow from their earliest embryonic stages to the adult, they undergo changes in shape and size. This life history of an individual organism is known as its ontogeny. The amount of growth that an organism experiences during its ontogeny can be more or less than its ancestor. This can apply to the organism as a whole or to specific parts.

Evolution can be viewed as a branching tree of modified ontogenies. Heterochrony that produces these changes in size and shape may be the link between genetics at one extreme and natural selection at the other.

Processes. If a character of one species in an evolutionary sequence undergoes less growth than its ancestor, the process is known as pedomorphosis. If it undergoes more growth, the process is known as peramorphosis. Each state can be achieved in three basic ways, by varying the timing of onset, offset, or rate of development.

Pedomorphosis. If development is stopped at an earlier growth stage in the descendant than in the ancestor (for example, by earlier onset of sexual maturity), ancestral juvenile features will be retained by the descendant adult (progenesis). If the onset of development of a particular structure is delayed in a descendant, the structure will develop less than in the ancestor (postdisplacement). The third process that produces pedomorphosis is neoteny, whereby the rate of growth is reduced. Few, if any, organisms grow isometrically. Different morphological features grow with varying allometries; some are positive, some negative. Changing growth rates change the allometric coefficients and produce a structure in the descendant that is different in shape from that in the ancestor.

Peramorphosis. The three ways that this type of growth can be achieved are just the opposite at pedomorphosis. Development can start earlier in the descendant than in the ancestor (predisplacement); or the rate of development can be increased, thus increasing the allometric coefficient (acceleration); or development can be extended by a delay in the onset of sexual maturity (hypermorphosis). In such cases, ancestral allometries are extended, producing descendant adults that are morphologically quite different from their ancestors.

Heterochrony can affect both the differentiation of meristic characters (differentiative heterochrony), such as, the timing of induction of limb bud growth in vertebrates or the formation of spines in a sea urchin, and also affect the structure once it has been formed (growth heterochrony). The complex interplay between these two aspects of heterochrony may have played a crucial role in the evolution of a number of major evolutionary novelties.

Cellular level. As an organism grows, the numbers of cells that it produces increases. Ultimately, changes to rate and timing of growth are reflections of changes to the timing of onset and rate of cell development, and the balance between cell growth and cell death. Morphogens and growth hormones play a major role in controlling development, in terms of initiation, rate of division, and migration. Therefore, changes to the timing of their expression affect the shape and size of the final adult structure. Inception of hormonal activity is under the control of genes that regulate the timing of its production.

Variations between individuals or species in the timing of induction of cellular development are manifested in differentiative heterochronic changes. Changes to the rates of cell division are reflected in growth heterochronies, affecting the shape and size of specific morphological structures. Any slight perturbations to cellular differentiation early in embryonic development can have major effects on the resultant adult form. Growth heterochronies occurring later in ontogeny produce relatively minor morphological differences. In addition, cell size can be a factor in determining heterochronic processes. Pedomorphic salamanders possess larger cells than nonpedomorphic forms. Likewise, trends toward the evolution of pedomorphic species of lungfishes over the last 350 million years are accompanied by an increase in cell size.

Variation within species. Much intraspecific variation within animals or plants involves slight differences in the shape or size of morphological features, or overall body size, and is likely to have arisen by heterochrony. The importance of change in growth proportions during ontogeny is seen by comparing the growth of a dog skull with that of a cat skull. In terms of overall body shape and proportions, breeds of dogs exhibit a far greater range of morphologies than do breeds of domestic cats. This greater range of morphologies arises from the much lower proportionate growth rate of the cat skull.

A number of species of insects show polymorphism, whereby a number of morphologically quite distinct forms occur within a single species. Many of these polymorphisms are due to heterochrony. For example, parthenogenetic thrips are always males and morphologically more advanced than nonparthenogenetic forms. This condition is accompanied by progenesis, that is, precocious onset of sexual maturity. Whether a female produces such young is determined by environmental factors, in this case photosensitivity that results in the production of a neurosecretory brain hormone, virginoparin. Aphids that lack wings are also an example of pedomorphosis. This condition is also attained by progenesis, and such forms experience rapid reproduction, produce more offspring, and have shorter generation times.

Sexual dimorphism in animals often arises from heterochrony. This frequently involves variations in body size, arising from either differences in maturation times or different growth rates. Thus, in *Homo sapiens* the generally smaller adult female body size arises from an earlier onset of sexual maturity, juvenile growth trajectories prior to maturation being similar in the two sexes.

Variations between species. Heterochrony is a crucial factor in speciation, providing the raw material upon which natural selection operates. Many examples of speciation, both from the fossil record and from living examples, are the product of heterochrony. Species are usually a product of a mosaic of heterochronic characters. Some features demonstrate a higher level of development than in the ancestor by peramorphic processes; others developed less by pedomorphic processes. *See* MOSAICISM; SPECIATION.

A classic example of evolution is the numerous species of finches first described from the Galápagos Islands by Charles Darwin. These finches possess a wide variety of bill shapes that have allowed a wide range of food types to be utilized. Some bills are used for crushing; others are for probing or biting. Such adaptations are generally regarded as having evolved in order for the species to feed on a particular object. However, heterochrony played a crucial role in the evolution of these bills. Because of small changes to the rates of growth of upper and lower bills, some reducing in size and others growing more and becoming much more robust, a wide range of species has evolved. If suitable food sources were available that could be exploited by a particular bill shape, this shape would be selected for and a new species would evolve.

Thus the evolution of species arises from the natural selection of particular morphologies that have arisen by heterochronic changes to the rates and timing of growth, and that confer an adaptive advantage to the organism. Because such adaptive changes often result in changes to life history strategies, resultant behavioral changes can in some instance be considered to have arisen ultimately from heterochrony.

Evolutionary trends. Many evolutionary trends documented from the fossil record were fueled by heterochrony. Through a succession of species, certain traits that evolved are shown to be either increasingly pedomorphic or increasingly peramorphic. Such trends are called, respectively, pedomorphoclines and peramorphoclines and have evolved along a particular environmental gradient. Many examples have been described from the fossil record in trilobites, echinoids, ammonoids, brachiopods, and bivalves.

A pedomorphocline develops when an ancestral species that passes through a number of ontogenetic stages (say, A to M) gives rise to a descendant species that passes through fewer (say, stages A to K). As the adult state K occurred in the ancestral juvenile, the descendant species is pedomorphic. If this species subsequently gives rise to a pedomorphic species that passes through only stages A to I, a pedomorphocline has been established. A peramorphocline exists when successive younger species pass through successively more morphological stages during their ontogeny. Often a mosaic of heterochronic characters occurs in such trends, with some traits forming a pedomorphocline, and others forming a peramorphocline, in the same sequence of species.

Major evolutionary novelties. While groups such as flightless birds, amphibians, and humans are thought to have arisen by pedomorphosis, it is more likely that a critical combination of pedomorphic and peramorphic characters (mosaic heterochrony) was more important in causing such drastic evolutionary changes. For instance, the evolution of horses through the Cenozoic in North America shows trends of increased body size (by hypermorphosis), while changes to the structure of the leg arose by a pedomorphic reduction in digit number, with a peramorphic increase in growth of the one remaining digit to produce a hoof.

Human development has for a long time been ascribed to the pedomorphic retention of many primate juvenile characters. However, while a few features, such as dental development, may be pedomorphic, the most distinctive features of the human species are relatively long life-span, long juvenile growth period, large brain size, and large body size. These are peramorphic features produced by a sequential hypermorphosis, wherein the time spent in successive preadult periods is extended, relative to the same time periods in other primates. *See* ANIMAL EVOLUTION. Kenneth J. McNamara

Bibliography. S. J. Gould, *Ontogeny and Phylogeny*, 1977; B. K. Hall, *Evolutionary Developmental Biology*, 1998; M. L. McKinney (ed.), *Heterochrony in Evolution: A Multidisciplinary Approach*, 1988; M. L. McKinney and K. J. McNamara, *Heterochrony: The Evolution of Ontogeny*, 1991; K. J. McNamara (ed.), *Evolutionary Trends*, 1990.

Heterocorallia

A small, extinct late Paleozoic order of fossil corals known from Europe, North Africa, Asia, North America, and Australia but limited to the Middle Devonian and to the Late Mississippian. They are found in calcareous shales and in limestones. The preserved material, called a corallite, is a calcareous substrate secreted by a scleractinian-like polyp. The corallites attained elongated shapes (**Fig. 1***a*) and show a bifurcational pattern of septal insertion during growth in the diameter. Nevertheless, the known materials show little change in diameter from one end to the other. There are rare indications of branching, which can be interpreted as budding of new polyps in the soft tissues of the original polyp. *See* DEVONIAN; MISSISSIPPIAN.

Each corallite is a framework of septa (radial longitudinal elements) and tabulae (horizontal skeletal elements). Each such element consists of fibrous calcium carbonate; the fibers are perpendicular to the growth lamellation, which is weakly apparent in this section (Fig. 1*b–d*). An external sheath, or epitheca, is absent or very rare.

The septa and tabulae of heterocorals justify their classification in the subclass Zoantharia of the class Anthozoa. However, the pattern of septal insertion indicates that they are a separate group (that is, a separate evolutionary lineage) aside from the Tabulata, Rugosa, and Scleractinia.

As interpreted by O. H. Schindewolf, there are four original septa. Each of these may split into two near the periphery; new septa are inserted only in the four spaces formed between the two split portions of each original septum. Of these four insertion spaces, two opposing ones regarded as lateral show

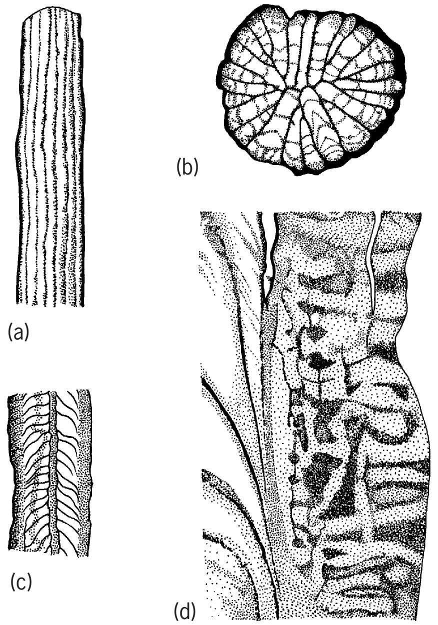


Fig. 1. Morphology of the heterocorallian skeleton. (*a*) External view. (*b*) Transverse thin section. (*c*) Median longitudinal thin section. (*d*) Enlarged part of c. (*After R. C. Moore, ed., Treatise on Invertebrate Paleontology, pt. F, Geological Society of America, University of Kansas Press, 1956*)

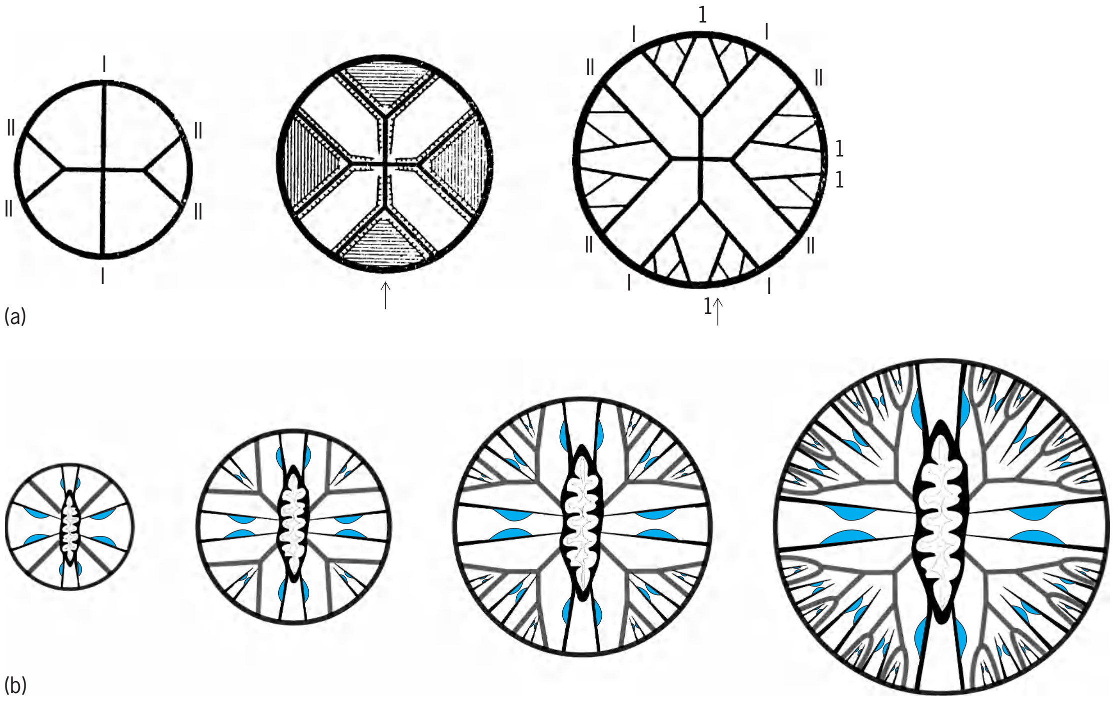


Fig. 2. Details of the heterocorallian body. (*a*) Generalized pattern of septal insertion (*from O. Schindewolf, Grundfragen der Paläontologie*, *Schweizerbart*, *Stuttgart, 1950*). (*b*) Reconstruction of the mesenterial in the soft body (*from M. Gudo, Konstruktion, Evolution und riffbildendes Potential rugoser Korallen*, *Courier Forschungsinstitut Senckenberg*, *228:1–153, 2001*).

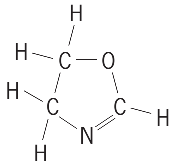
far more newly formed septa than the remaining two. This pattern of septal insertion, and the knowledge of the general functional design of cnidarian polyps, allows the reconstruction of the soft body of the heterocorals (**Fig. 2**). The soft body of the polyp must have had six original twins of mesenteries and four growth zones in which new twins of mesenteries were inserted in an exponential manner. As a result, the septa which mold the arrangement of mesenteries were formed in a bifurcational pattern.

The order contains only one family, the Heterophylliidae, with two genera, *Hexaphyllia* and *Heterophyllia*. *See* CNIDARIA. Michael Gudo; Dorothy Hill

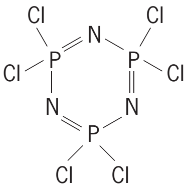
Bibliography. B. Berkowski, Famennian Rugosa and Heterocorallia from southern Poland, *Palaeontologia Polonica*, 61:3–88, 2002; M. Gudo, Konstruktion, Evolution und riffbildendes Potential rugoser Korallen, *Courier Forschungsinstitut Senckenberg*, 228:1–153, 2001; M. Montenari, U. Leppig, and D. Weyer, Heterocorallia from the Early Carboniferous of the Moldanubian Southern Vosges Mountains (Alsace, France), *Neues Jahrbuch für Geologie und Paläontologie, Abhandlungen*, 224(2):223–254, 2002; O. Schindewolf, *Grundfragen der Paläontologie*, Schweizerbart, Stuttgart, 1950; C. Teichert (ed.), *Treatise on Invertebrate Paleontology*, pt. F, suppl. 1, 1986.

Heterocyclic compounds

Cyclic compounds in which the rings include at least one atom of an element different from the rest. Homocyclic compounds are cyclic compounds in which all the ring atoms are the same. In organic homocyclic compounds the annular atoms are all carbons. If the molecule contains carbon atoms, then it is organic (most types of heterocyclic compounds studied to date are organic compounds). An example of an organic heterocyclic compound is oxazoline (**1**), which is composed of five annular atoms, two of which are not carbon. An example of an inorganic heterocyclic compound is phosphonitrilic chloride (**2**); the six-membered ring is composed of alternating atoms of nitrogen and phosphorus.

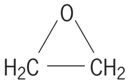


(**1**)



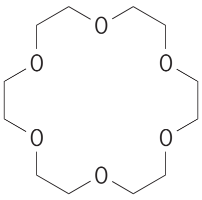
(**2**)

The smallest possible ring is three-membered, for example, ethylene oxide (**3**), but very large rings are



(**3**)

possible, as in the crown ethers, for example, 18-crown-6 (**4**). Thus, considerable diversity in the ring

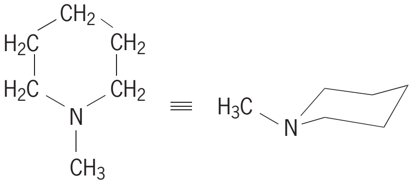


(**4**)

system is possible since not only may the size of the ring vary, but also the nature and number of heteroatoms in the ring can vary. The cycle may contain only single bonds and is thus saturated; it may include one or more double bonds; or it may possess aromatic unsaturation characteristics of benzene, that is, it is heteroaromatic. As with the homocyclic compounds, heterocyclic compounds can contain more than one ring, either heterocyclic or homocyclic.

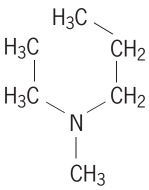
The more familiar heteroelements in an organic compound are oxygen, nitrogen, and sulfur, but many elements have been or could be incorporated into a heterocyclic system. Compounds containing a ring formed by internal hydrogen bonding are not classed as heterocyclic. There is no limitation as to what kind of heteroatom may participate in ring formation, provided it has the appropriate bond angles and geometry.

Properties. The chemical properties of saturated heterocyclic compounds are usually very similar to those of appropriate open-chain analogs, provided account is taken of conformational differences. For example, the chemistry of *N*-methylpiperidine (**5**)



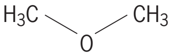
(**5**)

is not very different from that of ethylmethylpropylamine (**6**). Differences in properties originate in



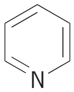
(**6**)

stereochemical factors. One such factor concerns four- and three-membered heterocycles in which the ring bonds are distorted with respect to bond angle and bond length, just as are those in cyclobutane and cyclopropane. These strained heterocyclic systems show enhanced activity in processes involving ring opening. Thus ethylene oxide (**3**), although formally an ether, is far more reactive than its open-chain analog, dimethylether (**7**).



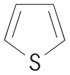
(**7**)

On the other hand, some polyunsaturated heterocyclic systems exhibit chemical properties implied in the term aromatic. For example, pyridine (**8**),



(**8**)

the parent of six-membered heteroaromatic nitrogen compounds, is a thermally stable material resisting oxidation, undergoing aromatic substitution (particularly nucleophilic substitution—the ring nitrogen atom makes electrophilic substitution difficult and the electrophile usually attacks the nonbonded pair of electrons on the ring nitrogen atom), and possessing appreciable resonance energy. Thiophene (**9**) is



(**9**)

a five-membered sulfur heterocycle, but it has aromatic properties similar to those of benzene. Such aromatic character may be expected in planar, unsaturated rings in which six electrons derived from ring unsaturation and from available electron pairs of the heteroatoms are delocalized over the whole ring system.

Nomenclature. Heterocyclic compounds may be named systematically. Many heterocycles, however, have nonsystematic names that are usually preferred by practicing chemists over the systematic ones. In the systematic approach to nomenclature the ring size is denoted by the appropriate stem. For example, three-membered saturated rings without nitrogen would have a name ending in -irane. The nature of the heteroatom is denoted by such prefixes as oxa-, thia-, or aza-, for oxygen, sulfur, or nitrogen, respectively. Thus, ethylene oxide (**3**) becomes oxirane. A five-membered unsaturated ring would have a name ending in -ole. Thiophene (**9**) thus becomes thiole. A six-membered unsaturated ring containing nitrogen would have a name ending in -ine according to this scheme. Pyridine (**8**) then becomes azine. Actually, the trivial names for these three systems are commonly accepted, and the systematic names are not often used. The degree of unsaturation is specified in the suffix as indicated in the **table**. It is important to note that the suffix is slightly modified when nitrogen is absent from the heterocyclic ring. The numbering of the ring begins with the heteroatom of highest priority and proceeds around the ring so as to give other heteroatoms or substituents the lowest number possible. The **illustration** shows how the systematic naming rules are applied and also how the ring atoms are numbered. An apex at which no

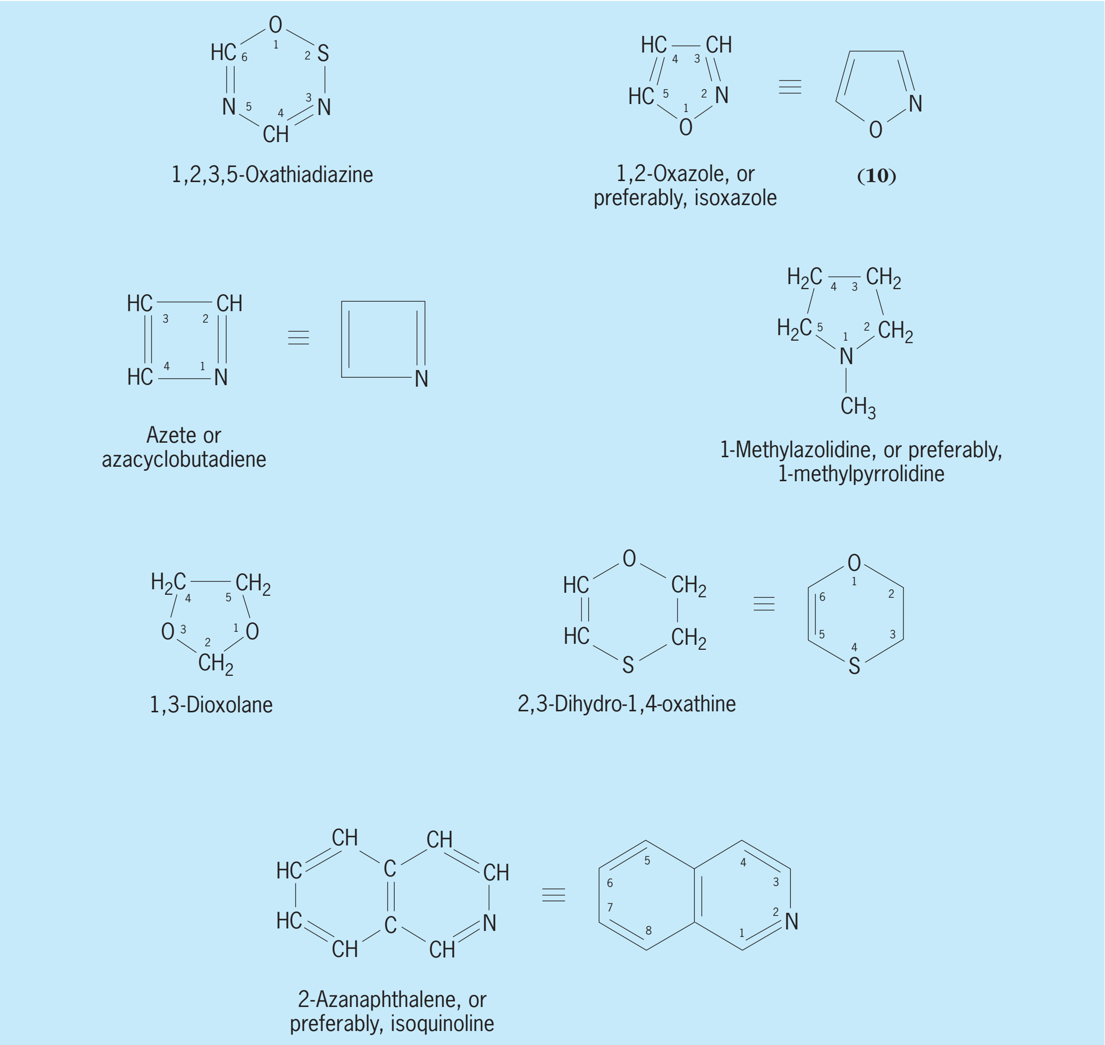
Suffixes for degree of unsaturation

	Number of double bonds in ring							
	With nitrogen				Without nitrogen			
Ring size	Max.	2	1	0	Max.	2	1	0
3			irine	iridine			irine	irane
4		ete	etine	etidine		ete	etene	etane
5	ole		oline	olidine	ole		olene	olane
6	ine	*	*	*	in	*	*	inane or ane
7	epine	*	*	*	epin	*	*	epane

* Add prefix, such as dihydro and tetrahydro, to the name of the ring with maximum unsaturation.

specific atom is shown [for example, the equivalent structure (**10**) for isoxazole] represents ═CH—.

Natural and synthetic compounds. Heterocyclic compounds are encountered in a very large number of groups of organic compounds. Inorganic heterocyclic compounds are also very common, but have not been studied systematically in terms of their heterocyclic characteristics in most cases. Naturally occurring heterocyclic compounds are extremely common as, for example, most alkaloids, sugars, vitamins, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), enzymic cofactors, plant pigments, many of the components of coal tar, many natural pigments (such as indigo, chlorophyll, hemoglobin, and the anthocyanins), antibiotics (such as penicillin and streptomycin), and some of the essential amino acids



1,2,3,5-Oxathiadiazine

1,2-Oxazole, or preferably, isoxazole

(**10**)

Azete or azacyclobutadiene

1-Methylazolidine, or preferably, 1-methylpyrrolidine

1,3-Dioxolane

2,3-Dihydro-1,4-oxathine

2-Azanaphthalene, or preferably, isoquinoline

Structural formulas of heterocyclic compounds. The ring positions are numbered.

(for example, tryptophan), and many of the peptides (such as oxytocin). Some of the most important naturally occurring high polymers are heterocyclic, including starch and cellulose. The major groups of natural products that are not mainly heterocyclic are the fats and most of the terpenes, steroids, and essential α-amino acids, though exceptions do exist.

Important synthetic dyes, such as phthalocyanins and phthaleins, as well as a large number of drugs, poisons, and medicinal products (both natural and synthetic) are heterocyclic. These include sulfothiazole, pyrethrin, strychnine, most of the antihistamines, the ergot alkaloids, morphine, the barbiturates, the tranquilizers Librium and Valium, and dihydrocannabinol (the main active constituent in marijuana). Many synthetic polymers (such as polyvinylpyridine, polyvinylpyrrolidone, and melamineformaldehyde), a number of industrial solvents (such as pyridine, dioxane, and tetrahydrofuran), and certain bulk chemicals (ethylene oxide, propylene oxide, and several aromatic nitrogen heterocycles obtained by synthesis from petroleum hydrocarbons or from coal tar) are heterocyclic.

For details about the specific heterocyclic systems *see* FURAN; INDOLE; PYRIDINE; PYRIMIDINE; PYRROLE; THIOPHENE. R. A. Abramovitch

Bibliography. D. J. Coffey (ed.), *Rodd's Chemistry of Carbon Compounds*, vol. 4, pt. G: *Heterocyclic Compounds*, 1977; G. P. Ellis, *Synthesis of Fused Heterocycles*, 1987; T. L. Gilchrist, *Heterocyclic Chemistry*, 3d ed., 1997; R. R. Gupta, *Physical Methods in Heterocyclic Chemistry*, 1983; A. R. Katritzky, *Advances in Heterocyclic Chemistry*, vols. 1–32, 1963–1982; A. R. Katritzky and C. W. Rees (eds.), *Comprehensive Heterocyclic Chemistry*: *The Structure*, *Reactions*, *Synthesis*, *and Uses of Heterocyclic Compounds*, 8 vols., 1984.

Heterocyclic polymer

Essentially, linear high polymers comprising heterocyclic rings, or groups of rings, linked together by one or more covalent bonds. As the search has continued for polymeric materials having useful properties at high temperatures (500°C or 930°F or higher), much attention has been given to heterocyclic polymers. The possibility of forming rigid molecules that can be ordered into anisotropic arrays having exceptional stiffness may become of even greater interest. As a group such polymers are often both mechanically rigid and inherently resistant to thermal degradation.

Some of these polymers form molecules in which the rings are fused together, as shown symbolically in the **illustration** (ladder polymers), and some form molecules in which fused rings are joined by single bonds (stepladder polymers). Similar considerations hold for simple aromatic systems (for example, linear polymers of benzene), but the heterocyclic systems have been more useful in application.

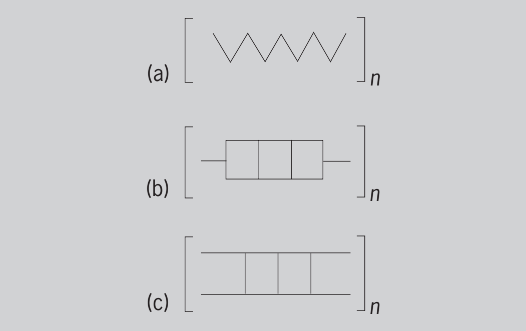
In practice, the heterocyclic resins have rather high molecular weights and high glass transition temperatures. Some cross-linking may also be introduced during curing. Because of the insolubility and infusibility of the unmodified polymers, processing and fabrication was originally accomplished in stages. First, soluble prepolymers were prepared and fabricated into the final form desired (film, molding, coating, impregnated glass cloth, and so on). In this stage, the heterocyclic rings are not yet closed. Closure of the rings by condensation reactions was then effected by heating, and volatile by-products were eliminated. In contrast, poly(amide-imide) and the unmodified resins can be processed by more conventional techniques.

Major applications for these polymers are as metal-to-metal adhesives and as laminating resins for fibrous composites for structural applications in the aerospace industry. Other applications requiring both strength and resistance to oxidation at elevated temperatures have developed, including valve seats, bearings, and turbine blades. *See* POLYMERIC COMPOSITE.

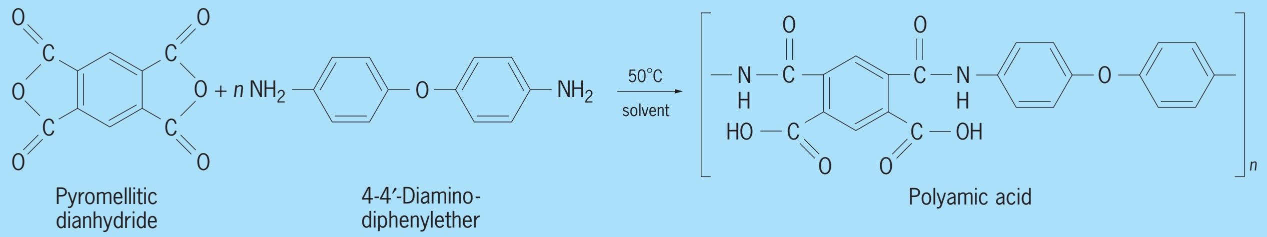
Polyimide resins. The basic synthetic reaction to form the prepolymer is the condensation of an aromatic dianhydride with an aromatic diamine. Thus, pyromellitic dianhydride may be added to 4,4′-diaminodiphenyl ether in an anhydrous medium to give a high-molecular-weight poly(amic acid), as shown in reaction (1). Solvents such as *N*,*N*-dimethylformamide are suitable media. The ingredients must be extremely pure, and must be present in equal molar amounts. Other dianhydrides and diamines may be used.

The prepolymer solution may then be used to impregnate fiber-glass cloth or other reinforcement, or applied to other substrates. Then, the solvent must be driven off and the rings closed by condensation, usually in stages [reaction (2)]. Frequently, most of the solvent is removed in a precuring stage and the major condensation effected later, at temperatures in the range 180–380°C (360–720°F). A subsequent postcuring step, also at elevated temperatures, may be used. The structure shown is of the stepladder type; ladder structures may be obtained by use of phenyl or fused-ring diamines.

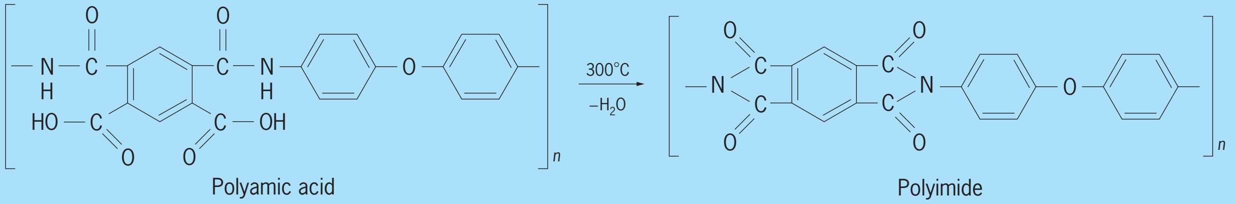
Properties depend on the structures of the ingredients and on the reaction and curing conditions. Fiberglass composites that are made by using



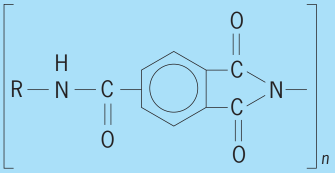
Structural units in linear polymers. (*a*) Simple linear polymer. (*b*) Stepladder polymer. (*c*) Ladder polymer.



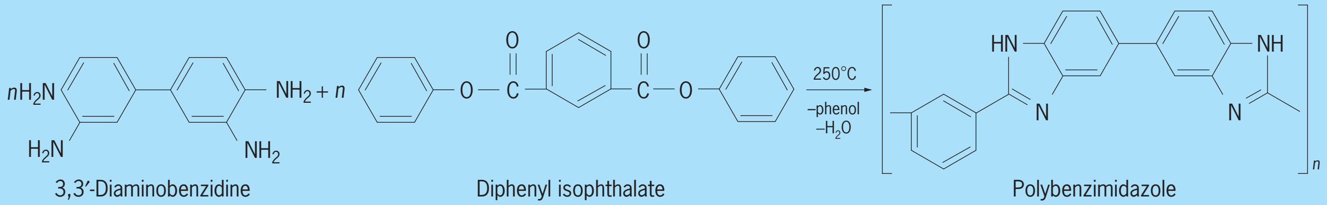
(1)



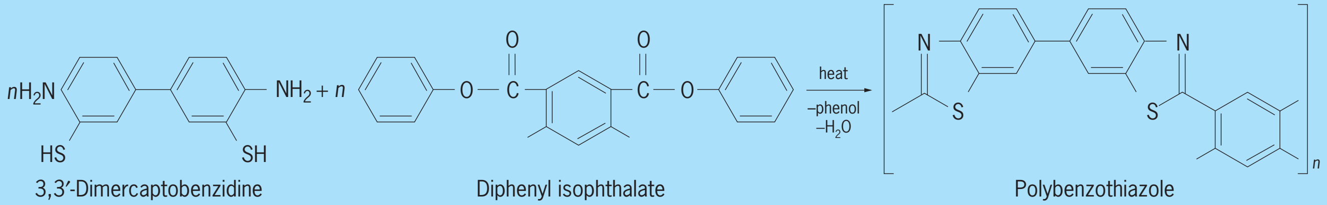
(2)



(3)



(4)



(5)

typical resins retain considerable strength after they are aged for 100 h at 500–600°C (930–1100°F). Adhesive bonds also show good resistance to aging at high temperatures. Certain polyimides are available as films and as molding powders, and some can be spun into fibers.

By introducing amide groups into a polyimide, some thermal stability is sacrificed, but improved processibility is gained. A typical structure is given in notation (3). Applications have developed in high-performance electrical connectors, engine parts, pumps, valves, and turbines.

Polybenzimidazoles. The basic reaction for the synthesis of aromatic polybenzimidazoles is the reaction of an aromatic tetramine with an aromatic diacid or diester. Although the details of the intermediate steps are not completely understood, the overall reaction for a typical example, the condensation of 3,3′-diaminobenzidine and the diphenylester of isophthalic acid, is as shown in reaction (4).

Polymerizations may be conducted in the melt, and to varying degrees of reaction. Partially polymerized resins may be dissolved in solvents such as *N*,*N*′-dimethylformamide and applied as a solution for laminating and adhesive applications. As with other heterocyclic polymers, final curing is effected at elevated temperatures (up to about 400°C or 750°F) under pressure.

Properties depend on the structures of the ingredients, and on reaction and curing conditions. In general, polybenzimidazoles are somewhat less stable in air at high temperatures than the polyimides. Applications are mainly as laminating and adhesive resins for composites and metals.

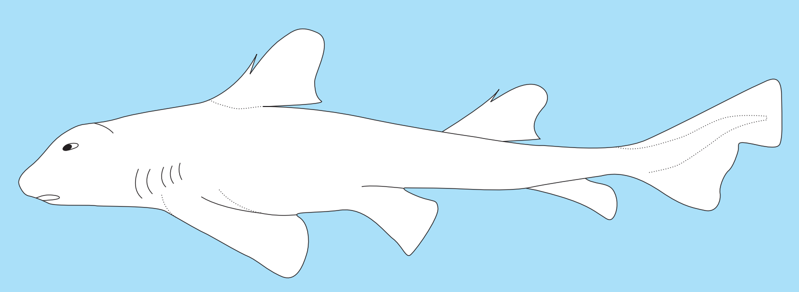
Polybenzothiazoles. Polybenzothiazoles are typically prepared by the reaction of a dimercaptobenzidine with an aromatic diacid, diester, or diacyl chloride, as shown in reaction (5).

A mainly soluble prepolymer is prepared by carrying the reaction forward to only a limited extent. As with the other resins in this family, the solvent is removed and the reaction is completed by heating after application to the substrate desired. Polybenzothiazoles find the same applications as other heterocyclic polymers and are intermediate in stability between polyimides and polybenzimidazoles. *See* POLYESTER RESINS; POLYETHER RESINS; POLYMER; POLY-P-XYLYLENE RESINS; POLYSULFONE RESINS. John A. Manson

Bibliography. H. F. Mark, *Encyclopedia of Polymer Science and Technology*, 12 vols., 3d ed., 2004; K. L. Mittal (ed.), *Polyimides and Other High Temperature Polymers: Synthesis, Characterization, and Applications*, 2005; G. Odian, *Principles of Polymerization*, 4th ed., 2004.

Heterodontiformes

An order of galeomorph sharks that are commonly known as the bullhead sharks. These sharks are distinguished by the following combination of characters: a typically sharklike (subcylindrical) trunk; a head that is elevated above the eyes, but with a blunt and flattened snout; a small mouth; teeth that are similar in both jaws; spiracles present; eyes that lack a nictitating fold or membrane (which is at the inner angle of the eye or below the eyelid and is capable of extending over the eyeball); two dorsal fins, each with a well-developed spine; anal fin present; oviparous development; and an egg case that is horny with spiral flanges but no tendrils. They comprise one family (Heterodontidae), one genus (*Heterodontus*), and nine species, which vary in maximum total length from 57 to 165 cm (22 to 65 in.) [see **illustration**].



***Heterodontus* sp. (*J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006*)**

Bullhead sharks occur in warm-temperate and tropical continental waters of the western Indian Ocean and western and eastern Pacific Ocean, but are absent from the Atlantic Ocean and from oceanic insular waters. They live on the bottom, often occupying crevices and caves by day and feeding at night on sea urchins, starfishes, crabs, mollusks, and small fishes. *See* ELASMOBRANCHII; SELACHII. Herbert Boschung

Bibliography. L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503–547 in W. C. Hamlett (ed.), *Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids, and Chimaeras*, Science Publishers, Enfield, NH, 2005; L. J. V. Compagno, Sharks of the world: An annotated and illustrated catalogue of shark species known to date, *FAO Species Catalogue for Fishery Purposes*, no. 1, vol. 2: *Bullhead, Mackerel, and Carpet Sharks (Heterodontiformes, Lamniformes, and Orectolobiformes)*, FAO, Rome, 2001; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Heterodyne principle

The principle that multiple frequencies applied to a nonlinear device produce new frequencies that are sums and differences of the applied frequencies and their harmonics. Heterodyning, or mixing, is the process of multiplying a weak signal by a strong sinusoidal carrier, sometimes called the local oscillator, to shift the frequency of the signal in such a way that the information carried by the signal is preserved. *See* CARRIER (COMMUNICATIONS).

Applications. Signal frequencies are changed in a wide variety of electrical and electronic equipment, including radio, television, communications, radar, and measuring instruments. In modulation, low-frequency signals, such as audio or video, are translated to higher frequencies to permit the multiplexing of signals in a common communication system, transmission by a waveguide incapable of handling low frequencies, efficient radiation by an antenna of reasonable size, or generation of frequencies appropriate for physical measurements by a particular device. In demodulation, low-frequency audio or video information is removed from a high-frequency carrier at a receiver. In a superheterodyne receiver a low-power incoming signal is translated to a fixed lower-frequency band, eliminating the need for tuning an intermediate-frequency amplifier. Input signals of different frequencies are accommodated by merely changing the frequency of the local oscillator. Parametric amplifiers provide power gain to a small signal at the same frequencies or at changed frequencies. *See* DEMODULATOR; MODULATION; MODULATOR; PARAMETRIC AMPLIFIER; RADIO RECEIVER; TELEVISION RECEIVER.

Nonlinear devices. New frequencies cannot be generated in linear elements; nonlinearity is required. The form of nonlinearity varies with technology, ranging from iron-cored devices at power frequencies to minute semiconductor devices at microwave or millimeter-wave frequencies and nonlinear optical materials in laser systems. These nonlinear devices may often be idealized as nonlinear resistors, capacitors, or inductors. *See* INDUCTOR; NONLINEAR OPTICAL DEVICES; RESISTOR.

The analysis of nonlinear devices is difficult and has usually been restricted to special cases. The Manley-Rowe relations are an exception, providing general relations between the powers at the different frequencies in nonlinear inductors and capacitors. These results are remarkable in that they are independent of the shape of the nonlinear characteristic, the magnitudes of the signals at the different frequencies, and the external circuit connected to the nonlinear elements. They delineate what is possible and what is not. For example, they show which devices are stable and which are not; and they indicate that up-converters, used to produce a high frequency from a low frequency in transmitters, can have gain, whereas down-converters, used to reduce the input frequency to a lower frequency in receivers, have loss. They have been applied to the Raman effect at optical frequencies and to a wide variety of physical systems, including rotating electrical machinery, nonlinear inductors and capacitors and equivalent semiconductor devices, and microwave ferrites. *See* ELECTRIC ROTATING MACHINERY; ELECTRICAL INSTABILITY; FERRITE DEVICES; GAIN; RAMAN EFFECT.

Harrison E. Rowe

Bibliography. J. M. Manley and H. E. Rowe, Some general properties of nonlinear elements, I. General energy relations, *Proc. IRE*, 44:904–913, 1956; M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed., 1990; H. Taub and D. L. Schilling, *Principles of Communication Systems,* 1986.

Heterogeneous catalysis

A chemical process in which the catalyst is present in a separate phase. In the usual case, the catalyst is a solid, and the reactants and product are in gaseous or liquid phases. *See* CATALYSIS.

Heterogeneous catalysis proceeds by the formation and subsequent reaction of chemisorbed complexes which can be considered to be surface chemical compounds. In the simple case where A → B is slow in the absence of catalyst, reaction (1) might

$$\begin{array}{ll} \left.\begin{array}{l} * + A \rightarrow A* \\ A* \rightarrow B* \\ B* \rightarrow B + * \end{array}\right\} & \text{Chain propagating steps} \\ \quad A \rightarrow B & \text{Overall reaction} \end{array} \qquad (1)$$

occur. might occur. Reaction A → B is fast if the three preceding steps are fast. Here, * represents a catalytic site on the surface of the catalyst, $A^* \rightarrow B^*$ is called a surface reaction, $^* + A \rightarrow {}^*A$ represents the chemisorption of A, and $B^* \rightarrow B + {}^*$ represents desorption of B. *See* ADSORPTION.

With most sets of reactants, more than one reaction will be thermodynamically possible. The degree to which a given catalyst favors one reaction compared with other possible reactions is called the selectivity of the catalyst for reaction (1). Two aspects of a catalyst are of particular importance: its selectivity and its activity, which can be taken as the rate of conversion of reactants by a given amount of catalyst under specified conditions. Ideally, the rate will be proportional to the amount of catalyst.

History. Heterogeneous catalytic processes were first recognized by Humphry Davy in 1817 (the oxidation of hydrocarbon gases on platinum) and by L. J. Thénard in 1818 (the decomposition of hydrogen peroxide on various solids). Industrial applications started about 50 years later with the Deacon process for oxidizing hydrogen chloride (HCl) to chlorine (Cl_2), reaction (2), with a copper chloride catalyst, and the contact process for making sulfuric acid (H_2SO_4), reactions (3), with a

$$O_2 + 4HCl \rightarrow 2Cl_2 + 2H_2O \qquad (2)$$

$$\begin{array}{l} 2SO_2 + O_2 \rightarrow 2SO_3 \\ SO_3 + H_2O \rightarrow H_2SO_4 \end{array} \qquad (3)$$

platinum catalyst. The original platinum catalyst in (3) has been replaced by vanadium pentoxide on silica gel.

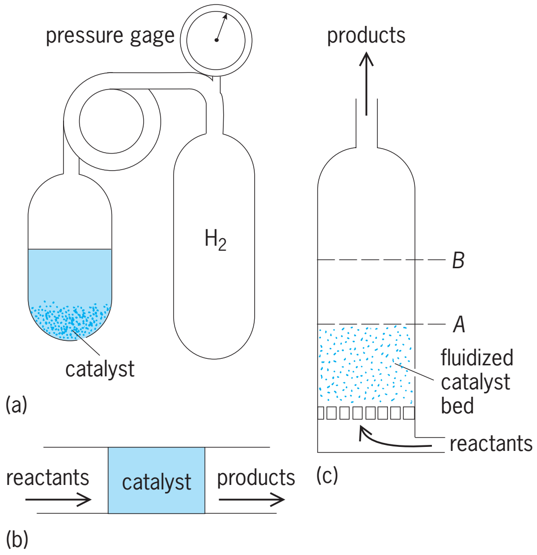
The rate of development of industrial applications of heterogeneous catalysis accelerated after about 1900. Manufacture of nitric acid (HNO_3) by the oxidation of ammonia (NH_3), reaction (4), began in

$$NH_3 \xrightarrow{O_2} NO \xrightarrow{O_2} NO_2 \xrightarrow{H_2O, O_2} HNO_3 \qquad (4)$$

1906, although in 1838 C. F. Kuhlmann had discovered that platinum catalyzed the first step of the process. In modern usage, the catalyst is platinum-rhodium gauze at 900°C (1652°F).

By the 1950s, heterogeneous catalytic processes had come to dominate the petroleum, petrochemical, and chemical industries. Heterogeneous catalysis is a critical feature in energy conservation and interconversion, is a key feature in the production of synthetic fuels from coal and oil shale, and is the primary strategy for control of automobile air pollution. *See* ALCOHOL FUEL; COAL GASIFICATION; FISCHER-TROPSCH PROCESS.

Reactors. Both industrially and in the laboratory, catalytic reactions are usually effected in one of three kinds of reactors: batch, tube-flow, and gradientless flow. Examples are shown in the **illustration**. In the flow reactors, there is usually a large change in concentrations of products and reactants between the entrance and exit of the bed. In the gradientless reactor, the changes are kept very small, either by recirculation of 99% of the exit gases from the catalyst bed in *b* back to the entrance to the bed, or by use of a stirred-flow reactor, of which the fluidized reactor in *c* is one form.



Catalytic reactors. (*a*) Batch. (*b*) Continuous fixed-bed. (*c*) Continuous fluidized reactor.

Some typical heterogeneous catalytic reactions

Catalyst	Reaction	*T*, °C*
	Hydrogenation	
Pt, Pd, Rh, Ni, as powders or supported on SiO_2, Al_2O_3, or C	$H_2 + C_2H_4 \rightarrow C_2H_6$	−100
	$H_2 + D_2 \rightarrow 2HD$	−180
	$C_3H_8 + D_2 \rightarrow C_3H_7D, C_3H_7D—C_3D_8 + HD$	50
	Cyclopropane + $H_2 \rightarrow C_3H_8$	40
	$C_2H_6 + H_2 \rightarrow 2CH_4$	250
Cr_2O_3 activated to generate coordinatively unsaturated Cr^{3+}	$H_2 + C_2H_4 \rightarrow C_2H_6$	−100
	$H_2 + D_2 \rightarrow 2HD$	−180
	$D_2 + C_3H_8 \rightarrow C_3H_7D + HD$	200
Fe†	$3H_2 + N_2 \rightarrow 2NH_3$	400‡
Cu/ZnO	$2H_2 + CO \rightarrow CH_3OH$	350‡
Pt, Cr_2O_3	Heptane → toluene + $4H_2$	450
Pt, Cu	Acetone + H_2 → 2-propanol	75
	Polymerization	
Cr^{2+}/SiO_2	C_2H_4 → linear polyethylene	50
	Olefin metathesis	
Mo^{4+}/Al_2O_3	$2C_3H_6 \rightarrow C_2H_4 + CH_3CH{=}CHCH_3$	50
	Oxidation	
Pt, many oxides of transition metals	$2H_2 + O_2 \rightarrow 2H_2O$	0−200
	$2CO + O_2 \rightarrow 2CO_2$	50−200
	$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$	200−350
Ag	$2C_2H_4 + O_2$ → ethylene oxide	200
Bismuth molybdate	$C_3H_6 + NH_3 + 3/2O_2 \rightarrow CH_2{=}CHCN + 3H_2O$	450
V_2O_5/SiO_2	Naphthalene + O_2 → phthalic anhydride	350
Fe_3O_4	$H_2O + CO \rightarrow H_2 + CO_2$	450

*The lowest temperature at which significant yields are obtained in a flow reactor. °F = (°C × 1.8) + 32.

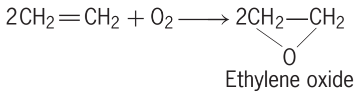
†In commercial practice, a few percent of a promoter such as potassium oxide plus aluminum oxide ($K_2O + Al_2O_3$) is added to the iron (Fe) to improve the performance and extend the life of the catalyst.

‡Because of the small value of the equilibrium constants at the operating temperatures, these reactions are run at about 100 atm (10 megapascals) in order to get adequate conversions.

Catalysts. A selection of heterogeneous catalytic processes of scientific or industrial interest is shown in the **table**.

Oxidation reactions of organic compounds may be divided into two categories, complete oxidation (or combustion) and partial (or selective) oxidation. The oxidation of methane (CH_4) shown in the table is an example of complete oxidation. One of the most widely applied uses of complete oxidation is the catalytic converter in motor vehicles. Here CH_4, other hydrocarbons, carbon monoxide (CO) in the engine exhaust, and oxygen are converted to carbon dioxide (CO_2) and water (H_2O) by passage over a supported platinum-palladium-rhodium catalyst.

Reaction (5) represents a particular catalytic partial oxidation reaction of commercial interest. The catalyst of choice is Ag/Al_2O_3 with various promoters, for example, Cs^+ and Cl^-. Reaction is carried out in a packed-bed flow reactor in the range of 200–300°C (392–572°F), and selectivites of the order of 80% are achieved.



Ethylene oxide

(5)

Since catalytic activity will ordinarily be proportional to surface area, most catalysts are used in forms with large specific areas, a_s. The low-area catalyst of reaction (4) is a rather unusual case. Higher-area metal powders are often used for liquid-phase reactions in batch reactors like that shown in *a*. For example, finely divided nickel, $a_s = 25–40\ m^2\ g^{-1}$, is used for the hydrogenation of unsaturated glycerides in the manufacture of margarine from vegetable oils. *See* HYDROGENATION.

Supported catalysts are widely used. In these, the catalytic ingredient is dispersed in the internal porosity of such supports as silica gel, γ-alumina, and charcoals. These supports have large areas in the internal porosity, 100–1000 $m^2\ g^{-1}$, and their average pore diameters are 2–20 nanometers. Tiny crystallites of such metals as platinum, palladium, rhodium, and nickel can be formed in the pore structure. Supported oxides and sulfides of transition metals are also used. Supported catalysts have the advantage that the area of the catalytic ingredient can be very large. Since, however, the support granules can have diameters in the 1–5 mm range, large flows of gas produce only moderate pressure drops across a catalyst bed. Supported catalysts are much more resistant to coalescence of the catalytic ingredient than are powders. Further, deposition of carbonaceous residues accompanies many catalytic reactions. In some cases, the catalyst can be regenerated by burning off the residues. Such regeneration would result in drastic losses of area with metal powders.

Supported catalysts are particularly prone to problems with diffusion. Reaction $^* + A \rightarrow {}^*A$ must be preceded by the diffusion of the reactant through the pore structure of the support to a catalytic site. Reaction $B^* \rightarrow B + {}^*$ must be followed by the diffusion of the product out of the support. Heat must also flow in and out of the granule of support. Such matters receive particular attention in the chemical engineering aspects of catalysis. *See* DIFFUSION.

One important type of catalyst exposes strongly acidic sites in its internal porosity. Such catalysts are used to crack larger molecules of hydrocarbon into smaller ones in petroleum refining. Other catalysts, called dual-functional catalysts, have a hydrogenating catalytic ingredient on an acidic support. These are also of major importance in processing petroleum. The acidic function for catalytic cracking is commonly provided by a zeolite, a crystalline porous material with strongly acidic sites. *See* CRACKING; HYDROCRACKING; ZEOLITE.

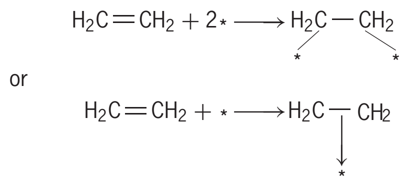
Another type of catalyst consists of an organometallic complex deposited on such supports as silica or alumina. These catalysts have been called heterogenized homogeneous catalysts, and they accompany a development in which the nature of surface sites on heterogeneous catalysts has been interpreted in terms of coordination chemistry and homogeneous catalysis. *See* COORDINATION CHEMISTRY.

Most catalysts exhibit coordinatively unsaturated surface sites (*cus*) which are capable of reacting with molecules in the gas or liquid phases to form chemisorbed intermediates, as in reaction $^* + A \rightarrow {}^*A$. For example, the atoms at the surface of a crystallite of platinum must be coordinatively unsaturated. Considerable progress has been made in applying the studies of surface chemistry and physics on particular crystal surfaces—(111), (110), or (100)—of such metals as platinum to the interpretation of catalytic reactions and chemisorption on metals. *See* ELECTRON DIFFRACTION; SURFACE PHYSICS.

High catalytic activity for a given reaction requires adsorption of the reactants at *cus* sites to be of intermediate strength. If the adsorption is too weak, the reactants are unlikely to be activated; if too strong, A^* will be unreactive. With good catalysts, then, it is likely that there will be other compounds which adsorb more strongly than the reactants, with partial or complete blockage of the catalytic reaction. Such compounds are said to be poisons. For example, transition metals are poisoned by soft bases like R_2S, R_3P, and CO, where R represents a functional group. Poisoning is a problem in all types of catalysis.

Mechanism. The kinetics of heterogeneous catalytic reactions are often rather complicated. In general, rate = (amount of catalyst)$f(T, P_i, \text{ or } C_i)$, where T is the temperature of the catalyst, P_i is the partial pressure of component i, and C_i is the concentration of component i. The index i covers reactants and products. A reaction having a rate-limiting process $A^* + B^* \rightarrow C^*$ will have a rate proportional to $\theta_A\theta_B$, (where θ_A is the fraction of surface sites represented by * which have been converted to A^* and θ_B) that converted to B^*. For $A^* + B^* \rightarrow C^*$ to be a rate-limiting process, A^* must be in equilibrium with $A(g)$ and B^* with $B(g)$, where g indicates the gaseous phase. Thus, θ_A and θ_B can be expressed in terms of P_A and P_B through the appropriate adsorption isotherms. If θ_A is small, it is apt to be proportional to P_A or $P_A{}^{0.5}$. If $\theta_A \approx 1$, it will be very nearly independent of P_A. *See* ADSORPTION.

Considerable work has been aimed at the determination of the chemical mechanism of particular catalytic reactions. Such work has depended upon kinetics, isotopic tracers, stereochemistry, and structure variation in reactants. A simple example is that which is commonly accepted for the hydrogenations of olefins, the Horiuti-Polanyi mechanism as illustrated for ethylene adsorption in reactions (6). It is as yet unclear which formulation for the chemisorption of ethylene is most appropriate. The geometries of the two forms would be nearly the same. The adsorption of ethylene is accompanied by the dissociative chemisorption of hydrogen, reaction (7), followed by reactions (8) and (9).Much still remains to be learned about the details of catalytic mechanisms. *See* HOMOGENEOUS CATALYSIS.



(6)

$$H_2 + 2{*} \rightarrow 2H{*} \tag{7}$$

$$*H_2C\text{—}CH_2{*} + H{*} \rightarrow *CH_2CH_3 + 2{*} \tag{8}$$

$$*CH_2CH_3 + H{*} \rightarrow CH_3CH_3(g) + 2{*} \tag{9}$$

Robert L. Burwell, Jr.; Gary L. Haller

Bibliography. J. R. Anderson and M. Boudart (eds.), *Catalysis: Science and Technology*, vols. 1–11, 1981–1996; J. A. Moulijin, P. Van Leeuwen, and R. A. Van Santen (eds.), *Catalysis: An Integrated Approach to Homogeneous, Heterogeneous, and Industrial Catalysis*, 1993; J. M. Thomas and W. J. Thomas, *Principles and Practice of Catalysis*, in G. Ertl, H. Knözinger, and J. Weitkamp (eds.), *Handbook of Heterogeneous Catalysis*, vols. 1–5, 1997.

Heteronemertini

An order of the class Anopla in the phylum Rhynchocoela, with an unarmed proboscis, a thick partly fibrous dermis, and a three-layered body musculature composed of outer longitudinal, median circular, and inner longitudinal strata. Cerebral organs, cephalic grooves and slits, and eyes are generally present. The alimentary system consists of a mouth, foregut, intestine with regular lateral diverticula but no cecum, and anus. The principal family, Lineidae, contains some of the most common and best-known rhynchocoelan genera of temperate seashores, such as *Lineus* ("bootlace worms"), *Micrura*, and *Cerebratulus* ("ribbon worms"); the last one widely used in studies on regeneration, embryology, and nutrition. Heteronemertini are mainly marine littoral in habit. *See* ANOPLA; ENOPLA; PALAEONEMERTINI; NEMERTEA.

J. B. Jennings

Heterophile antigen

The serologic reactions of the tissue and blood-cell antigens of most animals are normally characteristic of the species. Significant serologic cross reactions usually occur only with antisera to the corresponding antigens of closely related species. The numerous groups of heterophile antigens—of which the Forssman antigens are the best studied—constitute significant exceptions. Heterophile antigens link the species hog-ox-human (blood group A), cat-horse, and dog-hog-cat-human, while several heterophile groups link otherwise diverse microorganisms. Links between pneumococcus type XIV and the human blood groups are also known, as well as similarities between antigens in mammalian hearts and the cell walls of the group A hemolytic streptococcus—a bacterium that is a common cause of rheumatic fever in humans. The cross reactions between the *Proteus* bacillus and the *Rickettsiae* are important in the diagnosis of typhus fever.

In 1911 J. Forssman reported that the injection of tissues from the guinea pig into rabbits stimulated the formation of antibodies which, together with complement, lysed sheep red blood cells. The organs of the horse, dog, cat, mouse, fowl, and tortoise also stimulate the production of antibodies with affinities for sheep cells, while the organs of humans, the rabbit, ox, rat, goose, eel, and frog do not. The Forssman antigen is also widely distributed among microorganisms, for example, pneumococci, anthrax, and dysentery bacilli. Cross reactions between the Forssman antigen in human tissues and human blood-group-A substance are said to be due to the sharing of a common disaccharide, α-*N*-acetyl-galactosaminoyl-(1→3)-D-galactose. "Purified" Forssman antigen derived from pooled sheep red cells was found to contain as the principal ingredients galactose, galactosamine, lignoceric acid, and sphingosine. Some human sera react strongly with the erythrocytes of animals (sheep, horse, ox) that carry heterophile antigens. These reactions are commonly used for the diagnosis of infectious mononucleosis. *See* ANTHRAX; ANTIBODY; ANTIGEN; BACILLARY DYSENTERY; INFECTIOUS MONONUCLEOSIS; PNEUMOCOCCUS; RICKETTSIOSES; SEROLOGY.

Margaret J. Polley

Bibliography. R. M. Aloisi, *Principles of Immunology and Immunodiagnostics*, 1988; L. G. Schook and J. G. Tew, *Antigen Presenting Cells: Diversity, Differentiation, and Regulation*, 1987.

Heterosis

Hybrid vigor or increase in size, yield, and performance found in hybrids, especially if the parents have previously been inbred. The application of heterosis has been one of the most important contributions of genetics to scientific agriculture in providing hybrid corn, and vigorous, high-yielding hybrids in other plants and in livestock. *See* BREEDING (ANIMAL); BREEDING (PLANT); GENETICS.

It has been known for several centuries that hybrids between varieties and even between species are frequently of unusual size and vigor. The proverbial hardiness of the mule is often given as an example in animals. Charles Darwin devoted many years to a study of the effects of inbreeding and hybridization in several kinds of domestic plants. He reported, and his findings have since been abundantly confirmed, that inbreeding usually leads to decreased size and weakness, and to an increased frequency of abnormalities. The changes are cumulative so that with successive generations of inbreeding the deleterious effects are more and more pronounced. However, all the effects of inbreeding are immediately eliminated by outcrossing. Darwin also noted that many plants have elaborate mechanisms that prevent self-fertilization, thus avoiding the deleterious effects of close inbreeding.

A satisfactory explanation of heterosis awaited the development of mendelian genetics. It is now known that the effect of inbreeding in a population is to make homozygous many genes that had previously been heterozygous. Thus the decline in vigor and size must have its explanation in the increased homozygosity that accompanies inbreeding.

Likewise, heterosis may be explained by the reverse effect of increased heterozygosity following outcrossing, since a hybrid will in general be more heterozygous than its parents. *See* MENDELISM.

Hypotheses. There are two principal hypotheses to account for the association of size and vigor with heterozygosity.

Dominance. The dominance hypothesis notes that any noninbred population carries a number of recessive genes that are harmful to a greater or lesser extent, but which are rendered ineffective by their dominant alleles. As they become homozygous through inbreeding, they exert their harmful effect. With hybridization, some of the detrimental recessives contributed to the hybrid by one parent are masked by dominant alleles from the other, and an increase in vigor is the result. This hypothesis is supported by numerous experimental studies showing the large number of harmful recessives carried in actual populations, a number that is quantitatively adequate to account for the observed decline in vigor with inbreeding and recovery on crossing. *See* DOMINANCE.

Overdominance. The alternative hypothesis is that there are loci at which the heterozygote is superior in vigor to either homozygote. This, the overdominance hypothesis, also has the consequence that vigor is proportional to heterozygosity. It is argued that studies of individual genes have revealed only a very small number with this property, but on the other hand such genes would be expected to persist in a population and to make a disproportionate contribution to its variability. The dominance hypothesis has been more widely accepted, but the two are very difficult to distinguish experimentally, and it is likely that overdominant loci are playing an appreciable role in heterosis, particularly in determining why one hybrid is better than another.

In the fruit fly (*Drosophila melanogaster*) the decreased viability with inbreeding is about two-thirds caused by genes with drastic effects—most of them being lethal when homozygous—and about one-third by a much larger number of very mildly deleterious genes that act cumulatively. The evidence that these genes are numerous comes from two sources. One is the high rate of mutation; either each gene mutates very often or there are many of them, and the latter is much more likely. The second line of evidence is the discovery of a great amount of heterogeneity for proteins. Since these proteins are direct gene products, they reflect the genetic variability. *Drosophila* studies have shown that the average fly is heterozygous for about 12% of the gene loci. The values in vertebrates are smaller, being about 4-5%. Whether all this variability is maintained in the population by overdominance or by recurrent mutation is not yet clear. However, these are presumably the kind of genes that are responsible for one-third of the inbreeding effect, as well as for the differences between one hybrid and another.

Use. The most important example of the systematic use of heterosis has been the development of high-yielding hybrid corn. Almost all the corn grown in the north-central corn-producing states of the United States is now hybrid. About 1910 two American geneticists, E. M. East and G. H. Shull, discovered that by inbreeding corn, selecting the inbred strains, and then making hybrids between the selected inbred strains, they could obtain excellent plants. The hybrids not only had a high yield but also were extremely uniform, and by choosing the right inbred strains for hybridization, the breeder could incorporate into the hybrids other desirable traits such as disease resistance, straight stalks, and well-shaped ears.

The main difficulty in this scheme is that the seeds from which the hybrid plants develop are grown on ears of inbred plants. Hence the seed producer has a low yield, and the seed becomes expensive. This difficulty is circumvented by using the double cross, or four-way-cross, originated by D. F. Jones. The breeder starts with four inbred strains and makes two different crosses yielding two hybrids. The two hybrids are then crossed to produce the commercial double-cross seed. In this way the commercial seed is produced on high-yielding hybrid plants. The double-cross hybrids have about the same yield as the single-cross hybrids, though they are less uniform. It has become possible to produce higher-yielding inbred lines, so almost all hybrid corn is now single cross, with an improvement in uniformity.

Following the pattern set for hybrid corn, breeders have produced many other agricultural plants as hybrids. There has also been an increasing tendency to use heterosis in animal breeding, both by employing mating schemes that maximize heterosis and by hybridization between inbred lines, the latter being widely practiced especially in poultry breeding. James F. Crow

Bibliography. G. W. Burns and P. J. Bottino, *The Science of Genetics*, 6th ed., 1988; J. F. Crow, The rise and fall of overdominance, *Plant Breed. Rev.*, 17:225-257, 2000; V. Loeschcke (ed.), *Genetic Constraints on Adaptive Evolution*, 1987.

Heterostraci

Extinct, armored, jawless vertebrates with a single pair of external gill openings. Heterostraci first appeared in the Ordovician, according to some authorities, and became common in the Silurian and Devonian. Their bony dermal armor consists mainly of aspidine (bone with apparent growth layers, lacking cavities for bone cells) and is composed of three structural layers: a basal layer of lamellar bone, a middle layer usually with a honeycomb arrangement of cavities and partitions, and a superficial layer of dentine tubercles or ridges. *See* JAWLESS VERTEBRATES; OSTRACODERM.

Morphology. The head region was usually enclosed in spoon-shaped upper and lower bony plates or

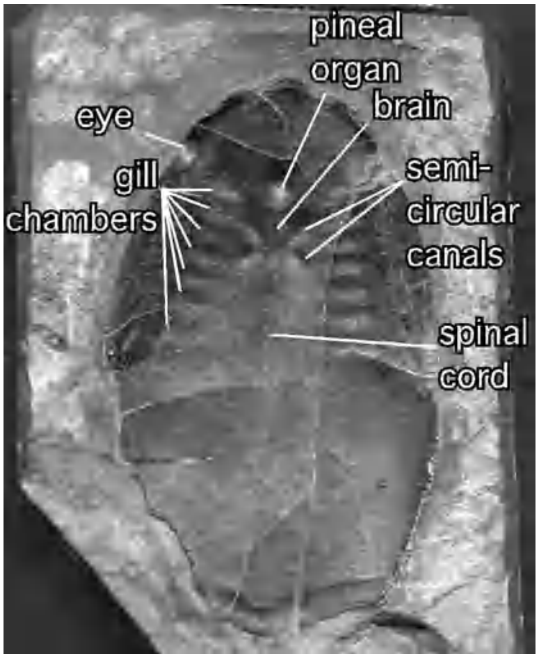


Fig. 1. Internal surface of the dorsal head shield of a cyathaspidiform heterostracan from northern Canada. Locations of organs are indicated by impressions in the surface of the shield. Length of shield, 3.5 cm (1.4 in.).

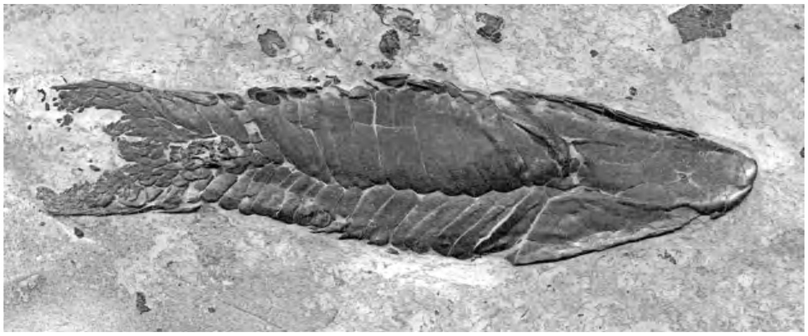


Fig. 2. *Dinaspidella elizabethae*, **a typical cyathaspiform heterostracan from the Early Devonian of northern Canada. Anterior is to the right; total length, 11 cm (4.3 in.).**

shields, joined to each other by smaller plates near the eyes, mouth, and gill chambers. The inner surface of the head armor often bore impressions of the brain, pineal organ, two pairs of semicircular canals, and the gill chambers (**Fig. 1**). The mouth lacked jaws, but small movable plates formed its rear margin. The eyes were widely separated on the sides of the head. The seven, paired, internal gill chambers emptied through a single pair of external gill openings. The body was covered in thick scales, each shaped like a diamond or parallelogram. The tail had a strong lower lobe into which the axis of the body continued, but an almost equally strong upper lobe was often present, with smaller scale-covered lobes between them (**Fig. 2**).

Phylogeny. Ordovician genera such as *Arandaspis* and *Astraspis* are often considered to be close relatives to or members of the Heterostraci, though they lack some of the group's unifying features. In the Silurian, primitive heterostracans such as *Athenaegis* and *Tolypelepis* occur. In Late Silurian and Early Devonian rocks, typical heterostracans in the Cyathaspidiformes (such as *Dinaspidella*, *Anglaspis*, *Poraspis*, *Ctenaspis*) are very diverse and among the most common fossil vertebrates, while other forms such as *Lepidaspis* and *Aserotaspis* are more poorly known. In the Late Silurian and during the Devonian, more advanced forms including Pteraspidiformes (such as *Errivaspis*, *Pteraspis*, *Phialaspis*) and Psammosteiformes (such as *Drepanaspis*, *Psammolepis*) diversified and became dominant. The Heterostraci declined in diversity and abundance toward the end of the Devonian and were extinct before the end of that period. *See* DEVONIAN; ORDOVICIAN; SILURIAN.

Fossil record. Heterostraci were common in North America and in Europe but are not known from China or the Southern Hemisphere. Their fossils are useful to geologists as indicators of the ages of the rocks in which they are found. Silurian representatives lived in marine habitats, while later forms occupied both marine and freshwater environments. The latest surviving Heterostraci in the Psammosteiformes were confined to fresh or brackish waters and reached body sizes up to 1.5 m (5 ft). *See* FOSSIL.

Mark V. H. Wilson

Bibliography. P. Janvier, *Early Vertebrates*, Clarendon Press, Oxford, 1996; J. A. Long, *The Rise of Fishes: 500 Million Years of Evolution*, Johns Hopkins University Press, Baltimore, 1995; J. A. Moy-Thomas and R. S. Miles, *Palaeozoic Fishes*, 2d ed., W. B. Saunders, Philadelphia, 1971.

Heterotardigrada

An order of the tardigrades, the majority of whose genera have widely varied structure. Cephalic appendages having a sensorial function are present, as well as cirrus lateralis and clava. Pharyngeal pockets are strengthened by uninterrupted ridges, except in *Pseudechiniscus islandicus*, which has interrupted ones. Toes or claws are uniform in structure and completely separated from one another. A preanal gonopore is present; excretory glands are absent. This order of tardigrades is divided into two suborders, Arthrotardigrada and Echiniscoidea.

Members of the suborder Arthrotardigrada have toelike terminations of the legs (**Fig. 1**). The tubular middle part of the leg telescopes into the broad proximal part. These animals are marine organisms found in sand or on algae. One species, *Tetrakentron*, is found on the buccal tentacles of a sea cucumber, *Leptosynapta*. Species of *Batillipes* were recorded from North Carolina and Florida.

In the suborder Echiniscoidea the legs terminate with claws (**Fig. 2**). The middle part of the leg is

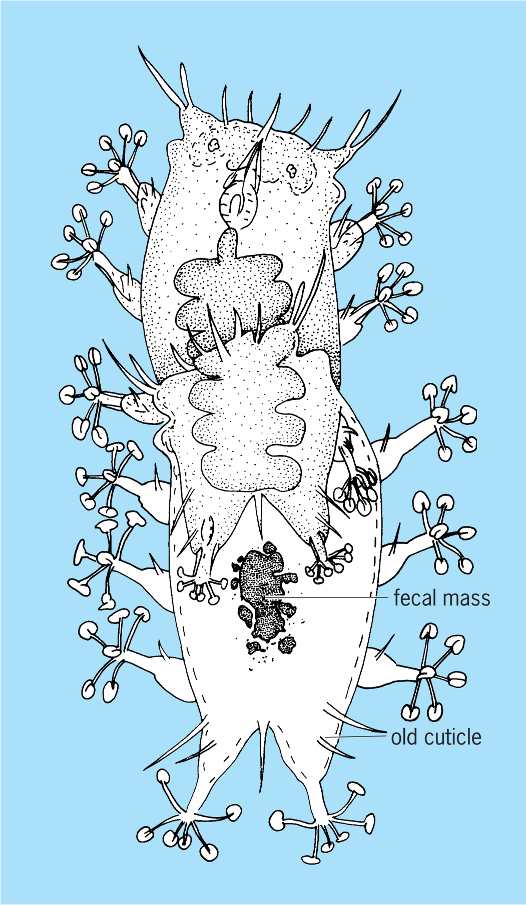


Fig. 1. *Batillipes*; **defecation during molt.**

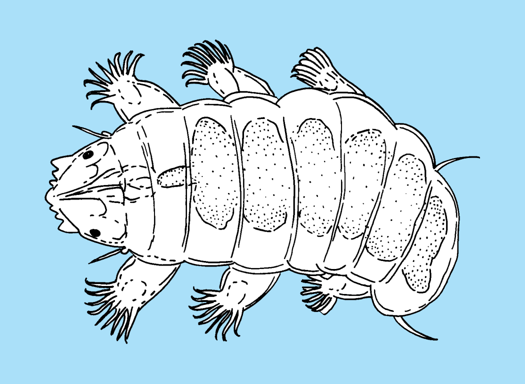


Fig. 2. *Echiniscoides sigismundi*, a representative of the suborder Echiniscoidea.

partially retractable into the proximal part. At least the fourth pair of legs has a distinct fold. There are two families: Nudechiniscidae, with a uniform cuticle; and Scutechiniscidae, with segmental and intersegmental thickenings (plates) of cuticle. Frequently these animals are red because of the presence of carotenoid pigments. Many species are active at relatively low humidity. *See* CAROTENOID; TARDIGRADA.

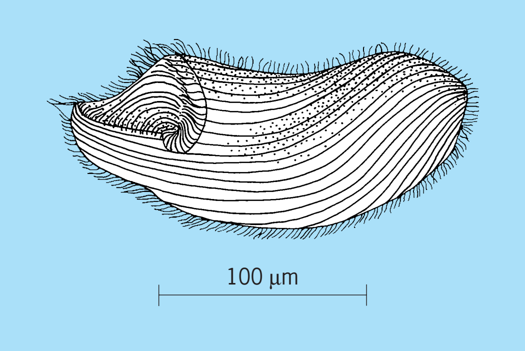
Eveline Marcus

Heterotrichida

A large order of the Spirotrichia containing many well-known, sizable species. The buccal ciliature, both membranes and the adoral zone of membranelles, is well developed, although in a number of families the somatic, or body, ciliature is really holotrichous in nature. Heterotrichs have become adapted to all sorts of habitats, including the digestive tracts of a variety of invertebrate and a few vertebrate hosts. Man is not involved. Some species contain pigments in their cytoplasm giving them such coloration as blue, green, pink, brown, or black.

Because of their size and amazing regenerative powers, a number of heterotrichs have been widely used in experimental research paralleling the studies by embryologists in the fields of morphogenesis and differentiation. *Stentor* (most popular in morphogenetic investigations), *Blepharisma*, and *Folliculina* are all genera with colored species. *Condylostoma*, *Climacostomum* (see **illus.**), and *Spirostomum*, which may reach a length of $^1/_8$ in. (3 mm), are common species. *Nyctotherus* is found in the digestive tract of amphibians and many invertebrates. *Balantidium*, long considered a heterotrich and important as the only ciliate parasitizing humans, is actually a trichostome. *See* CILIOPHORA; PROTOZOA; SPIROTRICHIA.

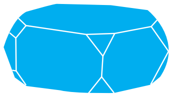
John O. Corliss



Climacostomum, an example of a heterotrich.

Heulandite

A mineral belonging to the zeolite family of silicates and crystallizing in the monoclinic system. It usually occurs in crystals with prominent side pinacoid, often having a diamond shape. There is perfect side pinacoid cleavage on which the luster is pearly; elsewhere the luster is vitreous. The crystals often have undulating faces, and are made up of subindividuals in nearly parallel position (see **illus.**). In polarized light they show optical anomalies of a sectoral nature. The hardness is $3^1/_2$ to 4 on Mohs scale; specific gravity is 2.18–2.20. The mineral is usually white or colorless but may be yellow or red.



Heulandite crystal habit. (*After C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*)

Heulandite is essentially a hydrous calcium aluminum silicate, $Ca(Al_2Si_7O_{18}) \cdot 6H_2O$. Small amounts of sodium and potassium usually substitute for calcium. Heulandite is a secondary mineral found in cavities in basalts associated with other zeolites and calcite. Notable localities are in the Faeroe Islands, India, Nova Scotia, and West Paterson, New Jersey. *See* SILICATE MINERALS; ZEOLITE.

Clifford Frondel;
Cornelius S. Hurlbut, Jr.

Hexacorallia

A subclass of cnidarian class Anthozoa that includes stony and black corals and sea anemones; also known as Zoantharia. With few exceptions, these animals are benthic as adults, living attached to firm substrata or burrowed into soft sediments; they occur at all latitudes, and from the high intertidal zone to the deepest parts of the oceans. As in all anthozoans, the adult form is a polyp—a cylindrical organism that has, at its free end, the single body opening, the mouth, which is surrounded by retractile, rarely branched, tubular tentacles **(Fig. 1)**. The tentacles number from six to several hundred, but typically the number is approximately a multiple of six, which gives the subclass its name, Hexacorallia. The opposite end of the polyp is attached to or burrowed into the substratum, or it emerges from the tissue that unites the members of the colony. *See* ANTHOZOA; SEA ANEMONE.

Fig. 1. Anthozoan anemone polyp. (*Photo by Eugene Weber, © California Academy of Sciences*)

Diversity. Orders Scleractinia (stony corals) and Zoanthidea (mat anemones) have some members that are colonial, some clonal, and some solitary; all members of Antipatharia (black corals) are colonial, and all members of Ceriantharia (tube anemones) are solitary; orders Actiniaria (sea anemones in the strict sense) and Corallimorpharia (mushroom anemones) contain some members that are solitary and some that are clonal (in at least one anemone species, clone mates may not separate completely, but whether this should be considered coloniality is debatable). Actiniaria and Scleractinia have the greatest number of species, each with somewhat more than a thousand; the other orders include a few tens to a hundred or so species. *See* ACTINIARIA; ANTIPATHARIA; CERIANTHARIA; CORALLIMORPHARIA; SCLERACTINIA; ZOANTHIDEA.

Previously, a seventh order of Hexacorallia was recognized, Ptychodactiaria. Comprising only three species, each in its own genus, this group is now considered a member of Actiniaria (where the two species known at the time had been placed until the middle of the twentieth century). The order Corallimorpharia, which is morphologically intermediate between Scleractinia (which it resembles in internal anatomy and the type of nematocysts it possesses) and Actiniaria (which it resembles in lacking a calcareous skeleton and lacking colonial members), may actually belong to Scleractinia. *See* PTYCHODACTIARIA.

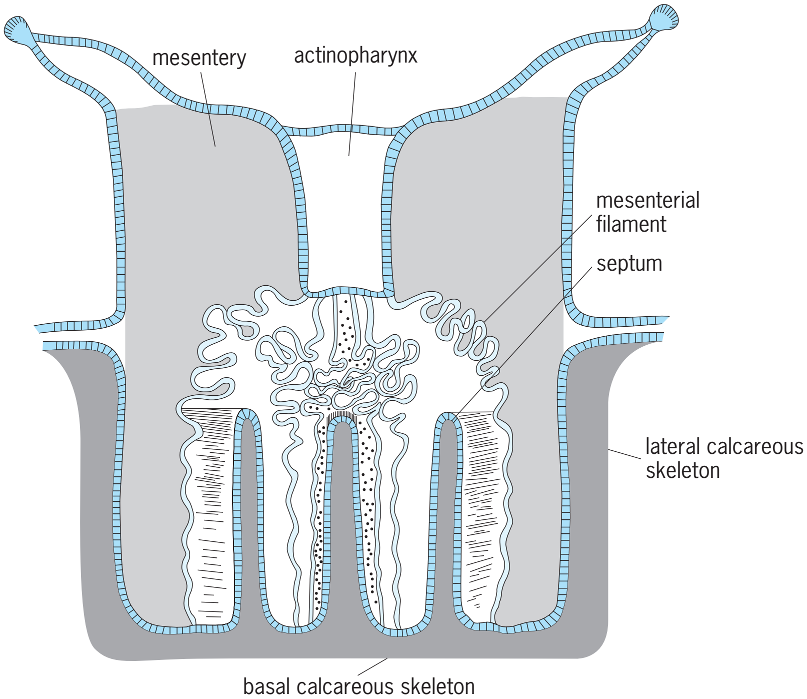


Fig. 2. Diagrammatic longitudinal section through a scleractinian polyp.

Morphology. As in all anthozoans, the internal body space (the coelenteron, or gastrovascular cavity) is divided by longitudinally oriented sheets of tissue, the mesenteries **(Fig. 2)**. Along the free edge of most mesenteries is a filament that contains cilia, nematocysts, and gland cells; typically a longitudinal retractor muscle runs along one side of each mesentery. Commonly, two mesenteries—the directives—attach to each siphonoglyph, a histologically specialized channel running the length of the actinopharynx. A typical member of Actiniaria and Antipatharia has two siphonoglyphs diametrically opposite one another; a member of Cerianthiaria and Zoanthidea has a single one, conventionally considered to be on opposite sides of the animal in the two orders, the position in the former termed asulcal and in the latter, sulcal. As far as is known, scleractinians lack siphonoglyphs, which are typically absent or poorly developed in corallimorpharians.

In many hexacorals, the number of mesenteries increases as the diameter of the polyp does. An antipatharian polyp does not grow much after it is formed; in some species, a polyp never has more than the six initial mesenteries; in others, four or six additional mesenteries develop. Mesenteries are added only in certain regions and bilaterally—they form in couples, the halves arising simultaneously on each side of the animal. In Actiniaria, Corallimorpharia, Scleractinia, and Zoanthidea, two mesenteries (which constitute a pair) arise side by side simultaneously. In Zoanthidea, one member of most pairs is complete and one incomplete; in the other orders, the members of a pair are typically the same size. Mesenteries are added around the perimeter of an actiniarian, corallimorpharian, and scleractinian polyp, with the pairs forming diametrically opposite one another and constituting a couple. In a zoanthidean polyp as well, a couple constitutes two pairs of mesenteries, but addition is only on each side of the directive mesenteries that connect to the siphonoglyph. In a cerianthid polyp, mesenteries are also added in a single growth zone, but on the side of the polyp opposite the siphonoglyph; and a couple constitutes two single (not paired) mesenteries. Thus, the oldest mesenteries flank the siphonoglyph and are sequentially younger toward the opposite side of the animal, whereas the youngest mesenteries in a zoanthidean flank the siphonoglyph and are sequentially older toward the opposite side of the animal.

Characteristics. As is true of all cnidarians, hexacorals are carnivores. Many shallow-water species, particularly those in warm waters, have intracellular dinoflagellate symbionts (zooxanthellae) that contribute fixed carbon to the animal and remove wastes from it. Gametes form and mature in the mesenteries, between the filaments and the retractor muscles, which rupture to release them when they are ripe; some kinds of hexacorals are hermaphroditic and some are gonochoric (with separate sexes). Typically gametes are spawned freely into the sea, where egg and sperm meet and the planula (ciliated, free-swimming larva) develops; in some hexacorals, however, embryos are brooded internally or on the surface of the parent, so a free-swimming stage of the life cycle is lacking.

Polymorphism of the sort that is common in octocorals is virtually absent in hexacorals. In both clonal and colonial species, some individuals may have distinctive morphological attributes, but in contrast to octocorals these differences appear to be situational, so an individual can develop or lose the attributes. For example, in some species of scleractinians, the polyps at the edge of a colony that encounters cnidarians of a different species can develop elongate tentacles with nematocysts specialized for aggression. In actiniarians of a few clonal species, individuals at the periphery of a clone have highly developed acrorhagi (marginal bulges with dense nematocysts that are used in aggressive encounters) and form few or no gametes, whereas individuals at the core of the clone have small acrorhagi and their gametogenic tissue is richly developed.

Reproduction and propagation. Sexual reproduction typically results in a planula larva, which eventually metamorphoses into a polyp. In clonal and colonial species, asexual (vegetative) reproduction occurs in several ways, typically in only one way per species. Budding involves an outgrowth of a new polyp from an existing one or from the tissue connecting members of the colony; other hexacorals divide into two, either longitudinally or transversely; some sea anemones regenerate from detached tentacles or from bits of tissue that tear from the edge of the base; and apparently some anemones and corals can develop from bits of tissue shed from the inside of the animal, although the precise mechanism for this mode of propagation is unclear.

Skeletal features. A skeleton is absent in Actiniaria, Ceriantharia, and Corallimorpharia; zoanthideans do not form skeletons either, but in many taxa a polyp can incorporate into the mesoglea of its column particles such as sponge spicules or sand grains to form a sort of adventitious skeleton. In life, an antipatharian skeleton is hidden under the thin layer of polyps that secrete it; after their death, the polyps rot away, leaving the bushy or whiplike black skeleton that gives the taxon its common name, black corals. Composed of organic material so it is somewhat flexible, it has traditionally been formed into bracelets thought to ward off illness (hence the name Antipatharia) and is made into other jewelry as well. All scleractinian polyps secrete around themselves a cuplike calcareous skeleton in the crystal form aragonite; in some taxa the animals are solitary; in some the colonial skeleton is massive, with the "cups" of adjacent polyps sharing walls. Each polyp also deposits radial calcareous septa within the cup in the space between members of each larger mesenterial pair (Fig. 2). This skeleton is the major structural component of tropical and subtropical coral reefs and of the structures known as deep-sea coral reefs. *See* REEF.

Fossil record. Several orders of Hexacorallia have excellent fossil records, but fossils of groups that are exclusively soft-bodied are rare and almost never show taxonomically informative features. The earliest-known fossil corals are of Middle Cambrian age, but their sporadic occurrence in the Cambrian and earliest Ordovician suggests that early hexacorals lacked skeletons and that skeletalization evolved several times, possibly not in the same lineage each time. Members of the two major orders of Paleozoic corals first appeared in the Ordovician—Tabulata in the Early Ordovician and Rugosa in the Middle Ordovician. Middle Cambrian corals seem to represent two or more "experiments" in skeletalization, and Tabulata a later one. Rugosa seems to be polyphyletic: it can be logically derived from Tabulata at least twice. A minor order commonly thought to have separated from Tabulata in the Middle Ordovician became common and even a reef-builder in the Silurian and Devonian but was extinct before the end of the Middle Devonian. Another minor order separated from the Rugosa during the Late Devonian but became extinct during the early Carboniferous.

Lineage. Whether the lineages of these corals survived beyond the Paleozoic is debated. Zoanthideans bear considerable anatomical resemblance to rugosans. It has also been suggested that Scleractinia descended from Rugosa, possibly polyphyletically, with various scleractinian groups arising from different rugosan families. However, because of differences in ontogeny (mode of septal insertion), skeletal composition (the calcareous rugosan skeleton has the crystal form calcite), and the time gap (there are no known Early Triassic corals), it is more likely that Scleractinia evolved from one or more groups of early Mesozoic soft-bodied hexacorals. These ancestral animals are often referred to as sea anemones, but there is no independent evidence that order Actiniaria is so ancient. Because of similarities between scleractinians and actiniarians, it is also possible that sea anemones are derived from a lineage of ancestral scleractinians through loss of the calcareous skeleton and other morphological modifications; preliminary molecular evidence favors this direction of evolution. *See* RUGOSA.

Further complicating hexacorallian phylogeny is the strong evidence that Scleractinia is polyphyletic. All hexacorals with paired, coupled mesenteries and a calcareous skeleton are considered scleractinians, but details of skeletal structure and of animal anatomy and biology vary greatly, and the groups with similar skeletons do not necessarily share other attributes, so it may be that Scleractinia is a grade, not a clade. Moreover, the skeleton is inferred to have been lost and regained in some lineages. The greatest difference between Scleractinia and Corallimorpharia is that a polyp of the former has a calcareous skeleton whereas one of the latter is askeletal; the relative ephemerality of the scleractinian skeleton has led to the suggestion that corallimorpharians are simply naked scleractinians. Daphne G. Fautin

Bibliography. D. F. Dunn, Cnidaria, pp. 669–706 in S. P. Parker (editor-in-chief), *Synopsis and Classification of Living Organisms*, vol. 1, McGraw-Hill, New York, 1982; H. Erhardt and D. Knop, Order Zoantharia, pp. 269–278 in *Corals: Indo-Pacific Field Guide*, Ikan, Frankfurt, 2005; D. G. Fautin and G. R. Allen, *Field Guide to Anemonefishes and Their Host Sea Anemones*, Western Australian Museum, 1992; L. H. Hyman, Class Anthozoa: Subclass Zoantharia, pp. 566–632 in *The Invertebrates: Protozoa through Ctenophora*, vol. 1, McGraw-Hill, New York, 1940.

Hexactinellida

A class of sponges whose skeletons are made of siliceous hexactine spicules. These exclusively marine sponges are widely distributed in modern oceans. Their fossil record extends from the late Precambrian to the Recent. The basic spicule type of the class is a triaxial hexactine, in which the three pairs of opposed rays are at right angles to each other and lie along one of the three axes of a cube. Proximal ray ends and axial filaments meet at the center of the cube. These principal spicules and variants of that form make up skeletons of the sponges.

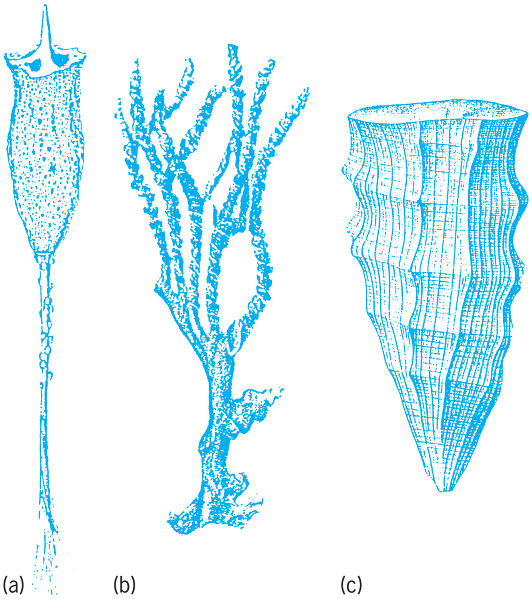


Fig. 1. Living hexactinellids (*a*) *Hyalonema* (*Cyliconema*) *thomsoni* Marshall, order Amphidiscosida (*after F. E. Schulze, 1887*) and (*b*) *Regadrella phoenix* Schmidt, order Lyssacinosida (*after E. A. Minchin, 1900*). Fossil hexactinellid (*c*) *Hydnoceras tuberosum* Conrad, order Reticulosa, Devonian (*after J. Hall and J. M. Clarke, 1900*).

Recent hexactinellid sponges are chiefly upper bathyal marine animals and are most common in depths of 200–2000 m (660–6560 ft), although many species are known to inhabit lower bathyal depths. Only a few species are known to range into hadal depths, below 6000 m (19,685 ft), but there they may form dense aggregations. Confirmed records of the shallowest living hexactinellids are at 25–30 m (80–100 ft) in the Pacific Ocean near Victoria, British Columbia. The most diverse bathyal hexactinellid faunas are those in Antarctic seas, Indo-west Pacific, and North Pacific regions.

Morphology and physiology. Living hexactinellid sponges are commonly goblet- or vase-shaped, although branched, massive, tubular, or ropy-appearing sponges also occur in the class (**Fig. 1**). Many have root tufts of long spicules that anchor them in place and support them above the sea floor.

The body wall of hexactinellid sponges is relatively open and cavernous. Soft parts are suspended on the mineralized skeleton, and consist of inner and outer layers of syncytial filaments or trabeculae and an internal layer of variously spaced, thimble-shaped flagellated chambers, which are suspended in the reticulate networks of the other layers (**Fig. 2**). These syncytial filaments are composed of protoplasm that contains scattered nuclei, but they are not subdivided into distinct cells. The trabecular network covers all surfaces that come into contact with water flowing through a sponge, and may be a continuous cytoplasmic network throughout a sponge.

Trabeculae have no regular direction in the inner and outer networks, but form bounding dermal and gastral membranes on the external and internal margins of the wall. The dermal membrane is perforated by pores, which allow water to enter into and circulate through the sponge. Sponges may have a flagellate chamber system only or may possess a canal system for water circulation. Small-diameter inhalant canals may extend in from the dermal pores to flagellate chambers, and moderately distinct, larger-diameter exhalant canals may lead from those chambers in the gastral part of the wall. Those canals commonly allow water to exit into a spongocoel or cloaca and out one or more large exhalant oscula. Such canals are seen as gaps in the trabecular structure or as interruptions in the mineralized skeleton.

Skeleton and spicules. Spicules of the skeleton are composed of hydrated amorphous silica, close to opal in composition, but may include some organic matter as well. The silica is deposited around an organic axial filament, which has a square cross section (**Fig. 3*a***). Such filaments meet at the center of the spicule and form an axial cross. The silica may show a layered structure, presumably a result of discontinuous secretion. Spicules may be formed in all parts of the trabecular network by multinucleate scleroblast syncytia, at least where origins are known.

Spicules are separable into two types: larger main elements termed megascleres, and usually smaller

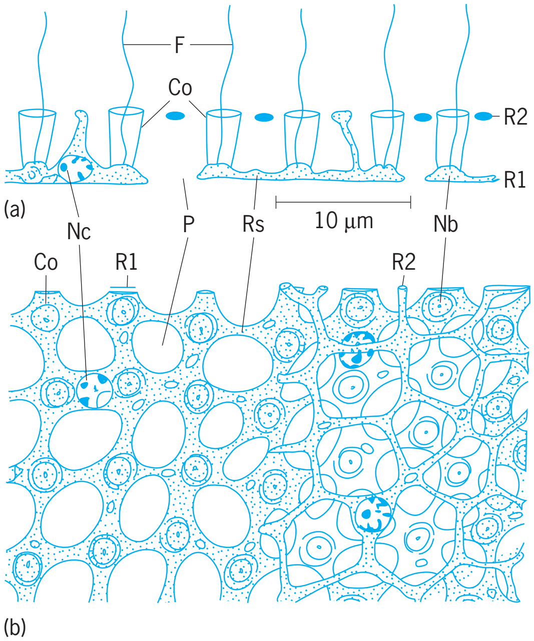


Fig. 2. Diagrammatic representation of sections through the choanosyncytium of *Aphrocallistes vastus* (*a*) in vertical section and (*b*) in tangential section from the interior of the chamber; the secondary reticulum is omitted for clarity. Co, collar; F, flagellum; Nb, nodal body; Nc, choanosyncytium nucleus; P, prosopyle; RS, connecting strands of main reticulum; R1, main reticulum of chamber wall; R2, secondary reticulum of chamber wall. (*Adapted from H. M. Reiswig, 1979*)

secondary elements termed microscleres (Figs. 3 and 4). Many variants of the basic triaxial hexactine megasclere (Fig. 3*c*) are known. For example, pentactines (pentacts) may result where one ray is reduced or lost (Fig. 3*d*), stauractines (stauracts) where two opposed rays are lost (Fig. 3*e*), or rhabdodiactines where all but two opposed rays are lost (Fig. 3*f*). Some such spicules still show a rudimentary axial cross, even in fossil examples, so their relationships to the basic orthotriaxial hexactine are clearly demonstrated. Other modifications may take place as well, such as in spicules where one ray becomes greatly elongated or enlarged, or where some rays are bent near their origins and curve away from the cubic axes. Spicules also may be modified by addition of solid spines that lack axial filaments (Fig. 3*b*, *g*, *i*). Such spines, for example, may be laterally directed on the distal ray of a pentactine to produce a conifer-appearing pinnule, or they may be thornlike to produce an uncinate of a rhabdodiactine.

Megascleres may be united in several ways to form rigid skeletal frameworks in some Hexactinellida (Fig. 3*j*–*l*). In the simplest skeletal framework, described as lyssacine, adjacent megascleres may be fused where rays cross one another or lie side by side. Commonly these rays are still recognizable as distinct elements. Where rays are close to one another but not in contact, they may be fused by siliceous rodlike synapticulae that bridge between them. Branched or interbraided siliceous filaments also may bridge between rays of associated spicules and fuse them together. Lyssacine frameworks may have loose, irregularly shaped meshes. In a dictyonine framework, overlapping rays or rays that meet tip to tip may be enveloped in a siliceous coating to form beams, in which individual rays are not identifiable except by their separated parallel axial filaments. This framework is usually a regular three-dimensional rectangular structure. The more complex lychniscose framework appears similar to a regular dictyonine framework, but has 12 small diagonal bracing struts developed at each axial megasclere node (Fig. 3*l*) to form an even more rigid skeletal framework.

Paleozoic hexactinellid sponges commonly have unfused megasclere spicules, but many still have their quadrate graph-paper-like skeletal structure preserved. Such fossilization may have been a result of their burial in relatively quiet environments, or it may suggest that their skeletons were held together by nonmineralized trabecular material or perhaps by unpreserved spongin. The skeleton of regularly arranged and ranked spicules like that in the Cambrian *Protospongia* is characteristic of early Paleozoic hexactinellids. Bundled unfused hexactines and pentactines, along with rhabdodiactines, produced the rectangularly three-dimensional skeletons of middle and late Paleozoic dictyosponges

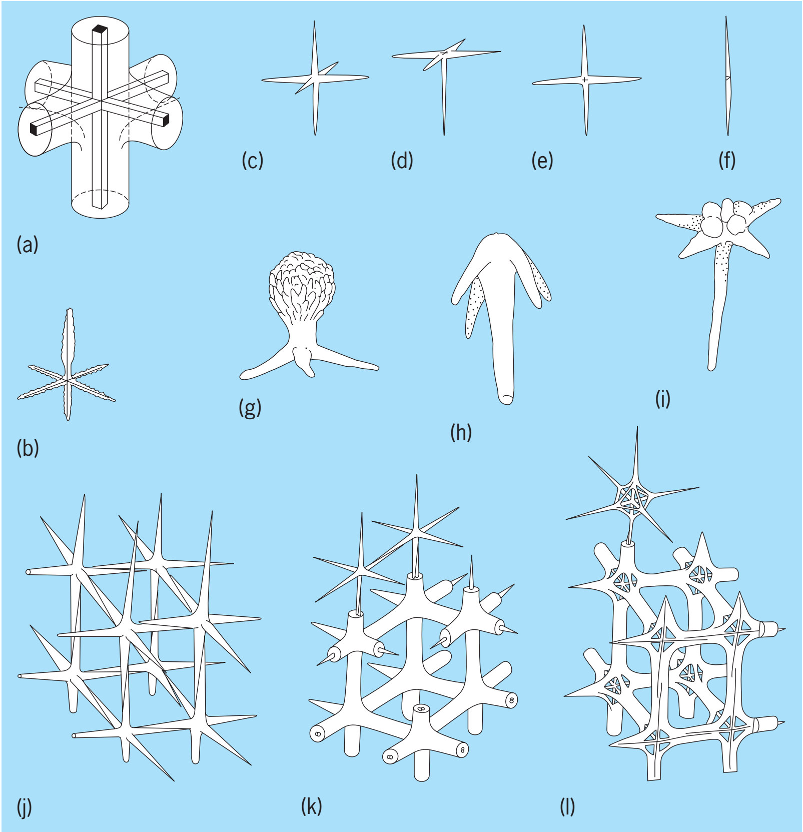


Fig. 3. Kinds of megasclere spicules and skeletal nets in the Hexactinellida. (*a*, *b*) Pattern of junction of coring axial filaments, with square cross sections, and spicule rays in pinnulated hexactine of *Sympagella nux* (*after H. M. Reiswig, 1971*). (*c*) Hexactine. (*d*) Pentactine. (*e*) Stauractine. (*f*) Rhabdodiactine. (*g*) Pinnular pentactine. (*h*) Anchorate root tuft pentactine. (*i*) Hexactine with nodes. (*j*) Lyssacinosid skeletal net with unfused hexactines. (*k*) Hexactinosid or dictyonine skeletal net with overlapping rays fused in siliceous coating and with simple spicule centers. (*l*) Lychniscosid skeletal net with fused rays and bracing struts at spicule centers. (*Parts j, k, l from J. K. Rigby, 1987*)

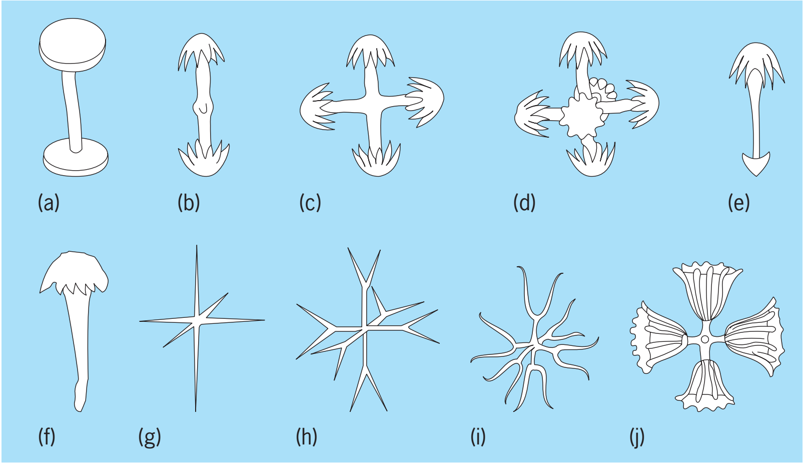


Fig. 4. Representative microsclere spicules found in the Hexactinellida. (*a–e*) Birotulates have umbrellalike terminations; (*f*) paraclavules have only one such termination; (*h*, *i*) hexasters are small hexactinal spicules. (*a*) Birotulate. (*b*) Amphidisc. (*c*) Staurodisc. (*d*) Hexadisc. (*e*) Hemidisc. (*f*) Paraclavule. (*g*) Microhexactine. (*h*) Hexaster. (*i*) Hexaster with curved terminations. (*j*) Floricome, with rays toward and away from the viewer omitted.

like *Hydnoceras* (Fig. 1*c*). Mesozoic and Cenozoic hexactinellid sponges, like Recent ones of the class, commonly have fused rigid skeletons.

Microsclere spicules are small (**Fig. 4**), commonly occur free within soft parts of the sponges, and are not united as part of the solid skeletal structure of the sponge. Microscleres typically range 10–100 micrometers long and 1 μm or less in diameter, only one-tenth the size of most megascleres. They may be small versions of megascleres, or they may be of two other basic types, hexasters or birotulates. Birotulates (Fig. 4*a–e*) are the small spicules with umbrellalike terminal expansions. Hexasters (Fig. 4*g–j*) are small hexactinal spicules that usually have branched ends, which may appear delicately flowerlike. Because microscleres are small and occur loose in the sponge, they are not as commonly preserved in fossils as are the megascleres. Taxonomically, microscleres are very useful spicules. The subclass Amphidiscophora is distinguished by having microscleres that are birotulates and never hexasters; whereas the subclass Hexasterophora has microscleres that are the small six-rayed hexasters, and lacks birotulates. Within the Amphidiscophora, sponges in the order Amphidiscosa have microscleres that are amphidiscs, and in some taxa also staurodiscs or hexadiscs, whereas sponges in the order Hemidiscosa have microscleres that are hemidiscs (Fig. 4*e*). Orders within the subclass Hexasterophora are differentiated on the basis of the firm rigid skeletal frameworks produced by the megascleres.

Taxonomy. The following classification is a combination of ones used in living and fossil sponges.

Class Hexactinellida
  Subclass Hexasterophora
    Order: Lyssacinosida
      Hexactinosida
      Lychniscosida
  Subclass Amphidiscophora
    Order: Reticulosa
      Amphidiscosa
      Hemidiscosa

The class Hexactinellida are exclusively marine sponges whose skeletons are made of siliceous hexactinal megasclere spicules with axial fibers or canals that have square cross sections, and microscleres that include hexaster or amphidisc microscleres; Precambrian–Recent.

The subclass Hexasterophora are hexactinellids in which microscleres are hexasters, and megascleres may be free and unconnected or, more commonly, fused to form a rigid skeleton; Ordovician–Recent. *See* HEXASTEROPHORA.

The order Lyssacinosida are hexasterophorans in which the megascleres are usually distinct and free, although some may be fused by supplementary siliceous elements; Ordovician-Recent.

The order Hexactinosida are hexasterophorans in which rows of hexactinal megascleres are fused by a supplementary coating of silica to form a rigid reticulate skeleton; Triassic-Recent. *See* HEXACTINOSA.

The order Lychniscosida are hexasterophorans in which the rigid framework of fused megascleres has each spicule node, termed a lychnisc, braced by 12 struts; Triassic-Recent.

The subclass Amphidiscophora are hexactinellids in which microscleres are birotules but never hexasters; megascleres are free and not fused to form a rigid skeleton; Precambrian-Recent.

The order Reticulosa are amphidiscophorans in which a dermal skeleton of parallel stauractines, pentactines, or hexactines forms a major part of the skeleton, and in which the microscleres include paraclavules; Precambrian–Permian. *See* RETICULOSA.

The order Amphidiscosa are amphidiscophorans in which the microscleres are equal-ended amphidiscs and not paraclavules; Ordovician-Recent. *See* AMPHIDISCOSA.

The order Hemidiscosa are amphidiscophorans whose principal microscleres are hemidiscs; Upper Pennsylvanian, Cretaceous-Recent. *See* HEMIDISCOSA.

There is general agreement on placement of most major taxa. However, the order Reticulosa is included in the subclass Amphidiscophora by paleontologists, based on the mode of occurrence of paraclavule microscleres (Fig. 4*f*) that indicates they functioned like amphidiscs in several Mississippian fossil sponges. The Reticulosa are considered by paleontologists to be an exclusively Paleozoic group of extinct sponges. However, some zoologists include the peculiar, branched *Sclerothamnus* as the only living and post-Paleozoic representative of the Reticulosa, and include the order in the subclass Hexasterophora.

Fossil record. Fossil hexactinellid sponges are known as impressions in the Precambrian Ediacaran beds of Australia and as isolated spicules from approximately equivalent beds in the Yangtze Gorge of China; but the earliest known body fossils with clearly preserved intact spicule nets are from lowermost Cambrian rocks in Hunan Province, China. The best example of an early sponge fauna is that of the Middle Cambrian Burgess Shale of western Canada, which includes not only well-preserved hexactinellid sponges but other types as well. *See* BURGESS SHALE.

Ordovician hexactinellid sponges document the beginning of thick-walled dictyosponges, which became major elements in Devonian and Mississippian faunas of North America when one major lineage developed a three-dimensional bundled skeletal structure. Essentially contemporary hexactinellids with disordered skeletons also developed but were not as diverse or as abundant as the dictyosponges. Both lineages persisted until the end of the Paleozoic.

The Permian-Triassic boundary marks a major break in the history of many groups of fossils, including the hexactinellid sponges. Reticulose amphidiscophoran sponges became extinct at the end of the Permian and, of the subclass, only the minor hemidiscosan sponges continued to the Recent. Hexasterophoran sponges, however, became abundant and diverse during the Mesozoic, although those with lyssacine skeletons have left only a poor record in Jurassic and younger rocks. However, dictyonine hexactinosid sponges, which have a fused reticulate skeleton, produced an abundant and diverse Jurassic and Cretaceous record, particularly in Europe, and are represented in modern seas by several surviving families. Lychniscosid hexasterophoran sponges also experienced great diversity and abundance in the Jurassic and Cretaceous of Europe, but only two of those families survived into the Cenozoic, and only one is known in modern seas. *See* PORIFERA. J. Keith Rigby

Bibliography. P. R. Bergquist, *Sponges*, University of California Press, Berkeley, 1978; J. Hall and J. M. Clarke, *A Memoir of the Paleozoic Reticulate Sponges Constituting the Family Dictyospongidae*, N.Y. State Mus. Mem. 2, 1888, 1900; J. N. A. Hooper and R. W. M. Van Soest (eds.), *Systema Porifera, A Guide to the Classification of Sponges*, vol. 2, Kluwer Academic/Plenum, New York, 2002; R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, pt. E (rev.), vols. 2 and 3, 2003, 2004; E. A. Minchin, Sponges: Phylum Porifera, pp. 1–78 in E. R. Lancaster (ed.), *A Treatise on Zoology*, pt. 2, chap. 3: The Porifera and Coelenterata, Adam and Charles Black, London, 1900; H. M. Reiswig, The axial symmetry of sponge spicules and its phylogenetic significance, *Cah. Biol. Mar.*, 12:505–514, 1971; H. M. Reiswig, Histology of Hexactinellida (Porifera), pp. 173–180 in C. Lévi and N. Boury-Esnault (eds.), *Biologie des Spongiaires—Sponge Biology*, Colloques Internationaux du Centre National de la Recherche Scientifique, 2901, Centre National de la Recherche Scientifique, Paris, 1979; J. K. Rigby, Phylum Porifera, pp. 116–139 in R. S. Boardman, H. Cheetham, and A. J. Rowell (eds.), *Fossil Invertebrates*, Blackwell Scientific, Palo Alto, CA, 1987; F. E. Schulze, Report on the Hexactinellida collected by H.M.S. "Challenger," *Challenger Report: Zoology*, 1887.

Hexactinosa

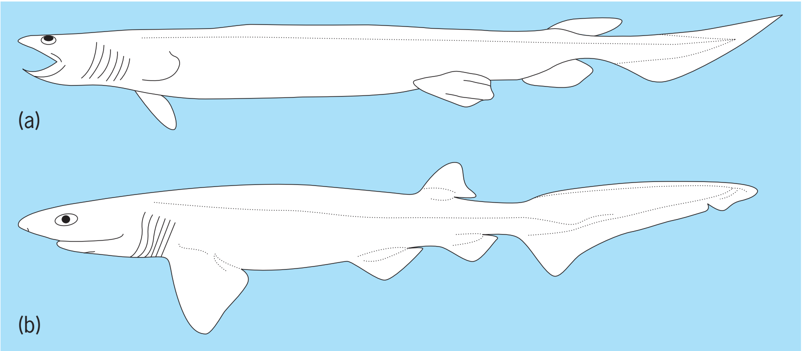
An order of sponges of the subclass Hexasterophora in the class Hexactinellida. The parenchymal megascleres in this order are united to form a rigid framework and consist wholly of simple hexactins which are arranged in parallel linear series. The members of each series are united one to another by a secondary envelope of silica. Examples include *Hexactinella*, *Aphrocallistes*, *Eurete*, and *Farrea*. *See* HEXACTINELLIDA; HEXASTEROPHORA. Willard D. Hartman

Hexanchiformes

An order of Squalomorpha that are known as the six-gill sharks. These sharks are distinguished by a combination of the following characters: six or seven gill slits, all anterior to the pectoral fins; a single dorsal fin without a spine; anal fin present (the only order of squalomorphs with an anal fin); the dorsal, anal, and pelvic fins placed far posteriorly; eyes lacking a nictitating fold or membrane (at the inner angle of the eye or below the eyelid that is capable of extending over the eyeball); spiracles present, but small and well behind the eyes; and ovoviviparous development. They comprise two extant families (chlamydoselachidae and Hexanchidae), four genera, and five species. *See* ELASMOBRANCHII; SELACHII.

Chlamydoselachidae (frill sharks). This family, comprising only one species (*Chlamydoselachus anguindus*), is characterized by six gill openings, with the margin of the first gill opening continuous across the throat; very elongate body; subcylindrical trunk; terminal mouth; and teeth with three elongate cusps that are similar in the upper and lower jaws. Maximum length is about 196 cm (77 in.) [**illus.** *a*].

The frill shark occurs on continental and insular slopes [usually at depths between 120 and 1280 m (393 and 4200 ft)] of the western North Atlantic, eastern Atlantic from Norway to South Africa, southwestern Indian, western Pacific, and eastern Pacific from California to Chile. It feeds on squids, bony fishes, and other sharks.



Examples of Hexanchiformes. (*a*) Frill shark. (*b*) Cow shark. (*J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006*)

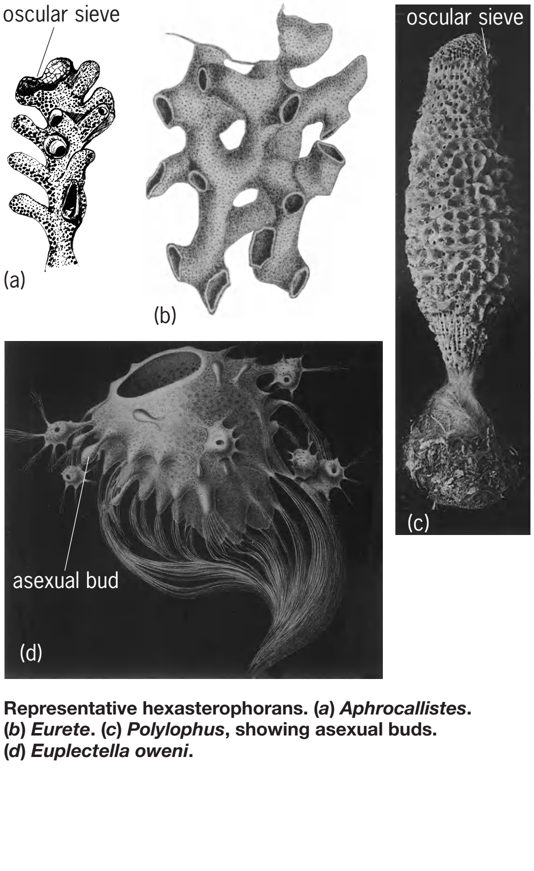
Hexanchidae (cow sharks). The cow sharks, comprising three genera and four species, are identified by six gill openings in the genus *Hexanchus*; seven gill openings in the genera *Heptranchias* and *Notorynchus* (each sometimes placed in its own family, Heptranchiidae and Notorynchidae, respectively); the margin of the first gill opening not continuous across the throat; inferior mouth (overhung by a long snout); upper and lower teeth different. Maximum length is about 470 cm (15 ft) [illus. *b*].

These sharks are circumglobal in temperate and tropical waters of continental and insular shelves and slopes. They feed on relatively large marine organisms, including other sharks, rays, bony fishes, crustaceans, and carrion. Herbert Boschung

Bibliography. M. R. de Carvalho, Higher-level elasmobranch phylogeny, basal squaleans, and paraphyly, pp. 35–62 in M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Hexasterophora

A subclass of sponges of the class Hexactinellida, in which the parenchymal microscleres are typically hexasters (small, six-rayed spicules, often with branched ends). This is a diverse assemblage of sponges commonly firmly fixed to the substratum by the base, less commonly anchored by means of basal spicule tufts or mats. The spicules of the body are sometimes free and unconnected, but the parenchymal megascleres are often fused to form a rigidly connected skeleton (see **illus.**) The following orders are recognized: Hexactinosa, Lychniscosa, Lyssacinosa, and Reticulosa. *See* HEXACTINELLIDA; HEXACTINOSA; RETICULOSA. Willard D. Hartman



Representative hexasterophorans. (*a*) *Aphrocallistes*. (*b*) *Eurete*. (*c*) *Polylophus*, showing asexual buds. (*d*) *Euplectella oweni*.

Hibernation and estivation

A general term applied to a condition of dormancy and torpor found in cold-blooded (poikilotherm) vertebrates and invertebrates. This rather universal phenomenon can be readily seen when body temperatures of poikilotherm animals drop in a parallel relation to ambient environmental temperatures. In the strict sense of the word, hibernation is a term which physiologists apply to relatively few species of mammals and birds (warm-blooded vertebrates).

Poikilotherm Animals

Hibernation occurs with exposure to low temperatures and, under normal conditions, occurs principally during winter seasons when there are lengthy periods of low environmental temperatures. Animals that hibernate naturally can be induced to do so under controlled laboratory conditions; consequently, hibernating animals are becoming useful as experimental subjects. The potential use of hibernators and animals in similary depressed metabolic states has attracted the attention of scientists in a variety of biological disciplines, including space biology and medical research. It has been found that hibernating animals resist lethal levels of radiation, that they show resistance to infection and parasitism, and that antibody and immune responses are also greatly altered. These examples represent only a few areas of active modern research utilizing hibernators.

A related form of dormancy is known as estivation. Many animals estivate when they are exposed to prolonged periods of drought or during hot, dry summers. For all practical purposes, hibernation and estivation in animals are indistinguishable, except for the nature of the stimulus, which is either cold or an arid environment.

There is no complete list of animals that hibernate; however, many examples can be found among the poikilotherms, both vertebrate and invertebrate. The poikilotherms are sometimes referred to as ectothermic, because their body temperatures are not internally regulated but follow the rise and fall of environmental temperatures. During hibernation and winter torpor, body temperatures reflect the environmental temperature, often to within a fraction of a degree. Among the classic examples of hibernators or estivators are reptiles, amphibians, and fishes among the vertebrates, and insects, mollusks, and many other invertebrates.

Reptiles and amphibians. Many terrestrial reptiles, such as lizards, snakes, and turtles, become dormant and hibernate by burrowing in crevices under rocks, logs, and in the ground below the frost line. Terraqueous turtles also become cold-torpid and may often be found completely submerged in mud and in ponds under ice.

Since the hibernating reptile is subject to the caprices of duration of seasonal low temperature, there is no well-defined period of dormancy. The period of hibernation may often be related to latitudinal positions, as evidenced by the turtle family Emydae. These are aquatic and semiaquatic species, some of which range in the United States from northern latitudes southward through the Mississippi Valley. Species that inhabit the northern climes will hibernate longer than their southern relatives, thus showing hibernation periods which are proportional to the length of the winter period. The physiological characteristics of these hibernating reptiles show many striking alterations of body chemistry. Hibernating reptiles show a loss of appetite and discontinue the ingestion of food. Although the metabolic rate is reduced as much as 95% in hibernating turtles, there is some utilization of stored food products. There are two principal types of reserve food: (1) lipids, which are often found in solid fatty masses in the viscera, under the skin, and distributed in tissues such as muscle, kidney, and particularly liver; and (2) glycogen, the animal starch, which is less stable and more rapidly used than fats. Glycogen is generally localized in tissues such as liver and muscle. There is evidence that these reserve foods are selectively utilized. In hibernating turtles, the tissue glycogen is used during the initial days and weeks of hibernation; later, the lipids are utilized.

The Crocodylia, such as caimans, crocodiles, and alligators, are least able to hibernate. These live in tropical regions where the winter temperatures are relatively uniform. The American alligator (*Alligator mississippiensis*) becomes lethargic and stops eating for several months during winter. This animal undergoes other seasonal physiological modifications as is evidenced by a winter hypoglycemia, that is, a reduction in blood sugar level. The blood sugar level of animals is a physiological index of metabolism and hormonal balance. A major hazard to hibernating poikilotherms is death from freezing; ice crystals form in free protoplasmic water and ultimately destroy the cells and tissues, causing the death of the animal. Frogs, salamanders, and turtles are often surrounded by ice while in hibernation. These animals are able to survive, despite the reduction in body temperatures to about 32 to 31°F (0 to −1°C). As winter approaches, the water content of the tissues becomes reduced and the blood more concentrated. Experiments with turtles have shown that, during winter, there is a lowering of the freezing point of blood. Insects are also credited with the ability to lose body water during winter. Therefore, any physical adjustments which these animals are able to make to rid themselves of water will aid their survival. Hibernating frogs die if, during a severe winter, the frost penetrates their mud hibernacula. Frequently, the uppermost layers of mud freeze and so do its inhabitants, while the animals buried further down in the unfrozen mud remain safe. Frogs, as well as fishes, often freeze to death in shallow artificial pools made of concrete and lacking a mud substrate.

In consideration of the geographical distribution of the amphibians, the limitations imposed by permafrost and the frozen tundra of the northern latitudes can readily be realized. Amphibians are not ordinarily found in permafrost climates.

Fishes. Hibernation in fishes does not occur. Many fishes do, however, spend much of the winter in a state of quiescence while partially frozen in mud and ice. Shallow ponds in temperate, subarctic, and Arctic latitudes often freeze solid in winter. Fishes such as the carp may become encased in the ice or partially buried in the bottom mud. Thus they remain torpid for lengthy periods. The Alaskan black fish (*Dallia pectoralis*) has received some fame, regarding its ability to "freeze solid." Actually, *Dallia*, or any other fish, does not freeze, but becomes encased in ice. If the fish were to freeze solid, with the production of ice crystals in its tissues, it would die. Fishes exposed to cold have a much-reduced metabolism; therefore their oxygen requirements are readily adjusted to the oxygen content of the cold waters. The situation is similar to that which occurs in hibernating aquatic turtles.

The phenomenon of estivation is best known in the dipnoans, that is, the lungfishes. These fishes (*Neoceratodus* in Australia, *Protopterus* in Equatorial Africa, and *Lepidosiren* in South America) are restricted to tropical regions marked by repeated seasons of drought. They survive the dry seasons by becoming dormant and torpid. The lungfishes and their relatives were abundant during the Paleozoic Era. They are, therefore, among the more primitive air-breathing animals possessing a lung which utilizes atmospheric oxygen. This lung becomes the primary organ of respiration during the torpidity of estivation. In general, the lungfishes follow a similar behavioral pattern as the dry seasons approach. *Protopterus*, for example, burrows in the bottom mud as the water begins to diminish during the dry season. Once in its burrow, the fish assumes the coiled posture of seasonal dormancy (**Fig. 1**). A lifeline of air is provided by the tunnel from the burrow to the surface. In preparation for estivation, *Protopterus* secretes a slimy mucus around itself which hardens in a tight cocoonlike chamber. This chamber forms an effective hygroscopic structure which prevents the desiccation of the fish. There is but one opening, formed around the mouth. Thus the air from the tunnel enters the mouth and passes to the lung apparatus. At the termination of the dry season, water slowly enters the burrow, softens the contents, and awakens the lungfish. The metabolism of the lungfish is at a low ebb during estivation, with the energy for its modest life processes provided by the utilization of tissue protein. This is evidenced by the large increase in the formation of urea, a product of protein metabolism. The urea is promptly excreted in large quantities when the fish returns to water. *Lepidosiren*, the South American member of the family of lungfishes, sometimes called the mud siren, also

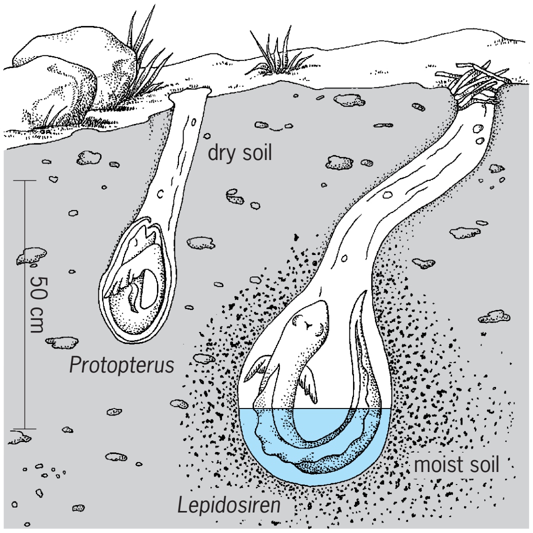


Fig. 1. Lungfishes in estivation.

burrows. The burrow entrance is loosely plugged with surface clay, perforated with air holes, each about 0.5 in. (1.25 cm) in diameter. Mud sirens burrow in the deepest and most concave bottoms, seemingly to ensure that they will receive the inflowing water when it first returns to the pond. The fishes' tunnellike burrows are about 2–3 in. (5–7.5 cm) in diameter and extend downward about 2–3 ft (0.6–0.9 m). The mud siren estivates in a gourd- shaped, gelatinous-coated burrow which may be partially filled with water. These fishes, therefore, differ from their African relatives in that they maintain a soft, moist medium. Lungfishes in estivation are readily handled and transported, indicating the extremely quiescent nature of their torpor.

Invertebrates. This is often more than a period of seasonal dormancy, because in some snails estivation may be extended for years at a time, and among the insects and spiders the period of hibernation becomes intimately associated with a phase in the life cycle. During the winter months and during a hot dry summer, the soil contains a remarkable variety of torpid invertebrates, for example, earthworms, snails and slugs, nematodes, insects and spiders, grubs, larvae, and pupae of many insects, egg cases, and cocoons. The hibernacula for dormant invertebrates are also provided by rotting logs and vegetation, the undersurfaces of bark, and crevices in rock and building foundations.

Insecta. Insects overwinter, for the most part, in the egg or larval stage of metamorphosis (egg→larva→pupa→adult). Hibernation frequently becomes integrated with the diapause, or arrested development, of the egg or larva which occurs during the winter. The egg of the silkworm (*Bombyx mori*) undergoes diapause in conjunction with hibernation. The eggs are laid in autumn and must be exposed to about 32°F (0°C) for several months to ensure hatching in the spring. The Syrphidae, commonly referred to as the hover flies, are ordinary garden inhabitants. They live in close association with plants, as commensals and parasites, and are also known to infest animals, producing myiasis. The larvae of the syrphids are particularly susceptible to hibernation and estivation. They become markedly desiccated but are quickly reactivated by exposure to water. The Lepidoptera, butterflies and moths, offer some of the most common examples of hibernating larvae and pupae. The familiar cocoon of the butterfly is the hibernaculum of the larva and pupa. The disappearance of mosquitoes in the Arctic, subarctic, and temperate zones in autumn and winter is also a reflection of insect hibernation. The adult hibernates, after its last autumnal blood meal, and presumably subsists on its stored food reserves. Provision for overwintering in *Culex* and *Anopheles* mosquitoes lies in their development of a reserve food material in the form of a fat body. *See* INSECT PHYSIOLOGY; MYIASIS.

Mollusca. Land mollusks, such as snails and slugs, are noted for their ability to hibernate and estivate. Land snails are particularly cold-hardy and have frequently been found encased in ice. Among the more cold-resistant land snails are species of *Physa*, which normally live in northern Siberia, 73°30′N latitude, where the mean annual temperature is below −4°F (−20°C). In the subarctic and temperate zones most fresh-water mollusks can live in water under the ice and are often enclosed in solid ice. The common garden snail (*Helix aspersa*) and the edible snail (*H. pomatia*) exhibit the chief characteristics of hibernation common to the Gastropoda. In autumn, these snails cease ingesting food and burrow into the ground under logs, leaves, and other forms of natural debris, where they remain in hibernation for about 6 months. The animal makes this burrow with its muscular foot and settles down to a winter sleep. The hibernating position is such that the shell aperture is turned upward. The hibernating snail may be found singly or in clusters, with numerous individuals in contact with each other. The animal seals its aperture by secreting a membrane, the epiphragma, chiefly composed of calcium phosphate. In the epiphragma, there remains a minute respiratory opening. The respiratory movements are drastically reduced and, as in other poikilotherms, the rate of heart beat is related to temperature; at 86°F (30°C) the heart rate is 50–60 beats per minute, whereas in hibernation the rate is reduced to about 4–12 per minute. The energy for the lowered metabolism during hibernation appears to be derived from the utilization of stored fats, with a subsequent loss of body weight.

The slugs, referred to as garden snails without shells, are also hibernants. Their hibernaculum is similar to that of the shelled forms. They hibernate in the ground under leaves, tree bark, and other forms of decomposing vegetation. Estivation of land mollusks occurs chiefly in tropical climates, where they bury themselves deeply in the ground or under rocks. They also form a type of epiphragma as a defense against evaporation and desiccation. Tropical snails are able to estivate for lengthy periods, months and possibly years at a time. A speciman of *Helix desertorum* is reputed to have awakened after 5 years of dormancy. *See* MOLLUSCA.

The two marine mollusks periwinkles (*Littorina*) and mussels (*Mytilus* and *Modiolus*) cannot be regarded as hibernators, yet they are extremely cold-hardy These animals are inhabitants of the intertidal zone. In such latitudes as Woods Hole, Massachusetts, and Hebron, Labrador, they may be exposed repeatedly to temperatures of −4°F (−20°C) in winter. At such temperatures, ice forms inside the shelled animals. These seashore mollusks can live through the winter season with as much as 70–75% of the body cavity being recurrently filled with ice. At such low temperatures the animals are being dehydrated, since a great deal of the body water is being converted into unusable ice and the remaining fluid becomes a more concentrated salt solution. In these animals, the body water which freezes is not tissue or cellular protoplasmic fluid, but is chiefly extracellular water.

Nemata. Nematodes are cosmopolitan and have been found in practically all ecological settings. However, the terrestrial and fresh-water forms are most apt to respond to winter exposure by becoming quiescent. Pasture puddles, which often freeze solid in winter, harbor a particularly rich nematode fauna. Literally millions per acre are found in the top few inches of soil. Among the "cushion" plants such as mosses and lichens one finds frequent nematode "parasitism." In addition to a rich fauna of nematodes, these plants also harbor tardigrades, rotifers, and even protozoans. These animals are characterized by their capacity to endure the extremes of drought, heat, desiccation, and cold. Under such forms of environmental stress they go into a dormant state, and revive when environmental conditions become less severe and again conducive to active life. The particular nematodes such as *Mononchus, Dorylaimus, Plectus, Monhystera, Cephalobus, Trilobus, Trypyla, Tylenchus,* and *Aphelenchus* are considerably alike throughout the world, and in spite of wide longitudinal and latitudinal separation their overwintering and drought behavior are similar.

Their resistance notwithstanding, nematodes do die and decline in numbers during the winter. *See* NEMATA (NEMATODA); ROTIFERA; TARDIGRADA.

Bryozoa. The Ctenostomes are the only Bryozoa with fresh-water representatives. These animals bear a resemblance to moss and seaweeds and are collectively called moss animals. The colonial fresh-water *Paludicella* and *Victorella* form hibernacula. These are specialized external buds which persist through the winter while the rest of the colony breaks down. At the termination of winter, the shell of the hibernaculum splits into two halves and a young colony is brought forth. *See* BRYOZOA.

Protozoa. The phenomenon of encystment is commonplace in the Protozoa, or single-celled animals. Encystment is remarkably similar to estivation and hibernation, and an encysted protozoon is extremely quiescent and almost nonmetabolizing. It is generally accepted that encystment is a protective measure against cold and desiccation, and yet it may coincide with some general physiological, protoplasmic reorganization or dedifferentiation in the species. For example, during encystment of the ameba *Pelomyxa carolinensis*, there is a reduction in number of nuclei, a concentration of protoplasm, and the formation of a cyst membrane and shell. The ciliate *Colpoda cucullus* during its encystment undergoes nuclear reorganization, and multiplication of the organism takes place. The ability of protozoa to encyst is at least one of the reasons for their worldwide distribution. The Arctic tundra, which remains frozen for 6–8 months of the year, teems with protozoa as well as other invertebrate life. These microscopic animals could not live through these winter rigors unless they had this capacity for self-preservation and protection.

The free-living protozoa offer many examples of encystment due to winter low temperatures and to evaporation due to drought. The cyst may be regarded as the hibernaculum, and it is usually found in mud or ice and soil. Cysts of the single-celled organisms can be recovered from ice and mud samples taken in midwinter and also from dried soil collected during the most arid summers. The protozoan cyst is often, as with many invertebrates, a function of reproduction and therefore a stage in the life cycle of the species. The ciliates *Colpoda* and *Didinium* are noted for their cyst formation. Encysted *Colpoda* can resist experimental freezing for as long as 2 months, and can be activated after being in a dried state for as long as 5 years. *See* PROTOZOA.

Hibernacula. The hibernacula of poikilotherm vertebrates and invertebrates are as varied as the animals themselves (**Fig. 2**). The minute cysts in protozoa, the cocoon and egg case of insects and spiders, the burrows and crevices of reptiles, and the dried mucous case of the lungfish, in all instances, protect the animal from evaporation or desiccation and freezing. X. J. Musacchia

Warm-Blooded Vertebrates

Many mammals and some birds spend at least part of the winter in hiding, but remain no more drowsy than in normal sleep. On the other hand, some mammals undergo a profound decrease in metabolic rate and physiological function during the winter, with a body temperature near 32°F (0°C). This condition, sometimes known as deep hibernation, is the only state in which the warm-blooded vertebrate, with its complex mechanisms for temperature control, abandons its warm-blooded state and chills to the temperature of the environment. Between the drowsy condition and deep hibernation are gradations about which little is known. The bear, skunk, raccoon, and badger are animals which become drowsy in winter. Although usually considered the typical hibernator, the bear's body temperature does not drop more than a few degrees. Body temperature, measured rectally, was 96°F (35.5°C) at an air temperature of 40°F (4.4°C). At 25°F (−3.5°C), body temperature was 88°F (31.2°C).

Deep hibernators. The deep hibernators are confined to five orders of mammals: the marsupials, the Chiroptera or bats, the insectivores, the rodents, and, probably, the primates. Some of the South American

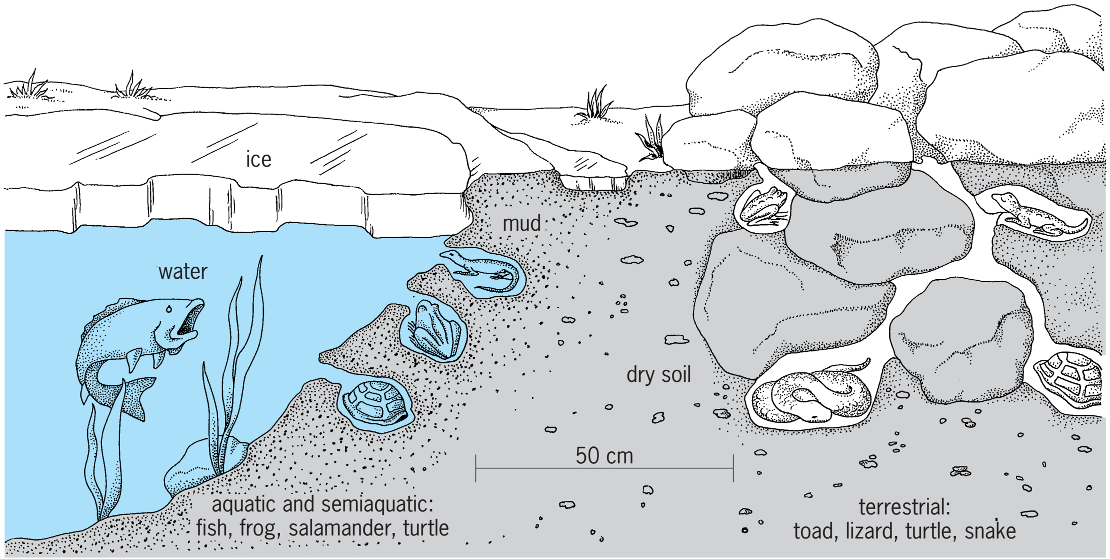


Fig. 2. Hibernacula of various cold-blooded vertebrates.

opossums are reported to hibernate. Most, if not all, of the insect-eating bats of temperate climates not only hibernate in the winter, but also drop their body temperature when they roost and sleep. The advantage of this for a small mammal with a disproportionately large heat-losing surface is obvious when conservation of energy is considered. Some monotremes and marsupials hibernate. Among the insectivores, the European hedgehog is a deep hibernator, and the Madagascan tenrec (*Centetes*) probably hibernates or estivates during the summer. Many rodents are deep hibernators, including ground squirrels, woodchucks, dormice, and hamster. The fat-tailed and mouse lemurs are primates that hibernate or estivate. Among birds, the poorwill (*Phalaenoptilus*) and some hummingbirds and swifts undergo a lowering of body temperature and metabolic rate in cold periods.

With all deep hibernators, except the bats, hibernation is seasonal, usually occurring during the cold winter months. In all cases, it occurs in animals which would face extremely difficult conditions if they had to remain active and search for food. A few desert-living species disappear during the driest, hottest months, and it is probable that they enter into a stage of deep hibernation in their cool burrows, though no studies have been made on this, or on the estivation of animals during the tropical rainy season. Hibernators are restricted typically to more temperate zones, but the Alaskan ground squirrel lives near the Arctic Circle.

Physiology. During a preparation period for hibernation, the animals either become fat, like the woodchuck, or store food in their winter quarters, like the chipmunk and hamster. Prior to hibernation, there is a general involution of the endocrine glands, but at least part of this occurs soon after the breeding season and is not directly concerned with hibernation. Animals such as ground squirrels become more torpid during the fall, even when kept in a warm environment, indicating a profound metabolic change which may be controlled by the endocrine glands. On the other hand, ablation of any of the glands has not brought on hibernation. Lack of food will cause hibernation in the spiny pocket mouse (*Perognathus*), but in most hibernators this has little if any effect, and the stimulus for hibernation is not known. It has been reported that an extract from the blood of an animal in hibernation will induce hibernation when infused into an active potential hibernator, indicating that the factor which produces hibernation may be bloodborne.

Neural mechanism. Hibernation in mammals is not caused by an inability to remain warm when exposed to cold, for hibernators are capable of very high metabolic rates and sometimes do not enter hibernation if exposed to cold for months at a time. When the animal is entering hibernation, heart rate and oxygen consumption decline before body temperature, indicating that the animal is actively damping its heat-generating mechanisms. The autonomic nervous system is involved in this process, for the heart is slowed to a lesser extent in animals entering hibernation when their parasympathetic system is blocked by atropine. As normal hibernation deepens, the heart rate, blood pressure, metabolic rate, and body temperature slowly drop, but in some animals periodic bouts of shivering and increased oxygen consumption occur, elevating the body temperature temporarily and causing a stepwise entrance into hibernation. Presumably this is caused by bursts of activity of the sympathetic nervous system, which otherwise remains in abeyance during entrance into hibernation. The preoptic-hypothalamic area of the brain is the center for temperature control in mammals. Ground squirrels with lesions in this area will enter into a state resembling hibernation, but detailed physiological comparisons of lesioned and normal hibernators have not been made. *See* AUTONOMIC NERVOUS SYSTEM.

Homeostasis. In deep hibernation at a steady state the body temperature is 33–35.5°F (0.5–2°C) above that of the environment, and it is a peculiarity of hibernators that the vital processes can function at lower temperatures than those of nonhibernators. The heart rate varies between 3 and 15 beats per minute, and blocking or stimulating the vagus nerve has no effect on the heart rate, indicating that the parasympathetic nervous system cannot be regulating the changes in heart rate. The metabolic rate is less than one-thirtieth of the warm-blooded rate at rest, and the main source of energy is fat, since the respiratory quotient (RQ) is about 0.7. An RQ above this figure would indicate that proteins or carbohydrates were being used. The torpid black bear, with its relatively high body temperature, also uses fat for energy, whereas the starved, active bear uses fat and protein.

In spite of its low body temperature, the hibernating animal retains a remarkably rigid control of its internal environment. Blood sugar levels are low in some hibernating animals, but there is no known consistent variation in any of the blood constituents which can be reasonably considered as a cause of the hibernating state. The hibernating animal senses changes in its environment and responds to increases in atmospheric CO_2 by increasing its respiratory rate. This sensitivity to the environment varies with time, even when body temperature remains the same, but the causes of this variation are not known. In its most sensitive state an animal in hibernation will respond with muscular activity to the movement of one of its hairs or to a light breeze playing over its back.

Temperature control. If the environmental temperature drops to 32°F (0°C), the hibernating animal may respond either by increasing its metabolic rate and remaining in hibernation or by a complete arousal from the hibernating state. Chilling the preoptic-anterior hypothalamic area of the brain in hibernating golden-mantled ground squirrels causes an increase in metabolic rate, which varies linearly with the decrease in the temperature of the brain, indicating that the hibernating animal is truly regulating its body temperature. Unlike animals in the warm-blooded state, hibernating hamsters and ground squirrels do not sense the increasing cold with peripheral receptors, for heart and metabolic rate do not increase as chilling proceeds if the brain temperature does not decline. However, dormice, which hibernate in nests exposed to rapid temperature variations, increase heart and metabolic rate when the periphery is chilled and brain temperature remains unchanged. Hibernating marmots also sense peripheral cold. In this case the temperature sensors in the brain are insulated from the cold by the curled position of a relatively large animal, and the marmot may need peripheral input to avoid freezing. Thus, some animals at the low temperatures of hibernation appear to have the same reactions to a cold environment as mammals in the warm-blooded state. Between an environmental temperature of about 39 and 59°F (4 and 15°C) the hibernating animal makes no attempt to regulate body temperature at a preferred level, and in this aspect it differs from the warm-blooded animal, which regulates its temperature between a low and a high point. If hibernation does involve lowering the "set point" in order to control body temperature, the method by which this is accomplished is unknown.

Temperature and pH effects. A hibernating mammal reduces its metabolic rate by nearly 30-fold and shifts from glycogen to lipid (that is, fat stores) as the major fuel source for metabolism. The magnitude of metabolic rate reduction is far in excess of what would be expected solely as a result of a hibernator's lowered body temperature (T_b, down from 98.6 to 44.6°F, or 37 to 7°C, in some species). Moreover, suppression of glycogen metabolism during hibernation must be poised for regular and rapid relaxation during periods of arousal (which are fueled by glycolysis) as well as at the end of the hibernation period.

Mechanisms controlling these aspects of hibernation metabolism appear to be the relative acidification of the intracellular fluids of the hibernator. This is a consequence of the hibernator's tendency to continuously regulate its blood pH (at about pH 7.4, termed pH stat) while its T_b drops, and of the adoption of a modified breathing pattern that, although variable among species, is typified by periods of apnea lasting up to 2 h that are interspersed between 3–30 min intervals of rapid ventilation.

Just as the neutral pH of water changes with temperature due to shifts in its dissociation constant, varying the T_b of an ectotherm causes shifts in the pH of its blood and intracellular fluids, at an average of −0.015 pH unit/°C. At 98.6°F (37°C) T_b, the blood pH of most ectotherms would be about 7.4, and about 7.85 at 44.6°F (7°C) T_b. Due to both the physical nature of intracellular fluids and the ectotherm's homeostatic and metabolic compensation for these thermally caused pH shifts (that is, mainly through the regulation of total CO_2), the histidine imidazole buffer groups of many of its regulatory proteins (enzymes) gain or release protons in order to maintain a constant dissociation ratio. By preserving the dissociation state (termed alpha stat regulation) of a key enzyme, its activity and in turn the metabolic process it catalyzes could remain near optimal at the specific combinations of intracellular temperature and pH imposed by the range of T_b experienced by the ectotherm. *See* PH REGULATION (BIOLOGY).

Since hibernating mammals tend to maintain their blood pH near 7.4, blood and most intracellular pH values (except those in the heart and liver which may have to function normally to sustain life at such low rates) are acidotic relative to the temperature-pH relationships known for ectotherms. Alpha stat regulation does not occur, and increased protonation of the imadazole buffers leads to changes in the kinetic and structural properties of enzymes, depending upon their pH and temperature sensitivity. Phosphofructokinase (PFK), a key enzyme in the control of glycolysis, becomes completely inactivated at the intracellular temperature (43°F or 6°C) and pH (7.0) conditions prevailing in the muscle of a hibernating ground squirrel *(Spermophilus)*. Glycogen metabolism is similarly reduced in hibernating bats *(Eptesicus)*, even though their rate of fat oxidation is unaffected by this state.

Inactivation of PFK by imidazole protonation conserves glycogen stores, and may account for the large decrease in the metabolic rate of hibernators. Also, since it occurs rapidly and is reversible, this deactivation would permit arousal metabolism. In addition to the temperature-pH effect, PFK is also inhibited by low amounts of urea. Also, apnea (through CO_2 retention) causes acidosis, which apart from a large drop in T_b could inactivate PFK. This suggests a glycogen-energy–serving strategy used by mammals that hibernate or enter daily torpor (for example, tropical bats) in warm habitats (where T_b would never drop to 44.6°F or 7°C). Through apnea or rebreathing of hibernacula air containing high CO_2, PFK inactivation might be induced. Finally, the different effects of the pH stat and alpha stat strategies on energy metabolism and tissue function have important biomedical implications for human hypothermic surgery. Jeffrey B. Graham

Arousal. The hibernator is capable of waking at any time, using self-generated heat, and this characteristic clearly separates the hibernating state from any condition of induced hypothermia. During the total period of hibernation, the hibernator spontaneously wakes from time to time, usually at least once a week. In the period of wakefulness the stored food is evidently eaten, but animals which do not store food rely on their fat for the extra energy during the whole winter. The cause of the periodic arousals has not been definitely determined, but it is theorized that the arousal is due to the effect of the accumulation of a metabolite or other substance which can be neutralized only in the warm-blooded state.

Arousal is a complex, coordinated event which results in warming in the shortest time, usually about 3 h when in an environmental temperature of 41°F (5°C). As soon as the animal starts to rouse, the heart speeds and oxygen consumption rises, followed shortly by an increase in body temperature. The anterior parts of the body warm rapidly, but because of vasoconstriction, the posterior parts remain cold until late in the warming process. Shivering can be detected early, using the electromyogram, and it becomes more obvious as the animal warms. The energy for warming, at least in rodents, is obtained from glycogen stored in the liver and other tissues. It has been calculated that one waking uses as much energy as does 10 days of hibernation, so that waking must use most of the winter stores, whether fat or fodder. Charles P. Lyman

Anaerobiosis

For many ectothermic vertebrates (fishes, amphibians, and reptiles) the ability to avoid seasonal and periodic environmental rigors by entering a state of metabolic inactivity is a crucial element in their survival. Specifically, winter dormancy and summer estivation—the usual context in which these terms are applied to ectotherms—permit these animals to survive and flourish, first, by reducing the impact of seasonal extremes (such as cold, heat, dryness, and ice) and, second, by significantly lowering the ectotherm's energetic costs during times that would not be favorable for activity (that is, when food is available).

Regulating factors. As in hibernating endotherms (birds and mammals), a key factor regulating seasonal torpor in ectotherms is the continuous internal monitoring of environmental cues (such as day length) which in turn triggers temporally precise seasonally adaptive changes in systemic function, metabolism, and behavior. Thus, both the desert-inhabiting spadefoot toad, *Scaphiopus couchi*, and the horned lizard, *Phrynosoma*, are obligatory hibernators; they enter winter dormancy at a fixed time each year, even if local conditions might favor their activity for a slightly longer period. A second important factor is the presence in ectotherms of a bioenergetic metabolic system that, when compared to mammals and birds, operates at a much lower intensity and has less absolute dependence on molecular oxygen. The metabolic energy adaptations for seasonal torpor in ectothermic vertebrates are to a large extent similar to those required by vigorous activity or prolonged diving, and thus involve the processing or storage of intermediate metabolitics such as lactic acid, the regulation of intra- and extracellular pH, and enduring periods without access to oxygen.

Energy production. Both endotherms and ectotherms acquire energy through the metabolic breakdown of stored food reserves (carbohydrates, fats, and protein) and the transfer of the released chemical energy into molecules of adenosine triphosphate (ATP). Both anaerobic (glycolysis) and aerobic (oxidative phosphorylation) metabolic pathways are used in ATP production, and in most cells these two processes are coupled. The product of glycolysis, acetyl coenzyme A (AcCoA), flows directly into the aerobic pathway, where it is fully oxidized to water and carbon dioxide. Glycolysis has a much lower rate of ATP yield per volume of substrate than does oxidative phosphorylation; but it will proceed in the absence of oxygen, whereas oxidative phosphorylation cannot. Cells metabolizing without oxygen convert AcCoA into lactic acid, the anaerobic storage product, until oxygen becomes available. For example, during vigorous activity, the short-term energy utilization by an animal's muscle far exceeds oxygen availability and, following activity, the animal repays the oxygen debt accrued by this tissue in the oxidation of the lactic acid. In ectotherms, glycolytic energy production contributes from as little as 50% to as much as 98% of the total energy utilized during activity, which is a much higher proportion than in endotherms. Even in the presence of oxygen, however, aerobic glycolysis occurs, and the AcCoA produced is not always shunted into oxidative phosphorylation. Except that constant low levels of lactic acid are present in the tissues, little is known about resting levels of aerobic glycolysis in animals. However, prolonged confinement in burrows or in a submerged hibernaculum may require an ectotherm to suspend oxidative metabolism and rely solely upon anaerobic metabolism. *See* ENERGY METABOLISM; METABOLISM.

Terrestrial ectotherms. Subterranean refugia permit escape from extreme surface temperatures, and cool subsurface temperatures reduce metabolism. During winter dormancy, spadefoot toads buried 20 in. (50 cm) in sand lower aerobic metabolism to 20% of the resting level. Some toads and frogs envelop themselves in cocoons made from many layers of unshed skin. In summer the yellow mud turtle *Kinosternon* leaves drying ponds to estivate in a dry burrow for as long as 3 months. During estivation, oxygen consumption declines by 75% and the animal becomes anaerobic. Upon emergence, the mud turtle quickly restores its considerable water losses, unloads a large volume of carbon dioxide, and then resumes a normal rate of oxygen consumption.

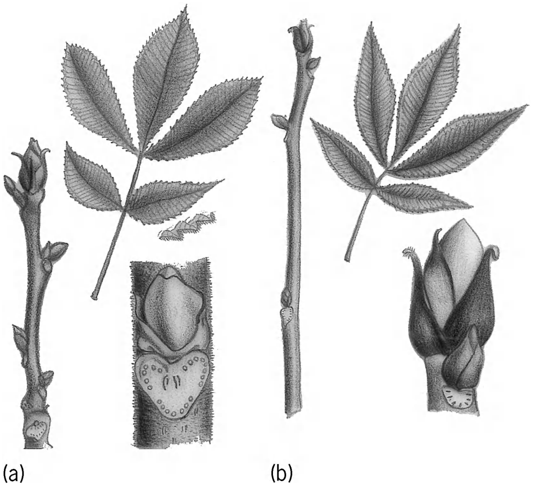
Aquatic ectotherms. Seasonal inactivity in aquatic ectotherms imposes an additional set of hardships in which specializations related to diving come into play. Winter-dormant turtles (*Pseudemys*) and bullfrogs (*Rana*) have subsurface burrows that penetrate deeply into anoxic mud; thus they are without access to oxygen. The tropical marine turtle *Chelonia* enters a temperature induced torpidity and remains submerged, sometimes in anoxic areas, for considerable periods. *Pseudemys* can survive long periods without oxygen.

If lactic acid is stored in the body, it must be neutralized or the tissue pH will be affected. This is usually done by bicarbonate and chloride, but in turtles (*Chrysemys*) kept submerged without oxygen for 3 months lactate levels become so high that calcium and magnesium must be mobilized from its bone and shell to mitigate the effect. In fact, anoxia stimulates a compensatory lowering of the amounts of cytochrome oxidase, an important aerobic enzyme, in its brain, heart, and muscle. A key glycolytic enzyme, pyruvate kinase, undergoes an increase in the turtle's muscles, but not in its heart. A great many fishes endure hypoxic and anoxic conditions under winter ice, in stagnant swamps that may also become dry, and in deep lakes and ocean basins. *Rasbora*, a cyprinid, can survive over 100 days without oxygen. Goldfish and carp (*Carassius*) can overwinter under ice without oxygen for from 2 to 5 months. Goldfish are seasonally adapted to these conditions. In anoxia this fish dramatically rearranges its anaerobic metabolism, and lactic acid does not concentrate. Rather, it is remetabolized to AcCoA and carbon dioxide. AcCoA is in turn broken down to ethyl alcohol which, along with carbon dioxide, is excreted by the anoxic goldfish. Even some air-breathing fish endure anoxia in their drying mud burrows. The aerobic metabolism of an estivating African lungfish (*Protopterus*) is reduced to 20% of resting. The South American lungfish (*Lepidosiren*) and the swamp eel (*Synbranchus*), both competent air breathers, have very high levels of glycogen and glycolytic enzymes in their hearts to serve them when they become anaerobic during prolonged burrow dives. Jeffrey B. Graham

Bibliography. A. F. Bennett and J. A. Ruben, Endothermy and activity in vertebrates, *Science,* 206:649–654, 1979; S. F. Hand and G. N. Somero, Phosphofructokinase of the hibernator *Citellus beecheyi*: Temperature and pH regulation of activity via influences on the tetramer-dimer equilibrium, *Physiol. Zool.*, 56:380–388, 1983; M. Harless and H. Morlock (eds.), *Turtles*: *Perspectives and Research*, 1989; C. Heller and G. W. Colliver, *Amer. J. Physiol.*, 227:538–589, 1974; P. W. Hochachka, *Living Without Oxygen*, 1980; J. Hudson, Torpidity in mammals, *Comparative Physiological Thermoregulation*, vol. 3, pp. 97–165, 1973; L. L. McClanahan, R. Ruibal, and V. H. Shoemaker, Rate of cocoon formation and its physiological correlates in a ceratophryd frog, *Physiol. Zool.*, 56:430–435, 1983; A. Malan, Respiration and acid-base state in hibernation, *Hibernation and Torpor in Mammals and Birds*, pp. 237–282, 1982; F. N. White, A comparative physiological approach to hypothermia, *J. Thorac. Cardiovasc. Surg.*, 82:821–831, 1981; F. N. White and G. N. Somero, Acid-base regulation and phospholipid adaptations to temperature: Time courses and physiological significance of modifying the milieu for protein function, *Physiol. Rev.*, 62:40–90, 1982; M. E. Yacoe, Adjustments of metabolic pathways in the pectoralis muscles of the bat, *Eptesicus fuscus*, related to carbohydrate sparing during hibernation, *Physiol. Zool.*, 56:648–658, 1983.

Hickory

Any species of the genus *Carya*, formerly known botanically as *Hicoria*. Hickories are mostly tall forest trees characterized by strong, terminal, scaly winter buds, pinnately compound leaves (see **illus.**), solid pith (not chambered), and fruit with an outer husk or exocarp which splits more or less readily into four parts, revealing a nut with a hard shell or endocarp. *See* FAGALES.



Twigs, buds, and leaves of hickories. (*a*) Shagbark (*Carya ovata*). (*b*) Pignut (*C. glabra*).

The shagbark hickory (*C. ovata*) grows to a height of about 120 ft (36 m) and is found in the eastern half of the United States and adjacent Canada. It is the most important species because of the commercial value of its nuts, the hickory nuts of commerce, and of its wood. It is easily recognized by the bark, which in older trees exfoliates from the trunk in long, curving, irregular plates, and by the leaves, which usually have five leaflets, of which the terminal one is larger.

The pecan (*C. illinoensis*) is also a valuable species because of its commercially popular, thin-shelled, sweet nuts. Although its native range is limited to the Mississippi Valley region and Mexico, many varieties are cultivated in the southern United States.

Other species are the mockernut and pignut hickories, widely distributed through the eastern half of the United States, and the shellbark hickory found in the east-central United States. The remarkably tough and strong wood of all species makes it the world's best wood for tool handles. It is also used for parts of vehicles, furniture, flooring, boxes, and crates, and for smoking meats. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Hidden variables

Additional variables or parameters that would supplement quantum mechanics so as to make it like classical mechanics. Hidden variables would make it possible to unambiguously predict (as in classical mechanics) the result of a specific measurement on a single microscopic system. In contrast, quantum mechanics can give only probabilities for the various possible results of that measurement. Hidden variables would thus provide deeper insights into the quantum-mechanical probabilities. In this sense the relationship between quantum mechanics and hidden variables could be analogous to the relationship between thermodynamics (for example, temperature) and statistical mechanics (the motions of the individual molecules). *See* STATISTICAL MECHANICS.

As an example, in Young's double-slit experiment the pattern observed on a distant screen consists of alternating bright and dark regions. This pattern is the average of the results for a large number of photons passing through the double-slit arrangement, and represents the quantum-mechanical probability that a photon will be detected in any specific region of the screen. Hidden variables would, in principle, make it possible to precisely predict where each photon would be detected. *See* INTERFERENCE OF WAVES.

The concept of hidden variables appeared during the early development of quantum theory, apparently the result of an instinctive response engendered by the deterministic nature of classical mechanics. Specifically, it was felt that if the same measurement was made on each member of a group of identically prepared systems, the result of each measurement should be the same. If several different results were observed, then those identically prepared systems must not have been identical; that is, it was believed that the systems must have been in different microstates that were unknown at the time of preparation. The different microstates would be specified by the hidden variables. The variables were hidden only in the sense that they had not been observed. There were no restrictions as to whether or not they would, in principle, be possible to observe.

F. J. Belinfante formulated a three-section classification scheme for hidden variable theories—zeroth kind, first kind, and second kind. Most interest, both theoretically and experimentally, has been focused on hidden variable theories of the second kind, also known as local hidden variable theories.

Hidden variables of zeroth kind. In 1932 J. von Neumann provided an axiomatic basis for the mathematical methods of quantum mechanics. As a sidelight to this work, he rigorously proved from the axioms that any hidden variable theory was inconsistent with quantum mechanics. This was the most famous of a number of proofs, appearing as recently as 1980 and purporting to show the impossibility of any hidden variable theory. In 1966 J. S. Bell pinpointed the difficulty with von Neumann's proof—one of his axioms was fine for a pure quantum theory which makes statistical predictions, but the axiom was inherently incompatible with any hidden variable theory. The other impossibility proofs have also been found to be based on self-contradictory theories. Such theories are called hidden variable theories of the zeroth kind.

An impossibility proof originally based on a theorem by A. M. Gleason is, however, instructive since it does rigorously rule out one form of hidden variable theories whose precise predictions for a single system do not depend on a specific measurement procedure. This is also consistent with N. Bohr's dictum that a measurement is a joint result of both the system being measured and the measuring apparatus. It is easy to forget this point and be seduced by the false expectation that the result of a quantum-mechanical measurement is a preexisting property having nothing to do with the measuring apparatus.

Hidden variables of the first kind. Hidden variable theories of the first kind are constructed so as to be self-consistent and to reproduce all the statistical predictions of quantum mechanics when the hidden variables are in an "equilibrium" distribution.

D. Bohm introduced the first important theory of this type in 1952. It was based on ideas that had already been introduced in 1927 by L. de Broglie in his description of the "pilot wave." In Bohm's theory the hidden variable for a particle is identified with the position, x, of the particle, and the hidden variable equilibrium distribution is set equal to the particle's position probability given by quantum mechanics. Basically, this theory is a reinterpretation of nonrelativistic quantum mechanics in terms of hidden variables. However, it is inherently nonlocal (as below), and no satisfactory relativistic formulation has been given. As yet no prescriptions have been found for experiments that would distinguish the theory from conventional quantum mechanics. Thus, as a description of nature, it is an unnecessary complication of quantum mechanics, and the principle of Occam's razor may be applied.

Two first-kind theories that can be tested experimentally are those of N. Wiener and A. Siegal, and of Bohm and J. Bub. These postulate an equilibrium distribution of hidden variables that is gaussian. Deviations from the statistical predictions of quantum mechanics can be observed only when this equilibrium distribution is disturbed. Consequently, these theories can be tested by performing an experiment that disturbs this equilibrium distribution and then performing a second experiment before the hidden variables can relax. C. Papaliolios did such a test by observing the passage of photons through two closely spaced linear polarizers. The results were in agreement with the statistical predictions of quantum mechanics. Hence, the hidden variables described by these theories would have to relax to their equilibrium distributions in this experiment in a time less than about 2×10^{-14} s. From a heuristic point of view, Bohm and Bub estimate a relaxation time of about 10^{-13} s. Thus, this experiment decreases the credibility of these theories but does not entirely rule them out. *See* DISTRIBUTION (PROBABILITY).

Hidden variables of the second kind. These theories predict deviations from the statistical predictions of quantum mechanics, even for the "equilibrium" situations for which theories of the first kind agree with quantum mechanics. They are generally called local hidden variable theories because they are required to satisfy a locality condition. Intuitively, this seems to be a very natural condition. Locality requires that an apparatus at one location should operate independently of any settings or actions of a second apparatus at a spatially separated location. In the strict Einstein sense of locality, the two apparatus must be independent during any time interval less than the time required for a light signal to travel from one apparatus to the other.

Einstein-Podolsky-Rosen thought experiment. The focus for much of the discussion of local hidden variable theories is provided by a famous thought experiment (a hypothetical, idealized experiment in which the experimental results are deduced) that was introduced in 1935 by A. Einstein, B. Podolsky, and N. Rosen. Their thought experiment involves an examination of the correlation between measurements on two parts of a single system after the parts have become spatially separated. They used this thought experiment to argue that quantum mechanics was not a complete theory. Although they did not refer to hidden variables as such, these would presumably provide the desired completeness. The Einstein-Podolsky-Rosen (EPR) experiment led to a long-standing philosophical controversy; it has also provided the framework for a great deal of research on hidden variable theories.

A version of the Einstein-Podolsky-Rosen experiment is based on the single system consisting of the two photons emitted in an atomic cascade. A state for two such photons moving in opposite directions along the z axis is given by the equation below, where $|x\rangle_i$ and $|y\rangle_i$ are linear polarization states for

$$|\Psi\rangle = \frac{1}{\sqrt{2}}\{|x\rangle_1|x\rangle_2 - |y\rangle_1|y\rangle_2\}$$

photon i in the x and y directions, respectively; x and y are arbitrary orthogonal directions in a plane perpendicular to the z axis. Quantum mechanics assumes this state remains the same regardless of the distance between the two photons, and it predicts very strong correlations between the measurements of the polarizations for the two spatially separated photons. In fact, if the measurement of polarization for photon 1 gives a result along the x (or y) axis, measurement of the polarization for photon 2 can be predicted with absolute certainty to also be along the same x (or y) axis. This is an obvious result in view of the conservation of angular momentum for the entire, spatially separated, two-photon system; but in view of the spatial separation, it is an extraordinary result that already suggests a hidden variable explanation (that is, each photon would carry with it a prescription that predetermines the result of any polarization measurement). Furthermore, this same result holds for any different set of axes, say x', y'. Thus, by simply measuring polarization for photon 1, the result of measuring polarization for photon 2 can be predicted with absolute certainty in either the primed or unprimed axes (or, in fact, any such set of axes) without disturbing it. Einstein, Podolsky, and Rosen argued that, for any axis, the polarization of photon 2 must therefore somehow be real (predetermined); and since quantum mechanics does not encompass this, it must be an incomplete theory. N. Bohr has supplied the most famous response to this argument by questioning the Einstein-Podolsky-Rosen definitions of real and complete, and emphasizing the meaning of measurement as a function of both the microsystem and the entire measuring apparatus.

The nonlocality aspect of the Einstein-Podolsky-Rosen experiment is especially important. In the above example, if photon 1 passes a polarizer in any orientation, photon 2 absolutely must also pass another polarizer in the same orientation. Similarly if photon 1 does not pass, photon 2 absolutely must not pass. This kind of behavior is easily understood if the photons carry prescriptions (hidden variables) telling them what to do in every measurement case. But suppose such prescriptions do not exist. How then does photon 2, together with the second polarizer, "know" whether or not photon 2 should pass if locality is enforced and it cannot "know" the polarizer orientation for the measurement being made on photon 1?

Bell's result. The Einstein-Podolsky-Rosen experiment appeared in 1935; nevertheless, the von Neumann "impossibility" proof was dominant for many years. However, new efforts were stimulated in 1952 when Bohm did the "impossible" by designing the hidden variable theory of the first kind described above. Bohm's theory was explicitly nonlocal, and this fact led Bell to reexamine the Einstein-Podolsky-Rosen experiment. He came to the remarkable conclusion that any hidden variable theory that satisfies the condition of locality cannot possibly reproduce all the statistical predictions of quantum mechanics.

Specifically, in Einstein-Podolsky-Rosen type experiments, quantum mechanics predicts a very strong correlation between measurements on the spatially separated parts. Bell showed that there is an upper limit on the strength of these correlations in the statistical prediction of any local hidden variable theory. Bell's result can be put in the form of inequalities which must be satisfied by any local hidden variable theory but which may be violated by the statistical predictions of quantum mechanics under appropriate experimental conditions.

Local hidden variable experiments. Experiments performed under conditions in which the statistical predictions of quantum mechanics violate Bell's inequalities can test the entire class of local hidden variable theories. It should be emphasized that the quantum-mechanical predictions resulting from most conceivable systematic errors tend to yield a weaker correlation that does not violate the inequalities. However, all existing experiments have required supplementary assumptions regarding detector efficiencies. The experiments considered to be most definitive involve polarization correlation between two photons from either an atomic cascade or from parametric down conversion in a nonlinear crystal. Some experimental highlights in the history of those two-photon correlation experiments are (1) the 1972 experiment by S. Freedman and J. Clauser, which was the first experimental test; (2) the 1976 experiment by E. Fry and R. Thompson, which provided a dramatic improvement in the signal-to-noise ratio by using laser excitation, and also employed an anisotropic initial-state density matrix; (3) the 1982 experiment by A. Aspect, P. Grangier, and G. Roger, which introduced two-channel polarizers; (4) the subsequent 1982 experiment by Aspect, J. Dalibard, and Roger, which used time-varying polarizers and was the first effort to enforce Einstein locality; (5) the experiments in 1987 by Y. H. Shih and C. A. Alley and in 1988 by Z. Y. Ou and L. Mandel, which were the first to use parametric down conversion to generate the photon pair; (6) the 1995 experiment by the groups of A. Zeilinger and Y. Shih, which introduced a new type II down conversion source and used all four possible Bell states for experimental tests; (7) the 1998 experiment by the group of A. Zeilinger, which was the first to rigorously enforce Einstein locality; and (8) the 1998 experiment by the group of N. Gisin, which required the lifetime of the entangled state (for example, $|\Psi\rangle$ above) to be greater than 30 μs, as compared to a few nanoseconds in previous experiments.

Nevertheless, due to the supplementary assumptions, small loopholes still remain, and experiments have been proposed to eliminate them. One, an exact experimental realization of Bohm's version of the Einstein-Podolsky-Rosen thought experiment, has been undertaken by the group of E. Fry and Th. Walther. It involves massive, spin-$^1/_2$ nuclei (fermions) of mercury-199 rather than massless photons (bosons), and requires the entangled-state lifetime to exceed several milliseconds.

In conclusion, the overwhelming experimental evidence is against any theory that would supplement quantum mechanics with hidden variables and still retain the locality condition; that is, any hidden variable theory that reproduces all the statistical predictions of quantum mechanics must be nonlocal. The remarkable Einstein-Podolsky-Rosen correlations have defied any reasonable classical kind of explanation. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Edward S. Fry

Bibliography. F. J. Belinfante, *A Survey of Hidden-Variables Theories*, 1973; J. F. Clauser and A. Shimony, Bell's theorem: Experimental tests and implications, *Rep. Prog. Phys.*, 41:1881–1927, 1978; B. D'Espagnat, The quantum theory and reality, *Sci. Amer.*, 241:158–181, 1979; E. S. Fry and Th. Walther, Fundamental tests of quantum mechanics, *Adv. Atom. Mol. Phy.*, 42:1–27, 2000; E. S. Fry, Th. Walther, and S. Li, Proposal for a loophole free test of the Bell inequalities, *Phys. Rev.*, 52:4381–4395, 1995; N. D. Mermin, Is the Moon there when nobody looks? Reality and the quantum theory, *Phys. Today*, 38(4):38–47, April 1985; J. A. Wheeler and W. H. Zurek (eds.), *Quantum Theory and Measurement*, 1983.

Higgs boson

A hypothetical massive scalar elementary particle, the avatar (embodiment) of electroweak symmetry breaking in the Glashow–Weinberg–Salam theory. Interactions with the Higgs boson endow the quarks, leptons, and weak gauge bosons with mass. Although experiments provide indirect evidence for the influence of the Higgs boson, no direct observation has yet been made (as of 2006).

Since the mid-1990s, experiments in laboratories around the world have elevated the electroweak theory to a law of nature that holds over a remarkable range of distances, from the subnuclear (10^{-19} m) to the galactic (10^{20} m). The electroweak theory offers a new conception of two of nature's fundamental interactions, ascribing them to a common underlying symmetry principle. It joins electromagnetism with the weak interactions—which govern radioactivity and the energy output of the Sun—in a single quantum field theory. Dozens of measurements have tested and confirmed the agreement between theory and experiment at the level of one part in a thousand. *See* ELECTROWEAK INTERACTION; FUNDAMENTAL INTERACTIONS; STANDARD MODEL.

Weak interactions and electromagnetism. Weak interactions and electromagnetism are linked through symmetry, but their manifestations in the everyday world are very different. The influence of electromagnetism extends to unlimited distances, while the influence of weak interactions is confined to dimensions smaller than an atomic nucleus, less than about 10^{-17} m. The range of an interaction in quantum theory is inversely proportional to the mass of the force particle, or messenger. Accordingly, the photon, the force carrier of electromagnetism, is massless, whereas the *W* and *Z* particles that carry the weak forces are heavyweights, with masses nearly a hundred times that of the proton, the nucleus of hydrogen. *See* ELECTROMAGNETISM; INTERMEDIATE VECTOR BOSON; PHOTON; QUANTUM ELECTRODYNAMICS; WEAK NUCLEAR INTERACTIONS.

The many successes of the electroweak theory and its expansive range of applicability are most impressive, but the present understanding of the theory is incomplete. We have not yet learned what differentiates electromagnetism from the weak interactions—what endows the *W* and *Z* with great masses while leaving the photon massless. It is commonplace in physics to find that the symmetries observed in the laws of nature are not manifest in the consequences of those laws. A liquid is a disordered collection of atoms or molecules held together by electromagnetism that looks the same from every vantage point, reflecting the fact that the laws of electromagnetism are indifferent to direction. A crystal is an ordered collection of the same atoms or molecules, held together by the same electromagnetism, but it does not look the same from every vantage point. Instead, a crystal displays ranks, files, and columns that single out preferred directions. The rotational symmetry of electromagnetism is hidden in the regular structure of a crystal. *See* CRYSTAL STRUCTURE; SYMMETRY BREAKING; SYMMETRY LAWS (PHYSICS).

Higgs condensation. The central challenge in particle physics today is to understand what hides the symmetry between the weak and electromagnetic interactions. The simplest guess goes back to theoretical work by Peter Higgs and others in the 1960s. According to this picture, the diversity of the everyday world is the consequence of a vacuum state that prefers not a particular direction in space but a particular particle composition. The vacuum of quantum field theory—and of the world—is a chaotic confusion of many kinds of virtual particles winking in and out of existence. The laws of physics do not change if we interchange the identities of different particles. At high temperatures, as in the disordered liquid, the symmetry of the laws of physics is manifest in the egalitarian throng of particles. But at low temperatures, as in the ordered crystal, one kind of particle is preferred, and condenses out in great numbers, hiding the symmetry. *See* QUANTUM FIELD THEORY.

The condensate of Higgs particles can be pictured as a viscous medium that selectively resists the motion of other particles through it. In the electroweak theory, the drag on the *W* and *Z* particles caused by their interactions with the Higgs condensate gives masses to the weak-force particles. Interactions with the condensate could also give rise to the masses of the constituent particles—quarks and leptons—that compose ordinary matter. Today's version of the electroweak theory shows how this could come about, but does not predict what the mass of the electron, the top quark, or other particles should be. *See* LEPTON; QUARKS.

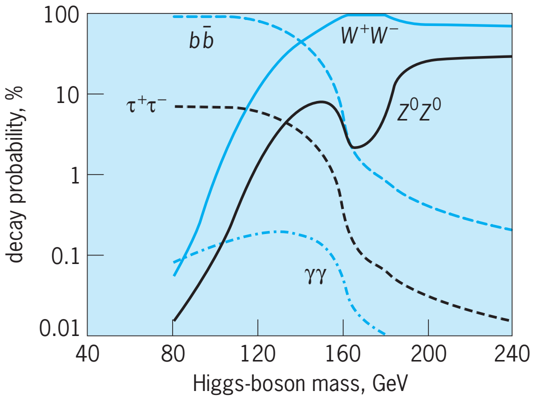
Experimental searches. If it were possible to heat up the vacuum enough, we could see the symmetry restored: all particles would become massless and interchangeable. For the present, it is beyond human means to heat even a small volume of space to the energy of 1 TeV (10^{12} electronvolts)—the temperature of 10^{16} kelvins. However, there is hope of succeeding on a smaller scale: it is possible to excite the Higgs condensate and see how it responds. The minimal quantum-world response is the Higgs boson: an electrically neutral particle with zero spin.

Although the Higgs boson has not been observed, experiments have begun to offer some evidence for its presence, beyond the shadowy role as giver of mass for which it was conceived. Experiments carried out over the past decade, using the Large Electron-Positron Collider (LEP) at CERN (European Laboratory for Particle Physics) and other instruments, are sensitive not just to the structure of the electroweak theory but also to quantum corrections. Long before the top quark was discovered at Fermilab in 1995, precision measurements detected its virtual quantum effects and anticipated its extraordinarily great mass. The influence of the Higgs particle is subtle, but increasingly incisive experiments strongly hint that a Higgs boson with mass less than about 200 GeV (1 GeV = 10^9 electronvolts) is required by precision measurements. *See* PARTICLE ACCELERATOR.

The electroweak theory predicts that the Higgs boson will have a mass, but it does not predict what that mass should be. (Consistency arguments require that it weigh less than 1 TeV.) Nevertheless, enough is known about the Higgs boson's properties—about how it could be produced and how it would transform into lighter particles or decay (as do all elementary particles except stable particles like the electron)—to guide the search.

LEP searches. The most telling searches have been carried out in experiments investigating electron-positron annihilations at energies approaching 210 GeV at LEP. The quarry of the LEP experimenters was a Higgs boson produced in association with the *Z*. A Higgs boson accessible at LEP would decay into a *b* (bottom) quark and a *b* antiquark, and have a total width (inverse lifetime) smaller than 10 MeV. The **illustration** shows how the decay pattern of a standard-model Higgs boson depends on its mass.

In the final year of LEP experimentation (before the machine was shut down in November 2000 to make way for construction of the Large Hadron [LHC]), accelerator scientists and experimenters made heroic efforts to stretch their discovery reach to the highest Higgs mass possible. At the highest energies explored, a few tantalizing four-jet events showed the earmarks of Higgs + *Z* production. A statistical analysis shows a slight preference for a



Dependence of the decay probabilities of a standard-model Higgs boson on its mass.

Higgs-boson mass near 116 GeV, but not enough to establish a discovery. Time will tell whether the LEP events were Higgs bosons or merely background events. The LEP observations do place an experimental lower bound on the Higgs-boson mass: the standard-model Higgs particle must weigh more than about 114 GeV.

Prospects. In 2006, experiments at Fermilab's Tevatron, where 1-TeV protons collide with 1-TeV antiprotons, were poised to extend the search, looking for Higgs + *W* or Higgs + *Z* production. The best hope for the discovery of the Higgs boson lies in experiments at the LHC at CERN, where 7-TeV beams of protons are designed to collide head-on. The LHC is scheduled to be commissioned in 2007, and it should make possible the study of collisions among quarks at energies approaching 1 TeV. A thorough exploration of the 1-TeV energy scale should determine the mechanism by which the electroweak symmetry is hidden and reveal what makes the *W* and *Z* particles massive.

Further questions. Once a particle that seems to be the Higgs boson is discovered, a host of new questions will come into play. Is there one Higgs boson or several? Is the Higgs boson the giver of mass not only to the weak *W* and *Z* bosons but also to the quarks and leptons? How does the Higgs boson interact with itself? What determines the mass of the Higgs boson? To explore the new land of the Higgs boson in more ways than the LHC can do alone, physicists are planning a TeV linear electron-positron collider.

Origins of mass. While the Higgs boson may explain the masses of the *W* and *Z*, and perhaps the masses of the quarks and leptons, it is not accurate to say that the Higgs boson is responsible for all mass. Visible matter is made up largely of protons and neutrons, and their masses reflect the energy stored up in the strong force that binds three quarks together in a small space. The origins of mass suffuse everything in the world around us, for mass determines the range of forces and sets the scale of all the structures we see in nature.

Extensions of electroweak theory. The present inability to predict the mass of the Higgs boson is one of the reasons many physicists believe that the standard electroweak theory needs to be extended. The search for the Higgs boson is also a search for extensions that make the electroweak theory more coherent and more predictive. Supersymmetry, which entails several Higgs bosons, associates new particles with all the known quarks, leptons, and force particles. Dynamical symmetry breaking interprets the Higgs boson as a composite particle whose properties we may hope to compute once its constituents and their interactions are understood. These ideas and more will be put to the test as experiments explore the 1-TeV scale. *See* ELEMENTARY PARTICLE; SUPERSYMMETRY.

Chris Quigg

Bibliography. F. Close, *Lucifer's Legacy: The Meaning of Asymmetry*, Oxford University Press, 2000; M. Kado and C. Tully, The searches for Higgs bosons at LEP, *Annu. Rev. Nucl. Part. Sci.*, 52:65–113, 2002; G.'t Hooft, Nobel Lecture: A confrontation with infinity, *Rev. Mod. Phys.*, 72:333–339, 2000; M. J. G. Veltman, Nobel Lecture: From weak interactions to gravitation, *Rev. Mod. Phys.*, 72:341–349, 2000; F. Wilczek, Masses and molasses, *New Scientist*, 162(2181):32–37, April 10, 1999.

High magnetic fields

Magnetic fields that are large enough to significantly alter the properties of objects that are placed in them. Valuable research is conducted at high magnetic fields. *See* MAGNETISM.

High-field magnets. Interest in the production of high magnetic fields, driven primarily by opportunities for new fundamental and applied science, extends back more than a hundred years. Until the late 1940s and early 1950s, magnets were confined to fields below 2 tesla due to the saturation of the iron cores used in laboratory magnets and the problems in cooling the conventional wire-wound coils. The attainable magnetic fields were extended more than 20 times by (1) the development of superconducting wire, fabricated from type II superconductors, which is capable of carrying high currents, and (2) the development of a new resistive magnet design by Francis Bitter using coils made of copper plates with the cooling water flowing through the plates themselves. Superconducting magnets have made fields up to about 15 T accessible to many laboratories, while higher-field magnets are restricted to a small number of facilities of 20 T. *See* SUPERCONDUCTING DEVICES.

Research and development efforts in magnets and magnet materials have led to gradual increases in the fields available for scientific research to fields near 20 T from superconducting magnets, 33 T in copper-core (resistive) magnets, and 45 T for hybrid magnets, which combine a superconducting outer core and a copper inner core. Superconducting magnets, which are used not only for research but also for magnetic resonance imaging (MRI), have the advantage that they use no electrical power once the field is established and the temperature is maintained at liquid-helium temperatures of 4.2 K (−452°F) or below.

The disadvantage is that there is a critical magnetic field, H_{c2}, determined by the type of conductor, that limits the attainable field to about 22 T in superconducting materials currently available. It is anticipated that new high-temperature superconductors may exceed this limit by 50% or more. Resistive magnets have no such limit, but are limited by the strength of the copper composite plates and the ability to cool them. The disadvantage is that they consume enormous amounts of power, in the 12–30-MW range. This limitation, plus the fact that they are very expensive to build and operate, has confined them to a few central facilities worldwide. *See* MEDICAL IMAGING; SUPERCONDUCTIVITY.

Another class of resistive magnets is pulsed magnets, in which the field is maintained for only brief periods of time. Advanced pulsed magnets that are not self-destructing provide fields beyond 70 T for about 0.1 s. Pulsed magnets using explosive magnetic flux compression have achieved fields above 500 T for periods of 10 microseconds. While the experiments that can be performed in the restricted times of the pulse are limited, much important research has been carried out with these magnets. *See* MAGNET.

Materials research. Research at very high magnetic fields spans a wide spectrum of experimental techniques for studies of materials. These techniques include nuclear magnetic resonance (NMR) in biological molecules utilizing the highest-field superconducting magnets, while the resistive magnet research is primarily in the investigation of semiconducting, magnetic, superconducting, and low-dimensional conducting materials. *See* NUCLEAR MAGNETIC RESONANCE (NMR).

Semiconductors. Much of the progress in semiconductor physics and technology has come from high-field studies. For example, standard techniques for mapping the allowed electronic states (the Fermi surface) of semiconductors and metals are to measure the resistance (in the Shubnikov–de Haas effect) or magnetic susceptibility (in the de Haas–van Alphen effect) as a function of magnetic field and to observe the oscillatory behavior arising from the Landau levels of the electron orbits. Measurements at low fields are limited to low impurity concentrations since the orbits are large and impurity scattering wipes out the oscillations. At high fields of 20–200 T, the orbits are smaller, and higher impurity concentrations (higher carrier concentrations) have been studied. *See* DE HAAS-VAN ALPHEN EFFECT; FERMI SURFACE; SEMICONDUCTOR.

In two-dimensional semiconductor sandwiches (quantum wells), quantum Hall effects are observed at very low temperature (a few millikelvin) and high field. In such samples the Hall voltage shows quantized steps as the field is increased. The integer quantum Hall effect was first observed by Klaus von Klitzing. Subsequently, fractional quantum Hall steps were observed by Daniel C. Tsui and Horst L. Störmer, and explained by Robert B. Laughlin. The fractional Hall effect studies in fields up to 30 T have led to the observation of a new quasiparticle, the composite fermion, consisting of an electron and two flux quanta bound together. Measurements at fields of 45 T and beyond are predicted to provide evidence for the existence of a quasiparticle consisting of an odd number of flux quanta bound to an electron, a composite boson, and to lead to connections between the quantum Hall effect and superconductivity. In addition to the new physics, the quantum Hall resistance is now the international resistance standard, and there are many potential applications in microelectronics. *See* ELECTRICAL UNITS AND STANDARDS; HALL EFFECT; RESISTANCE MEASUREMENT.

High-temperature superconductors. Another area in which very high magnetic fields have an important role is in high-temperature superconductors, which have great potential for high-field applications, from magnetic resonance imaging, to magnetically levitated trains, to basic science. In type II superconductors, the practical superconductors, nonsuperconducting quantized vortices penetrate the superconductor when a magnetic field is applied. The superconductor can be thought of as resembling Swiss cheese, with the vortices as the holes in the superconductor. The number of vortices increases linearly with increasing field until the vortices overlap and the superconductivity is destroyed throughout the superconductor at the critical magnetic fields (H_{c2}). Motion of vortices driven by magnetic field or current is an energy-loss process leading to a voltage across the superconductor. It is thus essential to understand the vortex dynamics and pinning mechanisms of superconductors for practical applications. Low-temperature superconductors, such as niobium-tin (Nb_3Sn), the most commonly used, have values of H_{c2} slightly above 20 T; in high-temperature superconducting materials, H_{c2} is several times greater. High-temperature superconductors also have the complexity of a vortex liquid phase in addition to the solid phase. The phase boundary is dependent on both field and temperature. Transport, specific heat, magnetization, and nuclear magnetic resonance studies up to 30 T showed that the phase boundary is independent of field above 9 T, an unexpected result. Further work at higher fields is necessary to clarify the behavior.

Magnetic materials. Magnetism is one of the most challenging subjects of condensed-matter physics, and studies at high magnetic fields have played an important role in advancing understanding of magnetic materials. For example, in many organic conductors the conduction electrons (or holes) are confined to one or two dimensions, leading to very rich magnetic phase diagrams. High-field phases above 20 T include spin-density waves, a modulation of the electron magnetic moments that can propagate through the crystal, modifying the conduction and magnetic properties. Magnetic-field-driven metal insulator transitions are another area of great interest. Studies of unusual metallic compounds of rare-earth and actinide elements with both localized electron magnetic moments and conduction electrons, at fields at or above the interaction energy of the electrons and moments, have greatly enhanced the understanding of this interaction. *See* ACTINIDE ELEMENTS; ORGANIC CONDUCTOR; PHASE TRANSITIONS; RARE-EARTH ELEMENTS.

Magnetic levitation. Another area of interest is the magnetic levitation of diamagnetic materials (the most common materials). For a diamagnetic material, an upward force is generated that is proportional to the product of the magnetic field and the magnetic field gradient. When the levitation condition is satisfied at high magnetic fields, zero gravity is simulated. Uses include growth of higher-purity crystals (aligned materials, including biological molecules, for detailed structural studies, much cheaper to produce in high magnetic fields than in satellite environments), studies of the effects of gravity and magnetic field on plant growth, and studies of the complex assembly processes of small particles under zero gravity and high magnetic field conditions. *See* SPACE BIOLOGY; SPACE PROCESSING; WEIGHTLESSNESS.

William G. Moulton

Bibliography. J. S. Brooks, J. E. Crow, and W. G. Moulton, Science opportunities at high magnetic fields, *J. Chem. Phys. Sol.*, 59:569–590, 1998; R.G. Clark (ed.), *Proceedings of the 5th International Symposium on Research at High Magnetic Fields*, North-Holland Publishers, 1997; Z. Fisk et al. (eds.), *Physical Phenomena at High Magnetic Fields II*, World Scientific Press, 1995; E. Manousakis et al. (eds.), *Physical Phenomena at High Magnetic Fields I*, Addison-Wesley, 1991; *National High Magnetic Field Laboratory*, H. Schneider-Muntau (ed.), *High Magnetic Fields: Applications, Generation, Materials*, World Scientific, 1997; M. Springford, L. J. Challis, and H. U. Karow (eds.), *The Scientific Case for a European Laboratory for 100 Tesla Science*, European Science Foundation, 1998.

High-pressure chemistry

Chemistry at very high pressures, arbitrarily chosen to be above 10,000 bars (10^9 pascals), and mainly concerned with solid and liquid states. A bar is 10^6 dynes/cm^2, or 1.0197 kg/cm^2, or 0.9869 atm, or 10^5 Pa. Multiples of the bar are the kilobar (1 kilobar = 10^3 bars = 10^8 Pa) and the megabar (1 megabar = 10^6 bars = 10^{11} Pa). At 25°C (77°F) and 10 kilobars (10^9 Pa), nearly all ordinary gases are liquid or solid, and only a few liquids are not frozen; thus most high-pressure chemistry involves either higher temperatures, at which chemical reactions can occur at appreciable rates, or studies of internal arrangements in solids.

Figure 1 illustrates the range of high pressures which exist in nature as well as those which can be attained in the laboratory. Three broad ranges of pressures are also shown. In the lowest range, from 1 bar (10^5 Pa) to about 10^5 bars (10^{10} Pa), normal low-pressure chemical behavior prevails, and only minor departures from the usual valence and coordination rules are found. However, as discussed later, many interesting changes in materials can be effected in this pressure range as atoms are forced into new bonding arrangements. In the second range, from 10^5 to 10^9 bars (10^{10} to 10^{14} Pa), the energy added by compression becomes comparable with chemical bond energies, so that outer-shell electronic orbits are distorted and atoms and molecules change in character. A general tendency toward more metallic behavior is observed as the electrons become less strongly fixed to particular atoms, and chemical bonds may be broken. In the third region, upward of about 10^9 bars (10^{14} Pa), the delocalization of electrons is extensive, and the material consists of a mixture of ions and electrons, so that chemical bonds are of little importance. The boundaries on these three pressure ranges are, of course, only approximate, and show some variation according to the temperature and the atoms involved.

Equipment. High-pressure chemical phenomena can be studied in a wide variety of types of equipment, depending on the pressure and temperature range and the object of the study. The highest laboratory pressures, several megabars, are achieved, albeit but for a few microseconds, by the accelerative forces generated by high explosives or high-velocity impacts in the so-called shock-wave techniques. *See* SHOCK WAVE.

Static pressures which may be exerted for minutes or hours have reached a maximum of about 400 kilobars (4 × 10^{10} Pa) between diamond faces on specially supported anvils of the type shown in **Fig. 2***a*, but the specimens are quite thin, with typical diameters of 1 mm, and the temperature range is limited. Larger specimens, 1 cm^3 (0.6 in.^{3}) or more in size, may be studied up to about 100 kilobars (10^{10} Pa) and 2000°C (3600°F) or higher in cylindrical, tetrahedral, or cubical apparatus of the types that are indicated in Fig. 2*a*, *c*, and *d*.

There is no theoretical upper limit to the static pressures that may be achieved, but practical limits are imposed by the magnitude of the forces required, the materials of construction available, and the stress gradients in them. The changes occurring in the compressed material may be monitored in place, for example, by optical, x-ray, or electrical techniques, for in many cases the material reverts to its original state upon release of pressure. Pressure apparatus is conveniently calibrated by observing definite changes in certain substances, for example, resistance changes in bismuth at about 25 kilobars (2.5 × 10^9 Pa) or in lead at 130 kilobars (1.3 × 10^{10} Pa). *See* HIGH-PRESSURE MINERAL SYNTHESIS.

High-pressure effects. The simplest effect of high pressure is the closer compression of atoms. The noble gases and alkali metals are quite compressible (potassium shrinks to half its original volume under a pressure of 100 kilobars or 10^{10} Pa), whereas most oxides and the stronger metals are considerably stiffer. However, at a pressure exceeding about 100 kilobars (10^{10} Pa), most of the easily compressed electronic clouds are tightened up, and the compressibilities of most substances approach each other. Usually, only minor amounts of energy compared with chemical bond energies can be added by compression to 100 kilobars. Nevertheless, the atoms of the substance can thereby be forced much closer together than they would be by cooling to low temperatures. An immediate consequence is an increase in melting temperature with pressure for most substances, since nearly all solids, with the exception of unusual ones like ice and bismuth, expand when they melt, and the process of melting absorbs heat as the entropy of the substance increases. The thermodynamic relationship shown in Eq. (1) expresses the

$$\frac{dT}{dP} = \frac{T\,\Delta V}{\Delta H} \qquad (1)$$

change in melting temperature T with pressure P as a function of the volume increase ΔV and the heat absorbed in melting ΔH for a given quantity of the substance. Thus, pressures of 50–100 kilobars (5–10×10^9 Pa) increase melting points about 100°C (212°F) for substances such as iron, whose ΔV of melting is small, to several hundred degrees for substances such as sodium chloride (NaCl), whose ΔV of melting is large. At 1 bar (10^5 Pa), NaCl melts in an iron crucible, but at 100 kilobars (10^{10} Pa) iron can be melted in NaCl crucible.

Substances that consist of large molecules are easily stiffened or frozen by high pressures. The mobility of the molecules is sharply decreased by a sort of interlocking and tangling effect; thus for the substance to be sheared, chemical bonds must be broken, a process which requires considerable energy. For example, ordinary oils become so stiff at a few dozen kilobars that they are useless for transmitting pressure, and a droplet of pressure-frozen oil is capable of denting a steel plate. This stiffening phenomenon limits the study of most reactions of organic molecules to low pressures because they are rather large and "freeze" easily, but are usually not stable enough to withstand the temperatures necessary for liquefaction or intermolecular reactions.

The thermodynamic relationship shown in Eq. (1) applies not only to melting but also to phase changes or internal structure changes in solids. In general, the higher-density forms are favored at high pressures, but since ΔH can be positive or negative, the effects of pressure and temperature may oppose or reinforce each other in determining the stability of a particular high-pressure phase. Phase changes between solids rarely run freely to follow the theoretical equilibrium line, but instead tend to be sluggish or exhibit a region of indifference. The stronger or more refractory the solid, or the greater the atomic displacements involved in the change, the broader the region of indifference. An outstanding example is diamond, which at room conditions persists at nearly 20 kilobars (2×10^9 Pa) out of its stability field. The existence of a region of indifference can hamper the study of some high-pressure phenomena, but on the other hand, it may permit the recovery of high-pressure phases for more detailed analysis at room conditions.

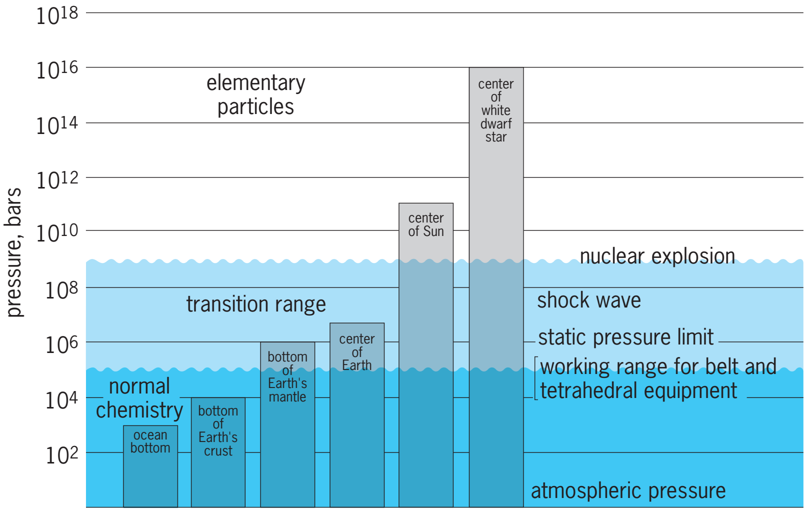


Fig. 1. Range of existing natural high pressures. 1 bar = 10^2 kPa.

The region of indifference may shrink drastically in the presence of solvents or catalysts that can promote the phase change. One of the more effective solvents, especially for systems containing oxides, is hot water. For chemically reducing systems such as carbon, the diamond-graphite transformations are assisted by the group VIII metals, iron, nickel, platinum, and so on. Other carbon solvents, such as silver chloride (AgCl) or cadmium oxide (CdO), exist but do not permit diamond to form, apparently because they do not favor carbon cations. For transformations in the boron nitride (BN) system, nitrides, either molten or in metallic solution, are effective catalysts. The rule "like dissolves like" is a useful, though not infallible, guide in these matters. *See* DIAMOND.

Some interesting chemical effects of pressure are related to shifts in chemical equilibria. The free-energy change ΔG determines chemical equilibria at temperature T through an expression of the

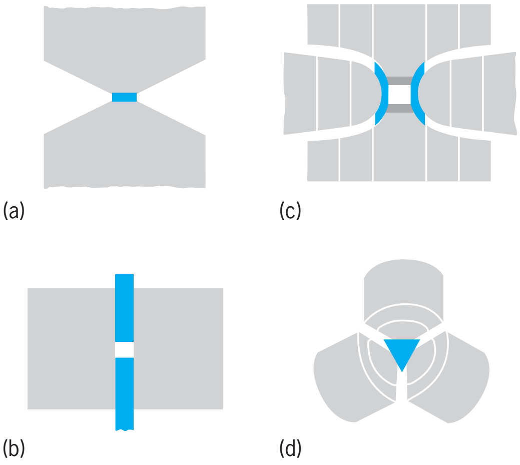


Fig. 2. Basic types of static high-pressure equipment. (*a*) Bridgman anvil. (*b*) Piston and cylinder. (*c*) Belt design. (*d*) Tetrahedral design.

form of Eq. (2), where R is the gas constant, and

$$\Delta G = RT\ \ln\left(\frac{a_m a_n \cdots}{a_x a_y \cdots}\right) \qquad (2)$$

a_m, a_n, . . . , are terms related to the concentrations of reactants and products in a reaction of the type $m + n + \cdots = x + y + \cdots$. For most reactions, ΔG ranges about 50 kcal (200 kilojoules) each side of zero. At 43 kilobars (4.3×10^9 Pa), a change in volume of 1 cm^3 (0.06 $in.^3$) corresponds to a change in energy of 10 kcal (40 kJ), and a change of 1 cm^3 (0.06 $in.^3$) per gram mole of reacting substance is not large, so that it is easy to find large shifts in chemical equilibria produced by pressure. Phase changes can also be regarded as special kinds of intramolecular reactions whose equilibria are shifted by pressure when interatomic bonds are broken and reformed.

Several rules exist regarding phase changes or electronic bonding changes in solids produced by pressure. The first rule recognizes that open structures stabilized by relatively weak ionic or van der Waals forces can easily collapse under pressure to denser structures. For example, eight crystalline forms of water ice have been identified. Potassium chloride (KCl) collapses from the NaCl structure to the more closely packed cesium chloride (CsCl) structure at about 25 kilobars (2.5×10^9 Pa). White phosphorus, made up of separate rings of four P atoms, is permanently pushed into the denser, more stable, semiconducting black form at pressures of a few tens of kilobars.

The second rule results from the compression of the outermost electronic shells and the delocalization of the electrons so that they are not as firmly fixed to particular atoms, and so that greater numbers of atoms can cluster around a given atom as the strongly directed valence forces are weakened. This is the same behavior observed as atomic number increases: Most of the lighter elements, with a few valence bonds per atom, are nonmetallic, whereas most of the heavier elements are metallic with higher valence numbers. Thus, pressure favors not only more metallic behavior but also behavior more like that of elements of higher atomic number. For example, upon compression to 100–130 kilobars ($1–1.3 \times 10^{10}$ Pa), silicon and germanium, normally semiconductors with the relatively open diamond structure, collapse to the white tin structure and become good metallic electrical conductors, equivalent to aluminum in that respect. As another example, silica, with four oxygen atoms about each silicon atom, changes at pressures of about 100 kilobars (10^{10} Pa) and moderate temperature in the presence of water to stishovite, which has the rutile structure of titanium oxide (TiO_2), with six oxygen atoms about each silicon. Many other examples of similar behavior in silicate systems are known wherein pressures of 100 kilobars or so force common crustal minerals into crystalline forms embodying higher coordination numbers, similar to those found in oxide compounds of heavier atoms. These effects are important in considerations of the nature of the deeper layers of the Earth's crust. *See* STISHOVITE.

The approach to the metallic state with increasing pressure may take different forms. Substances such as phosphorus, iodine, and selenium become substantially metallic at 100–150 kilobars ($1–1.5 \times 10^{10}$ Pa). On the other hand, when normally insulating organic compounds such as pentacene or hexacene, which consist of five or six aromatic rings fused together, are compressed to 150–200 kilobars ($1.5–2 \times 10^{10}$ Pa), they become semiconductors, since electrons are able to travel between the large molecules as the electronic clouds overlap. Certain complex compounds containing linear chains of metal atoms, such as Magnus's green salt, $Pt_2Cl_4(NH_4)_4$, increase in electrical conductivity up to 160 kilobars (1.6×10^{10} Pa) and then show a decrease; evidently, conductivity is favored only in particular ranges of interatomic spacing. Mössbauer studies have shown that ferric ion is reversibly reduced to ferrous ion in many compounds at pressures in the 100–200 kilobars ($1–2 \times 10^{10}$ Pa) range. Theoretical calculations indicate that even hydrogen could become metallic at pressures variously estimated to be 2–18 megabars ($2–18 \times 10^{11}$ Pa), but such pressures would be extremely difficult to generate and use, even with shock-wave techniques.

By changing the environment of atoms, high pressures can have strong effects on cooperative phenomena such as magnetism and superconductivity. For example, the high-pressure form of iron found above about 110 kilobars (1.1×10^{10} Pa) at 25°C (77°F) is not ferromagnetic. Superconducting transition temperatures are lowered by the application of high pressures to the metallic superconductors. Pressure studies guided the attainment of the ideal compositions of the high-temperature superconductors based on copper oxides reported in 1989, although pressure usually affects these superconductors adversely.

Many chemical reactions proceed through an intermediate state whose volume may be larger or smaller than that of the reactants. When the volume is smaller, the rate of reaction may be increased by pressure. Usually the intermediate state is more voluminous, especially in solids, and the rate of reaction is reduced by pressure. Here, as with chemical equilibria, the energy change due to pressure in the 30–100 kilobars ($3–10 \times 10^9$ Pa) range can be comparable with ordinary chemical bond or reaction-activation energies, and large pressure effects are possible. Certain reaction pathways may be effectively blocked so that new paths are followed. For example, at 130 kilobars (1.3×10^{10} Pa), where diamond is stable at temperatures up to 3000°C (5400°F) or more, the pyrolysis of some organic compounds, especially those consisting mainly of aromatic rings, produces graphite as the initial product whereas paraffin or polyethylene loses hydrogen to form waxy, dense solids of increasing microcrystalline diamond content as the pyrolysis temperature increases. A hexagonal form of diamond can be prepared by subjecting highly crystalline graphite to pressure of 130 kilobars (1.3×10^{10} Pa) and 1500°C (2700°F). The reaction proceeds to some extent at 25°C (77°F) but proceeds much further on heating. (Poorly crystalline or partly amorphous graphite yields ordinary cubic diamond.) The hexagonal diamond can be recovered at room pressure and temperature if it has been heated at high pressure. Hexagonal graphitic BN can be converted to a wurtzite form by the application of 130 kilobars (1.3×10^{10} Pa) at 30°C (86°F), but at high temperatures, or in the presence of a molten catalyst-solvent, the cubic form of BN, which is slightly more stable than the wurtzite form, is obtained. Robert H. Wentorf, Jr.

Bibliography. W. Holzapfel and N. Isaacs (eds.), *High Pressure Techniques in Chemistry and Physics: A Practical Approach*, 1997; N. Isaacs, *Liquid Phase High Pressure Chemistry*, 1981; W. F. Sherman and A. A. Stadtmuller, *Experimental Techniques in High Pressure Research*, 1988; R. van Eldik and F.-G. Klärner (eds.), *High Pressure Chemistry: Synthetic, Mechanistic, and Supercritical Applications*, 2002.

High-pressure mineral synthesis

A laboratory technique for studying the behavior of minerals under high-pressure conditions.

The structure and nature of minerals vary with temperature and pressure. An example is ice, an important "mineral" in the polar regions of the Earth and in the interior of some planets: the solid form of water (H_2O), ice changes into a liquid and then into a gas with increasing temperature. Pressure dramatically alters the structure and nature of this material as well. When liquid H_2O is compressed at 300 K (80°F), it freezes at approximately 1 gigapascal (10 kilobars). This high-pressure form of ice (known as ice-VI) is denser than the coexisting liquid; thus ice-VI sinks into the bottom of water, while the density of ice at atmospheric pressure is lower than that of liquid water (**Fig. 1**). In ordinary ice, the H_2O molecule possesses hexagonal symmetry, which is manifested in the striking and beautiful shapes of snow crystals. In ice-VI, however, the H_2O molecules are arranged in a different manner, and the crystal shape is also quite different. With further compression of ice-VI to a pressure of 2 GPa (20 kilobars), another high-pressure phase forms, known as ice-VII, which has cubic symmetry. All of these forms of ice are colorless, but it is believed that the ice becomes opaque and "metallic" at ultimate high pressure.

The nature of minerals as they exist at atmospheric pressure represents only a very limited part of their real nature. The range of pressure and temperature prevailing at the surface of the Earth is very limited compared to the ranges that exist in the other planets of the solar system. The bottom of the ocean, which is at the highest pressure that can be observed directly, is only 0.1 GPa (1 kilobar), while the pressure at the center of the Earth is 390 GPa (3900 kilobars). Pressures at the centers of large planets such as Saturn and Jupiter exceed 1000 GPa (10,000 kilobars). Therefore, to study the formation and structure of the Earth and other planets, it is essential to study the behavior of minerals under high pressure. It has become clear through high-pressure experiments that the minerals constituting the Earth's lower mantle (which extends from 650 to 2900 km or 400 to 1800 mi from the surface and occupies more than 50% of the entire volume of the Earth) are mostly so-called silicate perovskites that can never be formed on the surface of the Earth. *See* EARTH INTERIOR; JUPITER; SATURN.

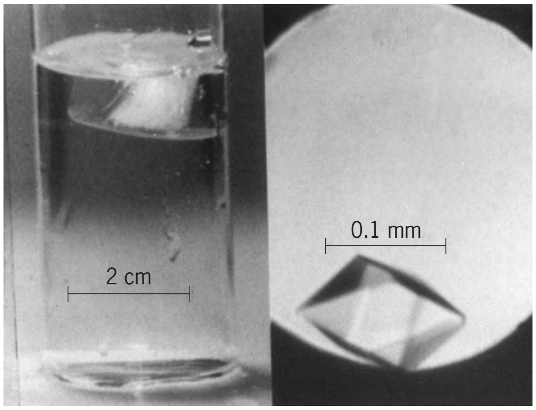


Fig. 1. **Differing ice density. Normal ice floats on water while high-pressure ice formed at 1 GPa (10 kilobars) sinks to the bottom. (*From N. Miura, T. Yagi, and H. Ishimoto, The world of extreme, Kotai Butsuri* (*Japanese*), *23:424–430, 1988*)**

Experiments and apparatus. Pressure is defined as a force per unit area; therefore, in order to apply a high pressure, it is necessary to concentrate a large force in a small area. Because of the limited strength of materials used to produce sample chambers, many different techniques are required, depending on the pressure range (**Fig. 2**).

Static compression. There is a specific type of equipment used to produce static high pressures. Mineral synthesis under conditions similar to those at the surface of the Earth is carried out by using a hydrothermal bomb apparatus. The starting materials are sealed in a high-pressure vessel (made of a hardened steel) together with a pressure-transmitting medium such as water or inert gas, and the whole vessel is heated in a large furnace. In this apparatus, the sample chamber is large (on the order of cubic centimeters), and the temperature throughout the sample is uniform, but the pressure is limited to about 1 GPa (10 kilobars).

The piston-cylinder arrangement is the high-pressure apparatus most commonly used to study the minerals in the Earth's crust. In this apparatus, the piston is subjected to the same pressure as that of the sample; it is difficult to achieve pressures higher than 3 GPa (30 kilobars), even when the piston and cylinder are fashioned from tungsten carbide, an exceptionally hard alloy. The multianvil apparatus was developed to overcome this difficulty.

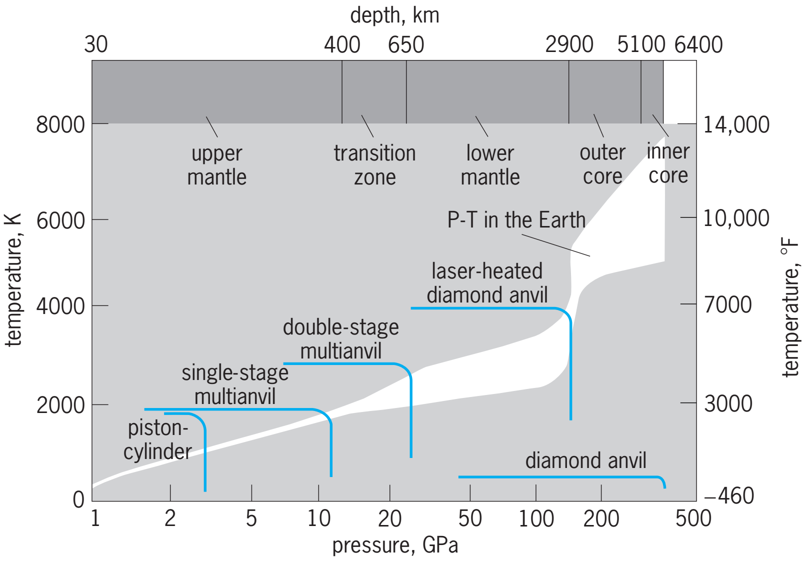


Fig. 2. Diagram of pressure (P) and temperature (T) within the Earth, showing capabilities of various types of high-pressure apparatus. 1 GPa = 10 kilobars. 1 km = 0.6 mi.

Very high pressure achieved on the top of the anvil is supported by a much larger area at the bottom. The upper portion of the anvil is also supported laterally, which increases the strength of the anvil. Consequently, pressures achieved by the apparatus are much higher than is possible with the simple piston-cylinder apparatus. However, when tungsten carbide is used for anvil material, 25 GPa (250 kilobars) is the practical upper limit of the pressure. Higher pressures have been achieved by using sintered diamond as the anvil material. In this multianvil apparatus, a sample and a small furnace are embedded in a solid pressure-transmitting medium, and the entire assembly is compressed. Typically, the sample chamber has a volume of several cubic millimeters, and it can hold a sample weighing several milligrams.

Although the amount of sample used is usually very small (typically a few micrograms), the diamond-anvil apparatus is often chosen for studying the behavior of minerals under a very wide range of pressures and temperatures. The principle of operation of the diamond anvil is very simple (**Fig. 3**). The sample and the pressure-transmitting medium are placed in a small hole (typically 0.3 mm or 0.012 in. in diameter and 0.1 mm or 0.004 in. deep) of a metal gasket and then squeezed between two flat faces of gem-quality diamonds. The metal gasket is deformed and high pressure is applied to the sample. Diamond is the hardest material known, and the highest pressure that has been achieved by this apparatus is more than 400 GPa (4000 kilobars), which is more than that at the center of the Earth. Although the size of the sample is very small, the development of many sophisticated experimental techniques, such as electron-probe microanalysis, very intense x-rays from synchrotron radiation, and Raman microanalysis, has made it possible to study the structure and the nature of minerals synthesized in such small quantities. The pressure in the sample chamber can be measured by using the wavelength shift with pressure of the ruby fluorescence line. Small chips of ruby are placed in a sample chamber and are irradiated by blue light from a laser; they emit red fluorescent light, and the pressure can be calculated precisely by measuring the wavelength of this light. *See* DIAMOND; ELECTRON SPECTROSCOPY; RAMAN EFFECT; SYNCHROTRON RADIATION.

The sample is heated by using strong laser light. Since diamond is transparent to visible and near-infrared light, the light from the laser can reach the sample through the diamond without much loss of energy. The energy of the light is absorbed and changed into heat either by the sample itself or by a dark material, such as graphite or platinum powder, that is mixed with the sample. Temperatures as high as 4000 K (6740°F) can be achieved, and formation of molten diamond has been reported. With this technique, it has become possible to produce in the laboratory conditions very similar to those believed to exist at the center of the Earth.

A disadvantage of the diamond-anvil apparatus is that while the heated area is very small (only 20 micrometers or less), a large temperature gradient occurs within the sample. In order to heat the whole area, it is necessary to scan the hot spot; thus, the thermal history of the sample becomes complicated. By contrast, the heating in multianvil apparatus is much more uniform, and a much larger sample can be treated. Therefore, the multianvil apparatus is a very powerful tool for studies of multicomponent systems or for synthesis of single crystals of high-pressure minerals.

Shock compression. In nature, shock compression caused by the impact of meteorites plays an

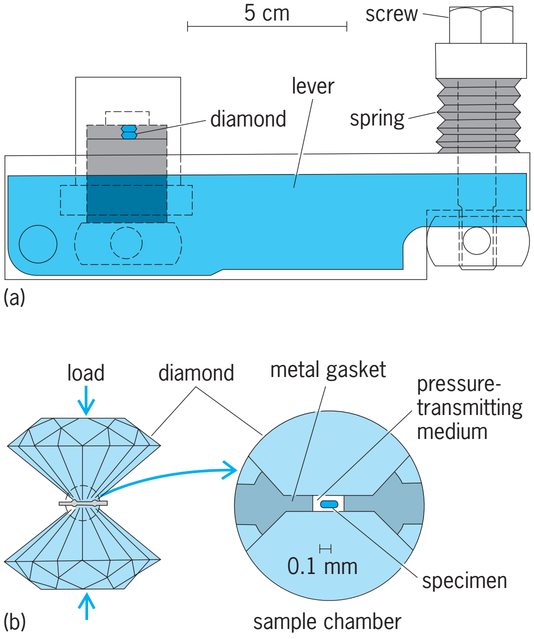


Fig. 3. Lever and spring type of diamond-anvil apparatus. (*a*) Cross section. (*b*) Expanded view, showing placement of the specimen. (*After T. Yagi, Diamond anvil and high-pressure experiments, Ouyou Butsuri* (Japanese), *50:1337–1341, 1981*)

important role in the formation of high-pressure minerals. Shock compression experiments are carried out in the laboratory by using either explosions or collisions of projectiles into sample material. An example of this type of equipment is a single-stage gas gun that is used to study equations of state (pressure-volume-temperature relationships) or to synthesize high-pressure minerals. The projectile is accelerated to 4–5 km/s (2.4–3 mi/s) and collides with the sample chamber; a very high pressure is generated for a few microseconds. Usually, the sample is broken into small pieces after the impact; thus, proper design of the sample chamber is required to facilitate recovery of the sample. *See* METEORITE.

In-place observations. It has been found that in many minerals the phases formed under high pressure and temperature can be recovered at a pressure of 1 atm (10^5 pascals) if the temperature is reduced rapidly while the sample is still maintained under pressure; this is known as quenching the high-pressure phase. However, some high-pressure phases are known to undergo retrogressive transformation when the pressure is released. Many calcium-bearing silicates and iron oxides are in this category. To study these high-pressure minerals, it is necessary to carry out the observations while the sample is still under high pressure. Such in-place observations at high pressures become more and more important when the pressure range is expanded, because more minerals become unquenchable under these extreme conditions. X-ray and optical techniques provide the means for such in-place observations. By using very strong x-rays from synchrotron radiation or strong light from a laser, it has become possible to perform high-quality observations even in a pressure range of 100 GPa (1000 kilobars). *See* LASER.

Effects on structure. Minerals are composed of anions [such as oxygen (O) and fluorine (F)] and cations [such as silicon (Si), magnesium (Mg), and iron (Fe)]. The basic building unit of a mineral is known as a coordination polyhedron, in which a cation is surrounded by several equidistant anions. The number of surrounding anions is known as the coordination number. In silicate, the fundamental SiO_4 tetrahedron is usually the basic unit of the structure, and silicates are classified into several different groups depending on how these SiO_4 tetrahedra are connected with each other in the structure. The effects of pressure on mineral structure also depend on how the polyhedra are connected. *See* CRYSTAL STRUCTURE; SILICATE MINERALS.

When high pressure is applied to a silicate in which the SiO_4 tetrahedra are isolated, vacancies among the polyhedra shrink first, because the polyhedra are much "harder" than the vacancies. When the polyhedra are connected to each other, by sharing corners or edges, the crystal will tend to reduce the total volume by rotating the polyhedra. When rotation is no longer possible, the polyhedra become compressed. When the compression of the crystal exceeds a certain value, rearrangement of the atoms occurs, and the crystal structure is transformed; this is accompanied by a sharp decrease in volume. Sometimes, the coordination number increases in these pressure-induced phase transformations. The reason for this is that a large anion is usually more compressible than a small cation; and by increasing the pressure, the ratio of the radius of cation to that of the anion changes. For example, when silicon dioxide (SiO_2) and tin dioxide (SnO_2) transform from coesite into stishovite (the rutile-type structure) and from the rutile-type structure into the fluorite-type structure, the coordination numbers of Si and Sn increase from 4 to 6 and from 6 to 8, respectively. *See* COESITE; FLUORITE; RUTILE; STISHOVITE.

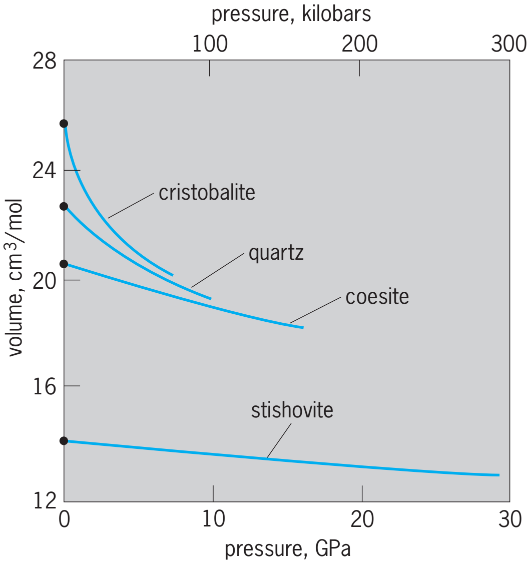


Fig. 4. Compression curves of various polymorphs of SiO_2. 1 cm^3 = 0.06 in.^{3}

Volume compression curves vary for various polymorphs of SiO_2 (**Fig. 4**). Cristobalite is a high-temperature polymorph of quartz. Silicate tetrahedra are connected by sharing oxygen at the corners, and the structure possesses many vacancies. The volume can be reduced by rotating tetrahedra; and consequently the compressibility of this structure is almost three times larger than that of quartz. When quartz transforms into coesite at about 3 GPa (30 kilobars), the density increases approximately 6%, but the coordination number of silicon remains unchanged. At about 10 GPa (100 kilobars), coesite transforms into stishovite with a density increase of 28%, and the coordination number of silicon is increased to six. This is a very close-packed structure, and the compressibility is very small. *See* QUARTZ.

A large number of phase transformations has been found in minerals under high pressure, but most of these structures have already been observed in other minerals existing under atmospheric pressure. For example, rutile-type SiO_2 (stishovite) is formed only above 10 GPa (100 kilobars), but the same structure is obtained at atmospheric pressure when the Si ion is replaced by the larger germanium (Ge) ion. This implies that crystal structure is determined mainly by the ratio of the cation radius to that of the anion.

When the very dense structure is compressed further, the bond length becomes shorter and shorter, and the orbitals of the electrons around the ions begin to overlap. This means that the orbital electrons can move freely in the material, which changes into a so-called metallic state. This metallic transition is believed to occur in all materials when they are subjected to high enough pressure. Even hydrogen, helium, and ice are believed to exist in the metallic state in the interiors of Jupiter and Saturn. In the laboratory, however, this transformation into the metallic state under pressure has been confirmed in only a limited number of materials such as Si, Ge, and gallium arsenide (GaAs). *See* BOND ANGLE AND DISTANCE; FREE-ELECTRON THEORY OF METALS.

The effects of shock pressure are very similar to those of static pressure. In the laboratory, however, because of the small size of the projectile, the duration of shock pressure is limited to the order of microseconds. When the synthesis of a high-pressure mineral requires rearrangement of atoms, the high-pressure state sometimes does not last long enough to accomplish phase transformation. In a naturally occurring impact of a meteorite, the size of the colliding material is much larger, and consequently the shock compression state lasts for a much longer time. This difference in time duration may explain the fact that many high-pressure minerals such as coesite and stishovite are found in natural craters while it is difficult to synthesize these minerals in the laboratory. Another important factor is that in shock compression it is almost impossible to control the temperature. When materials are compressed by shock waves, the temperature rises quickly and higher than that of the adiabatic compression curve. The cooling process is usually much slower, causing the retrogressive transition. Although shock compression is a very interesting and powerful means of synthesizing high-pressure minerals, the process is much more complicated than static compression. *See* METEORITE; SHOCK WAVE.

Transformations in mantle minerals. Several thousand minerals have been found to occur on the surface of the Earth, but the number of important minerals occurring in its deep interior is believed to be relatively small. The three major minerals that are believed to exist in the upper part of the mantle are olivine, pyroxene, and garnet. Many phase transformations are found to occur in these minerals up to 30 GPa (300 kilobars).

Olivine. Olivine is probably the most abundant mineral in the mantle, and the behavior of this material has been studied in detail. Based on the observations of the isomorphous structures of magnesium germanium oxide (Mg_2GeO_4), J. D. Bernal first pointed out the possibility that Mg_2SiO_4 olivine may transform into spinel structure in the Earth's mantle, and that the discontinuous density increase associated with this transition might be the cause of the seismic discontinuity in the mantle. The prediction of the olivine-spinel transformation in silicate was proved to be correct more than 30 years later by an experimental study on nickel silicate (Ni_2SiO_4). The coordination numbers of both silicon and nickel remain unchanged during this olivine-spinel transition; but because of the better packing of polyhedra the spinel structure is approximately 10% denser than olivine.

Natural olivine is a solid solution, and its composition is close to $(Mg_{0.9}Fe_{0.1})_2SiO_4$. Phase diagrams in the system Mg_2SiO_4-Fe_2SiO_4 have been studied in detail. With increasing magnesium content, the pressure from the olivine-to-spinel transition increases rapidly. When the magnesium content increases to more than 80%, a new phase appears that gives an x-ray diffraction pattern similar to that of spinel phase. The structure of this phase was elucidated by using an isostructural single-crystal cobalt silicate of (Co_2SiO_4) synthesized under high pressure; it was characterized as possessing a modified spinel structure. It was a unique structure never found in other minerals formed at atmospheric pressure, and it does not satisfy the Pauling's rule for the stability of minerals. Because of these observations, it was at first believed that this phase might not be thermodynamically stable but instead might be formed by the retrogressive transition from an unquenchable phase. This possibility was contradicted by the high-pressure in-place observation. Because of the existence of this phase, a sequence of the phase transformations expected to occur in the Earth became complicated. The properties of density and elasticity of this modified spinel structure are similar to those of spinel structure, and the major discontinuity at 400 km (250 mi) in the upper mantle can be explained by the transition from olivine to modified spinel. *See* MOHO (MOHOROVIČIĆ DISCONTINUITY); SOLID SOLUTION.

Many interpretations were advanced concerning the nature of postspinel transition; they included the transition to a potassium nickel fluoride (K_2NiF_4) structure and a strontium lead oxide (Sr_2PbO_4) structure, and the decomposition into MgO plus SiO_2. Experimental study confirmed that Mg_2SiO_4 spinel decomposes into a mixture of perovskite-type $MgSiO_3$ plus MgO at about 25 GPa (250 kilobars). In so-called ABX_3-type compounds, the most common structure at atmospheric pressure is that of perovskite. However, this structure is formed only when the larger cation (A) is similar in size to the X anion and the smaller cation (B) is large enough so that it can be sixfold-coordinated by X anions (**Fig. 5**) Therefore, under ambient conditions, this structure can be formed when cation A is very large, such as calcium (Ca), strontium, and barium (Ba), while smaller cation B is relatively large such as titanium (Ti), vanadium (V), and Fe. At atmospheric pressure, Mg and Si are far too small to satisfy these conditions. These silicate perovskites are stabilized only under very high pressure, where the oxygen anion is compressed much more than the cations and the ratio of the ionic radii has changed. An interesting observation is that when the large cation A is a transition-metal ion such as Fe or Co, this silicate perovskite is not formed but is decomposed into a mixture of a rocksalt-type AO compound plus a rutile-type BO_2 compound.

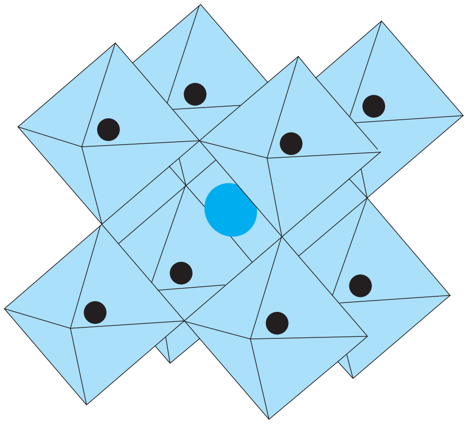


Fig. 5. Perovskite structure; the oxygen atoms (black circles) are on the corner of each octahedron, and a large cation (color circle) is surrounded by eight octahedra.

This is probably because the site for the cation A in perovskite structure has an unfavorable energy for such transition-metal ions. *See* OLIVINE; PEROVSKITE; SPINEL; TRANSITION ELEMENTS.

Pyroxene. Pyroxene is the second most abundant mineral in the Earth. The phase boundary between ortho- and clinopyroxene in $MgSiO_3$ is somewhat ambiguous; this is probably because the transition is extremely sensitive to the uniaxial component of the pressure. Above 18 GPa (180 kilobars) pyroxene decomposes into a modified spinel-type structure, Mg_2SiO_4 plus SiO_2. This modified spinel transforms into spinel structure at about 20 GPa (200 kilobars). Further compression forms a single phase of ilmenite structure, and this phase transforms into perovskite structure. At very high temperatures, a garnet phase is found to exist. *See* ILMENITE; PYROXENE.

Garnet. The first silicate perovskite was found in pyrope when this material was subjected to a pressure above 30 GPa (300 kilobars). The structure of this perovskite is believed to be complicated because of the existence of aluminum (Al) ion, and the structure has not been analyzed in detail. *See* GARNET.

Major minerals within the Earth. It has become clear that many of the major phases of silicate transform into the perovskite structure above 25 GPa (250 kilobars). Therefore it is believed that silicate perovskite is the most abundant mineral within the Earth, although it is an exotic mineral on the surface. In order to clarify the nature of the high-pressure minerals believed to be present in the interior of the Earth (**Fig. 6**) many studies have been made using various techniques. For this type of study, it is important to obtain a single crystal. The multianvil apparatus has been widely used for such experiments, and various single crystals of high-pressure minerals such as silicate perovskite, spinel, and stishovite have been synthesized.

Phase transformations in gases. Gases such as hydrogen, helium, and methane are important constituents in the outer planets. Since these gases are very compressible, it was difficult to study their behavior at high pressure in the laboratory. However, by using diamond-anvil apparatus it has become possible to liquefy these gases either by cooling or by applying pressures of a few thousand atmospheres and to load them into high-pressure sample chambers. All of these gases are found to solidify below 10 GPa (100 kilobars) at room temperature; and, as is found in other solids, many polymorphs are formed under very high pressure. In some materials, experiments are carried out up to 100 GPa (1000 kilobars), and the results of these studies provide important information for understanding the nature of the interiors of the planets.

Significance. A relatively narrow range for atmospheric pressure is an absolute requirement for the existence of most life forms on Earth, but this range is just a very small portion of the pressure range found in the universe. Therefore, high-pressure mineral synthesis is an indispensable tool for understanding the real nature of minerals and to study the interiors of planets. Development of experimental techniques has made it possible to perform such studies on various materials under practically the entire pressure range existing in the Earth. Many efforts are in progress to extend the range of pressures and temperatures for such experiments.

High-pressure synthesis is a powerful method not only for use in earth and planetary sciences but also for the creation of new materials. Many industrial diamonds are synthesized by using high-pressure techniques, and some new high-pressure materials, such as cubic boron nitride, have found wide application. Pressure is one of the most fundamental parameters that can alter the state of materials, and research in this field is also expected to expand in the future. *See* HIGH-PRESSURE CHEMISTRY; HIGH-PRESSURE PHYSICS; SILICATE PHASE EQUILIBRIA;

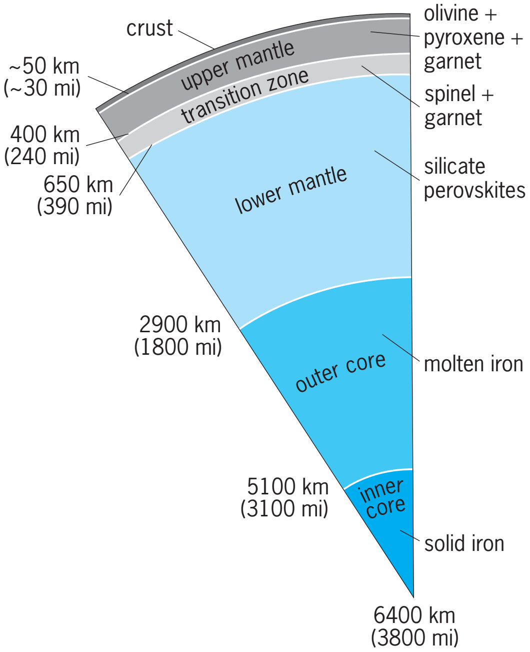


Fig. 6. Schematic representation of a section through the Earth, showing major minerals associated with each layer.

SOLID-STATE CHEMISTRY; SOLID-STATE PHYSICS; THERMODYNAMIC PROCESSES. Takehiko Yagi

Bibliography. H. Aoki, Y. Syono, and R. J. Hemley (eds.), *Physics Meets Mineralogy: Condensed Matter Physics in the Geosciences*, 2000; P. W. Bridgman, *The Physics of High Pressure*, 1949; L. G. Liu and W. A. Bassett, *Elements, Oxides, Silicates, High-Pressure Phases with Implications for the Earth's Interior*, 1986; M. H. Manghnani and T. Yagi, *Property of Earth and Planetary Materials at High Pressure and Temperature*, 1998; D. C. Rubie, T. S. Duffy, and E. Ohtani (eds.), *New Developments in High-Pressure Minerals Physics and Applications to the Earth's Interior*, 2005.

High-pressure physics

High-pressure physics is concerned with the effects of high pressure on the properties of matter. Since most properties of matter are modified by pressure, the field of high-pressure physics encompasses virtually all branches of physics. The justification for classifying high-pressure physics as a separate field is that rather specialized and often ingenious techniques are needed, both to produce high pressures and to make measurements of changes of physical properties of a material at high pressure. This field is therefore analogous to the fields of low- and high-temperature physics. Indeed, high-pressure experiments have been performed at temperatures approaching absolute zero and at temperatures as high as 9000°F (5000°C). *See* LOW-TEMPERATURE PHYSICS.

Pressure is defined as force per unit area; commonly used units of pressure include pounds per square inch (psi), kilograms per square centimeter (kg/cm^2), atmospheres (atm), bars, and pascals (Pa). The bar, equal to 10^6 dynes/cm^2 or about 14.5 lb/in.^{2} (that is, slightly less than 1 atm), is the basic unit that has been used by most researchers, although the approved unit of pressure in the International System (SI) of Units is now the pascal, which is equal to 10^{-5} bar. *See* PRESSURE.

The "high" of high-pressure physics connotes experimental difficulty. At liquid-helium temperatures, pressures of several hundred bars are considered high. In general, however, the high-pressure range may be arbitrarily regarded as extending from about 1 kbar (100 MPa or 14,500 lb/in.^{2}) upward to the present experimental limit. Prolonged static pressures in excess of 1 megabar (100 gigapascals or 1.45×10^7 lb/in.^{2}) can be achieved in very small samples weighing about 1 microgram. For perspective, a 100-micrometer-diameter (0.004-in.) grain of sand weighs about 1 μg. If simultaneous prolonged high temperatures in the range from 1800 to 3600°F (1000 to 2000°C) are required, the limit of static high pressures attainable is reduced to about 200 kilobars (20 GPa or 2.9×10^6 lb/in.^{2}). Additional experimental requirements, such as the need for nonmagnetic pressure vessels in magnetic resonance studies, further limit maximum attainable pressures.

Transient pressures as high as about 10^7 bars (1000 GPa or 1.45×10^8 lb/in.^{2}) have been attained in shock waves produced by high explosives or by projectile impact. This article is primarily concerned with static high-pressure experimentation, but it should be noted that the results of shock-wave experiments generally complement those of prolonged static high-pressure experiments. *See* SHOCK WAVE.

High-pressure effects. The major effects of high pressure on matter include diminution of volume, phase transitions, changes in electrical, optical, magnetic, and chemical properties, increases in viscosity of liquids, and increases in the strength of most solids. The magnitudes of these pressure effects may be illustrated by some examples, both general and specific. If a gas, initially at low pressure and a temperature below critical, is compressed, the first effect of pressure is to reduce the free space separating the atoms or molecules. At pressures on the order of several hundred bars, the volume will be about one-thousandth of the initial volume and a change of phase from gas to liquid will occur. Further compression of the liquid, by pressures up to about 50 kbar (5 GPa or 7.3×10^5 lb/in.^{2}), results in an additional volume decrement of only 20–50%. For liquids the greatest effect of pressure may be the increase in viscosity, which can be as high as a millionfold for a pressure increase of 10 kbar (1 GPa or 1.45×10^5 lb/in.^{2}).

For liquids that freeze to a solid phase of greater density than the liquid, the effect of pressure is to raise the freezing temperature. In the case of substances, such as water, having a solid phase of lower density than the liquid, the freezing point is lowered by pressure. The freezing point of water is lowered to −4°F (−20°C) by a pressure of about 2 kbar (0.2 GPa or 2.9×10^4 lb/in.^{2}). Above pressures of 2 kbar, however, water crystallizes in new solid forms which are denser than the liquid, and the freezing point is raised by further increases of pressure. At 45 kbar (4.5 GPa or 6.5×10^5 lb/in.^{2}) the freezing point of water is 374°F (190°C).

In general solids are less compressible than liquids, and the compressibility of both solids and liquids decreases with increasing pressure. However, solids and liquids show wide individual variations in compressibility. At a pressure of 200 kbar (20 GPa or 2.9×10^6 lb/in.^{2}) solid sodium is reduced to about half its initial volume, but diamond is reduced in volume by only a few percent at the same pressure. The electrical conductivity of metals is generally increased by pressure, but there are numerous exceptions to this rule. Changes in the electrical conductivity of metals are typically on the order of 10% for a 10-kbar (1-GPa or 1.45×10^5 lb/in.^{2}) pressure change. Phase transitions in metals generally result in discontinuities in the pressure-versus-resistance curves. These resistance discontinuities in various metals, for example, bismuth, iron, and lead, may be used as calibration points for high-pressure apparatus. *See* ELECTRICAL CONDUCTIVITY OF METALS.

At high pressure many solids exhibit polymorphic phase changes, that is, a rearrangement of the atoms or molecules in the solid. There are no universally applicable rules governing the number of phase changes or the kind of phase change to be expected at high pressure, but there is a thermodynamic requirement that the phase that is stable at high pressure must have a smaller volume than the phase that is stable at low pressure. Camphor has 11 probable solid phases, and seven solid phases of water are known to exist in the pressure range extending to 45 kbar (4.5 GPa or 6.5×10^5 lb/in.^{2}). On the other hand, many elements and compounds are known to exist in only one solid phase up to pressures of over 1000 kbar (100 GPa or 1.45×10^7 lb/in.^{2}). *See* THERMODYNAMIC PRINCIPLES.

Frequently, dramatic changes in physical properties result from phase changes. Ferromagnetic iron transforms to a paramagnetic form at pressures somewhat above 100 kbar (10 GPa or 1.45×10^6 lb/in.^{2}). In the same pressure range, the semiconducting element germanium transforms into a metallic phase that has an electrical conductivity greater than a million times that of the semiconductor. Similar semiconductor-to-metal transitions at high pressure have been observed in the cases of silicon, indium arsenide, gallium antimonide, indium phosphide, aluminum antimonide, and gallium arsenide. *See* SEMICONDUCTOR.

Many phases that form at high pressure transform back to low-pressure phases as the pressure is released. However, some high-pressure phases may be retained in a metastable condition at low pressures, and some low-pressure phases can persist metastably at high pressure. Diamond, the high-pressure form of carbon, is thermodynamically unstable at room temperature and pressures below about 12 kbar (1.2 GPa or 1.74×10^5 lb/in.^{2}). Nonetheless diamond persists indefinitely as a metastable phase at low temperatures; it transforms to the stable form, graphite, only when heated to temperature in excess of 1800°F (1000°C) at low pressure. Graphite can be exposed to pressures in excess of 200 kbar (20 GPa or 2.9×10^6 lb/in.^{2}) without transforming to diamond. Simultaneous high temperatures are required before diamond can form. Through use of a molten metal catalyst to lower the activation energy for the transformation, diamond is commercially synthesized at pressures of about 40 kbar (4 GPa or 5.8×10^5 lb/in.^{2}) and temperatures of about 2700°F (1500°C). Without the catalyst, pressures in excess of 100 kbar (10 GPa 1.45×10^6 lb/in.^{2}) and temperatures above 3600°F (2000°C) are required for diamond synthesis. *See* METASTABLE STATE.

An important area of research is the study of electronic transitions at high pressure. Mössbauer spectral measurements of iron compounds have demonstrated the reversible reduction of Fe^{3+} (ferric) ions to Fe^{2+} (ferrous) ions at high pressures. Similar reductions of Mn^{3+} to Mn^{2+} and of Cu^{2+} to Cu^{+} have been observed by optical spectroscopy at high pressures. This research is relevant to the problem of estimating the possible chemical composition of the interior of the Earth. Some researchers now hypothesize that only a negligible amount of ferric iron (Fe^{3+}) is present at depths below 600 mi (1000 km).

Experimental methods and problems. The simplest type of high-pressure apparatus is a thick-walled, hollow cylinder closed by sliding pistons to form a cavity. The sample is contained in the cylinder and force is applied to the pistons. Both the force and the change in volume can be measured by ordinary instruments. However, large corrections must be made for the pressure-induced distortion of the sample cavity and for the friction between pistons and cylinder. For liquids, the stresses will be hydrostatic, and the corrections for sample-cavity distortion may be calculated from elasticity theory. However, additional corrections, for example, for friction and for compression of the seals used to prevent leakage, must be determined empirically. Electrical leads may be introduced into the pressure cavity; the liquid may then serve as a pressure-transmitting fluid during measurements of the electrical properties of various materials. The electrical leads can also be connected to a coil of wire made from a metal or alloy having a well-known pressure variation of resistivity. The pressure in the liquid may then be determined independently by measuring the resistance change in the wire. Measurement of the compressibility of solid samples can best be made if the sample is immersed in a fluid of known compressibility.

At pressures above about 25 kbar (2.5 GPa or 3.6×10^5 lb/in.^{2}) at room temperature, however, most liquids have solidified, and solid pressure-transmitting media must be used. At these higher pressures, an important part of the experimental problem is to ensure that the nonhydrostatic, or shearing, components of stress are as small as possible in comparison with the hydrostatic component. The pressure-transmitting medium must be a soft solid, chosen empirically as one which does not become prohibitively stiff at high pressure.

The strength of materials presently available limits the simple piston-and-cylinder apparatus to a maximum pressure of about 50 kbar (5 GPa or 7.2×10^5 lb/in.^{2}). Apparatus designed for higher pressures takes advantage of the fact that the strength of many materials is greatly enhanced by pressure. Multistage apparatus has been built in which the sample is contained in a pressure vessel, which is in turn contained in a larger pressure vessel. Pressures up to about 90 kbar (9 GPa or 1.3×10^6 lb/in.^{2}) have been reached in two-stage apparatus, at the cost of greatly increased complexity of the apparatus and marked reduction in the accuracy with which measurements can be made. Through ingenious use of geometrical factors, the benefits of multistaging can be realized without having to construct concentric pressure vessels. However, the problem of measurement is enormously complicated, because only a small and imprecisely known fraction of the total force applied to the system actually serves to compress the sample. In some types of apparatus, over 98% of the total applied force is used to provide support for the highly stressed regions of the apparatus.

Temperature may be varied in high-pressure experiments by externally heating or cooling the entire pressure vessel. The temperature limit for externally heated pressure vessels is about 1800°F (1000°C), a limit determined by the reduction in strength of the vessel material at high temperatures. Higher temperatures may be obtained in internally heated pressure chambers in which heat is generated by passage of current through a resistance element; alternatively, the heat of a chemical reaction may be used. Temperature is measured by thermocouples within the pressure cavity, and suitable corrections must be made for the effect of pressure on thermocouple output. The effect of pressure on thermocouple output varies greatly according to the combination of metals or alloys chosen for the thermocouple junction. If two different thermocouple junctions are subjected to the same temperature and pressure, their outputs may be used to determine both temperature and pressure. *See* THERMOCOUPLE.

Sustained pressures in the megabar (100 GPa or 1.45×10^7 lb/in.^{2}) range are achieved in the diamond anvil apparatus. The sample cavity consists of a hole in a metal sheet compressed between two diamond surfaces. A typical sample cavity, having a diameter of 100 μm (0.002 in.) and a height of 50 μm (0.002 in.), can contain several ruby chips and a pressure-transmitting medium in addition to the material under investigation. The diamond anvils also serve as windows permitting direct optical observation of the sample and ruby chips. When ruby is illuminated by ultraviolet light, it fluoresces (emits light) of a characteristic wavelength that depends on pressure. This relationship between pressure and wavelength is often called the ruby pressure scale, and it is commonly used for determination of pressure in the diamond anvil apparatus. Because of their inertness and low shear strengths, solidified rare gases such as helium and neon are frequently used as pressure-transmitting media.

X-ray diffraction has become an important tool for studying volume changes and phase changes at high pressure. A portion of the pressure chamber must be constructed of a material of low atomic number, such as boron, beryllium, or diamond, to provide a "window" that is relatively transparent to x-rays. If an x-ray beam of suitable wavelength is passed through the sample, a portion of this beam is diffracted by the sample. The x-ray diffraction pattern, which may be recorded on film, is characteristic of the crystal structure and the atomic spacings of the material under investigation. If a calibrant (that is, a material that behaves in a known manner under pressure) is mixed with the specimen, the diffraction patterns of calibrant and specimen may be simultaneously recorded at high pressure. The diffraction pattern of the calibrant then serves as an internal standard of pressure. *See* X-RAY DIFFRACTION.

Previously, long exposure times (about a day) were required to record an x-ray diffraction pattern at a single pressure. Now many researchers transport their diamond anvil apparatus to very intense synchrotron x-ray sources where they can accomplish as much work in a week as would have taken a year or more with conventional x-ray sources. *See* SYNCHROTRON RADIATION.

Other ingenious techniques have been used for high-pressure studies of the Mössbauer effect, nuclear magnetic resonance, diffusion in solids, and many other effects of current interest in solid-state physics research. The primary problem of high-pressure physics continues to be accuracy of measurement.

No article on high-pressure research would be complete without mention of the late Percy W. Bridgman, who won the 1946 Nobel Prize in physics. For 40 years, beginning in 1908, Bridgman's work completely dominated the field of high-pressure physics. Virtually all high-pressure research depends heavily on apparatus designed by Bridgman and on his pioneering measurements of changes in physical properties at high pressure. *See* HIGH-PRESSURE MINERAL SYNTHESIS; HIGH-PRESSURE PROCESSES.

R. K. Linde; P. S. DeCarli

Bibliography. J. R. Asay and M. Shahinpoor (eds.), *High Pressure Shock Compression of Solids*, 1993; R. A. Graham, *Solids Under High Pressure Shock Compression*, 1992; R. M. Hazen, *The New Alchemists: Breaking Through the Barriers of High Pressure*, 1993; H. D. Hochheimer and R. D. Etters (eds.), *Frontiers of High Pressure Research*, 1992; C. Homan et al., *High Pressure in Science and Technology*, 3 pts., 1984; W. F. Sherman and A. Stadtmuller, *Experimental Techniques in High Pressure Research*, 1987.

High-pressure processes

Changes in the chemical or physical state of matter subjected to high pressure. The earliest high-pressure chemical process of commercial importance was the Haber synthesis of ammonia from hydrogen and nitrogen developed in Germany prior to World War I. The synthesis of diamonds from graphite developed in the early 1950s is a high-pressure physical process. Raising the pressure on a system may result in several kinds of change. It causes a gas or vapor to become a liquid, a liquid to become a solid, a solid to change from one molecular arrangement to another, and a gas to dissolve to a greater extent in a liquid or solid. These are physical changes. A chemical reaction under pressure may proceed in such a fashion that at equilibrium more of the product forms than at atmospheric pressure; it may also take place more rapidly under pressure; and it may proceed selectively, forming more of the desired product among multiple possible products.

Pressures higher than that of the atmosphere are expressed in bars and kilobars as well as in other units. A bar is 10^5 pascals, or 10^5 newtons per square meter, which are the units for pressure in the International System of units. These units are too small for convenient use in high-pressure processes, hence the bar is used. The bar equals 0.9869 standard atmosphere, 760 mmHg.

Physical processes. Increasing the pressure on a gas or vapor compresses it to a higher density and so to a smaller volume. If the pressure exceeds the vapor pressure, the vapor will condense to a liquid which occupies a still smaller volume. A vapor may be condensed at a higher temperature when it is under pressure; this permits the use of cooling water to remove the latent heat instead of more costly refrigeration.

Solids also change from a less dense phase to a more dense phase under the influence of increases in pressure. The density of diamond is about 1.6 times greater than that of graphite because of a change in the spatial arrangement of the carbon atoms. The temperatures and pressures used in the commercial synthesis of diamond range up to 5000°F (3000 K) and 100,000 atm. A molten metal is required as a catalyst to permit the atomic rearrangement to take place at economical rates of conversion. Metals such as tantalum, chromium, and iron form a film between graphite and diamond.

The highest static pressure attained in the laboratory is about 600,000 atm. At still higher pressures, the electrons are stripped from the atomic nuclei and matter loses its identity as recognizable atoms. This situation apparently exists in the centers of the white dwarf stars, where pressures of the order of 10^{16} atm prevail.

Chemical processes. In a manner similar to its effect during a physical change in which the volume of a system decreases, pressure also favors a chemical change where the volume of the products is less than the volume of the reactants. This is Le Chatelier's principle, which applies to systems in equilibrium. This general principle may be derived more precisely by thermodynamic reasoning, and thermodynamics is used to predict the effect of pressure on physical and chemical changes which lead to an equilibrium state.

Ammonia production. Ammonia is formed according to the reaction shown below. The ammonia content at

$$N_2 + 3H_2 \rightleftharpoons 2NH_3$$

equilibrium in a mixture which contains initially a ratio of 3 moles of hydrogen to 1 mole of nitrogen is shown in **Fig. 1**. At 1 atm only a fraction of 1% ammonia is formed. The ammonia content increases greatly when the pressure is raised. At 100 atm and 392°F (200°C) there would be about 80% ammonia at equilibrium. However, a very long time is required to form ammonia under these conditions, and consequently commercial processes operate at higher temperatures and pressures and use a catalyst to obtain higher rates of reaction. Many combinations of pressure and temperature have been used. The largest number of plants now operate in the region of 300 atm and 840–930°F (450–500°C). A higher pressure process is carried out at about 1000 atm and 930–1200°F (500–650°C).

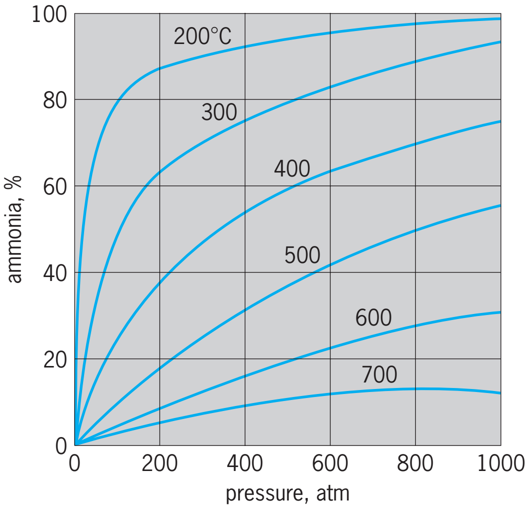


Fig. 1. Effect of combined temperature and pressure on the formation of ammonia. °F = (°C × 1.8) + 32; 1 atm = 1.01 bar. (*After E. W. Comings, High Pressure Technology, McGraw-Hill, 1956*)

A catalyst must be used or the reaction is too slow. Thus the process does not operate at 100 atm and 390°F (200°C) in spite of the favorable conversion at equilibrium because no catalyst has been found to provide an adequate rate of reaction under these conditions. Iron is the catalyst used at higher temperatures and pressures. The physical form of the catalyst is very important. Iron shavings or lumps are not effective. A suitable catalyst has been made from a fused mixture containing 66% Fe_2O_3, 31% FeO, 1.0% K_2O, and 1.8% Al_2O_3. The last two components are promoters and significantly increase the activity of the catalyst, but they alone are not catalysts. The mixture must not be contaminated with other impurities or the catalyst may be poisoned. Sulfur, phosphorus, and arsenic poison the catalyst permanently, whereas gases such as oxygen, water vapor, and carbon oxides reduce the catalyst's activity only while thay are in contact with it and are temporary poisons. Before the catalyst is used, the iron oxides are first reduced to iron by the hydrogen of the reacting mixture. The catalyst is activated by this treatment and must not be exposed to gases, which may contain poisons. Its life may be from several months to several years. Increasing the pressure increases the rate of reaction of the mixture in contact with a suitable catalyst.

A flow sheet for an ammonia process operating at 1000 atm is shown in **Fig. 2**. Nitrogen is obtained from the air, and hydrogen comes from natural gas. The process involves preparation of the nitrogen-hydrogen mixture, purification to exclude catalyst poisons, compression to high pressure, and circulation of the synthesis gas through a closed loop which contains two synthesis converters charged with catalyst. Because of the incomplete conversion in one converter, the unreacted gas is passed on to the next converter after ammonia has been condensed from the partially reacted gas mixture.

The temperature of the reaction is carefully controlled. Higher temperatures shorten the life of the catalyst and reduce the degree of conversion at equilibrium, whereas lower temperatures reduce the rate

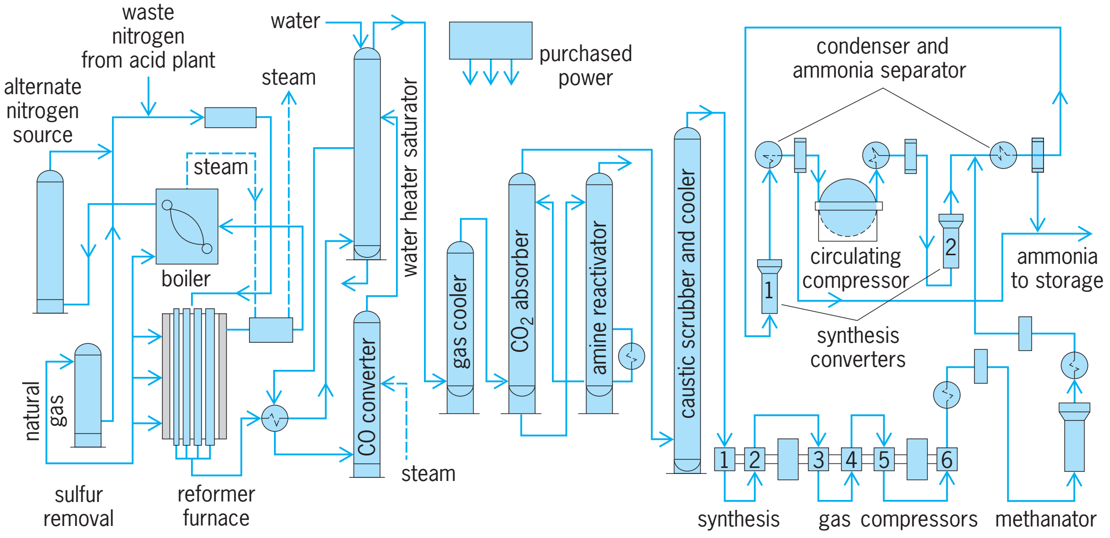


Fig. 2. Ammonia process at 1000 atm. 1 atm = 1.01 bar. (*After W. H. Shearon and H. L. Thompson, Ind. Eng. Chem., 44(2):254–264, 1953*)

of reaction. The reaction mixture entering the converter is relatively cool. It is heated to the reaction temperature inside the converter by heat exchange with the reacting gas and with the gas that has passed through the catalyst. The reaction is exothermic, and the converter is designed to remove some of the heat of reaction while the gas is passing through the catalyst.

The converters are large pressure vessels containing a heat exchanger and a catalyst basket. The converter for the 1000-atm process contains 16.5 ft^3 (0.5 m^3) of catalyst and produces 250–400 lb (110–180 kg) of ammonia per hour in each cubic foot of catalyst. This is comparable to 144 ft^3 (4 m^3) of catalyst in a converter for a 300-atm process, which produces 50 lb (22.5 kg) of ammonia per hour per cubic foot of catalyst. Molecular sieves (refractory compounds with controlled porosity) are coming into use as a base for catalysts to reduce the volume and increase the life of the catalysts. The high reaction temperature is confined to the catalyst basket inside the vessel, and the thick walls of the vessel are held at a lower temperature.

Methanol production. Methanol is synthesized from hydrogen and carbon monoxide at 200 atm and 600°F (315°C) in a similar manner. The catalyst contains aluminum oxide, zinc oxide, chromium oxide, and copper. Higher alcohols are produced at pressures of 200–1000 atm and temperatures up to 1000°F (538°C) with a similar catalyst to which potassium carbonate or chromate has been added.

Polyethylene production. Polyethylene has been produced at pressures in the ranges 3–4, 20–30, 40–60, and 1000–3000 atm. The latter is probably the highest pressure yet used in the commercial synthesis of an organic chemical product. The ethylene is polymerized in a stainless steel tubular reactor at 375°F (191°C) with small amounts of oxygen as a catalyst.

Phenol production. Phenol can be formed from chlorobenzene mixed with 18% sodium hydroxide solution at a pressure of 330 atm. Pressure is employed in this instance to maintain the mixture in the liquid phase at a temperature high enough for the hydrolysis reaction to proceed at an acceptable rate. A tubular reactor installed in a furnace provides for heating the mixture to 680°F (360°C).

Hydrogenation. Hydrocracking and hydrodesulfurization in the refining of gasoline and fuel oils are carried out at pressures up to 200 atm and temperatures of 800°F (427°C) and higher. The large volumes processed require very large reaction vessels, 8 ft (2.4 m) in diameter by 80 ft (24 m) long with 8-in.-thick (20-cm) walls, for example.

Other chemical reactions are carried out at elevated pressure. A reaction which has been observed to take place in the laboratory under high pressure may later be found to take place at lower pressure with a suitable catalyst. *See* HYDROCRACKING; HYDROGENATION.

Apparatus. High-pressure apparatus is carefully designed to provide for the measurement of pressure and temperature, a vessel with sufficient strength, a closure to prevent leakage, easy access to the interior, for the compression of gases and liquids at high pressures, and for the safety of people and equipment in case of ruptures or explosions. The deadweight gage is the principal primary gage for calibrating other pressure-measuring instruments. The deadweight gage consists of a free piston that is balanced between the force of gravity acting on weights at one end and the pressure of oil contained in the system at the other end.

P. W. Bridgman received the Nobel Prize in physics in 1946 for his pioneering work in the physics of high pressure. He was the originator of a self-sealing closure based on the unsupported area principle. This principle is illustrated in **Fig. 3**. The sealing gasket is initially under a low or moderate pressure. As pressure rises in the vessel, the pressure in the gasket is automatically maintained at a level higher than in

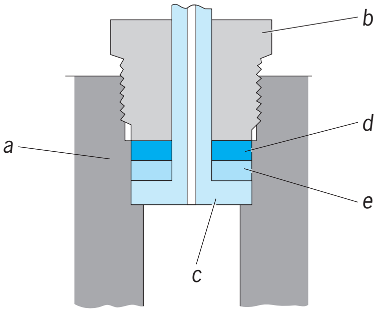


Fig. 3. Bridgman unsupported area gasket. Refer to the text for *a–e*. (*After E. W. Comings, High Pressure Technology, McGraw-Hill, 1956*)

the vessel. This ensures against leakage of the contents. Pressure, contained in the vessel *a*, acts upward on the steel disk *c*, which in turn acts on the gasket *e*. The gasket is supported by the steel ring *d*, which has a smaller area than *c* because of the unsupported area in the center. A balance of forces on the gasket requires that the forces acting downward balance those acting upward. The upward force is the product of the pressure in the vessel and the area of the lower face of *c*. This force is balanced by an equal force downward, which is the product of the pressure in the gasket and the area of the gasket. Because the area of the gasket is smaller than the area of the lower face of *c*, the pressure in the gasket must always be higher than the pressure in the vessel. *See* CHEMICAL EQUILIBRIUM; HIGH-PRESSURE CHEMISTRY; HIGH-PRESSURE PHYSICS; PHASE EQUILIBRIUM. Edward W. Comings

Bibliography. P. W. Bridgman, *The Physics of High Pressure*, reprint 1949; C. Homan et al., *High Pressure in Science and Technology*, 1984; K. Ishizaki, J. K. Tien, and E. Hodge (eds.), *Design Effects on Materials Processing and Design*, 1992; G. N. Peggs (ed.), *High Pressure Measurement Techniques*, 1983; I. L. Spain and J. Paauwe, *High Pressure Technology: Applications and Processes*, vol. 2, 1977; R. Van Eldik, *Inorganic High Pressure Chemistry: Kinetics and Mechanisms*, 1987; R. Van Eldik and J. Jonas (eds.), *High Pressure Chemistry and Biochemistry*, 1987.

High-temperature chemistry

The study of chemical phenomena occurring above 500 K (440°F). High temperatures represent one of the important variables available to scientists for increasing the variety of possible chemical reactions over that expected for classical ground-state atoms and molecules. One can enhance the relative population of excited rotational, vibrational, and electronic states by increasing the temperature and thus effectively create new species and new mechanisms for reaction. The potentialities of this approach are well illustrated by the three laws of high-temperature chemistry: (1) At high temperatures everything reacts with everything. (2) The higher the temperature, the faster the reaction. (3) The products may be anything. With an infinity of species available at high temperatures, the "golden age" of chemical synthesis is probably still in the future.

High temperatures also provide a common tie among the various options for energy production, conversion, or storage. For maximum thermodynamic efficiency, an energy production cycle should operate with a working fluid at as high a temperature as possible, and exhaust the spent fluid at as low a temperature as possible. Thus, in the combustion of coal to produce electric power or in the combustion of gasoline or diesel fuel to propel a car or an airplane, there is a need for materials of construction which allow operation of such devices at high temperatures. In the evaluation of new fuels or propellants, higher flame temperatures are among the desirable properties often sought.

It is convenient to discuss temperatures in terms of energy and to note that 11,500 K (20,200°F) corresponds to 1 electronvolt. In this sense, the particles emitted by radioactive nuclei or accelerated in cyclotrons and synchrotrons, which have energies in the keV, MeV, and BeV ranges, are effectively at temperatures of $\sim 10^7$ K, $\sim 10^{10}$ K, and $\sim 10^{13}$ K, respectively, and "high-energy physics" is synonymous with "ultrahigh temperature chemistry."

Traditional high-temperature chemistry has been mainly concerned with phenomena in the range of 500–3000 K (440–4900°F), although exotic flames can produce temperatures up to ~6000 K (10,300°F), shock waves can generate temperatures up to ~25,000 K (45,000°F), electric arcs can be operated in constricted modes to produce temperatures of ~50,000 K (90,000°F), and nuclear processes begin to occur at temperatures in the millions of degrees range. Laser excitation of selected energy states can produce species with effective temperatures in the range of 10^8 K. The major goals of high-temperature scientists include (1) the characterization of all important gaseous molecules, ions, and condensed phases—molecular formulas and structures, energy levels, thermodynamic properties, and the details of chemical bonding; (2) the establishment of reaction rate parameters and correlations with molecular properties; (3) the development of unique approaches to chemical syntheses and the preparation of new materials; and (4) the development of new techniques for generating or containing or utilizing high-temperature fluids in connection with energy production, conversion, and storage.

Gaseous species. Studies of thermal decomposition, vaporization, and/or sublimation of inorganic compounds have shown that complex gaseous species are much more common than usually realized. For example, condensed alkali metal sulfates vaporize appreciably as M_2SO_4 (gas); alkali metal carbonates vaporize as M_2CO_3 (gas); and even NH_4X (solid) yields NH_4X (gas). At high temperatures the relative fraction of complex species in the vapor actually increases with the increasing temperature for many systems. Vapors at high pressures and high temperatures can thus be complex. Rapid mass spectrometric sampling techniques have demonstrated unequivocally the existence of polymers such as $(Ar)_m$, $(CO_2)_n$, and $(H_2O)_q$ where m, n, and q are 2,3,4, . . . , 8,9,10, ... As a further complication, various oxidation states can exist in high-temperature systems and thereby a still greater variety of possible molecular species. Typical systems which have been characterized by a combination of mass spectrometric, optical, infrared, nuclear magnetic resonance, and electron spin resonance spectroscopy and chemical studies are the difluorides of group 14 (CF_2, SiF_2, GeF_2, SnF_2, PbF_2) and their polymers.

In particular, the technique of generating high-temperature gaseous species and then reacting them with low-temperature molecules on a surface which is maintained at low temperatures (∼−190°C or −310°F) has been very productive. Practical approaches to a great variety of organometallic syntheses have been developed through the use of molecules such as SiF_2, C_2, C_3, BF, SiO, and SiS, and various atoms—such as C, Si, Li, Mg, Fe, Cu, Ni, Pd, and Pt. Matrix-isolation spectroscopy at 4–100 K (−452 to −279°F) provides the opportunity for gaining detailed information about atomic/molecular parameters—energy levels, frequencies, bond angles—which are necessary for calculation of thermodynamic functions and the prediction of chemical reactions.

Condensed systems. Condensed systems at high temperatures provide additional versatility in the chemical world since one fails to observe almost uniquely in high-temperature systems that atoms do combine in the ratio of small whole numbers. Such observations on gases led to the formulation of the law of definite proportions and other basic rules of early chemistry, but these rules no longer hold exactly for condensed systems at temperatures in the 1000–3000 K (1300–4900°F) range. Experiments have established melting points of refractory solids (binary carbides, borides, silicides, and so forth) at 1 atm (10^2 kPa) pressure in the 2000–4000 K (3100–6700°F) range, and unless one goes to higher pressures than 1 atm there really is no sold-state chemistry at temperatures greater than ∼4000 K (6700°F). Obviously, there is a chemistry of liquid systems which might be of some technological importance, and ingenious techniques for generating and maintaining reactive high-temperature liquids have been described.

Phase diagrams have demonstrated conclusively that one can prepare crystalline solids in an almost infinite number of compositions. For example, there are "pure" compounds fitting classical valences, and then one can substitute, as in $MgAl_2O_4$, where Al^{3+} ions can occupy some of the Mg^{2+} sites or Mg^{2+} ions can occupy some of the Al^{3+} sites so that there is an almost continual variation in properties. One can prepare Ta-O-C phases in which the gross stoichiometry is $Ta_1O_xC_{1\text{-}x}$ and, further, find it extremely difficult to detect the fact that oxygen is even present since the sizes of the oxygen and carbon atoms are so nearly the same in this lattice. Discovery of oxynitride phases and the establishment of essentially continuous solid solutions over wide ranges of compositions for many oxide, sulfide, carbide, boride, and other systems have been reported. The variations in compositions of alloys are similarly extremely complex; the variety possible is only beginning to be appreciated. Techniques for presenting phase diagrams have been described in which an attempt is made to present a single diagram which summarizes the behavior of whole classes of alloy systems rather than that of individual binary or ternary combinations. *See* NONSTOICHIOMETRIC COMPOUNDS.

Properties of high-temperature solids are, of course, not reliably interpreted until the exact stoichiometric and structural data have been obtained. The status of high-temperature thermodynamic properties is acceptably characterized by saying that no material is described thermodynamically to better than ±0.5% up to 1500 K (2200°F), or better than to ±1–2% from 1500–2000 K (2200–3100°F). There are practically no reliable data at temperatures greater than 2000 K (3100°F) except for a handful of basic materials—tungsten, molybdenum, tantalum, and aluminum oxide.

There is a great need for experimental techniques by which thermodynamic measurements can be reliably extended into this high-temperature region. There is already experimental evidence which is not explainable by current theoretical viewpoints on solids and liquids at high temperatures. Vaporization, sublimation, and other heterogeneous equilibrium studies have been made for many systems, and these are, of course, sensitive to the exact nature of the surfaces, to variations in stoichiometry as gases are evolved from a condensed phase, and to other factors which may accompany deviations from equilibrium.

Electrochemical techniques. One of the unique ways in which high-temperature thermodynamic data have been obtained involves the use of electrochemical techniques of the same sort as those which were widely applied during the 1920–1940 period for the establishment of reliable thermodynamic reference data for aqueous systems and for simple solids at temperatures near 25°C (77°F). High-precision free energies of formation for nonstoichiometric phases can be derived over fairly wide ranges of temperature. However, eventually at 1500 K (2200°F) and higher, one begins to have difficulty with vaporization, melting, and high diffusion rates.

Calorimetry. Calorimetry at temperatures up to 3000 K (4900°F) became routine through the use of electron bombardment heating and levitation heating. In electron bombardment heating, one boils electrons out of a thermionic emitter, such as thoriated tungsten, and accelerates them across a potential drop of a few thousand volts to strike the conducting sample and raise its temperature by means of the kinetic energy transferred and also by the heating caused by the electric resistance of the material. Levitation heating of conducting samples is accomplished with standard radio-frequency induction heaters and a pair of oppositely wound (left-handed and right-handed) coils. Samples weighing up to 1000 g (35 oz) can be levitated, melted, and cast in a containerless, controlled atmosphere. Copper, gold, platinum, tantalum, graphite, and many other materials have been levitated and, in all cases except graphite, raised to temperatures above their melting points (3000 K or 4900°F and higher). Levitation calorimetry has provided heats of fusion and heat capacities for many metals, alloys, and conducting compounds in both solid and liquid states.

Chemical kinetics. From the viewpoint of heterogeneous kinetics, the high-temperature area has been studied extensively, with the practical concerns involving (1) the rates of interaction between various corrosive gases (oxygen, nitrogen, sulfides, halogens, and so on) and surfaces of various pure metals, alloys, and ceramic structural materials, and (2) the catalytic effects of solids of various stoichiometries on various gas reactions as in hydrocarbon refining, in the preparation of SO_3 or NH_3, or in the catalytic conversion of $CO + H_2$ to methanol, of CO to CO_2, or of NO_x to N_2 and O_2.

Homogeneous gas kinetics in the region above 1000 K (1300°F) has been mainly concerned with reactions of neutral atomic and molecular species which were stable at high temperatures, as in typical flames. The roles of H, O, OH, and of intermediate combustion products like HO_2, MO, and CHO have been established by direct mass-spectrometer probing of flames. The importance of electrons and atomic or molecular ions in flame kinetics has also been recognized. This knowledge is of interest in identifying the most efficient combustion systems for energy generation and in elucidating the chemical parameters which are crucial to the development of flame-retardant materials.

Another fertile area for high-temperature kinetics research has been the study of the reaction rates of atoms or molecules which have been produced by either photochemical, laser-pulse, or electric-arc excitation or by pulsed thermal dissociation processes and then allowed to react either in systems at relatively higher pressures by random collisions or in systems at low pressures by molecular-beam techniques. Thus, a hydrogen atom created from the H_2 molecule in an arc or by thermal dissociation requires 2.24 eV/atom, and therefore is a chemical species of the sort that might be expected in a very high-temperature system. The thermodynamic potential of such atoms for reaction is much greater than that for most molecular species, and reaction rates are usually measurable although relatively fast. Selective excitation to specific energy levels is possible with tunable dye lasers. *See* ULTRAFAST MOLECULAR PROCESSES.

Several types of monitoring techniques, rapidscan infrared spectroscopy, electron spin resonance spectroscopy, and mass spectrometry, are typically used in these studies. Gas reaction rates in mass spectrometers also have been widely explored by scientists interested in ion-molecule reactions. Of course, since ions are created endothermally by the impartation of several electronvolts of energy, they are high-temperature species. Many new types of high-temperature reactants can be created. It is possible that a synthetic chemistry making use of ion-molecule reactions may someday be of economic significance. Ion sputtering is a widely used technique for the preparation of electronic circuitry. *See* CHEMICAL DYNAMICS. John L. Margrave

Bibliography. L. Eyring (ed.), *Advances in High Temperature Chemistry*, vols. 1–4, 1967–1975; J. W. Hastie, *High Temperature Vapors*, 1975; J. L. Margrave (ed.), *Modern High Temperature Science*, 1984; E. T. Turkdogan, *Physical Chemistry of High Temperature Technology*, 1980.

High-temperature materials

Materials that serve above about 1000°F (540°C). In the broad sense, high-temperature materials can be identified by the following classes of construction solids: stainless steel (limited), austenitic superalloys, refractory metals, ceramics and ceramic composites, metal-matrix composites, and graphitic composites. The first three classes are well proven in industrial use, although stainless steels serve but slightly above 1000°F (540°C) and refractory metals are usually limited to nonoxidizing atmospheric conditions. The other classes are under extensive worldwide research to establish whether they can be utilized to replace and extend the capabilities of austenitic superalloys, which are the mainstay of high-temperature service.

The most demanding applications for high-temperature materials are found in the aircraft jet engines, industrial gas turbines, and nuclear reactors. However, many furnaces, ductings, and electronic and lighting devices operate at such high temperatures. In order to perform successfully and economically at high temperatures, a material must have two essential characteristics: it must be strong, since increasing temperature tends to reduce strength; and it must have resistance to its environment, since oxidation and corrosion also increase with temperature.

High-temperature materials have acquired their importance because of the pressing need to provide society with energy and transportation. Machinery that produces electricity or some other form of power from a heat source operates according to a series of thermodynamic cycles, including the basic Carnot cycle and the Brayton cycle, where the efficiency of the device depends on the difference between its highest operating temperature and its lowest temperature. Thus, the greater this difference, the more efficient is the device—a result giving great impetus to create materials that operate at very high temperatures. Following is a discussion of the characteristics of some of the more significant alloys, a brief review of strengthening mechanisms, and a description of several developments. *See* CARNOT CYCLE; EFFICIENCY; MECHANICAL ENGINEERING.

Metallic materials designs. Alloys used at high temperatures in heat engines are composed of several elements. High-temperature metallurgists, using both theoretical knowledge and application of empirical experimental techniques, depend upon three principal methods for developing and maintaining strength. The alloys are composed of grains of regular crystalline arrays of atoms and show usable ductility and toughness when one plane of atoms slips controllably over the next (dislocation movement); excessive dislocation movement leads to weak alloys. Thus, in terms of achieving mechanical strength, alloy design attempts to inhibit but not completely block dislocation movement.

The most common technique is solid solution strengthening. Foreign atoms of a size different from that of the parent group cause the crystal lattice to strain. This distortion impedes the tendency of the lattice to slip and thus increases strength. *See* ALLOY; CRYSTAL STRUCTURE.

A second significant mechanism is dispersion strengthening. This is the introduction into the alloy lattice of extremely hard and fine foreign phases such as carbide particles or oxide particles. Carbide dispersions can usually be created by a solid-state chemical reaction within the alloy to precipitate the particles, while oxide particles are best added mechanically, as described later. These hard particles impede slippage of the metal lattice simply by intercepting and locking the dislocations in place.

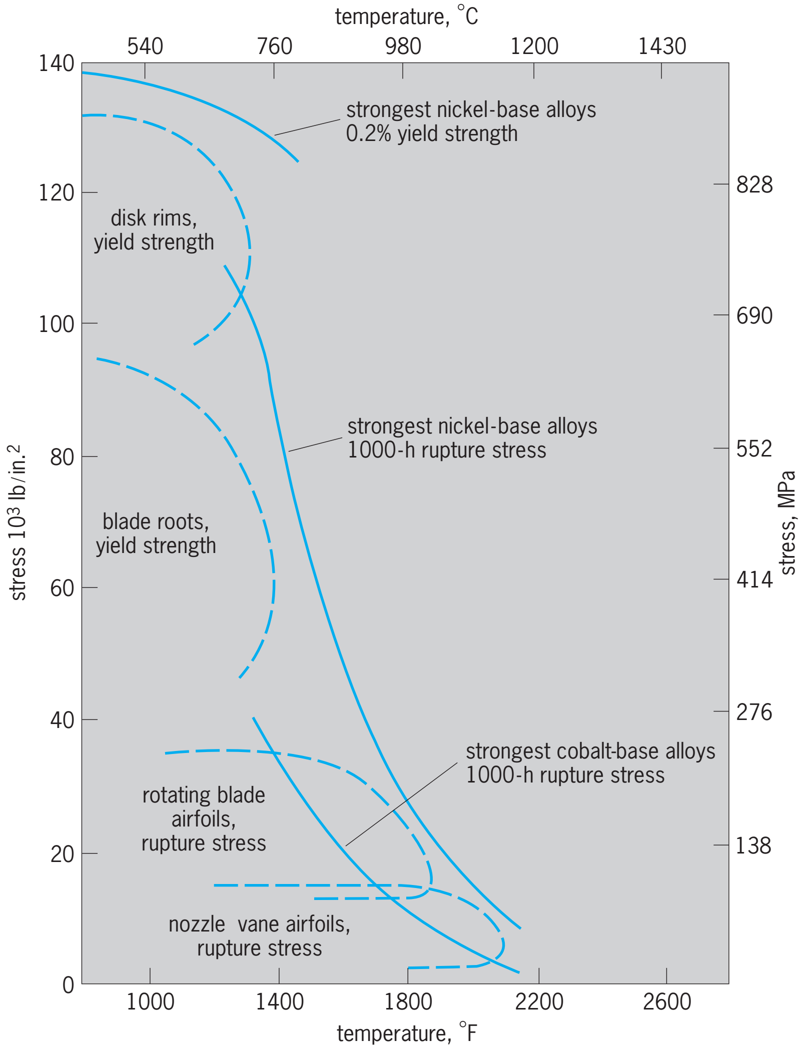


Fig. 1. **Capabilities (broken lines) of superalloys utilized in critical parts of aircraft jet engines. (*After C. T. Sims, N. S. Stoloff, and W. C. Hagel, eds., Superalloys II, John Wiley and Sons, 1987*)**

Still another strengthening technique is coherent-phase precipitation; a foreign phase component of similar but modestly differing crystalline structure is introduced into the alloy—always by a chemical solid-state precipitation mechanism. The phase develops significant binding strength with the mother alloy in which is resides (it is "coherent") and, like the other mechanisms, then impedes and controls dislocation flow through the metal lattice. Coherent phase precipitation is a special case within precipitation hardening.

However, certain advances in processing have had significant effects on these classical strengthening approaches in recent years, and they are covered separately below. *See* HEAT TREATMENT (METALLURGY); SOLID-STATE CHEMISTRY.

Metallic materials systems. These include superalloys, stainless steel, eutectic alloys, and intermetallic compounds.

Superalloys. High-temperature alloys also must be resistant to chemical attack by the atmosphere in which they exist, more often than not a high-speed high-pressure gas of highly oxidizing nature. Protection from the atmosphere is achieved by causing the material to develop a tough diffusion-resistant oxide film from various elements in the alloy, or by applying a protective coating to the material.

From all this, a high degree of success has been achieved with the superalloy class of materials. Superalloys serve under tough conditions of mechanical stressing and atmospheric attack to over 2000°F (1100°C), which often is within a few hundred degrees of their melting point. These alloys are capable of supporting modern aircraft jet engines (**Fig. 1**).

Undoubtedly, the most complex and sophisticated group of high-temperature alloys is the superalloys. A superalloy is defined as an alloy developed for elevated temperature service, where relatively severe mechanical stressing is encountered and high surface stability is frequently required. Superalloys, classically, are those utilized in the hottest parts of aircraft jet engines and industrial turbines; in fact, the demand of these technically sophisticated applications created the need for superalloys.

Superalloys are strengthened by all of the methods described above, as well as by some even more subtle ones, such as control of the boundaries between the grains of the crystalline metal by minor additives and other little-understood solid-state chemical reactions. **Figure 2** illustrates how these mechanisms combine to make nickel-base superalloys very strong. The superalloy René 77 (a nickel-base casting alloy used for turbine blades) is shown in two magnifications as a basis for illustrating some of these strengthening mechanisms. The continuous matrix of the alloy is the face-centered-cubic-phase austenite, referred to as γ. The most effective

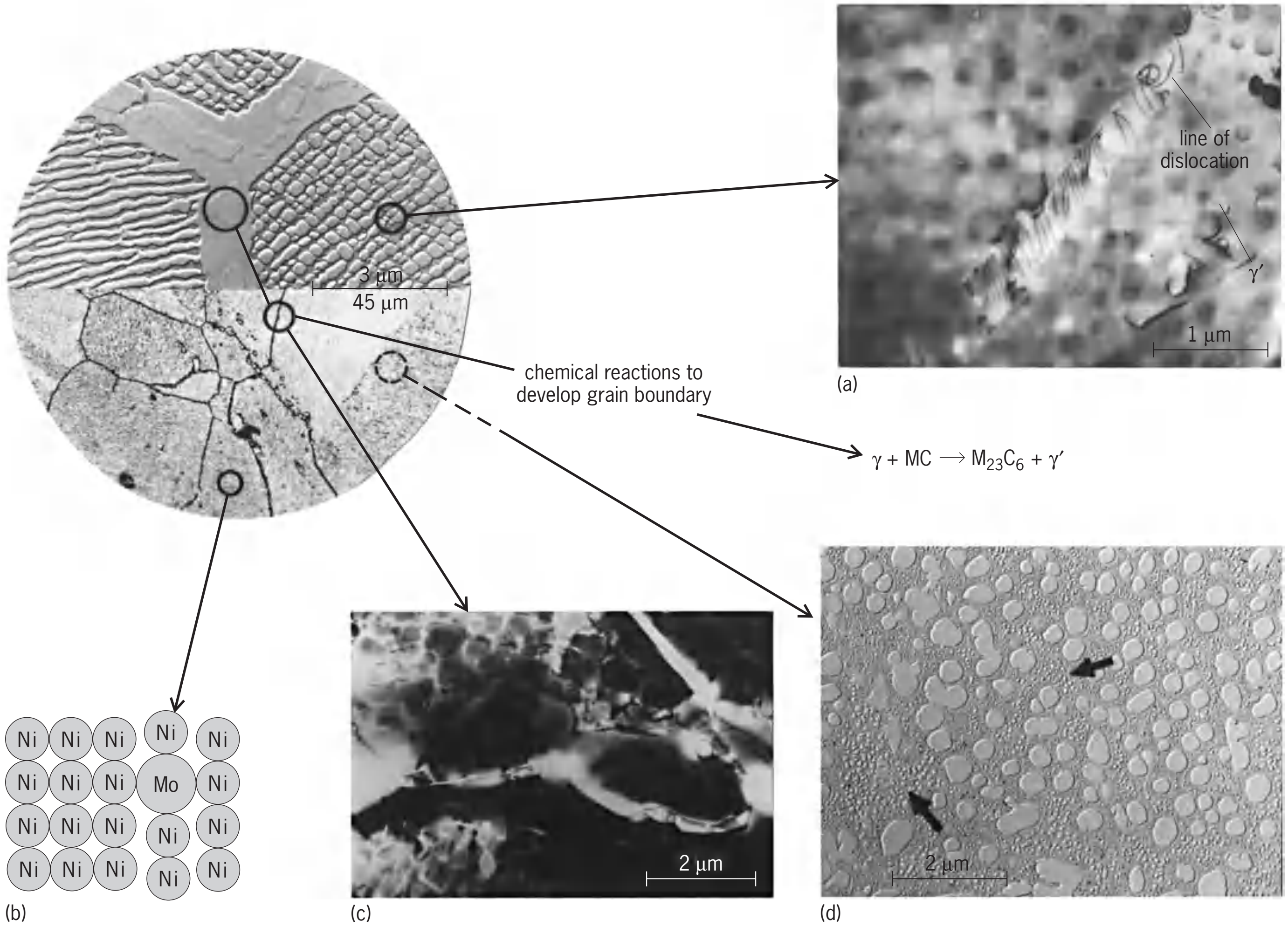


Fig. 2. Major features for achieving useful strength in nickel-base superalloys. (*a*) Dislocation interaction with γ', a coherent crystalline phase, shows dislocation movement. (*b*) Lattice after solid solution strengthening with molybdenum. (*c*) Carbide strengthening at grain boundaries. (*d*) Fine γ (arrows) for low-temperature strengthening; the big blobs are large γ'. (*From C. T. Sims and W. C. Hagel, eds., The Superalloys, Wiley-Interscience, 1973*)

strengthening effect comes from precipitation of γ', an intermetallic compound of nickel (Ni) and aluminum (Al), basic composition Ni_3Al. It is coherent and readily controlled by heat treatment. Chemical reactions between the γ and MC (a simple metallic carbide often present in superalloys) create $M_{23}C_6$ (a complex metallic carbide). These carbides, precipitated at boundaries between the grains, lock the grain boundaries in place and generate the dislocations. While these alloys were originally processed by melting in air, the large amounts and complex control of reactive metals needed eventually demanded processing in vacuum to maintain alloy cleanliness; these alloys contain as many as 15 different alloying additions.

Since the mid-1960s, advanced processing applied to superalloys has provided much additional capability. It was found that directional solidification could be utilized to reduce thermal fatigue failures and to create single-crystal components with still greater advantages. Dispersion of fine oxide particles in superalloys by powder metallurgy techniques also produced advantages.

Stainless steel. Stainless steels are strengthened principally by carbide precipitations and solid solution strengthening, since they cannot form the γ' coherent precipitate, their use is limited to about 1200°F (650°C), except where strength may not be needed. *See* STAINLESS STEEL.

Refractory alloys. Refractory metal alloys are based on elements that have extremely high melting points—greater than 3000°F (1650°C). These elements—tungsten, tantalum, molybdenum, and niobium (columbium)—are strengthened principally by a combination of solid solution strengthening, carbide precipitation, and unique metalworking. However, they cannot be used in modern heat engines because no commercially successful method of preventing their extensive reaction with oxidizing environments for broad application has been found. *See* REFRACTORY; SOLID SOLUTION.

Eutectic alloys. These metal-based systems have been under study since the late 1960s, but service applications have not been developed. They are usually similar to or based on superalloys, but they are processed by directional solidification so that a particularly strong stable phase forms as a needlelike structure, giving unusual strength. However, they are expensive to process, and industrial use is questionable.

Intermetallic compounds. These are metal-based systems centered on the fixed atomic compositions occurring in metallic systems of aluminum with nickel, titanium (Ti) and niobium (Nb), such as Ni_3Al, Ti_3Al, TiAl, and Nb_3Al. Intermetallic compounds are of interest because they often possess a lattice arrangement of atoms that leads to higher melting points and less ease of deformation. Some have shown ductility and toughness potential, and alloying to optimize properties is under significant study. For instance, Ti_3Al is fabricable and ductile, particularly when alloyed with niobium. None are in gas turbine service, but there is a possibility that Ti_3Al will become acceptable if the danger of titanium combustion and hydrogen embrittlement can be avoided. *See* INTERMETALLIC COMPOUNDS.

Oxidation and corrosion. In addition to possessing strength, high-temperature alloys must resist chemical attack from the environment in which they serve. Most commonly this attack is characterized by a simple oxidation of the surface. However, in machines which utilize crude or residual oils or coal or its products for fuel, natural or acquired contaminants can cause severe and complex chemical attack. Involved reactions with sulfur, sodium, potassium, vanadium, and other elements that appear in these fuels can destroy high-temperature metals rapidly—sometimes in a matter of a few hours. This is known as hot corrosion, and it is a major problem facing otherwise well-suited alloys for service in turbines operating in the combustion products of coal. *See* CORROSION.

Some nuclear reactors operate above 1000°F (540°C), and the working fluid is often not an oxidizing gas. For instance, high-temperature gas-cooled reactors in the development stages utilize high-temperature, high-pressure helium as the working fluid. However, the helium inevitably contains very low levels of impurities—oxygen, carbon, hydrogen, and others—and it does not contain enough oxygen to form a protective oxide film on the metals which contain it. Impurities enter these alloys and reduce strength by precipitating excessive amounts of oxide and carbide phases and by other deleterious effects. Refractory metals, nominally of high interest here, have shown long-term degradation by carbide-related mechanisms, and so metallurgists have been working to develop resistant superalloys. *See* HELIUM.

In other reactor applications such as the liquid metal fast-breeder reactor, construction metals must resist the high-temperature liquid metal used to transfer heat from hot uranium-plutonium fuel. The liquid metal is usually sodium or potassium. All classes of high-temperature alloys—stainless steels, superalloys, and refractory metals—have been under evaluation for service. As in helium gas, small amounts of impurities are the critical item. They can react with the metal at one temperature and transfer it to a component operating at a lower temperature. Stainless steels remain the prime construction materials for these breeder reactors. *See* NUCLEAR REACTOR.

The most significant problem, however, remains that of resistance to oxidation and high-temperature corrosion in present heat engines. Superalloys contain small-to-moderate amounts of highly reactive elements such as chromium and aluminum which react easily with oxygen to form a thick tenacious semiplastic oxide surface film. This prevents further reaction of the aggressive environment with the underlying metal. This is the reason why stainless steels are "stainless" all contain a minimum of 10% chromium. Superalloys follow approximately the same rule, but often also contain about 5% aluminum, which further enhances oxidation resistance. It has been found that small (<1.0%) additions of rare-earth elements (such as yttrium) further improve oxidation resistance by reducing oxide spalling during the inevitable thermal cycles which gas turbines experience. *See* RARE-EARTH ELEMENTS.

The natural protective system works well when oxidation is the only or primary type of attack. However, when the contaminants—vanadium and others—are present, they react in myriad ways in the 1000–2000°F (540–1100°C) temperature range to destroy the protective oxide and eventually the alloy. Coating is a method used to combat this problem.

Superalloy processes. Metallurgists have been seeking ways to increase the capability of superalloys through development of new processes.

Oxide dispersion strengthening. One technique to generate improved strength is known as oxide dispersion strengthening (**Fig. 3**). The objective is to distribute very fine, uniformly dispersed nonreactive oxide particles throughout the alloy. The processing usually involves starting with very fine particles of the metal itself, to which the oxide is added by a chemical or mechanical process step. The alloy is then consolidated by a mechanical pressing operation and forged into a final useful shape. Oxide dispersion strengthening materials are characterized by unusual creep resistance at very high temperatures; however, intermediate-temperature creep and all tensile properties are mediocre.

Success has been obtained in combining some of the classic strengthening factors described previously with oxide dispersion strengthening. This gives a balance of property enhancement which can be particularly useful to turbine metallurgists—good high-temperature creep strength from oxide dispersion strengthening and good intermediate-temperature strength from the other mechanisms. In this process, prealloyed superalloy powder (carbide- and γ'-strengthened) is hammer-milled in the presence of a very reactive metal, such as yttrium (Y), with oxygen (O) present. The result is a superalloy containing finely dispersed Y_2O_3 particles. *See* CREEP (MATERIALS).

Directional solidification. A technique adaptable to investment casting is that of directional solidification, which is particularly suitable to the complex shapes of airfoil parts used in gas turbines. It has been found

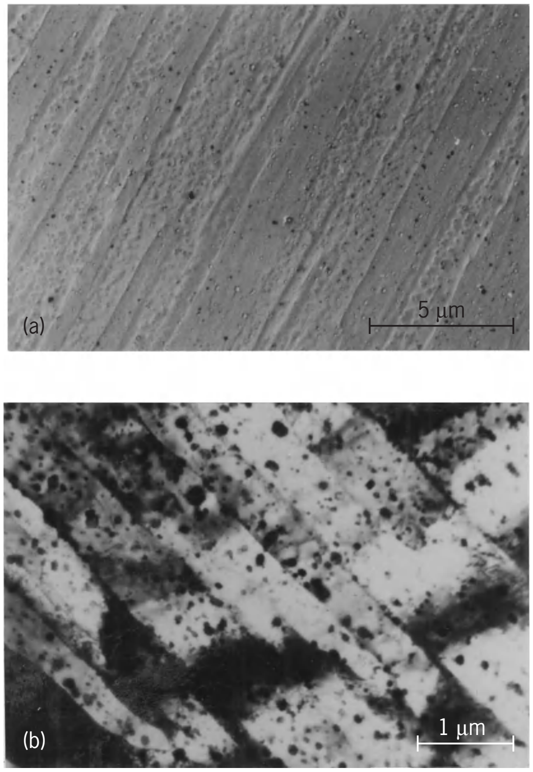


Fig. 3. Structure of thoria-strengthened nickel showing thoria particles. (*a*) Electron micrograph. (*b*) Transmission electron micrograph. (*From C. T. Sims and W. C. Hagel, eds., The Superalloys, Wiley-Interscience, 1973*)

that by commencing the freezing of molten superalloys (held in a ceramic mold of the shape desired) at the bottom, then allowing the freezing process slowly to proceed up through the shape to be cast, the grains of the structure acquire a long slender shape in the direction of freezing. This significantly increases the ability of the structure to withstand mechanical load in the freezing direction, and particularly increases resistance of the superalloy to a complex phenomenon involving stress and temperature cycling or temperature gradients, known as thermal fatigue. Thermal fatigue failures commence at transverse grain boundaries by this process.

Another approach to advanced airfoil materials is to eliminate grain boundaries entirely. In directional solidification, a number of grains nucleate in the first solid to freeze, and the few with the most preferred orientation for growth choke off solidification of the others and grow along the length of the airfoil. By reducing the cross section of the first solid to form, fewer grains are nucleated. If the ceramic mold also has a short length of spiral cavity before the airfoil shape, solidification of metal through this spiral "pigtail" will allow only a single grain to grow into the airfoil section. If a particular crystal orientation with respect to the airfoil is required, a seed crystal is used instead of a pigtail. The seed placed at the bottom of the mold is long enough that the very bottom portion is not melted in the directional solidification furnace. As the mold-containing molten alloy is withdrawn, solidification occurs on the seed, with essentially the same crystal orientation as the seed. The seed can be cut from the solidified airfoil and used to seed additional airfoils. These single crystals, also known as monocrystals, allow elimination of grain-boundary strengtheners such as zirconium, carbon, boron, or hafnium. This may allow more flexibility in the remainder of the composition, and higher heat treatment temperatures to optimize alloy mechanical properties and microstructure. Further, freezing occurs in the 001 crystallographic direction, preferred because it is in the low-elastic-modulus direction for the face-centered cubic alloys.

Powder metallurgy. An old process, powder metallurgy, has found increased use for high-temperature superalloys, and offers the possibility of added flexibility in alloy chemistry. Alloys originally developed for cast and wrought processing have been modified slightly for use as powder metallurgy materials. For large structures such as turbine disks, this has resulted in much more homogeneous chemistry and microstructure and achievement of outstanding thermal fatigue resistance. Metal flow resistance in isothermal forging is reduced because of nearly superplastic grain-boundary sliding. *See* POWDER METALLURGY; SUPERPLASTICITY.

Coatings. Superalloys and stainless steels must possess a balance of properties for high-temperature service. The most oxidation- and corrosion-resistant alloys do not have acceptable strength for most structural applications. Therefore, a viable solution is to utilize strong alloys for airfoil structures but then coat them to create environmental stability. This is done by adding elements such as chromium and aluminum into the surface of the alloy; these elements react with the aggressive environment to form highly protective oxides. The coating is applied by chemically reacting the turbine part with an atmosphere containing chromium or aluminum halides at very high temperature so that the active elements diffuse into the surface of the alloy to form the protective layer. Thus, ultimately, the coating layer is composed mainly of nickel from the alloy, and aluminum or chromium or both added in the coating process. Other methods develop overlayer coatings, which form not by reaction with the nickel substrate but by deposition from sources of the desired overlayer chemistry that are made by processes such as physical vapor deposition from an electron-beam–heated source, or by low-pressure plasma spraying of powders of desired coating compositions. Coatings formed by any of these methods are bonded to the substrate in the high-temperature heat treatment to cause the necessary interdiffusion. Such aluminum- and chromium-rich coatings can triple the life of industrial gas-turbine parts at temperatures such as 1600°F (870°C) in oxidizing atmospheres. *See* METAL COATINGS; SPUTTERING.

Eventually, however, the coatings fail by oxidizing away or by further interdiffusion with the superalloy underneath. If a very thin layer of platinum is used, the concentration of aluminum in the surface appears to be enhanced, creating an even greater measure of protection. A metallographic picture of a

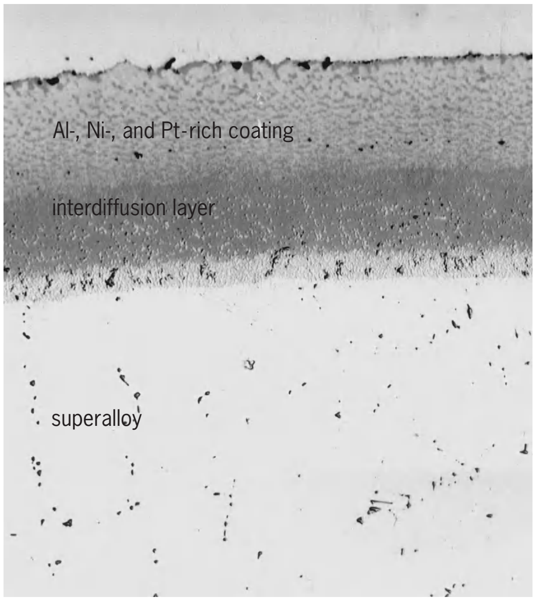


Fig. 4. Aluminide-platinide coating on a nickel-base superalloy. (*E. Buchanan, General Electric Co.*)

superalloy with a coating made by the electron-beam process, with aluminum increasing in concentration toward the surface, is shown in **Fig. 4**.

Nonmetallic materials. Considerable activity involving attempts to adapt ceramics to high-temperature applications has occurred since the mid-1960s. These are ceramics of the covalent-bonded type, such as silicon carbide and silicon nitride. Oxide ceramics, such as aluminum oxide (Al_2O_3) and zirconium oxide (ZrO_2), possess ionically bonded structures, and so tend not to possess usable high-temperature creep resistance. Turbine designers and materials engineers are struggling with the problems of utilizing the covalent ceramics, since they possess great strength. It has also become apparent that these ceramics possess great oxidation and corrosion resistance—features that would be useful in turbine equipment which must handle high-temperature products of combustion from coal. Data have shown that silicon carbide (SiC) and silicon nitride (Si_3N_4) are attacked to only 2 or 3 mils (0.05–0.08 mm) in depth after 6000–8000 h exposure in corrosive atmospheres; most high-temperature metals or alloys cannot meet this performance level. However, their complete lack of ductility means that new design techniques are required to prevent early and catastrophic failure, and only very small parts have been shown to be useful so far. It is not certain that such ceramics will be usable in heat engines.

To bypass the brittleness problem, the concept of ceramic composites has emerged. In these materials, a ceramic matrix is filled with ceramic fibers, which strengthen the whole body. When subjected to a high mechanical load, the fibers pull against the matrix and slip slightly, giving the material a certain level of tolerance. This is known as fiber pull-out. However, the effect is not elastic, so that it does not approach the deformation tolerance normally seen by metallic alloys that is essential for safe structures in tension.

The use of graphite and graphite/graphite compositions has also been explored. Graphite has a strikingly unique combination of very low density and high elastic modulus, and it demonstrates increasing mechanical strength with increasing temperature. It is capable of mechanical service at tempertures of 2200°C (4000°F) or higher. However, since graphite is a form of carbon, it has virtually no oxidation resistance. Therefore, attempts to utilize graphite must involve truly extraordinary success in protecting it from oxidation. So far, protective coatings have had but limited success, and it appears that use may well be limited to rather low temperatures. *See* CERAMICS; COMPOSITE MATERIAL; GRAPHITE; METAL, MECHANICAL PROPERTIES OF. Chester T. Sims

Bibliography. American Society for Metals, *Metals Handbook, Desk Edition*, 2d ed., 1999; American Society for Metals, *Sourcebook on Materials for Elevated-Temperature Applications,* ed. by E. F. Bradley, 1979; W. Betteridge and J. Heslop (eds.), *The Nimonic Alloys,* 2d ed., 1974; C. T. Sims, N. S. Stoloff, and W. C. Hagel (eds.), *Superalloys II,* 1987.

Highway bridge

A structure that crosses over a body of water, traffic, or other obstruction, permitting the smooth and safe passage of vehicles. In highway transportation systems, the term “bridge” is usually reserved for structures over bodies of water. However, many other structures are generally considered highway bridges. An overhead is a structure carrying a highway over a railroad, and an underpass is a structure providing passage of a highway under a railroad. An overcrossing is a structure carrying a county road or a city street over a state highway, and an undercrossing is a structure providing passage of a county road or a city street under a state highway. A separation is a structure separating into two state highways. A connector ramp is a structure connecting intersecting highways and roads. An interchange is the group of ramps and structures providing connections for traffic between intersecting highways (**Fig. 1**). A viaduct is an elevated structure carrying a highway over streets, railroads, or other features. *See* BRIDGE; HIGHWAY ENGINEERING.

Highway bridges can be made of steel (**Fig. 2**), concrete (**Fig. 3**), timber, stone, metal alloys, or advanced composite materials, and may have different structural systems such as girder (beam), truss, arch, cable-stayed, and suspension. *See* STRUCTURAL MATERIALS.

Design. Bridge design is a combination of art and science. Conceptual design is usually the first step. Before any theoretical analysis and detailed design proceeds, the designers visualize the bridge in order to determine its function and performance. The conceptual design process includes selection of bridge systems, materials, proportions, dimensions, foundations, esthetics, and consideration of the surrounding landscape and environment. A bridge may be straight or horizontally curved, or have skewed supports. The width of a highway bridge is determined by the number and width of the traffic lanes and the shoulder or sidewalk width, and is typically the same dimension as the approaching highway.

The selection of bridge type is influenced by many factors such as span length, site geology and foundation requirements, design loads, surrounding geographical features, width requirements, clearance requirement below the bridge, transportation of construction materials, erection procedures, and construction cost and duration. The **table** shows the span lengths appropriate for various bridge types. A bridge is required to fulfill its function as a thoroughfare while blending and harmonizing with its surroundings.

The final design process involves structural analysis, member and detail design, and preparation of construction drawings and specifications. Structural analysis commonly involves computer models, which use appropriate material properties, member discretization, boundary conditions, and loads. Members and connection joints are proportioned to carry all possible loads (permanent loads, vehicular live loads, wind loads, and earthquake loads), combined and factored in accordance with the requirements of applicable design standards and codes. Two standards are the American Association of State Highway and Transportation Officials Load and Resistance Factor Design (AASHTO-LRFD) Bridge Design Specifications (2004) and AASHTO Standard Specifications for Highway Bridges (2002) in the United States. **Figure 4** shows the AASHTO "design truck." The variable axle spacing between the 145-kilonewton loads is adjusted to create a critical condition for the design of each location in the structure. Well-prepared construction documents are biddable, constructable, and cost-effective. *See* STRUCTURAL ANALYSIS; STRUCTURAL DESIGN.

Construction. Bridge construction involves managing the fabrication and erection operations needed for safe and efficient building of the structures in accordance with the construction documents.

A fundamental part of construction engineering is project management (planning, scheduling, and controlling). Planning starts with analyses of the type and scope of work to be accomplished, as well as the identification of construction methods and equipment, and type and size of the work force. Scheduling establishes a sequence of operations and accounts for the interrelation of operations at the job site, as well as allocation of the work force and equipment. Controlling consists of procuring materials in accordance with construction staging, dictating overall construction procedures, engineering inspection, and maintaining complete records of daily operations, payments, and expenditures. Proper planning, scheduling, and controlling prevents costly delays and cost overruns. *See* CONSTRUCTION ENGINEERING.

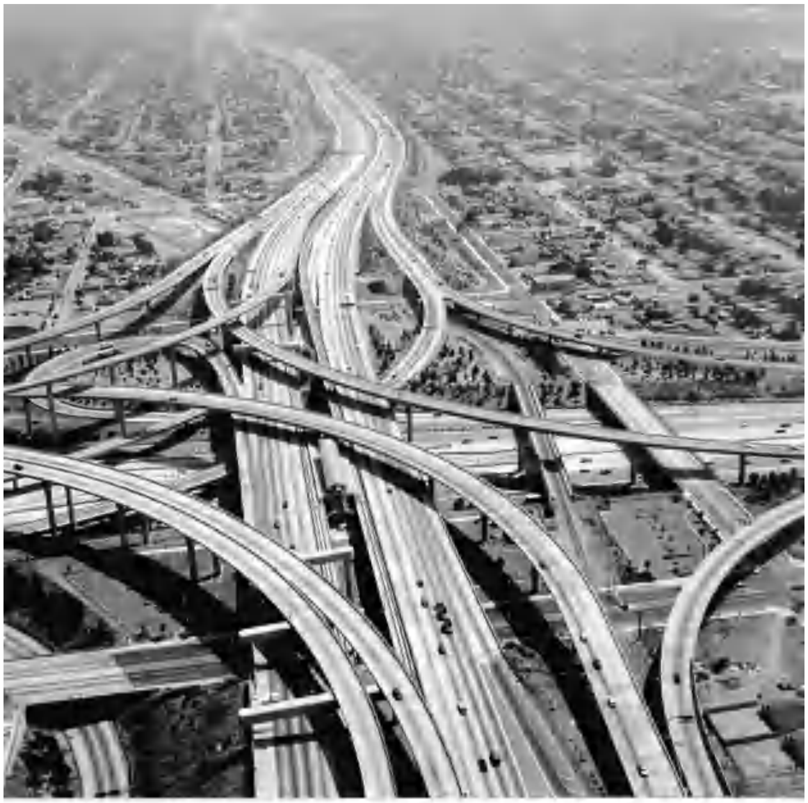


Fig. 1. Aerial view of I-105/605 Interchange in California. (*California Department of Transportation*)

Maintenance. The primary objective of bridge maintenance is to ensure public safety. Proper maintenance also reduces life cycle costs and earns public confidence. Before 1970, components of the United States' highway network exhibited rapid aging. Since then, bridge maintenance programs have evolved into a sophisticated bridge management system.

The Federal Highway Administration (FHWA) requires all state agencies to submit highway bridge

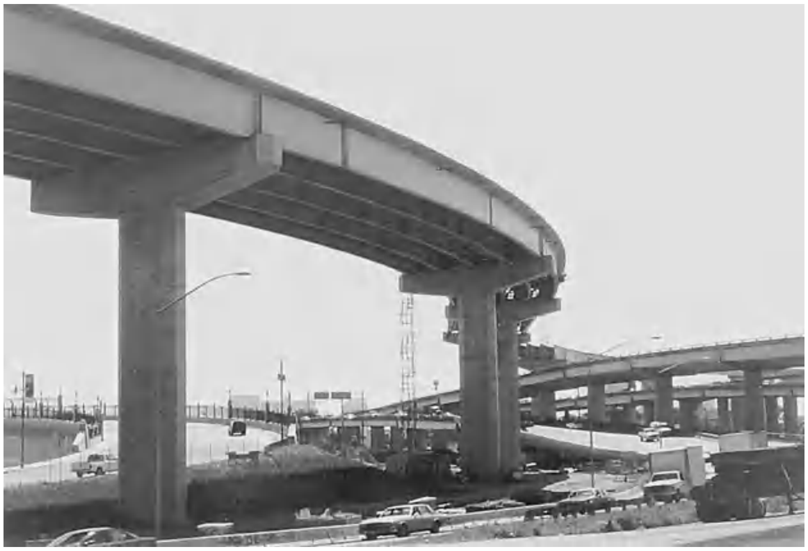


Fig. 2. Steel girder bridge.

Types of bridges and applicable span length

Bridge type	Span range	Leading bridge and span length
Prestressed concrete girder	10–300 m (33–984 ft)	Stolmasundet, Norway; 301 m (988 ft)
Steel I/box girder	15–376 m (49–1234 ft)	Sfalassa Bridge, Italy; 376 m (1234 ft)
Steel truss	40–550 m (131–1804 ft)	Quebec, Canada; 549 m (1801 ft)
Steel arch	50–550 m (164–1804 ft)	Shanghai Lupu, China; 550 m (1804 ft)
Concrete arch	40–425 m (131–1394 ft)	Wanxian, China; 425 m (1394 ft) [steel-tube-filled concrete]
Cable-stayed	110–1100 m (361–3610 ft)	Sutong, China; 1088 m (3570 ft)
Suspension	150–2000 m (492–656 ft)	Akaski-Kaikyo, Japan; 1991 m (6532 ft)

data on a Structure Inventory and Appraisal (SI&A) sheet. Bridge data include description of the structure, structural data and history, traffic information, load rating, condition and appraisal ratings, and inspection findings. The information on SI&A sheets is a valuable aid to establish maintenance and replacement priorities, as well as to project the maintenance cost of the highway bridges.

FHWA requires that each public bridge be inspected at a regular interval not exceeding 2 years. Underwater bridge components that cannot be visually evaluated during periods of low flow by tactile means are required to be inspected at an interval not exceeding 5 years. The purpose of the bridge inspection is to obtain the information necessary to properly evaluate the bridge capacity and the adequacy of the bridge.

The major tasks of the bridge inspection are to (1) identify minor problems that can be corrected before they develop into major repairs; (2) locate bridge components that require repairs in order to avoid replacement; (3) identify unsafe conditions, prepare accurate condition reports, and recommend corrective actions; (4) investigate more serious damage and its effect on the structure; and (5) generate accurate bridge records.

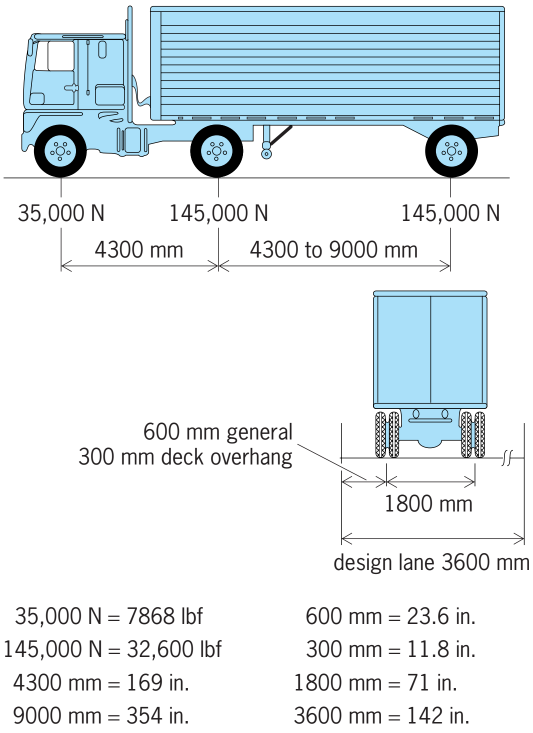


Fig. 4. AASHTO design truck (load).

Bridge ratings. The evaluation or rating of existing bridges is a continuous activity of the bridge owners to ensure the public safety. The evaluation provides criteria to repair and rehabilitate superstructures, to issue special permits, and to post, close, or replace the existing bridge.

It is generally believed that when a bridge is designed for the AASHTO design vehicles, and is constructed and maintained in accordance with the contact documents, the bridge will have adequate capacity to handle the actual present traffic. However, changes in details during construction, failure to meet minimum material strength requirements, unexpected settlements of foundation after construction, and unforeseen damage to a member could affect the capacity of the bridge. In addition, old bridges might be designed for a lighter vehicle than the present one, or a different design code.

Sometimes, an industry needs to transport their heavy machinery from one location to another location. These vehicles could weigh much more than the design vehicles, and thus the bridge owner may need to determine the current live-load-carrying

Fig. 3. Concrete girder bridge.

capacity, or rating, of the bridge to issue special permits. There are two levels of rating for bridges: inventory and operating. The inventory rating determines the load that can be safely carried by a bridge for an indefinite period. The operating rating reflects the absolute maximum permissible load that can be safely carried by the bridge.

When a bridge is found to have inadequate capacity for design vehicles, engineers look at several alternatives prior to closing the bridge. Possible alternatives include limiting allowable vehicle weight and speed, reducing vehicular traffic volume, restricting vehicles to certain lanes, and recommending repairs to alleviate the problem. In addition, when evaluations show the structure is marginally inadequate, frequent inspections may be recommended to monitor the physical condition of the bridge in the later stages of its life cycle. Lian Duan

Bibliography. AASHTO, *AASHTO LRFD Bridge Design Specifications*, 3d ed., American Association of State Highway and Transportation Officials, Washington, DC, 2004; AASHTO, *Standard Specifications for Highway Bridges*, 17 ed., American Association of State Highway and Transportation Officials, Washington, DC, 2002; R. M. Barker and J. A. Puckett, *Design of Highway Bridges*, Wiley, New York, 1997; W. F. Chen and L. Duan (eds.), *Bridge Engineering Handbook*, CRC Press, Boca Raton, FL, 2000.

Highway engineering

A branch of civil engineering that includes planning, design, construction, operation, and maintenance of roads, bridges, and related infrastructure to ensure effective and efficient movement of people, goods, and services that use this mode of transportation. Highway engineering has a profound impact on residential, commercial, recreational, and industrial locations and operations, and therefore has a far-reaching influence on the cultural and economic activities of a region. *See* CIVIL ENGINEERING.

Planning and project development. At the planning stage, the need for the highway development is established. Depending on the scale of the proposed development, highway planning may occur at any level of the administrative hierarchy such as the state, regional, or local levels, and public input and involvement is critical. The need for highway development may arise out of new construction, major structural repair or replacement, existing or projected future travel demand in excess of available capacity, and need for increased mobility through capacity enhancement along the planned or existing highway corridors. Other needs arise from safety concerns that can be addressed only through improvements in the geometric design of roads, such as curves and grades. The cornerstone of highway planning is the estimation of current and future traffic volumes on the road network and comparison of the expected volume and the capacity of each individual highway section to yield a measure of the expected level of performance or service at that section.

Traffic volume. This may be measured in terms of annual average daily traffic (AADT) and computed as the total yearly traffic count divided by the number of days in a year. AADTs are obtained by conducting 48-hour counts and adjusting these values by a factor derived from continuous counts. The 48-hour counts are traditionally done using mechanical devices such as pneumatic tubes laid across the road. Continuous counts are done using automatic traffic recorders permanently embedded in the pavement. Continuous counts provide factors by which values from 48-hour counts can be adjusted to reflect seasonal or temporal variations in traffic volumes.

For urban highway systems, traffic volumes are generally predicted using a four-step process. This begins with trip generation, where the study area is divided into zones and the number of trips generated by each zone is computed on the basis of socioeconomic characteristics of that zone. This is followed by a distribution of trips from and to each zone. The third step, modal split, predicts the proportion of trips that use the major available modes of transport. Finally, the trips are assigned on the various links of the network.

For purposes of design, traffic volumes are needed for a representative period of traffic flow. However, because traffic flow is not uniform, it is inappropriate to use traffic volumes per average hour or traffic volumes during the hour of highest flow, as these would result in underestimation and overestimation, respectively, of traffic flow. The actual traffic flow rate used for design is often a compromise value that falls within these two extremes. Traffic flow rate is the number of vehicles that pass a given spot in a specified time interval, such as 1 hour. In order to incorporate different vehicle classes in the traffic stream, vehicle flows are expressed in terms of passenger-car equivalents.

Capacity. This is the maximum theoretical traffic flow rate that a highway section can accommodate under a given set of environmental, highway, and traffic conditions. The capacity of a highway section is influenced by factors such as the number of lanes, lane width, presence of traffic control systems and their effectiveness, frequency and duration of traffic incidents, and efficiency of collection and dissemination of highway traveler information. To increase the capacity of an existing highway section, engineers traditionally widen existing lanes, provide additional lanes, or make horizontal curves and vertical grades smoother and gentler. Where physical, institutional, or environmental constraints preclude such geometric improvements, capacity-enhancing technologies that are collectively termed intelligent transportation systems (ITS) can be pursued. ITS increase highway capacity in a variety of ways, including enhanced traffic management, reduced probability of crashes, quick detection and clearance of incidents, and effective delivery of traveler information. *See* TRAFFIC-CONTROL SYSTEMS.

Level of service. Traffic conditions arising from the interplay of traffic volume and section capacity are perceived by road users in a way that is termed level of service (LOS). LOS ranges from an index of A (very good) to F (very poor). When traffic volume approaches highway capacity, vehicular flow becomes congested and speeds fall below the design speed. Under such conditions, queues are formed and operations within the queue are characterized by stop-and-go waves. Such a situation is indicative of a poor level of service (LOS F). When traffic volumes are much lower than highway capacity, vehicles travel at a freely flowing speed and are virtually unaffected by the presence of other vehicles in the traffic stream, indicating LOS A or B. In most cases, highway engineers design for an LOS of C. An important part of highway planning is to specify the desired minimum level of service.

Public participation. Public participation in highway project planning and development should not be limited to conventional public meetings where completed design alternatives are presented for public input, but should occur early in the project development process. A consensus is often reached among identified stakeholders regarding the existence of a problem, the need for the highway project, and to some extent, how it should be fixed.

Environmental considerations. Highway facilities often have adverse effects on the environment, such as noise pollution, air pollution, water pollution, and degradation of habitats and the flora and fauna they support. As such, the prediction and mitigation of imminent environmental impacts of a highway are a key aspect of transportation planning. Traffic noise arises from sources such as vehicle/air interaction, tire/pavement interaction, vehicle exhausts, and engines, and may be addressed at source, in the transmitting medium, or at the receptor. Engine insulation and mufflers are used to reduce engine and exhaust noise, respectively; noise absorbing pavement materials or longitudinal tining (texture) reduce tire/pavement noise; and traffic noise barriers protect highway neighbors from traffic noise. Highway air pollution is caused by the dispersion of gas or liquid droplets and solid particles emitted from vehicle tailpipes, such as carbon monoxide, nitrogen oxides, volatile hydrocarbons, sulfur oxides, and particulate matter. Pollutants are also released during refueling and from wear of tires and brake linings. Air pollution is generally associated with respiratory damage to humans and damage to structures and vegetation. In addition, highway system construction, operation, and maintenance degrade water quality by contributing a variety of pollutants to runoff, such as solids, heavy metals, oil and grease, and pesticides. In many countries, transportation-related legislation has been passed to protect the environment from adverse effects of highway usage. *See* AIR POLLUTION.

Geometric design. Highway engineering design involves selecting the appropriate location, alignment, and cross section of a planned highway by considering three major factors—human, vehicle, and roadway/environment—and how these factors interact to provide a safe highway. Human factors for design include reaction time for braking and steering, visual acuity for traffic signs and signals, and car-following behavior. Vehicle considerations include vehicle size and dynamics, which are essential for determining lane width and maximum slopes as well as for selecting the design vehicles, which range from passing cars to tractor trailers. The physical dimensions of vehicles affect several design elements, including road corner radii and clearance of highway overpasses. Highway engineers design road geometry to ensure that vehicles negotiating curves and grades are stable, and to provide adequate sight distances for vehicles that undertake passing maneuvers along curves on two-lane two-way roads.

Location planning. This involves fitting the road efficiently onto the surrounding terrain and environment. Given a set of specific terminal points, a number of locations are usually evaluated, and the best alternative is chosen according to criteria such as travel time, vehicle operation cost, section length, environmental impacts, total cost, ease of maintenance, accessibility/connectivity, and landscape aesthetics.

Horizontal alignment. This can be represented in an aerial view of the highway (**Fig. 1***a*). It consists of straight lines and curves. Curves are fitted to provide a smooth transition between straight highway sections. A transition curve may be used to connect the circular and straight sections. Elements of curve design include the selection of an appropriate radius and the degree of superelevation in the curve (raising one lateral end of the highway above the other). To ensure vehicle stability at high speeds, large curve radii and maximum superelevation are necessary. However, for curves at low-speed sections, such as local streets, flat curves of small radii may be acceptable. Intersection and interchange designs are part of horizontal alignment.

Intersections and interchanges. Where two or more highways cross each other at the same level, various vehicle maneuvers (turning, crossing, and through movements) occur within a limited area, and as the volumes of these movements increase there is increased likelihood of traffic conflicts and crashes. Channelization—the provision of islands or raised curbs—reduces such danger by limiting each stream to a unique path. In areas where traffic volumes are high, the movement of traffic streams can be separated in time using multiphase traffic signals that permit movement of only nonconflicting streams in each phase. Where highways cross each other at different levels, grade separation (overhead bridges with or without access) is used.

Vertical alignment. The vertical alignment of a highway is represented by its longitudinal profile, which illustrates the elevation of all points along the length of the highway. The purpose of vertical alignment design is to determine the level of the highway at each point in an overall bid to ensure adequate safety and drainage along the entire length of the section. Vertical alignment consists of grade tangents connected

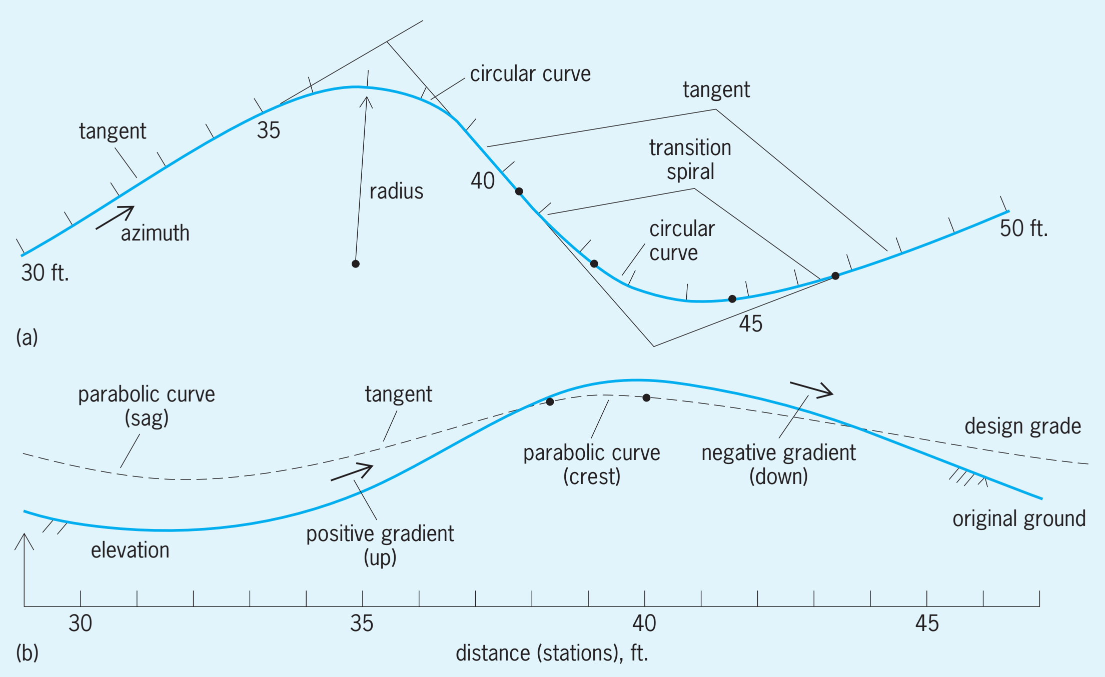


Fig. 1. Typical (*a*) plan view and (*b*) profile view of a highway section.

with parabolic curves. The desirable maximum slope of the tangent and rate of slope change depends on the road classification and characteristics of the design vehicle. The vertical alignment is generally kept as close as possible to the natural terrain to minimize earthwork volumes. However, challenging highway geometry, such as steep slopes, may be unavoidable at locations of difficult topography. Also, the combined effect of the longitudinal grade and the length over which it occurs is an important consideration. A vertical curve could be a crest or a sag curve (Fig. 1*b*). Construction of vertical curves is typically expensive because of the volume of earthwork involved. In the design of vertical alignment, it is desirable to balance cuts and fills, so that suitable material excavated in cuts can be used to fill low-lying sections. A challenge often encountered in highway engineering design is to minimize earthwork volumes while providing acceptable levels of safety. An alignment with acceptable level of safety is one that provides drivers with adequate sight distance to enable them to stop their vehicles in response to a perceived obstruction. Braking distance, the theoretical minimum distance that a vehicle travels while braking to a stop, depends on factors such as vehicle speed and friction between the vehicle tire and pavement. The distance traveled by the driver between the time of perception and braking is added to the braking distance to give the total required stopping sight distance. For sag curves, the critical concern is the headlight sight distance (length of road illuminated by vehicle headlights) for nighttime driving.

Highway cross section. This refers to the road profile perpendicular to the direction of travel and within the limits of the highway right-of-way. Highway cross section consists of driving lanes and may include other elements such as bicycle or pedestrian lanes, shoulders, medians, barriers, curbs, cross slope for drainage, and superelevation. **Figure 2** shows a typical cross section for a rural highway. The elements

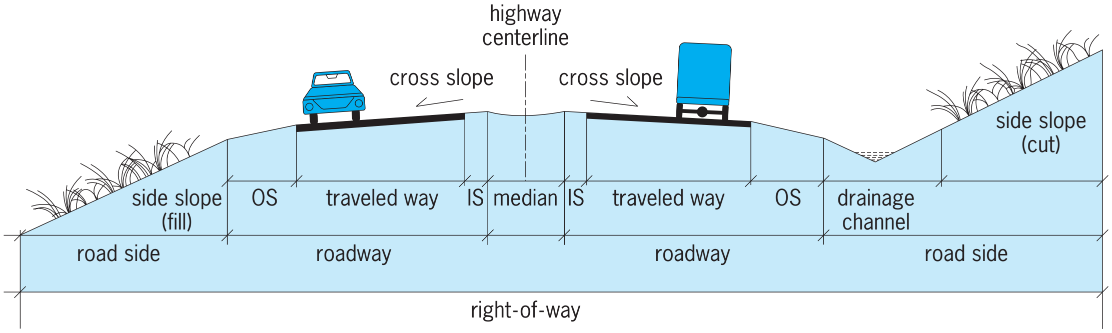


Fig. 2. Typical highway cross section. IS, inside shoulder; OS, outside shoulder.

and dimensions of highway cross section at a given location depend on road classification and land use (urban or rural).

Pavement design. This is the process of selecting appropriate types and thicknesses of pavement layers so that the pavement structure can withstand expected loads in a cost-effective manner. The original ground, or subgrade, typically lacks adequate bearing capacity and may need replacement or overlay using more competent material to reduce the stresses imposed by the vehicle loads to tolerable levels. This way, the natural ground can withstand loads without deforming.

Pavement materials. Pavement materials include surface and subsurface layers that consist mainly of mineral aggregate. Aggregate may be unbound (in the case of gravel roads) or bound using cementitious material such as ordinary portland cement (for rigid pavements) or asphaltic cement (for flexible pavements). Pavement materials have to meet specified standards before they are used in construction. A wide range of physical and chemical tests are carried out on ingredients as well as pavement mixes to ensure that they are sufficiently strong and durable.

Ingredients. Commonly used mineral aggregates include crushed rock and natural gravel and sand. An aggregate mixture whose particles are of a wide range of sizes, with the smaller particles filling the voids between the bigger ones, is described as a well-graded mix. A well-graded aggregate mix provides the maximum degree of aggregate interlock and generally yields competent material. In some cases, physical or chemical stabilizers such as lime are added to improve the properties of available material. In recent years, efforts have been made to recycle waste materials as pavement materials, including crushed glass, ground-up tires, and recycled bituminous mixes scraped from old pavements.

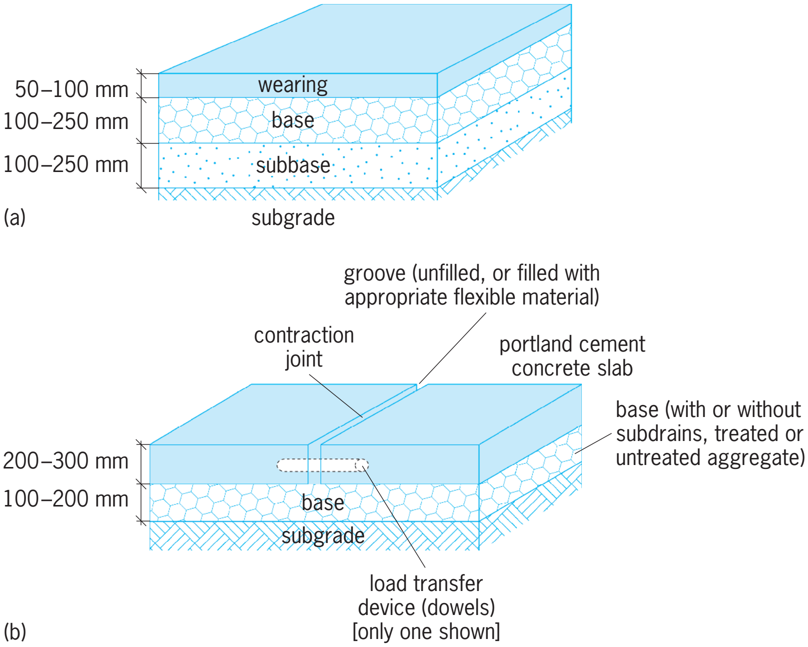

50–100 mm
100–250 mm
100–250 mm
wearing
base
subbase
subgrade
(a)
groove (unfilled, or filled with appropriate flexible material)
contraction joint
portland cement concrete slab
base (with or without subdrains, treated or untreated aggregate)
200–300 mm
100–200 mm
base
subgrade
load transfer device (dowels) [only one shown]
(b)

Fig. 3. Cross sections of (*a*) typical flexible pavement and (*b*) typical rigid pavement. Some designs include a subbase layer, between the base and subgrade.

Mixes. For rigid pavements, ordinary portland cement is mixed with water and aggregates to produce a viscous concrete mix that is poured into prepared forms and vibrated. The mix hardens, or cures, as the cement hydrates, and is usually watered to enhance the curing process. Special additives are usually added to the concrete mix to retard or to accelerate the curing process or to improve performance. For flexible pavements, asphaltic cement (bitumen) is liquefied by heating, solvent addition, or emulsification, and is mixed with mineral aggregates. The asphaltic concrete mix is then laid on the prepared base course in compacted layers. The laid material is cured by cooling or by evaporation of the solvent/emulsifying agent. *See* CEMENT.

Pavement types. There are generally three types of pavements: gravel, flexible, and rigid. The choice of any type depends on a variety of factors, including traffic, construction and maintenance costs, and availability of materials, labor, equipment, or funding. *See* PAVEMENT.

1. *Gravel pavement.* This is the simplest type of pavement and is often designed for lightly traveled roads. Material used for such pavements often consists of well-graded, naturally occurring silty or clayey gravel. This material is spread, watered, and compacted in layers to provide a crowned shape that is good for drainage purposes. Increased traffic may result in accelerated pavement deterioration and may require more frequent maintenance. *See* GRAVEL.

2. *Flexible pavement.* This is a multilayered structure that includes a subbase, a base, and an asphaltic wearing course (**Fig.** 3*a*). It is described as flexible because the main structural layer is based on asphalt, a viscous material that causes inability of asphaltic concrete layers to span a gap without sagging into it. *See* ASPHALT AND ASPHALTITE.

3. *Rigid pavement.* This pavement type consists of a plain or steel-reinforced portland cement concrete slab laid on a prepared surface such as crushed-stone base course (Fig. 3*b*). It is described as rigid because the slab is competent enough to bridge any minor gaps in the underlying layers without undue failure. *See* REINFORCED CONCRETE.

Highway construction. Highway construction usually follows planning and design, and involves the implementing facility designs to yield new or reconstructed facilities such as pavements, bridges, drainage structures, and traffic control devices. Road construction is often preceded by detailed stakeout surveys and preparation of the subgrade. *See* CONSTRUCTION ENGINEERING; CONSTRUCTION EQUIPMENT.

Surveys and subgrade preparation. The purpose of stakeout surveys is to lay out the designed position of all the road elements on the field as a guide for the contractor during construction. Stakeout surveys are carried out continuously during the construction period to monitor the work. After completion of the surveys, clearing and topsoil removal may be necessary where the road alignment passes through vegetated areas. Where the alignment passes through built-up areas, clearing includes demolition and relocation of existing structures and utilities within the road right-of-way. If the formation level (the level of foundation on which pavement will sit) is below the existing ground level, excavation is needed to lower the existing ground, thus forming a cut section. If the formation level is above existing ground, filling is carried out to raise the road and an embankment is subsequently formed. Embankments are constructed one layer at a time, and each layer is spread uniformly and compacted using an appropriate combination of rolling, watering, vibrating, and kneading, depending on the type of fill material.

Pavement. When the subgrade elevation (formation level) is reached either in a cut or in a fill section, it is compacted thoroughly to cross-fall shape, and the subbase and base courses are laid. These courses may consist of natural gravel, crushed rock, or stabilized material, depending on pavement design specifications. The type of wearing surface ranges from gravel, sand-seal, or chip-seal for low-volume roads, to asphaltic or portland cement concrete for high-volume roads.

Traffic operations and control. Traffic signals are important traffic control devices for smooth operation of highways. The typical traffic signal at an intersection displays a sequence of green, amber, and red. One complete signal sequence is called a cycle. Traffic signals are either pretimed or demand-actuated. Pretimed signals repeat a predetermined unchanging cycle. For demand-actuated signals, the lengths of display time for each color vary depending on the presence of vehicles or pedestrians. Flow-concentration controllers are capable of sensing detailed demand information and respond appropriately by revising the signal cycle length and phasing patterns. Advanced traffic control systems usually employ a central computer to monitor and control traffic flows on large highway networks. Special types of demand-actuated signals are able to recognize and give priority to particular types of vehicles such as buses, light-rail vehicles, and emergency vehicles.

Highway infrastructure management. The practice of highway engineering has evolved to include the effective and efficient management of existing infrastructure (such as pavement, bridges, and traffic equipment) and operations (such as safety and congestion) throughout the life cycle of the asset. In all functional areas of highway infrastructure management, the key is to arrive at optimal facility-level and network-level scoping and timing decisions so that the highest levels of service can be accrued at least cost to the agency and user over the life cycle. The levels of service are typically expressed in terms of a variety of performance measures that include facility condition, safety, and protection from accidental or natural disasters.

Pavement management. Pavement condition is monitored over a period of time using a condition index or serviceability rating. Most highway agencies have implemented pavement management systems for cost-effective rehabilitation and maintenance decision-making. The type of intervention for a particular section of highway depends on the severity and frequency of pavement distress as well as the road class. Severe and widespread occurrences of potholes on high-volume roads, for instance, may warrant extensive rehabilitation in the form of newly added pavement layers, with or without removal of the existing layers. However, scattered and light occurrences of cracking on a local road may require only routine maintenance in the form of crack sealing. For flexible pavements, maintenance is often needed to address common surface problems, including potholes, cracks, and ruts. Potholes and cracking are caused by a variety of factors, including the freezing/thawing of moisture beneath the pavement and traffic loads, and are exacerbated by lack of adequate lateral restraint to the pavement layers. Rutting in the wheel tracks may be due to excessive wheel loads, poor asphaltic concrete mixes, poor subgrade material, or subgrades weakened by the ingress of moisture. For rigid pavements, maintenance addresses problems such as flexural cracking caused by stresses induced by loss of subgrade or base support. Other problems on rigid pavements include spalling or faulting at the joints, surface deterioration, and corrosion of steel reinforcement, which is accelerated by the deicing salts used during winter maintenance.

Bridge management. Highway agencies responsible for bridge maintenance and repair are faced with the challenge of monitoring the condition of bridges and deciding what types of repairs need to be applied and when. Through the development and implementation of bridge management systems, many agencies establish decision support tools that analyze and summarize data, use mathematical models to make predictions of bridge conditions, and facilitate rational evaluation of alternative policies and programs, preferably on the basis of multiple performance measures.

Congestion management. In highway congestion management, a system-wide plan is developed for implementing measures to mitigate the magnitude and duration of traffic congestion within a given analysis period and with a given budget. The cost and effectiveness (benefits) of each mitigation measure are key inputs to the management process. Mitigation measures include a variety of approaches involving travel demand management as well as technological innovations such as intelligent transportation systems. The purpose of these approaches is to reduce demand, encourage off-peak travel, increase capacity, and enhance traffic flow. Provisions may include improved signalization, use of frontage roads, arterial access control measures, provision of special lanes for high-occupancy vehicles, carpooling, improved public transportation services, electronic toll collection, and incident response services.

Safety management. This is a systematic process to ensure that highway safety resources are appropriately allocated to reduce the frequency and severity of all types of highway crashes. Safety management involves identifying, considering, implementing as appropriate, and evaluating all safety-enhancing opportunities in all phases of highway development. As such, highway engineers seek to minimize the frequency and severity of crashes by improving highway planning, design, construction, maintenance, and operation. Highway safety management generally consists of several components, including analyzing and predicting crash trends, identifying hazardous links and intersections in highway networks, identifying crash causes, and evaluating effectiveness and cost-effectiveness of various countermeasures. Specific highway-safety countermeasures include the provision of traffic signalization, channelization, luminous median markers, road signs, and crash barriers. Well-maintained pavements with adequate surface friction can also contribute to improved highway safety.

Kumares C. Sinha; Samuel Labi

Bibliography. N. J. Garber and L. A. Hoel, *Traffic and Highway Engineering*, 3d ed., Thomson-Engineering, Toronto, 2001; F. L. Mannering, W. P. Kilareski, and S. S. Washburn, *Principles of Highway Engineering and Traffic Analysis*, 3d ed., Wiley, New York, 2004; C. H. Oglesby and R. G. Hicks, *Highway Engineering*, 4th ed., Wiley, New York, 1982; *A Policy on Geometric Design of Highways and Streets*, 5th ed., American Association of State Highway and Transportation Officials, Washington, DC, 2004.

Hilbert space

An abstract notion of great power and beauty, which has been central to the development of mathematical analysis and forms the backdrop for many applications of analysis to science and engineering. Its essence lies in the fact that the objects of primary interest in analysis (namely, functions) enjoy geometrical properties which are in important ways analogous to the geometry of physical space. Thus the highly developed human visual and spatial intuition can lead to significant truths about functions.

The prototype of Hilbert space is familiar three-dimensional euclidean space, but the general notion requires "space" to be understood in an abstract way. With the development of the axiomatic method, mathematicians have come to use "space" for any collection of mathematical entities possessing in their ensemble some structure analogous to physical space. The individual entities are then called points of this space, even though they may be elaborate constructs such as functions or infinite sequences. This device of focusing on the relationship between mathematical objects and temporarily ignoring their natures is one of the most fruitful constituents of modern analysis, allowing a dramatic extension of the range of application of geometrical ideas. *See* EUCLIDEAN GEOMETRY.

Generalizations of euclidean space. Vectors (or directed line segments) in euclidean space have a rich structure. It makes sense to multiply them by real numbers or to add two of them together. It is possible to speak of their lengths (or magnitudes) and the angles between them. Vectors may be perpendicular, and Pythagoras' theorem holds. *See* LINEAR ALGEBRA.

Suitable classes of functions have some, but not usually all, of these properties. Functions can be added, and there are various ways (appropriate in different contexts) of defining the magnitude of a function and even the angle between two functions. If certain of the desirable properties of euclidean space are isolated and adopted as postulates, a class of spaces is defined that may include spaces of functions that are of concern in analysis. During the early decades of the twentieth century it was gradually realized that it is possible to select a small number of properties in such a way that the resulting spaces possess virtually all the desirable features of euclidean space except those which are closely linked to finite dimensionality. Furthermore, many examples of such spaces arise naturally in analysis, and many classical results, processes, and problems can be formulated and analyzed within the framework of these spaces. They are named after David Hilbert, who took a decisive step toward their introduction in 1906, when he proved what is known as the spectral theorem. From a contemporary viewpoint his discovery is best described as a decoupling result for certain transformations of Hilbert space, but Hilbert himself did not use the term Hilbert space, and the geometric interpretation was developed by other mathematicians, notably E. Schmidt, F. Riesz, and J. von Neumann.

Role of Hilbert space. Hilbert's innovation occurred while he was investigating integral equations which arose from mathematical physics. These equations are continuous analogs of the systems of simultaneous linear equations which are encountered in elementary algebra, but the unknown entity, instead of being a finite set of numbers, is a function. Typical examples of functions are physical quantities which depend on position and time, such as temperature and pressure. Physical laws impose constraints on these functions and so give rise to equations. Just as graphical methods give insight into the solution of a pair of simultaneous linear equations in two unknowns, so the development of Hilbert-space geometry had great consequences for the understanding of integral equations. The geometry is infinite-dimensional in general; that is, elements of the space cannot be specified by a finite set of coordinates. In consequence, ordinary geometric imagination is a shaky guide, and it must be complemented by rigorous definition and proof. By the 1940s the appropriate concepts and basic theorems were well established, and it was clear that Hilbert space is a truly fundamental mathematical structure which appears in widely disparate branches of pure and applied mathematics. A striking instance is quantum mechanics, where observable quantities are modeled by linear transformations of Hilbert space. *See* INTEGRAL EQUATION; LINEAR SYSTEMS OF EQUATIONS;

NONRELATIVISTIC QUANTUM THEORY; OPERATOR THEORY.

Definition. The definition of Hilbert space requires use of the concept of a vector space over the real or complex numbers. This is any collection of objects (to be called vectors) for which are defined two operations satisfying certain natural laws. The operations are to add any pair of vectors and to multiply any vector by a real or complex number. The laws include commutative, associative, and distributive laws. Vector spaces are important in their own right, but they are too general to be of much use in analysis, and it is necessary to postulate further structure having to do with lengths and angles.

A Hilbert space is defined to be a vector space H over the real or complex numbers that possesses an inner product and is complete with respect to this inner product. Here an inner product means a scalar (that is, real or complex number) valued function (x, y) of the pair of vectors x, y in H which satisfies conditions (1), (2), and (3).

$$(x, y) \text{ is linear in } x \tag{1}$$

$$(y, x) = \overline{(x, y)} \tag{2}$$

$$(x, x) > 0 \text{ unless } x = 0 \tag{3}$$

The overbar in Eq. (2) denotes complex conjugation; it can be omitted for real Hilbert space. Completeness is a technical condition which is automatically satisfied in finite-dimensional spaces but is needed in infinite-dimensional ones to ensure the validity of limiting procedures. The inner product corresponds roughly to the orthogonal projection of one vector on another. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

N. J. Young

Bibliography. S. K. Berberian, *Introduction to Hilbert Space*, 2d ed., 1999; J. Dieudonné, *A History of Functional Analysis*, 1983; P. Halmos, *A Hilbert Space Problem Book*, rev. ed., 1994; F. Riesz and B. Szökefalvi-Nagy, *Functional Analysis*, 1955, reprint 1990; W. Rudin, *Functional Analysis*, 2d ed., 1991; N. J. Young, *An Introduction to Hilbert Space*, 1988.

Hill and mountain terrain

Land surfaces characterized by roughness and strong relief. The distinction between hills and mountains is usually one of relative size or height, but the terms are loosely and inconsistently used. Because of the prevalence of steep slopes, hill and mountain lands offer many difficulties to human occupancy. Cultivable land is scarce and patchy, and transportation routes are often difficult to construct and maintain. However, many rough lands, especially those readily accessible to centers of population, attract tourists because of their scenic quality and the opportunities they may afford for outdoor recreation.

High mountain ranges set up major disturbances in the broad pattern of atmospheric circulation, and thus affect climates over extensive areas. More locally, by inducing turbulence or forcing moist air to rise in crossing them, rough lands commonly induce condensation and precipitation that makes them moister than surrounding lowlands. Within the rough country, local differences in elevation and in exposure to sun and wind produce complex patterns of local climatic contrasts. These variations, in turn, are often accompanied by unusual local variety in vegetation and animal life.

Development of rough terrain. Uplift of the Earth's crust is necessary to give mountain and hill lands their distinctive elevation and relief, but most of their characteristic features—peaks, ridges, valleys, and so on—have been carved out of the uplifted masses by streams and glaciers. The upraised portions of the crust may have been formed as broad, warped swells, smaller arched folds or domes, upthrust or tilted blocks, or folded and broken masses of extreme complexity. Some limited areas owe their elevation and certain of their local features to the outpouring of thick sheets of lava or the construction of volcanic cones. Hill lands, with their lesser relief, indicate only lesser uplift, not a fundamentally different course of development.

In some rough lands, for example the Appalachians or the Scottish Highlands, the most intense crustal deformation is known to have occurred hundreds of millions of years ago, while in others, such as the Alps, the Himalayas, or the California Coast Ranges, intense deformation has occurred quite recently. Since mountains can be erosionally destroyed in relatively short spans of geologic time, however, the very existence of mountains and hills indicates that structural deformation, at least simple uplift, has continued in those areas until recent times. This conclusion is reinforced by the fact that the major cordilleran belts are currently foci of earthquake activity and volcanism.

Distribution pattern of rough land. Hill and mountain terrain occupies about 36% of the Earth's land area. The greater portion of that amount is concentrated in the great cordilleran belts that surround the Pacific Ocean, the Indian Ocean, and the Mediterranean Sea. Additional rough terrain, generally low mountains and hills, occurs outside the cordilleran systems in eastern North and South America, northwestern Europe, Africa, and western Australia. Eurasia is the roughest continent, more than half of its total area and most of its eastern portion being hilly or mountainous. Africa and Australia lack true cordilleran belts; their rough lands occur in patches and interrupted bands that rarely show marked complexity of geologic structure. *See* CORDILLERAN BELT.

Relationship to plate tectonics. The broad-scale pattern of crustal disturbance, and hence of rough lands, is unquestionably but complexly related to plate tectonics, that is, the relative movements of a worldwide set of immense crustal plates. Strong crustal deformation tends to be concentrated at or near plate margins, where adjacent plates are being jammed together, rifted apart, or moved laterally along a common border. Comparison of a world map of plate convergence zones with maps of earthquake frequency and high mountains indicates that plate convergence zones are generally both earthquake-prone and mountainous, suggesting probable causal relationships. However, large areas of hill and mountain lands lie far from plate boundaries and hence cannot be explained by current plate convergence alone.

Where crustal plates converge, dense, flexible oceanic crust turns downward into the underlying mantle and usually undergoes partial melting at depth, often giving rise to volcanic activity above the sinking slab. However, low-density, more rigid continental crust resists sinking and tends to become intensely folded, thrust-faulted, and thickened in the convergence (sometimes in these circumstances called the collision) zone. The extensive cordilleran systems of southern Eurasia have been carved from complex structures formed by collision of the Eurasian plate with the African-Arabian and Australian plates, the crust involved being largely continental. In contrast, the island arcs of the Aleutians, the Kuriles, the Marianas, the eastern West Indies, and southern Indonesia are largely volcanic, having developed where convergence has involved only oceanic crust. Japan, the Andes, and the Cascade Range of the northwestern United States represent zones of convergence where oceanic crust from one plate has underrun continental crust from the other, producing structures combining complex deformation with prominent volcanic features. *See* EARTH CRUST.

At present, most plate divergence or rifting is associated with the extensive system of mid-oceanic ridges and is of little direct concern to continental terrain character. In a few areas, however, notably in the Red Sea–Gulf of Aden–East Africa rift zone, rifting extends into the continental blocks. Here the characteristic downfaulted troughs are in part occupied by arms of the sea or by lakes and are flanked by dissected bands of high ground, in places marked by volcanic flows and cones. A less extensive example is the Gulf of California trough and adjacent highlands in northwestern Mexico. *See* MID-OCEANIC RIDGE.

The many areas of hill or mountain terrain far from current plate margins are less obviously related to movements of the existing plate system. In some of these areas the geologic structures are similar to those now found at plate margins, but are ancient and have clearly been produced by convergence in a much older and very different plate system. The Appalachian–New England–Laurentian highland of North America and its structural continuations in northern Britain and Scandinavia are a good example. The Urals, eastern Australia, many of the mountainous areas of northern and eastern China, and parts of the uplands of eastern Brazil owe their existence to some form of late reuplift of the ancient structures. Other examples, notably the Rocky Mountains of the western interior of the United States and southern Canada, are even more difficult to explain by plate tectonics, for although their structures date from times since the present plate system came into being, they have developed far from plate margins. The broad, gentle upwarps of ancient shields or sedimentary platforms in plate interiors, such as the Ozark, Lake Superior, and Great Plains uplands of the United States, although simpler in form, are similarly perplexing.

Among the factors suggested to account for such intraplate structures and uplifts, sometimes acting in combination, are (1) isostatic adjustment resulting from continued erosional stripping of the thick, low-density crust of ancient structures; (2) unusually strong compressional and torsional stresses transmitted into the interior from the plate margins; and (3) movement of the crustal plates across areas of upwelling currents (plumes or hot spots) in the underlying mantle, with accompanying uplift, more or less volcanic activity, and differential movements of small sections of the crust. *See* EARTH, CONVECTION IN; EARTH INTERIOR; HOT SPOTS (GEOLOGY); PLATE TECTONICS.

Predominant surface character. The feature of hill and mountain lands are chiefly valleys and divides produced by sculpturing agents, especially running water and glacier ice. Local peculiarities in the form and pattern of these features reflect the arrangement and character of the rock materials within the upraised crustal mass that is being dissected.

Stream-eroded features. The principal features of most hill and mountain landscapes have been formed by stream erosion together with landslides and slower forms of gravity-induced movements (mass wasting) on the valley sides. The major differences between one rough land and another are in the size, shape, spacing and pattern of the stream valleys and the divides between them. These differences reflect variations in the original form and structure of the uplifted mass and in the ways in which erosion has been produced. *See* LANDSLIDE; MASS WASTING.

In consequence of uplift, streams have a large range of elevation through which they can cut, and as a result usually possess steep gradients early in their course of development. At this stage they are swift, have great erosive power, and are marked by many rapids and falls. Because of the rapid downcutting, the valleys are characteristically canyonlike with steep walls and narrow floors. At the same early stage, however, divides are likely to be high and continuous, and broad ridge crests common. In hill lands such ridges often provide easier routes of travel than do the valleys. In the Ozark Hills of Missouri, most of the highways and railroads follow such broad ridge crests. In high mountain country the combination of gorgelike valleys and continuous high divides makes crossing unusually difficult. In the Himalayas and the central Andes, and to a lesser degree in the Rocky Mountains of Colorado, the narrow canyons are very difficult of passage, and the divides are so continuously high and steep as to provide no easy pass routes across the ranges.

As erosional development continues, the major streams achieve gentle gradients and their valleys continue to widen. Divides become narrower and deep notches develop at valley heads, with well-defined peaks remaining between them. At this stage, which is represented by much of the Alps and by the Cascade Mountains of Washington, the principal valleys and the relatively low passes at their heads furnish feasible routes across the mountain belts. If other conditions are favorable, the wide valleys may afford significant amounts of cultivable or grazing land, as is true in the Alps. Still further erosion continues the widening of valleys and reduction of divides until the landscape becomes an erosional plain upon which stand only small ranges and groups of mountains or hills. The mountains of New England and many of the mountains and hill groups of the Sahara and of western Africa represent late stages in the erosional sequence.

Glaciated rough terrain. Glacial features of mountain and hill lands may be produced either by the work of local glaciers in the mountain valleys or by overriding glaciers of continental size. Continental glaciation has the general effect of clearing away crags, smoothing summits and spurs, and depositing debris in the valleys. The resulting terrain is less angular than is usual for stream-eroded hills, and the characteristic glacial trademark is the numerous lakes, most of them debris-dammed, a few occupying shallow, eroded basins. Examples of rough lands overridden by ice sheets are the mountains of New England, the Adirondacks, the hills of western New York, the Laurentian Upland of eastern Canada, and the Scandinavian Peninsula.

In contrast, mountains affected by local glaciers are made rougher by the glacial action. The long tongues of ice that move slowly down the valleys are excellent transporters of debris and are able to erode actively on shattered or weathered rock material. Valleys formerly occupied by glaciers are characteristically steep-walled and relatively free of projecting spurs and crags, with numerous broad cliffs, knobs, and shoulders of scoured bedrock. At their heads they generally end in cliff-walled amphitheaters called cirques (**Fig. 1**). The valley bottoms are commonly steplike, with stretches of gentle gradient alternating with abrupt rock-faced risers. Especially in the lower parts of the glaciated sections are abundant deposits of rocky debris dropped by the ice. Sometimes these form well-defined ridges (moraines) that run lengthwise along the valley sides or swing in arcs across the valley floor. Lakes are strung along the streams like beads, most of them dammed by moraines but some occupying shallow eroded basins. *See* MORAINE.

Because of rapid erosion and steepening of valley walls, either that effected directly by the ice or that attendant upon the exposure of the rock surface by continual removal of the products of weathering, glaciated mountains are likely to be unusually rugged and spectacular. This is true not only of such great systems as the Himalayas, the Alps, the Alaskan Range, and the high Andes, but also of such lesser ranges as those of Labrador, the English Lake District, and the Scottish Highlands. *See* EROSION; WEATHERING PROCESSES.

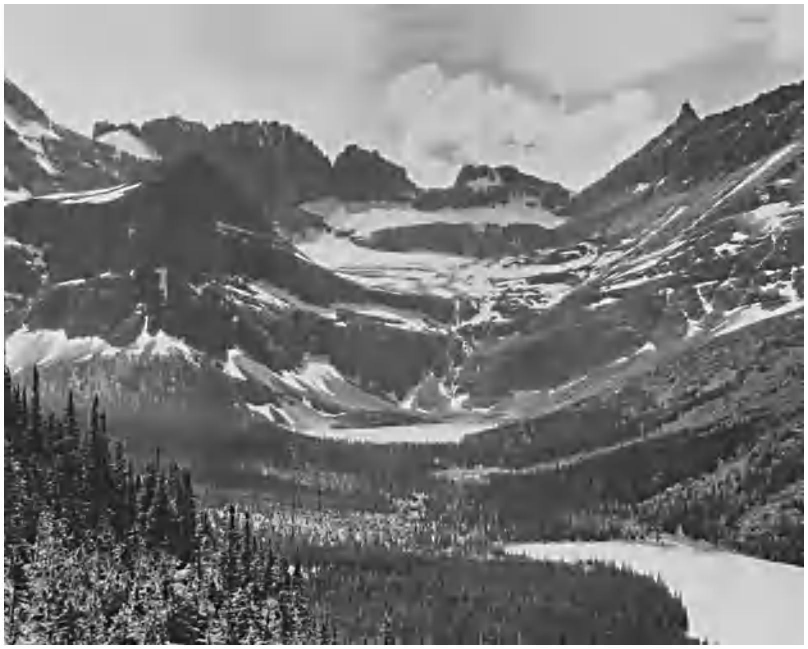


Fig. 1. Head of a glaciated mountain valley ending in a cliff-walled amphitheater, called a cirque, showing a steplike valley bottom with lakes and waterfalls. (*Photograph by Hileman, Glacier National Park*)

Most of the higher mountains of the world still bear valley glaciers, though these are not as large or as widespread as formerly (**Fig. 2**). Dryness and long, warm summers limit glaciers in the United States to a few large groups on the

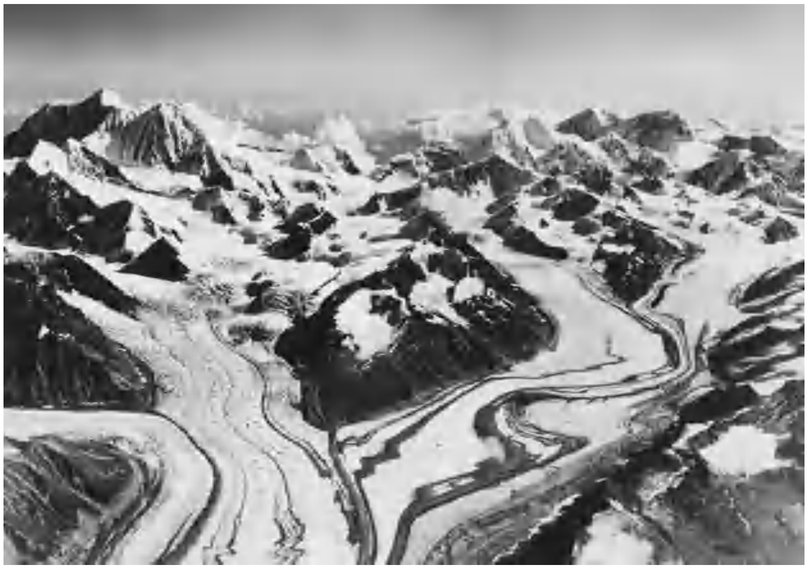


Fig. 2. Upper reaches of Susitna Glacier, Alaska, showing cirques and snowfields. The long tongues of glacier ice carry bands of debris scoured from valley walls. (*Photograph by Bradford Washburn*)

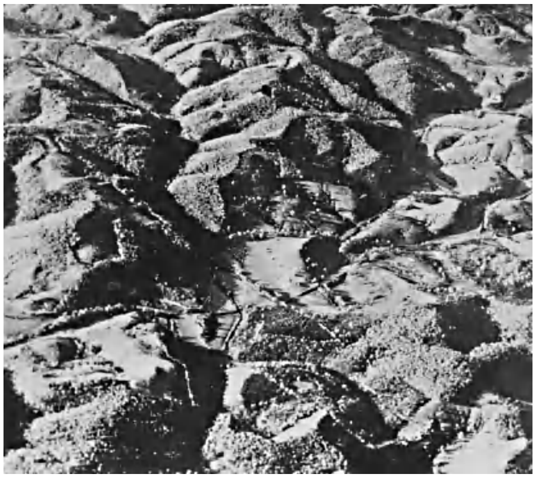


Fig. 3. Hills in the western Appalachians, West Virginia. (*Photograph by John L. Rich, Geographical Review*)

higher peaks of the Pacific Northwest and numerous small ones in the northern Rockies. *See* GLACIATED TERRAIN.

Effects of geologic structure. Form and extent of the elevated areas, pattern of erosional valleys and divides, and, to some extent, sculptured details of slope and crest reflect geologic structure.

Some areas, such as the Ozarks, the western Appalachians, and the coast ranges of Oregon, are simply upwarped areas of roughly homogeneous rocks that have been carved by irregularly branching streams into extensive groups of hills or mountains (**Fig. 3**). Others, like the Black Hills or the ranges of the Wyoming Rockies, are domes or arched folds, deeply eroded to reveal ancient granitic rocks in their cores and upturned younger stratified rocks around their edges. The Sierra Nevada of California is a massive block of the crust that has been uplifted and tilted toward the west so that it now displays a high abrupt eastern face and a long canyon-grooved western slope. The central belt of the Appalachians displays long parallel ridges and valleys that have been hewn by erosion out of a very old structure of parallel wrinkles in the crust. The upturned edges of resistant strata form the ridges; the weaker rocks between have been etched out to form the valleys. The Alps and the Himalayas are eroded from folded and broken structures of incredible complexity involving almost all varieties of rock materials.

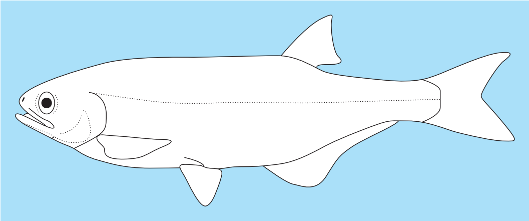
Most volcanic mountains, like the Cascades of the northwestern United States or the western Andes of Peru and Bolivia, are actually erosional mountains sculptured from thick accumulations of lava and ash. In these areas however, modern eruptive vents give rise to volcanic cones that range from small cinder heaps to tremendous isolated mountains. The greater cones, such as Fuji, Ararat, Mauna Loa, Rainier, or Shasta, are among the most magnificent features of the Earth's surface. *See* MOUNTAIN; MOUNTAIN SYSTEMS; TERRAIN AREAS; VOLCANO.

Edwin H. Hammond

Bibliography. K. W. Butzer, *Geomorphology from the Earth*, 1977; A. Doerr and J. F. Coling, *Fundamentals of Physical Geography*, 1993; G. H. Dury, *An Introduction to Environmental Systems*, 1981; P. Keary and F. J. Vine, *Global Tectonics*, 1990; D. P. McKenzie, The Earth's mantle, J. Francheteau, The Oceanic crust, and B. C. Burchfiel, The continental crust, *Sci. Amer.*, pp. 66–145, September 1983; G. T. Trewartha et al., *Fundamentals of Physical Geography*, 3d ed., 1977.

Hiodontiformes

An order formerly in Osteoglossiformes but now independent, with the two orders comprising the extant Osteoglossomorpha, a primitive group of teleost fishes. The Hiodontiformes consist of one family, one genus, and two extant species, *Hiodon alosoides* (goldeye) and *H. tergisus* (mooneye). Hiodontids superficially resemble clupeids and for many years were classified as clupeiforms. The body is deep and laterally compressed and has a ventral keel, but the keel is not serrated as in most clupeids; the eyes are far forward and large, with the diameter greater than the length of the snout; the scales are large and cycloid and either bright silvery or golden; the anal fin base is much longer than the dorsal fin base; and the caudal fin is forked. The goldeye differs from the mooneye in having 9 or 10 principal dorsal fin rays versus 11 or 12, and a ventral keel extending anteriorly past the pelvic fins versus not extending anteriorly past the pelvic fins. *See* CLUPEIFORMES; OSTEICHTHYES; OSTEOGLOSSIFORMES.



Example of a hiodontid. (*From J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006*)

Both species are limited to the freshwaters of North America between the Continental Divide and the Appalachians. *Hiodon alosoides* has the greatest north-south distribution of North American freshwater fishes, ranging from 69° to 31°N latitude and including the MacKenzie River delta and Great Plains of northwestern Canada, isolated populations in the Hudson River drainage in eastern Canada, and populations southward in large rivers of the Mississippi Basin to Louisiana and Alabama. The species is absent from the Great Lakes basin. *Hiodon tergisus* ranges between latitudes 60° and 30°N, including the Mississippi River basin east of the Great Plains and west of the Appalachians to the Gulf Coast, parts of the Great Lakes, the St. Lawrence River, and tributaries to Hudson Bay northward.

Both species are opportunistic feeders, foraging in late evening and at night, at the surface, primarily on aquatic insects, crustaceans, mollusks, small fishes, and frogs. *Hiodon alosoides* attains a total length of 52 cm (20.5 in.), and *H. tergisus* a total length of 47 cm (18.5 in.). Herbert Boschung

Bibliography. E. J. Hilton, Comparative osteology and phylogenetic systematics of fossil and living bony tongue fishes (Actinopterygii, Teleostei, Osteoglossomorpha), *Zool. J. Linn. Soc.*, 137:1–100, 2003; G.-Q. Li and M. V. H. Wilson, Phylogeny of Osteoglossomorpha, pp. 163–174 in M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Hippopotamus

The name for two species of even-toed (artiodactylid) ungulates which form the family Hippopotamidae. Both species occur in Africa; the great African hippopotamus (*Hippopotamus amphibius*) inhabits the rivers of tropical Africa, and the pygmy hippopotamus (*Choeropsis liberiensis*) lives near the rivers of western Africa but is more terrestrial.

The great African hippopotamus, the largest living artiodactylid, reaches a length of 18 ft (5.5 m), stands about 5 ft (1.5 m) at the shoulder, and may weigh up to 4 tons (3.6 metric tons; see **illustration**). The ears are small and flexible, and the nostrils and eyes protrude so that they are out of the water as the animal floats. The skin may be 2 in. (5 cm) thick in places and almost devoid of hair. The tail is quite short and has some hair, as does the muzzle and inside of the ears. The feet end in four toes enclosed in round hoofs. The lateral toes are almost as well developed as the median ones, and all four toes touch the ground—a primitive condition. Both canines and incisors continue to grow during the life of these animals, a condition similar to that of rodents. These teeth are few in number and are long only in the lower jaw. The canines, which are curved and resemble tusks, may reach a length of 2 ft (0.6 m) and weigh 5–6 lb (2.3–2.7 kg). The other teeth are hidden behind the lips, but there are numerous powerful molars. The dental formula is I 2/2 C 1/1 Pm 4/4 M 3/3 for a total of 40 teeth. These animals have a gestation period of about 34 weeks, and a single calf is born.

The pygmy hippopotamus is about the size of a large pig and may weigh up to 400 lb (180 kg). It is a more solitary species and does not live in large herds, as does the great African species. The dental formula differs from that of *H. amphibius*, being I 2/1 C 1/1 Pm 4/4 M 3/3, for 38 teeth. *See* DENTITION.

These animals migrate regularly, following the river upstream during the rainy season and downstream during the dry season to new pastures. They come onto land, especially at night, to feed on vegetation. The males of the great African hippopotamus occupy and maintain territories within which are small herds of females and juveniles. The closest living relatives are the pigs. *See* ARTIODACTYLA. Charles B. Curtin

Hippopotamus (*Hippopotamus amphibius*). (*Photo by Gerald and Buff Corsi; © California Academy of Sciences*)

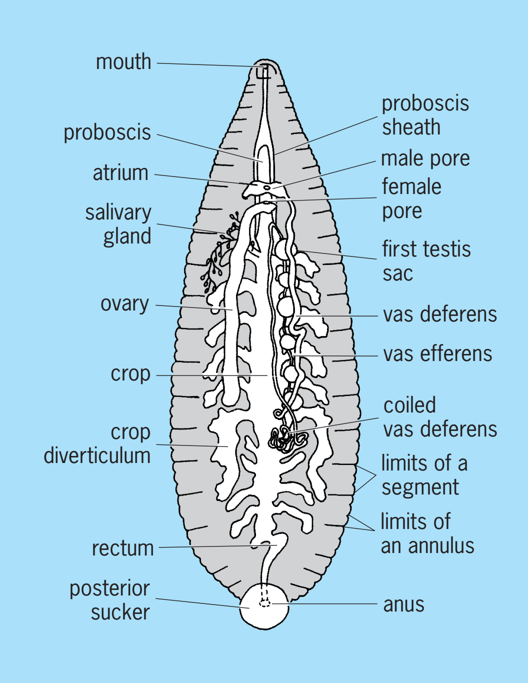
Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Hirudinea

A subclass of the annelid worms commonly known as leeches. These organisms are parasitic or predatory and have terminal suckers for attachment and locomotion. Most inhabit inland waters, but some are marine and a few live on land in damp places. The majority feed by sucking the blood of other animals, including humans.

Morphology. Leeches differ from other annelids in having the number of segments in the body fixed at 34, chaetae or bristles lacking, and the coelomic space between the gut and the body wall filled with packing (botryoidal) tissue. In a typical leech the first six segments of the body are modified to form a head, bearing eyes, and a sucker, and the last seven segments are incorporated into a posterior sucker (see **illus.**). Each segment is divided externally into 2–16 rings or annuli. The number per segment is constant for a particular species in the midbody region, but decreases toward the extremities.

The mouth of a leech opens within the anterior sucker, and there are two main methods of piercing the skin of the host to obtain blood. The rhynchobdellid leeches have the anterior part of the gut modified to form an eversible proboscis. The anterior sucker is planted firmly in position on the skin of the host, and the proboscis is forced out of the leech's mouth and into the host tissue. The arhynchobdellid leeches lack the proboscis, and in its place have three jaws, each shaped like half a circular saw, placed just inside the mouth. When the anterior sucker is placed in position on the skin of the host, the three jaws are pushed foward and rocked in such a way



General structure of a leech. Male reproductive system is shown on the right, the female on the left. (*After K. H. Mann, A key to the British freshwater leeches, Freshwater Biol. Ass. Sci. Publ., 14:3–21, 1954*)

as to cause a Y-shaped incision. Salivary glands are present in both orders of the leeches, and their secretion helps to prevent the clotting of blood. The jawed leeches are able to penetrate the skin of humans and other mammals, but a proboscis is less efficient, and rhynchobdellid leeches usually have to seek out softer tissues such as the lining of the nostril of a bird, or the gills of a fish. Some arhynchobdellid leeches have given up bloodsucking and feed instead on such animals as earthworms and insects. In these leeches, the jaws are often reduced to muscular ridges that are used merely to grasp the prey, which is normally swallowed whole. *See* ARHYNCHOBDELLAE; RHYNCHOBDELLAE.

The bloodsucking leeches have a large region of the gut, the crop, modified for the storage of blood. It consists of a central chamber with a number of paired diverticula, or side branches. When a meal of blood is taken, the crop becomes considerably distended, so that some leeches can store up to 10 times their own weight in blood.

The process of digestion is very slow, and a meal may last a leech for 9 months. The carnivorous forms have lost most or all of their gut diverticula and resemble earthworms in having a straight, tubular gut.

Life cycle. Leeches are hermaphroditic, having a single pair of ovaries and several pairs of testes. The latter are arranged on either side of the gut, and the sperms pass forward in paired ducts to a single midventral opening placed about one-third of the distance from the anterior to the posterior sucker. In many leeches there is an eversible penis, and sperms are transferred directly to the female pore of another leech. However, in others the penis is lacking, and the sperms are made up into packets, or spermatophores, which are attached to the body surface of another leech. From here the sperms migrate through the tissues to fertilize the eggs. The female pore is in the midventral line, a short distance behind the male pore, and at the time of egg laying a thickened, glandular region of the body wall, the clitellum, secretes a barrel-shaped cocoon around the body, over the genital pores. The fertilized eggs are passed into this cocoon, after which the leech works the cocoon over its head and then closes the ends. Some leeches loosely deposit the cocoons in a damp place, others cement them to stones or vegetation under water, while still others place them under their bodies and brood over them. In the last case the young, on hatching, attach themselves to the undersurface of the parent and are carried about for some time. The method of reproduction is very like that of the earthworm, and leeches may be regarded as earthworms which have become modified for the parasitic mode of life. *Acanthobdella* is regarded as a transitional form. It has only 30 segments in the body, of which 5 bear bristles, or chaetae. There is no anterior sucker, and the coelomic space is not entirely filled by packing tissue. *See* OLIGOCHAETA.

Economic importance. The importance of leeches as a means of making incisions for the letting of blood or the relief of inflammation is declining, and in civilized countries the bloodsucking parasites of mammals are declining, because of lack of opportunity for contact with the hosts. In other countries they are still serious pests. *Limnatis nilotica* of the Middle East enters the mouth or nostrils of animals or humans drinking at streams, and causes bleeding, nausea, and vomiting, often followed by death from anemia or suffocation. The leech may sometimes be removed, with difficulty, by the use of irritants such as salt water, chloroform water, or vinegar. *Hirudo* is widespread in Europe and Asia. Its attacks on humans and domestic animals are usually not serious, since they usually occur on the body or the limbs, but occasionally serious damage is caused by attacks on delicate tissues such as the eyes. In South America the attacks of *Haemadipsa chiliani* on horses and cattle can be fatal. *Theromyzon*, which enters the throat and nostrils of water birds, may cause serious losses of domestic ducks and geese when crowded conditions lead to an increase of the leech population. The fish parasite *Piscicola* may be a serious pest in hatcheries or heavily stocked ponds.

On the other hand, many of the leeches which are carnivorous in habit form an important contribution to the diet of fresh-water fishes. Records from the stomach contents of fish do not often show this, because of the rapid and complete digestion of the leeches. *See* ANNELIDA. Kenneth H. Mann

Bibliography. K. J. Muller et al. (eds.), *Neurobiology of the Leech*, 1981; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; R. T. Sawyer, *Leech Biology and Behavior*, 3 vols., 1986.

Histamine

A biologically active amine that is formed by the decarboxylation of the amino acid histidine. It is widely distributed in nature and is found in plant and animal tissues as well as in insect venoms. In humans, histamine is a mediator of inflammatory reactions, and it functions as a stimulant of hydrochloric acid secretion in the stomach.

Most tissue histamine is found stored in mast cells, where it can be released by a variety of stimuli. Once released, it can cause many effects, including constriction of bronchiolar, gastrointestinal, uterine smooth muscle, and lowering of blood pressure. If histamine is released in the skin, itching, a flare (area of redness) due to vasodilation, and a wheal due to leaking of fluid into the tissue are observed. The increase in vascular permeability that permits this leakage is due to an action on the endothelial cells of postcapillary venules.

All of these actions of histamine are mediated by the activation of histamine receptors, designated either H-1 or H-2. Antihistamine drugs exert their effects by blocking the combination of histamine with these receptors. *See* ANTIHISTAMINE.

Histamine release can be caused by tissue injury, by physical stimuli such as cold or pressure, by drugs such as heroin, and most importantly by immunologic mechanisms. Mast cells in the skin, the lung, the nasal passages, or other sites may become sensitized to antigens such as ragweed or other pollens, and then release histamine and other biologically active substances upon exposure to them. The released histamine may then cause wheals, bronchoconstriction, itching, runny nose, and other effects commonly associated with allergic responses. If the allergic reaction becomes generalized and severe, life-threatening anaphylactic shock may ensue. The prompt administration of epinephrine, which exerts effects opposite to those of histamine, can be life-saving in such cases. *See* ALLERGY; ANTIGEN; EPINEPHRINE; IMMUNOLOGY. Alan Burkhalter

Bibliography. M. Garcia-Caballero, L. J. Brandes, and S. Hosada (eds.), *Histamine in Normal and Cancer Cell Proliferation*, 1993; J. C. Schwartz and H. Haas, *The Histamine Receptor*, 1992.

Histocompatibility

A term used to describe the genes that influence acceptance or rejection of grafts. When grafts of tissue are exchanged between genetically dissimilar individuals, profound immunological rejection generally takes place. In contrast, grafts between genetically similar individuals, such as identical twins, are normally tolerated; they are histocompatible. Most known examples of histocompatibility (or H) genes encode polymorphic (that is, tending to differ between individuals) cell-surface proteins. The major histocompatibility complex (MHC) contains a set of histocompatibility genes, termed major because mismatching at these genes invokes rapid rejection. Other, minor antigens are revealed after precise matching of the major antigens.

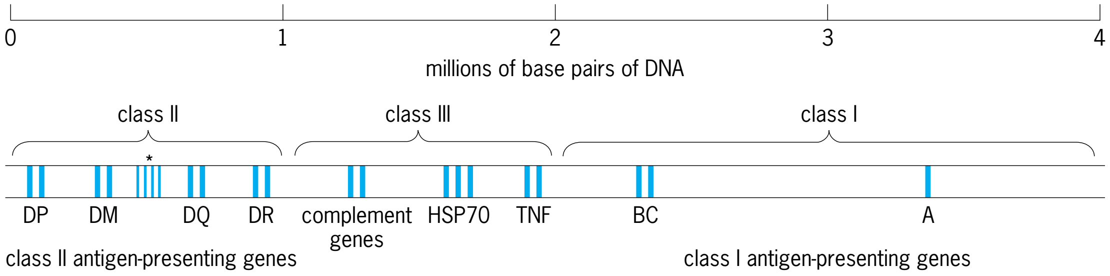
The main function of MHC genes involves distinguishing self from nonself in the immune system, as part of preventing the spread of infectious disease. The body employs special mechanisms to avoid rejection of the fetus, which is effectively an allograft, that is, a graft from a donor to a genetically dissimilar recipient of the same species; in this case, the mechanisms include a diminution of MHC gene expression.

Genetics. Histocompatibility antigens are encoded by genes that are inherited in a codominant, Mendelian fashion. The MHC, the locus to which the main barriers to tissue rejection map, contains a spectrum of genes, many of which influence processing and presentation of antigens to the immune system. In mice, the MHC is designated the H-2 complex; in humans, it is referred to as the HLA complex (for human leukocyte A system). Mice and other mammals seem to have a similar arrangement of genes in their MHCs. Other vertebrate MHCs have been identified, such as fish, chicken, and amphibian, and these are beginning to show some differences in arrangement of their genes. *See* MENDELISM.

The human MHC is normally divided into three regions, class I, II, and III. Located at the middle of the chromosome, the class II region contains several sets of related genes encoding the class II antigens (see **illus.**). The class II region also contains a group of genes that seem to be involved in processing antigens that are destined to be presented on class I molecules.

In the direction of the tip of the chromosome arm, the class III region consists of a mixture of over 40 genes, many of which bear no relationship to the immune system. Exceptions include genes for cytokines such as tumor necrosis factor. The class I region contains the *HLA*, *B*, and *C* genes as well as over 16 related loci. The class I genes are interspersed with a large number of genes of nonimmune function. *See* ANTIGEN; BLOOD GROUPS; CELLULAR IMMUNOLOGY; CHROMOSOME; CYTOKINE.

Polymorphism. Some of the MHC class I and class II genes have the remarkable property of being highly variable between individuals. Thus at each of the HLA loci there are multiple variant genes (alleles) in the population. Because humans are diploid, each individual can have two alleles at each locus. Each allele encodes one HLA antigen. Thus, an individual expresses two different HLA antigens from each series. The HLA system exhibits extreme polymorphism because there are so many different alleles (and corresponding antigens) in the population. When the alleles at each locus are considered in combination, millions of different HLA phenotypes can be generated, and it is rare to find unrelated individuals who share all the same antigens encoded at the HLA loci; hence the difficulties in tissue matching for transplantation. Siblings inherit one set of MHC genes (together called a haplotype) from each parent. Thus, their chances of having identical


0
1
2
3
4
millions of base pairs of DNA
class II
class III
class I
*
DP
DM
DQ
DR
complement genes
HSP70
TNF
BC
A
class II antigen-presenting genes
class I antigen-presenting genes

Arrangement of genes in the human major histocompatibility complex on the short arm of chromosome 6.

haplotypes is greater than for the general population. Transplant of a kidney from an HLA-identical (and ABO-identical) sibling has a high survival potential because there are no HLA incompatibilities.

Precise typing of HLA antigens or genes in tissue-typing laboratories aims to maximize matching organs from unrelated donors. Precise matching is critical in bone marrow transplantation, where there is a high load of lymphoid tissue and graft-versus-host reactions. It is progressively less important in transplantation of small bowel, kidney, heart, lung, and liver. However, in many cases, the use of powerful immunosuppressant drugs circumvents this problem. *See* IMMUNOSUPPRESSION; MUTATION; POLYMORPHISM (GENETICS).

Function. The MHC class I and class II products present fragments of pathogens to the immune system, in particular, to T cells.

In general, class I molecules bind peptides in the groove at the top end of the molecule. These peptides are derived from proteins that have been synthesized in the cytoplasm of the cell. The T cell contains a cell-surface molecule called a T-cell receptor which can bind to the class I molecule. Each T cell expresses a different receptor, that has been previously selected in the thymus to recognize a self MHC molecule only if it contains a nonself peptide. If the particular T-cell receptor binds to the MHC-peptide combination, the presenting cell is then killed.

An additional set of genes within the class II region, the *LMP* and *TAP* genes, are thought to be involved in processing antigen. The *LMP* gene products form part of a complex termed the proteasome, which cleaves proteins into peptides. These peptides then have to be transported across an internal membrane into the correct compartment where they encounter class I molecules. *TAP* gene products are responsible for this part of the process.

Class II molecules also bind peptides, usually those derived from proteins outside the cell. The foreign protein may be bound to the cell surface by means of an immunoglobulin molecule, expressed on a B cell. The foreign protein is then taken to a cellular compartment where it is broken into peptides capable of binding into the groove of the class II molecule. The class II molecule together with the foreign peptide, if recognized by an appropriate T cell, stimulates the T cell to produce cytokines which will cause the B cell to proliferate and produce immunoglobulin to neutralize the foreign protein.

The structure of class I and class II molecules clarifies their polymorphic quality. The amino acids that are most variable between alleles occupy those positions on the molecule that point into the peptide binding groove. This enables different individuals to bind different sets of peptides from diverse organisms. The inherent flexibility of the system is presumably important to impede the spread of pathogens since agents such as influenza virus or the human immunodeficiency virus (HIV) throw up altered protein sequences from time to time. Proteins from these escape mutants may not provide optimum peptides for binding into the grooves of certain MHC molecules, but their spread is checked by at least some of the many variant structures in the population that do elicit binding. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); IMMUNOLOGICAL ONTOGENY; INFLUENZA.

HLA and disease. Alleles of the various MHC genes are associated with a large number of diseases. These fall into three categories: autoimmune diseases, infectious diseases, and genetically linked diseases. Most common are the autoimmune diseases such as insulin-dependent diabetes mellitus, rheumatoid arthritis, and multiple sclerosis; these tend to occur more often with certain alleles. For example, multiple sclerosis tends to occur more frequently in individuals who possess the HLA-DR2 antigen. Ankylosing spondylitis, a disorder of the spine, is strongly associated with HLA-B27 individuals. Any explanation for these associations likely reflects the known biological function of HLA molecules in presenting antigens to the immune system. The second category of MHC-associated diseases is the infectious type; malaria is an example, but compelling evidence for others is lacking in humans. The third category consists of diseases for which the responsible gene is genetically linked to the MHC but is not thought to be involved with the immune system. There is a gene near the MHC alteration which disturbs iron uptake in the gut, causing hemochromatosis (that is, excessive accumulation of iron in the liver), for example. Another disease association is narcolepsy, in which affected individuals fall asleep involuntarily. It is almost exclusively associated with the HLA-DR2 antigen, although the reasons behind this remain obscure.

The associations between HLA and disease are important because of the insight that they provide as to the disease etiology. However, the HLA region provides just one of the factors influencing the course of these diseases, and important roles are also played by other unlinked genes located on other chromosomes, as well as by poorly understood environmental factors such as past and present microbial infection. *See* AUTOIMMUNITY.

Geography. HLA haplotypes are important demographic markers effective in tracing the spread of populations. For example, it is possible to monitor certain HLA alleles of ancient populations throughout the Americas, some of which are found in Asian populations. Their distribution is consistent with theories of migration across the Bering Straits from northeast Asia. In some isolated South American Indian populations, new *HLA-B* alleles seem to have arisen by recombination, possibly as a direct result of high pathogen loads during periods of their history. It is possible to trace so-called Celtic haplotypes through Europe, and to the west and north of the British Isles. By polymerase chain reaction, it is possible to analyze cadaveric samples, from bones or mummies, which may provide further clues to population movement. *See* GENETIC MAPPING.

Xenotransplantation. There is much interest in transplanting organs between species, using other mammals such as pigs as organ donors for humans. MHC incompatibility may not be as marked in this situation as in allotransplantation (that is, mismatching in the same species). This may be partly due not only to poor class I and class II recognition between species but also to incompatible surface adhesion molecules necessary for T-cell stimulation. There are a number of other problems, though, such as naturally occurring antibody to carbohydrate antigens, which may differ markedly between species. These antibodies can cause hyperacute rejection within minutes. It has been estimated that humans and Old World monkeys commit up to 1% of their antibody repertoire to a response against a carbohydrate present on other species studied for transplantation. *See* ANTIBODY; GENETICS; HUMAN GENETICS; TRANSPLANTATION BIOLOGY.

John Trowsdale

Bibliography. J. Klein and U. Klein (eds.), *Molecular Evolution of the Major Histocompatibility Complex*, 1991; R. Srivastava, G. P. Ram, and P. Tyle (eds.), *Immunogenetics of the Major Histocompatibility Complex*, 1991; A. R. Williamson and M. W. Turner, *Essential Immunogenetics*, 1987.

Histogenesis

The developmental processes by which the definite cells and tissues which make up the body of an organism arise from embryonic cells. Among animals, the ectoderm, endoderm, and mesoderm, also known as the primary germ layers, provide the stem cells which gradually transform into distinctive kinds of cells and tissues. In the higher plants, meristematic cells, which occur wherever extensive growth takes place, provide the basis for tissue formation. Parenchyma, a simple permanent tissue, may dedifferentiate and become meristematic. *See* APICAL MERISTEM; EMBRYOLOGY; GERM LAYERS; HISTOLOGY; LATERAL MERISTEM; PARENCHYMA; PLANT TISSUE SYSTEMS.

Charles B. Curtin

Histology

The study of the structure and chemical composition of tissues of animals and plants as related to their function. The primary aim is to understand how tissues are organized at all structural levels, including the molecular and macromolecular, the entire cell and intercellular substances, and the tissues and organs.

The four tissues of the animal body include cells and intercellular substances. They are (1) epithelium, in which the cells are generally closely applied to each other and separated by very little intercellular substance; (2) connective tissue, in which the cells are usually separated by greater amounts of intercellular substance, which may indeed form the great bulk of the tissue; (3) muscular tissue, whose cells are primarily concerned with contractility; and (4) nervous tissue, whose components are concerned primarily with rapid conduction of impulses.

The tissues are combined in various ways to form integrated patterns which are characteristic of every organ of the body. In this way, imperceptibly, the study of the minute structure and relations of organs merges into gross anatomy.

The structures studied in histology extend from those which are just too small to be seen with the hand lens to those which are beyond the resolution of the electron microscope, for example, those which measure about 2 nanometers. For this purpose, a wide range of instruments has been used. These include x-ray diffraction units, x-ray absorption microscope, electron microscope, polarization microscope, dark-field microscope, ultraviolet microscope, visible light microscope, phase contrast and interference microscope, and the dissecting microscope. *See* MICROSCOPE.

In earlier days, fresh material to be examined by the histologist was either teased or sliced freehand, scraped, or smeared so that it would be thin enough to be viewed with a microscope by transmitted light. Later, microtomes were introduced which could cut sections as thin as 1 micrometer; now they are capable of cutting sections 7 nm in thickness. The period of use of the microtome corresponds roughly with the use of fixatives, which were introduced to preserve and retain structure. The structural rearrangements, distortions, and losses caused by the use of fixatives and the use of irrelevantly esthetic stains eventually caused a reaction among histologists. This took the form of a return to the study of fresh material under strictly controlled conditions, the development of tissue culture and micromanipulation, microcinematography, intravital staining (of living cells), and supra vital staining (of surviving cells). Another practical development was fixation by freezing and drying. In this method of preservation specimens are frozen very rapidly by immersion in fluids cooled to −292°F (−180°C) or less, dried in a vacuum at a temperature of −58°F (−50°C) or less, and later infiltrated in paraffin. *See* MICROTECHNIQUE; TISSUE CULTURE.

The major fields of histological studies are (1) morphological descriptions; (2) developmental studies; (3) histo- and cytophysiology; (4) histo- and cytochemistry; and (5) fine (or submicroscopic) structure. Histo- and cytophysiology deal with correlations between morphological changes and functional activity. Histo- and cytochemistry deal with the chemical composition of morphological structures. The study of fine structure deals with the arrangements of structures which are below the resolution of the light microscope (about 0.2 μm). *See* CONNECTIVE TISSUE; EPITHELIUM; MUSCULAR SYSTEM; NERVOUS SYSTEM (VERTEBRATE). Isidore Gersh

Bibliography. D. H. Cormack, *Ham's Histology*, 9th ed., 1987; D. Fawcett, *Bloom and Fawcett: A Textbook of Histology*, 12th ed., 2001; L. C. Junqueira et al. (eds.), *Basic Histology,* 9th ed., 1998; T. S. Leeson, C. R. Leeson, and A. A. Paparo, *Text/Atlas of Histology*, 1988; I. R. Telford and C. F. Bridgman, *Introduction to Functional Histology*, 2d ed., 1994; L. Weiss (ed.), *Cell and Tissue Biology: A Textbook of Histology*, 6th ed., 1988.

Historadiography

The technique for taking x-ray pictures of cells, tissues, or sometimes the whole animal or plant, if it is a small one. Soft x-rays, those with low penetrating power and relatively long wavelengths, are required for this type of picture. The best pictures are obtained when the tissues contain deposits of metallic elements which have a high absorption capacity for x-rays. *See* X-RAYS.

In applying the technique to tissues, a relatively thin section is placed against an x-ray film and irradiated with a beam of x-rays. When the film is developed, a picture of the object or section of tissue shows on the film. Another method attempts to focus the x-rays after they pass through the specimen. In this arrangement, a lens is placed between the specimen and the film. X-rays are very difficult to focus. The lenses must be the reflecting type of mirror surfaces. A simple two-mirror system has two cylindrically curved surfaces set at right angles to each other. The glancing angle must be small to get reflection of the x-rays. The lenses also have certain aberrations that limit the resolution to about 0.2 micrometer, which is no better than the resolution of the best light microscope images. However, x-rays with their short wavelengths are potentially capable of much higher resolution if they could be focused without much aberration. The advantage of the x-ray picture is that it may resolve detail not visible with light in calcified bone tissues or in tissues injected with an x-ray- absorbing material.

An interesting use of historadiography in biology is the determination of the relative mass of cellular structures. The tissue is quick-frozen and dried at low temperatures. Then sections are prepared and photographed by the method described above. A relation exists between the mass of the various parts of the specimen and the contrast of the film, that is, the number of silver grains or the amount of silver deposited. Quantitative determinations can be made by scanning the photograph with a densitometer. *See* NUCLEAR RADIATION (BIOLOGY). J. Herbert Taylor

Hodgkin's disease

A malignant lymphoid neoplasm, usually arising in lymph nodes characterized by morphological heterogeneity and bizarre giant tumor cells referred to as Reed-Sternberg cells. The etiology of Hodgkin's disease is unknown, although current epidemiological data suggest an infectious etiology. The disease has been recognized in all or several members of "close-knit" groups, usually occurring in these persons several years after they have separated from one another. As in non-Hodgkin lymphomas, there is morphological and immunological evidence to suggest a viral etiology. *See* ANIMAL VIRUS.

Reed-Sternberg cells. Hodgkin's disease is diagnosed by finding Reed-Sternberg cells in histologic sections of involved lymph nodes. Classically, Reed-Sternberg cells are greater than 40 micrometers in diameter and have large bilobed nuclei and very prominent eosinophilic nucleoli surrounding by a halo. Several variants of Reed-Sternberg cells have been identified.

Classification. Until 1966 Hodgkin's disease was classified morphologically as paragranuloma, granuloma, and sarcoma. In 1966 R. Lukes and coworkers reclassified Hodgkin's disease into six categories using the following morphologic criteria: (1) the number of lymphocytes or histiocytes, or both, present in the involved lymph node; (2) the type, number, and distribution of the Reed-Sternberg cells; and (3) the amount and distribution of the associated connective tissue in the diseased tissue. The disease was then reclassified into four major groups: lymphocyte predominant; nodular sclerosis; mixed cellular; and lymphocyte depleted. Each of the four groups has a distinct microscopic appearance and different clinical course, Progression from one subtype to another is not uncommon.

Clinical features. Persons with Hodgkin's disease usually seek medical advice because of enlarged painless lymph nodes. They may also have fever, weight loss, anorexia, pruritus, and anemia. The clinical extent of the disease is determined by a process of staging based on physical examination, various biopsies, and usually a laparotomy for examination of the spleen and liver. The disease is divided into the following stages: stage 1—disease limited to one anatomic region or two contiguous anatomic regions on the same side of the diaphragm; stage 2—disease in more than two anatomic regions or in two noncontiguous regions on the same side of the diaphragm; stage 3—disease on both sides of the diaphragm but not extending beyond the involvement of lymph nodes, spleen, or Waldeyer's ring; stage 4—involvement of the bone marrow, lung parenchyma, pleura, liver, bone, skin, kidneys, gastrointestinal tract, or any organ in addition to lymph nodes, spleen, or Waldeyer's ring. Each stage is also classified as A or B depending on whether the patient has systemic symptoms of disease.

Susceptibility. Hodgkin's disease more commonly affects males than females, except for the nodular sclerosing variety, which occurs with equal frequency in both sexes. It generally is a disease of persons between the ages of 20 and 40, but it may affect the very young and the very old. Characteristically, persons with Hodgkin's disease exhibit a loss of cell-mediated immunity and become susceptible hosts for infection with a variety of microorganisms such as tubercle bacilli.

Cellular origin. The cellular origin of Hodgkin's disease remains uncertain. One hypothesis suggests that it is caused by a defective T-lymphocyte immune system which permits the development of neoplastic Reed-Sternberg cells. *See* CELLULAR IMMUNOLOGY.

Treatment. The treatment of Hodgkin's disease is dependent on its clinical stage and microscopic appearance. A combination of chemotherapy (anticancer drugs) and radiation therapy is commonly used. The best prognosis is in those persons with stage 1 disease who have the lymphocyte predominant or nodular sclerosing variety. Very encouraging therapeutic results have occurred, with many apparent cures. *See* CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; LYMPHATIC SYSTEM; RADIOLOGY.

Samuel P. Hammar

Bibliography. E. Braunwald et al. (eds.), *Harrison's Principles of Internal Medicine*, 15th ed., 2001; F. Cavalli, G. Bonodonna, and M. Rozencweig (eds.), *Malignant Lymphomas and Hodgkin's Disease: Experimental and Therapeutic Advances*, 1986; B. W. Dana (ed.), *Malignant Lymphomas, Including Hodgkin's Disease: Diagnosis, Management, and Special Problems*, 1993.

Hoisting machines

Mechanisms for raising and lowering material with intermittent motion while holding the material freely suspended. Hoisting machines are capable of picking up loads at one location and depositing them at another anywhere within a limited area. In contrast, elevating machines move their loads only in a fixed vertical path, and monorails operate on a fixed horizontal path rather than over a limited area. *See* ELEVATING MACHINES; MONORAIL.

The principal components of hoisting machines are sheaves and pulleys, for the hoisting mechanism; winches and hoists, for the power units; and derricks and cranes, for the structural elements.

Block and tackle. Sheaves and pulleys or blocks are a means of applying power through a rope, wire, cable, or chain. Sheaves are wheels with a grooved periphery, in appropriate mountings, that change the direction or the point of application of a force transmitted by means of a rope or cable. Pulleys are made up of one or more sheaves mounted in a frame, usually with an attaching swivel hook, eye, or similar device at one or both ends. Pulley systems are a combination of blocks.

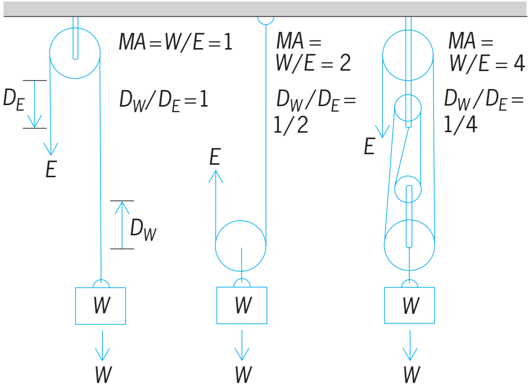


Fig. 1. Mechanical advantage of pulley systems.

The mechanical advantage *MA* of a pulley system is the ratio of weight lifted *W* to input effort exerted *E*. The mechanical advantage, neglecting friction, for any of various arrangements can be determined readily because it equals the number of strands that support the load (**Fig. 1**). The ratio of distance D_W through which the weight is lifted to distance D_E through which the effort is expended is inversely proportional to the mechanical advantage.

Sometimes used alone, sheaves and pulleys find their most usual application as the hoisting tackle of derricks and cranes.

Winches and hoists. Normally, winches are designed for stationary service, while hoists are mounted so that they can be moved about, for example, on wheel trolleys in connection with overhead crane operations.

A winch is basically a drum or cylinder around which cordage is coiled for hoisting or hauling. The drum may be operated either manually or by power, using a worm gear and worm wheel, or a spur gear arrangement. A ratchet and pawl prevent the load from slipping; large winches are equipped with brakes, usually of the external band type. Industrial applications of winches include use as the power element for derricks and as the elevating mechanism with stackers (**Fig. 2**).

Floor- and wall-mounted electric hoists are used for many hoisting and hauling jobs from fixed locations in industrial plants and warehouses. Heavy-duty types are standard equipment for powering ship's gear in cargo handling. They are also mounted on over-the-road carriers to facilitate the moving of heavy bulky loads, and serve as the power units of power cranes and shovels. A railroad car puller employs the same principle; however, the drum is mounted vertically and is used for spotting railroad cars in freight yards.

Hoists are designed to lift from a position directly above their loads and thus require mobile mountings.

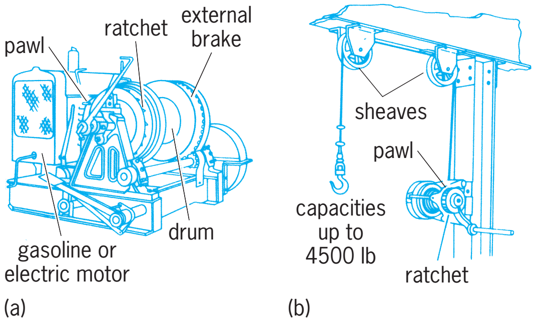


Fig. 2. Powered and hand winches. (*a*) Heavy-duty single-drum winch. (*b*) Wire-rope hand winch. 4500 lb = 2042 kg.

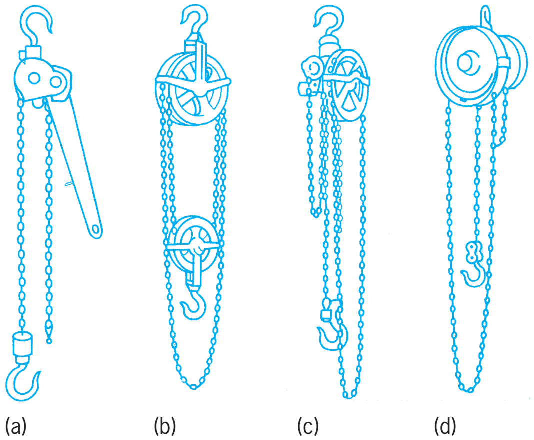


Fig. 3. Chain hand hoists. (*a*) Lever (ratchet). (*b*) Differential. (*c*) Worm gear. (*d*) Spur gear.

Hoists are classified by their power source, such as hand, electric, or pneumatic.

Hand hoists are chain operated. There are four types: spur geared, worm geared, differential, and pull-lift or lever (**Fig. 3**). The last, with its lever for operation and its ratchet for holding, is the simplest and most economical. However, since the operating lever is located on the anchor end of the hoist, it is not as convenient for vertical lifting as it is for horizontal pulling.

The spur-gear type costs the most but is the most economical to operate, with an efficiency as high as 85%. Where hoists are to be used frequently, it is the type most recommended.

Worm-gear hoists transform about one-third to one-half the input energy to useful work. Offsetting this low efficiency is the locking characteristic of the worm drive: The load cannot turn the mechanism; consequently the load is at all times restrained from running away. In contrast, with a ratchet the load is locked only at positions where the pawl engages a step of the ratchet. Differential hoists use only about one-third the energy input; they too prevent loads from running away during lowering. Spur-gear hoists are more efficient, but require a brake to restrain loads during lowering or holding.

Electric hoists lift their loads by either cable or chain (**Fig. 4**). The cable type has a drum around which a wire cable is coiled and uncoiled for hoisting and lowering. Chain models have either a roller chain and sprocket or a link chain and pocketed wheel for hoisting and lowering.

There are innumerable below-the-hook attachments, such as slings, hooks, grabs, and highly specialized devices to facilitate practically any handling requirement. Many of these devices are designed to pick up and release their loads automatically. All chain hoists are designed with the lower hooks as the weakest parts; not being interchangeable with the anchor hook, therefore, if the hoist is overloaded, the lower hook spreads or opens up. If the inside contour of the hook is not a true arc of a circle, this is an indication that the hook has been overloaded.

Pneumatic or air hoists are constructed with cylinders and pistons for reciprocating motion and air motors for rotary motion. Compressed air is the actuating medium in both. Various arrangements admit air to and discharge it from cylinders mounted to operate vertically or horizontally. Pneumatically operated hoists provide smooth action and sensitive response to control; these characteristics account for their wide use in handling fragile materials, such as molds in foundries. In addition, freedom from sparking makes them useful in locations where the presence of explosive mixtures make electrical equipment hazardous.

Derricks and cranes. A derrick is distinguished by a mast in the form of a slanting boom pivoted at its lower end and carrying load-supporting tackle at its outer end. In contrast, jib cranes always have horizontal booms.

Derrick masts are supported by guy lines or stiff legs; some are arranged to rotate 360°. Winches, hand or powered, usually in conjunction with pulleys, do the lifting. Derricks are standard equipment on construction jobs; they are also used on freighters for loading and unloading cargo, and on barges for dredging operations.

Jib cranes, when carried on self-supporting masts, are called pillar cranes; those mounted on walls are called wall-bracket cranes. Cranes with jiblike booms are frequently used in shops, mounted on columns

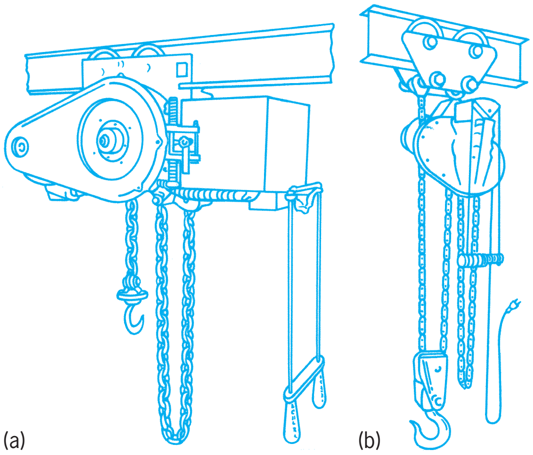


Fig. 4. Hoists. (*a*) Chain. (*b*) Roller chain.

or walls, but have limited coverage. They may have their own running gear or be mounted on trucks (**Fig. 5**). Mobile types for heavier service are called yard cranes or crane trucks. They may or may not be able to rotate their booms. More powerful machines of this type belong to the power crane and shovel group. *See* BULK-HANDLING MACHINES.

Overhead-traveling and gantry cranes. Hoisting machines with a bridgelike structure spanning the area over which they operate are overhead-traveling or gantry cranes. In the overhead-traveling type, the bridge is carried by, and moves along, overhead trackage which is usually fixed to the building structure itself. The gantry crane is normally supported by fixed structures or arranged for running along tracks on ground level. Gantry cranes are standard equipment in shipside operations. Basic arrangements of overhead-traveling cranes are top running and underhung. In the former, the bridge's end trucks ride on top of the runway rails; in the latter, the end trucks carry the bridge suspended below the rails (**Fig. 6**). Types for relatively light duty can be made of elements used in the construction of overhead track. *See* MONORAIL.

Both overhead-traveling and gantry cranes are called bridge cranes. These cranes span a rather large area and differ among themselves primarily only in the construction of the bridge portion of the crane, and in the method of suspension of the bridge. Where smaller areas are to be spanned, standard beams are used for the bridge structure; however, built-up girders or trusslike bridge structure are used for larger spans. A full gantry crane has both its supporting elements erected on the ground, usually riding on tracks. A half-gantry crane has a supporting structure at one end of the bridge that reaches to the ground, and the other end is carried directly on overhead tracks. Selection of either type depends primarily on building design and the areas in which the crane is to be used.

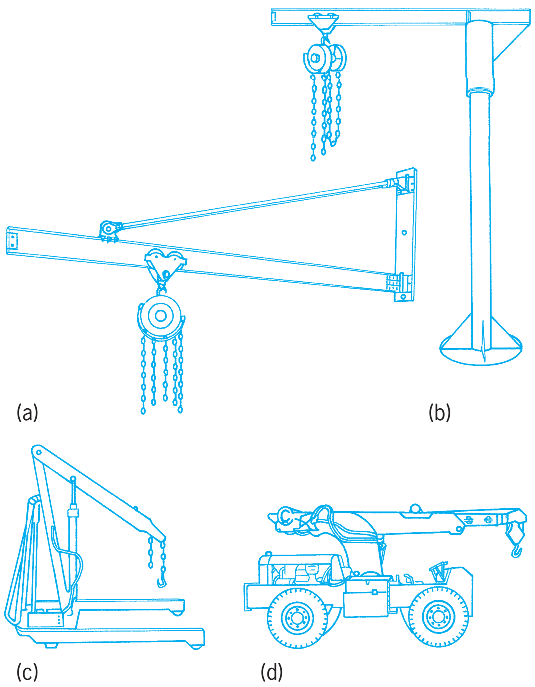


Fig. 5. Four types of jib cranes. (*a*) Wall crane. (*b*) Pillar crane. (*c*) Movable hydraulic crane. (*d*) Crane truck.

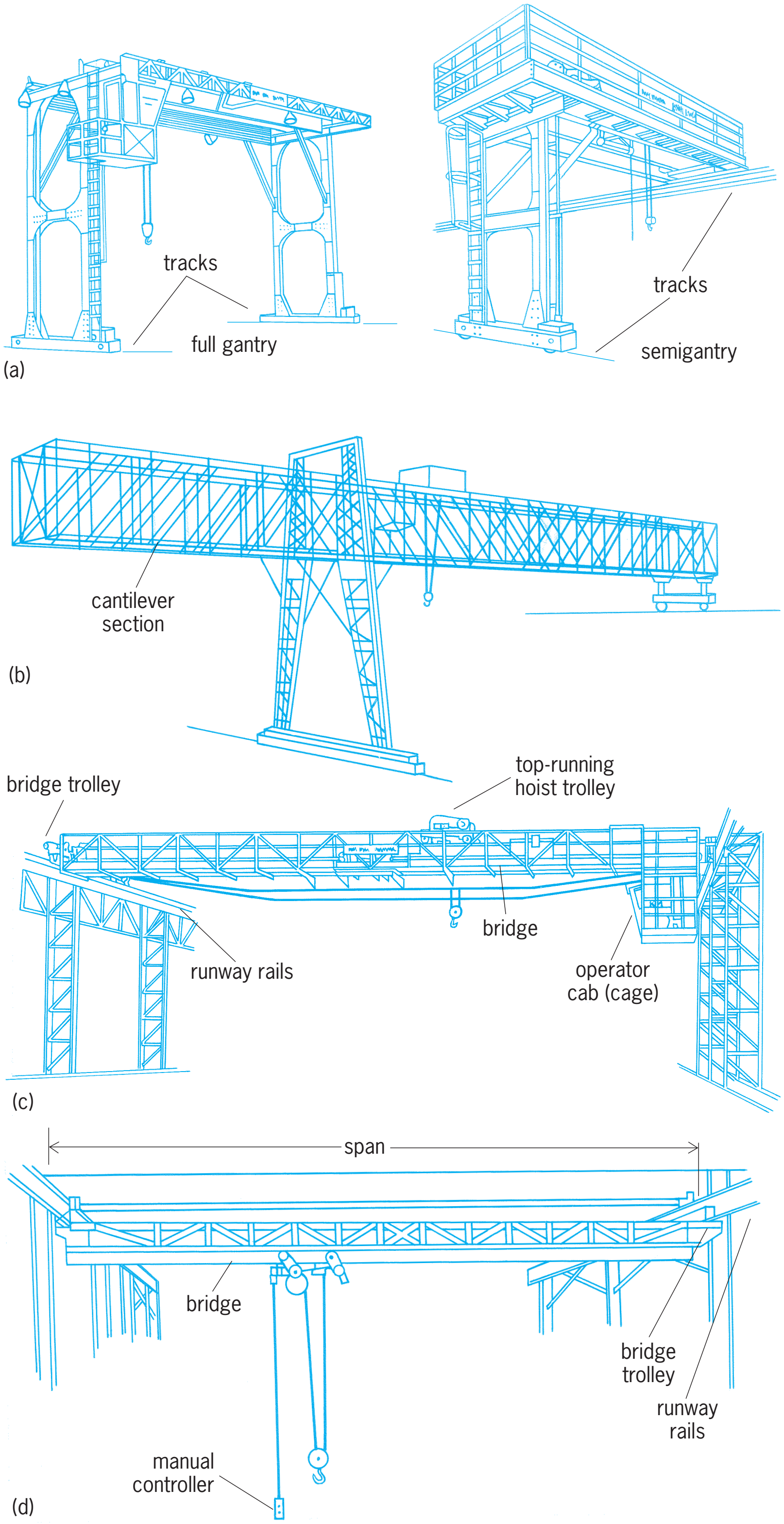


Fig. 6. Basic types of overhead-traveling cranes. (*a*) Traveling gantry cranes. (*b*) Cantilever crane. (*c*) Top-running crane. (*d*) Underhung crane.

Any one or several of the hoists described earlier may be attached to the bridge, usually suspended from a trolley attached to an I-beam track. The combination of a hoist on a track and the bridge crane itself moving on tracks provides for usable movement of equipment within a rectangular area governed only by the length of the bridge and the total horizontal movement of the bridge or gantry crane.

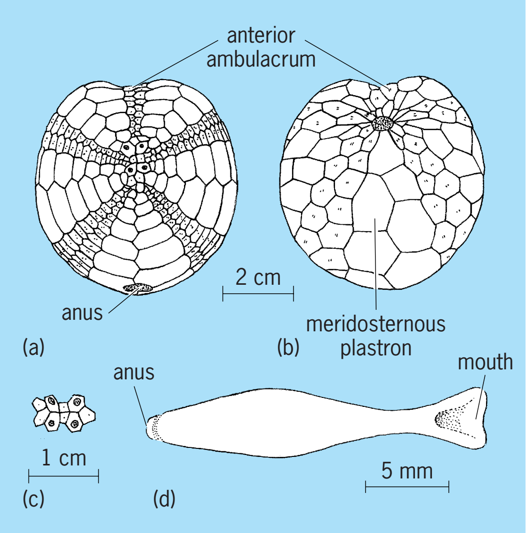
In simpler units the bridges and trolley hoists may be hand-propelled. In heavy-duty units hand operation is not practical and separate electric motors drive each motion. Controls for the motors vary from pendant-type push buttons operated from the ground to remote or automatic control. Pendant controls are satisfactory when a crane has intermittent use during the work day; in larger units, where the crane is in constant use and heavy loads are the rule, an operator may be stationed in a cab mounted to the bridge structure. A more sophisticated extension of this is the operatorless crane, which is worked by means of an electronic control.

Another type of lifting mechanism, used on a modified type bridge crane, is a fork-lift attachment suspended from the overhead truck or bridge; it is referred to as a stacker crane. This unit is especially used for such locations as stock rooms, die racks, or finished-goods storage because it has the advantage of being able to handle unit loads in narrow aisles. Arthur M. Perrin

Bibliography. J. M. Apple, *Plant Layout and Materials Handling*, 3d ed., 1977, reprint 1991; Electric Overhead Crane Institute, *Specifications and General Information for Standard Industrial Service: Electric Overhead Traveling Cranes*; W. A. Rossnagel, *Handbook of Rigging: For Construction and Industrial Operations*, 4th ed., 1988; H. I. Shapiro, *Cranes and Derricks*, 3d ed., 1999.

Holasteroida

An order of irregular echinoids (sea urchins) of the superorder Atelostomata, with a strongly bilaterally symmetrical test. They have a small oval mouth lacking buccal notches that lies close to the anterior on the lower surface: a lantern is never present. Ambulacral plating is simple. The anterior ambulacrum is often differentiated, and petals are developed adapically (see **illus.**). The anus lies on the posterior face. The elongate apical disc, in which ocular plates II and IV abut along the midline and separate the anterior and posterior pairs of genital plates, is the most distinctive characteristic. In a few genera, the apical disc is split, anterior and posterior parts being separated by intervening interambulacral plates. Phyllodes of penicillate feeding tube feet with their associated characteristically large pores are typically developed around the mouth.



Diagnostic features of holasteriods. (*a*) Aboral aspect. (*b*) Adoral aspect. (c) Apical system. (*d*) *Echinosigra paradoxa*, a deep-sea species of Pourtalesiidae, adoral aspect.

The 105 Recent species are divided into 13 genera. All are deep-sea forms, living at depths between 325 and 24,600 ft (100 and 7500 m), but are most commonly found between 4900 and 14,800 ft (1500 and 4000 m). Their fossil record indicates that they were much more common and diverse in the past, with 61 genera divided into seven families, most of these being shallow-water forms. The group first appeared in the Early Cretaceous, diversified greatly during the middle and Late Cretaceous, and then declined during the Tertiary to their present low levels. They are all deposit feeders, living either epifaunally or shallowly buried in unconsolidated substrata. *See* ECHINODERMATA. Andrew Smith

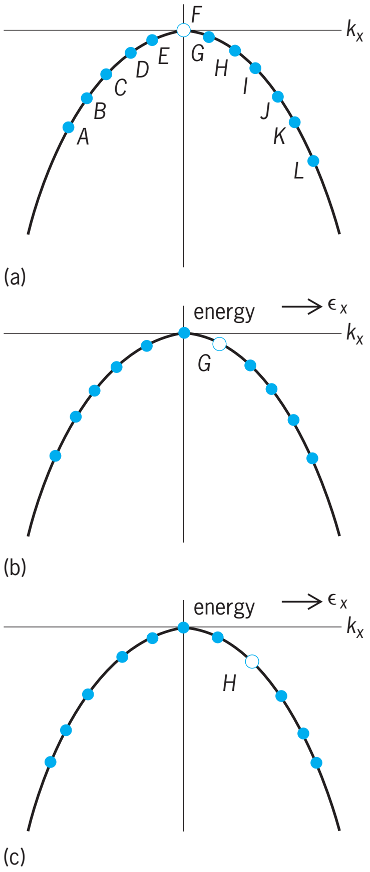
Hole states in solids

Vacant electron energy states near the top of an energy band in a solid are called holes. A full band cannot carry electric current; a band nearly full with only a few unoccupied states near its maximum energy can carry current, but the current behaves as though the charge carriers are positively charged. The situation can be understood in terms of the definition of the effective mass: if the energy band is specified by a function $E(k)$, where k is the magnitude of the wave vector $\mathbf{k}$, the effective mass for a spherical band is given by the equation below, where $\hbar$ is Planck's constant

$$m* = \hbar^2 \left(\frac{\partial^2 E}{\partial k^2}\right)^{-1}$$

divided by 2π. Near a maximum of the band, the second derivative of the energy is negative, so the effective mass is negative. States for which the effective mass is negative are defined as hole states. Carriers in such states behave under the influence of an external electromagnetic field as though they carry positive charge. *See* BAND THEORY OF SOLIDS.

The process of conduction in such a system may be visualized in the following way. An electron moves against an applied electric field by jumping into a vacant state. This transfers the position of the vacant



Process of hole conduction. (*a*) At time $t = 0$, energy states *A* through *L* are filled except *F*. (*b*) An electric field ϵ_x is applied in the $+x$ direction. The force on the electrons is in the $-k_x$ direction, and all electrons make transitions in the $-k_x$ direction, moving the hole to *G*. (c) After a further interval, the electrons move farther along, and the hole is now at *H*. (*After C. Kittel, Introduction to Solid State Physics, 3d ed., Wiley, 1956*)

state, or propagates the hole, in the direction of the field, as shown in the **illustration**. Whether conduction occurs by electrons or holes is determined experimentally from the sign of the Hall emf. If a current is carried in the presence of a magnetic field perpendicular to the current, an emf is developed perpendicular to the current and to the field. The sign of this emf depends on the charge on the carriers. *See* HALL EFFECT.

Hole conduction is important in many semiconductors, notably germanium and silicon. The occurrence of hole conduction in semiconductors can be favored by alloying with a material of lower valence than the "host." Semiconductors in which the conduction is primarily due to holes are called *p* type. Hole conduction is also observed in some metals, including iron and chromium. In other metals, including aluminum and bismuth, both holes and electrons may be present in equilibrium. *See* ELECTRICAL CONDUCTIVITY OF METALS; SEMICONDUCTOR.

Joseph Callaway

Bibliography. N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, 1976; C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996; A. Sutton, *Electronic Structure of Materials*, 1993.

Holectypoida

An extinct order of primitive irregular sea urchins of the class (Echinoidea), which retain a functioning lantern tern throughout life. Although their periproct opens on the oral surface in the posterior interambulacrum, the test still shows considerable radial symmetry. The mouth is large, circular, and centrally positioned and is indented by sharp buccal notches. The lantern has stout teeth that are wedge-shaped in cross section. Ambulacral plating is trigeminate adorally but usually simple adapically; the tube feet and their pore pairs are undifferentiated. Tubercles are arranged radially on the oral surface, and tuberculation is distinctly finer adapically. The holectypoids are distinguished from other primitive irregular echinoid groups by the presence of a fifth genital plate in the apical disc that may be perforated by a gonopore.

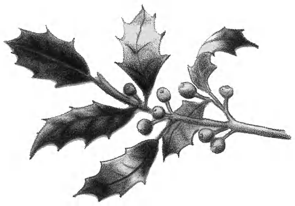
The approximately 130 nominal species are divided into nine genera and two families, Holectypidae and Discoididae. The families are differentiated on the presence or absence of internal buttressing. Holectypoids first appeared in the late Lower Jurassic and survived to the end of the Cretaceous (Maastrichtian), when they became extinct. They were mostly infaunal deposit feeders living in relatively coarse, permeable substrata. *See* ECHINODERMATA; ECHINOIDEA.

Andrew Smith

Holly

The American species of holly (*Ilex opaca*) attains a maximum height of 40–50 ft (12–15 m) and has evergreen leaves. It grows naturally in the eastern and southeastern United States close to the Atlantic and Gulf coasts, in the Mississippi Valley, and westward to Oklahoma and Missouri. It is best known for its bright red berries, which make a pleasing contrast with the deep-green spiny leaves, and for this reason it is valued for decorations at the Christmas season (see **illus.**). The wood is hard, tough, and close-grained. The heartwood is ivory white when first cut, but becomes brownish with age or on exposure, and takes a high polish. It is used for cabinet work and musical instruments; because it resembles ivory, it is sometimes used for keys for pianos and organs. Its fine grain makes it valuable for wood-engraving work.

The English holly (*I. aquifolium*) is cultivated extensively in the extreme northwestern United States, but is not hardy in the northeastern states. Its spiny



The American holly (*Ilex opaca*).

leaves are glossier than those of the American holly and have wavier margins. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Holmium

A chemical element, Ho, atomic number 67, atomic weight 164.93, a metallic element belonging to the rare-earth group. The stable isotope ^{165}Ho makes up 100% of the naturally occurring element. The metal is paramagnetic, but as the temperature is lowered, it changes to antiferromagnetic and then to the ferromagnetic system. *See* ANTIFERROMAGNETISM; PERIODIC TABLE; RARE-EARTH ELEMENTS.

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

Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner, Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.

Holocene

That portion of geologic time that postdates the latest episode of continental glaciation. The Holocene Epoch is synonymous with the Recent or Postglacial interval of Earth's geologic history and extends from 10,000 years ago to the present day. It was preceded by the Pleistocene Epoch and is part of the Quaternary Period, a time characterized by dramatic climatic oscillations from warm (interglacial) to cold (glacial) conditions that began about 1.6 million years ago. The term Holocene is also applied to the sediments, processes, events, and environments of the epoch.

In 1833 British geologist Charles Lyell applied the term Recent to the interval of time of human existence. Now it is known that humans have been around much longer than the last 10,000 years, but that fact does not discount the need for a term for the part of Earth's history closest to us. In 1846 Forbes proposed Postglacial as an alternative term for Recent. However, Postglacial is inappropriate in higher latitudes where the postglacial interval did not begin until several thousand years after it did in the middle latitudes.

In 1867 the term Holocene (meaning wholly recent) was proposed for the interval of time following the Pleistocene, and in 1885 it was formally submitted to the International Geological Congress. The U.S. Geological Survey officially adopted the term Holocene in 1967, and the U.S. Commission on Stratigraphic Nomenclature endorsed it in 1969.

The Holocene represents the current interglacial episode. The Holocene interval is treated differently in geologic time charts than are earlier interglacial episodes, however. It is designated as a separate epoch of the Quaternary Period. Although there is little reason to suspect that the present interglacial interval differs substantially from earlier ones and there is little doubt that another glacial interval will follow, a separate designation is appropriate for the Holocene. It holds a unique place in geologic history—it is the time of rapid development of human culture and the rise of humankind as a geologic agent whose activities have substantially modified Earth's environment.

As the interval of time closest to us, the Holocene Epoch is very convenient to study. Holocene sediments cover virtually every part of the Earth's surface and represent almost every environment of deposition. With the development of ^{14}C dating (a method of age determination based on the measurement of radioactive carbon decay), Holocene sediments are relatively easy to date. From a scientific standpoint, the Holocene Epoch is of great interest because it provides a recent analog for past environments and processes. Its sediments and landforms provide important clues to changes that occurred as a result of the last shift from the glacial to the nonglacial climatic mode. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; RADIOCARBON DATING.

The Pleistocene/Holocene transition was a time of dramatic environmental change. The huge ice sheets that had developed over the northern and western parts of North America (Laurentide and Cordilleran, respectively) and most of Scandinavia were at their maximum geographic extent about 18,000 ^{14}C years B.P. (before present, where present is defined as the

CENOZOIC	QUATERNARY	**Holocene**
		Pleistocene
	TERTIARY	Pilocene
		Milocene
		Oligocene
		Eocene
		Paleocene
MESOZOIC	CRETACEOUS	

year 1950) and in full retreat by 14,000 ^{14}C years B.P. By 10,000 ^{14}C years B.P., the Laurentide ice sheet had withdrawn from the Great Lakes. The ice sheets survived in the northern latitudes for another 3000 ^{14}C years or so. The progress of deglaciation was complex, because the overall glacial meltback was interrupted by intervals of glacier readvance. It remains unclear whether these readvances were synchronous on a hemispheric or global scale and what role ice sheet/oceanic interactions played in the deglaciation. *See* PLEISTOCENE.

Boundaries. Because the deglaciation began earlier in the middle latitudes than it did in the high latitudes, the boundary between sediments of the Pleistocene and Holocene epochs is time-transgressive (varies in age from place to place). A precise location for the boundary between sediments corresponding to the two epochs has never been agreed upon, although an arbitrary age of 10,000 years ^{14}C years B.P. (defined by the Holocene Commission for the International Quaternary Association) is commonly used as the age for the boundary.

The time-transgressive nature of the boundary between Pleistocene and Holocene sediments is verified by ^{14}C dating, and some geologists have argued that it is more realistic and practical to represent boundaries between late Quaternary events and sediments as diachronous (beginning and ending at different times in different places) than to use geologic time classifications that assume time-synchronous (same age everywhere) boundaries for events and sediments. Since 1983, when the North American Stratigraphic Nomenclature Committee introduced diachronic units, informal and formal diachronic classifications for late Quaternary time and event units have been proposed as alternatives for geologic time units with time-synchronous boundaries. An example of a diachronic unit is the Hudson Episode. It is based on postglacial sediments that occur stratigraphically above deposits of the Wisconsin (last glacial) Episode. The Hudson Episode, which began several thousand years earlier in Illinois than it did at Hudson Bay, is an alternative designation to the Holocene Epoch, a geologic time unit that began everywhere 10,000 years ago.

Study and subdivision. Most studies of the Holocene apply an interdisciplinary approach. Such an approach is necessary to address complex interactions among the various natural systems, which respond with different sensitivities and at different rates to change. Traditionally, climate changes in the Holocene have been studied chiefly through the technique of palynology (pollen analysis), but increasingly paleontology (for example, fossil plant remains, beetles, ostracods, diatoms), dendochronology (tree ring analysis), marine and ice core records, and fluctuations of alpine glaciers have been used to reconstruct past climates. *See* PALEOCLIMATOLOGY.

Although commonly viewed as a relatively stable climatic interval compared to the Pleistocene Epoch, the Holocene Epoch has experienced small, yet significant climatic changes. It is subdivided informally into an early phase (from about 10,000 to 8000 ^{14}C years B.P.) influenced by the waning effects of receding ice sheets, a middle phase (from about 8000 to 4000 ^{14}C years B.P.) characterized by the warmest interval of the present interglacial episode, and a late phase (the last 4000 ^{14}C years) that included climatic reversal at the onset of Neoglaciation (a time of renewed glacial activity that some assume to be the first phase of the next glacial episode).

Early Holocene. The early phase of the Holocene was geologically the most eventful. The periglacial (near the edge of the ice) landscape was unstable and very dynamic. As the Pleistocene ice sheets melted, enormous volumes of water, stored as glacier ice for many thousands of years, returned to the oceans via meltwater streams or by way of ice streams that flowed directly to the ocean.

As the ice sheets shrank, sea level rose an average of 130 m, drowning the continental margins and closing many land bridges, including the land bridge across the Bering Strait between Asia and North America that had enabled humans to migrate to the Americas. In parts of Canada and Scandinavia, temporary marine invasions occurred when the ice melted from low areas where the Earth's crust had been depressed by the weight of the ice sheets.

In natural catchment basins along the retreating Cordilleran and Laurentide ice margins, huge lakes formed. Some were on the order of several hundred thousand square kilometers in areal extent. Many of these lakes were ephemeral and flashy (there one year and gone the next over several thousands of years). When ice or sediment dams gave way, catastrophic floods scoured the landscape and the lakes drained quickly, only to refill and repeat the process. Examples include glacial Lakes Missoula and Columbia, whose torrents repeatedly swept across Washington state to the Pacific Ocean via the Columbia River valley, and glacial Lakes McConnell and Agassiz, whose floodwaters alternately spilled to the Arctic Ocean, the Gulf of Mexico, and the Atlantic Ocean. Some lakes shrank more gradually or dried up completely when they were no longer fed by glacial meltwaters. Other lakes, such as the Great Lakes, found new, lower outlets, which sometimes altered the drainage course of vast watersheds.

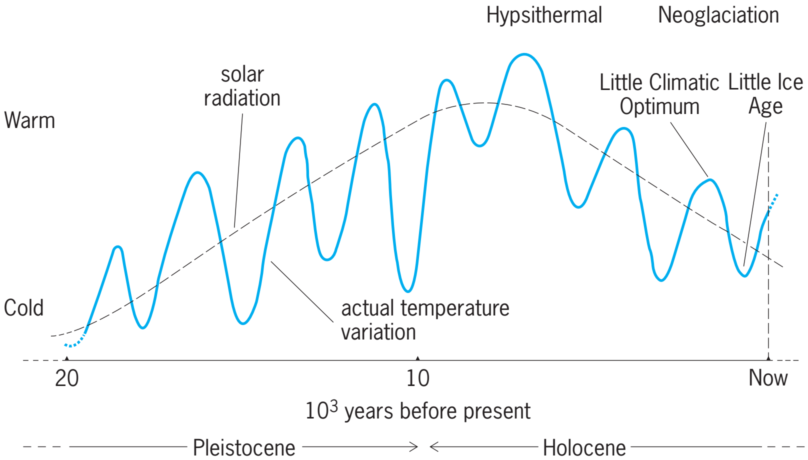
As the ice sheets waned, the Earth's crust rose, rebounding from the release of the weight of thousands of meters of glacier ice and creating uplifted shoreline features and sediments. Parts of Hudson Bay and Scandinavia were uplifted several hundred meters. Maximum uplift occurred in the early Holocene, but uplift continues even today although at much slower rates.

With the retreat of the Northern Hemisphere ice sheets came a wave of migration that eventually resulted in the present distribution of plant communities. Pollen studies of cores from lakes and bogs in glaciated areas of western Europe and North America indicate that tundra areas bordering the ice initially gave way to spruce forests. Later, the spruce were replaced by pine and birch, followed still later by broad-leaf deciduous trees such as oak and elm.

Animal migrations followed the changes in the distribution of plant communities. A dramatic phenomenon in the late Quaternary was the extinction of a large number of animals, especially large mammals, at the close of the last glacial episode. Most of the extinctions in North America took place before 10,000 ^{14}C years B.P., and almost all had occurred by 8000 ^{14}C years B.P. The cause of these extinctions is not known, but they coincide with the Pleistocene/Holocene transition, the time of maximum climate and vegetational change, the opening of an ice-free corridor between the Cordilleran and Laurentide ice sheets in Canada, and massive migrations of human hunters south to the margins of the ice sheets.

As the plant and animal communities adjusted to the changing climate of the early Holocene, other environmental adjustments also occurred. These adjustments were highly dependent on the geographic area and the direct or indirect impact of glaciation on the landscape. In the recently deglaciated areas, soils had begun to form, the effects of periglacial conditions on the landscapes had diminished, and slopes were becoming more stable by the close of the early Holocene interval. In the tropics and the equatorial belt, precipitation increased and lakes were filled to high levels, indicating a climate that was much wetter than that of today.

Middle Holocene. The middle phase of the Holocene has been called the hypsithermal, a name for the warmest interval of the present interglacial episode (see **illus.**). It has also been referred to as the climatic optimum, a term which is more appropriately applied to the peak warmth of the hypsithermal phase. At the climatic optimum, world temperature was probably 2 or 3°C higher than today. Like other late Quaternary phases, the hypsithermal reached the lower latitudes earlier than it did the middle and higher latitudes. In the tropics the climatic optimum may have corresponded more closely to the interval of maximum insolation (exposure to the Sun), about 10,000 ^{14}C years B.P., but in the middle latitudes of the midwestern United States it occurred about 7000 ^{14}C years B.P., whereas in Labrador it occurred about 4000 ^{14}C years B.P. Fossil finds reveal that the geographic ranges of numerous species of plants and animals were several hundred kilometers farther north during the hypsithermal than they are today. For example, broad-leaved trees reached farther north than they now do in northern Europe and North America. The descendants of fossil beetles found in central Ontario do not extend north of southern Ohio today. In the higher latitudes, permafrost (ground that remains frozen even during the summer) likely extended to a greater depth below the surface and was much less widespread than today. The climate was warm enough to melt much of the sea ice in the Arctic Ocean, as indicated by the occurrence of fossil driftwood (dated at 4000–6000 ^{14}C years B.P.) on uplifted beaches.



Climatic variations during the last 20,000 years. The broken line is solar radiation reaching the Earth at 65°N latitude. The solid line is a speculative reconstruction of actual temperature variation. (*From E. C. Pielou, After the Ice Age: The Return of Life to Glaciated North America, University of Chicago Press, 1991*)

In mountainous areas of the middle latitudes, the treeline and other vegetational boundaries shifted during the middle Holocene to several hundred meters higher than their present elevations. In the mountains of the western United States, many alpine glaciers disappeared. In the midwestern United States, the middle Holocene was a time not only of maximum warmth but of relatively dry conditions. Due to summer drought, the prairie/forest border extended more than 100 km east of its present position. Today relict prairie soils can be found in areas that became forested during the late Holocene. The Great Plains became more arid, which is believed to have caused a major reduction of human and animal populations in that area during the hypsithermal.

Late Holocene. After the climatic optimum, the Earth experienced climatic cooling. The shift to a cooler, moister climate began about 5000–4000 ^{14}C years B.P. in the midcontinent. In western North America at about 5000 ^{14}C years B.P., the mountain glaciers began to expand again. This renewed glacial activity is called Neoglaciation (see illus.). At least three intervals of glacial expansion have occurred in the late Holocene. The glacial advances are cyclic. In the mountains of the western United States, the three advances have been dated at about 5000, 2800, and 300 ^{14}C years B.P. The most impressive of the three glacial intervals is the last, called the Little Ice Age. It is well documented because it occurred in historic time. Between the intervals of glacier expansion were times of climatic warming. One, called the Little Climatic Optimum to differentiate it from the hypsithermal of the middle Holocene, peaked about 1800 ^{14}C years B.P.

Many Rocky Mountain glaciers advanced as much as 1–2 km during the Little Ice Age. Trees that were overrun by advancing glaciers now are being exposed as the ice recedes due to a climatic warming that began over 100 years ago. Little Ice Age moraines are well defined in the Rocky Mountains and in Europe. The moraines commonly occur only a few hundred meters downslope from the present glacier margin. Their surfaces are bare of vegetation, unlike older Holocene moraines. *See* MORAINE.

The late Holocene phase saw the spread of peatland (also called muskeg) across much of Canada and the northern Great Plains, particularly in areas of former lake plains. Peatland development is attributed to the return to a wetter, cooler climatic regime. Treelines that had migrated northward and to higher elevations during the hypsithermal interval returned southward and to lower elevations.

During the late Holocene, human populations expanded and human culture developed into the complex agricultural, industrial, and technological society of today. The result is that humans have become significant factors in altering the Earth's surface environment, including, most believe, Holocene climate. A number of scientists attribute recent global warming (nearly 0.5°C in the past 150 years) to the buildup of carbon dioxide and other greenhouse gases in the atmosphere as a result of human activity. Some scientists warn of a much greater increase to come in the next 50 years. Only geologic time will tell the impact of humankind on the Earth's environment, including the natural climatic cycles. *See* GEOLOGIC TIME SCALE; GREENHOUSE EFFECT; QUATERNARY.

Ardith K. Hansel

Bibliography. C. L. Matsch, *North America and the Great Ice Age*, McGraw-Hill, New York, 1976; E. C. Pielou, *After the Ice Age: The Return of Life to Glaciated North America*, University of Chicago Press, 1991; H. E. Wright (ed.), *Late Quaternary Environments of the United States*, vol. 2: *The Holocene*, University of Minnesota Press, Minneapolis, 1983.

Holography

A technique for recording, and later reconstructing, the amplitude and phase distributions of a coherent wave disturbance. Invented by Dennis Gabor in 1948, the process was originally envisioned as a possible method for improving the resolution of electron microscopes. While this original application has not proved feasible, the technique is widely used as a method for optical image formation, and in addition has been successfully used with acoustical and radio waves. This article discusses holography with electromagnetic waves in the optical and microwave regions of the electromagnetic spectrum, and its potential use with x-rays. For holography with sound waves. *See* ACOUSTICAL HOLOGRAPHY.

Optical Holography

Optical holography makes use of a highly coherent beam of light, such as supplied by a laser source. *See* LASER.

Fundamentals of the technique. The technique is accomplished by recording the pattern of interference between the unknown object wave of interest and a known reference wave (**Fig. 1**). In general, the object wave is generated by illuminating the (possibly three-dimensional) subject of concern with the coherent light beam. The waves reflected from the object strike a light-sensitive recording medium, such as photographic film or plate. Simultaneously a portion of the light is allowed to bypass the object, and is sent directly to the recording plane, typically by means of a mirror placed next to the object. Thus incident on the recording medium is the sum of the light from the object and a mutually coherent reference wave.

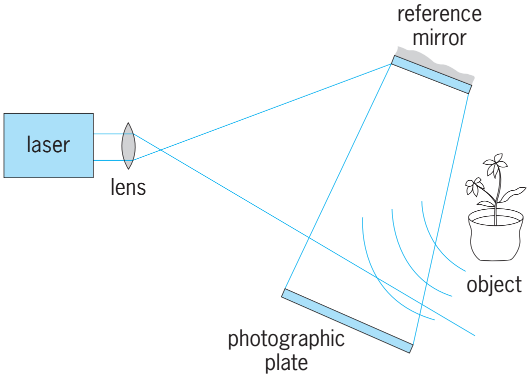


Fig. 1. Recording a hologram.

While all light-sensitive recording media respond only to light intensity (that is, power), nonetheless in the pattern of interference between reference and object waves there is preserved a complete record of both the amplitude and the phase distributions of the object wave. Amplitude information is preserved as a modulation of the depth of the interference fringes, while phase information is preserved as variations of the position of the fringes. *See* INTERFERENCE OF WAVES.

The photographic recording obtained is known as a hologram (meaning a total recording); this record generally bears no resemblance to the original object, but rather is a collection of many fine fringes which appear in rather irregular patterns (**Fig. 2**).

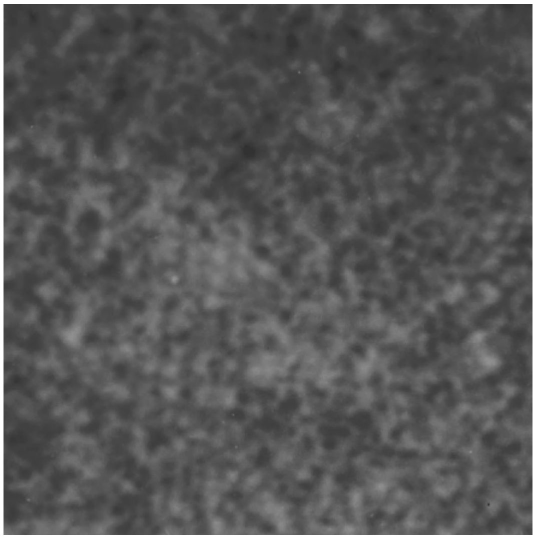


Fig. 2. Typical appearance of a hologram (under magnification).

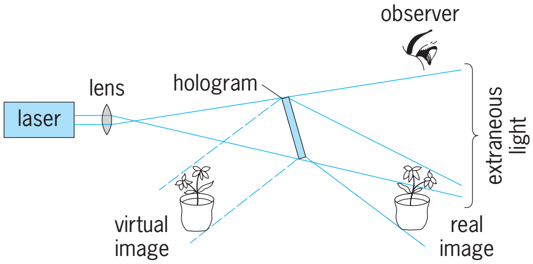


Fig. 3. Obtaining images from a hologram.

Nonetheless, when this photographic transparency is illuminated by coherent light, one of the transmitted wave components is an exact duplication of the original object wave (**Fig. 3**). This wave component therefore appears to originate from the object (although the object has long since been removed) and accordingly generates a virtual image of it, which appears to an observer to exist in three-dimensional space behind the transparency. The image is truly three-dimensional in the sense that the observer's eyes must refocus to examine foreground and background, and indeed can "look behind" objects in the foreground simply by moving the head laterally.

Also generated are several other wave components, some of which are extraneous, but one of which focuses of its own accord to form a real image in space between the observer and the transparency. This image is generally of less utility than the virtual image because its parallax relations are opposite to those of the original object.

Applications. The holographic technique has a number of unique properties which make it of great value as a scientific tool.

Microscopy. Historically, microscopy is the potential application of holography that has motivated much of the early work, including the original work of Gabor. The use of holography for optical microscopy has been amply demonstrated, but these techniques are not serious competitors with more conventional microscopes in ordinary microscopy.

Nonetheless, there is one area in which holography offers a unique potential for optical microscopy. This area might be called high-resolution volume imagery. In conventional microscopy, high transverse resolution is achieved only at the price of a very limited depth of focus; that is, only a limited portion of the object volume can be brought into focus at one time. It is possible, of course, to explore a large volume in sequence by continuously refocusing to examine new regions of the object volume, but such an approach is often unsatisfactory, particularly if the object is a dynamic one, continuously in motion. A solution to this problem is to record a hologram of the object by using a pulsed laser. The dynamic object is then "frozen" in time, but the recording contains all information necessary to explore the full object volume with an auxiliary optical system. Sequential observation is acceptable because the object (that is, the holographic image) is no longer dynamic. This approach has been fruitfully applied to the microscopy of three-dimensional volumes of living biological specimens and to the measurement of particle-size distributions in aerosols.

Interferometry. Holography has been demonstrated to offer the capability of several unique kinds of interferometry. This capability is a consequence of the fact that holographic images are coherent; that is, they have well-defined amplitude and phase distributions. Any use of holography to achieve the superposition of two coherent images will result in a potential method of interferometry.

The most powerful holographic interferometry techniques are based on the following property: When a photographic emulsion is multiply exposed to form several superimposed holograms, upon reconstruction the several corresponding virtual images are formed simultaneously and therefore interfere. Likewise the various real images interfere.

The most dramatic demonstrations of this type of interferometry use a pulsed ruby laser. Two laser pulses are used to record two separate holograms on the same transparency. Any changes of the object between pulses result in well-defined fringes of the interference in the reconstructed image (**Fig. 4**). The technique is particularly well suited for performing interferometry through imperfect optical elements (for example, windows of poor quality), thus making possible certain kinds of interferometry that could not be achieved by any classical means. *See* INTERFEROMETRY.

Memories. Optical memories for storing large volumes of binary data in the form of holograms have been intensively studied. Such a memory consists of an array of small holograms, each capable of reconstructing a different "page" of binary data. When one of these holograms is illuminated by coherent light, it generates a real image consisting of an array of bright or dark spots, each spot representing a binary digit. This image falls on a detector array, with one detector element for each binary digit. Thus to read a single binary digit at a specific location in the memory, a beam deflector causes light to illuminate the appropriate hologram page, and the output of the proper detector element is interrogated to

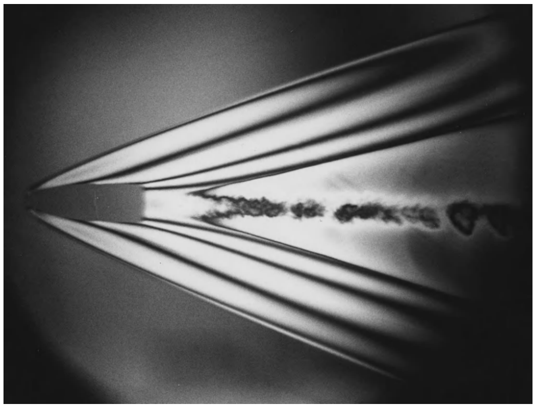


Fig. 4. Image taken by the technique of holographic interferometry, showing the compressional waves generated by a high-speed rifle bullet. (*Courtesy of R. E. Brooks, L. O. Heflinger, and R. F. Wuerker*)

determine whether a bright spot of light exists at that particular location in the image. *See* COMPUTER STORAGE TECHNOLOGY.

Display. There has been interest in the use of holography for purposes of display of three-dimensional images. Applications have been found in the field of advertising, and there is increased use of holography as a medium for artistic expression. A significant technical development in this area has been the perfection of a type of recording known as a multiplex hologram. Such a recording typically consists of a large number of separate holograms, all in the form of thin, contiguous, vertical strips on a single piece of film. Each of these holograms produces a virtual image of a different ordinary photograph of the subject of interest. In turn, each such photograph was originally taken from a slightly different angle. Thus when the observer examines the virtual image produced by the entire set of holograms, each eye looks through a different hologram and sees the subject from a different angle. The resulting stereo effect produces a nearly perfect illusion of three-dimensionality. Furthermore, as an observer moves the head horizontally, or as the collection of holograms is rotated, the observer's two eyes continuously see a changing pair of images. If the original set of photographs is properly chosen, the image can be made to move or dance about in nearly any desired fashion. Very dramatic three-dimensional displays of animated subjects can thus be constructed from a series of ordinary photographs. Such displays do not require a laser for viewing, but rather can be utilized with white-light sources.

Holographic optical elements. A hologram consisting of the interference of a plane reference wave and a diverging spherical wave, upon illumination by a reconstruction plane wave, will generate a diverging spherical wave (the virtual image) and a converging spherical wave (the real image), each traveling in a different angular direction. Thus such a hologram behaves as an optical focusing element, with properties similar to those of a lens (or, more accurately, a pair of lenses). More complex holograms can generate a multitude of foci, in virtually any pattern desired. Alternatively, by varying the periodicity of the gratinglike structure of the hologram, a small laser beam can be deflected through an angle that is controlled by the local period of the structure. Holograms which are used to control transmitted light beams, rather than to display images, are called holographic optical elements. Interest in such elements has grown substantially, and commercial applications have been found. Most notable is the use of holograms in supermarket scanners at checkout stands. Light from a helium-neon laser falls on a small region of a holographic optical element, which was recorded on a disk and is rotating continuously. As the hologram rotates, different portions of the hologram containing different grating periods are illuminated, and the angle of deflection of the laser beam sweeps through a pattern that was predetermined when the hologram was recorded. In this way the laser beam is caused to follow a complicated scan pattern, which ultimately allows the reading of information from the bar-code patterns recorded on each product. *See* CHARACTER RECOGNITION; GEOMETRICAL OPTICS; OPTICAL IMAGE.

Security applications. Holograms are widely used in various applications where authenticity is of the utmost importance, for example, in credit cards and in bank notes to prevent forgery. A holographic postage stamp is in use in the United States to prevent counterfeiting. Security applications have had the greatest commercial impact of any applications of holography.

Other applications. A variety of other applications of holography has been proposed and demonstrated, including the analysis of modes of vibration of complicated objects, measurement of strain of objects under stress, generation of very precise depth contours on three-dimensional objects, and high-resolution imagery through aberrating media. These and other applications of holography will be useful in future scientific and engineering problems. Joseph W. Goodman

Thick holograms. Thick, white-light reflection holograms were introduced by Y. Denisyuk in 1962. By illuminating a photosensitive plate so that the object and reference waves come from opposite sides of the plate, fringes can be formed that run parallel to the surfaces of the plate. These fringes act in concert as a multilayer mirror whose layers are warped. The warping is such that when the hologram is illuminated a reconstruction of the original object results. In addition, in polychromatic illumination the layers act as a thin-film stack to select the original wavelength. Multiple fringe patterns can be recorded to reproduce a color reconstruction of the object.

Rainbow hologram. A rainbow hologram, invented by S. A. Benton in 1969, is composed of vertically aligned narrow strips. Each strip reconstructs a single perspective of a three-dimensional (3-D) object. A strip is selected for viewing by locating the pupil of the eye on a line connecting the strip to a point in the light source. Because the light path to the left and right eyes will pass through different strips, the viewer has a stereoscopic impression of the object. Moving the head then gives a full 3-D impression. Object motion can also be incorporated by capturing the object at different times in different strips. For example, a 3-D object may appear to move as the viewer walks around a cylindrical hologram.

Embossed holograms. The embossed hologram is stamped into the medium. For this application, a mask is generated in a material such as photoresist. The mask is coated with a hard material to become a master for generating stamps. The stamps are then used on softer but durable materials to create the holographic emblems that are widely seen today. The technique of rainbow holography can also be applied to create pseudo-3-D images from white-light illumination.

Digital holography. In digital holography the recording is done by the combination of a digital sensor such as a charged-coupled device (CCD) and computer memory. The fringe pattern that is formed between the optical object and reference waves is stored digitally. The computer rather than wave propagation generates the reconstruction. *See* CHARGE-COUPLED DEVICES; DIGITAL COMPUTER.

Computer-Generated Holography

Information in a hologram is contained in patterns called fringes (**Fig. 5***a*). For optical holography, the fringe patterns are formed by the interference of two beams of light. In contrast, the fringes in a computer-generated hologram (CGH) are computer-calculated and then drawn, plotted, or printed on a transparent or reflective material. The computer-generated hologram was invented by A. W. Lohmann in 1966. It drew on the recent introduction of the fast Fourier transform, the recent performance gains of computers of that era, and the availability of relatively precise plotters. Insights into the relation of communications theory to holography developed by Lohmann around 1956 also played a crucial role in the development of computer-generated holography. Precursors to the computer-generated hologram can be found in the artificial holograms of W. L. Bragg (1939) and of G. L. Rogers (1952). A variation of the computer-generated hologram, the Kinoform, was described by L. B. Lesem in 1969 and has found important applications.

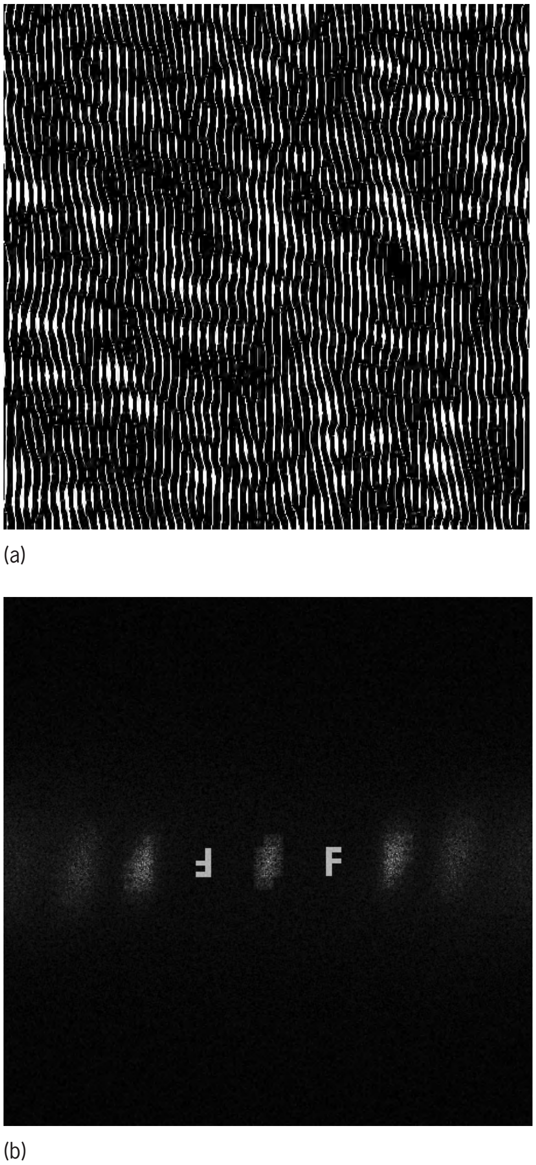


Fig. 5. Computer-generated hologram (CGH). (*a*) Mask. (*b*) Reconstruction from the hologram.

The computer-generated hologram, like the optical hologram, is an entire recording. Information is stored at each point on the computer-generated hologram's surface about both the energy of the light and the direction in which the light is moving. Figure 5*a* is a computer-generated hologram mask; the mask is photographically reduced onto transparent material such as a slide. If Fig. 5*a* is reduced to a size of about 5 mm × 5 mm onto slide film, the resulting slide will be a computer-generated hologram. If a point of red light, for example, a laser-point spot on a distant wall, is observed though the computer-generated hologram, the observer sees not a point but the letter F as illustrated in Fig. 5*b*. The letter appears to hover at the location of the point source of light. The computer-generated hologram has converted the simple optical wave emanating from the point into the more complicated wave that would have come from the F. The object in this illustration is flat, but a three-dimensional object could have been chosen just as easily. Because the hologram is artificially generated, there is almost unlimited flexibility with regard to what information is put into it.

In addition to the desired image, the F, there are other structures visible in Fig. 5*b*. The technical term for the desired image is the true image. An inverted image called the twin image is also seen, which comes from the fact that this type of computer-generated hologram is purely absorbing and does not change the phase of the light. There is a single point in the center of the reconstruction called the D.C. spike (D.C. is from the electrical term direct current), which is again an effect of the absorptive nature of this computer-generated hologram. The fuzzy clouds are known as high-diffraction-order noise and come from the fact that this type of computer-generated hologram is binary (that is, it contains only black and white) rather than graytone (containing shades of gray).

Constructing a computer-generated hologram. There are many ways of designing a computer-generated hologram. The three original Lohmann computer-generated hologram types seen in **Fig. 6** provide examples. A virtual grid is imposed on the surface that will become the computer-generated hologram. A rectangular aperture is inserted into each cell of the grid. For the type III computer-generated hologram, that aperture is centered vertically and its width is half the width of the cell. The height of the aperture is proportional to the amplitude A. The horizontal displacement is proportional to the phase Φ.

Applications. Computer-generated holograms can form images. The images can be located at any distance from the computer-generated hologram. The image location can be either in front of the computer-generated hologram or behind it. The images can be

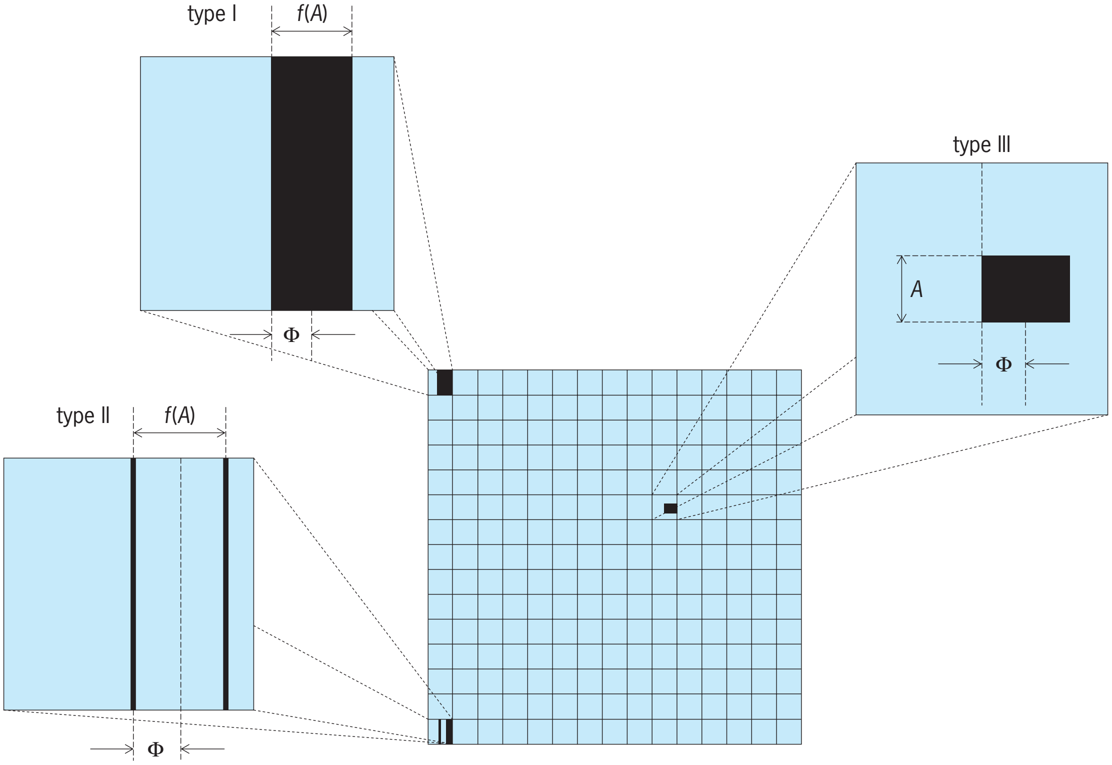


Fig. 6. The three original types of binary computer-generated holograms introduced by A. W. Lohmann.

three-dimensional and can be in color. Computer-generated hologram systems have even been built that produce 3-D television. They are presently limited to showing small images by the large amount of information needed to produce the images. The advantage of the computer-generated hologram over optical holograms is that the hologram may be of an object that has never existed. All that is needed is a mathematical description of the object.

Optical testing. In optical testing, an element under test, the test-piece, is compared to an element of known quality, the reference-piece. One way to perform such a test is to form overlapping images of the two pieces in laser light. The interference patterns formed show the difference between the pieces to a very high accuracy. The use of computer-generated holograms in optical testing was pioneered by A. J. MacGovern and J. C. Wyant in 1971. The basic idea in using computer-generated holograms for optical testing is to replace, or augment, the conventional reference-piece with a computer-generated hologram. Computer-generated holograms can be used in either transmission or in reflection. The primary advantage of a computer-generated hologram is its flexibility. Much more complicated reference pieces can be realized as computer-generated holograms than are practical to fabricate by traditional shaping and polishing of optical glass. *See* INTERFEROMETRY.

Diffractive optical elements. Just as holograms can be fabricated that serve as holographic optical elements, a computer-generated hologram can be created that acts as a diffractive optical element (DOE). The diffractive optical element mimics the optical properties of a glass element, such as a prism or lens. The advantages of such a diffractive optical element are that it is thin and weighs very little. The Kinoform is a type of computer-generated hologram with a structure that is similar to the glass door on a shower, though the dimples in the glass are very thin, about a micrometer, and perfectly controlled rather than random. Kinoforms have been widely applied in imaging systems as lenses. They are gaining acceptance in illumination systems, especially those that are based on light-emitting diodes. A stepped version of the Kinoform introduced by H. Dammann in 1970 has been widely used because of its relatively low cost. *See* LIGHT-EMITTING DIODE.

Diffractive optical elements have inherent dispersion; they act differently on different colors of light. Dispersion can be a direct advantage in devices such spectrometers that rely on it to function. It can be a disadvantage in lenses used for imaging. *See* DISPERSION (RADIATION).

Diffractive optical elements have combined with conventional glass lenses in camera objectives. The dispersion of the two types of elements cancel out, allowing commercial production of camera objectives that are significantly lighter and smaller than their all-glass counterparts. *See* CAMERA.

Optical vortices. Computer-generated holograms can be used to produce waves that rotate about the direction of propagation. These so-called helical waves can be used as propeller beams, to apply torque to components (cranks) of micromechanical devices; as optical vortices to trap and rotate microscopic particles; and in fiber-optic power transmission to maximize the transmission capabilities by minimizing the concentration of light at the center of the fiber. Helical waves are characterized by the

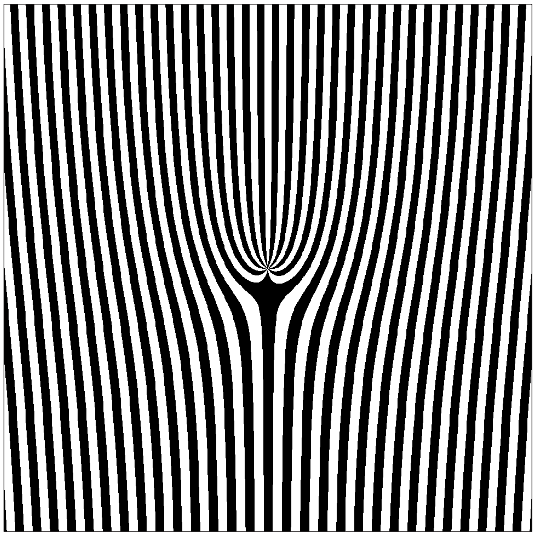


Fig. 7. Computer-generated hologram for topological charge of nine.

number of rotations per wavelength of propagation. This number is called the topological charge or the circular-harmonic order. **Figure 7** shows a computer-generated hologram that can be used to produce a wave with a topological charge of nine.

William J. Dallas

Microwave Holography

Microwave holography is microwave imaging by means of coherent continuous-wave electromagnetic radiation in the wavelength range from 1 mm to 1 m. As a long-wavelength imaging modality, it differs from techniques which employ echo timing (for example, conventional radar) by its requirement for phase information. In this respect it resembles optical holography, from which it has departed significantly. The technique usually involves small-scale systems, that is, systems in which the effective data acquisition aperture is of the order of tens or hundreds of wavelengths. Microwave holographic imaging is characterized by high lateral-resolution capability in comparison with images obtained from echo timing. The natural image format of the data it presents to the human observer enhances its diagnostic potential. In particular, it conveniently produces phase imagery which increases further its diagnostic capability. *See* MICROWAVE; RADAR.

Microwave holographic imaging originated from the two-stage optical process consisting of data recording in the form of an interferogram, and image reconstitution (from a transparency reduced in size because of the larger wavelength) by optical diffraction. The first microwave (and acoustic) holograms were recorded in 1951 before the availability of lasers. The first publicized demonstration of small-scale microwave imaging occurred in 1965 (**Fig. 8**). The object was a metal letter A with a height of approximately 7 ft (2 m), that is, 70 wavelengths, illuminated by X-band microwave radiation with a wavelength of approximately 30 mm. The hologram (approximately 10 × 10 ft or 3 × 3 m) was mapped by recording the field intensity and converting it to a small transparency for image construction by laser light. Subsequently the methods of data recording and the replacement of the optical diffraction process by digital computation transformed microwave holography into a diagnostic imaging technique in its own right.

Data recording. The replacement of the optical diffraction process by computer processing using a fast Fourier transform algorithm has important implications for the data-recording stage. Instead of obtaining the microwave interferogram analogously to the optical process, the microwave field scattered by the object is recorded directly in amplitude and phase by using a microwave receiver which compares the measured field at any point in space with a reference value. For the forward-scatter case (**Fig. 9***a*), the object (which may be semitransparent to microwaves) is illuminated from a microwave source and transmitting antenna T_x. A receiving antenna scans through known coordinates in the surface *S* and feeds the field values at each point to the receiver. Since a portion of the source energy is fed directly to the receiver by a separate reference channel (either a free space path or a waveguide), the receiver can generate the complex field values (phase and amplitude) at each point. An alternative recording geometry, the backscatter mode (Fig. 9*b*) is the usual configuration for radar systems. The transmitting and receiving antennas are either the same antenna or two antennas close together, as shown. The antennas scan as a unit over the desired surface *S*, and the complex field values are recorded. Because the illumination from the transmitting antenna also scans the object in this case, the resolution of the system is doubled in comparison with the forward-scatter case.

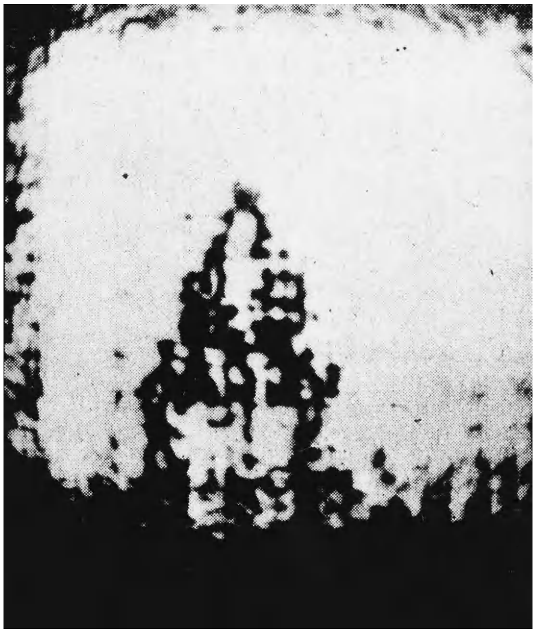


Fig. 8. Optically reconstructed image using laser light (wavelength equal to 0.6328 μm) of metal letter A 70 wavelengths high at the X band (wavelength of approximately 30 mm). (*From R. P. Dooley*, *X-band holography*, *Proc. IEEE.*, *53:1733–1755*, *1965*)

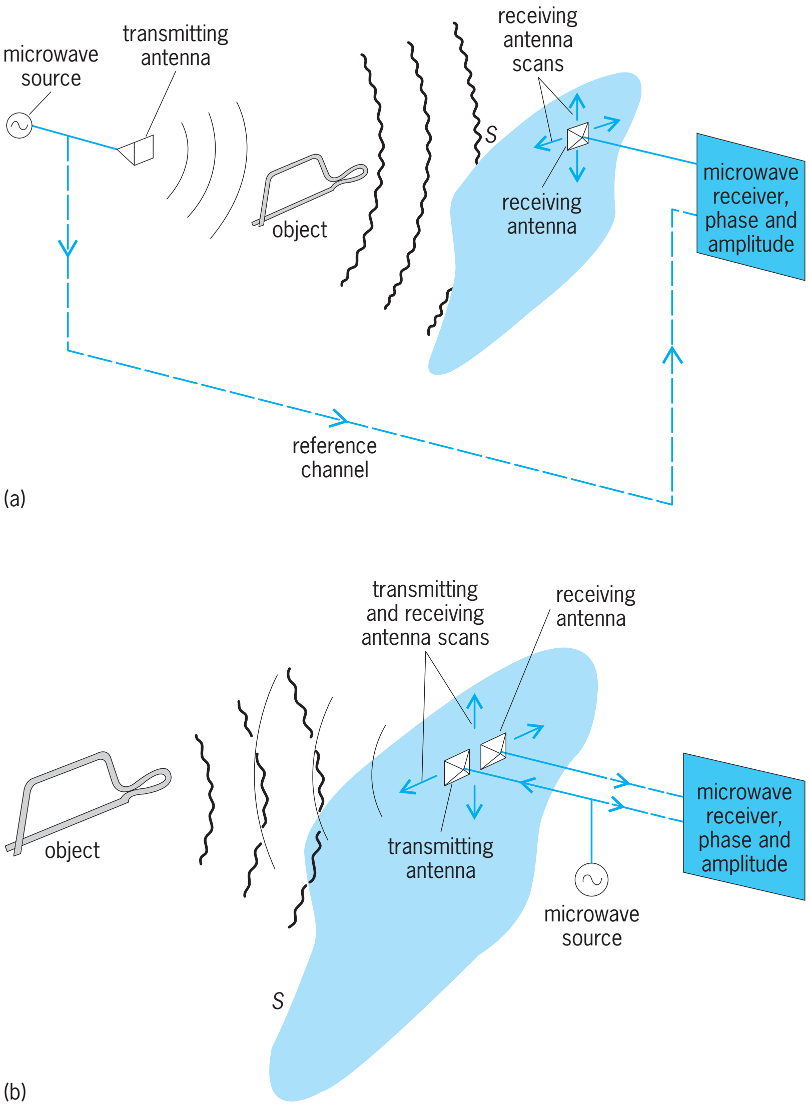


Fig. 9. **Typical scan geometries for recording complex field data. (*a*) Forward-scatter mode. (*b*) Backscatter mode.**

Computer processing. The sampled field values recorded over the surface *S* may be expressed as an array of complex numbers and are therefore suitable for computer processing. The computer algorithm is designed to reconstitute an image from the particular scan geometry used. The process can be thought of as effectively inverting the propagation process that brought the scattered waves to the surface *S*. The inversion process usually incorporates a version of the fast Fourier transform algorithm to convert the data recorded on *S* (not strictly a hologram) into the reconstructed object. The computer transfers its output to a memory and then a television monitor display. The important advantages of this digital microwave holographic process are (1) the availability of numerical field values with high accuracy and low noise; (2) the separate operations on the phase and amplitude values, and the separate display of these values; (3) the possibility of computer-observer interaction at any stage of the processing; and (4) the options of monochrome or false color display format. *See* HARMONIC ANALYZER.

Imaging applications. The efficacy of microwave holography as an imaging modality independent of optical holography is evidenced by a comparison of the microwave image of an object (**Fig. 10***a*) with the optical photograph (Fig. 10*b*). Considerable resolution of detail can be observed in an object that is only 20 wavelengths long at the microwave frequency. However, the role of microwave holography is not to mimic optical holography. Until 1979, perhaps the most useful diagnostic application of microwave holographic imaging was the metrology of large reflector antennas. The data acquisition procedure is a variant of the general forward-scatter case (Fig. 9*a*) since the object itself is scanned in both azimuth and elevation to synthesize the holographic aperture. In this arrangement (**Fig. 11***a*), the test antenna itself feeds the complex field values to the microwave receiver, and so functions simultaneously as the receiving antenna and the object. The transmitting antenna is located either on the ground, in the near field of the test antenna, or on board a synchronous satellite. The image, that is, the conventional notion of an image (Fig. 11*b*), is obtained by quantifying the amplitude distribution over the reflector, and also shows the support legs and the focal region “laboratory.” More important is the phase image (Fig. 11*c*), which corresponds to the errors in the reflector profile, that is, deviations from the ideal paraboloidal shape. Other important diagnostic information can be derived, for example, the astigmatism due to gravitational distortion which is apparent in Fig. 11*d*.

Microwave holography is also useful in applications where images of concealed structure are required. Microwave radiation penetrates a variety of dielectric media to a depth depending on the attenuation of a given wavelength in a particular medium. One such application is the mapping of subsurface pipes and cables. A scanning arrangement for this purpose (**Fig. 12***a*) uses the backscatter mode (Fig. 9*b*). The detection of the backscatter from buried pipes, which is very weak after suffering attenuation in the soil, is assisted by the polarization discrimination of the receiving antenna. The data acquisition and computer processing follow the normal procedure with compensation for microwave propagation in the soil. Plastic pipes as well as metal pipes can be imaged (Fig. 12*b*). Hence this noninvasive microwave technique has a diagnostic power greater than the normal metal detectors. *See* NONDESTRUCTIVE EVALUATION.

Microwave tomography. The major limitation of the microwave holographic techniques discussed above is that the images produced are essentially two-dimensional. This may seem surprising, given the fact that optical holography is a three-dimensional image construction process. The reason is that the microwave wavelength is so long (10^4–10^6 times that of light) that the depth of focus of the microwave hologram is prohibitive. This disadvantage is overcome by employing a tomographic mode of imaging which exploits the ability of microwaves to penetrate many materials and thereby characterize their three-dimensional structure more accurately. This development is analogous to the technique of computer-aided tomography used in x-ray scanning

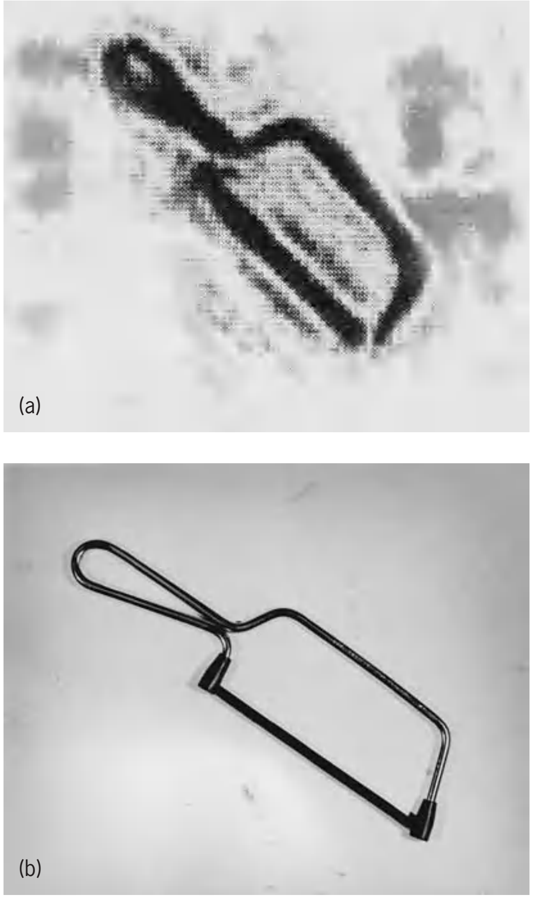


Fig. 10. Comparison of (*a*) digitally reconstructed microwave image of object 20 wavelengths long at the Q band (wavelength of 9 mm), showing reflections from surroundings, with (*b*) optical photograph of object.

systems. Microwave holographic tomography requires holograms to be recorded from different views of the object and synthesized. Again, the availability of phase imagery increases its diagnostic potential. *See* COMPUTERIZED TOMOGRAPHY. Alan P. Anderson

X-Ray Holography

Physicists and life scientists have been engaged in research that will ultimately allow three-dimensional imaging of living organisms with resolution and contrast far beyond the reach of optical microscopes. The impetus for this activity is the imminent availability of high-intensity coherent sources of electromagnetic radiation with wavelengths between 0.1 and 10 nanometers. Much of the study is concentrated on holographic imaging because it can eliminate the need for focusing elements which are difficult to fabricate with enough precision to achieve diffraction-limited resolution in the soft x-ray regime. Furthermore, several of these new sources promise extremely high intensity and subnanosecond pulses, and can circumvent the problem of killing and altering the specimen with the x-ray exposure by extracting an image from the specimen before it is obliterated. *See* X-RAY OPTICS; X-RAYS.

X-ray sources. To be suitable for holography, the x-radiation must be monochromatic and have a relatively high degree of coherence. Synchrotrons using magnetic undulators can generate a narrow band of intense radiation. Use of monochrometers and pinhole apertures can improve the coherence of this radiation at the sacrifice of intensity but with retention of sufficient intensity to image biological specimens on time scales from a few seconds to a few hours. *See* SYNCHROTRON RADIATION.

There are several promising sources. Nonlinear optical frequency multiplication techniques produce intense picosecond pulses of tunable coherent radiation, and have reached wavelengths as short as 40 nm. Similarly, multiphoton excitation can pump atoms to higher energy levels that have lasing transitions at wavelengths much shorter than the excitor laser. X-ray lasers driven by nuclear explosives and by more conventional laboratory sources are under development. X-ray and gamma-ray lasers will be inherently short-pulse, high-intensity devices because they will probably not have resonant cavities, so the radiation being amplified can make only a single passage through the active medium; and the creation and maintenance of a high density of excited atomic states of short lifetime and high quantum energy require enormous power, which

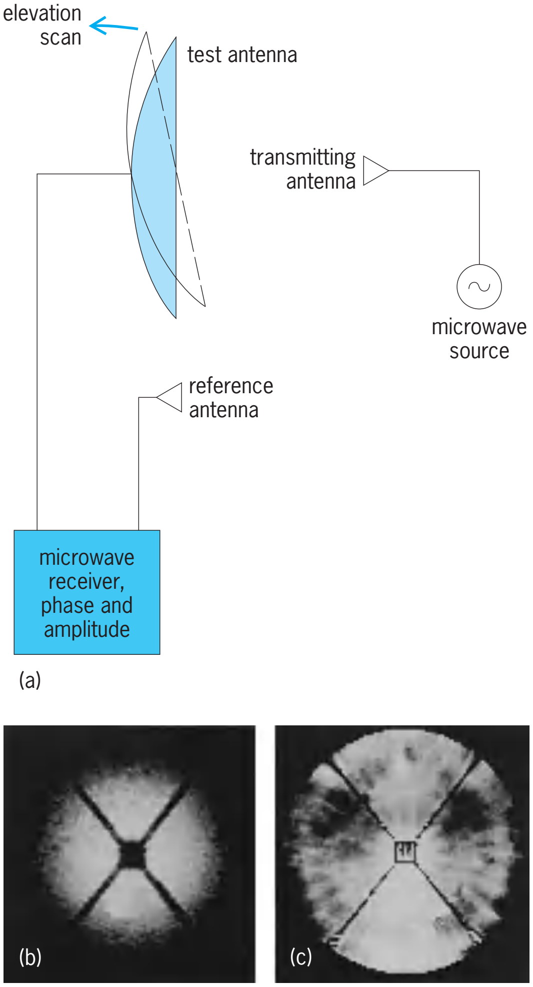


Fig. 11. Applications of microwave holographic imaging to metrology of an 82-ft-diameter (25-m) paraboloidal reflector antenna structure used in satellite communications. (*a*) Scan geometry used for data recording. (*b*) Amplitude image showing aperture illumination distribution. (*c*) Phase image showing deviations from the ideal paraboloid.

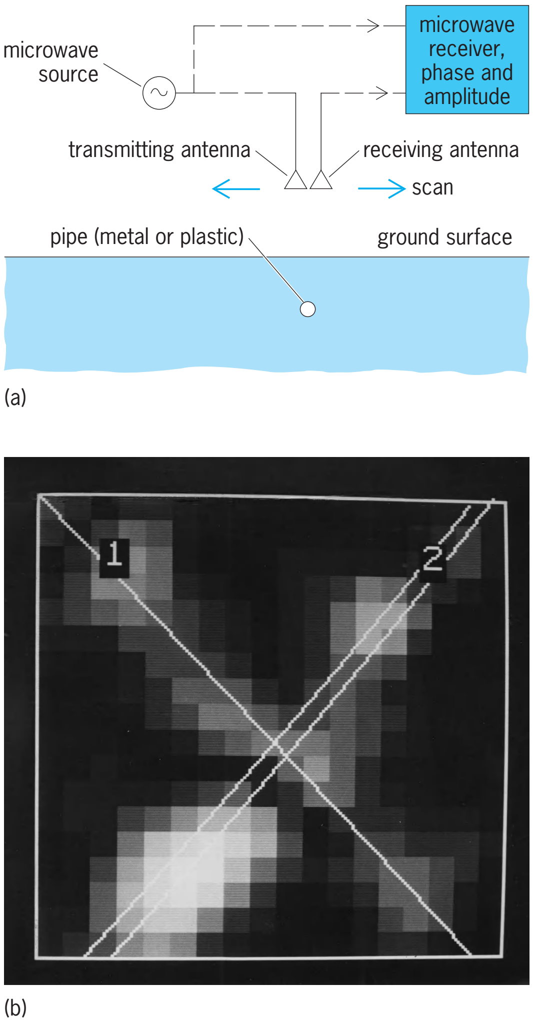


Fig. 12. Subsurface imaging. (*a*) Scan system. (*b*) Microwave amplitude image, at wavelength of 0.6 m (2 ft), of two crossed pipes whose positions are indicated by solid lines. Pipe 1 is plastic; pipe 2 is metal. Field of view is 7 × 7 ft (2 × 2 m).

terrestrial sources can supply only in the form of pulses. *See* NONLINEAR OPTICS.

Geometries. There are three principal geometries for holography (**Fig. 13**). The Fresnel transform techniques use planar reference waves and have resolution limited by the grain size of the recording medium. The on-axis (Gabor) form is inherently simple but suffers from overlap of the real and virtual images. The off-axis (Leith-Upatnieks) modification reduces the image overlap problem but requires a mirror and a broadened beam for system illumination; both forms may be difficult at x-ray wavelengths. The Fourier transform (Stroke) geometries, using curved wavefronts, achieve large fringe spacings and are therefore less sensitive to grain size.

Coherence is characterized by the effective finite length of a photon wave train in the transverse direction (spatial). Both coherence length and geometry limit the holographable volume of a specimen. For most specimens of biological interest, spatial and temporal coherence lengths of 10 micrometers to 1 nm are adequate.

Interactions of x-radiation. The interaction of x-radiation with matter is quite different from the interaction of visible light with matter. Whereas the extinction of a visible beam traversing matter is mainly due to scattering, the extinction of an x-ray beam is mainly due to absorption. X-rays can also be scattered, but usually the cross section for coherent scattering is very much smaller than for absorption. In the visible regime, holographic images are primarily formed by refraction or reflection, whereas in the x-ray regime they are dominated by diffraction. The greatest contrast in x-ray absorption between water, which composes most of the cytoplasm, and protein (or the nucleic acids) occurs between the *K* edges of oxygen and nitrogen.

Snapshot x-ray holography. Existing x-ray sources, in particular, synchrotron radiation sources, have been used to make holograms. However, they require long exposures, limiting their usefulness for research on living specimens. More coherent sources may also be developed, but those of low intensity will be similarly limited, since ionization will have decomposed molecules, modified compositions, and altered biological functions before enough radiation can be received to form a useful hologram. Snapshots are essential for x-ray holography of living specimens. Fortunately, it is likely that x-ray sources producing brief intense bursts will be developed.

With an intense pulsed coherent source (such as an x-ray laser), hydrodynamic expansion, initiated by sudden heating, rather than normal biological activity, chemical change, or thermal agitation, will limit the time during which recording of the hologram must be accomplished. Analytical expressions for the explosion of a semiopaque feature (such as a protein

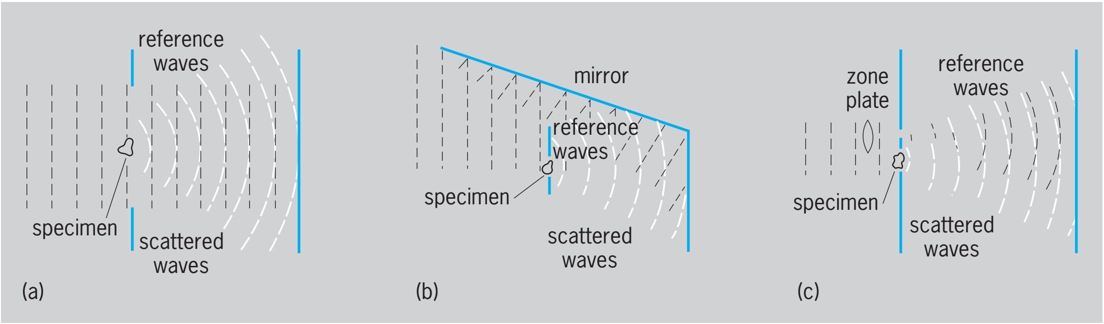


Fig. 13. Geometric configurations of x-ray holographic techniques. (*a*) On-axis Fresnel transform (Gabor holography). (*b*) Off-axis Fresnel transform (Leith-Upatnieks holography). (*c*) Planar Fourier transform (Stroke holography).

globule) are useful for estimating the radiation requirements for typical cases. They are based on the criterion that, to achieve a linear resolution δ, a specified minimum number of photons must have been coherently scattered in a volume δ^3 and that, during the exposure time Δt, no dimension of the specimen should have increased by more than δ. For most biological specimens, intensities on the order of 10^{12} W · cm^{-2} with pulse lengths on the order of 10^{-11} s will be required to obtain an x-ray hologram with resolution of 10 nm.

Recording. An x-ray hologram can be registered by radiation-induced prompt or latent chemical change, or by photoelectron emission. Photographic emulsion is unsatisfactory for Fresnel transform x-ray holography because the resolution is limited by grain size. If an electron microscope could be used to image the points of electron emission from a photocathode reference surface, time-gated holography might be possible. However, the continuous distributions in energy and in angles of emission of electrons from a photocathode preclude the formation of sharp electron-optical images, imposing a tradeoff between quantum efficiency and resolution, unless image-deblurring analysis can be applied. Photoresists (materials that lose resistance to chemical etching at points exposed to radiation) have grain sizes that approach 5 nm, which is entirely adequate for x-ray holograms with resolutions of 10 nm. To reconstruct a photoresist hologram, a transmission electron micrograph could be formed and viewed with visible laser illumination, or a transmission electron microscope could be used to scan and digitize the photoresist for analysis by computation, which can also mitigate nonlinearities that may be troublesome in optical reconstruction. *See* ELECTRON MICROSCOPE; PHOTOEMISSION; PHOTOGRAPHIC MATERIALS.

By using Fourier transform x-ray holography, it is possible to arbitrarily adjust the fringe spacing at the sacrifice of intensity and thereby record with common photographic emulsions. However, when compared with photoresists on the basis of number of quanta required to produce a developable speck, it is not clear that the greater sensitivity of photographic emulsions offers any advantage, and consequently there is no clear advantage of Fourier over Fresnel methods.

Practical considerations. The realization of x-ray holography as a practical research tool still awaits the solution of some challenging technical problems:

1. Development of sources that can generate intense coherent radiation at the precise wavelengths to optimize contrast among specimen constituents. Perhaps nonlinear mixing with tunable visible radiation will be necessary.

2. Termination of exposure within a sufficiently brief time interval and with sufficient intensity to achieve the desired resolution, as discussed above. Frequency multiplication techniques and multiphoton excitation lasers can achieve these short pulses because the optical laser driving them can be mode-locked. Corresponding schemes are difficult to envisage for x-ray or gamma-ray lasers, and their pulse lengths are likely to be much longer. A shutter or gate, somewhere in the system, that operates when full intensity is reached will be essential. *See* OPTICAL PULSES.

3. In principle, photoelectric recording could be time-gated. However, complexity and precision required of the electronics, and blurring associated with initial electron velocity distribution make this approach unattractive. On the other hand, exposure control in photoresist recording is not likely to be managed by a gate; therefore exposure control must be provided elsewhere in the system.

4. Leith-Upatnieks holography may be necessary to avoid image overlap obscuration, and this requires an x-ray mirror. A synthetic Bragg crystal may suffice, and thermal expansion, if sufficiently uniform, can provide automatically time-gated reflection.

Johndale C. Solem

Bibliography. M. F. Adams and A. P. Anderson, Synthetic aperture tomographic imaging (SAT) for microwave diagnostics, *Proc. IEEE*, 129:83–88, 1982; A. P. Anderson, Microwave holography, *Proc. IEEE*, 124:946–962, 1977; S. A. Benton, Hologram reconstructions with extended incoherent sources, *J. Opt. Soc. Amer.*, 59:1545–1546A, 1969; B. R. Brown and A. W. Lohmann, Complex spatial filtering with binary masks, *Appl. Opt.*, 5:967–969, 1966; H. Dammann, Blazed synthetic phase-only holograms, *Optik*, 31:95–104, 1970; Y. Denisyuk, Photographic reconstruction of the optical properties of an object in its own scattered field, *Sov. Phys. Doklady*, 7:543–545, 1962; J. W. Goodman, *Introduction to Fourier Optics*, 3d ed., 2005; P. Hariharan, *Basics of Holography*, 2002; J. E. Kasper and S. A. Feller, *The Complete Book of Holograms: How They Work and How To Make Them*, 1987, reprint 2001; T. Kreis, *Handbook of Holographic Interferometry*, 2005; S. H. Lee (ed.), *Selected Papers on Computer-Generated Holography and Diffractive Optics*, 1992; E. N. Leith and J. Upatnieks, Photography by laser, *Sci. Amer.*, 212(6):24–35, 1965; L. B. Lesem, P. M. Hirsch, and J. A. Jordan, Jr., The Kinoform: A new wavefront reconstruction device, *IBM J Res. Dev.*, 13:150–155, 1969; A. W. Lohmann, Optische Einseitenbanduebertragung angewandt auf das Gabor-Mikroskop (Optical single-sideband transmission applied to the Gabor microscope), *Opt. Acta*, 3:97–99, 1956; J. Ludman, H. J. Caufield, and J. Riccobono (eds.), *Holography for the New Millennium*, 2002; A. J. MacGovern and J. C. Wyant, Computer generated holograms for testing optical elements, *Appl. Opt.*, 10:619–624, 1971; G. Saxby, *Practical Holography*, 3d ed., 2003; U. Schnars and W. Jüptner, *Digital Holography*, 2005; J. Solem and G. Baldwin, Microholography of living organisms, *Science*, 218:229–235, 1982.

Holostei

An unranked group of the fish subclass Neopterygii consisting of several fossil orders and the extant orders Lepisosteiformes and Amiiformes. The holosteans are descended from the older

Chondrostei and in turn are ancestral to the great mass of modern bony fishes, the Teleostei. It is not certain whether holosteans evolved from a single stock or multiple stocks of the Chondrostei. *See* TELEOSTEI.

Phylogeny. Holosteans made their first appearance in the Upper Permian as the order Semionotiformes; three additional orders arose in the Triassic Period, and the fifth and last order, the Aspidorhynchiformes, evolved in the Middle Jurassic. In the Jurassic and Lower Cretaceous, holosteans dominated actinopterygian fish life, but by the Late Cretaceous they had been largely replaced by teleosts. The specialized Pycnodontiformes persisted until the Eocene, but the spindle-shaped, predacious Aspidorhynchiformes and the Pholidophoriformes, which are likely ancestors of the Teleostei, died out in the Cretaceous. Fragmentary remnants of two large orders persist to the present time as survivors of the Mesozoic: Lepisosteiformes, as the North American and Middle American gars (family Lepisosteidae), and Amiiformes, as a single species, the bowfin (family Amiidae) of eastern North America. Understandably, both groups have been intensively studied by biologists in search of clues to the life of the past.

Morphology. Holosteans, although highly varied in body form, were structurally as well as temporally intermediate between chondrosteans and teleosts, to which group they passed on substantial advances. Mouthparts were improved by horizontal suspension of the hyomandibular from the skull, a more forward positioning of the angle of the gape, and the development of a strong coronoid process on the mandible. The maxilla was freed posteriorly, and the entire feeding mechanism became more mobile and was strengthened, thus permitting diversification in food habits, though most holosteans were predacious. The caudal fin is typically abbreviate heterocercal, with the posterior vertebrae upturned but not forming a long upper lobe. Typically the anterior upturned centra each support a single, slender hypural. Dorsal and anal fin rays are strengthened and reduced in number to approximate serial equivalence with the internal supports. In early forms scales are often thick and rhomboidal, as in chondrosteans, but in certain advanced types are thin and rounded; they retain an enamellike outer layer (ganoine) that is lost in all but the earliest teleosts. In living holosteans the swim bladder is highly vascularized, and auxiliary aerial respiration is possible, a sometimes essential faculty in oxygen-poor waters of swamps. *See* ACTINOPTERYGII; AMIIFORMES; ASPIDORHYNCHIFORMES; PYCNODONTIFORMES; SEMIONOTIFORMES.

Reeve M. Bailey

Holothuroidea

A class of Echinozoa characterized by a cylindrical body and smooth leathery skin, and known as sea cucumbers. There are no arms, but a ring of five or more tentacles may surround the mouth, which is usually at one end of the body. There are no pedicellariae. Tube feet may be present or lacking. There are no ambulacral grooves, although they are represented by internal epineural canals overlying the radial nerves. E. Deichmann (1957) regards them as the most aberrant group of extant echinoderms. *See* ECHINOZOA.

Holothurians resemble worms because the pentamerous symmetry is largely concealed by a secondary bilateral symmetry, and the general absence of external spines distinguishes them from the other extant echinoderms. They tend to rest on one side, so that the axis of radial symmetry becomes horizontal. This habit leads to the differentiation of an upper (dorsal) surface and a lower (ventral) one. The dorsal side corresponds to the interradius which contains the madreporite, and therefore the ventral side is one of the radii. Each radius runs from the anterior to the posterior end. If tube feet are developed, their disposition indicates the radii (**Fig. 1**).

The 1100 living species have been grouped in 170 genera arranged in six orders: the Dendrochirotida, Dactylochirotida, Aspidochirotida, Elasipodida, Molpadida, and Apodida. Species range in size from 1.2-in. (3-cm) body length up to 5 ft (1.5 m); the largest types are tropical Synaptidae. Colors vary widely; the most brilliant colors are found among the Synaptidae. Yellow, red, violet, and fawn tints occur, but many species are somber shades or black. *See* ECHINODERMATA.

Relation to humans. Some holothurians are esteemed as food, in particular about 20 species of the genera *Stichopus* and *Holothuria*. These are fished mainly in the Indian and Pacific oceans. The animals are boiled for 20 min in seawater, smoked, and dried. The product is then called chekin, trepang, or bêchede-mer. The annual export trade to the Far East and Mediterranean countries exceeds 10,000 tons (9072 metric tons). In Italy also, fishermen take local species, but these are not esteemed.

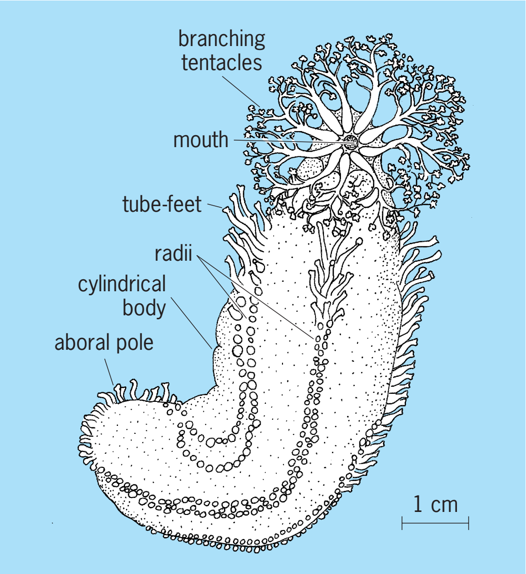


Fig. 1. ***Cucumaria*, a representative holothurian.**

A few holothurians are poisonous, especially to fish. In the Pacific some species of *Actinopyga* and *Holothuria* yield a secretion used by the islanders as an aid in fishing rock pools, in the same way as rotenone.

Ecology. The Dendrochirotida and Apodida have more restricted ranges than the other orders, which are believed to be global. Holothurians occur in all seas, from low-tide level down to the greatest depths explored. The Vitiaz expedition in 1957 took a holothurian from the Tonga Trench at a depth of $6^1/_2$ mi (10,415 m). At depths below $5^1/_2$ mi (8850 m) holothurians comprise 90% of the total mass of living matter, the rest being mainly starfishes. Two pelagic genera are known.

Parasites include protozoa, flatworms, nematodes, and annelids. Some crabs, such as *Pinnotheres*, inhabit the cloaca or respiratory trees, and another, *Lissocarcinus*, lurks between the tentacles. Gastropods such as *Entoconcha* and *Entocolax* bore into the skin, body cavity, or foregut. A fish, *Fierasfer*, inhabits the cloaca of large Aspidochirotida.

Skeleton. This structure usually comprises no more than a ring of 10 (or 5) calcareous plates around the esophagus, and numerous small platelets or spiculae scattered in the skin. In Psolidae the skin plates are large and overlap like scales, but they do not form radial and interradial series. The skin platelets may assume distinctive forms, such as anchors, wheels, or other shapes, and may be useful as taxonomic characters.

Muscular system. The muscular system is well developed, and the body can assume a variety of shapes. The chief muscles are five radial longitudinal bands in the body wall; outside these are transverse fibers. Some species have pharyngeal retractor muscles which invert the anterior part of the body. This feature is used as an aid to classification.

Alimentary system. The viscera lie in the coelom, which contains a fluid. The mouth leads into a short esophagus, which connects to the long and usually looped intestine. Its front part may be differentiated as a stomach, and the posterior loop serves as a rectum, opening at the cloaca (**Fig. 2**).

Most holothurians are mud swallowers, living on the bacteria and other organic material present in the substrate. The Dendrochirotida, however, feed on plankton, which is caught on their sticky tentacles and transferred to the mouth. Some Aspidochirotida have paired cloacal glands, the Cuvierian organs, which extrude a mass of sticky threads to trap small animals and to repel predators.

Respiration. Respiration occurs partly in the skin and tube feet and sometimes in tubular branches of the gut, the respiratory trees. Rhythmic contractions of the cloaca cause seawater to flow to and fro in them.

Hydrocoele. Although poorly developed in the Molpadida, and vestigial in the Apodida, the water-vascular system is typical in the other three orders. The one or more madreporites lie in the dorsal interradius, internal in nearly all cases, and open into the coelom. The single stone canal which arises from them runs downward to the ring vessel around the esophagus. A single polian vesicle lies in one interradius (Fig. 2). The radial vessels run forward to the base of the tentacle ring and then backward in the body wall, following the radii. The oral tentacles represent specialized tube feet, and may have ampullae.

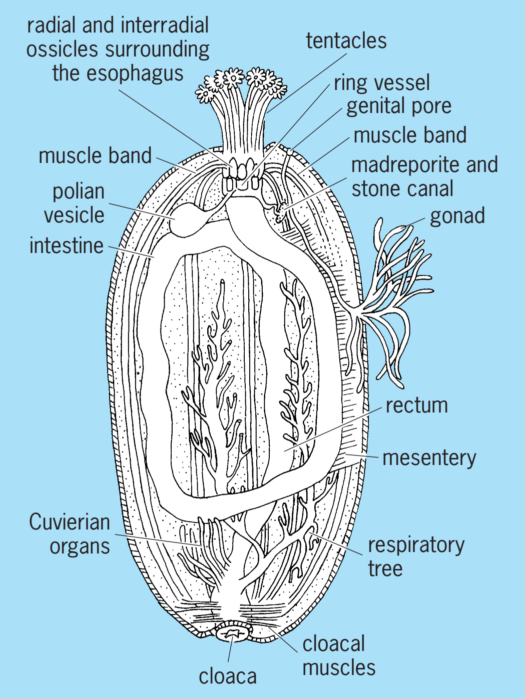


Fig. 2. Holothurian anatomy.

Nervous system. This system is typical for the phylum, following the pattern of the water-vascular system, lying between the canals and the body wall. The whole surface of the body is sensitive to light, but some Synaptidae have in addition an eyespot at the base of the tentacles. These are analogous to the eyespots of starfishes and, like them, do not form an image, but are able to detect shadow movements and thus indicate the presence of potential enemies or prey.

Life history. Holothurians seem to attain sexual maturity after about 3 years and continue to grow for several more years. The life span may therefore be longer than in sea stars and sea urchins, perhaps 10 years.

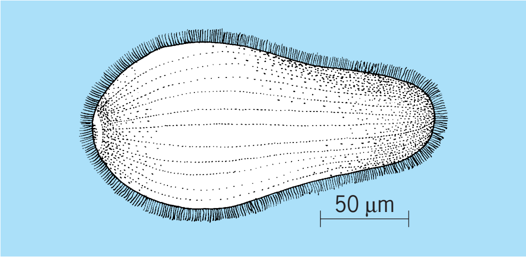
Reproductive system. The reproductive system comprises either one or two gonads, with their ducts. The gonopore is usually anterodorsal (Fig. 2). The sexes are usually distinct. Hermaphroditism is common in the Apoda, but individuals are either male or female at any given time, so self-fertilization does not occur. The life history may include a free-swimming larva (auricularia). Many cold-water species incubate their young in various ways. *Leptosynapta minuta* carries the developing embryos in the coelom until they escape by way of the cloaca. Some species of *Psolus* carry the young in dorsal pockets in the skin, covered by calcareous plates. Others shelter the eggs and young by lying on them. A few species, such as *Cucumaria planci*, habitually reproduce asexually by spontaneous transverse fission. Most species can regenerate lost organs. Autoevisceration is a peculiar protective habit which is often displayed in response to interference. The whole gut, the gonads, and respiratory trees are ejected through the cloaca and cast off. Regeneration of the lost structures occurs from the torn mesenteries, and requires several months. During regeneration the animal cannot feed; instead, the muscles are slowly digested in place. Howard B. Fell

Holotrichia

A major subclass of the class Ciliatea. These protozoa have a fairly uniform body ciliation, as the name implies. Separate articles appear on the groups listed in the following classification:

Subclass Holotrichia
  Order: Gymnostomatida
    Trichostomatida
    Chonotrichida
    Apostomatida
    Astomatida
    Hymenostomatida
    Thigmotrichida

The cilia are typically arranged in longitudinal rows over the body, although scattered exceptions exist. Cirri are absent, and buccal membranelles are absent or inconspicuous, except in some members of two groups. A mouth is often, although not always, present. Ciliary organelles associated with it are generally not conspicuous. Trichocysts are present in species belonging to a number of families.



***Prorodon*, a primitive holotrich.**

The seven orders contain several thousand species and include groups of, presumably, the most primitive ciliates living today. The prototype of the Holotrichia is exemplified by the form portrayed in the **illustration**. *See* CILIATEA. John O. Corliss

Homalozoa

A subphylum of echinoderms, made up of members having a flattened theca or body lacking pentameral symmetry. Homalozoans (also called carpoids) include four extinct classes of relatively uncommon primitive echinoderms ranging in age from the Early or Middle Cambrian to the Late Carboniferous. Homalozoans have a flattened, asymmetrical to bilaterally symmetrical theca often composed of a marginal frame of large elongate plates surrounding top and bottom central areas that had numerous smaller plates and were probably flexible. One homalozoan class (Ctenocystoidea) has no attached appendages; two classes (Stylophora and Homostelea) have a single plated armlike or taillike appendage used for feeding or swimming; and one class (Homoiostelea) has two plated appendages at opposite ends of the theca, one used for feeding, the other for swimming. The mouth is located either on one edge of the theca or at the base of the armlike appendage, while the anal pyramid is usually at the opposite margin or corner of the theca. All homalozoans were apparently mobile, benthic, detritus or suspension feeders that had adopted a flatfish way of life. Because they have a typical echinoderm skeleton with single-crystal, multiporous, sutured, calcite plates, most researchers consider homalozoans to be true echinoderms. However, they seem only distantly related to other echinoderms that have well-developed pentameral symmetry. It has been argued that one class of homalozoans (stylophorans or calcichordates) is intermediate between echinoderms and several groups of chordates, thus making the class directly ancestral to the vertebrates. However, this proposal has not been widely accepted by other echinoderm or vertebrate paleontologists. *See* CARPOIDS; ECHINODERMATA. James Sprinkle

Bibliography. R. P. S. Jefferies, *The Ancestry of the Vertebrates*, 1986; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. S, 1968.

Homeostasis

Either the relatively constant conditions within organisms or the physiological processes by which such conditions are maintained in the face of external variation.

Similar homeostatic controls are used to keep factors such as temperature and blood pressure nearly constant despite changes in an organism's activity level or surroundings. Such servosystems, which are of wide technological use as well, operate by detecting changes in the variable that the system is designed to hold constant and initiating some action that offsets any change. All incorporate a sensor within the system that responds when the actual condition differs from the desired one, a device to ensure that any action taken will reduce the difference between actual and desired, and an effector to take the needed action as directed. The crucial aspect is that information is fed back from effector to sensor and action is taken to reduce any imbalance—hence the term negative feedback.

Biological mechanisms. Blood pressure is, at least on a moment-to-moment basis, regulated by a system for which the sensors are stretch-sensitive cells located in the neck arteries that carry blood from heart to brain. An increase in blood pressure triggers sensor activity; their signal passes to the brain; and, in turn, the nerve supplying the heart (the vagus) is stimulated to release a chemical (acetylcholine) that causes the heart to beat more slowly—which decreases blood pressure.

The volume of the blood is subject to similar regulation. Fluid (mainly plasma) moves between the capillaries and the intercellular fluid in response to changes in pressure in the capillaries. A decrease in blood volume is detected by sensors at the base of the brain; the brain stimulates secretion of substances that cause contraction of tiny muscles surrounding the blood vessels that lead into the capillaries. The resulting arteriolar constriction reduces the flow of blood to, and the pressure within, the capillaries, so fluid moves from intercellular space into capillaries, thus restoring overall blood volume.

Body temperature in mammals is regulated by a sensor that consists of cells within the hypothalamus of the brain. If that area is experimentally heated, the rest of the body cools off. Several effectors are involved, which vary among animals. These include increasing heat production through nonspecific muscle activity such as shivering; increasing heat loss through sweating, panting, and opening more blood vessels in the skin (vasodilation); and decreasing heat loss through thickening of fur (piloerection) and curling up. Humans (but not pigs, cats, or dogs) sweat, but they retain only a vestige of piloerection ("goose flesh"). *See* THERMOREGULATION.

Regulation using negative-feedback servosystems occurs in more localized neuromuscular systems as well. When stimulated by nerves, muscles pull harder—how much shorter the muscle becomes depends on the load it is pulling against. But one often makes a muscle shorten by a fixed and load-independent amount. It is done with sensors within the muscle that effectively monitor its length. Extra load (for example, a book laid on an outstretched arm) stretches muscles such as the biceps; that stretch is detected and stimulates the brain to send more frequent nerve impulses to the muscle, restoring its original length—but at the higher tension needed for its greater load. Similar machinery permits a person standing with slightly bent knees to lift one leg off the ground without collapsing—the tensions of the extensor muscles of the opposite leg are doubled to offset the doubling of their load. *See* MUSCLE.

While the homeostatic mechanisms first described involved the neural and endocrine systems of mammals, it is clear that such arrangements pervade systems from genes to biological communities, and that they are used by the simplest and the most complex organisms. If a bacterium such as *Escherichia coli* is grown in a culture medium that lacks the amino acid tryptophan, it reacts to that absence by turning on the appropriate genes and making the enzymes that it needs to synthesize tryptophan. If the same bacterium is transferred to a medium in which lactose rather than glucose is the sole energy source, it begins to make the enzymes necessary to live on lactose. In either case the bacterium does what is necessary to maintain the level of some substance within itself.

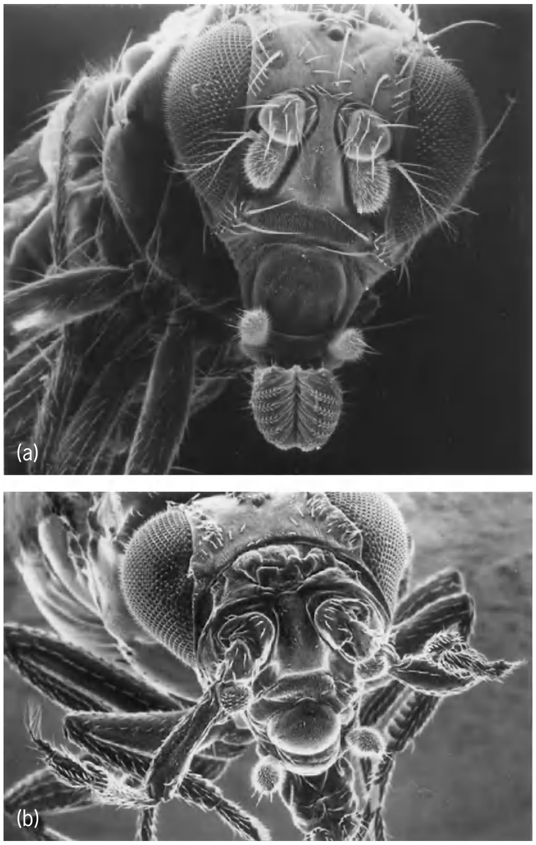
Other mechanisms. Servosystems of the kind described above are limited in their response speed by the necessity for the external change to affect the internal sensor. In practice, organisms commonly overlay such systems with anticipatory devices. The initial heat-conserving response to a sudden chill is triggered by information from cold-sensitive receptors in the skin. Since alteration in blood flow to the skin is part of the response to cold and since skin temperature is not closely regulated, these receptors cannot function as the primary source of the feedback signal—but they do initiate a rapid response. Similarly, when given access to water, a mammal deprived of water drinks rapidly. The animal stops drinking when sufficient water has been consumed to restore its normal content. But that happens well before the main detectors for water content (detectors that monitor blood composition) have been affected.

Organisms of every kind develop, mature, and even shift physiological states periodically—between day and night, with seasons, or as internal rhythms. Thus organisms cannot be considered constant except over short periods. However, all such changes appear to involve the same basic sensing of the results of the past activity of the system and the adjusting of future activity in response to that information. Development of an organism from a fertilized egg is far from a direct implementation of a genetic program; probably no program could anticipate all the variation in the external context in which an organism must somehow successfully develop. *See* BIOLOGICAL CLOCKS; ENDOCRINE MECHANISMS; NERVOUS SYSTEM (VERTEBRATE); SERVOMECHANISM. Steven Vogel

Bibliography. C. A. R. Boyd and D. Noble (eds.), *The Logic of Life: The Challenge of Integrative Physiology*, 1993; D. D. Chiras, *Human Biology: Health, Homeostasis and the Environment*, 3d ed., 1999; S. Vogel, *Vital Circuits*, 1993.

Homeotic (Hox) genes

The formation of a normal plant or animal body structure or organ in place of another at an abnormal site. Examples of homeosis (also called homeotic transformation) are most obvious in insect appendages, where an appendage that is characteristic of one segment, for example the antennae on an insect head segment, are transformed into insect legs that normally develop only on trunk segments (see **illus**.). Similar examples of homeotic transformations can also occasionally be found in vertebrates where lumbar vertebrae are transformed into thoracic vertebrae which then extend into rib processes, or in floral organs where petals are transformed into sepals. Homeotic transformations rarely occur in nature in living organisms, and are due to genetic defects in a class of proteins called homeotic proteins, the products of homeotic genes. In addition, homeotic



Scanning electron micrographs of the head of *Drosophila melanogaster*. (*a*) Normal individual; the antennae are the small, paired, bulbous structures between the large, faceted eyes. (*b*) An *antennapedia* mutant individual; leglike appendages replace the normally small, bulbous antennae.

transformations are due to the accidental or deliberate manipulation of homeotic gene expression so that homeotic proteins are produced in the wrong place or at the wrong time in developing plants and animals.

Animals. Homeotic transformation is best understood in animals, particularly in the fruit fly *Drosophila*. During early embryogenesis, the anterior-posterior axis of most animals is divided into segmental units or metameres. These metameres may be thought of as the basic building blocks of body form, usually incorporating large numbers of individual cells. During normal development, metameric building blocks of cells require additional instructions to indicate whether they are fated to develop as head, thoracic, or abdominal regions of the body. Thus, the cells of different metameres express unique combinations of the different homeotic proteins that act as developmental switches that assign different fates to different members of this serial array of embryonic developmental fields. For example, if a combination of homeotic proteins that usually assigns abdominal development is produced throughout the entire *Drosophila* embryo, the result is an animal that consists of reiterated series of abdominal segments (and the animal rapidly dies for obvious reasons) in place of normal head and thoracic segments. *See* FATE MAPS (EMBRYOLOGY).

There are eight *Drosophila* homeotic genes, which are specialized to assign head, thoracic, or abdominal development. These genes map in clusters on chromosome 3 of the fly in an array where the order of the genes in the cluster mimicks the order of the metameres in the body they control, with the head genes on the left, the thoracic genes in the middle, and the abdominal genes on the right. Each of these genes is similar at the molecular level in that they all contain a homeobox DNA sequence, which specifies a similar but distinct homeodomain of 60 amino acids in the different homeotic proteins. The homeodomain allows homeotic proteins to bind to the regulatory sequences of many other genes, and the regulation of these other genes presumably explains the global effects that homeotic proteins can induce during development.

Most animals use a similar genetic system to control which structures will develop in which positions on the head-tail axis of the body. Mice and humans have four clusters of homeobox-containing genes (called Hox, for homeobox genes) that are very similar to the *Drosophila* homeotic gene clusters. The order of the genes in the fly versus mammal is conserved, and the mammalian genes express homeodomain proteins that are produced in unique combinations on the developing head-tail axis of the embryos.

The common ancestor to present-day fruit flies and humans (probably a wormlike creature that existed 600 million to 1 billion years ago) probably used a cluster of proto-homeotic genes to assign head, trunk, and tail regional identities to embryonic cells of the body plan. Apparently, after this system evolved, it proved so useful (or so difficult to eliminate) for the animals with it that the system is still in use in most or all present-day animals to assign embryonic axial positional identities. In fact, at an early stage of embryogenesis (but not the earliest stages), invertebrate and vertebrate embryos are quite similar in at least these developmental control molecules.

Plants. A flower consists of four organ types arranged in whorls which can be thought of as developmentally similar to, though not evolutionarily homologous to, animal metameres. From innermost to outermost, these whorls are carpels, stamens, petals, and sepals. Plant homeotic genes that are involved in flower development are not structurally similar to animal homeotic genes. Many of the plant homeotic genes belong to the MADS-box gene family, which shares some important characteristics with the animal homeotic genes. For example, different floral MADS-box homeotic proteins are produced early in flower development only in a subset of the whorls of the floral primordia, and the combination of floral homeotic proteins determines whether cells in a certain whorl will develop as petal or carpel. Another shared characteristic is that the MADS-box proteins are also deoxyribonucleic acid-binding proteins that regulate the expression of other genes, and so produce their effects on the development of the flower by modulating which of a number of other genes are active or inactive in various subregions of the developing flower. *See* CELL DIFFERENTIATION; DEVELOPMENTAL BIOLOGY; DEVELOPMENTAL GENETICS; MUTATION. William McGinnis

Bibliography. W. Bateson, *Materials for the Study of Variation*, 1894; T. Jack, L. L. Brockman, and E. M. Meyerowitz, The homeotic gene APETALA3 of *Arabidopsis thaliana* encodes a MADS box and is expressed in petals and stamens, *Cell*, 68:693–697, 1992; W. McGinnis and R. Krumlauf, Homeobox genes and axial patterning, *Cell*, 68:283–302, 1992.

Homing

A process of navigation by which a destination is approached by keeping some navigation parameter constant. In its early uses, the most commonly chosen parameter was the relative bearing from the vehicle to destination as determined from a signal emitted at or near the destination point. The vehicle then steers to travel in the direction of its destination. The signal can be of many forms, ranging from a visual image to a radio wave or even an odor. This simple form of homing requires minimal on-board equipment, but the path taken by the vehicle over the Earth's surface is influenced by vehicle drift due to winds, currents, or other causes. *See* DIRECTION-FINDING EQUIPMENT.

Use of navigation aids. A higher level of homing is available through certain radio navigation aids (navaids), such as the very high frequency omnidirectional range (VOR) and the aircraft instrument landing system (ILS). While the signals from these aids define a path in space that can be followed by the user, the signals themselves do not indicate the courses that should be selected to remain on the path. In the case of VOR, the signals define the azimuth to or from the station, with the user being required to adjust the vehicle heading to acquire the desired path and to compensate for any drift in order to remain on the path. In common usage, the user specifies the azimuthal radial to be followed, and then the user equipment presents indications of any deviations from the desired path. When treated properly, the ILS signals define a precise fixed path in space, which the user equipment processes to yield deviation indications in the lateral and vertical directions. These aids require only limited on-board databases for effective use. *See* ELECTRONIC NAVIGATION SYSTEMS; INSTRUMENT LANDING SYSTEM (ILS).

Direct-to paths. The advent of affordable on-board digital processing accompanied by extensive geographical databases has resulted in supporting direct-to paths from present position to destinations that need not radiate a signal. These paths can be defined to the precision supportable by the navigation aids in use, and modern equipment generally presents the user with guidance indications to acquire and remain on the desired path. The user interface is essentially indistinguishable from that presented by classic homing devices. Almost any geographically referenced navigation aid can be used to support this technique, but any errors in the database are reflected directly into destination or path errors. There is some disagreement as to whether traveling a direct path to a destination that does not radiate distinctive signals is a true homing technique.

The navigation aids used most for establishing direct-to paths include VOR/DME (distance-measuring equipment), DME/DME, Loran-C, and the Global Positioning System (GPS). GPS has the advantage that it is the only global, all-weather precision navigation aid that can be used for all phases of navigation. *See* DISTANCE-MEASURING EQUIPMENT; LORAN; SATELLITE NAVIGATION SYSTEMS.

This form of navigation requires that the geographical location of the destination be known to the user. The requirement rules out use of the technique to home on vehicles in motion unless the location in time of the vehicle is also known to the user.

Vehicle-to-vehicle homing. One of the most useful navigation features developed after the advent of radionavigation is vehicle-to-vehicle homing, where a vehicle homes on signals radiated by another vehicle. This is especially valuable in rescue operations. For example, when a ship responds to an emergency distress signal from a foundering ship, it uses its radio direction finder to determine the direction of the transmission from the ship in distress, and then can proceed to the location of the ship. Additionally, lifeboats are equipped with special rescue beacons to furnish a transmission on which the rescue vessel can home.

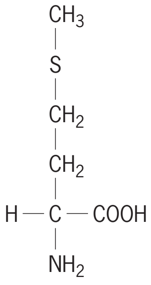
In the United States, aircraft are required to carry emergency locator transmitters (ELTs). These small transmitters, operating at 121.5 MHz for civil and 243 MHz for military aircraft, automatically turn on when the aircraft crashes. Search-and-rescue teams, both airborne and ground based, can then home on the crash site by using direction-finding techniques. For ships, there is a comparable location transmitter, the emergency position indicating radio beacon (EPIRB). These transmitters are augmented by a satellite-based search-and-rescue (SAR) network named SARSAT, which operates at 121.5, 243, and 406 MHz. Retransmission of the original ELT or EPIRB transmissions allows ground stations to process the data from these transmissions to estimate approximate locations of these transmitters. This can allow direction of search vehicles to the area where they can home in on the transmitters directly, thus reducing search times greatly. There is a Russian counterpart to SARSAT called COSPAS. SARSAT involves placing transponders in National Oceanic and Atmospheric Administration (NOAA) polar-orbiting satellites and Geostationary Operational Environmental Satellites (GOES). At present, signals intercepted by the GOES satellites are used for alerting and identification of 406 MHz signals only; they cannot be used for position fixing.

Military missions. The most serious military homing mode is that which enables a missile to home on any source of radiation. Typical sources are radio transmitters of all kinds, including radar and navigation aids, and the infrared exhausts of jet engines, tanks, and ships. Countermeasures against such activity include the use of short bursts of transmission, frequency hopping, high physical mobility, and decoys and chaff to attempt to attract the missile away from its intended target. *See* ELECTRONIC WARFARE; GUIDED MISSILE; MISSILE. Eugene O. Frye

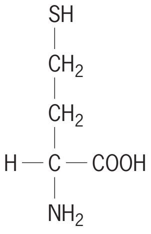
Bibliography. *Air Navigation: Homing*, U.S. Air Force Manual 51–40, 1954; N. Bowditch, *The American Practical Navigator*, 2 vols., reprint 2002; M. Kayton and W. Fried, *Avionics Navigation Systems*, 2d ed., 1997; U.S. Department of Transportation, *Federal Radionavigation Plan*, biennially.

Homocysteine

A sulfur-containing amino acid that is structurally and metabolically related to the essential amino acid methionine (see structures). Discovered by Vincent du Vigneaud in 1932, homocysteine is found in both prokaryotic and eukaryotic organisms. It is a sensitive marker of nutritional status, vitamin deficiency, and kidney function.



Methionine

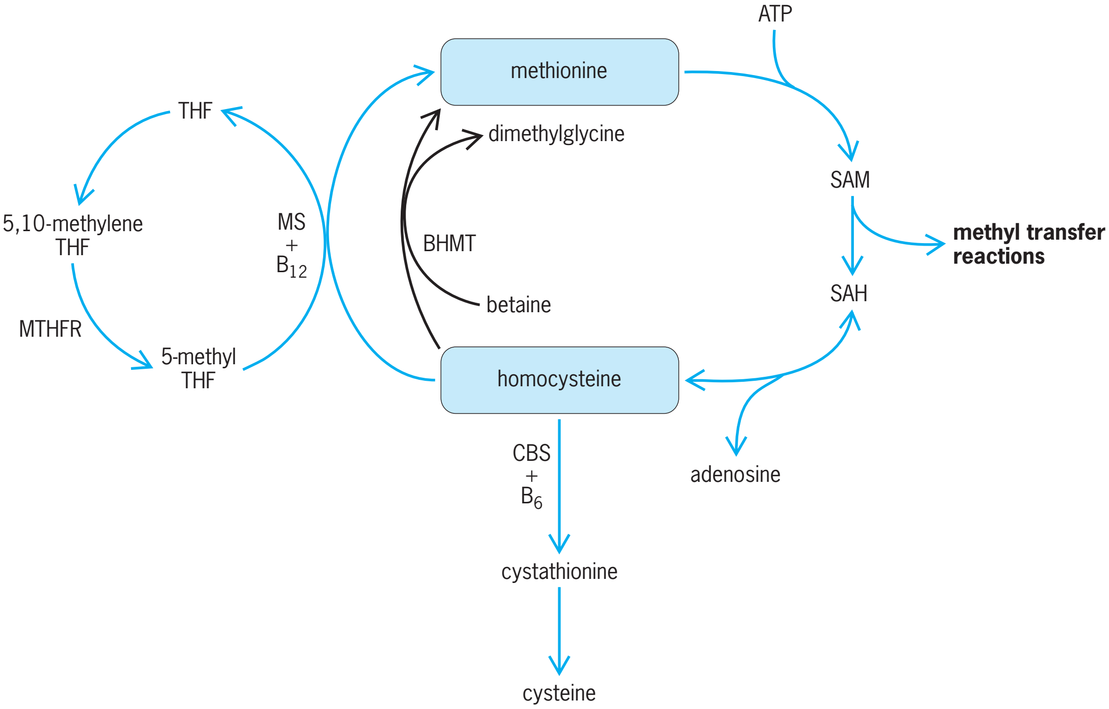


Homocysteine

It has been known since the 1960s that elevated levels of homocysteine and homocystine (the disulfide form of homocysteine) are found in the blood and urine of children with certain hereditary diseases of methionine metabolism. Clinical features of these rare diseases include skeletal deformities, abnormalities of the lens of the eye, mental retardation, abnormal blood clots, and premature death. Recent interest in homocysteine has been stimulated by emerging evidence that elevated blood levels of homocysteine are linked to coronary heart disease, stroke, Alzheimer's disease, and birth defects.

Homocysteine metabolism. Unlike most amino acids, homocysteine is not incorporated into polypeptide chains during protein synthesis but functions as a key intermediate in the metabolism of methionine (see **illustration**). Methionine is derived from dietary protein and is the direct precursor of the ubiquitous methyl donor, *S*-adenosylmethionine (SAM). By donating one-carbon methyl groups to hormones, neurotransmitters, nucleic acids, phospholipids, proteins, and other substrates, SAM serves an essential function in mammalian homeostasis. Homocysteine is produced as a by-product of SAM-dependent methyl transfer reactions.

Homocysteine occupies a critical branch point in the methionine metabolic cycle. Depending on the metabolic need for methionine and SAM, homocysteine either can serve as a substrate for the regeneration of methionine or can be diverted from the methionine cycle to produce cystathionine and, ultimately, cysteine (see illustration). The major



Homocysteine metabolism. CBS, cystathionine ß-synthase; MTHFR, 5,10-methylene tetrahydrofolate reductase; MS, methionine synthase; BHMT, betaine:homocysteine methyltransferase; SAM, *S*-adenosylmethionine; SAH, *S*-adenosylhomocysteine; 5-methyl THF, 5-methyltetrahydrofolate; 5,10-methylene THF, 5,10-methylene tetrahydrofolate; B_6, vitamin B_6; B_{12}, vitamin B_{12}. Black arrows indicate an alternate pathway that does not require B vitamins.

pathway for the regeneration of methionine from homocysteine in most tissues is catalyzed by the enzyme methionine synthase (MS). The MS reaction requires two B vitamins, methylcobalamin (a form of vitamin B_{12}) and 5-methyl tetrahydrofolate (a form of folate). 5-Methyl tetrahydrofolate, in turn, is produced by the enzyme 5,10-methylene tetrahydrofolate reductase (MTHFR). An alternative pathway for the regeneration of methionine from homocysteine is catalyzed by the enzyme betaine:homocysteine methyltransferase (BHMT). This enzyme is found mainly in liver and kidney and does not require B vitamins.

The rate-limiting reaction in the conversion of homocysteine to cysteine is catalyzed by the enzyme cystathionine β-synthase (CBS), which requires pyridoxal phosphate (a form of vitamin B_6). Thus, the level of homocysteine in a given tissue is determined by the amount of methionine in the diet, the bioavailability of B vitamins, and the relative activities of the enzymes MS, MTHFR, BHMT, and CBS. *See* AMINO ACIDS.

Hyperhomocysteinemia. Homocysteine contains a free thiol (SH) group, but only a small fraction (<2%) of the total homocysteine in blood is found in this (aminothiol) form. The remainder is a mixture of disulfide derivatives, including homocystine, homocysteine-cysteine mixed disulfide, and protein-bound disulfides. These various forms of homocysteine in blood are all derived from homocysteine that has been exported from the liver and other tissues. Sensitive and reliable methods for measurement of total homocysteine in blood have become widely available during the past decade. The normal concentration of total homocysteine in blood plasma is 5 to 15 μmol/L. Levels of total homocysteine tend to increase with age and are higher in men than in women. The term "hyperhomocysteinemia" refers to elevation of the blood level of total homocysteine above 15 μmol/L.

Severe hyperhomocysteinemia. Severe hyperhomocysteinemia, which is defined as a total homocysteine concentration greater than 100 μmol/L, occurs classically in patients with a homozygous defect of the gene that encodes CBS. Because these patients also have elevated levels of total homocysteine in their urine, they are said to have homocystinuria. Severe hyperhomocysteinemia also can be caused by hereditary defects in vitamin B_{12} metabolism or deficiency of vitamin B_{12} due to pernicious anemia.

Moderate hyperhomocysteinemia. Moderate hyperhomocysteinemia is defined as a blood level of total homocysteine between 15 and 100 μmol/L. This moderate degree of elevation of total homocysteine can be caused by genetic defects in CBS or MTHFR, kidney failure, or dietary deficiencies of folate or vitamin B_{12}. Moderate hyperhomocysteinemia also can be produced by drugs that interfere with the bioavailability of B vitamins, including certain medications used for the treatment of cancer, epilepsy, infections, asthma, and Parkinson's disease.

Hereditary homocystinuria. Hereditary homocystinuria (a cause of severe hyperhomocysteinemia) is quite rare, occurring in approximately one in 100,000 live births. The classic form of this disease, homozygous CBS deficiency, produces defective transsulfuration of homocysteine to cysteine, leading to accumulation of high blood levels of both homocysteine and methionine. When untreated, patients with homozygous CBS deficiency may develop skeletal deformities, dislocation of the lens of the eye, mental retardation, and premature death due to blood clots in the arteries or veins. Treatment with pharmacological doses of vitamin B_6, vitamin B_{12}, and folate, in conjunction with methionine restriction and supplemental betaine, lowers the plasma concentration of total homocysteine in most patients and markedly decreases the risk of vascular complications.

Hereditary homocystinuria also can be caused by genetic defects in the transport or metabolism of vitamin B_{12} or folate. Like homozygous CBS deficiency, these disorders cause markedly elevated levels of total homocysteine in the blood and urine, but levels of methionine are usually lower than normal. The clinical features of these disorders are very similar to those of homozygous CBS deficiency and include a high rate of cardiovascular events. It was the observation that vascular disease is a characteristic feature of homocystinuria caused by distinct metabolic defects that led the pathologist Kilmer McCully to postulate in 1969 that homocysteine may be a causative agent in atherosclerosis. *See* VITAMIN; VITAMIN B_6; VITAMIN B_{12}.

Birth defects and pregnancy complications. Moderate hyperhomocysteinemia is associated with an increased risk of birth defects resulting from incomplete closure of the neural tube during early embryogenesis. Although the overall prevalence is low (approximately one per 1000 live births), over 400,000 children are born each year with neural tube defects worldwide, and many of the cases are disabling.

The risk of a neural tube defect can be decreased by oral supplementation with folic acid during pregnancy, but only if the supplement is started before pregnancy or near the time of conception. To help prevent neural tube defects, enriched cereal grain products in the United States have been fortified with folic acid since 1998.

Moderate hyperhomocysteinemia also has been implicated as a possible risk factor for other complications of pregnancy, including intrauterine growth retardation, severe preeclampsia, birth defects other than neural tube defects, and premature separation of the placenta from the wall of the uterus. It is not yet known, however, whether treatment with folic acid or other homocysteine-lowering therapies will protect pregnant women from these complications. *See* CONGENITAL ANOMALIES.

Vascular disease. The hypothesis that homocysteine may cause cardiovascular disease was proposed over 30 years ago by Kilmer McCully, a pathologist who observed advanced vascular lesions in children with hereditary homocystinuria. McCully's pioneering observations were confirmed in 1985 by an analysis of over 600 patients with severe hyperhomocysteinemia due to CBS deficiency. By the age of 30, approximately half of these patients had suffered from a cardiovascular event (stroke, myocardial infarction, or abnormal blood clots). It is now known that the risk of cardiovascular events in patients with hereditary homocystinuria can be markedly reduced by homocysteine-lowering therapy.

The mechanisms by which elevated levels of homocysteine predispose to vascular pathology are incompletely understood. Homocysteine may damage the endothelial cells that line the surface of blood vessels, interfering with their ability to regulate blood flow and prevent blood clots and inflammation. Homocysteine also may produce an oxidative stress that leads to atherosclerosis through the generation of oxidized lipoproteins.

Although severe hyperhomocysteinemia is a proven risk factor for adverse vascular events, the importance of moderate hyperhomocysteinemia as a cardiovascular risk factor is still undefined. Since the 1980s, a large number of epidemiological studies have suggested that moderate hyperhomocysteinemia may be a very prevalent risk factor, occurring in approximately 30% of patients with stroke, myocardial infarction, poor circulation in the arteries, or blood clots in the veins. In fact, some recent epidemiological studies suggest that blood levels of total homocysteine within the high-normal range (10–15 μmol/L) may confer an increased risk of cardiovascular disease.

An association between moderate hyperhomocysteinemia and future cardiovascular events also has been observed prospectively, although a few prospective studies have failed to demonstrate this association. It is still uncertain, therefore, whether moderate hyperhomocysteinemia is an independent risk factor for cardiovascular disease or simply a marker for (that is, associated with) another factor. Nevertheless, because total homocysteine in blood can be lowered by oral administration of folic acid or combinations of B vitamins, there is growing enthusiasm for treatment of moderate hyperhomocysteinemia as a strategy for prevention of cardiovascular disease and its complications.

Guidelines from the American Heart Association emphasize that it is not yet known whether reduction of blood levels of total homocysteine through increased intake of folic acid or other B vitamins will decrease cardiovascular risk in patients with moderate hyperhomocysteinemia. Several large clinical trials to test this hypothesis are ongoing. *See* VASCULAR DISORDERS.

Neurological disease. Regulation of homocysteine metabolism appears to be especially important in the central nervous system, presumably because of the critical role of methyl transfer reactions in the production of neurotransmitters and other methylated products. Abnormal accumulation of homocysteine in the brain may lead to increased levels of *S*-adenosylhomocysteine (see illustration), which can inhibit methyl transferase reactions. Another metabolite of homocysteine, homocysteic acid, may directly damage the brain by activating a specific receptor on neurons.

It has been known for decades that mental retardation is a feature of severe hyperhomocysteinemia due to hereditary homocystinuria. It also is well known that impaired cognitive function can result from pernicious anemia, which causes hyperhomocysteinemia due to deficiency of vitamin B_{12}. Hyperhomocysteinemia also may be linked to several other neurological disorders, including depression, schizophrenia, multiple sclerosis, and dementia (including both Alzheimer's disease and other types of dementia). *See* ALZHEIMER'S DISEASE; METABOLIC DISORDERS; NERVOUS SYSTEM DISORDERS.

Steven R. Lentz

Bibliography. R. Carmel and D. W. Jacobsen (eds.), *Homocysteine in Health and Disease*, Cambridge University Press, UK, 2001; R. Clarke, Homocysteine-lowering trials for prevention of heart disease and stroke, *Semin. Vasc. Med.*, 5:215–222, 2005; S. R. Lentz, Mechanisms of homocysteine-induced atherothrombosis, *J. Thromb. Haemost.*, 3:1646–1654, 2005; J. Selhub, Homocysteine metabolism, *Annu. Rev. Med.*, 19:217–246, 1999; S. Seshadri et al., Plasma homocysteine as a risk factor for dementia and Alzheimer's disease, *N. Engl. J. Med.*, 346:476–483, 2002; D. S. Wald, M. Law, and J. K. Morris, Homocysteine and cardiovascular disease: Evidence on causality from a meta-analysis, *BMJ*, 325:1202, 2002.

Homogeneous catalysis

A process in which a catalyst is in the same phase as the reactant. A homogeneous catalyst is molecularly dispersed (dissolved) in the reactants, which are most commonly in the liquid state. Catalysis of the transformation of organic molecules by acids or bases represents one of the most widespread types of homogeneous catalysis. In addition, the catalysis of organic reactions by metal complexes in solution has grown rapidly in both scientific and industrial importance. *See* CATALYSIS.

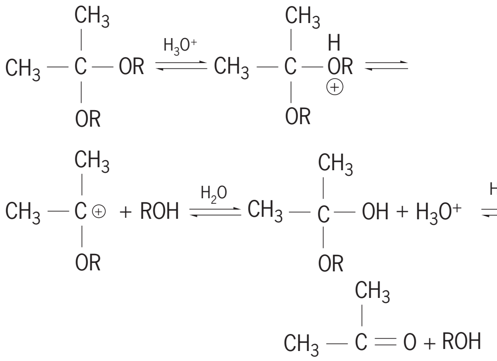
It is worth noting that sulfuric acid manufacture began in the eighteenth century using the lead chamber process in which oxides of nitrogen are used as homogeneous catalysts for the oxidation of sulfur dioxide. The basic chemistry involved is shown in reactions (1) and (2).

$$2NO + O_2 \longrightarrow 2NO_2 \qquad (1)$$

$$NO_2 + SO_2 + H_2O \longrightarrow H_2SO_4 + NO \qquad (2)$$

See SULFUR.

Acid-base catalysis. The two principal areas are specific acid (or base) catalysis and general acid (or base) catalysis. Specific acid catalysis refers to reactions in which only the oxonium ion (H_3O^+) can act as the catalyst. A common example is the hydrolysis of simple acetals, reaction (3). Specific



(3)

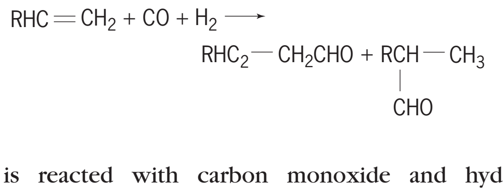
acid catalysis is found to be characteristic of reactions in which there is rapid, reversible protonation of the substrate before the slow, rate-limiting step.

Reactions that are catalyzed by proton donors in general are considered to be subject to general acid catalysis. General acid catalysis often becomes important only at higher acidity levels. The proton is a convenient and powerful agent for the distortion of the electronic configuration of a substrate in order to facilitate reaction. The mechanism by which this occurs has many variants. For example, a covalent bond may be more easily broken after protonation of one of the bonded atoms; the reaction, $ROH_2^{\oplus} \rightarrow R^{\oplus} + H_2O$ is easier than $ROH \rightarrow R^{\oplus} + OH^{\ominus}$.

Exactly the same distinction can be made in catalysis by bases as was made above for acids. Thus, in specific base catalysis the reaction rate is proportional to the concentration of $OH^{\ominus}$.

Metal complexes. In homogeneous catalysis by coordination compounds of transition metals, the catalyst is usually deployed in solution and most commonly exists in a molecularly dispersed form. Thus, all sites are potentially active for catalysis, and in many cases catalysis is observed under much milder reaction conditions than found with heterogeneous catalysis by metals and metal oxides.

The catalysis of the incorporation of carbon monoxide into organic substrates by transition metal complexes is technologically important. The hydroformylation or oxo reaction (4) in which an olefin



(4)

is reacted with carbon monoxide and hydrogen to generate a mixture of linear and branched aldehydes, was discovered in 1938. The first catalyst found was dicobalt octacarbonyl, $Co_2(CO)_8$, and this is still used extensively today in commercial operations. The steps involved in this reaction are summarized in reactions (5).

$$Co_2(CO)_8 + H_2 \rightleftharpoons 2HCo(CO)_4 \qquad (5a)$$

$$HCo(CO)_4 + \text{olefin} \rightleftharpoons HCo(CO)_3(\text{olefin}) + CO \qquad (5b)$$

$$HCo(CO)_3(\text{olefin}) \rightleftharpoons RCo(CO)_3 \qquad (5c)$$

$$RCo(CO)_3 + CO \rightleftharpoons RCo(CO)_4 \qquad (5d)$$

$$RCo(CO)_4 + CO \rightleftharpoons (R{-}CO)Co(CO)_4 \qquad (5e)$$

$$(RCO)(CO)_4 + HCo(CO)_4 \longrightarrow RCHO + Co_2(CO)_8 \qquad (5f)$$

Some of these steps represent transformations that are common to many sequences found in homogeneous catalysis. Thus, step (5*c*), the insertion of a coordinated olefin into the metal-hydride bond to generate a metal-alkyl bond, is a frequently encountered method of metal-carbon bond formation. Step (5*e*), the formation of a metal-acyl bond by alkyl migration to a coordinated carbon monoxide, is a key step in most catalytic (and stoichiometric) syntheses involving the incorporation of carbon monoxide into organic molecules.

While the reaction steps above can be conducted in a stoichiometric manner under very mild reaction conditions, in order for the system to function catalytically at rates which are desirable for industrial processes, the reaction temperature is maintained at greater than 248°F (120°C) and the reaction pressures are usually in excess of 200 atm (20 megapascals).

Other catalyst systems have been discovered which can perform hydroformylation reactions under much milder reaction conditions than cobalt. In particular, rhodium complexes containing triarylphosphine ligands can catalyze hydroformylation reactions at very rapid rates at ~212°F (~100°C) and 30 atm (3 MPa) pressure of synthesis gas ($CO + H_2$). Another important difference between the rhodium and cobalt catalysts is that the rhodium system can generate a much higher proportion of linear aldehyde product [reaction (4)]. This effect is related to the greater steric crowding around the metal when triarylphosphines are present in the coordination sphere. This is an example of the influence of stereochemistry around the metal on the stereochemical course of the catalytic reaction, and this phenomenon is an important feature of many homogeneously catalyzed reactions. *See* STERIC EFFECT (CHEMISTRY).

Another reaction involving the catalysis of the incorporation of carbon monoxide which has assumed considerable commercial importance is the synthesis of acetic acid from methanol, reaction (6).

$$CH_3OH + CO \longrightarrow CH_3CO_2H \qquad (6)$$

The reaction is catalyzed by both cobalt and rhodium complexes in the presence of an iodide cocatalyst or promoter. The mechanism of the rhodium-catalyzed reaction is reasonably well understood, as shown in reactions (7). The reaction which

$$CH_3OH + HI \rightleftharpoons CH_3I + H_2O \qquad (7a)$$

$$[Rh(CO)_2I_2]^- + CH_3I \rightleftharpoons [Rh(CO)_2(CH_3)I_3]^- \qquad (7b)$$

$$[Rh(CO)_2(CH_3)I_3]^- \longrightarrow [Rh(CO)(COCH_3)I_3]^- \qquad (7c)$$

$$[Rh(CO)(COCH_3)I_3]^- + CO \rightleftharpoons [Rh(CO)_2(COCH_3)I_3]^- \qquad (7d)$$

$$[Rh(CO)_2(COCH_3)I_3] \longrightarrow [Rh(CO)_2I_2]^- + CH_3COI \qquad (7e)$$

$$CH_3COI + H_2O \longrightarrow CH_3CO_2H + HI \qquad (7f)$$

generates the metal carbon bond, that is, step (7*b*), is rate-determining in the catalytic cycle.

The commercial reactors utilizing rhodium catalysts are operated at temperatures in the range of 300–390°F (150–200°C) and pressures of less than 40 atm (4 MPa). The rate of the reaction is sufficiently rapid that the amount of the very expensive rhodium catalyst required is very small.

A wide range of olefin transformation reactions are catalyzed by transition-metal complexes. Some of the more important reactions are isomerization, dimerization, polymerization, and metathesis. Olefin polymerization and oligomerization reactions involve a variety of homogeneously catalyzed processes and represent extremely large industrial applications of such systems (for example, polyethylene and polypropylene). Several different types of catalysts can be employed, including those based on free radicals, acids, carbanions, and transition metals. The properties of the polymer can be markedly influenced by the choice of catalyst. For example, free-radical polymerization of ethylene generates low-density polyethylene, whereas transition-metal catalysts give so-called high-density polyethylene. Examples of the transition-metal catalysts are the Ziegler-Natta systems produced by reacting titanium chloride ($TiCl_4$) with alkyl-aluminum compounds, and they are usually heterogeneous. Such transition-metal catalysts frequently introduce stereoregularity into the polymer, most likely because the reaction occurs at a crystal face. While detailed mechanistic information about these transition-metal catalyst systems is scarce, it is likely that the reaction occurs after formation of a metal-hydride either in solution or on a surface, followed by multiple olefin insertions to give polymer, as shown in reaction (8).

$$Ti{-}H \xrightarrow{C_2H_4} Ti{-}CH_2{-}CH_3 \xrightarrow{nC_2H_4} TI(CH_2CH_2)_nCH_3 \qquad (8)$$

See POLYMERIZATION.

Oxidation. Transition-metal complexes act as homogeneous catalysts in many different types of oxidation process. Two main categories of reaction can be recognized, involving either one-electron or two-electron processes.

Autoxidation. The involvement of transition-metal complexes in one-electron, radical processes is most evident in the so-called autoxidation reactions whereby hydrocarbons are oxidized to various oxygen-containing compounds by radical chain processes. The general scheme is shown in reactions (9).

$$\text{Initiation:}\quad \text{Initiator} + RH \longrightarrow inH + R\cdot \qquad (9a)$$

$$\text{Propagation:}\quad R\cdot + O_2 \longrightarrow RO_2\cdot \qquad (9b)$$

$$RO_2\cdot + RH \longrightarrow RO_2H + R\cdot \qquad (9c)$$

$$\text{Termination:}\quad R\cdot + RO_2\cdot \longrightarrow RO_2R \qquad (9d)$$

$$2RO_2\cdot \longrightarrow RO_4R \qquad (9e)$$

$$RO_4R \longrightarrow \text{nonradical products} + O_2 \qquad (9f)$$

While metal species can enhance the rate of several of the above steps, the most common pathway for catalysis of liquid-phase autoxidations involves the metal-catalyzed decomposition of alkyl hydroperoxides, of which reactions (10) and (11) are examples.

$$RO_2H + Co^{II} \longrightarrow RO\cdot + Co^{III}(OH) \qquad (10)$$

$$(RO)_2Co^{III} \rightleftharpoons RO_2\cdot + Co^{II} \qquad (11)$$

Cobalt and manganese salts are particularly effective in promoting autoxidation processes. The oxidation of *p*-xylene to terephthalic acid (the key monomer involved in the manufacture of polyester) is carried out on a very large scale using a cobalt-bromide catalyst.

Indirect oxidation. Transition-metal complexes find utility in the catalysis of various types of indirect, two-electron oxidations. Examples of these indirect processes are the so-called Wacker reaction, in which olefins are oxidized to aldehydes or ketones by palladium(II) compounds, with concomitant reduction of the palladium. The palladium is then reoxidized in a separate reaction by a combination of a copper salt and oxygen. The best-known example of the Wacker reaction is the oxidation of ethylene to actetaldehyde, reaction (12).

$$HC_2{=}CH_2 + {}^1/_2\,O_2 \longrightarrow CH_3CHO \qquad (12)$$

The reaction is conducted in an aqueous medium in the presence of palladium and copper chlorides as the catalyst system. The generally accepted mechanism is shown in reactions (13).

$$PdCl_4{}^{2-} + C_2H_4 \rightleftharpoons [PdCl_3(C_2H_4)]^- + Cl^- \qquad (13a)$$

$$[PdCl_3(C_2H_4)]^- + H_2O \rightleftharpoons [PdCl_2(H_2O)(C_2H_4)] + Cl^- \qquad (13b)$$

$$[PdCl_2(H_2O)(C_2H_4)] + H_2O \rightleftharpoons [PdCl_2(OH)(C_2H_4)]^- + H_3O^+ \qquad (13c)$$

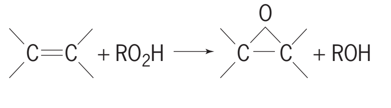
$$[PdCl_2(OH)(C_2H_4)]^- + H_2O \rightleftharpoons [HOCH_2CH_2PdCl(H_2O)] + Cl^- \qquad (13d)$$

$$[HOCH_2CH_2PdCl(H_2O)] \longrightarrow CH_3CHO + Pd + HCl + H_2O \qquad (13e)$$

$$Pd + 2CuCl_2 \longrightarrow PdCl_2 + 2CuCl \qquad (13f)$$

$$CuCl + {}^1/_2O_2 + HCl \longrightarrow CuCl_2 + {}^1/_2H_2O \qquad (13g)$$

Another important indirect oxidation process is the metal-catalyzed epoxidation of olefins with alkyl hydroperoxides, reaction (14).



(14)

Various molybdenum, vanadium, and chromium complexes act as catalysts for this reaction, by pathways that are still rather poorly understood.

Adiponitrile, $NC(CH_2)_4CN$, is produced as a precursor of hexamethylenediamine, one of the building blocks of Nylon 66. The selective addition of two moles of hydrogen cyanide to butadiene has been developed into a valuable new synthesis of adiponitrile. In one variant of this process, a zero-valent nickel catalyst, $Ni[P(OAryl)_3]_4$, can be used to bring about a series of reactions including HCN addition to butadiene, isomerization of cyanolefins, and hydrocyanation of 4-pentene-nitrile, reactions (15)–(17). A key step in reactions (15) and (17) appears to

$$HC_2{=}CH{-}CH{=}CH_2 + HCN \longrightarrow HC_3{-}CH{=}CH{-}CH_2{-}CN \quad (15)$$

$$HC_3{-}CH{=}CH{-}CH_2{-}CN \rightleftharpoons HC_2{=}CH{-}CH_2{-}CN \quad (16)$$

$$HC_2{=}CH{-}CH_2{-}HC_2CN + HCN \longrightarrow NC(CH_2)_4CN \quad (17)$$

be generation of a nickel hydride species capable of reacting with an olefin. An outline of the mechanism is shown in reactions (18)–(22), where L represents

$$NiL_4 + HCN \rightleftharpoons HNi(CN)L_3 + L \quad (18)$$

$$HNi(CN)L_3 + RCH{=}CH_2 \longrightarrow HNi(CN)(olefin)L_2 + L \quad (19)$$

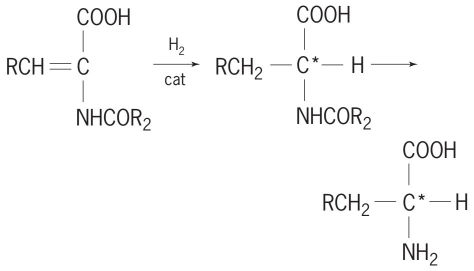
$$HNi(CN)(olefin)L_2 \longrightarrow RCH_2CH_2Ni(CN)L_2 \quad (20)$$

$$RCH_2CH_2Ni(CN)L_2 \longrightarrow NiL_2 + RCH_2CH_2CN \quad (21)$$

$$NiL_2 + 2L \longrightarrow NiL_4 \quad (22)$$

phosphite or phosphine.

Perhaps the most elegant illustration of the selectivity achievable with homogeneous catalysts is found in the asymmetric hydrogenation of unsymmetrical olefins in the presence of rhodium complexes containing optically active phosphine ligands. Through this process, it is possible to prepare a number of optically active α-amino acids from the corresponding unsaturated precursors, reaction (23).



(23)

(The carbon marked with an asterisk is an asymmetric center.) The energy difference between the optical enantiomers is very small, but nevertheless, with suitable optically active phosphine ligands, one isomer can be produced with greater than 90% selectivity. This approach is used in the synthesis of L-dopa, the drug that is used in the treatment of Parkinson's disease. *See* ASYMMETRIC SYNTHESIS; OPTICAL ACTIVITY; STEREOSPECIFIC CATALYST.

Enzymes are naturally occurring catalysts that are responsible for the myriad biological transformations involved in the synthesis and breakdown of biomaterials. The enzymes, which are usually macromolecular proteins, frequently operate homogeneously. They use many types of homogeneous catalysis, such as acid, base, free-radical, and oxidation and reduction catalysis. *See* ENZYME; ORGANIC SYNTHESIS.

Denis Forster

Bibliography. I. N. Levine, *Physical Chemistry,* 5th ed., 2000; J. A. Moulijin, P. Van Leeuwen, and R. A. Van Santen (eds.), *Catalysis: An Integrated Approach to Homogeneous, Heterogeneous, and Industrial Catalysis*, 1993; G. W. Parshall and S. D. Ittel, *Homogeneous Catalysis: The Applications and Chemistry of Catalysis by Soluble Transitic Metal Complexes*, 1992.

Homoptera

An order of the class Insecta related to the order Hemiptera. This is a major group of sucking insects, with more than 30,000 species, even though in Asia and Africa the number of undiscovered species probably still exceeds the discovered ones. Common examples are the cicadas, aphids, and leafhoppers. The group is difficult to characterize because of the large number and diverse forms of the species it contains.

The head of these insects is hypognathous or opisthognathous, the beak appearing to arise from the ventral posterior margin of the head or even from the prosternum. The gula is membranous or absent. As in the Heteroptera, the beak consists of two pairs of stylets, formed by the maxillae and the mandibles, ensheathed in the labium. The maxillary stylets fit together to form a double tube, one channel serving for the passage of food and the other for saliva.

Most winged species have four wings, but male scale insects have only two. In most forms both pairs of wings are membranous and transparent, but in some, the forewings are somewhat thickened and may then be either coriaceous and translucent, or opaque, and with or without an apical membranous area. When the insects are at rest, the forewings are usually held, rooflike, over the dorsum, with the apex of one of them slightly overlapping the apex of its complement (**Fig. 1**).

The digestive tract in a vast majority of species is peculiarly complex in that it forms a filter chamber, a structure consisting essentially of a close association

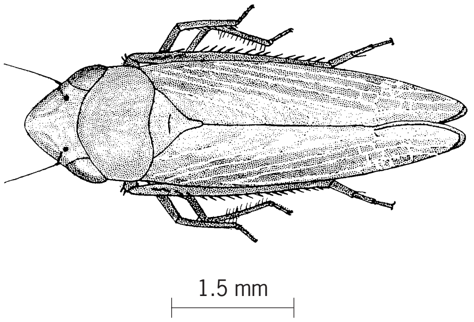


Fig. 1. A cicadellid in dorsal view.

of the posterior end of the midgut with the posterior end of the foregut. The approximated portions of the loop thus formed are believed to permit certain elements of the food to bypass most of the digestive portion of the gut. This anatomical feature, coupled with the prodigal feeding habits of most of the species, has led to the theory that certain factors in plant sap are present in such small quantities that large amounts of the sap must be imbibed, and the unneeded component shunted across the loop of the filter chamber, in order to obtain a sufficient quantity of the factors needed to sustain the insect. All species for which the food habits are known are phytophagous, deriving their nourishment from plant sap, except a few which are believed to be mycetophagous as immatures.

In most species metamorphosis is gradual, but in a few it is practically holometabolous. The adults and nymphs of most species are terrestrial, but a few species are subterranean in all stages and others are subterranean only in the immature stages. A number of species are vectors of virus diseases of plants.

Series Coleorrhyncha. This group is characterized by the origin of the beak, formed at the anteroventral extremity of the face, and by the fact that the propleura form a sheath for the base of the beak. The hindwings are absent, and the forewings are held flat over the abdomen in repose. The flight function has been lost. The prothorax is provided with dorsolateral expansions, the paranota, absent in other modern insects, but similar to structures found in Permocarboniferous fossil insects. There are additional anatomical features which have been cited as evidence of primitiveness. These serve as a basis for placing the Peloridiidae, the only known family, in the lowest position among the Homoptera. The species lack a filter chamber in the digestive tract, and are rare. They occur in Tasmania, New Zealand, and South America.

Series Auchenorrhyncha. This series and the Sternorrhyncha are the major groups of the Homoptera. In the Auchenorrhyncha the beak arises at the anteroventral extremity of the face and is not sheathed by the propleura. The labium arises well in front of the anterior legs. The antennae usually have one to three basal segments surmounted by a seta. In repose, the forewings are usually placed in a rooflike manner over the abdomen. They are more heavily sclerotized than the hindwings in some species, and may have a terminal membrane, thus somewhat resembling the forewings of the Heteroptera. Many species are good flyers. The hindwings nearly always have a jugum. The tarsi are nearly always three-segmented in the adults. Hindlegs are often adapted for jumping.

A filter chamber is apparently a constant feature of the digestive tract, although it may be greatly reduced in some species. There are usually four Malpighian tubules, originating in the anterior portion of the hindgut in the filter chamber. These frequently have an enlarged portion to which a glandular function has been attributed. Some species are cryptonephridial. Sound production by the males has been demonstrated to be of widespread occurrence in the Auchenorrhyncha and probably occurs throughout the group. A sound-producing apparatus has also been demonstrated in the females of some species.

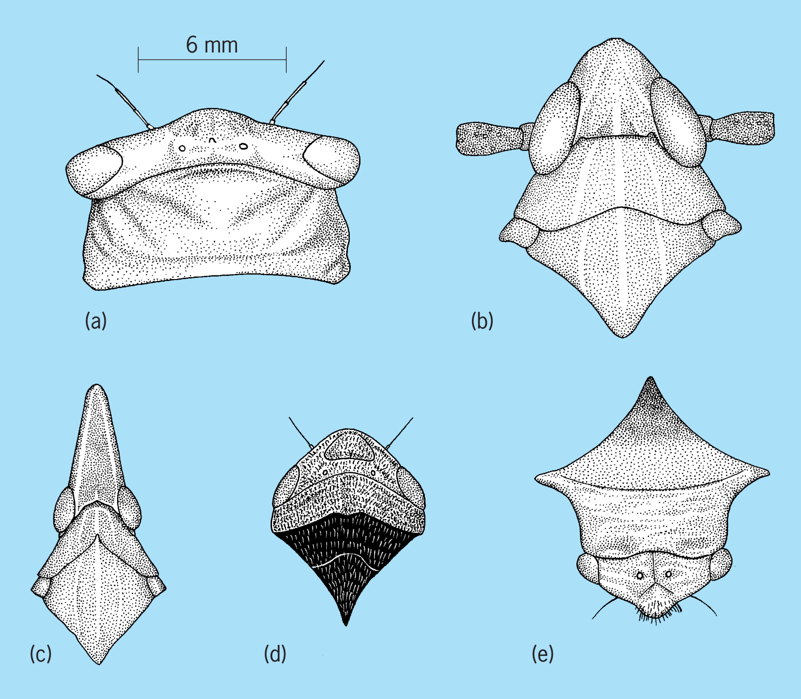


Fig. 2. Anterior dorsum of various homopterans: (*a*) Cicada, (*b*) fulgoroid (apical part of antenna omitted), (*c*) another species of fulgoroid, (*d*) cercopid. (*e*) A membracid, anterior view of face and pronotum.

The Auchenorrhyncha includes a large number of species. A number of classifications have been proposed, most of them differing chiefly in the rank assigned to the higher categories. The classification adopted here is a common one. It divides the series into the superfamily Fulgoroidea and the families Cicadidae, Cercopidae, Membracidae, and Cicadellidae. These families are not subordinate to the superfamily Fulgoroidea.

Superfamily Fulgoroidea. These are insects commonly known as lantern flies. They are distinguished from other Auchernorrhyncha by the following characteristics (**Fig. 2***b* and *c*). The middle coxae are the same length as the anterior coxae and are joined to the body at some distance from the median line. Tegulae, small, scalelike sclerites, are usually present at the base of the forewings. The hindwings usually lack a submarginal vein parallel to the wing margins.

The antennae, situated beneath the eyes, have two well-developed basal segments, of which the second is enlarged and often provided with numerous sensillae. Longitudinal carinae are often found on the face. There may be two or three ocelli, also located on the face. The forewings may be tectiform, vertical, or horizontal with their apices overlapping in repose. The anal, or claval, veins are usually fused to form a Y-shaped pattern. Contradictory statements have been made regarding the presence of a filter chamber in the digestive tract.

The popular name, lantern flies, resulted from old reports of bioluminescence in the head of a large South American species. This phenomenon has not been observed since, but the common name has survived.

This group is subdivided into 20 families and includes many species which are important because of the economic damage they do while feeding or because they carry virus diseases of plants.

Family Cicadidae. Included in this group are the cicadas, harvest flies, and jar flies. The insects in this family are probably better known to the layman than any other homopterous family because of their large size and the strident songs of the males. The following combination of characteristics will separate them from other families of Auchenorrhyncha. The short middle coxae differ from the anterior coxae and are joined to the body near the medial line. Tegulae are absent. The hindwings have a submarginal vein parallel to the wing margin. There are three ocelli arranged in a small triangle (Fig. 2*a*). The immature forms have digging legs.

The head of the adult is large, with protuberant eyes (Fig. 2*a*). The upper median portion (postclypeus) of the face is swollen, and this area is bounded throughout its length by the lora, or mandibular plates. The anterior and posterior tentorial arms are connected in the head. The forewings are membranous, tectiform in repose, having a single anal vein or two more or less parallel anal veins. The anterior legs have dilated femora which are spiny beneath. The posterior legs are not saltatorial. In the digestive tract, there is a long filter chamber in which the two parts of the gut are spirally interwound, with the hindgut issuing from its anterior end. There are four Malpighian tubules.

Adults are usually found on trees or shrubs. At least in some species, the songs of the males assemble local populations. After mating, the female lays eggs in the shoots of plants, making a slitlike incision for them by means of an ovipositor which is provided with sawlike blades or valves. These incisions frequently result in the death of the shoots. After hatching, the young cicadas fall to the ground and take up a subterranean existence which may last 1–17 years. During this period they feed on plant roots. Several species of North American periodical cicadas, with life cycles of 13 or 17 years, attract much attention because of occasional appearance in large numbers. These are often referred to as 13- or 17-year “locusts” in popular accounts, possibly as a result of association by early North Americans with the biblical plagues of locusts. There is no basis for such an association other than large numbers of individuals, and the use of the term locusts to refer to cicadas is inadvisable. *See* POPULATION ECOLOGY.

Family Cercopidae. Spittle bugs and froghoppers are common examples of this group. Insects in this group most often attract attention in the immature stages, during which they surround themselves with a mass of froth or spittle. The family has the following distinguishing characters. The middle coxae, the tegulae, and the hindwings are as described for the Cicadidae. There are usually two ocelli, never three, occasionally none. When present they are located on the crown of the head, of which the median apical portion is usually distinctly delimited from the remainder by sulci (Fig. 2*d*).

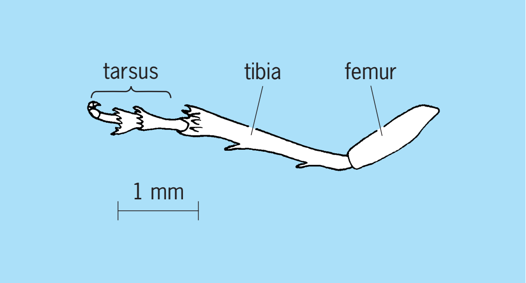


Fig. 3. Hindleg of a cercopid.

The head of the adult is proportionately smaller than in the Cicadidae, but the structure of the tentorium is similar. The pronotum, which does not extend posteriorly, is horizontal or sloped, but not vertical. The forewings, tectiform in repose, are usually somewhat sclerotized and different in texture from the membranous hindwings. The hind legs are saltatorial, with the coxae short and conical, not transversely dilated. The tibiae have one or a few stout spurs and a cluster of small outgrowths at the apex (**Fig. 3**). In the digestive tract, there is a complicated filter chamber in which a number of convolutions are found. The hindgut issues from its posterior end. There are four cryptonephridial Malpighian tubules arising from two basal stalks.

The froth, or spittle, which covers the immature insects consists of extruded anal fluid with which air bubbles are mixed. The frothy mass does not evaporate readily, a fact which seems correlated with the reported inability of the young cercopids to survive in a dry atmosphere. There are many more species in the tropics than in temperate climates. One species, *Philaenus spumarius*, the meadow spittle bug, is very common in the temperate portion of the Northern Hemisphere, and its masses of spittle are familiar sights. Some species of Australia and the East Indies live within a calcium carbonate tube attached to stems or leaves.

Family Membracidae. The treehoppers are small to medium in size and seldom attract attention. Most of them feed on woody plants and are found on the stems in sunny locations. Frequently a number of specimens are arranged in a vertical row on the stem, all with their heads downward. Membracidae may be recognized by the following combination of characters. The middle coxae and the tegulae are as described above for the Cicadidae. There are two ocelli, or none. The pronotum extends backward over the abdomen, sometimes almost completely covering the wings (**Fig. 4***a*). The upper portion of the head is vertical (Fig. 2*e*).

In the adult head, the anterior and posterior arms of the tentorium are not connected. The antennae consist of two larger basal segments and a

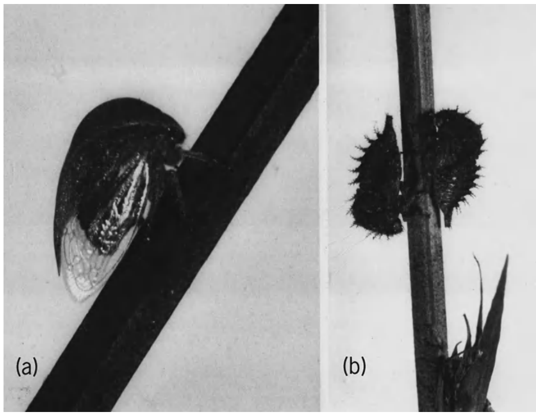


Fig. 4. Membracids on stems. (*a*) Adult. (*b*) Nymphs. (*Courtesy of C. H. Hanson*)

terminal seta with a large number of subsegments. The forewings are usually membranous but may be partly sclerotized, and are tectiform in repose. The radial and medial veins are not fused basally. There is often a submarginal vein in the hindwings. The anterior and middle tibiae are often dilated. The hind legs are saltatorial and spiny. The digestive tract has a filter chamber with fewer coils than found in many Homoptera. There may be two or four Malpighian tubules which are cryptonephridial, at least in some species.

The greatest number of species occurs in the warmer regions of the world. In many species the enlarged pronotum has adornments, excrescences, and processes which are astonishing in appearance, some of them nearly as large as the remainder of the insect. Most species are not active, leaping only to take flight, seldom flying, and then only for short distances. In North America, *Stictocephala bubalus*, the buffalo treehopper, causes economic damage to fruit and other trees by slitting the twigs during oviposition. The nymphs (Fig. 4*b*) leave the trees and feed on herbaceous plants, often occurring in great numbers in pastures. The adults return to woody plants before oviposition.

Family Cicadellidae. The leafhoppers are included in this large family. These usually small insects are known to many people by sight but not by name, because of their common occurrence in great numbers at night near lights. The species may be distinguished from other families by the following combination of characters. The coxae and tegulae are as described for Cicadidae. The pronotum does not extend backward over the abdomen and does not have a median ridge. The upper portion of the head is never vertical. There is usually a submarginal vein parallel to the wing margin in the hindwings (**Fig. 5**). There are two ocelli, or none. When present, they may occur on the face, on the crown of the head, or on the margin between the face and crown. The hind tibiae have many spines arranged in rows (Fig. 1).

On the face, the lora border the postclypeus for only a short distance. The anterior and posterior arms of the tentorium are not connected in the head. The forewings may be membranous or heavily sclerotized and, in the latter case, often have a membranous apical portion. They are usually tectiform in repose. Radial and medial veins arise from a common stalk. The hind legs are saltatorial. The digestive tract has a filter chamber, but reports indicate a considerable degree of variation in its detailed structure. There are four cryptonephridial Malpighian tubules.

Probably the greatest number of species occurs in tropical areas, but the majority of these have not been described. Temperate North America also has a large number of species. Leafhoppers occasionally bite man, but apparently they have never been seen taking blood. Several species have been found to be vectors of virus diseases of plants. Some of the more important of these virus diseases are phloem necrosis of elms, curly top virus of sugar beets, aster yellows virus, rice dwarf virus, and papaya bunchy top virus, in addition to virus diseases of potatoes, clover, alfalfa, grapes, peaches, sugarcane, eggplant, corn, maize, wheat, and cranberries. In a few cases, the virus is transmitted transovarially. *See* PLANT PATHOLOGY.

Series Sternorrhyncha. In this group of homopterous families, the beak appears to arise either between the fore coxae or behind them. The antennae are usually long, filamentous, and have no well-differentiated terminal setae. Wingless forms are common. The wings, when present, are usually membranous, with reduced venation, and usually without closed cells. The hindwings have no jugum. The tarsi of the adults are 1- or 2-segmented. A filter chamber is usually present in the digestive tract.

The Sternorrhyncha includes a large number of species, many of them of great economic importance. The winged forms are not strong flyers, but they are so light that they may be borne considerable distances by air currents. There have been several classifications proposed for the Sternorrhyncha. The following subdivision into superfamilies and families appears to be the most widely used classification.

Family Psyllidae. The Psyllidae are known as jumping plant lice. This family has also been known as the Chermidae. Its representatives resemble cicadas in appearance, but are much smaller. About 1000 species are known. Adult insects have a

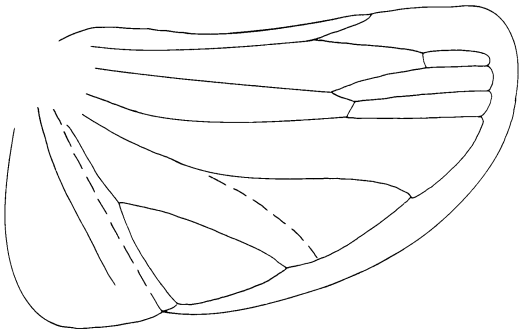


Fig. 5. Cicadellid wing. The more basal broken line represents the jugal fold.

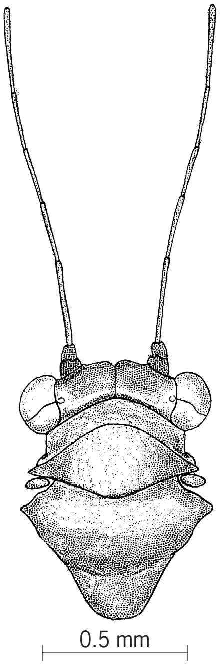


Fig. 6. Anterior dorsum of a psyllid.

transverse head, usually emarginate anteriorly, with protuberant eyes (**Fig. 6**). The face is often in the form of two conspicuous cones which are directed ventrally. Three ocelli occur. The antennae are 6- to 10-segmented, with the latter condition being common. They consist of two wider basal joints and a terminal portion which bears sensillae. The apical antennal segment bears two fine setae. The winged forms have four wings which are usually membranous, but the forewings are thicker than the hindwings. The wings are tectiform in repose. The venation is reduced, but conspicuous, and cross veins are seldom present. In the forewings, the radius, media, and cubitus arise from a common stalk. The tarsi are 2-segmented with a pair of apical claws. The hind legs of the adult are saltatorial, with the coxae greatly dilated.

The hindgut and the posterior portion of the esophagus are looped around each other. Possibly this is a functional filter chamber, although a specialized clearly delimited filter chamber, similar to that found in some other groups of Homoptera, is apparently absent. There are four Malpighian tubules opening separately into the midgut.

The species of psyllids may be monophagous or polyphagous. There may be an alternation of host plants, but the nymphs and adults occur on the same plant. Hibernation may occur in the egg, nymphal, or adult stages. The nymphs have flattened bodies and do not have saltatorial hind legs. They are frequently covered by a secretion which may be wooly, waxy, or viscous.

Some psyllids produce severe damage to their food plants. This damage may result from the mere feeding by tremendous numbers of individuals, from the resulting yellowing or rolling of leaves, or from galls produced on the leaves. Indirect damage may result from the growth of fungi on leaves which have become coated with the sugary excrement, the honeydew, of the psyllids. *Psylla pyricola*, the pear psylla, is a species of considerable economic importance in North America. Its damage includes premature falling of leaves and fruit, reduced quality of fruit, and blackening of the leaves from fungus.

Family Aleyrodidae. The whiteflies are 0.3 in. (7 mm) or less in length and usually lightly covered with a white, powdery, waxy material which has led to their common name. The head varies in shape, and there are two ocelli. The antennae are usually 7-segmented, with the two basal segments thicker than the other segments, which may have sensoria and may terminate in a bristle. The eyes may be reniform, or divided, in which case they appear as four eyes. Adults have four wings which are opaque, being covered with a whitish powder. The wings are tectiform in repose. The venation is greatly reduced and there are no cross veins. The tarsi are 2-segmented, with a pair of terminal claws between which is a median structure which varies in form from a blade to a bristle. The thorax is separated from the abdomen by a constriction. There is a filter chamber in the gut, but the first ventriculus of the gut is not greatly dilated where the filter chamber occurs. There are only two Malpighian tubules. The stylets of the mouthparts, when not in use, are held in the crumena, a pouch within the thorax.

Several species of Aleyrodidae are polyphagous. The eggs are attached to the leaves of plants. The newly hatched insects are ambulatory, but they soon choose a feeding site and remain virtually sessile until reaching the adult stage. The sessile forms, with their appendages greatly reduced, resemble scale insects, and the last larval instar is anchored to the plant by a waxy secretion. Both bisexual reproduction and parthenogenesis occur within the group.

Whiteflies directly damage plants by their feeding. Indirectly, damage results from spotting at the feeding site, growth of fungus on the excreted honeydew, or from increased susceptibility of leaves to winter damage. Two species are important pests in the United States. *Dialeurodes citri*, the citrus whitefly, an Asian species, is a pest of citrus and other plants in California, Florida, and elsewhere. *Trialeurodes vaporarium*, the greenhouse whitefly, is a pest in greenhouses and on several cultivated crops. Other species carry virus diseases to cassava, cotton, cucumber, sunflower, and tobacco.

Superfamily Aphidoidea. Members of this large superfamily of small insects have four wings, or none. The wings, usually membranous or whitish and opaque, are held tectiform over the abdomen in repose. The forewings are much larger than the hindwings. The tarsi are 2-segmented and usually have two claws. The hind legs ordinarily are not saltatorial. The intersegmental lines are distinct in the abdomen. Both parthenogenetic females and sexual females occur. The life history may be extremely complicated, involving several host plants as well as several forms of

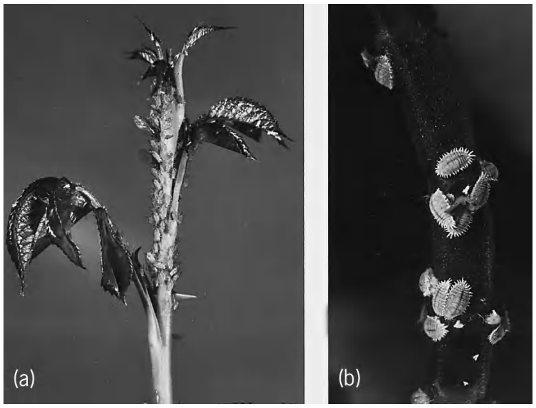


Fig. 7. Plant pests. (*a*) A colony of aphids on a rose shoot. (*b*) The citrus mealybug (*Ohio Agricultural Experiment Station*).

the insects. Honeydew is usually produced. This superfamily includes the families Aphididae and Chermidae.

Family Aphididae. The true aphids (**Fig. 7*a***) are members of this family in which the sexual females are oviparous and the parthenogenetic forms are viviparous. The sexual females, and usually the males, have a functional beak and a continuous digestive tract. Cornicles are usually present.

Compound eyes are present in the adults and three ocelli occur in the winged forms. Usually the antennae are 5- or 6-segmented with the apical segment ending in a tapering terminal process. Sensoria are present on some antennal segments. The beak has five segments. The wings extend caudad beyond the apex of the abdomen. In the forewing, veins Rs (radial sector) separates from the stigma and reaches the wing apex. The last abdominal segment is extended to form a cauda, with the anus opening beneath it. A poorly developed filter chamber is said to occur in the gut of some aphids, but a number of aphids lack it. Malpighian tubules are absent throughout the family (**Fig. 8**).

Many species of aphids have both an alternation of generations and an alternation of hosts, one of which is usually obligatory. In temperate climates, the eggs are laid most commonly on a woody host in autumn where they overwinter. The females which hatch from the eggs produce, by viviparous parthenogenesis, a variable number of generations on the primary host. Winged forms are eventually produced which migrate to a herbaceous alternate host. Here a number of generations are produced by viviparous parthenogenesis, some of the individuals having wings, but most of them being wingless. In autumn, a generation of winged females occurs which migrate to the primary host. Here they produce parthenogenetically the winged sexual males and females. After mating, the females lay eggs. There are many departures from this generalized life history. Some species are monophagous. In others, alternation of hosts appear optional, the aphid species being able to reproduce on the primary host, secondary host, or both.

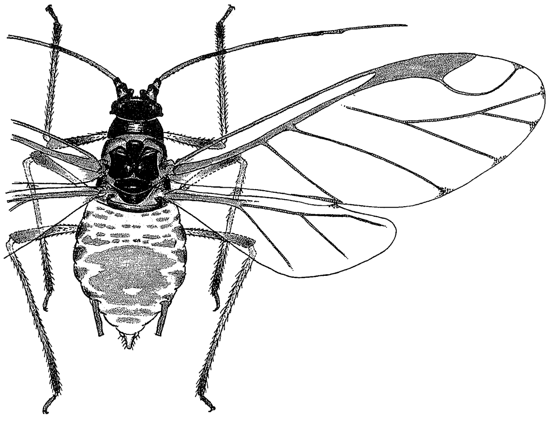


Fig. 8. An aphid, dorsal aspect.

Many species of aphids that produce honeydew are very attractive to ants, and colonies of very small aphid species which otherwise might escape notice can often be located by observing the attending ants. In a few species, this relationship has progressed to the point where ants are necessary for the survival of the aphid species, as in the corn root aphid, *Anuraphis maidiradicis*. The eggs of this species are cared for by the corn field ant, and the young aphids are transferred to corn roots early in the growing season. *See* SOCIAL INSECTS.

Many aphid species are important because of damage done in feeding. The pea aphid, *Acyrtosiphum pisum*, is a pest to peas and other crops in the United States. The cotton aphid, *Aphis gossypii*, is a pest of cotton, melons, and other crops in the same region. Other species are important because of their ability to transmit virus diseases to such plants as potatoes, sandalwood, squash, beets, cauliflower, celery, clover, cucumber, groundnut, alfalfa, onion, pea, sugarcane, tobacco, turnip, and citrus.

Family Chermidae. This is a small family of minute insects, the adelgids and phylloxeriids, in which both the sexual and parthenogenetic females are oviparous. Cornicles are absent. The antennae are 3- to 5-segmented and bear sensilae. Both winged and wingless forms occur. The following are subfamilies of Chermidae.

In the subfamily Cherminae all forms have a beak, and the digestive tract is not closed; the wings are tectiform in repose, and there is no branched vein extending toward the posterior margin of the forewing. There is frequently a waxy, flocculent secretion. The life cycle is extremely complicated and may involve an alternation of plant species and an alternation of generations in the insects, with several morphological forms occurring in the life history of one species. The primary host is always spruce. Secondary hosts are other conifers.

In the subfamily Phylloxerinae, the sexual forms lack mouthparts. The parthenogenetic females have a beak but the digestive system is closed (not continuous), and there is no honeydew. In repose, the membranous wings lie flat over the abdomen. There is a branched vein extending toward the posterior margin of the forewing. Waxy secretions, rarely present, are not flocculent. Although winged migrants occur commonly, there is no secondary host.

The grape phylloxera, *Phylloxera vitifoliae*, has been a severe pest of cultivated grapes and once threatened the entire wine industry of France. It was discovered that native North American grape roots were not damaged greatly by the phylloxera, and consequently these were used as grafting stock in Europe to reduce damage.

Superfamily Coccoidea. The scale insects and mealy bugs, which are members of this large and important superfamily, are usually small. More than 4000 species have been described. In the males, the hindwings are reduced to clublike halteres. The wings are usually held flat over the back in repose, and the venation is greatly reduced. The females are wingless. When legs are present the tarsi are usually 1-segmented, but in some males the tarsi are 2-segmented. There is a single tarsal claw. The hind legs are not saltatorial. In some species, the abdominal segmentation is much modified or obliterated.

The males are usually very small, even in species where the female is much larger. They have no beak, do not eat, and have a nonfunctional digestive tract. Young males resemble corresponding female stages, but they molt more often, eventually passing through a stage so much like the pupal stage of holometabolous insects that they have presented an obstacle to systematists of higher categories of insects who used the type of metamorphosis as a fundamental criterion in classification. Adult males normally have long antennae consisting of 10 or fewer segments, although written accounts have stated, as a result of erroneous observations, that as many as 25 segments were present. The eyes of the adult male are compound in some species, but they usually consist of a series of isolated facets. The legs are well developed. The caudal end of the abdomen often bears an elongate process. Males are unknown in some species, and apparently of rare occurrence in some others. A few species have wingless males.

Females are usually sedentary, with the legs absent or reduced and nonfunctional. The body may be soft, or gall-like, or covered with powdery or tufted wax, or with scales or other hardened secretions. Some species form cysts which resemble small pearls. Others inhabit plant galls. In some groups, the females superficially resemble aphids, but cornicles are never present. The females usually have a 1- to 3-jointed beak which arises behind the bases of the anterior legs and varies considerably in shape and size. Its stylets are coiled in a ventrally located internal crumena when not in use. Antennae are present in adult females and may be reduced to flat disks, or may be elongate, with as many as 11 joints. The eyes are very simple and somewhat resemble ocelli. There are usually a number of glands, pores, and ducts in the integument and these features are useful in taxonomy. There is no pupalike stage, as occurs in males. The gut may be either continuous, and with a complex filter chamber, or discontinuous. There are two, three, or four Malpighian tubules.

In the first nymphal stadium, both sexes are mobile and known as crawlers. At this period in their development, they spread over the plant and to other plants.

Scale insects injure the host plant by their feeding on leaves, stems, or roots, and a number of species are very important economically. Most species produce large quantities of honeydew. A few species have been shown to be vectors of virus diseases of cucumbers, tobacco, and cacao. The sedentary habits and small size of the females have contributed to their wide distribution on economically important plants. Monophagous and polyphagous species occur.

Although specialists in scale insects generally agree that the included groups of insects form a superfamily, there seems to be little agreement on the constitution of the several included families. For this reason the taxa discussed here can be treated most conveniently, at present, as subfamilies. As many as 20 taxa in the family group have been accepted by some authors.

In the Margarodinae, abdominal spiracles are present in all stages of development. The adult males have compound eyes, and usually 10-jointed antennae. Adult females have at least some segmentation of the legs, and usually two bristles on the tarsal claw. The antennae of the adult female are not contiguous basally, and the anal tube is apical if well developed. No stages have a flat anal ring which bears pores and setae. Some *Margarodes* and related species live underground and have become lawn pests in some parts of the southern United States, usually in areas where the soil is sandy. The females form glassy cysts in which they are able to live quiescent for a considerable time without food. The cysts, or "ground pearls," are lustrous and are used as jewelry by natives in some parts of the world.

In the Ortheziinae, abdominal spiracles are present in all stages. Immature forms of both sexes and all females have a flat anal ring which bears pores and setae. Adult males have 9-jointed antennae with an apical seta on the terminal segment. Adult males have a strongly bivalved penis sheath. The body of the female is covered with hard white waxy plates. *Orthezia insignis* has a widespread distribution in greenhouses.

The Monophlebinae posses the characters listed above for the Margarodinae but the anus is dorsal. This subfamily includes the cottony-cushion scale, *Icerya purchase*, a species introduced to California from Australia and which at one time threatened the citrus industry in California. It was brought under control by an introduced Australian coccinellid, *Rodolia cardinalis*, in one of the earliest known instances of biological control.

In the Aspidiotinae, the adult male does not have compound eyes. Adult females have greatly reduced antennae and are without legs. There are no abdominal spiracles in any of the stages. There are no setae at the anal opening. In the females and nymphs, the apical abdominal segments are fused to form a compound pygidium. The scales are shieldlike and nearly circular. The San Jose scale, *Aspidiotus perniciosus*, is one of the most important members of the subfamily and is thought to be a native of China. It received its common name from the California locality where it was first found in the United States. It attacks many kinds of fruit trees and shrubs, as well as woody ornamentals. It occurs throughout the United States and Canada and in other parts of the world in temperate regions.

The species of Lepidosaphinae have dark-colored scales and features similar to those of the Aspidiotinae, except that the scale is not circular. Three species of *Lepidosaphes* are of economic importance, two on citrus, and one, the oystershell scale, a cosmopolitan species which infests many kinds of woody plants.

In the Lacciferinae, the adult male does not have compound eyes. The abdomen is without spiracles in all stages, and in the nymphs and females its apical segments are not coalesced to form a pygidium. The abdominal apex has a tubular projection. The body is not covered by a scale, but the insects are enclosed in a mass of resin. The Oriental species, *Laccifer lacca*, or lac insect, found on a number of host plants, as well as related species, secretes the resinous material from which shellac is produced. *See* SHELLAC.

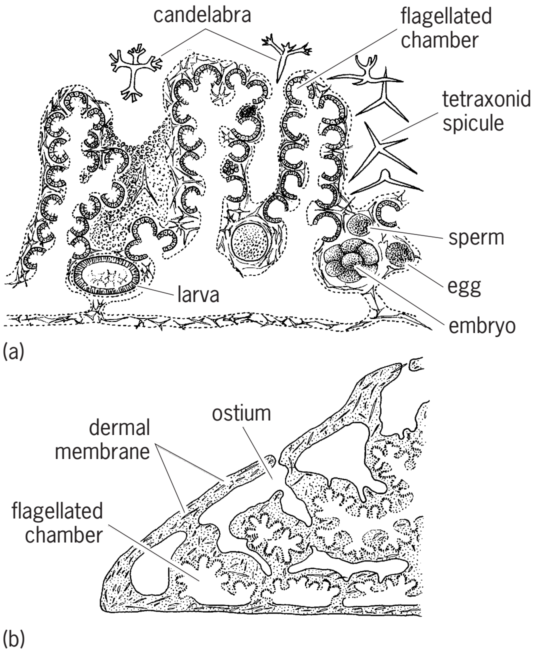
The Eriococcinae have many characteristics like those of the Lacciferinae, but the abdominal apex is without the long tubular projection. The adult females and late instar nymphs have an anal ring. The beneficial cochineal insects, *Dactylopius coccus*, belong to this group. The harmful mealybugs, species of *Pseudococcus* (Fig. 7*b*), and other genera also belong here. These soft-bodied insects are covered with flocculent wax secretion which suggested their common name. A number of species are harmful to cultivated plants, and some are serious pests of greenhouse and house plants. *See* ENTOMOLOGY, ECONOMIC; HEMIPTERA; INSECT PHYSIOLOGY; INSECTA.

David A. Young

Homosclerophorida

An order of primitive sponges of the class Demospongiae, subclass Tetractinomorpha, with a skeleton consisting of equirayed, tetraxonid, siliceous spicules and their derivatives formed through reduction in number of rays (see **illus.**). In some species the ends of the rays of the tetraxons branch many times to form spicules known as candelabra. Sponges of the genus *Oscarella*, considered to be primitive, lack spicules but are reinforced by a mesenchymal collagen bearing elastin fibers. The genus *Corticium* and related forms have a cartilagelike dermal region.

Homosclerophorid sponges are mostly small in size and encrusting to massive in shape. They occur in tidal and shallow waters, down to depths of at least 1640 ft (500 m). Fossil sponges with spicules suggesting homosclerophorid affinities are scattered through the fossil record from Carboniferous strata upward. *See* DEMOSPONGIAE.



Homosclerophorida morphology. (*a*) Section through *Plakina*. (*b*) Section through *Plakortis* with cortex.

Willard D. Hartman

Honey Dew melon

A long-keeping cultivar of muskmelon, *Cucumis melo inodorus*, of the gourd family, Cucurbitaceae (see **illus.**). *See* VIOLALES.

Honey Dew melon, *Cucumis melo inodorus*.

Description. Vines are vigorous and prolific and have large leaves and stems. The vines usually bear andromonoecious flowers which are pollinated by bees. The fruits are large, 5–7 lb (2.3–3.2 kg), slightly oval (length 7–8 in. or 18–20 cm, diameter 6–7 in. or 15–18 cm), smooth, creamy yellow to ivory when ripe, and with little or no net. They usually remain attached to the stem at harvest; some later introductions develop an abscission layer (slips) at harvest maturity. The flesh is thick, light green, tender, juicy, and very sweet with mild aroma and flavor; it contains 10% or more sugar when ripe and is rich in potassium and vitamin C, but not as rich in vitamin A as the orange-fleshed cantaloupe.

Quality. Honey Dew melons can be stored for 2–3 weeks at temperatures of 45–50°F (7–10°C). Melons do not ripen satisfactorily at temperatures below 50°F (10°C). Commercially, Honey Dew melons after harvest are often treated with ethylene gas (500–1000 parts per million) to induce and hasten ripening. As the fruits mature, the skin changes in color from greenish to creamy white, skin texture changes from a fuzzy feel to a smoothness, and the flesh increases in sugar, decreases in acidity, and releases the characteristic aroma. Sugars do not increase after the fruit is removed from the vine. Flavor and texture improve for a few days after harvesting and attain their highest quality in fruits harvested at their maximum sugar content. In California, state laws require that the edible portion contain at least 10% soluble solids before Honey Dews can be marketed.

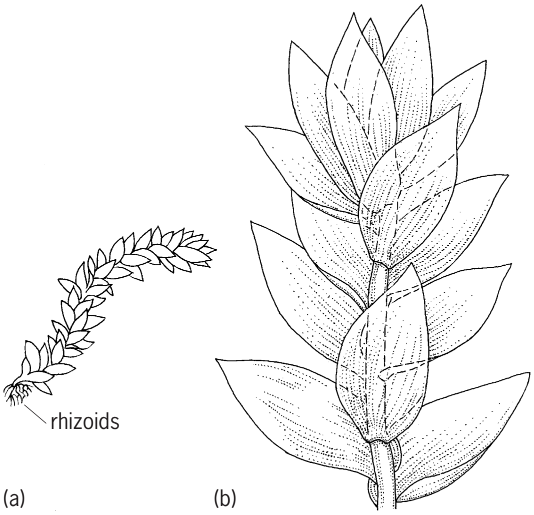
Cultivation. Honey Dew is of African origin, has been grown in France for many years as the cultivar White Antibe, and was introduced into the United States in 1911. It was named Honey Dew in 1915. It requires a long frost-free season of 110–125 days and grows well at average temperatures of 70°F (21°C) or higher. Most of the crop is harvested in the summer in California; almost all of the remaining crop is harvested in the spring in Texas, with less than 5% harvested in the fall in Arizona. The chief growing areas are the Sacramento and San Joaquin valleys of California. In these areas high sunlight, low humidity, and the absence of rain tend to prevent fungus diseases that often defoliate the plants in humid areas. Honey Dew is propagated by seeds planted directly in the soil or by transplants. Plants are spaced about 12 in. (30 cm) in rows 6–8 ft (2–2.5 m) apart.

Diseases. Diseases include powdery mildew, which occurs under high humidity and is controlled by sulfur dusts, and Fusarium and Verticillium wilts, which inhabit the soil and cause wilting of the foliage and early death. Virus diseases, including watermelon mosaic, cucumber mosaic, and squash mosaic, are transmitted by insects. Major insect pests include wireworms, cutworms, leaf miners, cucumber beetles, aphids, and spider mites. Nematodes attack the roots of plants grown in infested soils. *See* MUSKMELON; PLANT PATHOLOGY. Oscar A. Lorenz

Bibliography. California Melon Research Board, *Informational Bulletin*, quarterly; N. B. Childers, *Modern Fruit Science*, 9th ed., 1995; R. A. Seelig, *Honeydews*, Fruit and Vegetable Facts and Pointers series, 1973.

Hookeriales

An order of the true mosses (subclass Bryidae), found mostly in the tropics (see **illus.**). The Hookeriales are no doubt heterogeneous, but most have well-developed double costae. The plants are usually shiny and often yellowish. The stems are creeping to spreading or ascending, and sparsely to freely and irregularly branched, or frondose. The leaves are often asymmetric and complanate, and commonly bordered and toothed. The cells are often large and lax, and the alar cells are not differentiated. The sporophytes are lateral, with setae short to elongate and often rough, even bristly. The capsules may be erect to inclined, but are essentially symmetric. The double peristome consists of 16 teeth which are long and slender, gradually tapered, papillose throughout or cross-striolate below, bordered, and often furrowed, with an endostome consisting of 16 keeled segments from a basal membrane. The calyptrae may be cucullate or mitrate and often ciliate-fringed.



***Hookeria acutifolia*. (*a*) A portion of an entire plant. (*b*) Apical portion enlarged. (*After W. H. Welch*, *Mosses of Indiana*, *Indiana Department of Conservation*, *1957*)**

As presently constituted, the order consists of two or three families and about 40 genera. The strong double costa, well-developed double peristome, with teeth slender and often furrowed, and mitrate calyptrae, which are often somewhat hairy and sometimes fringed at the base, allow some clustering of genera around a natural center. However, the relationships of the Daltoniaceae and Ephemeropsidaceae (with papillose exostomes) to the Hookeriaceae, (with cross-striate exostomes) are unclear. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA. Howard Crum

Hooke's law

A generalization applicable to all solid materials, stating that stress is directly proportional to strain and expressed as

$$\frac{\text{Stress}}{\text{Strain}} = \frac{S}{\epsilon} = \text{constant} = E$$

where E is the modulus of elasticity, or Young's modulus, in pounds per square inch. The constant relationship between stress and strain applies only to stress below the proportional limit. For materials having a nonlinear stress-strain diagram, the law is an approximation applicable to low stress values. *See* STRESS AND STRAIN; YOUNG'S MODULUS.
W. J. Krefeld; W. G. Bowman

Bibliography. R. G. Budynas, *Advanced Strength and Applied Stress Analysis*, 2d ed., 1998; H. E.

Davis, G. E. Troxell, and F. W. Hauck, *The Testing of Engineering Materials*, 4th ed., 1982.

Hop

A plant (*Humulus lupulus*) belonging to the family Cannabaceae, also including nettles and hemp. The plant is a rough-stemmed, tall-twining dioecious perennial herb. It is propagated by cuttings of underground stems. The male flowers are borne on loose stalks. The anthers burst to produce large quantities of pollen, which is windborne. The female flowers occur in a cluster consisting of a compressed central axis that holds many flowers. As the clusters mature, the central axis elongates and papery leaves replace each flower, growing into a conelike structure, which is the hop of commerce (see **illus.**).

Mature hop cones.

The commercial value of the hop is in the golden-yellow lupulin glands on the bracteoles. The glands contain the resins, which impart bitterness, and essential oils, which impart hop flavors, to beer and ale. Dried hops are also used as soporifics. *See* MALT BEVERAGE.

Hop is grown primarily in the temperate areas of the world. In the United States, commercial production is in Washington, Oregon, and Idaho. Hop is planted on a 6.5–8-ft (2–2.5-m) spacing. The vine is trained by hand to strings fastened to a trellis 16.5–18 ft (5–5.5 m) in height. The female inflorescences are picked mechanically and dried in forced-air, direct-fired kilns at temperatures of 115–149°F (46–65°C). Hops are marketed in 190–200-lb (85–90-kg) bales. Other major hop-producing countries include Germany, the Czech Republic, Ukraine, China, United Kingdom, and Slovenia.

The principal pests are the two-spotted spider mite (*Tetranychus urticae*) and the hop aphid (*Phorodon humuli*). Major fungal diseases are downy mildew (caused by *Pseudoperonospora humuli*), powdery mildew (caused by *Sphaerotheca humuli*), Verticillium wilt (caused by *Verticillium albo-atrum*), and Phytophthora crown rot (caused by *Phytophthora citricola*). The fungal diseases are controlled by either using fungicides or growing resistant varieties. The severity of the fungal diseases varies between growing regions because of differences in environment. *See* FUNGISTAT AND FUNGICIDE.

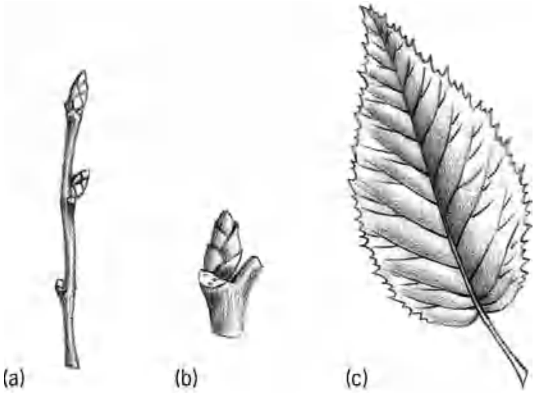
Viruses causing diseases in hop include the *Prunus* necrotic ringspot virus and apple mosaic virus (both ilarviruses); and hop latent virus, hop mosaic virus, and American hop latent virus (all carlaviruses). Viruses are controlled by planting virus-free rootstock or resistant varieties. Two viroids that infect hop are hop stunt viroid and hop latent viroid. *See* PLANT PATHOLOGY; PLANT VIRUSES AND VIROIDS.

C. B. Skotland; Stephen T. Kenny

Bibliography. R. A. Neve, *Hops*, 1991; V. Rybacek (ed.), *Hop Production*, 1991.

Hophornbeam

The genus *Ostrya* of the birch family, represented in North America by two species. *Ostrya virginiana*, a small tree which may reach a height of 60 ft (18 m), is widely distributed in the eastern half of the United States and in the highlands of southern Mexico and Guatemala. It can be recognized by its fruit, which closely resembles that of the hop vine, and by its very scaly bark. The scales usually occur in narrow, more or less parallel, vertical strips. The winter buds are usually tinged with green showing about six striate scales, and the leaves are sharply and doubly serrate (see **illus.**). This is one of several trees known as ironwood because of its hard, strong wood, and like the hornbeam, it is used for fence posts, tool handles, mallets, and other articles requiring hardness and strength. *See* FAGALES; TREE.

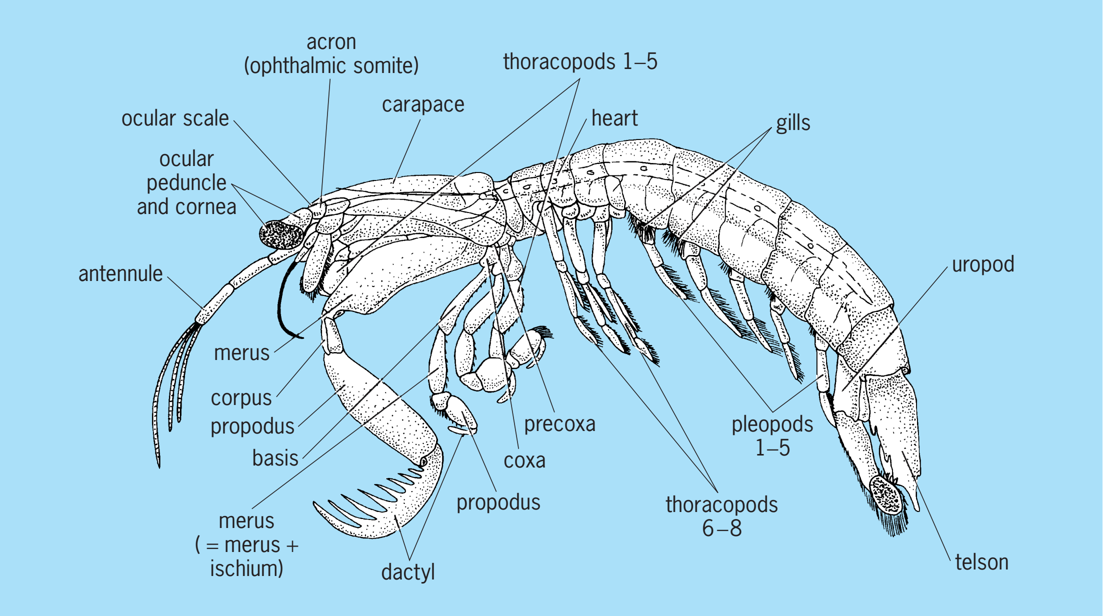


American hophornbeam (*Ostrya virginiana*). (*a*) Twig. (*b*) Terminal bud. (*c*) Leaf.

Arthur H. Graves; Kenneth P. Davis

Hoplocarida

A subclass of Crustacea, with a single extant order, Stomatopoda, commonly known as mantis shrimps. The Haplocarida was formerly included as a taxon



Haploid morphotype. (*After P. A. McLaughlin, Comparative Morphology of Recent Crustacea*, W. H. Freeman, 1980)

within the Eumalacostraca, and disagreement regarding its independent origin still persists. The controversy is centered on the question of whether those elements of the eumalacostracan caridoid facies observed in haplocarids represent examples of homology or convergence. Investigations of fossil (Paleostomatopoda) and Recent Stomatopoda suggest that a distinct set of morphological features, sometimes referred to as a hoploid morphotype, clearly delineate the Hoplocarida from the Eumalacostraca (see **illus.**). Those characters include cephalic kinesis, triflagellate antennules, thoracopods with trisegmented protopods, five pairs of subchelate thoracic appendages ("maxillipeds"), abdominal gills, and an elongate tubular heart that extends the length of the body and has segmentally arranged ostia and arteries. Morphological aspects of feeding observed in modern-day species, as well as features of the digestive system and abdominal musculature, tend to support the hypothesis of a distinct origin. *See* CRUSTACEA; STOMATOPODA. Patsy A. McLaughlin

Bibliography. F. R. Schram (ed.), *Crustacean Phylogeny*, vol. 1 of *Crustacean Issues*, 1983; F. R. Schram, Polyphyly in the Eumalacostraca?, *Crustaceana,* 16:243–250, 1969.

Hoplonemertini

An order of the class Enopola of the phylum Rhynchocoela, characterized by possession of an elaborate armed proboscis consisting of an anterior thick-walled tube, a median portion armed with stylets, and a posterior blind tube. The alimentary system possesses a cecum, and in some species (such as *Amphiporus lactifloreus*) the foregut (stomach) can be everted into the prey to achieve extracorporeal digestion prior to ingestion. There are two suborders, the Monostylifera and Polystylifera, separated on the number of stylets in the proboscis. Monostylifera include fresh-water, terrestrial, and symbiotic species, as well as the more common marine littoral forms. *See* ANOPLA; BDELLONEMERTINI; ENOPLA; NEMERTEA. J. B. Jennings

Horizon

Traditionally, the apparent boundary between the sky and the Earth or the sea. Without complicating factors, it would be a line located 90° from the zenith or overhead point; because of trees, hills, or buildings, the visible horizon differs from the ideal one. *See* ASTRONOMICAL COORDINATE SYSTEMS; CELESTIAL SPHERE.

Black holes. The term horizon also refers to the imaginary boundary of a black hole. The boundary, often referred to as the event horizon, is a spherical surface, with a radius equal to 1.8 mi (3 km) times the number of solar masses in the mass of the hole. A black hole forms when an object becomes so small that its gravity prevents nearby objects from escaping from its gravitational attraction. Objects close enough to a black hole, within its horizon, would have to move faster than light in order to escape its gravitational field. Because this is impossible according to the theory of relativity, escape cannot occur. *See* BLACK HOLE.

Cosmology. In cosmology the term horizon takes on still another meaning. According to the widely accepted big bang theory, cosmic evolution began from an explosion 10 to 20 billion years ago. Because nothing can travel faster than light, there is a limit to how far an observer can see or influence the surroundings; this limit is referred to as the horizon or horizon length. For example, when the universe is a million years old, the horizon length is a million light-years or about 3×10^5 parsecs. In classical cosmology the horizon length is the distance that light can travel since the beginning of time.

The uniformity of the universe is difficult to understand, considering the small size of the horizon when the universe was young. How is it possible that twos of the universe which were separated by several million light-years or several horizon lengths when the universe was a million years old were at the same temperature at that time? Neither light nor heat could be transmitted from one region to the other in the short time that had passed since the big bang, and yet temperature uniformity is shown by observation. The so-called inflationary scenario presents a slightly modified version of the early stages of the big bang model and provides an explanation. *See* BIG BANG THEORY; COSMOLOGY; INFLATIONARY UNIVERSE COSMOLOGY; UNIVERSE.

Harry L. Shipman

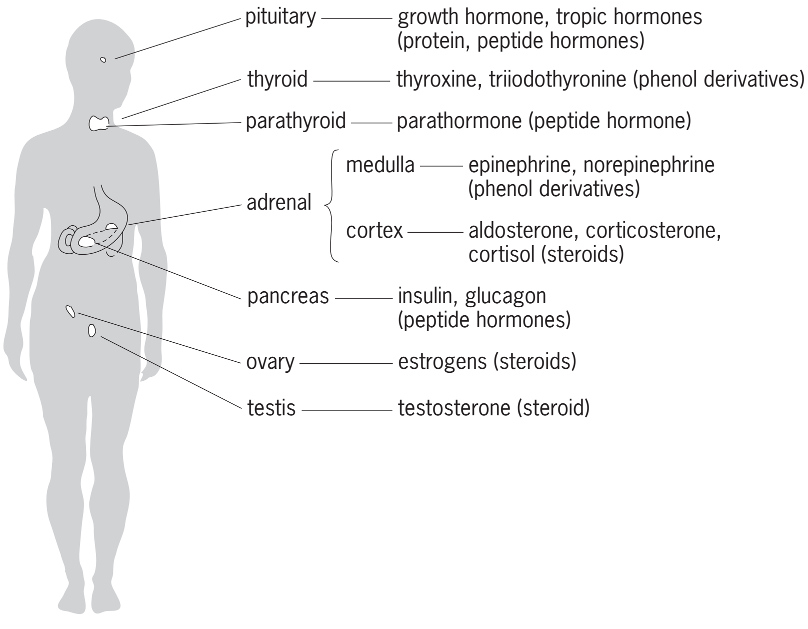
Hormone

One of the chemical messengers produced by endocrine glands, whose secretions are liberated directly into the bloodstream and transported to a distant part or parts of the body, where they exert a specific effect for the benefit of the body as a whole. The endocrine glands involved in the maintenance of normal body conditions are pituitary, thyroid, parathyroid, adrenal, pancreas, ovary, and testis (see **illustration**). However, these organs are not the only tissues concerned in the hormonal regulation of body processes. For example, the duodenal mucosa, which is not organized as an endocrine gland, elaborates a substance called secretin which stimulates the pancreas to produce its digestive juices. The placenta is also a very important hormone-producing tissue.

Classification. The hormones obtained from extracts of the endocrine glands may be classified into four groups according to their chemical constitution: (1) amino acid derivatives, such as epinephrine, norepinephrine, thyroxine, and triiodothyronine; (2) proteins, such as many of the anterior pituitary hormones; (3) peptides, such as insulin, glucagon, ACTH, vasopressin, oxytocin, and secretin; and (4) steroids, such as estrogens, androgens, progesterone, and corticoids (see **table**). Hormones, with a few exceptions like pituitary growth hormone and insulin, may also be classified as either tropic hormones or target-organ hormones. The former work indirectly through the organs or glands which they stimulate, whereas the latter exert a direct effect on peripheral tissues. Many of the target-organ hormones are steroids whose production and secretion are controlled by tropic hormones, which themselves are either proteins or peptides; for example, the target-organ steroid cortisol, from the adrenal gland, is released by the action of the tropic peptide hormone pituitary ACTH. *See* PEPTIDE; PROTEIN; STEROID.

Production and circulation. Hormones are responsible for the normal functioning of the body; thus, the organism's recurring needs call for a continuing cycle of hormonal production and circulation. The various hormones are secreted chiefly under the control of the anterior lobe of the pituitary gland, known as the master gland, and in terms of their physiological role they are all either metabolic or gonad-stimulating; that is, they are involved either in the regulation of body chemistry or in the sex cycle. A key to the physiological importance of hormones is that these substances can manifest an effect even when present in very minute or trace quantities. As little as 0.05 mg of the estrogenic hormone of the ovary will induce uterine bleeding in a woman.

Regulation. This maintenance of the normal state of body with respect to the hormonal functioning of the various endocrine glands depends upon a feedback mechanism which tends to preserve an internal balance. For example, the anterior pituitary gland and the adrenal cortex are parts of such a system. The pituitary hormone ACTH stimulates the adrenal cortex to secrete its steroid hormones. If one adrenal gland is removed, the ACTH stimulation causes the cortex of the remaining adrenal to increase in size. If, on the other hand, the anterior pituitary is removed, the consequent lack of ACTH results in considerable atrophy of the cortices of both adrenal glands. If ACTH is then injected, the adrenals regain their usual size. Under normal conditions, the concentration of adrenal cortical hormones in the bloodstream controls the secretion of ACTH by the pituitary; when this concentration is high, less ACTH is released; when the concentration is low, the output of ACTH increases. The same sort of push-and-pull mechanism is believed to operate in the secretion of sex hormones by the ovaries and testes.


pituitary
growth hormone, tropic hormones (protein, peptide hormones)
thyroid
thyroxine, triiodothyronine (phenol derivatives)
parathyroid
parathormone (peptide hormone)
adrenal
medulla
epinephrine, norepinephrine (phenol derivatives)
cortex
aldosterone, corticosterone, cortisol (steroids)
pancreas
insulin, glucagon (peptide hormones)
ovary
estrogens (steroids)
testis
testosterone (steroid)

Major endocrine glands and the hormones they secrete.

Classification of hormones released from principal endocrine glands[*]

Endocrine gland and hormone	Target tissue	Principal actions	Chemical nature
Posterior lobe of pituitary			
Antidiuretic hormone (ADH)	Kidneys	Stimulates reabsorption of water; conserves water	Peptide (9 amino acids)
Oxytocin	Uterus	Stimulates contraction	Peptide (9 amino acids)
	Mammary glands	Stimulates milk ejection	
Anterior lobe of pituitary			
Growth hormone (GH)	Many organs	Stimulates growth by promoting protein synthesis and fat breakdown	Protein
Adrenocorticotropic hormone (ACTH)	Adrenal cortex	Stimulates secretion of adrenal cortical hormones such as cortisol	Peptide (39 amino acids)
Thyroid-stimulating hormone (TSH)	Thyroid gland	Stimulates thyroxine secretion	Glycoprotein
Luteinizing hormone (LH)	Gonads	Stimulates ovulation and corpus luteum formation in females; stimulates secretion of testosterone in males	Glycoprotein
Follicle-stimulating hormone (FSH)	Gonads	Stimulates spermatogenesis in males; stimulates development of ovarian follicles in females	Glycoprotein
Prolactin (PRL)	Mammary glands	Stimulates milk production	Protein
Melanocyte-stimulating hormone (MSH)	Skin	Stimulates color change in reptiles and amphibians; unknown function in mammals	Peptide (two forms; 13 and 22 amino acids)
Thyroid gland			
Thyroxine (thyroid hormone)	Most cells	Stimulates metabolic rate; essential to normal growth and development	Iodinated amino acid
Calcitonin	Bone	Lowers blood calcium level by inhibiting loss of calcium from bone	Peptide (32 amino acids)
Parathyroid glands			
Parathyroid hormone	Bone, kidneys, digestive tract	Raises blood calcium level by stimulating bone breakdown, stimulates calcium reabsorption in kidneys; activates vitamin D	Peptide (34 amino acids)
Adrenal medulla			
Epinephrine (adrenaline) and norepinephrine (noradrenaline)	Smooth muscle, cardiac muscle, blood vessels	Initiate stress responses; raise heart rate, blood pressure, metabolic rate; dilate blood vessels; mobilize fat; raise blood glucose level	Amino acid derivatives
Adrenal cortex			
Aldosterone	Kidney tubules	Maintains proper balance of Na^+ and K^+ ions	Steroid
Cortisol	Many organs	Adaptation to long-term stress; raises blood glucose level; mobilizes fat	Steroid
Pancreas			
Insulin	Liver, skeletal muscles, adipose tissue	Lowers blood glucose level; stimulates storage of glycogen in liver	Peptide (51 amino acids)
Glucagon	Liver, adipose tissue	Raises blood glucose level; stimulates breakdown of glycogen in liver	Peptide (29 amino acids)
Ovary			
Estradiol	General	Stimulates development of secondary sex characteristics in females	Steroid
	Female reproductive structures	Stimulates growth of sex organs at puberty and monthly preparation of uterus for pregnancy	
Progesterone	Uterus	Completes preparation for pregnancy	Steroid
	Mammary glands	Stimulates development	
Testis			
Testosterone	Many organs	Stimulates development of secondary sex characteristics in males and growth spurt at puberty	Steroid
	Male reproductive structures	Stimulates development of sex organs; stimulates spermatogenesis	
Pineal gland			
Melatonin	Gonads, pigment cells	Function not well understood; influences pigmentation in some vertebrates; may control biorhythms in some animals; may influence onset of puberty in humans	Amino acid derivative

[*] Reproduced with permission from P. H. Raven and G. B. Johnson, *Biology*, 6th ed., McGraw-Hill, New York, 2002.

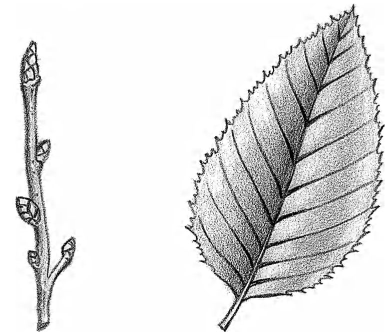
Molecular mechanisms of action. Hormones exert their effects on peripheral tissues by regulating the behavior of cells within tissues. They do this by interacting with discrete structures in cells called hormone receptors. These hormone receptors are also generally proteins, and they bind to a select hormone with a high degree of specificity. Hormone receptors are generally of two types: cell surface receptors or intracellular receptors. For a cell surface receptor, binding of its specific hormone generates a cascade of intracellular signaling messages that regulate cell behavior. This is how the insulin receptor works and how most of the tropic hormones work. For an intracellular receptor, the specific hormone must first enter the cell and bind to the intracellular receptor. After that, the hormone and the receptor travel together to the nucleus of the cell to bind to deoxyribonucleic acid (DNA) and, thereby, regulate the expression of specific genes. Thyroid hormone and the steroid hormones generally work in this manner. *See* ADENOHYPOPHYSIS HORMONE; ALDOSTERONE; ENDOCRINE MECHANISMS; EPINEPHRINE;

INSULIN; GENE; NEUROHYPOPHYSIS HORMONE; SIGNAL TRANSDUCTION. Choh Hao Li; Theodore Mazzone

Bibliography. F. F. Bolander, *Molecular Endocrinology*, Elsevier Science & Technology, 2004; F. F. Greenspan and G. J. Strewler (eds.), *Basic & Clinical Endocrinology*, 5th ed., Appleton & Lange, Stamford, CT, 1997; A. W. Norman and G. Litwack, *Hormones*, Academic Press, San Diego, 1997; C. C. Quirk and J. H. Nilson, Hormones and gene expression—Basic principles, in L. J. DeGrool and J. L. Jameson (eds.), *Endocrinology*, pp. 3–14, Saunders, New York, 2001.

Hornbeam

The genus *Carpinus* of the birch family, represented in the United States by *C. caroliniana*, the American hornbeam or blue beech. Hornbeam is a small tree sometimes attaining a height of 35 ft (10.7 m), and it has a smooth, steel-gray, fluted bark. It grows throughout the eastern half of the United States, especially in moist soil along banks of streams; it is sometimes called water beech. When mature, it is easily recognized by its peculiar bark, by the doubly serrate leaves resembling those of sweet birch, and by the small, pointed, angular winter buds with scales in four rows (see **illus.**). The fruit is a small nutlet subtended by a three-lobed serrate bract. The wood is very hard, giving rise to the name ironwood. The bole of this tree is rarely used for sawn products, because it is short and usually crooked. Leaves turn scarlet or orange in fall.



American hornbeam *Carpinus caroliniana*.

The European hornbeam (*C. betulus*) is often cultivated in parks and estates. It can be distinguished by its larger size, larger winter buds, and larger three-lobed, almost entire fruiting bracts. *See* FAGALES; FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Hornblende

The name that was traditionally assigned to common calcic amphiboles of metamorphic and igneous rocks. However, a nomenclature scheme for amphiboles was introduced in 1997 in which the names now carry strict compositional restrictions. Magnesiohornblende (contains magnesium) and ferrohornblende (contains iron) are monoclinic amphiboles with end-member compositions $Ca_2(Mg_4Al)(Si_7Al)O_{22}(OH)_2$ and $Ca_2(Fe^{2+}_4Al)(Si_7Al)O_{22}(OH)_2$, respectively (Ca = calcium, Mg = magnesium, Al = aluminum, Si = silicon, O = oxygen, Fe = iron, OH = hydroxyl). Most natural compositions differ significantly from these ideal end members. In particular, they show solid solution toward pargasite and ferropargasite with ideal end-member compositions $NaCa_2(Mg_4Al)(Si_6Al_2)O_{22}(OH)_2$ and $NaCa_2(Fe^{2+}_4Al)(Si_6Al_2)O_{22}(OH)_2$ [Na = sodium]. Significant deviations from these compositions are denoted by the addition and replacement of prefixes and adjectival modifiers characteristic of the compositions involved. Thus fluorohornblende (contains fluorine, F) has the end-member composition $Ca_2(Mg_4Al)(Si_7Al)O_{22}F_2$, in which all of the OH in hornblende has been replaced by F. When used to denote an amphibole of known chemical composition, the term hornblende is never used without a prefix or adjectival modifier. The unmodified term hornblende specifically refers to a calcic amphibole identified by physical or optical properties without characterization of the chemical composition. *See* AMPHIBOLE.

Physical and optical properties. In hand specimen, hornblende varies in color from dark green to brown to black. Crystals are normally anhedral to subhedral and prismatic, with prominent {110} faces meeting at 56° and 124°; end sections show {110} cleavage, and prism faces show longitudinal cleavage traces. In thin section, hornblende is strongly pleochroic (changes color on rotation in plane-polarized light) in various shades of green and brown. Electronic absorption spectroscopy shows that this is due primarily to intervalence charge-transfer reactions among the octahedrally coordinated transition-metal cations in the structure. The complexity of chemical substitutions possible in hornblende precludes characterization of chemical variation by optical methods. However, the textural information from this technique is important in the interpretation of hornblende paragenesis. *See* COORDINATION CHEMISTRY; CRYSTALLOGRAPHY.

Occurrence. Hornblende is a common rock-forming mineral in medium- and high-grade metamorphic rocks, particularly those of mafic and ultramafic composition. In mafic rocks, it first appears in the upper part of the low grade by a chemical reaction involving the disappearance of actinolite, a nonaluminous calcic amphibole. This change is extremely noticeable in thin sections, very pale-green actinolite giving way to blue-green hornblende. With prograde metamorphism, the composition of the hornblende gradually changes in a highly complex manner that is a function of temperature, pressure, oxygen fugacity (a measure of the activity of oxygen), and the chemical composition of the rock. This causes a gradual color change from blue-green through various shades of green to olive green and brown. At the middle of the high grade, hornblende becomes unstable and breaks down to form pyroxene (plus other minerals). The prominence of hornblende in medium-grade metabasic rocks has led to these rocks being called amphibolites. *See* PYROXENE.

Hornblende is commonly found as a minor phase in a wide variety of igneous rocks. Magnesium-rich hornblendes do occur as primary phases in basic and ultrabasic rocks, but this is not common. Igneous amphiboles are most abundant in calcic-alkaline diorites, granodiorites, and granites, becoming more iron-rich with increasing acidity of the host rock. This compositional trend is also characterized by a progressive increase in the alkali content of the amphibole, and hornblende grades into hastingsite, riebeckite, and arfvedsonite in granitic rocks. *See* GRANITE; GRANODIORITE; IGNEOUS ROCKS.

Due to its complex structure and chemistry, hornblende contains much information on its formation. Its behavior is understood reasonably well, and hornblende is of considerable use in interpreting the geological history of the rocks in which it occurs. *See* AMPHIBOLITE; METAMORPHISM.

Exsolution. For temperatures at which the amphiboles form, amphiboles of intermediate composition occur. These are unstable at lower temperatures and unmix, forming spectacular exsolution textures that are very useful in deciphering details of the thermal history of the host rock. Hornblende shows partial to complete miscibility with other amphibole subgroups, in particular the ferromagnesian and the alkali amphiboles. In medium-grade metamorphic rocks, coexisting hornblende and cummingtonite-grunerite is very common, and hornblende is characterized by prominent exsolution lamellae of cummingtonite-grunerite. Details of the orientation of these exsolution lamellae can give considerable information about the cooling history of the rock. *See* CUMMINGTONITE.

Frank C. Hawthorne

Bibliography. F. C. Hawthorne, The crystal chemistry of the amphiboles, *Can. Mineral.*, 21:173–480, 1983; F. C. Hawthorne, The crystal chemistry of the amphiboles, *Rev. Mineral.*, 9A:1–102, 1982; B. E. Leake et al., Nomenclature of amphiboles, *Can. Mineral.*, 35:219–246, 1997; P. Robinson et al., Phase relations of metamorphic amphiboles: Natural occurrence and theory, *Rev. Mineral.*, 9B:1–227, 1982.

Hornfels

A metamorphic rock that has been subjected to heating during contact metamorphism around intrusive igneous rocks. Hornfels is typically fine-grained, although where it is subjected to high temperatures, large crystals called porphyroblasts can form. In outcrop, hornfels is notoriously tough and can be difficult to sample. Mineral grains in hornfels are randomly oriented, with no preferred alignment of crystals to form foliation or cleavage planes. This texture indicates that the hornfels was not subjected to significant stresses during contact metamorphism.

Hornfels generally originates from sediments that undergo mineralogical changes, the nature of which depend on the magnitude of heating. The types of minerals that form are strongly dependent on the bulk composition. Minerals in hornfels formed from metamorphism of limestones, which are rich in calcium oxide (CaO), carbon dioxide (CO_2), and various amounts of magnesium oxide (MgO), iron oxide (FeO), and aluminum oxide (Al_2O_3), include (from high to low temperature) fosterite, diopside, tremolite, talc, and brucite. Other minerals that may be present include wollastonite, vesuvianite, anorthite, and grossular garnet, depending on the bulk composition of the rock. These minerals are also common in skarn-type deposits commonly found within contact metamorphic environments. *See* LIMESTONE; SKARN.

Pelitic sediments are rich in chemical constituents such as silicon dioxide (SiO_2), Al_2O_3, MgO, FeO, potassium oxide (K_2O), and water (H_2O), with relatively minor amounts of CaO, sodium oxide (Na_2O), manganese oxide (MnO), and titanium dioxide (TiO_2). Metamorphism of these sediments to form hornfels results in formation of minerals such as chlorite, muscovite, biotite, andalusite, sillimanite, cordierite, garnet, staurolite, and K-feldspar. At extremely high temperatures ($>800^\circ$C or 1470°F) aluminum-rich minerals such as sapphirine, spinel, and corundum form. Deposits of emery, utilized for abrasives, are aluminum-rich hornfels that are products of high-temperature contact metamorphism. Chemical study of emeries indicates a general lack of alkali elements (K, Na, and Ca), which has been used to argue that they form as a result of extraction of a melt phase during high-temperature contact metamorphism. Examples of emery deposits include the Cortlandt Complex in New York; the Martinsville Intrusive Complex in Virginia; and localities in Turkey, Siberia, and Australia. *See* EMERY.

During contact metamorphism, hornfels typically forms in the highest-temperature part of aureoles adjacent to the pluton. Further away from the pluton, metamorphism of sediments results in development of schists and phyllites. For example, the Ballachulish Igneous Complex in Scotland is a composite pluton consisting of granite and granodiorite. Around the pluton, low-grade chlorite-bearing slates are progressively metamorphosed, resulting in the systematic appearance from low to higher temperature of cordierite + biotite + muscovite phyllite to cordierite + K-feldspar + biotite hornfels. In hornfels of a slightly different composition, muscovite is preserved, resulting in a hornfels with the composition andalusite + K-feldspar + cordierite + biotite + muscovite. Adjacent to the contact with the pluton, these muscovite-bearing hornfels undergo partial melting, resulting in the segregation of K-feldspar + plagioclase + quartz from the metamorphosed sediment as a result of partial melting. *See* PLUTON; SLATE.

Metamorphic studies of hornfels provide an important avenue to documenting the temperature and, in particular, the pressure (that is, the depth) during emplacement of the intrusive igneous rock that provides the heat. For example, for hornfels metamorphosed at low pressures [<2 kilobars (200 megapascals) or ∼7 km (4 mi)] the predominant porphyroblasts present in hornfels are predicted to be cordierite + K-feldspar + andalusite or sillimanite. The presence of andalusite versus sillimanite is a function of temperature, with sillimanite favored at higher temperatures. At pressures greater than 2 kbar (200 MPa) but less than 3.5 kbar (350 MPa), staurolite is found in addition to cordierite + K-feldspar + andalusite or sillimanite. A pressure zone of 3.5–4 kbar (350–400 MPa) is distinctive because of the presence of garnet, and at still higher pressures (>4 kbar or 400 MPa) is marked by the lack of andalusite and the predominance of kyanite or sillimanite along with garnet and staurolite. The utility of this pressure scheme is that the depth at which sedimentary rocks were subjected to contact metamorphism can be evaluated by determining the metamorphic minerals in hornfels. This is a routine endeavor completed during field mapping and petrographic examination of hornfels. One difficulty in this technique is that the mineralogy used to distinguish between pressure zones is also a function of bulk rock chemistry, which means that the growth of a particular metamorphic mineral may vary within a sequence of metamorphosed sedimentary rocks. *See* METAMORPHIC ROCKS; METAMORPHISM; MINERALOGY.

Matthew W. Nyman

Bibliography. D. M. Carmichael, Metamorphic bathozones and bathograds: A measure of the depth and postmetamorphic uplift and erosion on the regional scale, *Amer. J. Sci.*, 278:769–797, 1978; D. M. Kerrick (ed.), *Contact Metamorphism*, Mineralogical Society of America, 1991.

Horology

Measurement of the time dimension. In practice, horology is the search for a steady or repetitive action, and the design of an instrument to perform that action and to indicate (read out) a measure of the action. Until early in the twentieth century, horology dealt with mechanical instruments, with effort distributed between improving accuracy and decreasing size of timepieces. Increasingly, however, electronic instruments have provided means for meeting these objectives. *See* ATOMIC CLOCK; ATOMIC TIME; CHRONOMETER; CLOCK (MECHANICAL); DYNAMICAL TIME; EARTH ROTATION AND ORBITAL MOTION; ESCAPEMENT; MASER; OSCILLATOR; PENDULUM; PULSAR; QUARTZ CLOCK; SATELLITE NAVIGATION SYSTEMS; SUNDIAL; TIME; WATCH.

Frank H. Rockett; Donald C. Backer

Bibliography. J. E. Burnett, *Time's Pendulum*, 1998; J. T. Fraser et al. (eds.), *The Study of Time,* vols. 1–10, 1972–2000; J. L. Jesperson and D. W. Hanson, Special issue on time and frequency, *Proc. IEEE*, 79(7):894–1079, 1991; J. L. Jesperson and J. Fitz-Randolph, *From Sundials to Atomic Clocks: Understanding Time and Frequency*, 2d ed., 2000; P. Kartaschoff, *Frequency and Time*, 1978; J. D. Kraus, *Radio Astronomy*, 2d. ed., 1986; A. R. Thompson, J. M. Moran, and G. W. Swenson, *Interferometry and Synthesis in Radio Astronomy*, 2d ed., 2001; M. E. Whitney, *The Ship's Chronometer*, 1984.

Horse production

The science of breeding, raising, and caring for the horse. Successful horse production depends upon the use of the most recent advances in breeding methods such as artificial insemination and embryo transfer, feeding, physical conditioning, genetics, health care, marketing, and general management.

Breeds and types. Horses are grouped in four classes—ponies, light horses, draft horses, and coach horses—based upon their height and size. Height is measured at the horse's shoulder by the standard unit of the hand, which is 4 in. (10 cm). Ponies are less than 14.2 hands, and most weigh less than 900 lb (408 kg). Horses are over 14.2 hands and are classified further as light or draft based upon their size: light horses usually weigh 900–1400 lb (408–635 kg), and draft horses weigh over 1400 lb (635 kg). Light horses are also classified according to their type and use, including the polo, racing, stock, hunter, jumper and gaited horses. Race horses include harness horses, such as Standardbred trotters and pacers; flat racers, such as Thoroughbreds and Quarter Horses; steeplechasers; and endurance racers. Driving horses include the heavy harness, fine harness, and roadster types, which are all used to hitch to some vehicle.

Within each of the classifications there are usually several breeds, each of which has certain distinguishing characteristics such as conformation, color, and sometimes function.

Pony breeds. There are four common pony breeds in the United States—Welsh, Shetland, Connemara, and Pony of America. The Welsh or Welsh Mountain Pony originated in Wales and was mainly used for work in coal mines prior to importation (**Fig. 1**). The Shetland Pony was developed in the Shetland Islands and is of small size (9–11 hands). Connemaras are the largest pony breed and were developed in Ireland. The Pony of America was developed in the United States by crossing a Shetland stallion with an Appaloosa mare; it must retain the Appaloosa color pattern to be eligible for registration.

Light horses. The Arabian horses, all of which descend from five foundation mares known as the Al-Khamesh, are the foundation for most of the light horse breeds. Arabians are distinguished from other breeds by a thin dished face and prominent forehead, and have gained recognition for their endurance.

Thoroughbred horses have been bred for their racing ability for several hundred years (**Fig. 2**). All Thoroughbreds trace back to three stallions—Byerly Turk, Darley Arabian, and Godolphin Barb—which were imported to England in the late 1600s and early

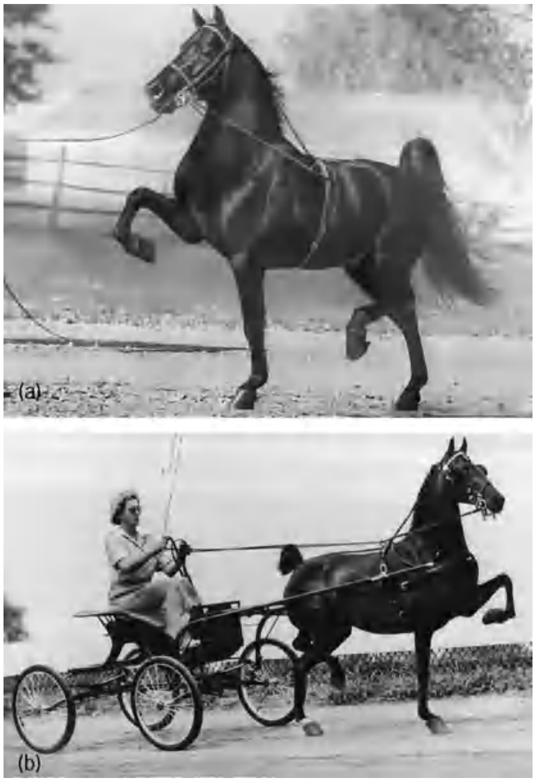


Fig. 1. Two popular pony breeds. (*a*) Welsh pony stallion. (*b*) Hackney pony mare. (*Photographs by McClasky*)

1700s, and are responsible for the Thoroughbred's refinement and speed. The Thoroughbred is now recognized for its versatility and is used for many purposes.

The American Quarter Horse is the most popular horse breed in America (**Fig. 3**), as proved by the registration of over 2,000,000 head since the establishment of the breed association in 1941. The history of the breed traces to the colonial era when these horses were used to run short races, but later they became the most popular horse for the cowboys working cattle on the western ranches. Within the breed there are several types, including working cow horses, racers, hunters, and jumpers.

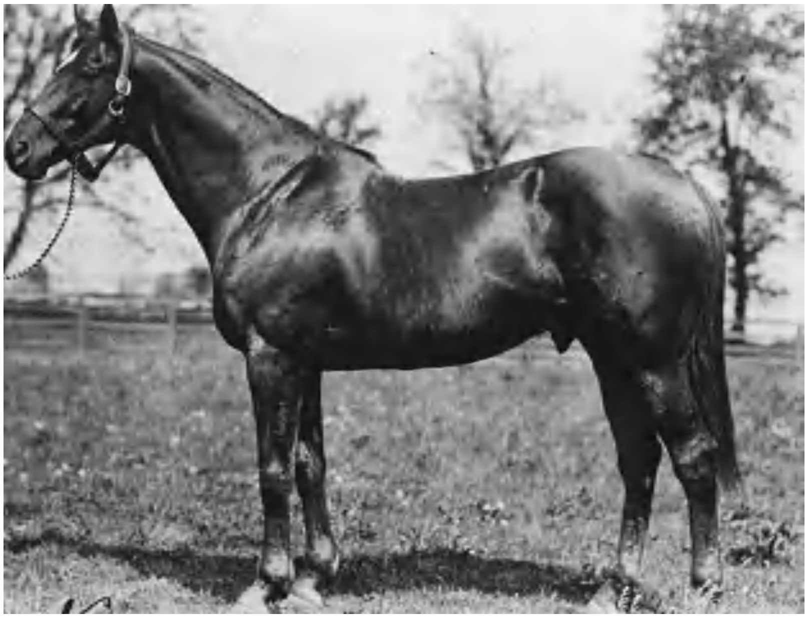


Fig. 2. The Thoroughbred. (*Photograph by The Thoroughbred Record*)

Standardbred horses were developed for their ability to race as trotters or pacers, and their name was derived from the original registration requirement of racing a mile in a standard time (**Fig. 4**). Morgan horses, known for their versatility, trace their development to one sire, Justin Morgan, and became popular on New England farms.

Various other breeds were developed for their ease of riding (**Fig. 5**). The American Saddlebreds excel as show horses and are shown as three- and five-gaited and as fine harness horses. (Five-gaited horses are shown at the rack and slow gait in addition to the walk, trot, and canter required of the three-gaited horse.) Tennessee Walking horses were developed for their ability to perform the running walk, an easy gliding gait. Breed associations whose only requirement is a specific color pattern include the White-Horse Registry, Cream Horse Registry, America Buckskin Registry Association, International Buckskin Horse Association, and Appaloosa, Pinto, Paint, and Morocco Spotted Horse associations.

Draft horses. Draft horses were developed for their pulling power, and are characterized by their size (1500–2000 lb or 680–907 kg), heavy muscling and bone, and deep, wide body (**Fig. 6**). The Percheron originated in France and has more refined characteristics than some of the other draft breeds. Suffolks were developed in England and were used exclusively for farm work; they are the only draft breed that breeds true for color—chestnut. Clydesdales are noted for their long hair on the back of their legs. Belgians, because of their action, endurance, and strength, were the primary horses used by the Roman cavalry. Shires, which also have long hair on the backs of their legs, were the least popular of the breeds used in the United States for farming.

Coach horses. Heavy harness or coach horses were used to pull coaches, and included breeds such as Hackney, Cleveland Bay, French Coach, Russian Orloff, Yorkshire, and American Carriage Horse.

Breeding practices. The horse breeding industry has become very specialized and technical. Computerized records of performance events and pedigrees have allowed the breeder to have instant access to necessary information. This information and the individuals conformation are used by breeders for selection or culling of breeding stock.

The natural breeding season of horses is during the spring and early summer. Most mares do not have estrous cycles during the winter. Fillies reach puberty at about 1 year of age and at that time start to have estrous cycles, but they are usually not bred until they are at least 3 years of age. Each estrous cycle lasts about 21–23 days. The period of estrus when a mare will accept a stallion is 5–6 days; ovulation occurs about 2 days before the end of the estrous period. If the mare is bred naturally, she is bred every other day during estrus, but if she is bred by artificial insemination, she is bred only once, just prior to ovulation. Mares are bred an average of two estrous cycles per conception. *See* ESTRUS.

Colts reach sexual maturity when they are about 15 months old. They may be used to breed 6–8 mares when they are 2 years old, but most are not used as breeding stallions until they are 4–5 years old. During a breeding season, stallions can successfully breed about 60 mares, unless artificial insemination is utilized in which case 200 or more mares may be bred by a stallion. Stallions produce 50–75 ml of semen, which contains about 15 billion spermatozoa, per ejaculation. When a mare is inseminated, 500 million actively motile spermatozoa are used.

Embryo transfer is sometimes used to obtain a foal from some mares that are infertile, or from mares that need to remain in competitive training. By this procedure the mare is bred and the embryo is flushed from the uterus when it is 6–8 days old. The embryo, which sometimes is stored frozen, is then transferred to a host mare that carries the fetus to term. Identical twins can be produced by dividing the embryo shortly after conception and placing each half in a host mare.

Colts that do not have the desired conformation or fail to meet performance expectations are castrated, after which they are known as geldings. Most colts are castrated between 6 and 12 months of age but can be castrated at any age. Very few stallions are used for pleasure or recreational purposes because of their unpredictable behavior; geldings, however, are usually docile and easily handled.

Feeding. Even though many scientific advances have been made in the understanding of the nutrient requirements of horses, the "eye of the master" determines the success of feeding a horse properly. Since there is a wide variation in the nutrients required to maintain proper physical condition of the same types of horses kept under similar circumstances, most horses are fed as individuals. Horses evolved on the open grassy plains and can obtain their nutrient requirements (as well as good exercise) from good pastures. Today, however, most horses are kept in paddocks and are fed hay for the roughage portion of their diet. *See* ANIMAL FEEDS.

Mature horses receiving a moderate amount of exercise can be maintained exclusively on hay. When exercised heavily, however, their energy requirement increases, so they are fed less hay and grain is added to their diet. The ratio of hay to grain depends on the amount of exercise a horse gets; more active horses are fed more grain. The standard hays fed to horses are alfalfa, oat, timothy, and various grasses, while the most commonly used grains are oats, corn, milo, and barley.

Young growing horses, lactating mares, and pregnant mares (in their last 90 days of pregnancy) have increased requirements for protein, minerals, and vitamins. Higher protein needs are met by feeding legume hays or protein supplements such as linseed or soybean meals, and mineral supplements such as dicalcium phosphate provide the extra calcium and phosphorus that is usually needed. If adequate vitamins are not present in the natural feeds, artificial sources are added. Salt should always be available. The average horse drinks 8–10 gallons (30–38 liters) of water per day, a requirement which increases during hot weather and with intensive exercise.

J. Warren Evans

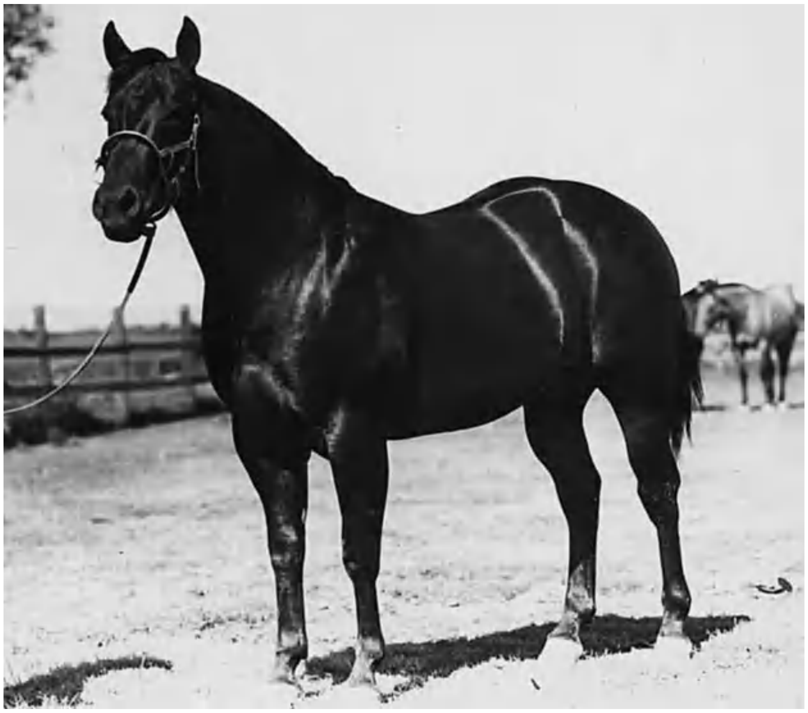


Fig. 3. American quarter horse. (*Photograph by J. A. Stryker*)

Disease and parasite control. Control measures frequently require the services of a professional veterinarian. However, good management practices can

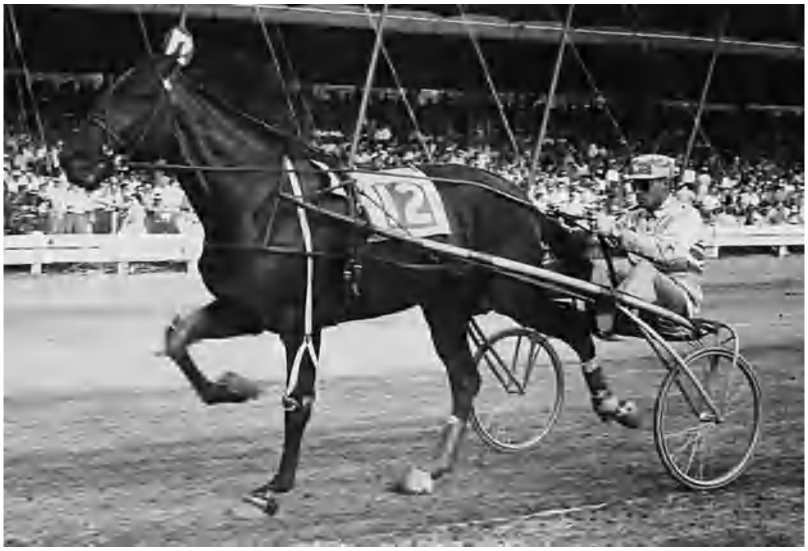


Fig. 4. Standardbred trotter. (*Photograph by The Harness Horse*)

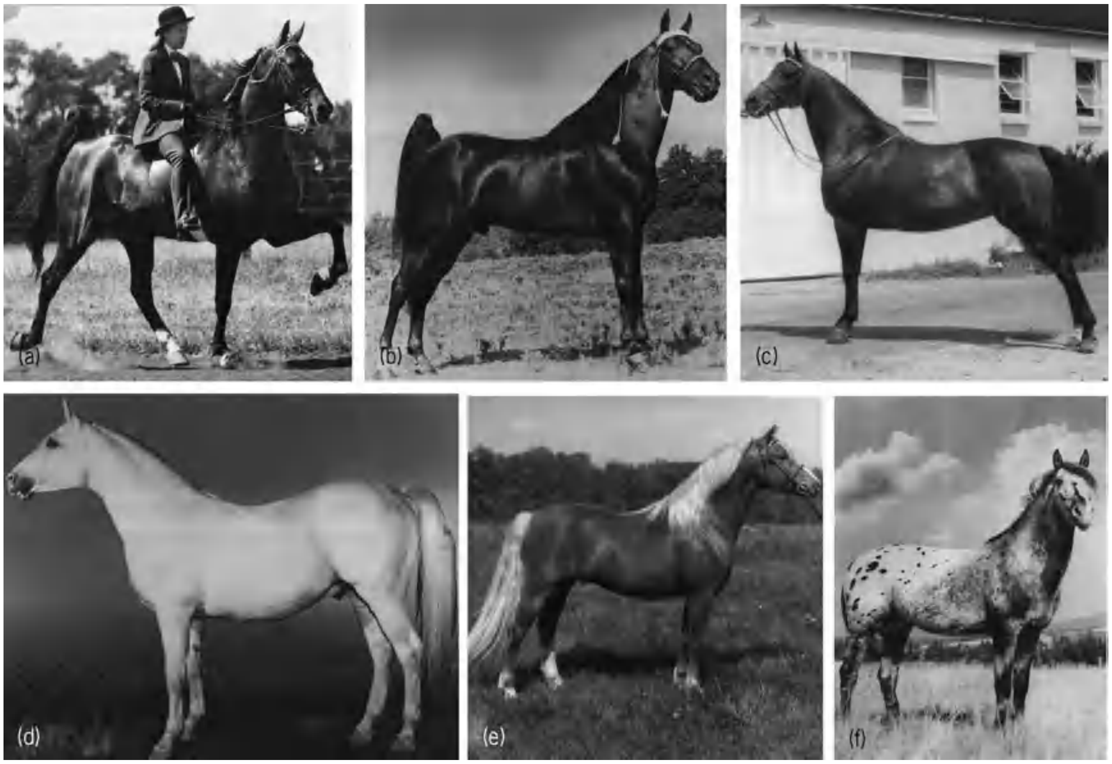


Fig. 5. Saddle horse breeds. (*a*) Five-gaited American saddle horse (*photograph by McClasky*). (*b*) Tennessee walking horse (*USDA photograph*). (*c*) Morgan mare. (*d*) Arabian stallion (*USDA photograph*). (*e*) Palomino gelding (*photograph by J. L. A. duPont*). (*f*) Appaloosa stallion (*photograph by H. H. Sheldon*).

prevent some troubles. Daily inspection may permit early detection of disease and treatment before the condition becomes serious.

The breathing of a horse should be free, soft, and noiseless. Normal respiration rate for a horse at rest is 8–16 per min. Normal pulse rate for a horse at rest is 36–40 per min. The pulse may be taken by palpating an artery that crosses the jawbone just in front of the large cheek muscle. Normal body temperature (rectal) for a horse at rest is 100°F (38°C). A horse's coat should be smooth and glossy, and its skin loose and supple. The weight should be borne equally on both front feet. The hindlegs, however, are rested alternately.

Common preliminary symptoms of disease include high temperature; fast, irregular, or noisy respiration; rapid or weak pulse; loss of appetite; sweating without known cause; stiffness; lameness; coughing; inflamed mucous membranes; discharges from nose, eyes, or genitals; diarrhea; constipation; dejection; restlessness; rolling; groaning; and heat or swelling in any part.

Internal parasites of horses must be controlled by treatment. Effective prevention of infestation is impossible. Conditions that favor good pasture also favor the completion of the life cycle of horse parasites and the likelihood of infestation.

Strongyles, or bloodworms, are the most damaging of the internal parasites of horses. These intestinal worms range from $^1/_3$ to $2^1/_2$ in. (8 to 64 mm) in length. When present in large numbers they cause anemia, weakness, and emaciation. Veterinarians often prescribe thiabendazole or cambendazole to eliminate strongyles.

Ascarids, or large roundworms, are particularly harmful to foals and young growing horses. Yellowish-white in color and up to 1 ft (30 cm) in length, ascarids stay in the small intestine. Toxins produced by these worms may kill a foal. Carbon disulfide given by means of a stomach tube eliminates roundworms. However, a piperazine–carbon disulfide complex or combendazole are more commonly used.

Bots, the larvae of the botfly, attach to the wall of the horse's stomach and interfere with the normal work of the stomach. Treatment is the same as for ascarids. *See* MYIASIS.

Lice are the most common of the external parasites of horses. Lice can be eliminated by several insecticides, of which coal tar dips and rotenone are the safest. *See* ANOPLURA; INSECTICIDE.

Sales. Selling horses profitably is the end and aim of most horse breeders. Young stock must be well bred, well fed, and well managed to command high prices. In addition, older horses must give evidence of good training and ability to perform well. Although most horses are sold privately, horse auctions are common throughout the United States. The most noted of these auctions are those for Thoroughbred yearlings held each summer at the Keeneland racetrack, Lexington, Kentucky, and at Saratoga, New York. Only a very select group of yearlings from

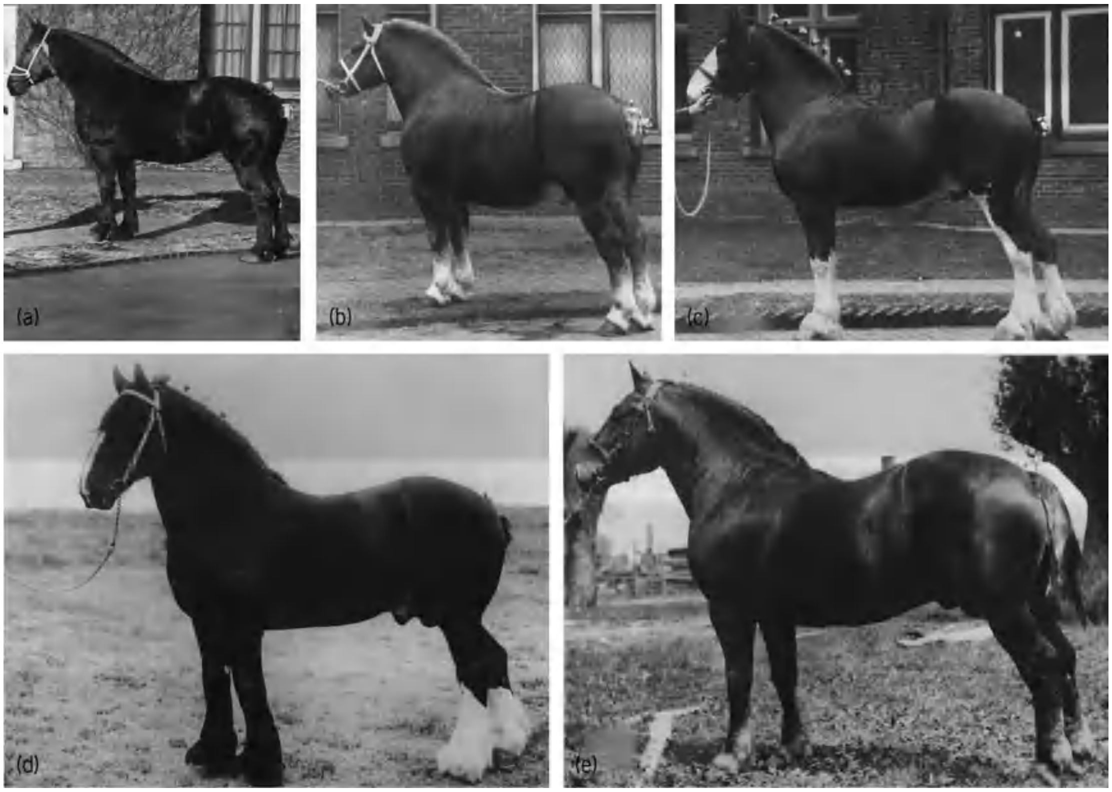


Fig. 6. Breeds of work horse. (*a*) Percheron mare; (*b*) Belgian stallion (*J. F. Abernathy Live Stock Photography Co.*). (*c*) Clydesdale stallion; (*d*) Shire stallion; (*e*) Suffolk stallion (*USDA photographs*).

the best studs is handled at these sales. Standardbred yearling sales patronized by the leading market breeders are held each fall at Lexington, Kentucky, and at Harrisburg, Pennsylvania. John M. Kays

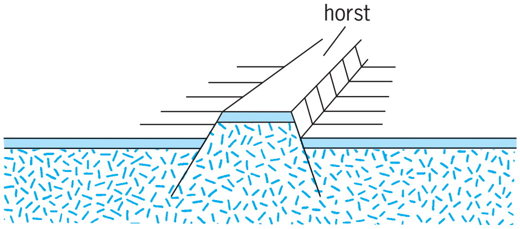
Horseradish

A hardy perennial crucifer, *Armoracia rusticana*, of eastern European origin belonging to the plant order Capparales. Horseradish is grown for its pungent roots, which are generally grated, mixed with vinegar and salt, and used as a condiment or relish. Propagation is by root cuttings, and the crop is grown like an annual. The individual roots or sets are uncovered by hand usually twice during the summer and stripped of all side roots (see **illus.**). Maliner Kren is a common variety. Harvesting of the roots occurs in the fall, usually $3^1/_2$ to 4 months after planting. Production in the United States is limited to northern areas; Illinois, Wisconsin, and Missouri are important producing states. *See* CAPPARALES. H. John Carew

Roots of horseradish stripped of side roots.

Horst

A segment of the Earth's crust, generally long as compared to its width, that has been upthrown relative to the adjacent rocks (see **illus.**). Horsts range in size from those that have lengths and upward displacement of a few inches to those that are tens of miles long with upward displacements of thousands of feet. The faults bounding a horst on either side commonly have inclinations of 50–70° toward the downthrown blocks, and the direction of



Simple horst with associated faults. (*After A. K. Lobeck, Geomorphology, McGraw-Hill, 1939*)

movement on these displacements indicates that they are gravity faults. These relationships suggest that horsts develop in regions where the crust has undergone extension. They may form in the crests of anticlines or domes, or may be related to broad regional warpings. *See* EARTH CRUST; FAULT AND FAULT STRUCTURES; GRABEN. Philip H. Osberg

Horticultural crops

Intensively managed plants cultivated for food or for esthetic purposes. Plant agriculture is divided traditionally into the fields of agronomy (herbaceous field crops, mainly grains, forages, oilseeds, and fiber crops), forestry (forest trees and products), and horticulture (garden crops, particularly fruits, vegetables, spices and herbs, and all plants grown for ornamental use). Most horticultural plants are utilized in the living state, with water essential to quality; thus most horticultural plants and products are highly perishable. *See* AGRICULTURAL SCIENCE (PLANT); AGRONOMY; FLORICULTURE; FOREST AND FORESTRY.

Classification. Custom has defined the classification of many crops. Thus field corn (maize, usually consumed as an animal feed grain) is agronomic; sweet corn (maize with sweet immature kernels consumed by humans) is horticultural. Pines grown for timber or naval stores are a forest crop; pines grown as landscape trees are a horticultural crop. Some crops that could be considered horticultural are not treated as such; tobacco is an example. Finally, many tropical species which are sometimes treated in a special category as plantation or estate crops are included as horticultural crops. These include coffee, tea, and cacao (beverage crops), plants grown for drugs and medicines (medicinal crops), tree crops grown for oils as oil palm and coconut (oil seed crops), and condiments and flavorings (spice crops).

Horticultural crops are usually classified as edibles or ornamentals. Edible crops which are used for direct human consumption are commonly subdivided into fruits or vegetables, but this classification is traditional and difficult to define precisely.

Fruit crops. Fruit crops in the horticultural sense are cultivated for tissues associated with the botanical fruit, that is, seed-bearing structures derived from the flower, which are usually pulpy and tasteful.

Fruit plants are typically woody perennials, but there are exceptions (strawberry, banana). Temperate fruit crops are deciduous; subtropical and tropical fruits are usually evergreen. Fruits borne on low-growing shrubs, vines, or herbaceous plants are known as small fruits. Trees or shrubs bearing nuts, characterized by a hard shell separated by a firm inner kernel (the seed), are often treated as a special category of fruit crops.

A traditional classification of fruit crops (with some examples) is presented below.

- Temperate fruits
  - Small fruits
    - Berries (blueberry, cranberry, strawberry)
    - Brambles (blackberry, raspberry)
    - Vine fruits (grape, kiwifruit)
  - Tree fruits
    - Pome fruits (apple, pear, quince)
    - Stone fruits (apricot, cherry, peach, plum)
    - Nuts (almond, chestnut, filbert, pecan, walnut, pistachio)
- Subtropical and tropical fruits
  - Small fruits
    - Vine fruits (passion fruit)
    - Herbaceous fruits (banana, papaya, pineapple)
  - Tree fruits
    - Citrus fruits (grapefruit, lemon, lime, orange, pummelo, tangerine)
    - Other fruits (avocado, date, fig, mango, mangosteen)
      - Nuts (Brazil nut, cashew, macadamia)

See FRUIT; FRUIT, TREE; NUT CROP CULTURE.

Vegetable crops. Vegetable crops in the horticulture sense are commonly herbaceous plants grown as annuals or biennials and occasionally as perennials that have edible parts (including, confusingly, the botanical fruit). Examples of edible parts include the root (sweet potato), tuber (potato), young shoot (asparagus), leaf (spinach), flower buds (cauliflower), fruit (tomato), and seed (pea). Temperate-tropical distinctions are not as important in vegetables as in fruit crops. Many tropical perennials (for example, tomato) are grown as annuals in temperate areas. A classification of vegetables based on use and botanical affinity is listed below.

Cole crops (broccoli, cabbage, cauliflower)
Legumes (pulse) crops (bean, lentil, pea)
Solanaceous fruit (chili pepper, eggplant, tomato)
Vine crops or cucurbits (cucumber, muskmelon, squash, watermelon)
Greens (chard, spinach)
Salad crops (celery, lettuce, parsley)
Root crops (beet, carrot, sweet potato)
Tuber crops (Jerusalem artichoke, potato)
Bulb and corms (garlic, onion, shallot)
Herbs (rosemary, sage)
Mushrooms [*Agaricus, Lentinus* (shiitake)]

Ornamental crops. Plants grown for ornamental use, such as cut flowers, bedding plants, interior foliage plants, or landscape plants, represent an enormous group and include thousands of species. They may be grouped as follows.

- Flowers and bedding plants:
  - True annuals (marigold, petunia, zinnia)
  - Biennials (English daisy, foxglove)
  - Perennials (daylily, delphinium, iris, peony, rose)
- Bulbs and corms (crocus, gladiolus, narcissus, tulip)
- Foliage (interior) plants (philodendron, sansevieria)
- Ground covers and vines (English ivy, Japanese spurge, myrtle)

Lawn (turf) plants (Bermuda grass, bluegrass, fescue, perennial ryegrass)
Evergreen shrubs and trees:
Broadleaf (holly, rhododendron)
Narrowleaf (fir, juniper, yew)
Deciduous shrubs (dogwood, forsythia, lilac, viburnum)
Deciduous trees (ash, crabapple, magnolia, maple)

See ORNAMENTAL PLANTS. Jules Janick

Hospital infections

Infections acquired during a hospital stay; also known as nosocomial infections. They may be recognized during or after hospitalization. They usually appear during hospitalization, but as many as 25% of infections related to surgery occur after discharge. Data suggest that in the United States 4.5% of all hospitalized patients can expect at least one such infection, which translates into almost 2 million infections annually. Thus this major public health problem accounts for 299 million patient-days in United States hospitals each year.

The mere presence of microorganisms in or on a patient does not indicate an infection and may only represent colonization. Colonization implies establishment, growth, and multiplication of organisms (such as normal bacteria in the mouth), but without clinical symptoms. Although viruses, bacteria, and fungi are always present in the human body, infection occurs as a result of some alteration in the normal balance between an infectious organism and a susceptible person. Factors that may influence the development of an infection include the individual's defense mechanisms and the site of entry of the microorganism. Infectious agents may enter the body through breaks in the skin and mucous membranes or by way of the respiratory, gastrointestinal, or urinary tract. Resistance to infection may be influenced by age (very young or very old), chronic disease (such as diabetes and sickle-cell disease), or cancer. Genetic, hormonal, nutritional, and hygienic factors also play a role in modifying the body's defense mechanisms. Most nosocomial infections (93%) are caused by bacteria; fungi account for about 6%; and viruses, protozoa, and parasites account for the remaining 1%.

Nosocomial infections occur most frequently in the urinary tract, in surgical wounds, as a complication of pneumonia, and in association with bacteremia.

Urinary tract infection. Urinary tract infection is an inflammatory process occurring in the kidney, ureter, bladder, or adjacent structures. Microorganisms usually enter through the urethra. The infection increases in severity as it moves through the urinary tract; at any point it may lead to kidney damage, disseminated infection, bacteremia, and associated morbidity and mortality. Urinary tract infection is of major importance, not only because it accounts for almost half of all nosocomial infections, but because of the potential for serious complications. Urine is normally sterile, and so the presence of bacteria in a properly collected urine specimen is usually evidence of infection.

The most common cause of nosocomial urinary tract infections is an indwelling urinary catheter. Bacteria are believed to enter at the time of catheter insertion or later by migration along the catheter wall into the bladder. *Escherichia coli*, an aerobic gram-negative bacillus, is responsible for about 80% of these infections. They can be prevented by maintaining asepsis during insertion and removal of urinary catheters and removing them at the earliest opportunity. *See* URINARY TRACT DISORDERS.

Surgical wound infections. Surgery is associated with more than 40% of all hospital admissions but with more than 70% of all nosocomial infections, which are not necessarily confined to the surgical wound itself. Almost 75% of all pneumonias, 56% of all urinary tract infections, and 54% of all cases of bacteremia occur in surgical patients.

Approximately 5% of all incisions become infected, but the incidence varies widely with the type of operation and the health of the patient. The risk of a surgical wound infection increases with increasing age of the patient, length of the operation, and duration of hospital stay before surgery. Infection rates are lowest in incisions made through clean tissue, that is not inflamed, infected, or contaminated, and highest when the tissue is already infected or contaminated, as in cases of trauma or ruptured appendix.

Staphylococcus aureus is the most common bacterium found in surgical wound infections, but *E. coli* is also common in such wounds. Other bacilli, such as *Pseudomonas aeruginosa* and *Serratia marcescens*, are seen in debilitated patients. The principal source of infection is the patient's own flora, and these microorganisms probably gain entrance during surgery. Some microorganisms are, however, transmitted from operating-room personnel, possibly through microscopic holes in surgical gloves. Airborne infections occur rarely and only under unusual circumstances. *See* STAPHYLOCOCCUS; SURGERY.

Respiratory infections. Pneumonia accounts for about 10% of all hospital-acquired infections, in part because the normal respiratory defense mechanisms are often weakened in critical illnesses. Anesthesia, alcohol intoxication, or convulsions may suppress the normal cough reflex and permit material to be aspirated into the lungs. Fluid accumulations in the lungs from congestive heart failure, chest injuries, or viral infection may encourage bacterial growth and lead to pneumonia.

Shortly after admission to a hospital, the patient's upper respiratory tract becomes colonized with microorganisms unique to the hospital environment and, in large measure, resistant to antibiotics commonly used there; that obviously complicates treatment. Furthermore, ventilatory support equipment may harbor bacteria that increase the risk to the patient. Such respiratory infections are commonly caused by *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*. *See* PNEUMONIA.

Bacteremia. Bacteremia, the invasion of the normally sterile bloodstream by bacteria, is a serious condition that may rapidly lead to shock and death. Invasion from an infection elsewhere in the body is termed secondary bacteremia. If no such source can be found, the bacteremia is classified as primary. Although primary bacteremia accounts for no more than 5% of all nosocomial infections, it is important because of its seriousness and its preventability. Many if not most cases are related to intravenous fluid therapy, which is administered to more than half of all patients at some time during their hospital stay.

Some outbreaks of bacteremia in the United States have been caused by contamination of commercially prepared intravenous fluids, but the principal source of infection is the insertion site of the needle or catheter. Small, short-bore steel needles are safer than the longer and larger plastic cannulae, but in either case the risk is correlated with the duration of use. Sterile insertion and proper maintenance of intravenous needles and their sites are of the utmost importance.

Other causes and consequences. Some antibiotics have direct toxic effects. Some drugs and (rarely) antibiotics injure the bone marrow, reducing the production of white blood cells needed to fight infection.

About 25% of hospital patients are given antibiotics, all of which cause major alterations in the microbial populations of the skin and the gastrointestinal and respiratory tracts. Infections with antibiotic resistant organisms may be more difficult to treat. *See* ANTIBIOTIC.

Acquired immune deficiency syndrome. Viral infections, although sometimes difficult to recognize, are becoming an increasing concern in hospitals, particularly because of the possibility of transmitting the human immunodeficiency virus (HIV), the causative organism for acquired immune deficiency syndrome (AIDS). The virus may be transmitted within the hospital through transfusion of infected blood or blood products, although the screening of all donated blood for HIV antibodies makes it extremely unlikely. Hospital workers are at risk for acquiring the virus from needle pricks only if the needle has contacted HIV-infected blood. Since the virus is not transmitted by the respiratory route or by casual contact, no one is endangered simply by being in the same hospital with HIV-infected patients. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); INFECTION; MEDICAL BACTERIOLOGY. Bruce B. Dan

Hot spots (geology)

The surface manifestations of plumes, that is, columns of hot material, that rise from deep in the Earth's mantle. Hot spots are widely distributed around the Earth (**Fig. 1**). One of their characteristics is an abundance of volcanic activity which persists for long time periods (greater than 1 million years). When the lithosphere (the rigid outer layer of the Earth) moves over a plume, a chain of volcanoes is left behind that progressively increases in age along its length. Hot spots are believed to be fixed with respect to each other and the deep mantle so that the age and orientation of these chains provide information on the absolute motions of the tectonic plates. *See* LITHOSPHERE; PLATE TECTONICS.

Volcanic expression. The Hawaiian-Emperor seamount chain in the central Pacific Ocean is a good example of a volcanic chain that was generated at a hot spot. The 3400-mi-long (5700-km) chain is made up mainly of tholeiitic lavas and ash tuff and pumice deposits. The lavas may have evolved from an initial submarine shield-building stage, through an explosive stage as they build up to sea level and finally to a subaerial posterosional stage. The largest volumes of lava formed during the shield-building stage, when extrusive processes built massive pillow and sheet lava flows, up to 3 mi (5 km) in height and 60 mi (100 km) in width, above the mean sea-floor depth. *See* LAVA; SEAMOUNT AND GUYOT.

Not all hot-spot volcanism is expressed in terms of highly lineated, multistage, volcanic chains. The Réunion hot spot, for example generated the Deccan Traps, which extend laterally for thousands of kilometers over the Indian subcontinent. Large basalt provinces have also been mapped at sea: the seaward-dipping reflector sequences of the Norwegian and Greenland passive margins and the deep-water volcanic sill complexes of the eastern Atlantic and Pacific oceans are examples. Aseismic ridges that extend up to or close to the axes of mid-oceanic ridges are another example of hot-spot volcanism. When a hot spot (for example, Iceland) is centered on the axis, pairs of ridges such as the Iceland-Faeroes Rise and the Greenland Rise are formed. Sometimes the plate (for example, Africa) has migrated off the hot spot (such as Tristan da Cunha), leaving behind ridge systems that no longer extend to the ridge axis (such as Rio Grande Rise and Western Walvis). *See* MID-OCEANIC RIDGE; VOLCANO.

Topographic swells. Another characteristic of hot spots is their association with broad swells in the Earth's topography. At Hawaii, for example, the swell reaches widths of 900 mi (1500 km) and heights of up to 0.7 mi (1.2 km) above normal depths of the sea floor (**Fig. 2**). The Hawaiian hot-spot swell is believed to have been formed in response to either thermal or dynamic effects in an underlying mantle plume. The crustal and upper-mantle structure, which is constrained by seismic refraction data, shows that the oceanic crust is of uniform thickness beneath the swell. The long-wavelength correlation that is observed between the gravity anomaly and the topography (about 37 mGal mi^{-1} or 22 mGal km^{-1}; Fig. 2) indicates that the mass excess of the swell is compensated by a low-density, high-temperature region below the crust. Swells of similar dimensions are found around Bermuda, Cape Verde, and Réunion.

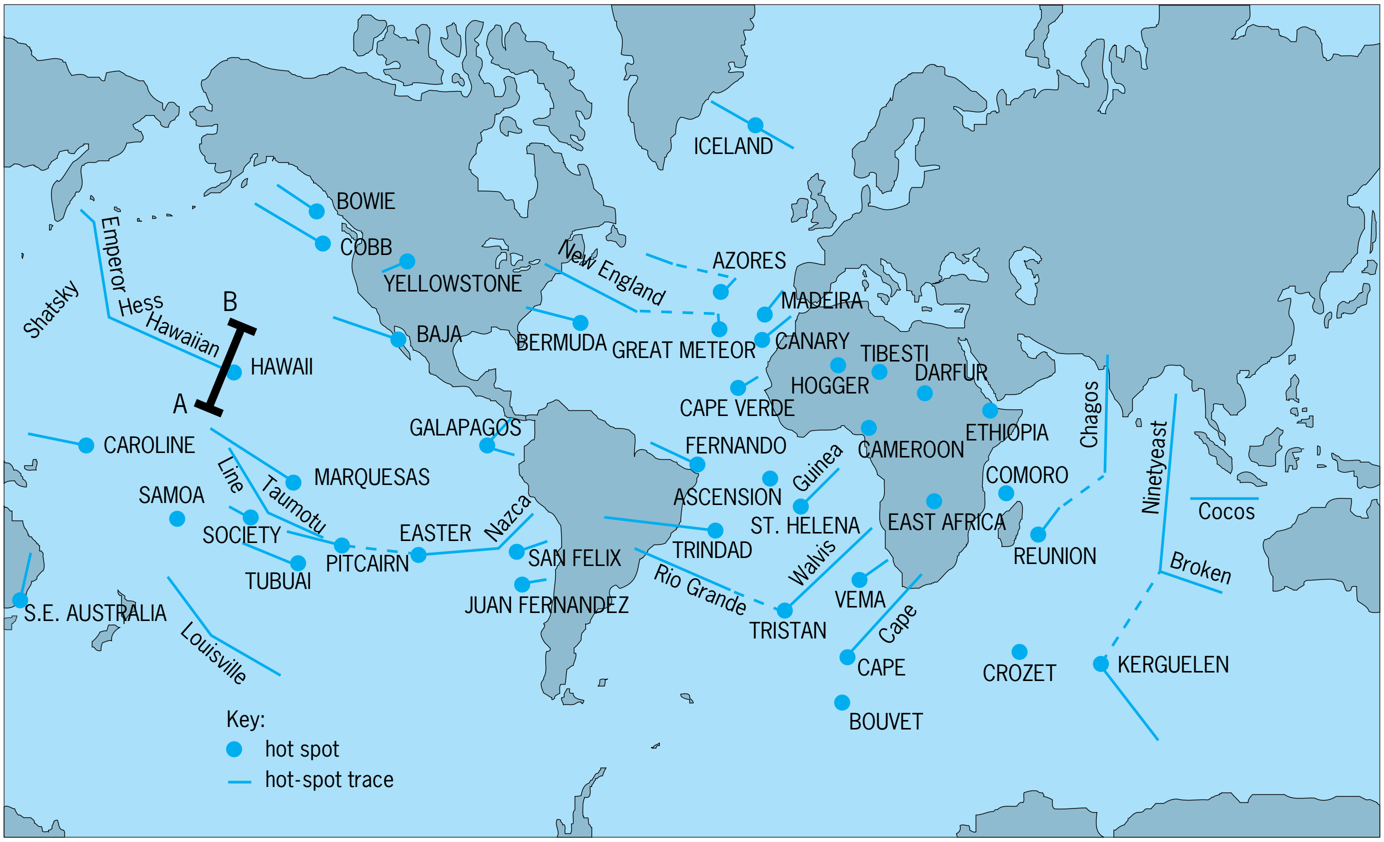


Fig. 1. Global distribution of hot spots and hot-spot traces. The broken lines indicate gaps in the traces, usually because of movement of a ridge over a hot spot. Line AB indicates the approximate location of the profile shown in Fig. 2. (*After R. S. Detrick and S. T. Crough, Island subsidence, hot spots, and lithospheric thinning, J. Geophys. Res., 83: 1236–1244, 1978*)

Some swells (for example, French Polynesia) are very large, reaching more than 1200 mi (2000 km) across, and may involve more than one hot spot (for example, Tubai, Society, Pitcairn, and possibly the Marquesas). The lack of a suitable reference complicates the identification of swells in continental regions. Swells similar in size to Hawaii have, however, been described around the volcanic centers of Dafur, Hoggar, and Tibesti in northern Africa. Others have been described from elsewhere in Africa and North America, where they may have influenced the development of sedimentary basins.

The uplift of hot-spot swells is believed to result from thermal perturbations in the underlying plume. The excess heights of swells suggest, on isostatic grounds, that temperature differences of about 450°F (250°C) occur between the plume and the surrounding mantle. Theoretical studies suggest that plumes are relatively narrow (about 180 mi or 300 km) when they begin their ascent through the mantle, but they widen to about 600–1200 mi (1000–2000 km) as they spread out beneath the lithosphere (**Fig. 3**). Hot ascending plumes may raise the temperature of the overlying lithosphere, thereby thinning it. The mechanism of thinning is not well understood, however, and it may involve either conductive heating of preexisting lithosphere or the removal of lithospheric material by some form of mantle flow.

Models. Two classes of models have been proposed to explain hot-spot swells: the reheating and dynamic models. In the reheating model swell, uplift is produced by thermal expansion that is confined to the conducting portion of the lithosphere (the thermal boundary layer). In the dynamic model, however, there is a contribution to the uplift that is produced by vertical normal stresses exerted to the seismically defined base of the lithosphere (the mechanical boundary layer) by convection. Unfortunately, it is difficult to use surface observations to distinguish between these models. Geoid anomaly data suggest, for example, that the excess mass of swells is compensated by low-density, high-temperature material at depth. However, these data indicate only the depth of the greatest temperature differences. They are therefore indeterminate as to whether the temperature differences are confined to within the lithosphere as predicted by the reheating model, or whether they extend below as implied by the dynamic model.

The main distinguishing feature between the different uplift models is that the reheating model predicts a higher heat flow than the dynamic model. Discrimination between these models therefore depends on how the subsidence history, heat flow, and long-term strength (which is controlled mainly by the temperature) differ from those for unperturbed lithosphere of the same age. Many seamounts and oceanic islands show a subsidence history that follows closely what would be expected for the age of the lithosphere on which they were emplaced.

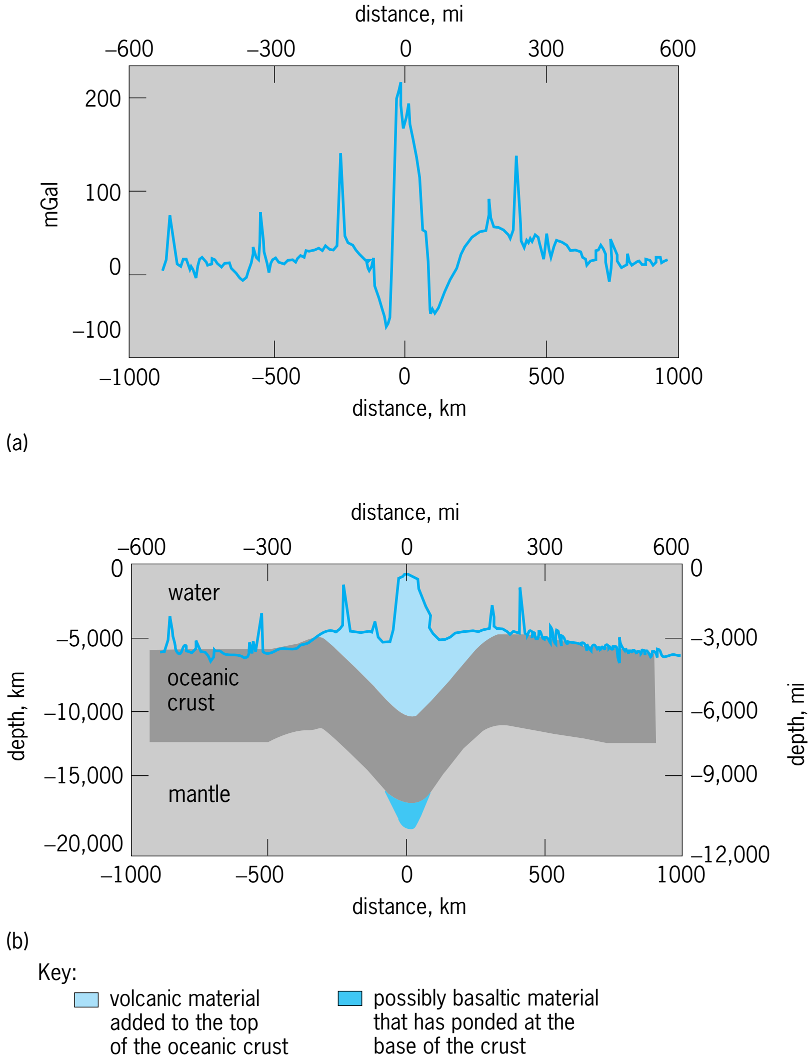


Fig. 2. Profile of the Hawaiian hot spot that crosses the Hawaiian-Emperor seamount chain in the region of Oahu. (*a*) Free-air gravity anomaly. (*b*) Crustal and upper-mantle model.

Others, however, show a much greater subsidence. The subsidence history of Eniwetok and Bikini atolls in the Western Pacific, for example, is more typical of 25-million-year-old oceanic lithosphere than the 90-million-year-old lithosphere on which they were emplaced. This is supportive of a reheating model.

Heat-flow and elastic-thickness data, however, are not so easily explained by the reheating model. At Hawaii, for example, the measured heat flow is not significantly higher over the swell than the surrounding sea floor. Furthermore, the elastic thickness [a measure of the long-term (greater than 1-million years) strength of the lithosphere] is close to what would be expected on the basis of the tectonic-plate age. One explanation is that the perturbed heat has been stored in the lithosphere but has moved away with it as the Pacific plate (which is moving at rates greater than 100 mm yr^{-1}) migrated over the Hawaiian plume. If this is so, the effects of reheating should be more apparent on slow-moving plates such as Africa, since there should been enough time for the plate to have been heated. The Cape Verdes, for example, have a relatively low elastic thickness (about 6 mi or 10 km less than expected), and high heat flow (about 20 mW m^2 more than expected). These data can be explained if the 125-million-year-old oceanic lithosphere that underlies the Cape Verdes has been thermally reset by a hot spot to a thermal age of about 60–80 million years. The corresponding change in the depth of the sea floor, however, amounts only to about 1650–2700 ft (500–820 m), which is significantly less than the observed swell height of 6900 ft (2100 m).

The reheating model is therefore unable to explain the full height of the Cape Verdes swell, suggesting that other factors, such as dynamic effects, may be involved in its support. Finally, some hot-spot swells (for example, French Polynesia) show both normal and low elastic-thickness values. These observations are difficult to explain by either of the uplift models, suggesting instead a high degree of volcano individuality and the possibility of local rather than regional controls on some of the geophysical parameters that are used to characterize hot spots.

Chemical characteristics of volcanoes. In contrast to the chemically depleted basalts of the mid-oceanic ridges, hot-spot volcanoes have a higher proportion of those elements that tend to enter the liquid phase first when the solid mantle is heated, as compared to mid-oceanic ridge magmas that have experienced the same degree of crystal fractionation. This observation suggests that mid-oceanic basalts are derived from a mantle source that has previously undergone a melt extraction event or that the hot-spot magma source has been enriched in these so-called incompatible elements by some other mechanism. A higher concentration of radiogenic elements in hot-spot volcanoes also points to a magma source that is chemically distinct from that of mid-oceanic ridges. The chemical variation within hot-spot volcanoes, which reflects the heterogeneous nature of mantle sources and the interaction of chemically enriched melts with depleted mantle and crustal-derived melts, has been described in terms of a number of end-member types, which include depleted mantle, enriched mantle, and material with a component possessing a high ratio of uranium-238 to lead-204 ($^{238}U/^{204}Pb$) known as HIMU. One of the best known of the enriched mantle components, known as DUPAL, occurs in the French Polynesia superswell region, suggesting a link between them. HIMU components are also found within the superswell region. Possible sources for these components include subcontinental lithosphere (DUPAL) and subducted oceanic lithosphere (HIMU) now residing at the lower-upper mantle or core-mantle boundary that has become entrained in the convecting upper mantle before being involved in melting. *See* EARTH INTERIOR.

Material additions. Although the temperature differences in plumes are relatively small and imply small amounts of melting, the fact that mantle material is continually being fed through the melting region of the plume means that substantial volumes of magma can be generated by hot spots. Seismic refraction data, which indicate the amount of volcanic material that has been added to the top and bottom of the crust, indicate large variations in the production rates between hot-spot volcanoes. For example, rates that vary from 0.005 mi^3 (0.02 km^3) yr^{-1} for the Canary Islands to about 0.07–0.1 mi^3 (0.3–0.5 km^3) yr^{-1} for Réunion. The Canary Islands and Réunion were formed on old and young lithosphere, respectively, suggesting that the plume flux may depend on the thermal age of the overlying lithosphere: thick lithosphere inhibiting melt production, thin lithosphere encouraging it. One explanation of this observation is that plumes may be sites of decompression melting. If the lithosphere is thick, little melt is produced because insufficient decompression melting can occur. When the lithosphere is thin, a plume may develop beneath young oceanic lithosphere (that is, at a ridge crest), and then large quantities of melt can be produced. *See* MAGMA.

There is a close association between magmatism and extension in the Earth's crust, such as continental margins that formed as a result of continental break-up. Some magmatism is to be expected at continental margins because of decompression melting. However, some margins (for example, Norway, Greenland, and Britain) are associated with such large thicknesses of volcanic material (up to 6 mi or 10 km) that they require high temperatures and high upwelling rates in order to explain them. Because of their association with Iceland, it has been proposed that the volcanic material at these so-called volcanic margins was generated at a hot spot. *See* CONTINENTAL MARGIN.

An important problem is the relative timing of initial rifting and the volcanism. The earliest volcanism may occur as the plume is rising, and before rifting begins; but the volume of such volcanism is likely to be small. According to the decompression hypothesis, a bulk of the melting will follow initial rifting when the continental crust is thinned. Seefloor spreading would therefore be expected to follow volcanism. Indeed, the separation of the Seychelles and Saya da Malha Banks from India followed the formation of the Deccan Traps at the Réunion hot spot 60–65 million years ago, and the opening of the North Atlantic followed the formation of the Brito-Arctic volcanic province at the Iceland hot spot 55–65 million years ago. The impingement of rising plumes on the lithosphere did not, however, lead to continental rifting in all cases (for example, Siberian Traps). Also, in some (for example, Karoo) there has been a long delay before rifting begins, suggesting that other controls may, in some cases, be more important.

Role of mantle plumes. Ever since W. J. Morgan introduced the concept of hot spots, there has been considerable discussion on the role of mantle plumes in the driving mechanism of plate tectonics. Early discussions were concerned with how ascending plumes interacted with convective motions in the mantle, especially the large-scale flow between mid-oceanic ridges and subduction zones. They now focus on whether hot spots play an active or passive role in the generation of plate motions. The uplift of plume-generated midplate swells may be sufficient, for example, to contribute to the forces that drive plate motions. However, plumes are passive, as is seen by their tendency to become entrained at mid-oceanic ridges (for example, Iceland) and large-offset fracture zones (for example, Louisville). *See* SUBDUCTION ZONES.

A. B. Watts

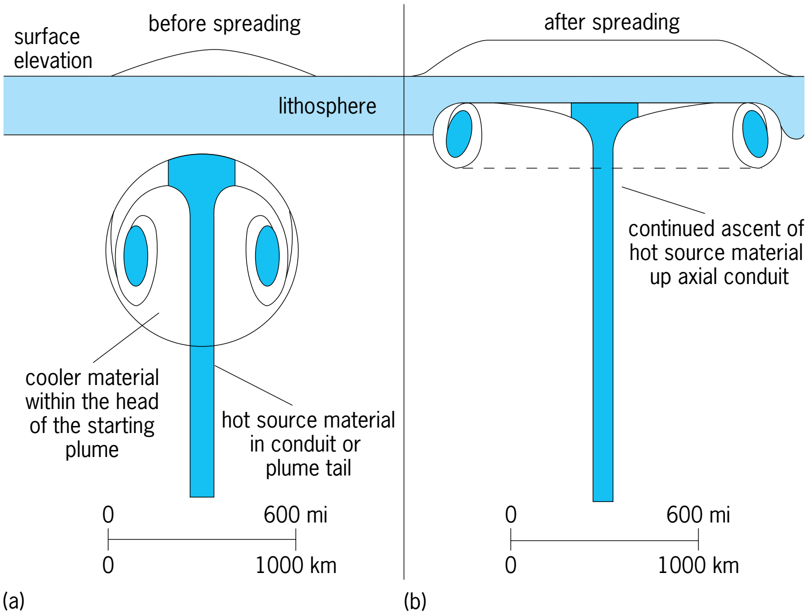

surface elevation
before spreading
after spreading
lithosphere
cooler material within the head of the starting plume
hot source material in conduit or plume tail
continued ascent of hot source material up axial conduit
0
600 mi
0
1000 km
0
600 mi
0
1000 km
(a)
(b)

Fig. 3. **Ascent of a plume (*a*) before and (*b*) after near-surface spreading. Darker areas indicate higher temperatures. The upper panels indicate the horizontal scale over which significant uplift would be expected to occur above such a plume. (*After R. I. Hill, Starting plumes and continental break-up, Earth Planet. Sci. Lett., 104: 398–416, 1993*)**

Bibliography. S. T. Crough, Hotspot swells, *Annu. Rev. Earth Planet. Sci.,* 11:165–193, 1983; R. S. Detrick and S. T. Crough, Island subsidence, hot spots, and lithospheric thinning, *J. Geophys. Res.*, 83:1236–1244, 1978; R. I. Hill, Starting plumes and continental break-up, *Earth Planet, Sci. Lett.*, 104:398–416; N. H. Sleep, Hotspot volcanism and mantle plumes, *Annu. Rev. Earth Planet. Sci.*, 20:19–43, 1992; R. S. White and D. P. McKenzie, Magmatism at rift zones: The generation of volcanic continental margins and flood basalts, *J. Geophys. Ref.*, 94:7685–7729, 1989.

Hot-water heating system

A heating system for a building in which the heat-conveying medium is hot water. Heat transfer in British thermal units (Btu) equals pounds of water circulated times drop in temperature of water. For other liquids, the equation should be modified by specific heats. The system may be modified to provide cooling.

A hot-water heating system consists essentially of water-heating or -cooling means and of heat-emitting means such as radiators, convectors, baseboard radiators, or panel coils. A piping system connects the heat source to the various heat-emitting units and includes a method of establishing circulation of

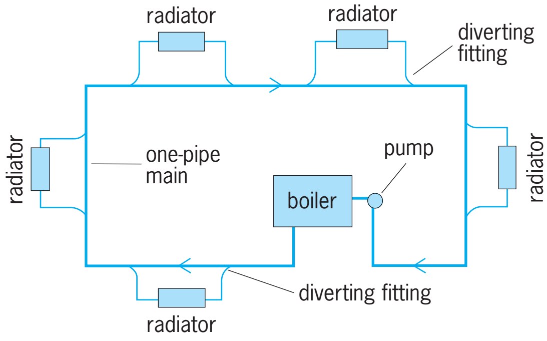


Fig. 1. One-pipe hot-water heating system.

the water or other medium and an expansion tank to hold the excess volume as it is heated and expands. Radiators and convectors have such different responses that they should not be used in the same system.

Types. In a one-pipe system (**Fig. 1**), radiation units are bypassed around a one-pipe loop. This type of system should only be used in small installations.

In a two-pipe system (**Fig. 2**), radiation units are connected to separate flow and return mains, which may run in parallel or preferably on a reverse return loop, with no limit on the size of the system.

In either type of system, circulation may be provided by gravity or pump. In gravity circulation each radiating unit establishes a feeble gravity circulation; hence such a system is slow to start, is unpredictable, and is not suitable for convectors, baseboard radiation, or panel coils because circulating head cannot be established, and circulation cannot be supplied to units below the mains. The pipes must be large. For these reasons gravity systems are no longer used.

In forced circulation a pump is used for motivation. Circulation is positive and units may be above or below the heat source. Smaller pipes are used.

Operation. For perfect operation it is imperative that the friction head from the heat source through each unit of radiation and back to the heat source be the same. This usually requires careful balancing after installation and during operation.

Expansion tanks are open or closed. Open tanks are vented to the atmosphere and are used where the water temperature does not exceed 220°F or 123°C (at sea level). They provide the safest operation, almost free from explosion hazards. Closed tanks, used for higher water temperatures, are provided with safety devices to avoid possible explosions.

One outstanding advantage of hot-water systems is the ability to vary the water temperature according to requirements imposed by outdoor weather conditions, with consequent savings in fuel. Radiation units may be above or below water heaters, and piping may run in any direction as long as air is eliminated. The system is practically indestructible. Flue-gas temperatures are low, resulting in fuel savings. The absence of myriads of special steam fittings, which are costly to purchase and to maintain, is also an important advantage. Hot water is admirably adapted to extensive central heating where high temperatures and high pressures are used and also to low-temperature panel-heating and -cooling systems.

Circulating hot-water pumps must be carefully specified and selected. On medium-size installations, it is recommended that two identical pumps be used, each capable of handling the entire load. The pumps operate alternately but never together in parallel. On large installations three or more pumps may be used in parallel, provided they are identical and produced by the same manufacturer. The casings and runners must be cast from the same molds, and the metals and other features that affect their temperature characteristics must be identical. All machined finishes must be identical, and the pumps must be thoroughly shop-tested to operate with identical characteristics.

When the system is in operation, the pump can be disconnected from the boiler by throttling down the valve at the boiler return inlet; it should not be closed completely. This permits the water in all boilers to be at the same temperature so that when a boiler is thrown back into service, the flue gases do not impinge on any cold surfaces, thus producing soot and smoke to further contaminate the outdoor atmosphere. *See* COMFORT HEATING; DISTRICT HEATING; OIL BURNER. Erwin L. Weber

Bibliography. American Society of Heating, Refrigerating, and Air Conditioning Engineers, *Handbook*: *Applications*, 1993, *Equipment*, 1983, *Systems*, 1992, *Fundamentals*, 1994; J. E. Traister, *Residential Hot Water Systems: Repair and Maintenance,* 1986.

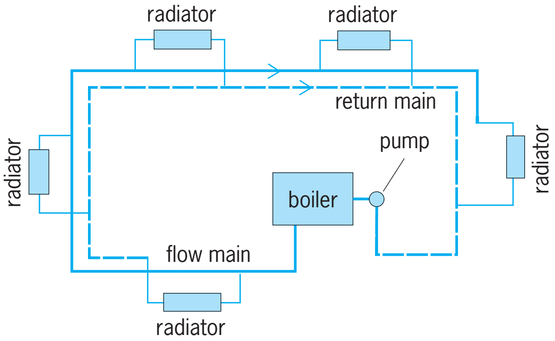


Fig. 2. Two-pipe reverse return system.

Hubble constant

A number that characterizes the expansion rate of the universe and is required to determine its age. In the standard big bang model, the local universe expands uniformly according to the Hubble law, $v = H_0 d$, where v is the velocity of a galaxy at a distance d, and H_0 is the Hubble constant. The wavelength of radiation is stretched due to the expansion of space so that the spectra of objects become progressively redder at greater distances. (For nearby objects, the observed redshift can be described as a Doppler effect.) The constant is named after Edwin P. Hubble, who discovered that the velocity of recession of a galaxy is proportional to its distance. A reliable and accurate measurement of the Hubble constant, an independent estimate of the ages of the oldest objects in the universe, and a further measurement of the average mass-energy density in the universe are all separately required in order to test and ultimately provide strong constraints on cosmological models. Measuring an accurate value of H_0 was one of the motivating reasons for building the *Hubble Space Telescope*. *See* DOPPLER EFFECT; HUBBLE SPACE TELESCOPE; REDSHIFT.

Although measurement of the Hubble constant is extremely simple in principle, it is much more difficult in practice. For decades, it remained an outstanding problem in cosmology following Hubble's original discovery in the 1920s. There were two main reasons for the difficulty: First, measuring distances turned out to be immensely challenging. Second, while the velocities can be measured very simply and accurately (from measurements of the positions of spectral lines in galaxies), galaxies are known to interact gravitationally with their neighbors. In so doing, their velocities are perturbed (inducing so-called peculiar motions superimposed on the general expansion). Hence, while an accurate measurement of the Hubble constant requires that an accurate extragalactic distance scale be established, this already-difficult task must be done at distances great enough that peculiar motions of galaxies are small compared with the overall cosmic expansion velocity, the Hubble flow. Measurements from two major efforts, the Hubble Space Telescope Key Project and the *Wilkinson Microwave Anisotropy Probe* (*WMAP*), have led to a convergence on the value of this important parameter. *See* ASTRONOMICAL SPECTROSCOPY.

Distances to galaxies and measurement. Astronomy is unique among the physical sciences in that most length scales cannot be measured directly. The size scales, especially in a cosmological context, are too vast. In general, the basis for estimating distances in astronomy is the inverse-square radiation law. If objects can be identified whose luminosities are either constant (standard candles), or perhaps related to a quantity that is independent of distance (for example, period of oscillation, rotation rate, velocity dispersion, or color), then, given an absolute calibration, their distances can be gauged. The standard candles must be independently calibrated (to absolute physical units) so that true distances (in meters or megaparsecs; 1 Mpc $= 3.08 \times 10^{22}$ m $= 3.26 \times 10^6$ light-years) can be determined.

Cepheid variables. Primary among the distance indicators are the Cepheid variables, stars whose outer atmospheres pulsate regularly with periods ranging from 2 to about 100 days. Empirically it has been established that the period of pulsation (a quantity independent of distance) is very well correlated with the intrinsic luminosity of the star. High resolution is the key to discovering Cepheids in other galaxies. The Cepheids must be identified against the background of fainter, resolved and unresolved stars that contribute to the light of the galaxy. From the Earth, turbulence in the atmosphere degrades the image resolution, smearing the light from the stars of interest and decreasing their contrast against the background. The resolution of the *Hubble Space Telescope* is about 10 times better than can be generally obtained through the Earth's atmosphere, and, moreover, it is stable. As a result, the volume of space made accessible by the *Hubble* increased by a factor of 1000. *See* CEPHEIDS.

Other distance indicators. The reach of Cepheid variables as distance indicators is limited, however, even with the *Hubble Space Telescope*. For distances beyond 20 Mpc or so, brighter objects than ordinary stars are required, for example, luminous supernovae or the luminosities of entire galaxies. The principle for measuring all of these distances is the same. Whether it is the rotation speed or velocity dispersion of a galaxy, or the rate at which a supernova fades, all of these distance-independent quantities are indicators of intrinsic luminosity. Given the apparent luminosity (corrected for intervening extinction by dust), a simple application of the inverse-square law then yields a distance. The absolute calibration for all of these methods is presently established using the Cepheid distance scale.

One of the most promising cosmological distance indicators is the luminous supernovae classified as type Ia. These objects have luminosities comparable to entire galaxies of moderate luminosity, and hence can be observed to distances of hundreds of megaparsecs. Unfortunately, the exact mechanism for the ignition of the explosion has not yet been theoretically or observationally established, nor are the progenitors known with any certainty. The basis for most of the relative distance indicators remains empirical. Ultimately, the confidence in these empirically based methods will be strengthened as the theoretical basis is more firmly established. *See* SUPERNOVA.

For spiral galaxies, the total luminosity shows an excellent, empirical correlation with the maximum rotation velocity of the galaxy (the Tully-Fisher relation). Independent of distance, galaxy rotation rates can be measured spectroscopically (from Doppler shifts of spectral features of hydrogen at radio or optical wavelengths). This relation has been measured for hundreds of galaxies within clusters, and in the general field.

For elliptical galaxies, a correlation exists between the stellar velocity dispersion and the intrinsic luminosity, analogous to the relation between rotation velocity and luminosity for spirals. Moreover, the surface brightness of an elliptical galaxy inside a given radius is tightly correlated with the velocity dispersion of the galaxy. Another method useful for measuring distances to elliptical galaxies makes use of the fact that the resolution of stars within galaxies is distance-dependent. *See* GALAXY, EXTERNAL.

Key project. The examples described above provide a means of measuring relative distances to galaxies, while the absolute calibration for these methods relies on the Cepheid distance scale. A key project of the *Hubble Space Telescope* has provided Cepheid distances to a sample of 18 galaxies useful for setting the absolute distance scale for galaxies, in addition to other methods. All of these methods can be applied at distances where the peculiar motions of galaxies contribute less than 10% of the overall cosmic expansion velocity.

A controversy existed for many decades over the value of the Hubble constant, with published values disagreeing at times by a factor of 2. However, the Cepheid distances from the Hubble Space Telescope Key Project have provided a means of calibrating and comparing a number of relative distance methods; and for the first time, to within an uncertainty of ±10%, all of these methods are consistent with a value of the Hubble constant of 72 kilometers per second per megaparsec.

Microwave anisotropy measurements. The detection in 1965 by Arno Penzias and Robert Wilson of the cosmic background radiation provides one of the strongest foundations for big bang cosmology. In 2002, the *WMAP* satellite measured small fluctuations in the microwave sky to a few parts in 100,000. Encoded within these fluctuations is information about the geometry, the Hubble constant, as well as the matter-plus-energy density of the universe. By themselves, these data provide only a weak limit on the Hubble constant. But combined with measurements of the large-scale distribution of galaxies, the *WMAP* results yield a value of the Hubble constant of 71 kilometers per second per megaparsec, with an uncertainty of only 5%. This value is in excellent agreement with that from the Hubble Key Project, and is a completely independent determination based on very different underlying physics. *See* COSMIC BACKGROUND RADIATION.

Age of the universe. The Hubble constant sets the expansion time scale for the universe. In order to measure the time since the big bang, it is necessary to determine the expansion rate and the matter-plus-energy density of the universe. Increasing evidence suggests that the total matter density of the universe is only about 30% of the total mass-energy density of the universe. The remaining 70% appears to be in a mysterious form with a repulsive pressure, which causes an acceleration of the universe (and results in an older universe). Measurements of type Ia supernovae provide evidence for an acceleration of the universe, consistent with a dark-energy component. This phenomenon is also consistent with the *WMAP* observations. *See* DARK ENERGY.

With a Hubble constant of 72 kilometers per second per megaparsec, and a universe composed of one-third matter density and two-thirds dark energy, the expansion age is calculated to be 13.5 billion years. The expansion age of the universe can be compared to the ages of the oldest stars within the Milky Way Galaxy, those in globular clusters. Results from the *Hipparcos* satellite, applied to models of stellar evolution, yield ages of about 12 billion years with an uncertainty of about ±20%. Given the uncertainties in the two independent measurements of the time scale (one based on the theory of the evolution of stars and the other based on the dynamics of the big bang), there is a remarkable agreement. *See* ASTROMETRY; BIG BANG THEORY; COSMOLOGY; STAR CLUSTERS; STELLAR EVOLUTION; UNIVERSE.

Wendy L. Freedman

Bibliography. J. D. Barrow, *The Origin of the Universe*, Basic Books, 1994; K. Ferguson, *Measuring the Universe*, Walker, New York, 1999; W. Freedman, The expansion rate and size of the universe, *Sci. Amer. Quart.*, 1:92–97, Spring 1998; W. Freedman and M. Turner, Cosmology in the new millennium, *Sky Telesc.*, 106(4):31–41, October 2003.

Hubble Space Telescope

The *Hubble Space Telescope* is the largest visible-light observatory ever placed into space. *Hubble*'s orbit, some 612 km (380 mi) above Earth's surface (**Fig. 1**), keeps it above almost all of Earth's atmosphere, at a location where its view of the heavens is much clearer than that of ground-based telescopes. The superior view afforded by *Hubble*'s orbit has made the telescope a unique resource for astronomers worldwide and has led to fundamental discoveries about the size and age of the universe, the birth and death of stars, and the development of galaxies.

The spacecraft. The *Hubble Space Telescope* is 13.2 m (43.5 ft) long, is 4.2 m (14 ft) in diameter at its widest, weighs 11,110 kg (24,500 lb), and orbits Earth once every 96 minutes. At its core is a reflecting telescope with a primary mirror 2.4 m (94.5 in.) in diameter (**Fig. 2**). The primary mirror directs light from astronomical objects to a 30-cm (12-in.) secondary mirror, which then bounces it back through a hole at the center of the primary mirror to the scientific instruments. The optical design is a Ritchey-Chrétien variant of a Cassegrain telescope.

Hubble's optical system enables the telescope to record astronomical images with unprecedented precision in the optical, ultraviolet, and infrared spectral bands. In order to take full advantage of the clearer view above Earth's atmosphere, *Hubble*'s mirrors had to be polished until they were extremely smooth: The largest bumps on *Hubble*'s primary mirror are analogous to the height of a baseball on a surface as wide as the continental United States. A flaw in the overall shape of the primary mirror hampered observations for several years after launch. However, because the primary mirror was so smooth, corrective optics installed in 1993 were able to realize *Hubble*'s expected performance (**Fig. 3**). Its angular resolution of 0.05 arcsecond at optical wavelengths is equivalent to being able to distinguish two fireflies 1 m (3 ft) apart at a distance of 5000 km (3000 mi). *See* INFRARED ASTRONOMY; TELESCOPE; ULTRAVIOLET ASTRONOMY.

Recording sharp images during exposure times that can approach 1 hour requires very precise and stable pointing of the telescope. Guide stars are used to keep *Hubble* locked on target. Each observation typically relies on two guide stars that *Hubble*'s guiding system keeps fixed at preselected locations in

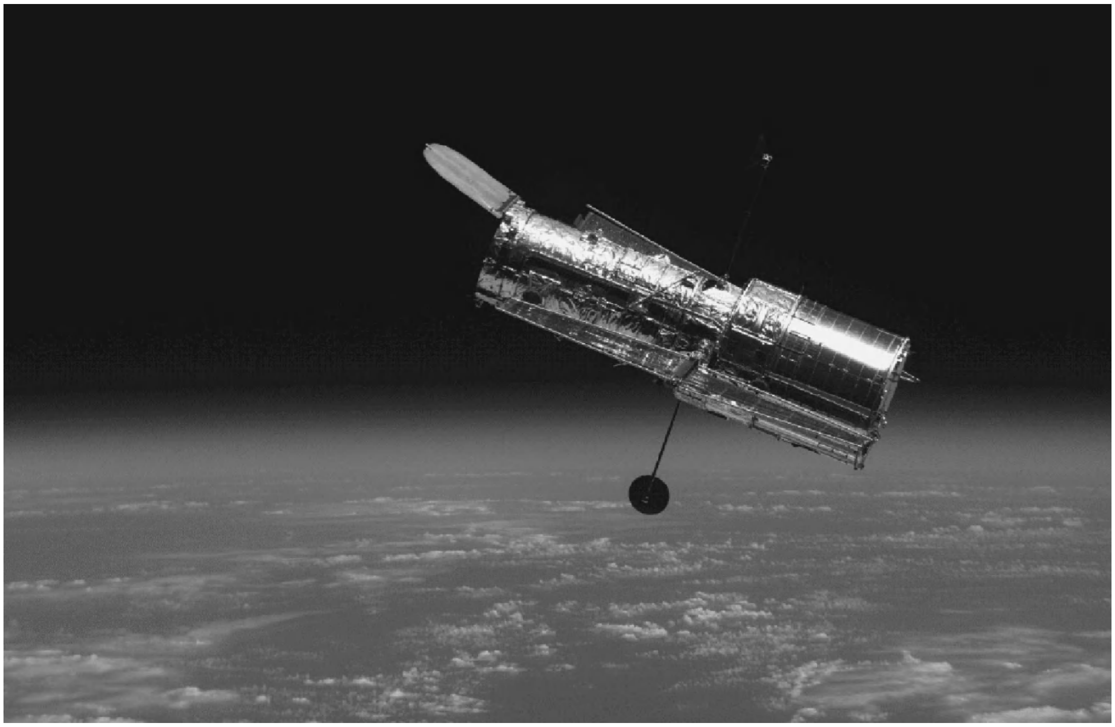


Fig. 1. ***Hubble Space Telescope* in orbit around Earth.**

two of its three Fine Guidance Sensors. Whenever *Hubble* drifts slightly off target, the guiding system returns these stars to their preselected locations in the telescope's field of view, placing the telescope back on target. This system keeps *Hubble* pointed to within 0.007 arcsecond, equivalent to keeping a laser in New York locked on a small coin in Washington, DC.

Hubble's complement of scientific instruments handles a wide range of observational tasks. Its cameras have recorded images of astronomical objects at wavelengths ranging from 115 to 2500 nanometers. *Hubble*'s spectrographs have analyzed the spectra of these objects between wavelengths of 115 and 1030 nm. Because these instruments can be removed and replaced, they have been upgraded several times during *Hubble*'s stay in orbit (see **table**). A diverse assortment of support equipment surrounds the telescope and its scientific instruments. To supply power, the spacecraft has two solar arrays and a set of storage batteries that keep the spacecraft operating when it passes through Earth's shadow. To rotate the telescope toward different spots on the sky, *Hubble* has four reaction wheels; increasing their spin rate in one direction causes the spacecraft to rotate in the opposite direction. To store and communicate the data gathered, *Hubble* has a set of solid-state data recorders and communications antennas that link to NASA's Tracking and Data Relay Satellite System (TDRSS). Governing the whole spacecraft is an onboard computer that interprets and executes the instructions relayed from the ground.

History. Two disadvantages of ground-based observing were well known to early-twentieth- century astronomers: (1) Because of the turbulent motions of Earth's atmospheric gases, the paths of light rays passing through the atmosphere are constantly shifting, distorting our view of astronomical objects. To human eyes these distortions are rather subtle; for example, they are responsible for the twinkling of stars. However, this effect limits the sharpness of most images taken with ground-based telescopes to a resolution of no better than 1 arcsecond (Fig. 3). (2) The Earth's atmosphere is quite transparent to visible light but blocks much of the infrared and ultraviolet light from the cosmos. Both of these wavelengths are scientifically important. *See* TWINKLING STARS.

In the 1920s, the decade in which the American scientist Robert Goddard launched the first liquid-fueled rockets, the German scientist Hermann Oberth published the first serious papers describing the advantages of a space-based telescope. Development of the V2 rocket in Germany during World War II made Oberth's speculations seem much more realistic. After the war, in 1946 the American astronomer Lyman Spitzer produced a detailed report for the RAND Corporation on what a large space-based telescope might accomplish. This report is widely regarded as the birth of the Hubble Space Telescope.

Even though the rationale for a space-based telescope was clear in 1946, the technology was not yet ready. Launch vehicles capable of placing a large telescope into orbit were not developed until the 1960s, when the planning of the *Hubble Space Telescope* began in earnest. Several smaller precursors, such as the *Orbiting Astronomical Observatory* (*OAO*) and the *International Ultraviolet Explorer* (*IUE*), preceded *Hubble* into space, demonstrating both

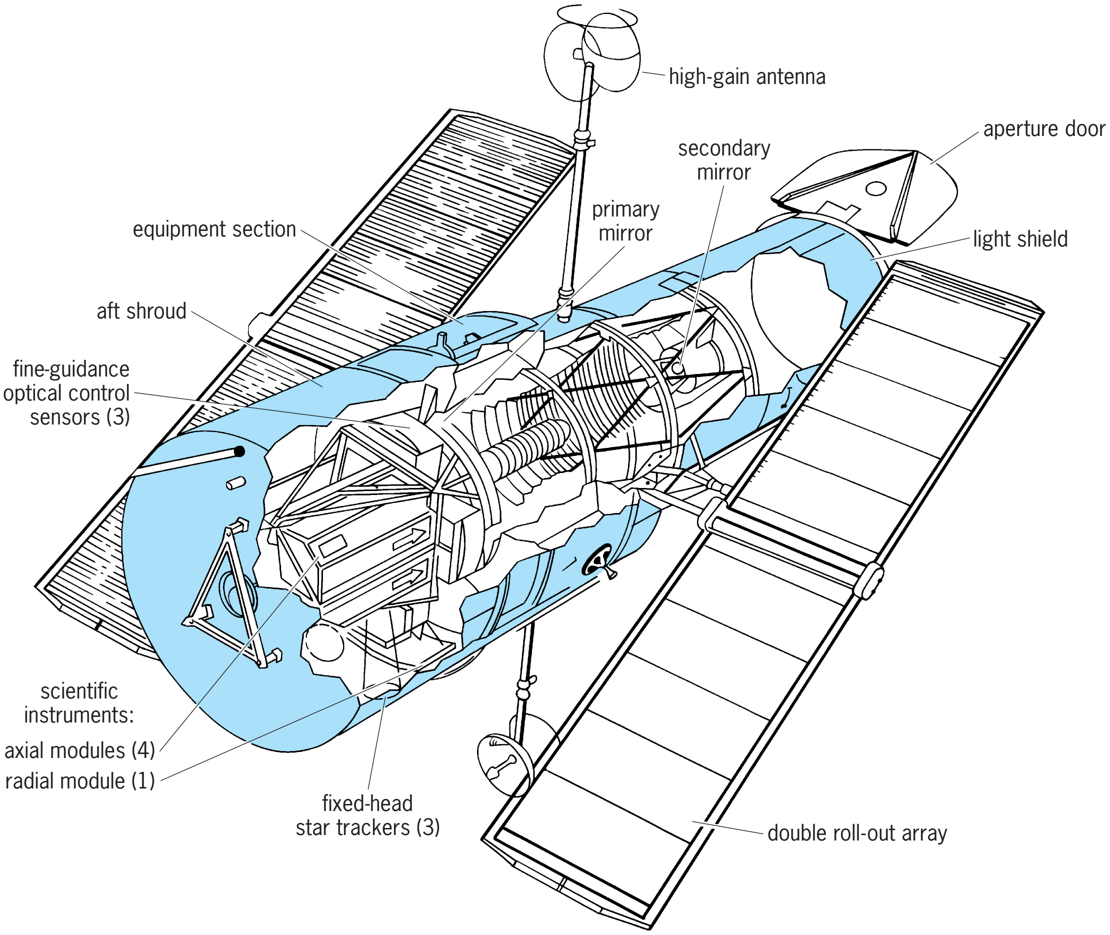


Fig. 2. Diagram of the *Hubble Space Telescope*.

the feasibility and the rewards of astronomical observing from above Earth's atmosphere. Meanwhile, NASA and the astronomy community worked to build political support for a much larger, much more expensive telescope in space. *See* SATELLITE (ASTRONOMY).

Congress approved the so-called Space Telescope in 1977. Lockheed Corporation was chosen to build the spacecraft, and Perkin-Elmer Corporation to grind and polish the primary mirror. The spacecraft, renamed the *Hubble Space Telescope* in 1983 in honor of American astronomer Edwin Powell *Hubble*, was ready for launch in 1986, but the tragedy of the *Challenger* space shuttle explosion delayed *Hubble*'s launch until 1990.

The space shuttle *Discovery* finally carried *Hubble* aloft on April 24, 1990, forty-four years after Spitzer's seminal report. The shuttle's robotic arm released the spacecraft into orbit the following day. However, the ultrasharp pictures expected from *Hubble* did not begin to arrive until $3^1/_2$ years later.

Shortly after launch, astronomers realized that *Hubble* could not be properly focused. The source of the problem was a flaw in the primary mirror which

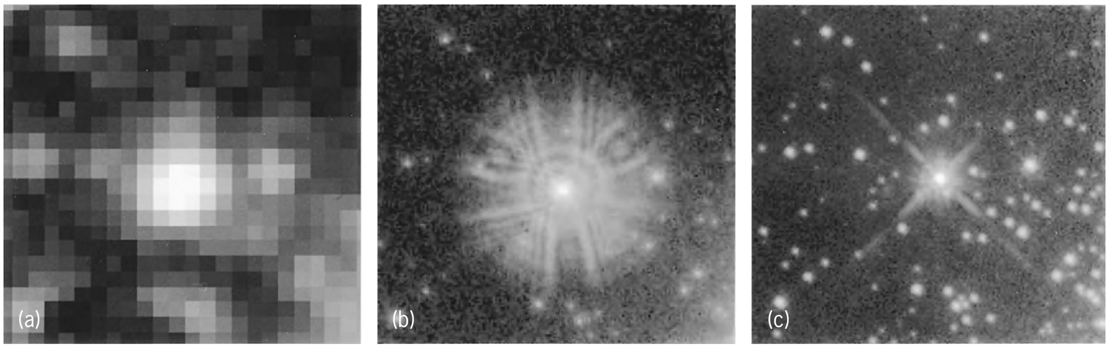


Fig. 3. Comparison of images of a field of stars in the 30 Doradus Nebula. (*a*) Ground-based image obtained under conditions of good seeing. (*b*) Same star field at the same scale taken with the *Hubble Space Telescope* before the first servicing mission. (*c*) *Hubble* image after the first servicing mission when the spherical aberration was corrected. (*NASA*)

History of *Hubble*'s instruments*

	Instruments	Dates
Original instruments	Wide Field/Planetary Camera (WF/PC), *Hubble*'s original visible/ultraviolet light camera	1990–1993
	Faint Object Camera (FOC), *Hubble*'s highest-resolution camera	1990–2002
	Faint Object Spectrograph (FOS), for analyzing light from faint objects	1990–1997
	Goddard High Resolution Spectrograph (GHRS), designed to perform detailed analyses of spectra	1990–1997
	High Speed Photometer (HSP), for measuring rapid variations in the brightness of astronomical objects	1990–1993
Instruments from first servicing mission	Wide Field Planetary Camera 2 (WFPC2), a visible/ultraviolet-light camera with corrective optics to compensate for the flaw in the primary mirror	1993–present
	Corrective Optics Space Telescope Axial Replacement (COSTAR), a device that placed corrective optics over *Hubble*'s original instruments to compensate for the flawed primary mirror	1993–present (no longer in use; current instruments designed with corrected optics)
Instruments from second servicing mission	Near Infrared Camera and Multi-Object Spectrometer (NICMOS), *Hubble*'s primary camera for observing infrared light	1997–present (inoperable 1999–2002)
	Space Telescope Imaging Spectrograph (STIS), a much more efficient optical/UV spectrograph than the FOS or GHRS; can also be used as an optical/UV camera	1997–present (inoperable since 2004)
Instrument from fourth servicing mission	Advanced Camera for Surveys (ACS), an optical-light camera much more sensitive than WFPC2 and with a wider field of view	2002–present
Planned instruments	Cosmic Origins Spectrograph (COS), an extremely sensitive UV spectrograph	Awaiting installation
	Wide Field Camera 3 (WFC3), an advanced optical-infrared camera	Awaiting installation

*As of 2006.

caused an effect known as spherical aberration. *Hubble*'s mirror was exquisitely polished, but its overall shape was incorrect.

Fortunately, NASA had designed *Hubble* for periodic servicing and upgrades (see table). During the first servicing mission in December 1993, astronauts were able to install a new camera (WFPC2) with optics that corrected for the flaw in *Hubble*'s mirror, as well as a device (COSTAR) that placed corrective optics in front of *Hubble*'s original instruments. That mission also installed new solar arrays on *Hubble*, eliminating some troublesome vibrations of the spacecraft produced by the original arrays. With these improvements, *Hubble*'s performance surpassed the original specifications, enabling *Hubble* to fulfill its scientific promise (Fig. 3).

Subsequent servicing missions performed additional upgrades. The second servicing mission, in February 1997, installed two new instruments: STIS, a vastly improved spectrograph, and NICMOS, *Hubble*'s primary infrared camera. The third mission, in December 1999, replaced four of *Hubble*'s guiding gyroscopes and fixed some of *Hubble*'s multilayer insulation. The fourth mission, in March 2002, installed a state-of-the-art visible-light camera, the Advanced Camera for Surveys (ACS), a new set of solar arrays, and a cooling system that revived the NICMOS camera, which had been inoperable since 1999.

A fifth servicing mission, originally scheduled for 2005, has been delayed by the loss of the shuttle *Columbia* and ongoing problems with the Space Shuttle program. The mission will install a more sensitive spectrograph (COS) and an upgraded optical/infrared camera (WFC3), along with new batteries and gyroscopes that are essential for extending *Hubble*'s useful lifetime. Each month that passes without an upgrade now places the *Hubble* spacecraft in greater danger of an equipment failure that would prematurely end its mission.

Highlights of Hubble science. The *Hubble Space Telescope*'s contributions to astronomy are numerous and wide-ranging. The following are a few of the most notable.

Measuring the size and age of the universe. The primary method for measuring distances to other galaxies is to measure the brightness of stars whose intrinsic light outputs are known by other means. A class of stars known as Cepheid variables is particularly useful because the period over which they vary in brightness is directly related to their total light output. By comparing the apparent brightness of these stars to the total light output inferred from their brightness variations, astronomers can calculate their distance. *Hubble*'s high resolution has enabled the study of individual Cepheid variable stars in distant galaxies and the measurement of distances to galaxies up to 100 million light-years away. From the distances to these galaxies and the speeds at which they are moving away from each other, astronomers can calculate how long it took for galaxies to reach their current positions. Current estimates put this amount of time—the age of the universe—about 13.5 billion years. *See* CEPHEIDS; COSMOLOGY; HUBBLE CONSTANT.

Observations of young galaxies. *Hubble* has supplied the first clear pictures of what the universe was like when it was only 1 or 2 billion years old. Images

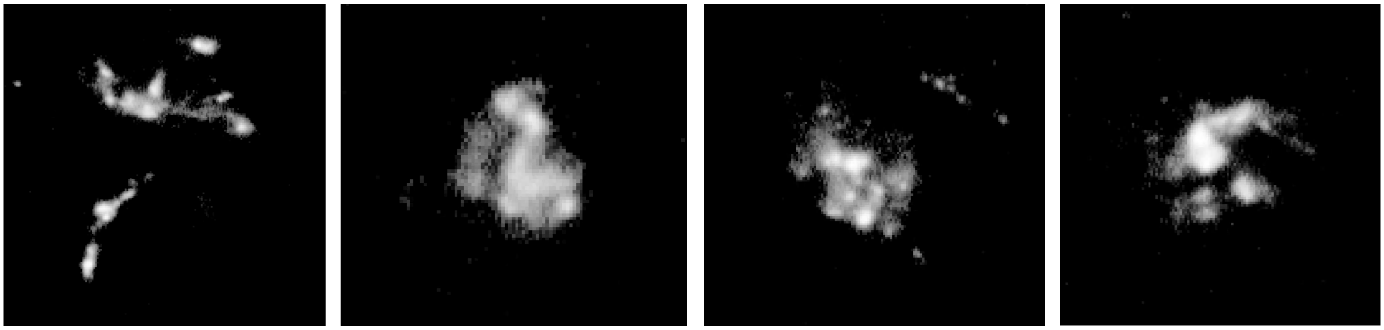


Fig. 4. Distant galaxies from the Hubble Ultra Deep Field. The distorted appearances of these galaxies suggest that they collided with other galaxies shortly before their light began the long journey to Earth.

such as the Hubble Deep Field and Hubble Ultra Deep Field reveal a multitude of galaxies at a wide range of distances, some of them over 10 billion light-years away. Light collected from these most distant galaxies took over 10 billion years to reach Earth. Thus, the images of these galaxies show how they looked over 10 billion years ago. Many of them appear quite strange and distorted in comparison to present-day galaxies, leading astronomers to believe that disruptive collisions between galaxies were much more common early in time than they are now (**Fig. 4**). *See* GALAXY, EXTERNAL.

Disks around young stars. The motions of planets in the solar system, which generally follow roughly circular paths sharing the same orbital plane, have long led astronomers to speculate that the solar system formed from a disk-shaped collection of matter with the Sun at its center. *Hubble* has helped to verify this suspicion through its observations of young stars in the process of formation. All stars, including the Sun, were born in clouds of interstellar hydrogen gas. Observations of clouds in which stars are currently forming show that brand-new stars are generally surrounded by disks of gas and tiny, solid particles called dust grains. The gas and dust in at least some of these disks are eventually expected to clump into planets similar to those that orbit the Sun. *See* SOLAR SYSTEM; STAR; STELLAR EVOLUTION.

Mark Voit

Bibliography. E. J. Chaisson, *The Hubble Wars*, 1994; C. C. Petersen and J. C. Brandt, *Hubble Vision: Further Adventures with the Hubble Space Telescope*, 1998; R. W. Smith, *The Space Telescope*, 1989; R. W. Smith and D. DeVorkin, *Hubble Space Telescope: Imaging the Universe*, 2004; M. Voit, *Hubble Space Telescope: New Views of the Universe*, 2000.

Hudson Bay

A horseshoe-shaped bay, approximately 1050 km (650 mi) wide and 1370 km (850 mi) long, stretching from the Canadian provinces of Quebec and Ontario in the south to the Northwest Territories. With a surface area of approximately 827,000 km^2 (320,000 mi^2), Hudson Bay ranks twelfth in surface area among the Earth's seas and oceans. The bay is connected to the Arctic Ocean by the Foxe Channel and Roes Welcome Sound and to the Atlantic Ocean by the Hudson Strait (see **illus.**). Several islands are located within the bay, the largest of which is Southampton (41,212 km^2; 15,912 mi^2). *See* ATLANTIC OCEAN; NORTH AMERICA.

Early exploration of Hudson Bay was prompted by interest in finding an Arctic shortcut through North America to Asia. The bay was discovered by English navigator Henry Hudson. In 1610, Hudson directed his ship, the *Hopewell*, on a 3-month investigation of the bay's eastern shore and islands. Another early explorer, Thomas Button, reached the site of present-day Churchill on the bay's western shore in 1612. The following summer, Button passed Southampton Island on his way to the Hudson Strait. Others who explored the bay include William Baffin in 1615, and Luke Fox and Thomas James in 1631.

Geologic history. Hudson Bay is located within a depression of the Canadian Shield and is underlain with Precambrian rocks more than 500 million years old. The shield covers 4.8 million square kilometers (1.8 million square miles) in the United States and Canada and encompasses portions of Wisconsin, Minnesota, New York, Michigan, the Northwestern Territories, Saskatchewan, Ontario, Quebec, Baffin Island, and Labrador. The shield is the oldest region within the North American crustal plate, and is considered a craton—an area of the Earth's crust that has been stable for millions of years. Bordering the bay are gently dipping Paleozoic limestones, sandstones, and dolomite rocks. *See* CRATON; PALEOZOIC; PRECAMBRIAN.

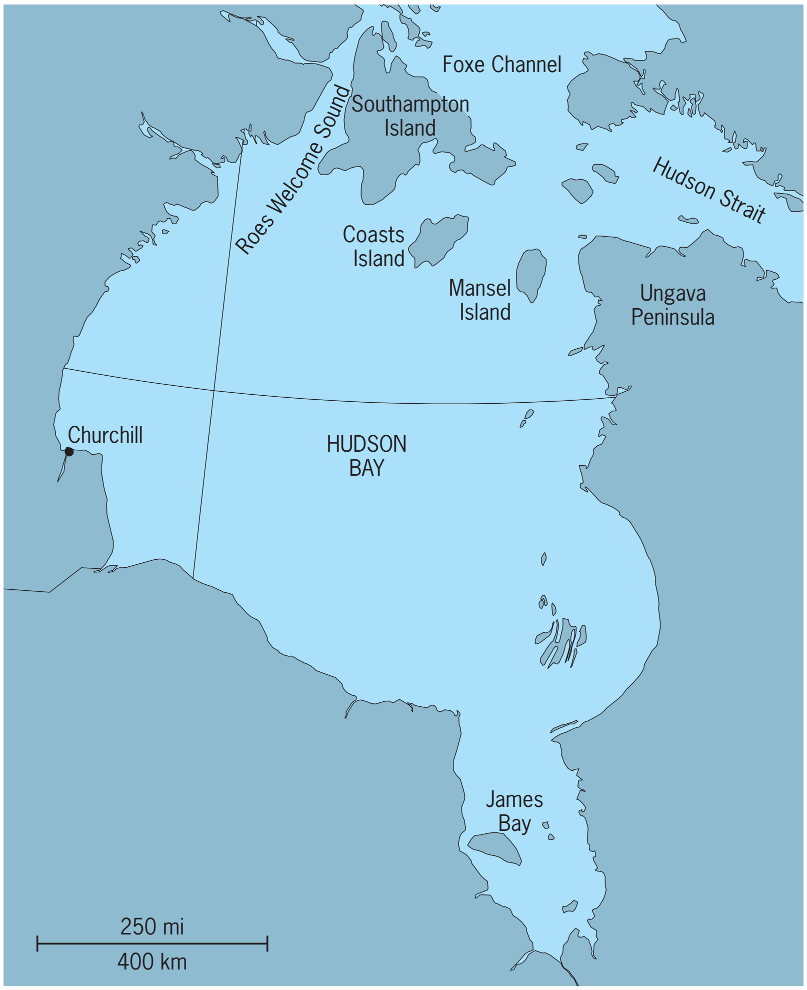
The bay was formed 2.4 million years ago during the Pleistocene geologic epoch. Centered on what is now Hudson Bay, the Laurentide ice sheet stripped away soil and deposited rocks and sediments as it moved across the landscape. The enormous weight of the accumulated ice caused areas of the Earth's crust to compress and sink. Melting and ice retreat began about 13,000 years ago during the Wisconsin glacial period. As the ice melted, relatively warm seawater invaded the Hudson Bay through the Hudson Strait, causing additional shrinking of the ice mass. Eventually, the ice sheet separated into the Keewatin and Labrador ice centers, which disappeared completely between 6500 and 5000 years ago. As the ice sheet melted, the crust began to slowly rebound due to isostatic adjustment. This process is not yet complete, as the area surrounding the bay continues to rise about 0.6 m (2 ft) each century. Uplifting of the land surface is most obvious along portions of the coast where lines representing former beaches run parallel to the shore. A rapid rate of uplift suggests that the Hudson Bay will become shallower over time and may disappear completely when isostatic equilibrium is reached. *See* GLACIAL EPOCH; ISOSTACY; PLEISTOCENE.

Physical geography. Hudson Bay's underwater physiography is represented by broad contours that are mostly concentric with its periphery. The floor of the bay is predominantly smooth but has been incised in some places. Its average depth is 128 m (420 ft) with a deepest known point of 183 m (600 ft). The basin containing the bay has a north-south running elongation, a reflection of its bedrock structure. Marine currents are the most important agent of sediment transport in the bay. Snow and ice, covering the bay for several months of the year, inhibits sediment discharge from streams. Sediment, coarse gravel, and even large boulders are moved away from shore in a process known as ice rafting. *See* BASIN.

Shoreline areas within the northern reaches of the bay are low in elevation, rocky, and indented with numerous small islands and inlets. Some areas surrounding the bay are characterized by broad, flat plains. To the north of Southampton Island and Quebec's Ungava Peninsula are areas with elevations exceeding 305 m (1000 ft). Hudson Bay's drainage basin is nearly 4 million square kilometers (1.5 million square miles). On the west and southwest shorelines are extensive areas of drowned swampland. Deltas, estuaries, and tidal flats are common through the area. The tundra surrounding much of the bay is a cold desert (moisture is scarce). Tundra plant cover includes lichens, grasses, and scattered shrubs. Plants must complete their annual cycles during brief summers in soils that are poorly drained as a result of the underlying permafrost. Soils on land areas surrounding the bay include infertile entisols, inceptisols, histosols, and spodosols. *See* SOIL.

Water properties and biology. The waters found within Hudson Bay are uniform in temperature, averaging near freezing with temperatures slightly cooler near the center of the bay in deeper water. The general pattern of water circulation is counterclockwise. River outflow inhibits Atlantic waters from entering the Hudson Strait, lowering the salinity of the bay water, especially in the spring and summer months. Tides range from just less than 1 m (3 ft) to greater than 4 m (13 ft) in western portions of the bay. Pack ice covers the open water from October to June.

Hudson Bay is connected to the Atlantic Ocean by the Hudson Strait. Compared to the bay, the strait is predominantly deep, in some places extending to 500 m (1640 ft). The strait forms a zone where water from the West Greenland Current mixes with water from the bay, creating an environment that supports plankton and diverse fish species.



Map of the Hudson Bay.

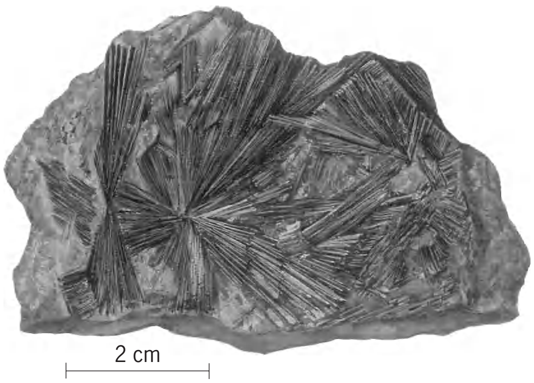
Climate. The climate of Hudson Bay is controlled by cold, dry, and stable continental polar and Arctic air masses. During much of the year, the bay is covered by snow and ice. On the land area surrounding the bay in January, the temperatures range from −30 to −20°C (−22 to −4°F), while in July the temperatures range from 10 to 20°C (50 to 68°F). The coldest land areas adjacent to the bay are located on its northwest side, extending from Manitoba's border with the Northwest Territories to Southampton Island. Rainfall is highest during summer months.

Thomas A. Wikle

Bibliography. A. W. Bally and A. R. Palmer (eds.), *The Geology of North America: An Overview*, Geological Society of America, Boulder, 1989; J. F. Lewry and M. R. Stauffer, *The Early Proterozoic Trans-Hudson Orogen of North America, Newfoundland*, Geological Society of Canada, 1990; B. MacLean (ed.), *Marine Geology of Hudson Strait and Ungava Bay, Eastern Arctic Canada: Late Quaternary Sediments, Depositional Environments, and Late Glacial–Deglacial History Derived from Marine and Terrestrial Studies*, Geol. Surv. Canada Bull. 566, 2001.

Huebnerite

A mineral with the chemical composition $MnWO_4$. Huebnerite is the manganese member of the wolframite solid-solution series. It commonly contains small amounts of iron. It occurs in monoclinic, short, prismatic crystals. Fracture is uneven. Luster varies from adamantine to resinous (see **illus.**). Hardness is 4 on Mohs scale and specific gravity is 7.2. Huebnerite is transparent and yellowish to reddish-brown in color; streak is brown. It is fusible with difficulty. *See* WOLFRAMITE. Edward C. T. Chao



Radiating groups of huebnerite crystals in quartz vein, Silverton, Colorado. (*Specimen from Department of Geology*, *Bryn Mawr College*)

Human biological variation

Anthropologists study human biology to better understand the extent of human biological variability, to explain the mechanisms that create and pattern this variability, to relate variability to health and disease, and to understand the sociocultural factors that interact with, and are influenced by, our biology. The major areas of study of human biological variation are growth and development, genetic variation, variation related to climate, infectious and noninfectious diseases, and demography. *See* ANTHROPOLOGY.

Biological versus biocultural variation. For a species with worldwide distribution, humans are remarkably similar in biology, behavior, emotions, and cognition. All humans are bipedal, have a relatively narrow range of body size and physical features (compared with other primate species), require the same essential nutrients, can eat a wide range of foods, depend on complex technology, organize their social lives via formal kinship relations and social networks, use symbolic language, and have religion. To account for the fundamental unity of the human species, anthropologists adopt an evolutionary perspective to explain variation through time and across space and a biocultural perspective that considers the interactive effects between human culture and biology.

Current evidence from the fossil record and genetic linkages among living people indicate that modern humans, *Homo sapiens sapiens*, had their origin in Africa about 125,000 years ago. Some of those African people migrated to the Middle East, Asia, Europe, Australia, and eventually the Americas. Modern humans carried with them a sophisticated array of technology, social organization, and ideological beliefs and practices. In other words, they had culture, and they used their cultural skills, as well as their biology and technology, to adapt to and exploit the environments of new lands as they migrated. The original human pattern of biology and culture was naturally selected because it conferred greater reproductive success on our ancestors. Biocultural variations that developed in the past 125,000 years allowed for continued adjustments to local climate, food, and lifestyle and continued reproductive success. *See* FOSSIL HUMANS.

History of human variation study. Until the mid-twentieth century, human variation and its biocultural nature were scarcely appreciated. From the time of Plato and Aristotle, living people and their cultures were considered to be imperfect copies, more or less, of an ideal type of human being and culture. In 1758, Carl von Linné (Linnaeus) formalized a system of human types in the 10th edition of his *Systema Naturae*. This book provided European science with the system of nomenclature for living things still used today. Linnaeus divided the human species into four types based on geography, skin color, temperament, clothing style, and other traits of material culture. In essence these four groups were Americans, Asians, Africans, and Europeans. A sense of superiority of Europeans to other "types" was inherent in this grouping and was elaborated into the "race science" of the nineteenth and early twentieth centuries. The typological approach of Linnaeus and the "race science" that followed ignored the true nature of human variation. "Races," it was believed, were fixed types in terms of size, shape, and other physical and behavioral traits. Moreover, these types were believed to be immutable, that is, not responsive to changes in the environment. *See* ANIMAL SYSTEMATICS; SPECIES CONCEPT.

A countercurrent of thought was evident in the late nineteenth century as migration of people from Europe to the United States was shown to alter the physical features of the migrants. At the forefront of the antiracial movement was Franz Boas, the founder of American anthropology. In a series of studies based on large samples of migrants to New York City, Boas demonstrated that the migrants changed their body size and shape, becoming taller and longer-legged, with narrower skulls, in response to the new environment. He referred to this process of change as plasticity in human biology. Boas noted that the earlier in life migration took place, the greater the plasticity of response. Those born in the United States showed the greatest change from their parents' "type" and often looked more like long-term residents than their own parents. Boas correctly concluded that such plasticity was due to the adjustment of growth and development of the body to the new environment, especially in relation to improvements in nutrition and health.

Full appreciation of Boas' work waited until after the genocide committed during World War II, in the name of "racial purification," and new discoveries in genetics after 1950. Anthropologists then began to reject the typological approach and "race" in favor of a population approach to the study of human variation and adaptation. By the 1960s, a series of research programs, coordinated by the International Biological Program, were implemented to study human adaptability and variability, and similar research continues today.

Growth and development. An interest in human growth variation is natural for anthropologists because the way a human being grows is the product of an interaction among the biology of our species, the physical environment in which we live, and the social, economic, and political environment that every human culture creates. The basic pattern of human growth and development is shared by all people and is the outcome of the evolutionary history of our fossil ancestors and modern people. That genetically encoded pattern includes approximately three years of infancy, four years of childhood, several juvenile years, and eight adolescent years prior to reaching reproductive and social maturation. Given the relatively long period of growth, the variety of human environments, and the plasticity of human biology, it is not surprising that people achieve significant variation in size, shape, and body composition (amount of muscle, fat, and bone). *See* ANIMAL GROWTH.

The shortest population is the Efe pygmies of central Africa; men average 145 cm (57 in.) and women 136 cm (54 in.). The tallest population are the Dutch of the Netherlands; men average 184 cm (72 in.) and women 172 cm (68 in.). There is much variation between populations, as some Efe are taller than some Dutch. Within populations there is also much variation—at least 60 cm (24 in.) of difference between the shortest and tallest "normal" people. Genetic factors are involved in some of this variation; however, inadequate nutrition and high rates of disease are the two most important environmental factors that lead to poor growth. Recent immigrants from Latin America and Asia (like Boas' migrants) become both taller and longer-legged due to improved nutrition and health in the United States. Over the past 200 years there have been general improvements in health in the wealthier nations due to better public hygiene, safer drinking water, and better diet. Along with these changes there has been a general increase in average stature, 1–2 cm (0.5 in.) per decade. There has also been a faster rate of maturation, as the average age of menarche (first menstruation) declined by 3–4 months per decade. Archeological studies show that the growth trend has returned us to the stature of our Paleolithic, hunting and gathering ancestors. It seems that cultural "advances" such as agriculture and urbanization were harmful to human health and growth in the past 12,000 years. Even today, urban dwellers are less muscular and have less dense bones than our Paleolithic ancestors. This is probably due to the sedentary nature of modern urban life compared with the hunting and gathering exercises of our ancestors. Because the human body changes so rapidly in response to the biocultural environment, many researchers use growth and development of populations as a "mirror" reflecting the material and moral conditions of the society. *See* PHYSICAL ANTHROPOLOGY.

Genetics, geography, and human variation. Genetics is concerned with the structure of deoxyribonucleic acid (DNA), how DNA is organized, how it operates, and how it is transferred across generations. Anthropological genetics focuses on the spatial distribution of genes within breeding populations, that is, groups of people who are likely to mate with each other. Geography, language, religion, nationality, and other sociocultural factors usually delineate human breeding populations. Anthropologists also study how changes in the DNA of breeding populations, brought about by mutation, migration, nonrandom mating, and random mortality in small populations, may be preserved or eliminated by natural selection, which may lead to new genetic variation and adaptation to local environments.

The well-known case of the sickle-cell hemoglobin allele (form of a gene) and its protection against malaria in the heterozygous individual is an example of genetic adaptation to the environment in the human species. Briefly stated, the introduction of farming in tropical Africa created excellent breeding environments for malaria-carrying mosquitoes. Malaria was then, and is still today, a major cause of death in children and teens in Africa. People carrying a single copy of the mutant sickle-cell allele paired with a normal red blood cell allele (heterozygotes) have resistance to malaria. Natural selection, then, increased the frequency of the mutant gene, even though it is lethal to those who inherit a double dose.

Richard Lewontin made the most important discovery of anthropological genetics when in 1972 he reported that more than 90% of human genetic variation is found within populations and less than 10% is found between populations. This is true for large continental populations, such as Africans compared with Europeans and Asians, and also for small isolated human groups, such as South American Indians. Many other geneticists have repeated Lewontin's finding, and together these studies totally disprove the notion of geographic "races." More recently, geneticists found that African populations express the greatest absolute range of genetic variation, while populations in the rest of the world show narrower ranges of variation. The non-African population variation appears to be a genetic subset of Africa. These observations support the African-origin hypothesis of modern human evolution. *See* EARLY MODERN HUMANS; HUMAN GENETICS; MOLECULAR ANTHROPOLOGY.

Human adaptation to climate. Human beings evolved in Africa, which has a range of climates from tropical to moderately cold, from sea level to high altitude, and a high to moderate exposure to ultraviolet (UV) radiation.

Temperature. Outside Africa, people encountered extreme cold, as well as regions with seasonally low ultraviolet radiation. These new climates were stressful and required biocultural adaptations for survival. Populations with long histories of exposure to extreme cold or heat may have evolved some biological adaptations. Humans conform to some extent to the ecological rules of relatively larger body mass, rounder heads, and short extremities in cold environments and relatively more linear bodies with longer arms and legs in hot environments (Bergmann's and Allen's Rules). These variations in body size and shape help to conserve heat in cold climates and dissipate body heat in hot climates. There are, however, many exceptions to these rules. Behavior and culture mitigate many of the stresses of temperature. Nutritional status, workload, and health can significantly modify human body size, shape, and behavior. Cold-adapted peoples (Eskimo, Inuit, Siberians) can modify the flow of blood to the extremities, especially hands and feet, so as to reduce heat loss without suffering frostbite. Growth and development in the cold as well as some genetic mechanisms are implicated in this adaptation. *See* BIOMETEOROLOGY.

UV radiation. All normal humans have the same skin pigment, melanin. Natural selection acted to produce and maintain the continuous variation (called a cline) in melanization from relatively dark near the Equator to relatively light at higher or lower latitudes (**Fig. 1**). Intense UV radiation destroys folate, an essential nutrient, in the human body, which can prevent successful pregnancy or cause severe damage to the fetus. Moderate UV exposure is needed for the body to produce vitamin D_3, which is also essential for normal reproduction and growth. Heavy melanization prevents folate destruction at tropical latitudes but permits sufficient vitamin D_3 production. At temperate and arctic latitudes, folate destruction is not a problem, and lighter skin color allows for sufficient vitamin D_3 synthesis, even in winter. Some arctic peoples are relatively dark-skinned and get their vitamin D_3 from their animal food diets. The lightest skin color is found in Central and Northwestern Europe and seems to be an adaptation to long, cold winters and cloudy skies, both of which severely limit UV radiation exposure. *See* VITAMIN D.

Altitude. The earliest human ancestors in Africa lived at a moderately high altitude of 1500–2500 m (5000–8000 ft) above sea level. Today, many millions still live at this altitude, and about 25 million people live at high altitude (above 3000 m or 10,000 ft). Life at high altitude is difficult due to the combined stresses of cold temperatures, aridity, nutritional inadequacy, steep terrain, heavy workloads, and hypoxia (low partial pressure of oxygen in the air). There is evidence that some Himalayan populations (for example, the Sherpa) have a genetic adaptation to increase the saturation of oxygen in their arterial blood, which can mitigate hypoxia. Other high-altitude populations (such as the natives in the Andes Mountains) adapt to hypoxia via developmental plasticity (for example, growing larger lungs and chests). All high-altitude peoples use behavioral modification (pacing work) and culture (layered clothing and support from social networks) to accommodate to the other stresses. Even so, short stature due to nutritional inadequacy and heavy work is common.

Human adaptations to disease. Infectious and noninfectious diseases, parasites, trauma, the strains of physical activity, and the lack of physical activity are major factors influencing the direction of human adaptation, variation, and evolution. Some variations in disease are genetic (for example, malaria and sickle-cell alleles), some are the result of developmental plasticity (such as the rates of type II diabetes as a consequence of obesity and physical inactivity),

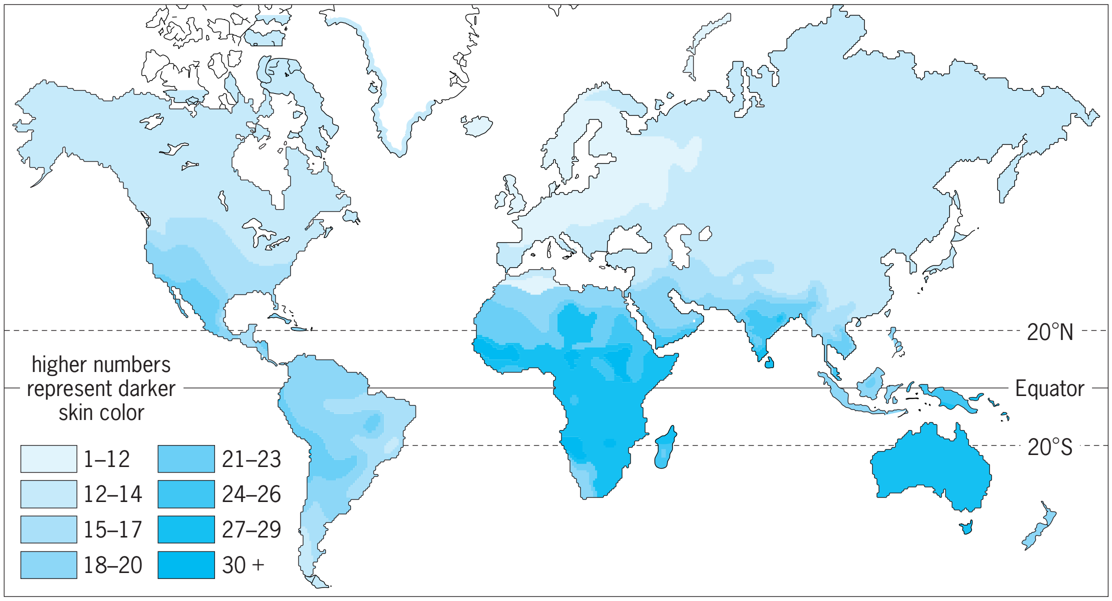


Fig. 1. Map of human skin color distribution around A.D. 1500. (*From D. O'Neil, Distribuzione della Varia Intensitá del Colore della Pelle, in Renato Biasutti, Le Razze e i Popoli della Terra, vol. 1: Razze, Popoli e Culture, 1959*)

while others are due to environmental contaminants (most cancers).

Epidemiology is the study of how, when, where, and why diseases occur. Anthropological epidemiologists focus on the biocultural nature of human diseases, that is, the behavior of humans (the host) and disease-causing agents (pathogens). To better understand the nature of disease, anthropologists also focus on the age, the sex and gender, the social status, lifestyle, ethnic, religious, and other biocultural traits of the hosts. Understanding and combating the worldwide epidemic of human immunodeficiency virus (HIV)/acquired immune deficiency syndrome (AIDS) requires such an approach. The transmission of HIV between people depends on intimate contact between hosts, such as sexual intercourse or blood transfusion. Preventing such contact stops the spread of the disease, but changing human behavior is difficult. Strongly held cultural beliefs and practices regarding sexuality, marriage, and body image (tattooing, scarification) may be vectors of transmission. Some people may believe they are immune to infection because of special social status, and some cultures without an understanding of viral infection may deny the existence of HIV or may ascribe it to spiritual causes. The goal of the anthropologist is to understand these cultural variations and at the same time apply the concepts and methods of human biology to variation in, and problems of, health and disease. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); EPIDEMIOLOGY.

Demography. Demography is the study of fertility, mortality, and migration. Anthropologists study these factors in relation to population growth, family formation, aging, and human ecology.

Fertility. Teenaged girls and women in their late thirties and forties have lower fertility than other age groups (**Fig. 2**). Boys under age 20 rarely produce offspring, but the fertility of men peaks rapidly after age 20, remains level into the fourth decade, and then declines. Biological growth and maturation dictate much of these age trends, but cultural constraints are at work as well. High levels of physical activity and low energy intake decrease female fertility. The rules regarding the formation of reproductive unions and the patterns of intercourse within unions affect the fertility of women and men. Contraceptives are used to some extent in nearly all populations, but their effectiveness varies by the technology employed, social availability and acceptability, and ideological motivation to use them. These factors result in a similar ∩-shaped curve of fertility from ages 15–45 but with differences in amplitude, with some populations achieving fertility rates more than five times greater than other populations (for example, in Fig. 2, the Hutterites versus the Portuguese).

Mortality. Mortality follows a ∪-shaped age pattern in all human populations. Death rates decline from birth to about age seven and remain relatively low throughout the next 20 years of life. Mortality of males rises in the late teens and twenties in militaristic and violent societies, such as the Yanamamö of Venezuela and the United States. By age 30, mortality begins to rise at an ever-increasing rate. In the prehistoric period the major causes of death were likely accidents and other trauma, starvation, and wear and tear on the body from rigorous lifestyles. After the advent of agriculture and settled living, the major causes of death included infectious diseases (interacting with malnutrition) and warfare. With industrialism came greater rates of death from accidents at work, alcohol abuse, and pollution. With the advent of better public health and medicine by the mid-twentieth century, the predominant causes of death in the wealthy nations became chronic degenerative diseases associated with overeating and sedentary lifestyle, automobile accidents, homicides and suicides, and drug abuse (including tobacco and alcohol). *See* AGING.

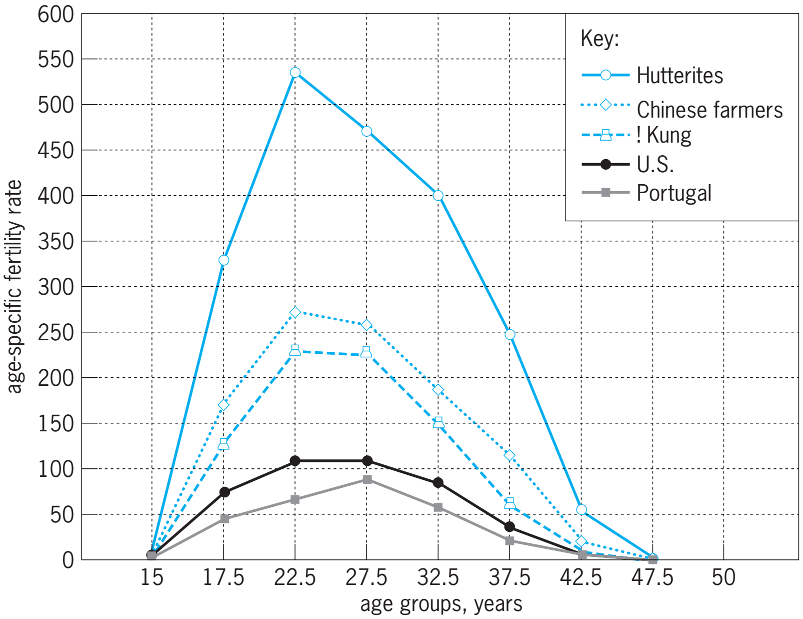


Fig. 2. Age-specific fertility rate of women in several human populations expressed as births per 1000 women. The Hutterites are a religious isolate of North America who prize high fertility and are in generally good health. The "Chinese farmers" represent a group of rural villagers, and the "!Kung" are a society of hunters and gatherers living in the Kalahari desert of Botswana in southern Africa. Lower fertility among the rural Chinese and the !Kung, compared with the Hutterites, may also be due to nutritional stress and disease. The United States is a wealthy industrialized nation in which cultural desire for smaller families combined with effective birth control technology reduces fertility at most ages. The U.S. population is slightly above replacement level, that is, growing very slowly due to in-migration. Portugal is a more extreme case of controlled fertility. The effect of out-migration of young women combined with strong desire for small families lowers fertility and results in a shrinking population over time. (*Adapted from B. Bogin, The Growth of Humanity, 2001*)

Migration. Migration moves people from place to place and redistributes human biocultural variation. Migration injects new genetic, physiological, and morphological variability into the recipient populations. It also may deplete these sources of biological variation from the nonmigrating donor population. Migrants carry diseases from their home to their new places of residence, plus they are exposed to new diseases en route and once they arrive at their destination. Migrants also develop new diseases due to the special circumstances of their life and behavior. Migration rates change due to economic, social, and political circumstances (for example, the slave trade or warfare). Migration rates tend to be highest in young adults (20 to 30 years old). All of these social and biological factors make the study of human migration difficult, intellectually exciting, and of much practical importance for human welfare. *See* POPULATION DISPERSAL.

Population growth. The balance of fertility versus mortality is responsible for population growth. By the late Neolithic Period (8000 BP) there were about 5 million people on Earth. Today there are more than 6 billion. The exponential growth of the global population is predicted to peak at about 9 billion people by the end of the twenty-first century. *See* POPULATION ECOLOGY.

Conclusion. Today anthropologists and human biologists appreciate the essentially adaptive nature of human variability and the importance of the concept of biocultural adaptability to the study of biological variation. This understanding opens opportunities for greater tolerance, reduced conflict, better health, and greater welfare for the entire human species.

Barry Bogin

Bibliography. B. Bogin, *The Growth of Humanity*, 2001; R. Lewontin, *The Triple Helix: Gene, Organism and Environment*, 2000; J. Marks, *Human Biodiversity: Genes, Race and History*, 1995; S. Stinson et al. (eds.), *Human Biology: An Evolutionary and Biocultural Perspective*, 2000.

Human-computer interaction

An interdisciplinary field focused on the interactions between human users and computer systems, including the user interface and the underlying processes which produce the interactions. The contributing disciplines include computer science, cognitive science, human factors, software engineering, management science, psychology, sociology, and anthropology. Early research and development in human-computer interaction focused on issues directly related to the user interface. Some typical issues were the properties of various input and output devices, interface learnability for new users versus efficiency and extensibility for experienced users, and the appropriate combination of interaction components such as command languages, menus, and graphical user interfaces (GUI). Since the late 1990s, the field of human-computer interaction has broadened and become more attentive to the processes and context in which the users' experience with human-computer interactions takes place. The focus of research and development is now on understanding the relationships among users' goals and objectives, their personal capabilities, the social environment, and the designed artifacts with which they interact. As an applied field, human-computer interaction is also concerned with the development process used to create the interactive system and its value for the human user.

Example of interaction. The interfaces and processes that make up human-computer interaction are understood and advanced through a variety of methods reflecting the field's interdisciplinary nature. Consider, for example, a recent university graduate using a personal computer to search for a job on the Internet. At one level, this interaction can be characterized by the capabilities and processes of the human and the computer to accept input, process that input, and generate output. The computer capabilities include the hardware (input and output devices) such as the monitor, mouse, keyboard, and Internet connection. These devices reflect contributions from computer science and engineering, whereas the human capabilities, both mental and physical, are understood through cognitive science and ergonomics. *See* COMPUTER PERIPHERAL DEVICES.

At another level, the interaction between the computer and the human consists of user interface software which governs the meanings of the inputs and outputs for the computer, as well as the corresponding rules and expectations that the user applies to generate meaningful actions. The software in this example includes an Internet browser with a graphical user interface that reflects recent advances in software engineering and multimedia design. The user's internal model of the interaction is supported by visual cues in the interface and designed in accordance with principles of human factors. *See* HUMAN-FACTORS ENGINEERING.

At a higher level, this interaction includes the context of goals, motivations, and other people and resources that determine what the person is doing, as in this example, searching the Internet for job openings. Understanding the process at this level requires insights from social and organizational sciences.

Advances. Recent advances in knowledge about human-computer interactions across all these levels, from technology to social context, have contributed to the ongoing evolution of the field and to the resulting products and processes available.

Technology. Advances in computer science have significantly increased the processing power of computers while decreasing their size. These advances have provided the underlying technology for creating a wider variety of human-computer interactions. For example, streaming audio and video over the Internet, now common, would not be possible without the increased processing power and network connectivity of computers. These technological developments were influenced by the discovery of useful applications in human-computer interaction. Increasingly sophisticated software has become available to address input through natural speech and immersive environments, providing a virtual reality experience. *See* VIRTUAL REALITY.

Human capability. There is a growing body of knowledge that is providing a good understanding of how humans learn and work with computers. Many factors need to be considered, including models of cognition, motivation, individual differences, and human diversity. Studies on memory, perception, motor skills, attention, learning and skill acquisition, and vigilance are important. Also, there are ergonomic issues that must be considered, such as sensory limits, fatigue, health, temperature, and environmental noise.

User interface. The user interface includes the input and output devices—such as the monitor, keyboard, and mouse—and the methods by which the users interact with them to carry out tasks. For example, the graphical user interface on most personal computers currently uses a windowing screen designed so that the user can have more than one software application active at any given time. The interface design must provide ways to share both technological and human resources, such as by allocating space on the monitor screen between the active applications and those not in use. These user interface issues are often addressed by interdisciplinary teams. For example, the concept of a windowed environment for multiple applications was investigated by an interdisciplinary team lead by Alan Kay at the Xerox Palo Alto Research Center and popularized in the Macintosh and Windows operating systems. Such empirical studies collect data on how people make use of these facilities, and guide the design of the next generation of the interface to better fit human capabilities and expectations, which are constantly evolving as computer use becomes more ubiquitous. Of particular importance in recent years have been advances in applications fostering human-human collaborations, in visualizing and manipulating multidimensional problem spaces, and in supporting navigation and information retrieval in complex interactive spaces.

Contextual design. Methods from fields such as management sciences, sociology, and anthropology have provided insights into the overall context in which interaction is taking place. For example, people often work in teams to reach goals, so that the interactions taking place between an individual team member and a computer must be considered within the larger framework of human-human communications. The issues involved would be significantly different from the situation of a user who is physically alone but who is participating in a game with fellow players on the Internet, even though the underlying technologies and the software processes might be similar.

Development process. Developing human-computer interactions involves design on both sides of the interaction. On the technology side, the designer must have a thorough understanding of the available hardware and software components and tools. On the human side, the designer must have a good understanding of how humans learn and work with computers, including envisioning new modes of working. The designer's task is to create effective, efficient, and satisfying interactions by balancing factors such as cost, benefits, standards, and the environmental constraints in which the interaction will take place.

Modern prototyping tools allow for the use of an iterative development model where a representative portion of the interface is designed and implemented with each iteration. Feedback from testers is used to enhance the design with each iteration. The final design consists of many elements: the resulting artifacts for use by the target population, as well as supporting elements such as an analysis of needs and tasks, descriptions of the dialog rules and users' conceptual models, expected scenarios of use, and the designer's rationale and reflections from the development process.

Evolution. With the rapid technological advances in interactive computer systems, it is inevitable that new technologies will raise new human-computer interaction issues. For example, the explosive use of multimedia over the Internet has raised issues at various levels. Computer scientists are concerned with providing the highest-quality video possible, given the current computer processing capabilities. User-interface designers are addressing new issues, such as how to index digital video. The types and varieties of human-computer interactions and human-human conversations mediated by the Internet have increased significantly as it has gained wide acceptance at work, school, and home. Understanding the context of these interactions has become more complex as computer use has become more widespread. Perhaps the major challenge facing human-computer interactions is the speed of technological change. The field of human-computer interactions must build the foundations for users to experience incremental change, even though it may not know where those changes will lead. Tom Carey; Kevin Harrigan

Bibliography. M. Helander et al. (eds.), *Handbook of Human-Computer Interaction*, 2d ed., 1997; J. Preece (ed.), *Human Computer Interaction*, 1994; B. Shneiderman, *Designing the User Interface*, 3d ed., 1997.

Human ecology

The study of how the distributions and numbers of humans are determined by interactions with conspecific individuals, with members of other species, and with the abiotic environment. Human ecology encompasses both the responses of humans to, and the effects of humans on, the environment. The lineage leading to modern humans arose in Africa millions of years ago. The oldest fossil remains of members of the genus *Homo* are estimated to be about 2 million years old. The tools associated with human fossils, as well as cave paintings created by early humans, reveal the existence of culture. Cultural learning greatly facilitated the spread of domestic plants and animals and the development of pastoral and agricultural societies. Agriculture, in turn, led to an increasingly sedentary life, greatly expanded food supplies, the development of cities, and the rapid growth of the human population. Human ecology today is the combined result of humans' evolutionary nature and cultural developments. *See* BIOSPHERE; ECOLOGICAL COMMUNITIES; ECOSYSTEM.

Responses to environments. Humans' strong positive and negative emotional responses to components of the environment evolved because our ancestors' responses to environmental information affected survival and reproductive success. Early humans needed to interpret signals from other organisms and the abiotic environment, and they needed to evaluate and select habitats and the resources there. These choices were emotionally driven.

A vital step in the lives of most organisms, including humans, is selection of a habitat in which to live, gather resources, and seek shelter. Evolutionary theory predicts that habitats that evoke strong positive responses should be those in which survival and reproductive success have been high. Conversely, habitats in which survival and reproductive success have been low should evoke weak or negative emotional responses. The responses of people on several different continents conform to these predictions. Food is one of the most important resources provided by the environment. Gathering food requires decisions of where to forage and what items to select. Anthropologists often use the theory of optimal foraging to interpret how these decisions are made. The theory postulates that as long as foragers have other valuable ways to spend their time or there are risks associated with seeking food, efficient foraging will be favored even when food is not scarce. This approach has facilitated development of simple foraging models and more elaborate models of food sharing and gender division of labor, symbolic communication, long-term subsistence change, and cross-cultural variation in subsistence practices.

Social systems. Humans rely heavily on culture to adapt to their environments. Cultural evolution is a process in which individuals copy others or are taught and then become models for others. But people do not imitate others at random; they pick and choose whom and what to imitate. Thus, cultural evolution proceeds by a process of biased transmission, guided by behavioral rules that have been molded by natural selection. Studies of the diffusion of technical innovations are used to understand the pragmatic decision-making techniques that people use when considering whether to adopt an innovation.

Prior to domestication of plants and animals, humans lived in small bands that shifted locations in response to changing distributions of food resources in the environment. Some of these nomadic people, such as Australian aboriginals, developed elaborate stories and songs that transmitted complex geographical knowledge to young people. Pastoral societies continued with similar nomadic traditions and knowledge.

Domestication of plants and the rise of agriculture, which began about 11,000 years ago in the Mideast, increasingly led to a more sedentary existence because crops needed to be tended and agricultural products could be stored for consumption during lean periods. In addition, by replacing complex ecological communities dominated by species of no food value to humans, with communities dominated by a few species of plants and animals that humans can eat, agriculture increased from 10 to 100 times the amount of food produced per unit land area. Food surpluses made possible the rise of cities, hierarchical societies, and specialized professions. Large domestic mammals revolutionized human society by replacing human backs as the major mode of land transport of goods. Eurasia's horses played a major role in wars of conquest, and maintained their role in military assault until World War I.

Ecological conditions have exerted a powerful influence on human history. The evolution of complex technologies by Eurasians was favored by the locations of continents and mountain ranges. Eurasia offered more readily domesticated plants and animals than other continents. Diffusion of innovations was easier in Eurasia than in other regions because the east-west-oriented mountain ranges facilitated dispersal of domesticated organisms and technological advances across regions at comparable latitudes and, hence, with similar climates. Also, as a result of millennia of living at high densities in association with large domestic animals, Eurasians developed immunity or resistance to an array of diseases with which people on other continents had no experience. Thus, when Europeans explored the other continents they introduced diseases that were lethal to the inhabitants. People from other continents had no comparable diseases to transmit to Europeans.

The rise of cities, a dominant feature of recent human history, was a major result of domestication of plants and animals. A city is supported by resources gathered from regions beyond its area. For example, 29 cities in the Baltic Sea region appropriate the production, in the form of wood, paper, fiber, and food, of ecosystems that occupy about 200 times the total area of the cities themselves. If cities are to be sustainable, extensive supporting ecosystems must continue to provide renewable resources. Most governmental policies pay insufficient attention to these services. Indeed, existing subsidies often deliberately or inadvertently impair the functioning of supporting ecosystems.

Influences on environments. Significant modification of the environment by people was initiated by the domestication of fire, used to change vegetation structure and influence populations of food plants and animals. Vegetation burning is still common in the world, particularly in tropic regions.

By about 1 million years ago, humans had spread from Africa to eastern Asia; Europe was colonized about 500,000 years ago. About 50,000 years ago, humans invented sophisticated tools for capturing and subduing a wide array of prey, including large mammals. Tool-bearing people arrived suddenly in Australia about 40,000 years ago; they crossed the Bering Land Bridge into North America about 12,000 years ago. Within 2000 years humans had spread across the entire New World. Finally, people colonized the islands of the Pacific Ocean, a process that was completed a little more than 1000 years ago.

The arrival of humans with sophisticated tools precipitated the next major transformation of Earth, the extinction of large vertebrates. When humans arrived, Australia and New Guinea had a suite of large birds and mammals, including giant kangaroos, rhinolike marsupials the size of cows, large marsupial carnivores, a 400-lb (182-kg) flightless bird, and giant lizards, snakes, and land-living crocodiles. All were extinct within a few thousand years of human arrival. Today the largest native mammals of the region are 100-lb (45-kg) kangaroos. The large birds and mammals of New Zealand, Madagascar, and the Hawaiian Islands also became extinct shortly after those islands were colonized by humans.

When humans arrived in North America about 12,000 years ago, they encountered a rich and varied fauna of large mammals that included elephants, horses, giant bison, camels, giant ground sloths, lions, and cheetahs. Within a few centuries, all had become extinct. Debate still surrounds the cause of those extinctions, but those mammals had survived many advances and retreats of glaciers and massive climate changes. Their disappearance precisely when humans arrived strongly suggests that humans sealed their fates. Extinctions of large mammals were far fewer in Africa and Eurasia, where human culture evolved slowly and large animals had ample time to adapt to the presence of this curious predator.

Whatever the cause of the extinctions of the megafaunas of the New World and Australia, humans there lost the opportunity to domesticate large mammals. When horse-riding Spaniards arrived in the New World, they defeated vastly greater numbers of Aztec and Incan foot soldiers. The conquest of the New World was probably the first major sociological consequence of a human-caused loss of biodiversity.

Agriculture drove the third major human modification of environments. Agricultural lands are manipulated so that their productivity is channeled primarily to human needs. Humans appropriate terrestrial production directly as food and fiber and indirectly as food for domestic animals. Today about 35–40% of terrestrial primary production is appropriated by people, and the percentage is rising. Soon, only about half of terrestrial primary production will be available for use by all other species on Earth.

Determining the percentage of aquatic primary production appropriated by people is more difficult. The world fisheries catch was about 131 million metric tons per year in 1990, about 85% of which came from wild stock. By assigning harvested species to trophic levels and using models of the efficiency of trophic energy transfer, ecologists calculated that about 8% of the Earth's primary aquatic production is required to support current fisheries. However, humans use 24–35% of the primary production of continental shelves and upwelling areas where most fish are harvested. Although about 75% of aquatic production occurs in the open oceans, those regions are so unproductive that it is profitable to harvest only the largest predators, such as tuna.

Humans appropriate an even higher percentage of the world's fresh water. Only about 0.77% of the Earth's water exists as fresh water in the atmosphere, aquifers, lakes, streams, soils, and organisms. About 97.5% of global water is salty; the remainder exists as permanent ice. Humans currently use about 54% of available runoff for agriculture, industry, municipal needs, and disposal of wastes. At current per capita rates of usage, the growing human population will need more than 70% of available runoff by 2025.

The high level of human appropriation of ecosystem production, combined with the landscape modifications it engenders, is currently the most important cause of species extinctions. The habitats required by some species are being destroyed. Habitats such as old-growth forests, natural grasslands, and estuaries are being reduced to widely separated patches that are too small to sustain populations of many of the species. Small habitat patches cannot maintain populations of species that require large areas, and they support only small populations of many species that can survive in them. In addition, the fraction of a patch that is influenced by effects originating on adjacent habitats rapidly increases as patch size decreases. As human appropriation of ecosystem production increases, more land will be converted to agriculture, thereby increasing the fragmentation of natural habitats.

The increasingly mobile human population is introducing, either deliberately or inadvertently, thousands of species into some areas. These exotic species often cause the extinction of native species upon which they prey, with which they compete, or with which they share diseases. Biogeographers may come to refer to the current period in human history as the Homogocene.

The magnitude of the current human enterprise is great enough to modify the Earth's biogeochemical cycles. Primarily as a result of burning fossil fuels, the concentration of atmospheric carbon dioxide has risen from a pre–industrial revolution value of about 265 parts per million to 350 ppm today. If current trends continue, concentrations are expected to reach 580 ppm by the middle of the twenty-first century. The buildup of carbon dioxide, which is transparent to sunlight but opaque to radiated heat, is expected to increase the Earth's mean temperature by about 3–5°C (5–9°F), with greater increases at high latitudes. Climate patterns will shift latitudinally, new climates will come into existence, polar icecaps will melt, and sea levels will rise, flooding coastal regions. As world climate changes, the ranges of many species must shift. Individuals will need to disperse in increasingly fragmented landscapes in which only a small fraction of the area is suitable for them. In the past, range shifts took place in landscapes that offered much easier dispersal routes. *See* BIOGEOCHEMISTRY; GREENHOUSE EFFECT.

Other biogeochemical cycles are also being substantially modified. Currently the magnitude of the global sulfur cycle is about twice the preindustrial level. Increases in sulfur emissions increase atmospheric concentrations of dimethyl sulfide, the major source of condensation nuclei, causing increases in cloud cover and additional climate changes. As a result of the massive increases of fixed nitrogen being introduced into the Earth's ecosystems by human activity, concentrations of nitrous oxide and nitric oxide are increasing in the atmosphere. Nitrous oxide absorbs infrared radiation and contributes to global warming. Nitric oxide is involved in reactions that create peroxyacyl nitrate, a constituent of smog. Nitric oxide is also converted to nitric acid, a principal component of acid precipitation. Acid precipitation causes decreases in soil fertility, acidifies lakes and streams, and percolates into ground water where it causes health problems in communities that consume that water. *See* ACID RAIN.

The future. Humans will continue to exert powerful influences on the functioning of the Earth's ecological systems. The human population is destined to increase for many years. Rising affluence will be accompanied by increased consumption of resources and, hence, greater appropriation of the Earth's primary production. Nevertheless, many future human ecology scenarios are possible, depending on how much the human population grows and how growth is accommodated, the efficiency with which humans use and recycle resources, and the value that people give to preservation of biodiversity. Some of the most important ethical decisions faced by human society today concern relationships with and responsibilities to the other species that share the Earth. The ethical positions that people adopt during the next few decades will determine to a substantial degree the actual ecological scenarios experienced by our distant descendants. *See* ANTHROPOLOGY; ECOLOGY; ENVIRONMENT; SOCIOBIOLOGY. Gordon H. Orians

Bibliography. S. C. Bourassa, *The Aesthetics of Landscape*, Belhaven Press, 1991; J. M. Diamond, *The Third Chimpanzee*, HarperCollins, 1992; J. M. Diamond, *Guns, Germs, and Steel*, Norton, 1997; P. R. Ehrlich, *A World of Wounds: Ecologists and the Human Dilemma*, Ecology Institute, 1997; T. F. Flannery, *The Future Eaters*, Reed Books, 1994; R. T. T. Forman, *Land Mosaics: The Ecology of Landscapes and Regions*, Cambridge University, 1995; G. Hardin, *Living Within Limits*, Oxford University, 1993; E. A. Smith and B. Winterhalder (eds.), *Evolutionary Ecology and Human Behavior*, Aldine de Gruyter, 1992; E. O. Wilson, *Biophilia*, Harvard University, 1984.

Human-factors engineering

The application of experimental findings in behavioral science and physiology to the design and operation of technical systems in which humans are users or operators. This includes design of hardware, software, training, and documentation as well as manufacturing and maintenance. Human-factors professionals are trained in some combination of experimental or cognitive psychology, physiology, and engineering—typically industrial, mechanical, electrical, or software engineering. Human-factors engineering seeks to ensure that humans' tools and environment are best matched to their physical size, strength, and speed and to the capabilities of the senses, memory, cognitive skill, and psychomotor preferences. These objectives are in contrast to forcing humans to conform or adapt to the physical environment.

Human-factors engineering has also been termed human factors, human engineering, engineering psychology, applied experimental psychology, ergonomics, and biotechnology. It is related to the field of human-machine systems engineering but is more general, comprehensive, and empirical and not so wedded to formal mathematical models and physical analysis. *See* HUMAN-MACHINE SYSTEMS.

Among the problems of human-factors engineering are design of visual displays for ease and speed of interpretation; design of tonal signaling systems and voice communication systems for accuracy of communication; design of seats, workplaces, cockpits, and consoles in terms of humans' physical size, comfort, strength, and visibility. Human-factors engineering addresses problems of physiological stresses arising from such environmental factors as heat and cold, humidity, high and low atmospheric pressure, vibration and acceleration, radiation and toxicity, illumination or lack of it, and acoustic noise. Finally, the field includes psychological stresses of work speed and load and problems of memory, perception, decision making, and fatigue. *See* HUMAN-MACHINE SYSTEMS.

History. The foundations of human-factors engineering were laid by Fredrick W. Taylor, who demonstrated that, by proper design of workplaces and procedures, the productivity of workers could be greatly increased. Early systematic studies were made by Frank B. Gilbreth and Lillian M. Gilbreth, whose therblig (an anagram after Gilbreth) system of categorizing hand movements is still a standard motion-and-time analysis technique. With World War II came a great demand for psychologists and physiologists to help engineers design aircraft, ships, tanks, and other weapons to ensure that these devices could be operated under stress by people with relatively little training. All too often it was found that engineers had designed the equipment around themselves and, as a result, large sailors could not fit in the required spaces, small pilots could not reach the required controls, or an appreciable percentage of operators were lacking in sufficient strength or were confused by procedures for operating the complex implements of war. After World War II, formal courses in human-factors engineering were introduced in psychology and engineering departments of colleges and professional organizations in the United States and in many European countries.

Early efforts were directed to providing experimental data for the more obvious gaps in human-factors knowledge, such as the variations in the physical dimensions of men and women of different ages, the strength and speed capacities of humans, and the physical conditions under which they could just barely detect or read certain visual images and hear or discriminate certain sounds. As more powerful techniques come to be developed, other sensory capacities of touch, taste, smell, and motion, as well as the more complex and subtle problems of training, stress, and fatigue, became amenable to laboratory experiment.

Anthropometry. The specialized field dealing with the physical dimensions of the human body is called anthropometry. In dimensioning a seat, console, workplace, or special piece of personal equipment (such as a space suit), it is important that as large a fraction of the intended users as possible be accommodated. But fitting a console to the largest person usually means the smallest person cannot reach the controls unless various features of the equipment are adjustable. Moreover, people are not shaped in simple proportion; the person with the largest head often does not have the longest legs or greatest girth. There is no average human: the person who is average in one dimension is not usually average in another (**Fig. 1**). To accommodate everyone, many features of a machine need to be adjustable independently. Because this is not always possible, it is important in sizing the equipment to know what percentage of large people and what percentage of small people will not be accommodated, and to weigh the relative costs and advantages of adding adjustments or building the equipment in different sizes. *See* ANTHROPOMETRY.

Displays and controls. Problems of visual displays have been of considerable interest in designing aircraft, spacecraft, ships, submarines, nuclear reactors, electronic equipment, and tools of all kinds. Human-factors engineers have sought to specify the light intensity required for reading different dials and signs in relation to the illumination level of the background. They have recommended, for specific tasks and environments, optimum shape and spacing of numerals and indicator marks on dials and specified what colors give the best contrast and how far away markings of certain size can be read.

Electronics have made possible a number of new visual display techniques: Self-illuminating electroluminescent numerals and holographic photographs, when illuminated by coherent (laser) beams, appear strikingly three-dimensional. Cathode-ray tubes driven by computer can provide a tremendous variety of images, including letters and numbers, graphs, and television-type pictures. In aircraft and other applications, single-purpose dials and indicators have been replaced by relatively few computer displays, which are completely flexible and can display a great variety of kinds of information. Sometimes these displays can be called up by the operator by keying in certain code letters. Alternatively, a number of different indications, both qualitative and quantitative, can be integrated on one display, such as the integrated aircraft landing display. *See* AIRCRAFT INSTRUMENTATION; CATHODE-RAY TUBE; COMPUTER GRAPHICS; ELECTRONIC DISPLAY.

With respect to sound signaling systems (auditory displays), human-factors engineers can specify how loud various sounds must be to be heard against background noises of differing loudness. Because consonants, which have relatively higher frequencies and lower loudness, are known to carry much more information than vowels, the fidelity of voice communication systems at these higher frequencies (2000–4000 Hz) is recognized to be more important than the fidelity at lower frequencies. Electronic reading machines and mobility aids for the blind make considerable use of auditory signals to replace visual senses.

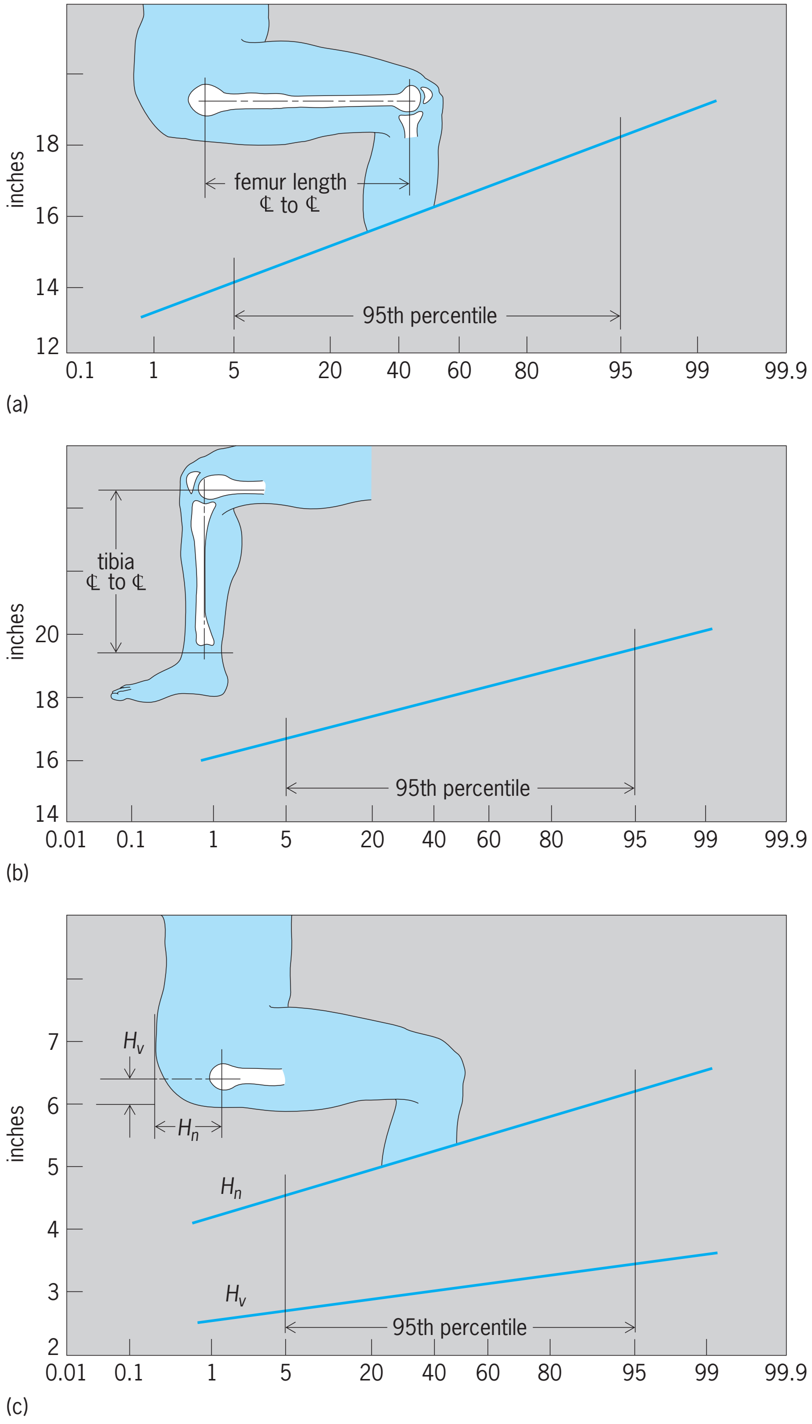


Fig. 1. Variations in the distributions of body dimensions as studied in military personnel. (*a*) Femur length. (*b*) Tibia length. (*c*) Femur-to-seat distances. (*After L. J. Fogel*, *Biotechnology*, *Prentice-Hall*, *1963*)

An important human-factors engineering problem is coordination of displays with the controls by which responses to these displays are made. In practice, there are many violations of even simple commonsense design principles such as locating controls adjacent to displays when they are used together, arranging for controls to move in the same direction as the associated display, or having all displays and controls move in the same direction to turn them off. Although the human operator can adapt to a reasonably broad range of forces and sizes of controls, some controls even now are installed that require either more strength or sensitivity than many operators can exercise. Human-factors handbooks are available that summarize various experimental data and give recommendations for display and control design.

Environmental stress. Human-factors engineers are concerned with the design of the environment, both to enhance humans' information-processing ability and to keep the body functioning normally. These concerns require an understanding of human tolerance ranges to heat and cold, to high and low pressures, to varying atmospheric concentrations, and to acceleration and vibration. Engineering solutions can take the form of space suits, diving suits, special clothing for extreme climates, and atmospheric-control units for sealed cabins and tanks. Alternatively, working conditions can be so arranged as to limit exposure to hazard, for example, by determining the least fatiguing distance from the open door of a blast furnace or by recommending the maximum time that can safely be spent at certain altitudes or depths without pressure-breathing equipment.

Bodily comfort is dependent upon a combination of ambient temperature, pressure, humidity, air movement, and amount of clothing—all of which interact to keep the internal body temperature constant at 98.6°F (37°C). Proper respiration depends upon the combination of ambient pressure and oxygen concentration that keep the lungs at the equivalent of breathing air (20% oxygen) at sea level. Constant high accelerations, experienced by astronauts in boost and reentry, can drain the blood from eyes and brain and temporarily prevent the body's proper functioning; such adverse effects can be prevented by orienting the body properly or using g suits, which keep blood from pooling in the veins of the trunk and legs. Vibration is likely to cause motion sickness if it occurs at certain frequencies at which body parts resonate sympathetically (worst at 2–5 Hz). Levels of thermal, pressure, atmospheric, acceleration, and vibration stresses, which are individually experienced as mild, can have violent effects when they occur in combination. *See* AEROSPACE MEDICINE; SPACE BIOLOGY.

Safety. Human-factors considerations often determine whether a particular vehicle, tool, or environment is safe. The ability of the human body to withstand various sudden acceleration forces is an important aspect of highway safety. Seat-belt designs for automobiles are based on rocket sled experiments originally conducted to design crash harnesses and parachute harnesses for aircraft pilots.

Human reaction time is also a large factor in accidents of all types. It has been shown that human reaction under the best conditions, such as those that exist when an alert, expectant test subject is required to push a button in response to a light signal, is about a quarter second. As the number N of response choices from among which the subject must choose increases, the reaction time increases as the logarithm of N. Reaction time can increase many times as a function of boredom and fatigue, but reliable quantitative predictors for those factors are not available.

Work load. Much research has tried to specify the optimum workload for humans. This research is motivated by the gradual disappearance of routine tasks which can be performed at an even pace and somewhat independently of the environment. Humans are used as a monitor of complex semiautomatic systems where the workload for most of the time is relatively light but where detection of certain low-probability contingencies, or failure of the system, requires rapid and dependable overt action. Examples occur in industrial inspection processes, monitoring of nuclear reactors, piloting of spacecraft, and rapid transit trains.

Early studies on ship's watchkeeping and radar watching indicated that humans could remain vigilant only about a half hour for signals that occur rarely if ever (such as a ship on a collision course with another at night or in fog or the approach of enemy aircraft or missiles). After this interval the chances of rare events being detected diminish markedly. One technique employed to keep watchkeepers and monitors alert is to introduce artificial signals that the observer initially cannot discriminate from the real ones. For example, it was reported that in an experiment one soft-drink-bottle inspector passed fewer dirty or defective bottles when, to keep him alert, cockroaches were added to a small percentage of otherwise clean bottles!

The Manned Spacecraft Center of NASA has conducted simulations of long space flights where not only was the workload negligible but in some cases sound, light, temperature, and gravity sensations were reduced to a minimum. Under such conditions humans become decreasingly alert and even hallucinate to substitute imagined sensation in compensation for the lack of real sensation.

At the other end of the workload scale, human-factors research has determined the upper limits on what humans can do with their sense organs and muscles, both in transmitting information and in performing mechanical work. In terms of the unit of information measure, the bit-per second (the average logarithm of the number of choices or binary decisions made per second), humans can transmit up to 35 bits/s in such tasks as piano playing and speaking, but for most routine manipulation skills their rate is much lower. As an engine for mechanical work, humans are rather inefficient because, even while operating at peak efficiency (for example, in pedaling a racing bicycle), experiments show that they can produce 1 hp (750 W) for only about 1 min—and in order to do so their bodies produce several times that amount of energy in wasted heat.

Humans versus machine. A fundamental problem of ever-increasing importance for human-factors engineers is what tasks should be assigned to people and what to machines. It is a fallacy to think that any given whole task can be accomplished best either by a human or by a machine without the aid of the other, because often some elements of both provide a mixture superior to either alone.

Machines are superior in speed and power; are more reliable for routine tasks, being free of boredom and fatigue; can perform computations at higher rates; and can store and recall specific quantitative facts from memory faster and more dependably. Humans, by contrast, have remarkable sensory capacities which are difficult to duplicate in range, size, and power with artificial instruments (the ratio of the greatest to the least energy which people can either see or hear is about 10^{13}). Humans' ability to perceive patterns, make relevant associations in memory, and induce new generalizations from empirical data remains far superior to that of any computer existing or planned. Thus, while people's overt information-processing rate in simple skills is low, their information-processing rate for these pattern recognition and inductive- reasoning capabilities (of which little is understood) appears far greater.

Problem areas. Below are described four different problem areas of human-factors engineering, and within each, two different specific problems are described. These illustrate the variety of challenges encountered in human-factors engineering. For each problem the design solutions require that the hardware, software, and training considerations be integrated.

Information display. Problems of information display include those of space telepresence and those of displays in nuclear and chemical plants.

Space telepresence. Crewed space flight is very expensive, and currently is not practical beyond the Moon for safety reasons. Exploration of the planets may be accomplished more economically by using robotic space vehicles and telepresence. Telepresence means that a human, in this case on Earth, can look around, see, and have the sense of being located in an environment other than her actual physical location. This is accomplished by means of a head-mounted display, which measures the orientation of the head and communicates this to a video camera in the remote environment, directing it to point in the same direction. The human then sees what she would see were she in that environment looking in that direction. This sense of telepresence is very compelling, and human-factors engineers are studying its limitations and refining this equipment for use in space. One limitation, for example, is the time delay in video and control signal communication between Earth and space, which means the video feedback lags the corresponding head direction by seconds to minutes, depending how far out in space is the planet being explored.

Displays in nuclear and chemical plants. Another type of display problem is the design of computer displays for nuclear power plants and chemical process plants. Because there are literally thousands of variables that must be monitored, for safety reasons it is important that these displays combine information about related variables, and give an overall picture of the health of the plant. There are several types of information to be displayed. First there are alarms, both auditory and visual, indicating what has failed and providing some direction about how to respond. Then there are status lights, showing a pump is on or off, a valve is open or closed. There are also indications of numerical values of key variables, pressures, temperatures, flow rates, and so on. Finally there are electrical and piping diagrams and written information, much of which is computerized to save space and make it more quickly accessible. *See* NUCLEAR REACTOR; PROCESS CONTROL.

Control. Problems of control involve aircraft automation and robots.

Aircraft automation. Modern commercial aircraft have a sophisticated control system called a flight management system. The flight management system is a computer-based system which includes the autopilot, the navigation subsystem, a radar-based traffic collision avoidance subsystem, and many other functions. Mostly the pilot flies the aircraft through the flight management system, and especially on long flights uses the manual controls (yoke and pedals) only for taxiing, takeoff, and landing, and not necessarily even then. A properly programmed flight management system has the capability to fly the aircraft automatically from takeoff to landing without the pilot ever touching the controls. The flight management system is programmed through a keyboard resembling a pocket calculator. However, there remain many problems of pilot error in using the flight management system, requiring significant human-factors effort.

Control of robots. Control of a robot, whether for painting, welding, or parts assembly in a manufacturing plant; for performing inspections and repairs on deep-ocean wellheads or pipes; or for working in chemically toxic or other hazardous environments, is similar to the flight management system in many ways. A human supervisor must plan a task, program its special features, and monitor the robot after it is put into automatic mode, updating its instructions as task conditions change. *See* ROBOTICS.

Safety research. Problems of safety research include those involving field studies in automotive safety and those involving the use of a human-in-the-loop simulator for evaluation.

Field studies in automotive safety. Human-factors specialists are engaged in a variety of field studies to investigate automobile accidents, examine roadway configurations to recommend signing, and test new designs of automobiles and trucks from the driver's point of view. What a normal driver can be expected to see or hear, and how he or she is expected to respond, are issues which must be backed up by statistics.

Human-in-the-loop simulator. Often vehicle safety research must be performed which is too dangerous to do in actual circumstances. In that case, use of a human-in-the-loop simulator is warranted. This is a device that is operated by a real person but in other respects is an electromechanical simulation. The operator uses realistic controls and views a computer-generated display which provides a realistic dynamic view of how the vehicle responds, giving the driver the feeling of being in an actual vehicle. The flight simulator is the best-known example of such a system. These have become so realistic that pilots are checked out in new aircraft types on simulators, and the first time they fly the aircraft is with a load of passengers. Such simulators are used for automobiles as well, enabling study of driver response in crash situations without actual danger to the driver. *See* AIRCRAFT TESTING.

Health care. Problems of health care include those involving doctors' instructions and those of in-home monitoring for the chronically ill.

Doctors' instructions. A well-known source of human error in hospitals has been the misreading of handwritten physician orders to nurses, hospital laboratory technicians, and pharmacists. A solution is to have physicians enter their instructions into computers. Then not only is there immediate feedback to the physician as to what was typed and an unambiguous record which can easily be transmitted to wherever it is needed, but also the computer can perform certain checks to see whether the medicine in the prescribed dosage is appropriate for what is known (in the computerized hospital database for the patient).

In-home monitoring. People who have serious chronic medical conditions such as congestive heart problems often are not aware of when they are in danger and when there is no immediate cause for concern. Some people are not treated until it is too late, while others unnecessarily crowd the emergency rooms of hospitals. An avenue of solution which can avoid time and expense for both the patient and medical personnel is to have monitoring devices at home which are connected by telephone modem to a hospital. Either the patient can attach such devices, or in more serious cases a nurse or health care worker visiting the home regularly can assist the patient in making certain measurements. The measuring device can make a simple analysis, comparing the measurement to readings made on the same patient on previous days, and decide if the situation is safe, that is, not getting worse. Otherwise the data can be telemetered to a hospital database and analyzed there by a specialist for appropriate action. To make such home care systems safe, effective, and user-friendly poses many human-factors challenges.

Thomas B. Sheridan

Cognitive engineering. One of the major issues in human-factors engineering is the concern that modern sophisticated hardware and software technology may be too complex for the people who will eventually use it. Requiring people to perform difficult or cognitively complex tasks is perhaps the leading cause of human–machine errors or accidents. Tasks can be cognitively difficult for a number of reasons. The number of steps required to use the system may exceed people's memory limitations, the user may be required to divide his or her attention between several different sources of information, or the person may be required to perform difficult mental operations. All of these factors burden an individual's cognitive capacity and, if that capacity is exceeded, errors may occur.

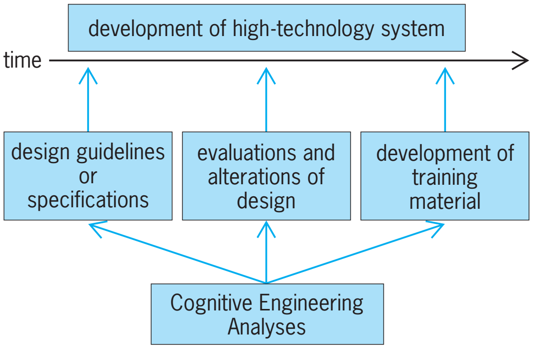

development of high-technology system
time
design guidelines or specifications
evaluations and alterations of design
development of training material
Cognitive Engineering Analyses

Fig. 2. Cognitive engineering and the development of new technology applications.

Two major trends have led to increased emphasis on the cognitive complexity of human-machine systems. One of these is the move toward larger and more complex systems where human error can have serious consequences for the systems' users and the general public. The other trend is the rapid development of modern information technology based upon powerful yet inexpensive microcomputers.

An important aspect in addressing this problem is the early identification and control of the cognitive complexity or mental difficulty of performing a task required of the new technology application. Identification and control of the mental difficulty of tasks is the goal of cognitive engineering. This aspect of engineering is accomplished in different ways depending on the development status of the new technology application (**Fig. 2**).

The best procedure (in terms of cost and effectiveness) for addressing this problem is to use cognitive analysis to develop specifications or guidelines that can be used during the initial design technology application. Such specifications might include instructions such as "do not ask the user to remember more than two items simultaneously" or "put all relevant information in a single location on the equipment." Successful designs are achieved through planning and sensitivity to user needs. Detailed qualitative models of the cognitive processes of the system's users serve important roles in these analyses. They can suggest the overall complexity of different design options for users, and indicate those aspects of each option that may exceed human cognitive capabilities. Through such design guidance, human cognitive limitations are controlled early in the technology application design when it is easiest and most cost-effective to make changes.

If the technology application has developed to the point where design guides would no longer be useful (for example, much design work has already been completed), an alternative approach is to use cognitive engineering to evaluate the design as it exists. The results of the cognitive analysis will indicate which aspects of the design may be too difficult for people to perform and could lead to human–machine errors. Those aspects of the techonology application can then be redesigned in order to reduce the likelihood of such error. The disadvantage of this approach is that redesigning hardware or software is often more expensive than designing it correctly in the first place.

The final use of cognitive engineering analysis is in preparing training materials. Cognitive analyses can identify the aspects of a person–machine interface that will be most difficult for people to perform. These aspects can then be given special training aids or more intensive hands-on training in order to reduce the potential of human–machine error. Paul G. Rossmeissl

Engineering psychology. The science of psychology has always been concerned with the basic functions that enable humans to interact effectively with their environment: sensing, attending, perceiving, remembering, choosing among actions, acting. Researchers have typically worked under carefully controlled laboratory conditions, using experimental methods derived from the older physical sciences.

Topics of greatest interest to engineering psychology at any given time are dictated by technological developments and the problems they pose for the designer and operator. Many issues revolve around computer-based systems with their almost limitless potential for automation, information handling, and modes of interfacing with human operators.

Automation. Since almost every system operation can be automated, designers must decide which functions should be left for the human operator. Although the choice seems to depend on the particular system, the trend is clearly toward giving human operators more overseeing and decision responsibility and less moment-to-moment control of system processes. Today's commercial airline pilot, for example, manages and monitors the flight to a great extent rather than actually flying the airplane.

Information processing. Given their immense capability for gathering, manipulating, storing, and displaying information, modern systems can easily swamp the human operator with information. This abundance of information raises important research issues about human capacity limits and ways of handling information, and about the mental effort required by particular tasks and system designs. Systems engineers must determine how mental workload should be measured, what is an acceptable amount of it, and how performance suffers as that amount is exceeded. Ultimately, the goal is to reduce mental workload without sacrificing information that the operator needs in order to perform well. Engineering psychologists have gained considerable insight and made some valuable practical innovations regarding these issues. For example, sound techniques have been developed for measuring the subjective mental workload associated with particular tasks, and these measures can be used to estimate the relative merit of alternative designs from a human performance perspective.

Human-machine interface. One way of coping with task complexity and workload is through improved design of the human-machine interface, that is, the way that information is displayed to the operator and responses are executed. Modern technology offers a host of interface options—from ordinary text to sophisticated graphics on the display side; from keyboards, to menus and touch screens, to "mice," joysticks, trackballs, and speech sensors on the response side. Some combinations of these design options (including applications of techniques such as virtual reality) can make the operator's task considerably easier, in effect lessening the complexity and the resulting mental workload. *See* VIRTUAL REALITY.

The best combinations of such tools, however, are not immediately apparent from a human-performance perspective. What works best is heavily dependent on how the operator is to process the information—the task requirements. The key relationship between the representation of to-be-processed information and processing requirements is known as compatibility. For information-rich, complex tasks, the compatibility of design features with the mental representation of the task (the operator's mental model) is what matters most. Perceived complexity and mental workload are lowest (hence operator and system performance are best) when the system design is compatible with these mental models.

Decision making. Human decision making under the conditions of high stress, uncertainty, and information load that are experienced by firefighters, air-traffic controllers, fighter pilots, and even automobile drivers in an unfamiliar city, requires processing a lot of information quickly and choosing the right course of action when a wrong one could have serious, even catastrophic, consequences.

There are major differences in the way that effective and ineffective decision makers approach this kind of task. Effective ones compare their past experience in similar situations when faced with a new crisis. Rather than trying to analyze the presently available information in depth, they simply scan their memory for a similar pattern and act accordingly. Poor decision makers, however, focus on only the current information and tend to become over-whelmed. Knowing how effective performance is accomplished, designers can build features into new systems that encourage operators to use the proper strategy and that aid them in doing so. Moreover, training can be focused on this approach. William C. Howell

Bibliography. S. Andriole and L. Adelman, *Cognitive Systems Engineering for User-Computer Interface Design, Prototyping and Evaluation*, Lawrence Erlbaum Associates, Mahwah, NJ, 1995; R. W. Bailey, *Human Performance Engineering*, 3d ed., Prentice Hall, Upper Saddle River, NJ, 1996; A. Chapanis, *Human Factors in Systems Engineering*, John Wiley, New York, 1996; W. C. Howell, Engineering psychology in a changing world, *Annu. Rev. Psychol.*, 44:231–263, 1993; J. Rasmussen, A. M. Pejtersen, and L. P. Goodstein, *Cognitive Systems Engineering*, John Wiley, New York, 1994; G. Salvendy (ed.), *Handbook of Human Factors and Ergonomics*, 2d ed., John Wiley, New York, 1997; M. S. Sanders and E. J. McCormick, *Human Factors in Engineering Design*, 7th ed., McGraw-Hill, 1993; B. Shneiderman, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3d ed., Addison-Wesley, Reading, MA, 1997; C. D. Wickens and J. G. Holland, *Engineering Psychology and Human Performance*, 3d ed., Prentice Hall, Upper Saddle River, NJ, 1999; C. E. Zsambok and G. Klein (eds.), *Naturalistic Decision Making*, Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

Human genetics

A discipline concerned with genetically determined resemblances and differences among human beings. The idea that certain physical and mental characteristics, normal or abnormal, can "run in families" goes back to ancient times, though the mechanism by which heredity operates remained mostly unknown until the twentieth century. Formerly, genetics was thought to be concerned only with the familial transmission of rare and insignificant characteristics, but its fundamental biological role is now apparent. Genes, the units of heredity, have two unique properties: they are self-replicating, and they carry in their biochemical structure the codes for protein synthesis. Consequently, genes play the double role of transmitting genetic information from generation to generation and of governing all the activities of living cells. *See* GENE.

Expansion of human genetic knowledge has come from several different directions and has had major consequences for human health and for the understanding of the place of humans in nature. Technological advances in the visualization of human chromosomes have shown that abnormalities of chromosome number or structure are surprisingly common and of many different kinds, and that they account for birth defects or mental impairment in many individuals as well as for numerous early spontaneous abortions. Progress in molecular biology has clarified the molecular structure of chromosomes and their constituent genes and the mechanisms of deoxyribonucleic acid (DNA) synthesis and protein synthesis, as well as the ways in which change in the molecular structure of a gene can lead to a disease. Concern about possible genetic damage through environmental agents, particularly ionizing radiation, and the possible harmful effects of hazardous substances in the environment on prenatal development has also stimulated research in human genetics. The medical aspects of human genetics have become prominent as nonhereditary causes of ill health or early death, such as infectious disease or nutritional deficiency, have declined, at least in developed countries. It has been estimated, for example, that half of all admissions to pediatric hospitals involve disorders that are completely or partly genetic. Detailed knowledge of the distribution of many genetic traits in individuals, families, and populations has expanded, providing fresh insights into human origins and the process of evolution.

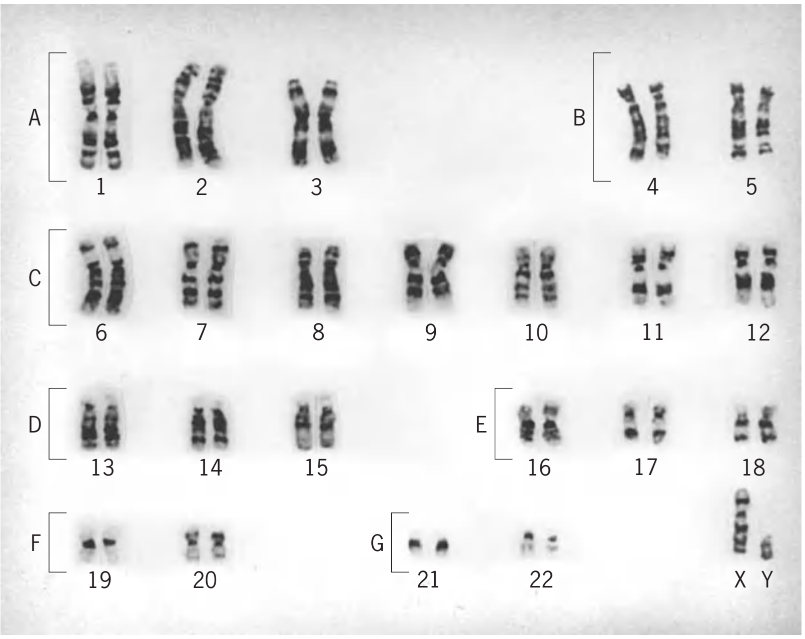


Fig. 1. Normal male karyotype. The chromosomes are classified in seven groups labeled A–G. The chromosome pairs are individually labeled. (*Photomicrograph courtesy of R. G. Worton*; *from J. S. Thompson and M. W. Thompson*, *Genetics in Medicine*, *3d ed.*, *W. B. Saunders*, *Philadelphia, 1980*)

Chromosome and gene structure. In normal humans, the nucleus of each normal cell contains 46 chromosomes, which comprise 23 different pairs. A karyotype (**Fig. 1**) may be prepared from any type of cell that will divide in cell culture, such as white blood cells, skin, bone marrow, or cells from amniotic fluid, for example. Of each chromosome pair, one is paternal and the other maternal in origin. In turn, only one member of each pair is handed on through the reproductive cell (egg or sperm) to each child. Thus, each egg or sperm has only 23 chromosomes, the haploid number; fusion of egg and sperm at fertilization will restore the double, or diploid, chromosome number of 46. *See* CHROMOSOME; FERTILIZATION (ANIMAL).

The segregation of chromosome pairs during meiosis allows for a large amount of "shuffling" of genetic material as it is passed down the generation. With 23 pairs of chromosomes, the total possible number of different chromosome combinations in the gametes is 2^{23}, or about 8 million. Two parents can provide $2^{23} \times 2^{23}$ different chromosome combinations. This enormous source of variation is multiplied still further by the mechanism of crossing over, in which homologous chromosomes exchange segments during meiosis. *See* CROSSING-OVER (GENETICS); MEIOSIS.

Twenty-two of the 23 chromosome pairs, the autosomes, are alike in both sexes; the other pair comprises the sex chromosomes. A female has a pair of X chromosomes; a male (Fig. 1) has a single X, paired with a Y chromosome which he has inherited from his father and will transmit to each of his sons. Sex is determined at fertilization, and depends on whether the egg (which has a single X chromosome) is fertilized by an X-bearing or a Y-bearing sperm. *See* SEX DETERMINATION.

All chromosomes are composed of DNA. It is estimated that the total length of DNA in the haploid set of human chromosomes is 1.7 m (5.7 ft), which is 10,000 times the length of the chromosomes at metaphase. Genes are segments of DNA. There may be as many as 50,000–100,000 genes on the 46 human chromosomes, of which about 3000 have been identified and about 1500 have been mapped to specific chromosomal locations. *See* DEOXYRIBONUCLEIC ACID (DNA); GENETIC MAPPING.

The genetic information is coded in DNA in the form of triplets of four bases: two purines, adenine (A) and guanine (G), and two pyrimidines, thymine (T) and cytosine (C). Each triplet combination (codon) codes for a specific amino acid. The sequence of bases in a specific gene dictates the sequence of amino acids in the specific protein coded by that gene. *See* GENETIC CODE.

A gene initiates synthesis of a protein by transcription of the DNA into messenger ribonucleic acid (mRNA), which is a single-stranded molecule complementary to the DNA. After processing within the cell nucleus, the mRNA moves into the cytoplasm where it binds to ribosomes. The translation of RNA into protein takes place on the ribosomes, which are small particles in the cytoplasm composed of protein and a special type of RNA, ribosomal RNA. Another type of RNA, transfer RNA, of which there is one type for each amino acid, moves amino acids from the cytoplasm to the mRNA molecule, where they are aligned in the sequence dictated by the genetic code carried in the mRNA. The sequence of amino acids forms a polypeptide, which is released from the ribosome when complete. An average protein contains about 500 amino acids; thus an average gene contains about 1500 base pairs. Since the haploid chromosome set contains 3 billion base pairs, the chromosomes must contain a large excess of DNA in addition to the DNA in the constituent genes. *See* PROTEIN; RIBONUCLEIC ACID (RNA); RIBOSOMES.

A typical gene is not a simple uninterrupted length of DNA, but most genes are made up of coding sequences (exons) separated by noncoding regions (intervening sequences, or introns). Transcription of the gene into RNA begins at an initiation site in advance of the first coding sequence and terminates beyond the end of the last sequence. After the entire DNA segment has been transcribed into mRNA, the intervening sequences are removed and the coding sequences are spliced together before the mRNA is translated into a polypeptide (**Fig. 2**). *See* EXON; INTRON.

Any gene occupies a specific chromosomal position, or locus. The alternative genes at a particular locus are said to be alleles. If a pair of alleles are identical, the individual is homozygous; if they are different, the individual is heterozygous. *See* ALLELE.

Mutation. Genetic variation has its origin in mutation. Although in broad terms any change in DNA is a mutation, whether it is a microscopically detectable change in the structure of a chromosome or a single base change in the genetic code, the term is usually applied to stable changes in DNA that alter the genetic code and thus lead to synthesis of an altered protein.

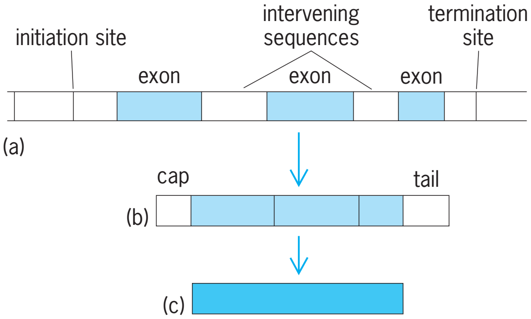


Fig. 2. Simplified diagram showing (*a*) the structure of a gene, (*b*) the mature messenger RNA after the two intervening sequences have been spliced out and the "cap" and "tail" added, and (c) the polypeptide translated from the messenger RNA.

Mutation can occur in reproductive cells or somatic cells, but the genetically significant ones are those that occur in reproductive cells and can therefore be transmitted to future generations. Natural selection acts upon the genetic diversity generated by mutation to preserve beneficial mutations and eliminate deleterious ones.

A very large amount of genetic variation exists in the human population. Everyone carries many mutations, some newly acquired but others inherited through innumerable generations. Though the exact number is unknown, it is likely that everyone is heterozygous at numerous loci, perhaps as many as 20%. *See* MUTATION.

Single-gene inheritance. The patterns of inheritance of characteristics determined by single genes or gene pairs depend on two conditions: (1) whether the gene concerned is on an autosome (autosomal) or on the X chromosome (X-linked); (2) whether the gene is dominant, that is, expressed in heterozygotes (when it is present on only one member of a chromosomal pair and has a normal allele) or is recessive (expressed only in homozygotes, when it is present at both chromosomes). *See* DOMINANCE.

Margaret W. Thompson

Quantitative inheritance. A quantitative trait is one that is under the control of many factors, both genetic and environmental, each of which contributes only a small amount to the total variability of the trait. The phenotype may show continuous variation (for example, height and skin color), quasicontinuous variation (taking only integer values—such as the number of ridges in a fingerprint), or it may be discontinuous (a presence/absence trait, such as diabetes or mental retardation). With discontinuous traits, it is assumed that there exists an underlying continuous variable and that individuals having a value of this variable above (or below) a threshold possess the trait.

A trait that "runs in families" is said to be familial. However, not all familial traits are hereditary because relatives tend to share common environments as well as common genes. It is the major task of quantitative genetics to disentangle the effects of environment and heredity, but this task is not easy in humans where environment and heredity are often confounded.

The variability of almost any trait is partly genetic and partly environmental. A rough measure of the relative importance of heredity and environment is an index called heritability. R. A. Fisher showed that the total variance (a statistical measure of variability) of a trait can be partitioned into a genetic variance and an environmental variance, at least in simple cases. The quantitative model can be expressed as:

$$\text{Variance (phenotypic)} = \text{variance (genetic)} + \text{variance (environmental)} \qquad (1)$$

Heritability is then defined as:

$$\text{Heritability} = \frac{\text{variance (genetic)}}{\text{variance (phenotypic)}} \qquad (2)$$

In humans, the heritability of height is about 0.75. That is, about 75% of the total variance in height is due to variability in genes that affect height and 25% is due to exposure to different environments. Expressed in another way, the difference in height between two individuals is, on the average, 75% genetic and 25% environmental.

There are actually two kinds of heritability. The one described above, usually called broad heritability, measures the total effect of heredity. The other, narrow heritability, is the proportion of total phenotypic variance resulting from the additive effects of genes. Narrow heritability is a more subtle concept but is actually more useful. For example, the correlation between parent and child is equal to one-half the narrow heritability plus their environmental correlation. Equations of this sort relate heritability to observable phenotypic correlations, thus providing a means of estimating the relative importance of heredity and environment. Carter Denniston

Hereditary Diseases

Medicine is an important field for the practical application of human genetics; medical genetics has become an integral part of preventive medicine (that is, genetic counseling, including prenatal diagnostics). It has contributed increasingly to systematics of disease, diagnostics, and even therapy. Many external causes of disease, such as infections, have been brought under control in the twentieth century; therefore, doctors can devote much of their skill to treating diseases from internal sources, that is, hereditary diseases, or those brought about by interaction of genetic predispositions with certain stress factors in the environment. Widespread use of genetic knowledge in medical practice has important consequences for basic science: problems posed by the numerous and often unexpected observations in medical genetics help in developing basic theory, and suggest new approaches in research.

Hereditary diseases may be subdivided into three classes: chromosomal diseases; hereditary diseases with simple, mendelian modes of inheritance; and multifactorial diseases.

Chromosomal diseases. One out of 200 newborns suffers from an abnormality that is caused by a microscopically visible deviation in the number or structure of chromosomes. Such chromosomal aberrations are much more common (≈40%) among spontaneous miscarriages. About 10–20% of all recognized pregnancies terminate in spontaneous miscarriage and many more embryos die during the first weeks of pregnancy, when fetal loss goes unnoticed. Hence, a large fraction of all human zygotes are abnormal chromosomally and die early. *See* CHROMOSOME ABERRATION.

The most important clinical abnormality among the survivors is Down syndrome—a condition due to trisomy of chromosome 21, one of the smallest human chromosomes. The term trisomy means that this chromosome is present not twice but three times; the entire chromosome complement therefore comprises 47, not 46, chromosomes. Down syndrome occurs one to two times in every 1000 births; its pattern of abnormalities derives from an imbalance of gene action during embryonic development. Down syndrome is a good example of a characteristic pattern of abnormalities that is produced by a single genetic defect. Such patterns, recognizable to the experienced observer as syndromes, are found not only in chromosomal diseases but in hereditary diseases with simple (mendelian) modes of inheritance as well. *See* DOWN SYNDROME.

Other autosomal aberrations observed in living newborns that lead to characteristic syndromes include trisomies 13 and 18 (both very rare), and a variety of structural aberrations such as translocations (exchanges of chromosomal segments between different chromosomes) and deletions (losses of chromosome segments). Translocations normally have no influence on the health status of the individual if there is no gain or loss of chromosomal material (these are called balanced translocations). However, carriers of balanced translocations usually run a high risk of having "unbalanced" offspring—children in whom the same translocation causes gain or loss of genetic material, and who suffer from a characteristic malformation syndrome.

Clinical syndromes caused by specific aberrations vary, but certain clinical signs are common: low birth weights (small for date); a peculiar face; delayed general, and especially mental, development, often leading to severe mental deficiency; and multiple malformations, including abnormal development of limbs, heart, and kidneys. Single malformations in children who otherwise develop normally are not typical for a chromosomal aberration. *See* CONGENITAL ANOMALIES.

Less severe signs than those caused by autosomal aberrations are found in individuals with abnormalities in number (and, sometimes, structure) of sex chromosomes. This is because in individuals having more than one X chromosome, the additional X chromosomes are inactivated early in pregnancy. For example, in women, one of the two X chromosomes is always inactivated. Inactivation occurs at random so that every normal woman is a mosaic of cells in which either one or the other X chromosome is active. Additional X chromosomes that an individual may have received will also be inactivated; in trisomies, genetic imbalance is thus avoided to a certain degree. However, inactivation is not complete; therefore, individuals with trisomies—for example, XXY (Klinefelter syndrome), XXX (triple-X-syndrome), or XYY—or monosomies (XO; Turner syndrome) often show abnormal sexual development, intelligence, or behavior.

In some individuals, chromosomal aberrations are found only in some cells. Clinical signs in these cases are often milder. Sometimes, a new mutation giving rise to a structural chromosomal aberration may occur in a somatic tissue, leading, for example, to a translocation only in one cell and its descendants. Sometimes, especially when a chromosome break involved in this translocation has affected an oncogene, these cells may have a selective advantage, and develop into a malignant tumor. An example is the translocation between chromosomes 22 and 9 found in chronic myelonic leukemia. *See* MOSAICISM; ONCOGENES.

Diagnosis of chromosomal aberrations, especially in newborns with multiple malformations, individuals with disturbances of sexual development, and parents suffering from multiple miscarriages, are of practical importance, since in many cases monitoring of future pregnancies by prenatal diagnosis is possible.

Diseases with mendelian inheritance. In contrast to chromosomal aberrations, the genetic defects in hereditary diseases with simple, mendelian modes of inheritance cannot be recognized by microscopic examination; as a rule, they must be inferred more indirectly from the phenotype and the pattern of inheritance in pedigrees. The defects are found in the molecular structure of the DNA. Often, one base pair only is altered, although sometimes more complex molecular changes, such as deletions of some bases or abnormal recombination, are involved. Methods of molecular biology have permitted in some cases direct analyses of such defects at the gene level.

Approximately 1% of all newborns have, or will develop during their lives, a hereditary disease showing a simple mendelian mode of inheritance.

Mendel called an allele “dominant” when the homozygote and the heterozygote were indistinguishable phenotypically; in experimental genetics, the same convention is still being used. In medical genetics, the terms dominant and recessive are not used so strictly. A condition is called dominant if the heterozygotes deviate in a clearly recognizable way from the normal homozygotes, in most cases by showing an abnormality. Since such dominant mutations are usually rare, almost no homozygotes are observed and their clinical condition is, in most cases, unknown. In exceptional instances, homozygotes for dominant conditions have been described; they have usually shown more severe clinical signs than the heterozygotes.

Cross between homozygotes of the same allele A

	AA	Gametes	
AA		A	A
Gametes	A	AA	AA
	A	AA	AA

Genotypes of children: AA

Backcross between homozygote AA and heterozygote AA′

	AA′	Gametes	
AA		A	A′
Gametes	A	AA	AA′
	A	AA	AA′

Genotypes of children: 1AA : 1 AA′

Cross between two heterozygotes AA′

	AA′	Gametes	
AA′		A	A′
Gametes	A	AA	AA′
	A′	AA′	A′A′

Genotypes of children: 1 AA: 2 AA′ : 1A′A′

Cross between homozygotes AA and A′A′

	AA	Gametes	
A′A′		A	A
Gametes	A′	AA′	AA′
	A′	AA′	AA′

Genotypes of children: AA′

Fig. 3. Examples of mendelian crosses using two alleles, A and A′.

In **Fig. 3** there are four mendelian crosses for one pair of alleles A and A′. Cross no. 2, the backcross between the normal homozygote and the heterozygote, which leads to 1:1 segregation, is found most commonly in autosomal dominant inheritance. Crosses 3 and 4 are extremely rare, since homozygotes for an abnormal allele are usually rare. In autosomal-recessive inheritance, backcrosses between normal homozygotes and heterozygotes (cross no. 2) are also common, but since heterozygotes are normal phenotypically, such crosses will go unnoticed in most instances. The most common cross leading to homozygous, clinically affected offspring is the intercross between two heterozygotes, AA′, which results in 1:2:1 segregation.

Autosomal-dominant inheritance. This type of inheritance pattern is often determined by studying the history of a trait among a group of relatives. Such a history, called a pedigree, can be represented in a standard chart (**Fig. 4**). A pedigree in which a rare autosomal-dominant condition is transmitted through four generations is shown in **Fig. 5**. The mutant allele is located on an autosome; the affected individuals are heterozygous. Since one of the two alleles at this gene locus is altered by mutations and each of the two alleles has a 50% probability of being transmitted to a child (Fig. 3, cross no. 2), each child has a 50% risk of being affected. Since this mutation is autosomal it is transmitted independently of the sex chromosomes and the risk of being affected is not influenced by sex of parents or child.

The pedigree in Fig. 5 demonstrates the characteristic features of autosomal-dominant inheritance. For conditions such as achondroplastic dwarfism, however, such large pedigrees are the exception rather than the rule. In most cases, known pedigrees extend over two generations only. It is also common that one person is the only affected individual in a family. These individuals owe their abnormal allele to a fresh mutation in the germ cell of one of their parents, but the risk for their children to become

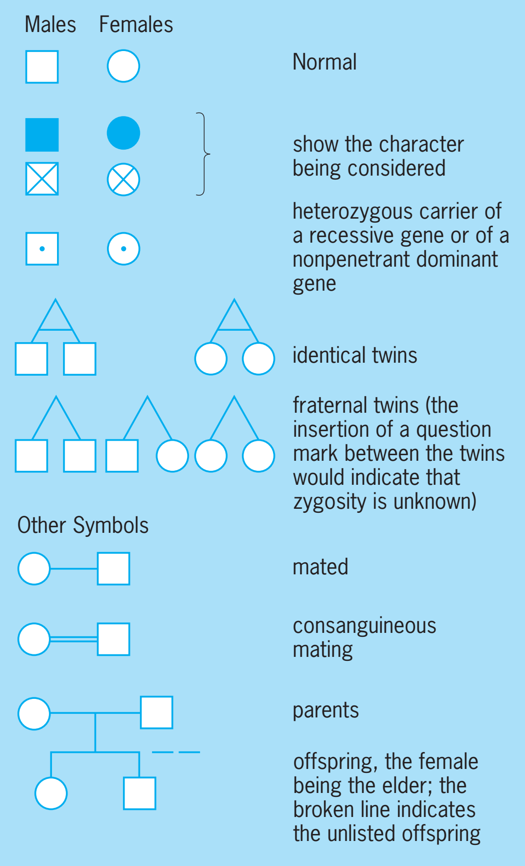

Males
Females
Normal
show the character being considered
heterozygous carrier of a recessive gene or of a nonpenetrant dominant gene
identical twins
fraternal twins (the insertion of a question mark between the twins would indicate that zygosity is unknown)
Other Symbols
mated
consanguineous mating
parents
offspring, the female being the elder; the broken line indicates the unlisted offspring

Fig. 4. Some symbols in common use in the United States for presenting human pedigrees.

affected is, nevertheless, 50% (**Fig. 6**). The fraction of new mutants among all carriers of a certain dominant disease must be high if the disease impairs average reproduction of its carriers: the harmful alleles are eventually eliminated from the population since many carriers have no children. When the carriers are so severely affected that they have no children at all, each new mutant will be eliminated in the first generation, and all cases in a population are new mutants. On the other hand, large pedigrees with many affected individuals are usually observed when the anomaly is relatively harmless or manifests itself later in life, at a time when the carriers already have had their children. An example is Huntington's disease, a severe degenerative disease of the brain: in most gene carriers, the first clinical signs become visible only between 40 and 50 years of age; they die after many years of progressive deterioration of brain function.

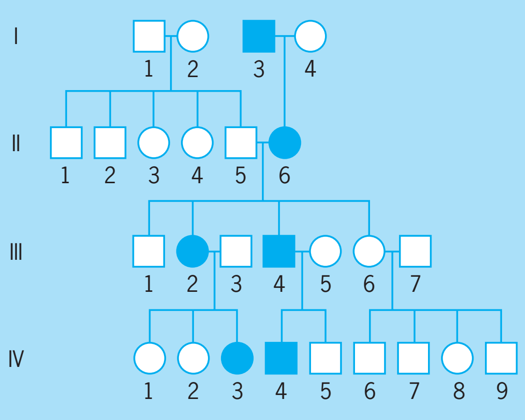


Fig. 5. Typical pedigree pattern of a rare autosomal dominant trait: classical achondroplastic dwarfism. Those with this anomaly, caused by a defect in growth of the long bones, have extremely short arms and legs, but are otherwise normal. I–IV are generations; shaded individuals manifest the trait.

In some dominant conditions, the harmful phenotype may not be expressed in a gene carrier (this is called incomplete penetrance), or clinical signs may vary in severeness between carriers (called variable expressivity). Penetrance and expressivity may be influenced by other genetic factors; sometimes, for example, by the sex of the affected person, whereas in other instances, the constitution of the "normal" allele has been implicated. Environmental conditions may occasionally be important. In most cases, however, the reasons are unknown.

Autosomal-recessive inheritance. A pedigree for albinism, that is, with an autosomal-recessive trait, is shown in **Fig. 7**. The affected individuals (IV, 2 and 6) have two mutant alleles, one from each parent; they are homozygous for the albinism gene. Hence, their unaffected parents, III, 4 and 5, must both be heterozygous for this allele. The probability of two individuals having the same allele increases when some of their genes come from a common ancestor (that is, when they are relatives). Indeed, the mothers of the parents in Fig. 4 are sisters; hence, the parents themselves are first cousins. An increase of matings between relatives, for example, first cousins, in comparison with the population average is characteristic for rare autosomal-recessive diseases. There is a segregation of 1:2:1 between homozygotes AA, heterozygotes Aa, and homozygotes aa. This means that every child from a mating of two heterozygotes (Fig. 3, cross no. 3) has a 25% risk of being homozygous for the mutant allele, and thus affected.

In the first decades of the twentieth century, when the first autonomal-recessive conditions such as albinism were discovered, both criteria—an increase of matings between relatives, and appearance of the disease in 25% of children—were often observed, and were useful for genetic analysis. In the 1980s, both indicators have become much rarer: with the average number of one to two children per marriage that is observed in many industrialized countries, there will be one affected sib only in the great majority of all sibships. Since the fraction of first-cousin marriages is only two to three per 1000 marriages, even a tenfold increase does not lead to an appreciable fraction among parents of such patients. Hence, recognizing an autosomal-recessive mode of inheritance has become more difficult.

As a rule of thumb, autosomal-dominant mutations often lead to structural anomalies such as malformation syndromes, whereas autosomal-recessive mutations can often be traced back to a biochemical abnormality, for example, an enzyme defect.

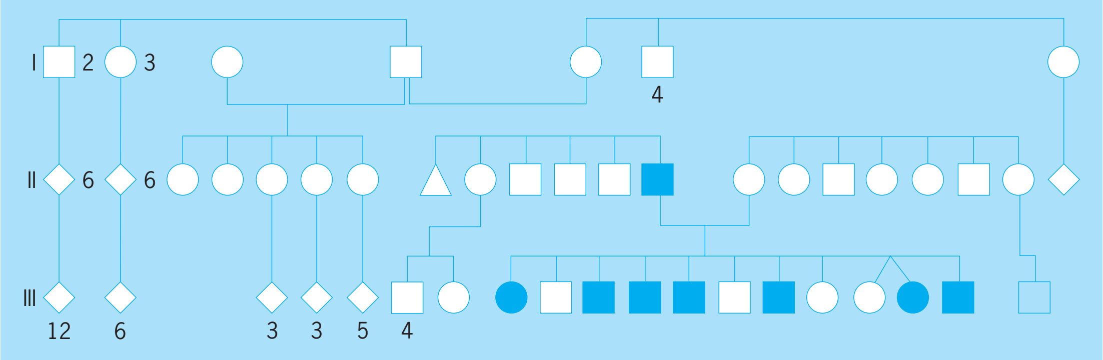


Fig. 6. Pedigree with new mutation to autosomal dominant aniridia (defect of irises). The mutation must have occurred in the germ cell of the father or mother of the patient in generation II. Diamonds represent the indicated number of unaffected offspring. (*After F. Vogel and A. G. Motulsky, Human Genetics: Patterns and Approaches, Springer-Verlag, 1979*)

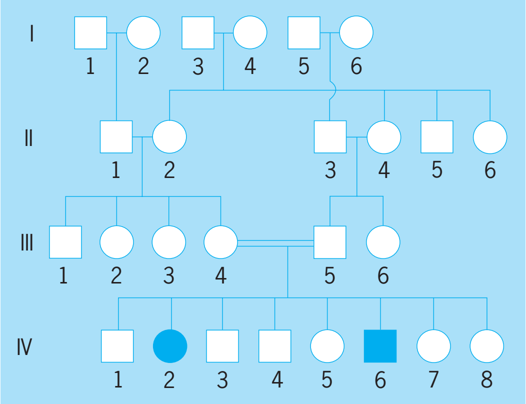


Fig. 7. Typical pedigree pattern of a rare autosomal recessive trait, one form of albinism. Note, however, that sibships with so many sibs as well as marriages between first cousins have become rare in populations of industrialized countries.

X-linked inheritance. X-linked modes of inheritance occur when the mutant allele is located on the X chromosome. The family patterns result from the mechanism of phenotypic sex determination: children receiving a Y chromosome from their fathers become males, and those receiving the paternal X become females. The mothers contribute an X chromosome to children irrespective of their sex. Therefore, all daughters and no sons receive an X-linked mutant allele from their fathers, whereas sons as well as daughters may receive such an allele from the mothers, the chance always being 50%, if the mother is heterozygous for the trait.

The most important X-linked mode of inheritance is the recessive one (**Fig. 8**). Here, the males (referred to as hemizygotes since they have only one allele) are affected, since they have no normal allele. The female heterozygotes, on the other hand, will be unaffected, since the one normal allele is sufficient for maintaining function. A classical example is hemophilia A, in which one of the serum factors necessary for normal blood clotting is inactive or lacking. (The disease can now be controlled by repeated substitution of the deficient blood factor—a good example for phenotypic therapy of a hereditary disease by substitution of a deficient gene product.) As shown in Fig. 8*a*, male family members are affected whereas their sisters and daughters, while being unaffected themselves, transmit the mutant gene to half their sons. Only in very rare instances, when a hemophilic patient marries a heterozygous carrier, are homozygous females observed (Fig. 8*b*). As for autosomal dominant diseases, the classic pedigrees are rare for severe X-linked conditions. Again, many of them are new mutants, and most actual pedigrees are small.

Some X-linked conditions are dominant, such that female heterozygotes express the abnormal trait. As a rule, however, their clinical signs are milder than those found in male hemizygotes. In quite a few such conditions, male hemizygotes are so severely damaged that they die even before birth, and are aborted. Hence, (almost) exclusively female patients are observed; there is a 1:1 ratio of normal and affected daughters of affected mothers, sex ratio among liveborn children is shifted in favor of girls, and

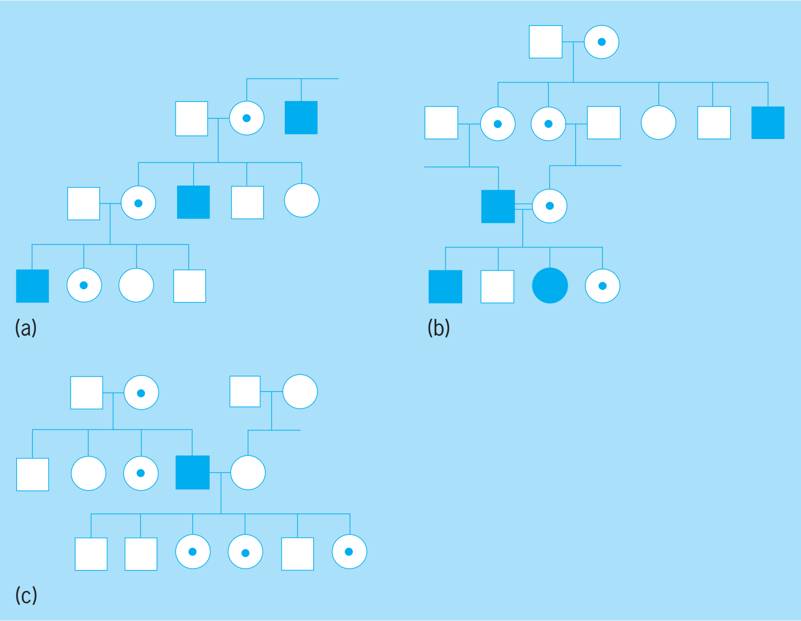


Fig. 8. Typical pedigree patterns of a sex-linked recessive trait, hemophilia A. (*a*) Only males are affected and usually come from unaffected mothers who often have affected fathers or brothers. (*b*) An affected female can result from an affected father and a carrier woman. (*c*) All daughters of affected males are carriers.

miscarriages are common. Mutant genes located on the Y chromosome would be expected to be transmitted from the father to sons only, but to all sons. No such example has been confirmed. *See* SEX-LINKED INHERITANCE.

Multifactorial diseases. There are thousands of hereditary diseases with simple mendelian modes of inheritance, but most common anomalies and diseases are influenced by genetic variability at more than one gene locus. Most congenital malformations, such as congenital heart disease, cleft lip and palate, neural tube defects and many others, fall into this category, as do the constitutional diseases, such as diabetes mellitus, coronary heart disease, anomalies of the immune response and many mental diseases, such as schizophrenia or affective disorders.

All of these conditions are common and often increase in frequency with advanced age. In industrialized societies, almost everybody dies sooner or later of such a multifactorial disease, in which neither a chromosome aberration nor a simple mode of inheritance can be detected. The influence of genetic constitution can be indicated in a variety of ways, including a higher concordance of monozygotic as compared with dizygotic twins; a higher incidence of the same disease among relatives of affected probands than in the general population; and similarity of adopted children with their biological, not their adoptive, parents. However, there is no simple 1:1 relationship between genotype and phenotype in these cases; a variety of genetic factors may be involved. Moreover, the environment usually contributes significantly to the disease risk, which can be concluded from the observation that monozygotic twins, while being affected concordantly more often than dizygotics, are discordant in many cases.

In the absence of a more penetrating analysis, such a situation is described preliminarily by the genetic model of multifactorial inheritance in populations with a threshold effect (**Fig. 9**). The conceptual background of the multifactorial model may be described as follows: many undefined genes cooperate in creating a disease liability, which is distributed normally among individuals of a population. When this liability exceeds a threshold, the individual will be clinically affected. Instead of a sharp threshold, as shown in Fig. 9, a threshold area may be assumed in which manifestation depends on additional environmental factors. From such a genetic model, and the theory of quantitative genetics, some conclusions may be derived that are often found to apply more or less to empirical data; for example, that close relatives are affected more often than remote relatives, or that relatives of more severely affected probands run a higher risk than those of less severely affected ones. However, the multifactorial model does not attempt to identify, much less to characterize, the individual genes involved in a disease liability. The genotype is treated as "black box."

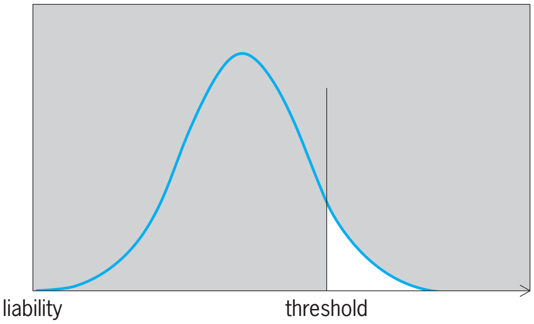


Fig. 9. Multifactorial inheritance in combination with a threshold effect. The figure shows the distribution of disease liabilities in a population (ordinate: number of individuals; abscissa: degree of disease liability). This distribution is assumed to be normal: many individuals have an average liability; in a few, liability is low, and in some, it is high. Individuals having liabilities on the right-hand side of the threshold (white area) are affected. (*After F. Vogel and A. G. Motulsky, Human Genetics: Problems and Approaches, Springer-Verlag, 1979*)

The multifactorial model poses questions rather than answering them. A meaningful answer can come only from analysis of the contribution of single genes (and specific environmental factors) to such disease liabilities. Some such analyses have been successful. The ABO blood groups, for example, influence liability to many common diseases, such as stomach cancer and cancers of some other organs; variants of the major histocompatibility gene (HLA) contribute to the liability for some rheumatic and autoimmune disease; and alpha$_1$-antitrypsin variants are involved in chronic obstructive pulmonary disease. Here, the risk increases when the carriers of such alleles suffer from repeated bronchial infections—for example, heavy cigarette smokers or laborers working in a dusty environment. This is a good example of a specific interaction between a genetic liability and an environmental stress factor. Such examples are studied in a special branch of medical genetics that is known as ecogenetics.

Heterogeneity for autosomal-recessive mutations, while not leading normally to disease, may contribute to disease liabilities under certain external conditions. Often, specific disease units with clearcut mendelian modes of inheritance have been identified by a combination of clinical, biochemical, and genetic methods in a minority of individuals and families within a bulk of conditions hitherto described as multifactorial. There are good chances that future research will help to open the "black boxes" of multifactorial inheritance step by step by analysis of single genes, the mechanisms of their action, and the ways in which they interact with each other and with specific stress factors in the environment. Friedrich Vogel

Population Genetics

Population genetics is the mathematical basis of evolutionary theory. It is concerned with the frequency of genes and genotypes in a population, their relationship, and how they change over time. A major concern is the frequency of alleles, which are different forms of a gene or DNA sequence. Many principles of population genetics are easily understood using a simple model of a locus with two alleles, A and a. There are three possible genotypes: AA, Aa, and *aa*. A starting point in population genetics is to determine the relative frequency of each allele. It is conventional to use the symbol p to refer to the relative frequency of the *A* allele, and the symbol q to refer to the relative frequency of the *a* allele. If each genotype can be uniquely identified, the allele frequencies can be determined by simple counting. The number of individuals with each genotype are first counted and expressed as a relative frequency. If, for example, there are 18 individuals with genotype *AA*, 24 with genotype *Aa*, and 8 with genotype *aa*, there will be a total of 50 individuals, and the relative frequencies will be $AA = 18/50 = 0.36$, $Aa = 24/50 = 0.48$, and $aa = 8/50 = 0.16$. The frequency of the *A* allele is determined by taking the frequency of genotype *AA* plus half the frequency of genotype *Aa*, giving $p = 0.36 + 0.48/2 = 0.6$. Likewise, the frequency of the *a* allele is determined by taking the frequency of genotype *aa* plus half the frequency of genotype *Aa*, giving $q = 0.16 + 0.48/2 = 0.4$. Note that $p + q = 1$.

Genotype frequencies. Given allele frequencies p and q, it is possible to predict the genotype frequencies in the next generation under different assumptions regarding mating.

Panmixis. Panmixis refers to random mating with respect to genotype. Under this model, any male of reproductive age is equally likely to mate with any female of reproductive age in terms of the particular gene or trait being analyzed. In other words, a male with genotype *AA* is just as likely to mate with a female of genotype *AA* as with a female of genotype *Aa* or *aa*. Under these conditions, the distribution of genotype frequencies in the next generation is a function of probability, known as Hardy-Weinberg equilibrium.

Hardy-Weinberg equilibrium. Given allele frequencies p and q, corresponding to alleles *A* and *a*, the expected genotype frequencies under panmixis in the next generation are $AA = p^2$, $Aa = 2pq$, and $aa = q^2$. Under Hardy-Weinberg equilibrium, the genotype and allele frequencies will remain constant generation after generation. Hardy-Weinberg equilibrium provides a baseline model from which to make predictions about nonrandom mating and evolutionary change.

Nonrandom mating. If individuals are not mating at random with respect to genotype, then there is no panmixis. There are two basic forms of nonrandom mating: inbreeding and assortative mating. Inbreeding is the mating of individuals that are closely related. In humans, this usually means a couple that are more closely related than third cousins are. Inbreeding changes genotype frequencies by increasing the proportion of homozygotes (*AA, aa*) and decreasing the proportion of heterozygotes (*Aa*). Assortative mating refers to nonrandom mating based on phenotype. Positive assortative mating refers to mating between individuals that are phenotypically similar, such as between two tall people or two people with blond hair. Positive assortative mating also increases the frequency of homozygotes. Negative assortative mating occurs when mating is between individuals that are phenotypically different, and will result in a decrease in homozygotes. Overall, nonrandom mating affects the genotype frequencies but does not change the allele frequencies.

Evolutionary forces. Allele frequencies will remain the same, generation after generation, under both panmixis and nonrandom mating. Evolution, defined here as a change in allele frequency over time, can be caused by four evolutionary forces: mutation, natural selection, gene flow, and genetic drift. Nonrandom mating does not lead to allele frequency changes, although it can affect the rate at which allele frequencies change.

Mutation. Mutation is a random change in the genetic code. It can consist of a small point mutation, where a single nucleotide is changed in the DNA, or larger units of DNA may be changed. Mutation rates are often very difficult to estimate, but most seem to be rather low, with probabilities ranging from about 10^{-4} to 10^{-8} per locus per generation. Mutation introduces new genetic variants into a population and is thus required for any evolutionary change. By itself, however, mutation does not cause rapid evolutionary change; the other evolutionary forces act upon this new variation to increase or decrease the frequency of a mutant allele.

Natural selection. Each genotype can be characterized by its relative probability of survival and reproduction, known as fitness (w), and the probability of not surviving or reproducing, known as the selection coefficient (s), such that $w + s = 1$. For example, if the probability of an individual with genotype *aa* surviving and reproducing is one-fourth that of individuals with genotypes *AA* or *aa*, then the fitness of $aa = 0.25$ and the selection coefficient of $aa = 1 - 0.25 = 0.75$. Fitness values depend on characteristics of the specific locus and the local environment. Hardy-Weinberg equilibrium assumes that the fitness values of all genotypes are the same. When this is not the case, then natural selection can cause changes in allele frequency through the process of differential survival. There are several different types of selection, each with different outcomes.

Selection against recessive homozygotes. This occurs when the allele *a* is recessive and the recessive homozygous genotype (*aa*) has a lower fitness relative to the other genotypes (*AA*, *Aa*), which have equal fitness. Under this condition, the frequency of the *a* allele will decline over time to approach a value of zero (although balanced to some extent by continuing mutation from *A* to *a*). The change in allele frequency per generation is a function of the selection coefficient (s) for *aa* and the initial frequency (q) of the *a* allele. The frequency of *a* will decline each generation by the amount $-spq^2/(1 - sq^2)$. When the allele *a* is lethal, as is the case with many genetic diseases, then $s = 1$ (all individuals with *aa* are selected against), and the decrease in *a* each generation is equal to $-q^2/(1 + q)$. The lethal allele will not be eliminated in a single generation because even though all individuals with *aa* are selected against, individuals with genotype *Aa* continue to contribute *a* alleles into the population.

Selection for recessive homozygotes. In this case, the fitness of the recessive homozygote is higher than that of the other two genotypes, and the frequency of the dominant allele will decline over time. If the dominant allele is completely lethal, then it will be eliminated in a single generation. If it is not completely lethal, then it will take time for the dominant allele to be reduced to near zero.

Selection against the heterozygote. In this case, the fitness of the heterozygote (*Aa*) is lower than the fitness of the homozygous genotypes (*AA*, *aa*). Both alleles will be selected against, and the frequency will change in the direction of the initially most common allele. For example, if selection against the heterozygote starts from a condition where $p = 0.7$ and $q = 0.3$, then selection will drive the population to the state of $p = 1.0$ and $q = 0.0$. If, however, p is initially less than q, then the reverse will occur. For example, if selection against the heterozygote starts with $p = 0.2$ and $q = 0.8$, then selection will ultimately result in $p = 0.0$ and $q = 1.0$. In the unlikely event that the initial allele frequencies are exactly equal ($p = q = 0.5$), the allele frequencies will stay the same until another evolutionary force changes them initially.

Balanced polymorphisms and selection for the heterozygote. All of the above examples involve allele frequencies moving toward 1 or 0. In some cases, the ultimate fate of natural selection is a balance between selective forces producing a set of intermediate allele frequencies that remain in equilibrium. Such balanced polymorphisms result when the heterozygote is selected for because the fitness of the heterozygote is higher than that of the two homozygous genotypes. Selection against *AA* results in the elimination of some *A* alleles, while selection against *aa* results in the elimination of some *a* alleles. At the same time, the selection for the heterozygote results in maintaining some *A* alleles and some *a* alleles. Unlike the previous examples of selection, selection for the heterozygote simultaneously selects for and against both alleles. The result is a balance in allele frequencies determined by the fitness values of the homozygotes. When the heterozygote is the most fit, it has, by definition, a relative fitness of 1. Given a fitness of $1 - s$ for genotype *AA* and a fitness of $1 - t$ for genotype *aa*, selection for the heterozygote will reach an equilibrium where $p = t/(s + t)$ and $q = s/(s + t)$. These values are the allele frequencies where overall survival is maximized. The classic example of a balanced polymorphism is human hemoglobin, a locus with a normal hemoglobin allele (*A*) and the sickle-cell allele (*S*). Individuals with the genotype *SS* have the genetic disease sickle-cell anemia, which is frequently fatal early in life. In many populations, this disease leads to selection against the *SS* genotype and the removal of the *S* allele. In environments with epidemic malaria, heterozygous people (*AS*) are the most fit because they have greater resistance to malaria due to the presence of one *S* allele, but do not suffer from sickle-cell anemia. As a result, the *S* allele is maintained at a relatively high frequency (usually 5–20%) in malarial environments.

Gene flow. Gene flow occurs when populations share genes through the process of mating with someone in another population. An allele absent in one population can be introduced from migrants from another population where the allele is present. Over time, gene flow acts to make populations increasingly genetically similar. Gene flow helps keep populations within a single species; for new species to form, gene flow must be eliminated or severely reduced. Further genetic change must occur for a new species to form, since reduction of gene flow is a necessary, though not sufficient, cause of speciation.

Genetic drift. Genetic drift refers to random fluctuations in allele frequency from one generation to the next because of chance. As an example, consider that if you flip a coin 10 times you expect to get, on average, 5 heads and 5 tails. Because of chance, you might actually get 3 heads and 7 tails, 8 heads and 2 tails, or any other possible combination adding up to 10. Reproduction in a population works in the same manner, so that alleles may not be passed along to the next generation in the exact same frequencies as in the parental generation. If p and q are the allele frequencies in the parental generation, the average expectation in the next generation is also p and q, but the variance around these values is equal to $pq/2N$, where N is the number of reproductive adults. When population size is very large, the variance will be small, so that there is little chance for genetic drift in a single generation. When N is relatively small (generally less than 200), the variance is large and there is a greater chance for genetic drift. Genetic drift is not directional; an allele frequency can increase, decrease, or remain the same. The probability of change is related to population size. Over time, all populations drift, and will continue to do so until, by chance, an allele becomes extinct (frequency = 0) or fixed (frequency = 1).

Interaction of evolutionary forces. The four evolutionary forces have been discussed one at a time. In reality, all four forces can operate simultaneously and can interact with each other in different ways. For example, a new mutation can increase or decrease in frequency because of natural selection and/or genetic drift, and it can be transmitted to another population via gene flow. In some cases, a harmful allele can actually increase in frequency because of genetic drift. New neutral mutations are frequently eliminated by genetic drift, but may occasionally increase because of random chance and become fixed in a population. Some evolutionary forces, such as drift and selection, will act to make populations genetically different, whereas gene flow can counter such differences. The interaction of evolutionary forces, and their net effect on allele frequency change, is the focus of many methods and studies of population genetics. John H. Relethford

Biochemical Genetics

Biochemical genetics began with the study of inborn errors of metabolism. These are diseases of the body chemistry in which a small molecule such as a sugar or amino acid accumulates in body fluids because an enzyme responsible for its metabolic breakdown is deficient. This molecular defect is the result of mutation in the gene coding for the enzyme protein. The accumulated molecule, dependent on its nature, is responsible for the causation of a highly specific pattern of disease. There was explosive growth of knowledge in this field with the discovery of many inborn errors of amino acid metabolism when widespread methods of amino acid metabolism followed the recognition that amino acids could be separated chromatographically and that they could be identified and quantified because they became purple when reacted with ninhydrin. More recently the development of gas chromatography–mass spectrometry led to logarithmic growth in the elucidation of organic acidemias (abnormal acidity of the blood) and disorders of fatty acid metabolism.

The field of biochemical genetics expanded during this period with the recognition that similar heritable defective enzymes interfere with the breakdown of very large molecules, such as mucopolysaccharides and the complex lipids that are such prominent components of brain substance. The resultant storage disorders present with extreme alterations in morphology and bony structure and with neurodegenerative disease.

Advances in the methodology of molecular biology have permitted the study and detection of disease at the level of the mutation in the DNA. This capability continues to broaden the scope of biochemical genetics. It has been particularly rewarding in the analysis of mutations in the mitochondrial genome. This technology is elucidating a previously unrecognized spectrum of mitochondrial disorders, in which abnormal energy metabolism is often characterized clinically by lactic acidemia.

Inheritance mechanism. The majority of hereditary disorders of metabolism are inherited in an autosomal recessive fashion. In these families, each parent carries a single mutant gene on one chromosome and a normal gene on the other. Most of these mutations are rare. In populations where consanguinity is common, rare recessive diseases are seen with relative frequency, and affected individuals are homozygous for the same mutation. In populations with more genetic diversity, most affected individuals carry two different mutations in the same gene. Some metabolic diseases are coded for by genes on the X chromosome. Most of these disorders are fully recessive, and so affected individuals are all males, while females carrying the gene are clinically normal. Some genes, such as that for ornithine transcarbamylase (an enzyme of the urea cycle), express as X-linked dominants in which most females are detectable as metabolically abnormal, and some are affected as severely as males. The disorders that result from mutations in the mitochondrial genome are inherited in nonmendelian fashion because mitochondrial DNA is inherited only from the mother. Those that carry a mutation are heteroplasmic; that is, each carries a mixed population of mitochondria, some with the mutation and some without. Expression of clinical disease and its degree of severity are functions of the numbers of abnormal genomes, and they may differ greatly among siblings, because the ovum results from a funneling that may lead to concentration or dilution of the prevalence of the mutation in maternal cells.

Inborn errors of amino acid metabolism. Phenylketonuria (PKU) is a prototypic biochemical genetic disorder. It is an autosomally recessive disorder in which mutations demonstrated in a sizable number of families lead, when present in the genes on both chromosomes, to defective activity of the enzyme that catalyzes the first step in the metabolism of phenylalanine. This results in accumulation of phenylalanine and a recognizable clinical disease whose most prominent feature is severe retardation of mental development. This was the disease with which programs of neonatal screening were developed throughout the world. Determination of the concentration of phenylalanine in a spot of blood obtained from the heel of the baby permits a diagnosis of phenylketonuria early enough to initiate a diet sufficiently restricted in its content of phenylalanine that mental retardation is prevented. This was a wonderful development in public health and preventive medicine. Programs of neonatal screening are expanding throughout the developed world as new methodology permits the detection and treatment of many more inborn errors of metabolism.

The mutations that cause phenylketonuria result in deficient activity of the enzyme phenylalanine hydroxylase, which normally catalyzes the conversion of phenylalanine to tyrosine. Thus, levels of phenylalanine soar and those of tyrosine diminish. When phenylalanine concentrations are high, the amino acid is converted to a number of products including phenylpyruvic acid, a phenylketone. A green color results when ferric chloride is added to solutions containing phenylpyruvic acid, and this once led to the detection of the compound in the urine of two mentally retarded siblings, the index cases of phenylketonuria. Patients with this disease are typically blond and blue-eyed, because the abnormal chemistry interferes with the development of pigment, in addition to affecting brain development. *See* PHENYLKETONURIA.

X-linked metabolic disorders. The most commonly encountered of the inborn errors of purine metabolism is the Lesch-Nyhan disease. The gene is on the end of the long arm of the X chromosome. Very many mutations have been discovered, a distinct one for virtually every affected family, and most mutations lead to a complete absence of enzyme activity. The enzyme is hypoxanthine, guanine phosphoribosyl transferase (HPRT), which catalyzes the conversion of hypoxanthine and guanine to their respective nucleotides, inosinic acid and guanylic acid. Inosinic acid is also converted to guanylic acid and to adenylic acid, and adenylic acid and guanylic acid are converted to adenosine triphosphate and guanosine triphosphate and their deoxynucleotides—the building blocks of the nucleic acids, RNA and DNA. HPRT is called a salvage enzyme because it is used in the reclaiming of purines resulting from the breakdown of cellular DNA and RNA. The salvage pathways contrast with the de novo pathway of purine nucleotide synthesis in which inosinic acid is made in stepwise fashion from small molecules such as carbon dioxide, ammonia, and glycine. When HPRT is deficient, the phosphoribosylpyrophosphate that serves as the ribose phosphate to form inosinic acid and guanylic acid accumulates and drives the de novo pathway into overactivity; the excess purine synthesized ends up as uric acid.

In this way, patients with HPRT deficiency have high concentrations of uric acid in blood and urine, as do patients with gout. Like patients with gout, these patients may develop acute arthritis, kidney stones, and renal failure. They also have a distinct neurologic disease in which retardation of motor development is associated with spasticity like that of patients with cerebral palsy, and involuntary movements and posturing called chroeoathetosis and dystonia. An even more striking feature of this disease is that the patients develop compulsive-aggressive behavior, the most prominent aspect of which is self-injury through biting. Most patients have loss of tissue about the lips, and most bite their fingers, sometimes with partial amputation. The disease promises to be rewarding in providing chemical explanations of behavior, but so far the linkage between the enzyme defect and the behavioral phenotype has been elusive.

Mitochondrial disease. The diseases that result from mutation in mitochondrial DNA have been recognized as such only since the 1990s. They result from point mutations, deletions, and other rearrangements. It seems likely that there are many more entities yet to be discovered. A majority of these disorders express themselves chemically in elevated concentrations of lactic acid in the blood or cerebrospinal fluid. The importance of the processes of energy metabolism to the central nervous system is underscored by the fact that levels in the cerebrospinal fluid are often higher than those in the blood and may be elevated in cerebrospinal fluid when the blood is normal. Many of the disorders are known as mitochondrial myopathies (diseases of muscles) because skeletal myopathy or cardiomyopathy are characteristic features. The histologic hallmark of these mitochondrial myopathies is the presence of ragged red fibers when the tissue is processed with the Gomori stain. These fibers result from the aggregation of mitochondria, which increase in quantity as compensation for diminished function, and which appear abnormal in size and shape when viewed with electron microscopy. Many of these disorders are known by acronyms, such as MERRF, which stands for mitochondrial encephalomyelopathy with ragged red fibers. An example of disease that results from mutation in mitochondrial DNA is mitochondrial encephalomyelopathy, lactic acidemia, and strokelike episodes (MELAS).

MELAS present classically with strokelike episodes. Myopathy may have been present earlier, as elucidated by a careful history, or the patient may have been unaware of it until it is discovered upon physical examination. The episode is clinically a typical stroke with hemiplegia (paralysis on one side of the body) or other neurologic manifestation dependent on the area of brain infarcted. They are called strokelike episodes because there is no demonstrated occlusion of a vessel. Consistent with this, the symptoms are sometimes transient, but they also may leave permanent neurologic evidence. Patients are characteristically short in stature. Levels of lactic acid, while elevated, are often not very high; levels in the blood may be normal, but those in the cerebrospinal fluid are seldom normal. Some patients have convulsions, and ultimately there is evidence of encephalopathy and neurodegeneration. Dementia is a late consequence. Additional features are migraine and noninsulin-dependent diabetes mellitus. The disease is generally recognized by the occurrence of the classic picture in a family member. Once the mutation is identified, it is then usually found in other members of the family, some of whom are asymptomatic, while others may have diabetes or migraine and no other signs of disease. The disease is caused by a point mutation in the mitochondrial gene for the transfer RNA for the amino acid leucine. The mutation generally interferes with protein synthesis in the mitochondria, disrupting the activity of a number of enzymes of the electron transport chain.

Significance. Biochemical genetics has been a rich source of insight into fundamental principles in biology and medicine. The concept that biochemical and clinical abnormality could result from a defect in a single enzymatic step in metabolism was first enunciated by A.E. Garrod around the turn of the century. This was the first statement of the one gene–one enzyme hypothesis in which genes function to determine the structure of proteins, whose structure determines function. These concepts, developed long before the elucidation of the DNA as the genetic material, have stood the test of time. The functions of mitochondrial DNA and the variety of disease its mutation causes represent a recent chapter in the growth of knowledge. William L. Nyhan

Chromosome Mapping

Human chromosome mapping, the localization of human genes to specific chromosomes or regions of chromosomes, has undergone explosive growth since the mid-1970s. Nearly 1500 genes have been assigned to their respective autosomes, well over a hundred genes to the X chromosome, and several functional genes to the Y chromosome. Assuming that the total number of human genes is about 50,000, this represents slightly more than 3% of the total number. There seems no reason in principle why a complete human gene map should not eventually be achieved. The current state of the map is summarized in **Fig. 10**. For each chromosome only a few genes, usually those of some clinical interest, have been named. A finding of general interest which has emerged is that the genetic maps of the great apes are almost identical to those of humans, and

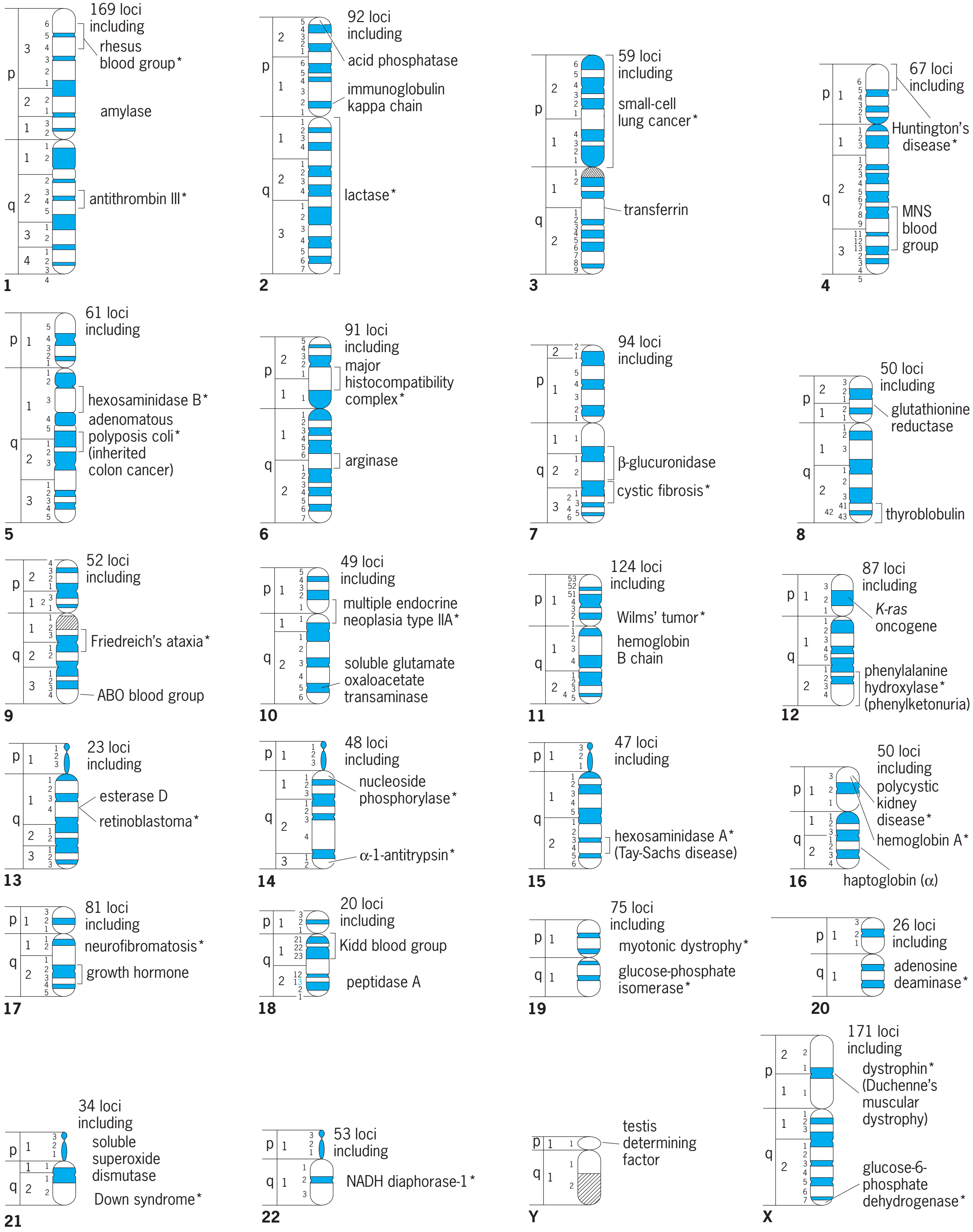


Fig. 10. Diagrammatic representation of the human chromosomes as observed by banding techniques. The number of genes already mapped to each chromosome is indicated beside it, and some examples of particular gene loci are shown. The exact position of these loci is marked where this is known. An asterisk indicates that the locus so marked is the site of one or more mutations associated with inherited disease or is of major clinical importance, such as ABO blood groups.

many groups of genes which are syntenic (on the same chromosome) in humans are syntenic even in the mouse. *See* PROTEINS, EVOLUTION OF.

Most of the information about localization of human genes has been obtained either by somatic cell hybrids or by family studies. Advances in DNA technology have greatly increased the potential of both methods, and have also allowed the development of in situ hybridization, which now accounts for at least one-third of new gene assignments.

Somatic cell hybrids. If human and rodent somatic cells are grown together in tissue culture and treated with certain viruses or chemicals, it is possible to select hybrid uninucleate cells which have resulted from the fusion of human with rodent cells. The most commonly used combinations are human and mouse, or human and Chinese hamster. In both these types of hybrids, on extended subculture, some of the human chromosomes are usually lost preferentially. It is therefore possible to obtain a set of clonal (derived from a single cell) hybrid cell lines, each containing a full set of rodent chromosomes together with a unique subset of human chromosomes.

Any gene whose product is distinguishable in human and rodent and which is expressed in cultured cells can be assigned to a particular chromosome by testing a relatively small number of hybrids for the presence or absence of a given human gene product. This can then be correlated with the presence or absence of some other human gene product or chromosome. The mouse and human gene products, most of which are proteins, can often be separated by electrophoresis. This relies on the fact that the homologous proteins in the two species are not identical but carry slightly different charges owing to differences in their amino acid sequences. **Figure 11** shows an example of electrophoresis of adenylate kinase in human and hamster controls and in six independent human-hamster hybrids. This experiment shows that the two major forms of adenylate kinase on the human cells are coded by different genes on different chromosomes. By chromosome analysis of a number of independently obtained hybrids, one can correlate the synthesis by the hybrid cells of either or both forms of the enzyme with the presence of two particular human chromosomes: actually, one gene (AK_1) is on chromosome 9, the other (AK_2) on chromosome 1. Many gene products distinguished in other ways, such as cell surface antigens, have also been mapped by using such hybrids.

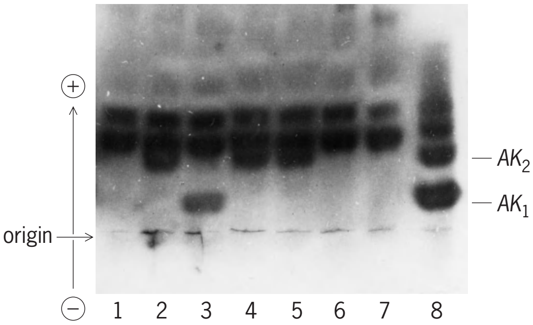


Fig. 11. Photograph of adenylate kinase isozymes separated by starch gel electrophoresis in human control (channel 8), hamster control (channel 7), and six independent hybrid clones, showing independent segregation of human AK_1 (positive in the hybrid in channel 3) and AK_2 (positive in the hybrids in channels 2, 4, and 5).

Originally the only genes that could be mapped in hybrids were those whose products could be detected in these cells. These would not, for example, include genes such as those coding for hemoglobin or insulin which are only produced in specialized cell types, although presumably the genes themselves are present in all cells. But it is now possible to isolate many of the genes themselves. The DNA sequence can then be labeled, usually with radioactivity, and used as a probe to search for similar sequences in the hybrids. The presence of the human sequence is then correlated with that of a particular chromosome in exactly the same way as described above for gene products.

This approach is not confined to DNA sequences of known function, and more than 3300 random DNA fragments have also been assigned to particular chromosomes. Although these are of less intrinsic interest they are proving invaluable as genetic markers in family studies. If random fragments from some particular chromosome are needed, these can be generated either from a hybrid containing the appropriate single human chromosome, or by direct sorting of the human chromosomes using a fluorescent activated cell sorter. In both cases it is necessary to check with a hybrid panel that the fragments obtained are indeed from the right chromosome.

The use of human parental cells containing rearranged chromosomes such as translocations or deletions, or the observation of spontaneous or induced chromosome rearrangements arising in hybrid cells, also allows the assignment of genes to particular regions of chromosomes. *See* SOMATIC CELL GENETICS.

In situ hybridization. The chromosomes of normal human lymphocytes can be readily visualized during the metaphase stage of cell division. In the case of genes involving localized repeated sequences, a DNA probe radioactively labeled to high activity can be used directly for hybridization to specific chromosomes in metaphase spreads, which are then autoradiographed. It has also become possible to label probes with sufficient specific activity to detect sequences present only as single copies (such as most of those coding for proteins), and thus, in one experiment, to localize the gene to a particular region of a particular chromosome. Developments in nonradioactive labeling of DNA with fluorescent dyes and the use of a confocal laser scanning microscope have greatly improved the precision of gene localization by this technique.

Family studies. Unlike somatic cell hybrids, gene mapping by family studies depends on finding genetic differences between people and observing the way these differences are inherited. A special case is the X chromosome because most genes on the X chromosome are inherited in a distinctive way—they never pass from father to son. It is therefore relatively easy to establish that a particular disease or distinctive trait is X-linked; for a long time, many such diseases have been recognized. A well-known example is the hemophilia gene that was carried by Queen Victoria and caused the disease in many of the royal families of Europe. The assignment of genes to particular autosomes by family studies presents a more formidable problem and it was only in 1969 that the first confident assignment, that of the Duffy blood group to chromosome 1, was made.

For many years, workers have analyzed the patterns of inheritance of two or more gene loci in the same family to establish whether the genes are linked, that is, carried on the same chromosome pair close enough together to show nonindependent segregation. If the genes are linked, a measure of the intergene distance can be derived by the frequency with which they are separated by crossing over. A linkage group found in this way was that of the ABO blood groups with one of the forms of adenylate kinase. The assignment of the adenylate kinase gene to chromosome 9 by somatic cell hybrids thus allowed the indirect assignment of the gene determining the ABO blood group to the same chromosome. *See* CROSSING-OVER (GENETICS).

Direct assignment of genes to chromosomes by using family studies can be accomplished by studying concurrently the inheritance of genetically determined traits and of chromosome polymorphisms or rearrangements. Individuals carrying unbalanced chromosome abnormalities involving deletions or duplications of specific regions of chromosomes can also be useful for mapping by gene dosage. However, quantitative differences indicating one, two, or three copies of a gene are difficult to distinguish from secondary effects, and from unusual normal alleles which code for enzymes of different activity. This method has been most successful in regional localization when the gene is already known to lie on a particular abnormal chromosome.

The efficiency of family studies in assigning genes to chromosomes depends very much on the proportion of the human genome which is within measurable genetic distance of “good” genetic markers. In this context a good genetic marker might be defined as one determined by a single gene locus, easily typed and showing a large amount of normal variation, so that a high proportion of individuals are heterozygous. Before the advent of DNA technology, the genetic markers available have been limited and in practice have been confined to those expressed in readily obtained tissues such as blood. Thus the chance of assigning any new gene to its chromosome even from study of a large number of families was small, one estimate being about 8%. Even in this early work, however, family studies had an essential role in establishing the fine details of gene arrangement, and in the construction of a true genetic map where the known amount of recombination between two genes gives a definite value or map distance. In most chromosome regions it has been found that recombination, and hence map distance, between genes is almost twice as great in females as in males.

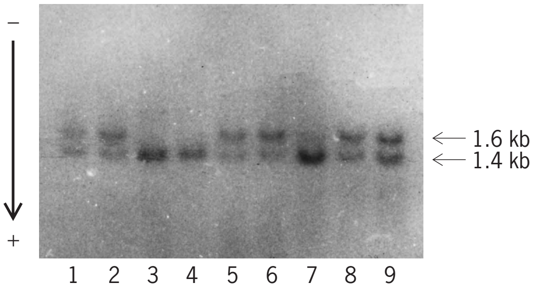


Fig. 12. **Autoradiograph showing polymorphism defined by a random DNA fragment from human chromosome 7 and the restriction enzyme *HinfI*. Three individuals (tracks 3, 4, 7) are homozygous for the common allele producing a fragment 1400 bases long (1.4 kb). The remaining samples are from heterozygotes in which one chromosome produces the common pattern and the other a longer fragment (1.6 kb). (*Photograph courtesy of Dr. Ben Carritt*)**

In the 1980s there was a dramatic increase in the number of genetic markers available, allowing the construction of at least partial genetic maps of all the human chromosomes. As a result, most of the human genome is now within reach of known genetic markers. This has come about through the use of DNA probes in conjunction with bacterial enzymes known as restriction endonucleases. These enzymes recognize certain specified short sequences of DNA, usually 4 or 5 bases long, and then cut the DNA at all the places where this sequence occurs. Many different enzymes can be extracted from different bacteria, each recognizing a different sequence. Human DNA from different individuals is treated with an appropriate enzyme which breaks it into many fragments. These are then separated by size on an agarose gel and the fragments containing sequences that are complementary to the DNA of the radioactively labeled probe can be identified (**Fig. 12**). The number and size of the fragments depend on the distribution of sites for that particular restriction enzyme in the region of the genome to which the probe hybridizes, and the sizes commonly detected are between 500 bases and 15,000 bases (15 kilobases, or kb). In this way variation in the base sequences can be detected not only in the region of the gene which codes for the protein but also in the large regions of noncoding DNA which form intervening and flanking sequences. Variants that are detected in this way (restriction fragment length polymorphism) behave as single codominant alleles, and it appears that for almost any DNA probe (whether a random fragment or a known gene) such variation can be found if a sufficient number of restriction enzymes are tried. *See* RESTRICTION ENZYME.

Another very useful type of genetic marker usually found in noncoding regions of DNA is the VNTR (variable number of tandem repeats). In this case, sequences recognized by one probe occur at the same place but in variable numbers on the chromosomes of different individuals. If the DNA is digested with any enzyme that does not cut within the repeat, the size of the fragment recognized by probing after agarose gel electrophoresis will vary between individuals and between the two homologous chromosomes of the same individual chromosomes. At some of these hypervariable loci, 99% of individuals are heterozygous, so every family is informative for linkage analysis with that marker.

Family studies for mapping disease genes. The frequency of the different polymorphisms described above means that virtually the whole genome is within reach of good genetic markers. Family studies on diseases with clear mendelian inheritance, but in which the basic biochemical defect is unknown, have become a very practical undertaking. Such studies have allowed the localization of the genes that cause many autosomal-dominant disorders such as Huntington's disease and adult-onset polycystic kidney disease. It is still much more complex to localize a gene that causes a disease inherited in an autosomal-recessive manner if the heterozygotes cannot be identified, but in the case of cystic fibrosis, a common disease in populations of European origin, this has been accomplished (Fig. 10).

There are several ways in which localizing such genes may help in understanding a disease, in some cases by demonstrating that more than one gene is involved and in others by providing a starting point for reverse genetics. Once the approximate position of a disease gene has been found, it is possible to generate more genetic markers in the same region of the chromosome and to "walk" or "jump" along the chromosome to find the actual disease gene and the mutations that cause the disease, allowing more certain diagnosis and better genetic counseling. A study of the protein coded for by the gene in question should then offer a deeper understanding of the disease and how it may be treated. Understanding a disease by reverse genetics was first successful for Duchenne muscular dystrophy and chronic granulomatous disease, both of which are inherited on the X chromosome. The most dramatic success of this approach has been the identification of the gene that is defective in cystic fibrosis.

A few diseases are inherited only through the mother, but are equally severe in males and females. (This is in contrast to X-linked inheritance, in which genes can pass from father to daughter but not from father to son.) Some of these diseases, including Leber's optic atrophy, are due to mutations of the mitochondria. The mitochondrion is a small circular piece of DNA, only about 16 kb long, that is found in many copies in each cell. The whole DNA sequence of human mitochondrial DNA is known and all the genes have been identified, and so in some ways it can be regarded as a very small chromosome.

Ongoing research. Several techniques should greatly speed up progress in human gene mapping. One is the polymerase chain reaction, which allows the selective amplification of a small region of DNA of interest from a single drop of blood—or even a single cell—without complex purification. One or more particular genes can then be examined quickly and easily in very large numbers of people or in many individual sperms from a single informative individual. The latter technique is the genetic equivalent of sampling hundreds of children from the same father. Another approach, which can be used in conjunction with polymerase chain reaction, is the direct preparation of DNA from a small piece, such as a single band, of a chromosome dissected from a metaphase spread. Therefore the chromosomal origin of clones that are derived from such pieces is precisely known.

Another exciting field of study is that of oncogenes. These are sequences similar to those of ceres, and in some circumstances these DNA sequences can cause changes in mouse cells similar to those seen in cancer cells. Although their relationship to human cancer is still not clear, the coincidence of the map positions of some of these oncogenes and the breakpoints of characteristic chromosome rearrangements seen in some tumors is under investigation. *See* ONCOGENES.

It is important to realize the different scale of mapping that can be achieved by the different approaches. Although the banding patterns shown in Fig. 1 are those usually obtained during the metaphase stage of cell division, higher resolution giving at least 800 bands for the whole genome can be obtained by examining the chromosomes earlier, during prophase, when they are much longer. This has led to a more precise definition of breakpoints and of map positions, although the smallest visible band still represents approximately 1 million base pairs, and two genes that appear to be quite close together in family studies may be several million base pairs apart. Therefore, there is a considerable difference in scale between molecular mapping (for example, of the various defects in beta-hemoglobin) and the physical and genetic chromosome mapping described here. This gap is bridged by a technique known as pulsed-field gel electrophoresis in which so-called rare cutting restriction enzymes are used to generate very large fragments of DNA (up to 1 million base pairs long), which are separated in gels by applying alternating cycles of electric fields in different directions. Some of these fragments are found to carry more than one gene, and so it gives an exact measure of the distance between them. Eventually, overlapping fragments of DNA spanning all the human chromosomes should be identified and most of it should be sequenced. The human genetic map will then be essentially complete and should lead to better diagnosis, genetic counseling, and, eventually, treatment of inherited diseases. The human genetic map should also allow a greater understanding of the genetic changes in cells that underlie other diseases, especially cancers. *See* GENETIC MAPPING; GENETICS.

Sue Povey

Bibliography. K. E. Davies (ed.), *Genome Analysis: A Practical Approach*, 1988; D. S. Falconer, *Introduction to Quantitative Genetics*, 3d ed., 1986; F. C. Fraser and J. J. Nora, *Genetics of Man*, 2d ed., 1986; H. Harris, *The Principles of Human Biochemical Genetics*, 3d ed., 1981; D. L. Hartl and A. G. Clark, *Principles of Population Genetics*, 3d ed., 1997; *Human Gene Mapping 10*, 1989; A. J. Jeffreys, V. Wilson, and S. L. Thein, Hypervariable "minisatellite" regions in human DNA, *Nature*, 314:67–73, 1985; B. S. Kerem et al., Identification of the cystic fibrosis gene: Genetic analysis, *Science*, 245:1073–1080, 1989; M. Levitan, *Textbook of Human Genetics*, 3d ed., 1988; V. A. McKusick, *Mendelian Inheritance in Man*, 8th ed., 1988; *Mapping Our Genes, The Genome Projects: How Big, How Fast*, 1988; J. J. Nora, *Medical Genetics: Principles and Practices*, 4th ed., 1993; W. L. Nyhan and P. T. Ozand, *Atlas of Metabolic Diseases*, 1998; R. R. Race and R. Sanger, *Blood Groups in Man*, 6th ed., 1975; J. H. Relethford, *The Human Species: An Introduction to Biological Anthropology*, 4th ed., 2000; J. R. Riordan et al., Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA, *Science*, 245:1066–1073, 1989; C. R. Scriver et al., *The Metabolic and Molecular Basis of Inherited Disease*, 1995; T. B. Shows et al., Mapping the human genome, cloned genes, DNA polymorphisms, and inherited disease, *Adv. Hum. Genet.*, 12:341–452, 1982; H. E. Sutton, *An Introduction to Human Genetics*, 4th ed., 1988; F. Vogel, Clinical consequences of heterozygosity for autosomal-recessive diseases, *Clin. Genet.*, 25:381–415, 1984; F. Vogel and A. G. Motulsky, *Human Genetics*, 2d rev. ed., 1986.

Human genome

On April 14, 2003, the International Human Genome Sequencing Consortium, led in the United States by the National Human Genome Research Institute and the Department of Energy, announced the successful completion of the Human Genome Project (HGP) more than 2 years ahead of schedule. The DNA sequence produced by the HGP covered about 99% of the human genome's gene-containing regions and was sequenced to an accuracy of greater than 99.99%. In addition, to help researchers better understand the meaning of the human genetic instruction book, the project had taken on a wide range of other goals, from sequencing the genomes of model organisms to developing new technologies to study whole genomes. All of those goals were met or surpassed (see **table**).

Scientific strategy. The scientific strategies that were employed evolved over the years, but the basic concept for the project that was initially proposed by the National Research Council (NRC) of the U.S. National Academy of Sciences in the mid-1980s turned out to be effective. Because the human genome is so big (human DNA consists of about 3 billion nucleotides connected end to end in a linear array; **Fig. 1**), it was necessary to break the task down into manageable chunks. The first step was to create a genetic map of the whole genome. Such a map is generated by analyzing how frequently markers are inherited together in families. Markers that are in proximity are inherited together more frequently than those that are far apart. The resulting map shows a series of signposts along the DNA (**Fig. 2**).

The second step was to create a physical map of the DNA. The DNA was broken into small pieces that could be cloned and replicated in bacteria to generate enough material for study. The pieces were then fitted together by studying how they overlapped with each other. The genetic markers, as well as other types of landmarks, were used for this purpose. Eventually the whole genome was covered with overlapping clones.

The final step was to sequence the clones and piece together the entire genome sequence. Before this could be done, much research and development was needed to make sequencing more efficient and less costly. With a major investment in research, the technology did improve substantially and the costs came down dramatically. The improvements included streamlined methods, the use of robots

Comparison of the first two human chromosomes to be sequenced

Category	Explanation	Chromosome 21	Chromosome 22
Base pairs sequenced	Only the long arms of the chromosomes, representing the euchromatic regions, were sequenced. The heterochromatin cannot be cloned with current technology.	33.5×10^6	33.4×10^6
Number of gaps	The gaps contain sequences that cannot be cloned in any known cloning vector.	3	10
Number of known genes	Genes for which there is evidence that they are expressed in humans, or which are very similar to known genes in model organisms.	127	397
Number of predicted genes	Genes predicted by computer analysis, but unlike any other known genes.	98	148
Number of pseudogenes	Genes that cannot function because they lack a key feature. They are probably evolutionary remnants.	59	134
Percent of repeats	DNA sequences that are repeated over and over throughout the genome.	40	42
Percent of Alu repeats	Two common kinds of repeats.	9.5	16.8
Percent of LINE1 repeats	Two common kinds of repeats.	12.9	9.7
Percent of G + C bases	DNA consists of four bases, abbreviated A, T, G, and C. The frequency of the bases varies between organisms and also in different parts of the genome. Since G always pairs with C and A with T, the composition of particular DNA regions is expressed as % G + C.	41	48

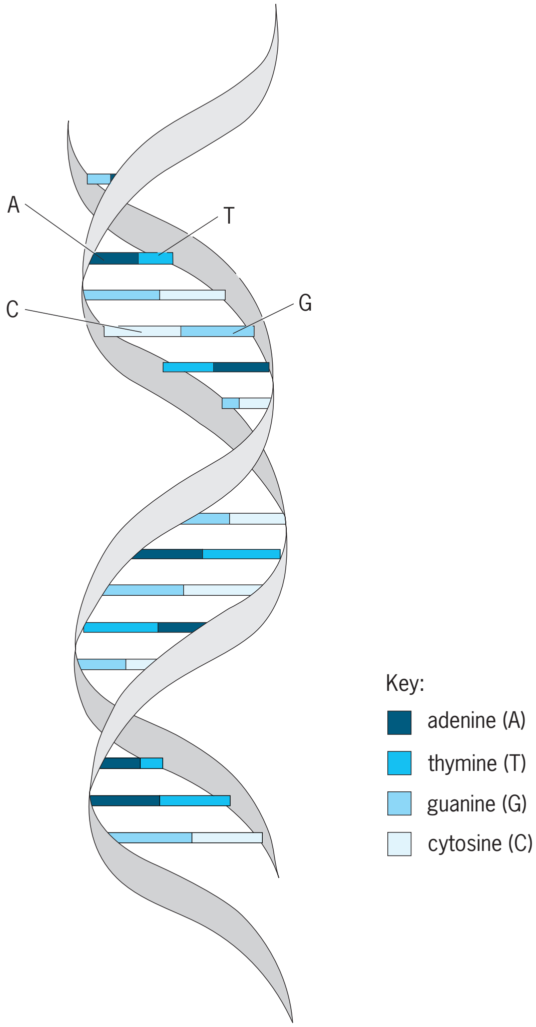


Fig. 1. Structure of DNA. DNA is a long thin molecule made up of chemical units called nucleotide bases. The four bases are adenine, thymine, guanine, and cytosine (A, T, G, and C). They are arranged in two parallel strands that are complementary to each other. An A always appears opposite a T, and a G always appears opposite a C. Sequencing the DNA involves determining the order of the bases in the molecule.

to handle large numbers of samples simultaneously, more accurate and faster sequencing machines, and better computer programs for tracking and assembling the data.

Model organisms. An important element of the overall strategy was to include the study of model organisms in the HGP. There were two reasons for this: (1) Simpler organisms provide good practice material. (2) Comparisons between model organisms and humans yield very valuable scientific information. All life forms have much in common, including some of their DNA sequences. Without the information that can be gained from the study of model organisms, it would be very difficult to know what the human DNA sequence means.

The HGP initially adopted five model organisms to have their DNA sequenced: the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the roundworm *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster*, and the laboratory mouse *Mus musculus*. Subsequently, the HGP supported the sequencing of additional organisms, including the rat *Rattus norvegicus*, the roundworm *Caenorhabditis briggsae*, and the fruitfly *Drosophila pseudoobscura*.

In December 2002, a study comparing the human genome sequence with the mouse genome sequence found that 99% of human genes have a counterpart in mice, even though the two mammals diverged from a common ancestor more than 75 million years ago. In addition to sharing most gene functions with humans, the mouse is a valuable model for biomedical research because it can be selectively bred and other experiments can be conducted that are not possible on humans.

Project evaluation. Originally, the proponents of the HGP suggested that a 15-year time frame was appropriate for obtaining the human genome sequence. In order to make sure that this schedule was followed, 5-year plans were developed to set intermediate milestones. Remarkably, all the milestones were met, many times well before the planned deadlines. The sequence of yeast was released in April 1996, the sequence of *E. coli* was published in September 1997, *C. elegans* was completed in December 1998, the *Drosophila* sequence was released in early 2000, a first draft of the rat sequence was released in November 2002, and the mouse sequence was published in December 2002.

A first draft of the entire human sequence was released in June 2000 with publications describing it following in February 2001. A high-quality, comprehensive version of the human genome sequence was completed in April 2003, more than 2 years quicker than the 15-year period and coinciding with the 50th anniversary of the discovery of the double-helical structure of DNA. From the outset, sequence information from the HGP was immediately and freely distributed to scientists around the world, with no restrictions on its use or redistribution. In parallel with the publicly funded effort, a private United States company called Celera also generated a draft of the human sequence, which was completed about the same time as the HGP's first draft, though not made freely accessible. The Celera strategy used a different approach. Instead of mapping the cloned DNA pieces first and sequencing them later, random pieces were sequenced directly and subsequently correlated with maps. Some aspects of this approach were incorporated into the strategy for sequencing additional organisms, such as the mouse and the rat.

Findings. How many genes are there is probably the most common question regarding the human genome. Although a definitive count of human genes must await further experimental and computational analysis, the HGP's initial analysis of the human genome in 2001 led to an estimate that humans have only about 30,000 genes. That was quite a surprise because previous estimates were 80,000 to 100,000 genes. The lower gene count does not necessarily mean that the human genome is less complex, because many genes can produce more than one protein by alternate splicing of their exons (protein-encoding regions of the gene) during translation into the constituents of proteins.

Another fascinating feature of the human genome sequence, as well as the genome sequences of other mammals, is the distribution of genes on the chromosomes. It turns out that mammalian chromosomes have areas with many genes in proximity to one another, but these regions are interspersed with vast expanses of DNA devoid of protein-coding genes. This uneven distribution of genes stands in marked contrast to the more uniform distribution of genes throughout the genomes of many other organisms, such as the roundworm and the fruitfly.

In their analysis, HGP scientists also found that about 50% of the DNA in the human genome is made of repetitive sequences, which is much greater than the percentage of repetitive DNA found in the roundworm (7%), the fruitfly (3%), and a variety of other nonmammalian species. These repeated sequence elements provide a rich record of clues to the evolutionary past of humans. It is possible to date various types of these repeats to the times when they appeared in the evolutionary process and to follow their fates in different regions of the genome and in different species. Because repeated sequences are more likely to misalign when DNA is being copied or repaired, areas with a high concentration of repeats are more prone to mutation than other regions of the genome. Consequently, the various types of repeats have helped to reshape the genome in a multitude of ways by rearranging and modifying existing genes, as well as by creating entirely new genes.

Such repeats and the relatively predictable patterns of DNA shrinkage and growth associated with them are also being used as “DNA dating” tools to explore evolution in a rapidly emerging area of research known as phylogenetics. The pace of evolutionary processes, sometimes referred to as a molecular clock, appears to vary among different types of repeats. However, researchers in some instances have been able to select a specific type of repeat, examine its prevalence and length among the genomes of a wide range of species, and then use that genetic information to construct a family, or phylogenetic, tree of various species showing when in evolution that repetitive element was “born” and how it “moved” among species during the course of evolution.

For example, two types of repeats, called DNA transposons and long terminal repeat (LTR) transposons, present in the genome of the mouse are much rarer in the human genome. Consequently, a phylogenetic tree built using data on DNA transposons and LTR transposons would branch in a manner that reflects how the prevalence of these repeat elements has changed since rodents and primates diverged from a common ancestor an estimated 75 million years ago.

Future research. The availability of large amounts of DNA sequence information as well as other genomic resources has profoundly affected biomedical research and thinking. Information about whole genomes that was previously gained one gene at a time has now been obtained. With the complete sets of genes of organisms available, how genes are turned on and off and how genes interact with each other can be studied. The possibilities are endlessly exciting, and the demand for more sequence is increasing all the time. Many more organisms will be sequenced. But having the sequence is not enough. What the different genes do and how they affect human health must also be learned.

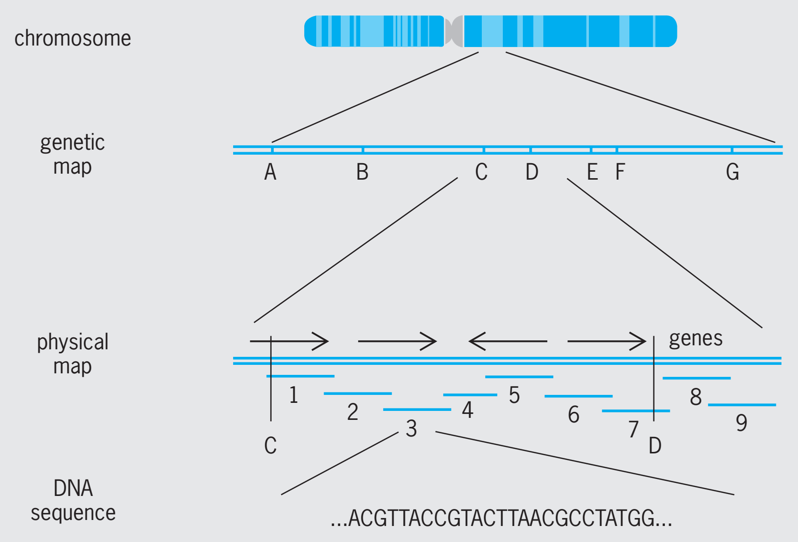


Fig. 2. Steps in analyzing a genome. (1) Markers are placed on the chromosomes by genetic mapping, that is, observing how the markers are inherited in families. (2) A physical map is created from overlapping cloned pieces of the DNA. (3) The sequence of each piece is determined, and the sequences are lined up by computer until a continuous sequence along the whole chromosome is obtained. Steps 2 and 3 can be reversed or done in parallel. As the pieces are sequenced, the sequences at the overlapping ends can be used to help order the pieces. If the sequencing is done before the pieces are mapped, the process is called whole-genome shotgun sequencing.

Among the most important large-scale opportunities for scientists in the genome era are the development of innovative technologies and the creation of informational frameworks to make the results of genomic research applicable to individual health. Although all humans are 99.9% identical in their genetic makeup, the 0.1% variation in DNA sequences is thought to hold key clues to individual differences in susceptibility to disease.

The International HapMap Project was launched in October 2002 with the goal of building a catalog of human genetic variations and determining how these variations are organized into neighborhoods, or haplotypes, along the human chromosomes. The HapMap will serve as a tool of researchers trying to discover the common genetic variations associated with complex diseases, as well as variations responsible for differences in drug response.

In March 2003, scientists also set out to develop efficient ways of identifying and precisely locating all of the functional elements contained in the human DNA sequence. The ultimate goal of the Encyclopedia of DNA Elements (ENCODE) project is to create a reference work that will help scientists mine and fully utilize the human sequence, gain a deeper understanding of human biology, and develop new strategies for the prevention and treatment of disease. In the first phase of ENCODE, researchers will work cooperatively to develop high-throughput methods for rigorously analyzing a defined set of DNA target regions constituting approximately 1% of the human genome. It is hoped this pilot project will pave the way for scaling up this effort to characterize efficiently and effectively all of the protein-coding genes, nonprotein coding genes, and other sequence-based functional elements contained in human DNA. As has been the case with the public effort to sequence the human genome, data from the ENCODE project will be collected and stored in a database that will be freely available to the entire scientific community.

Other frontiers include the establishment of publicly available libraries of chemical compounds for use by basic scientists in their efforts to chart biological pathways, as well as genomic initiatives to understand the life processes of single-cell organisms, or microbes, with the ultimate aim of using the capabilities of these organisms to address needs relating to health, energy, and the environment.

Many challenges lie ahead. Getting the full sequence of human DNA is a dramatic achievement. But what was viewed as an end point has turned out to be just a beginning: a new era of genomic biology lies ahead. *See* DEOXYRIBONUCLEIC ACID (DNA); GENE; GENETIC CODE; GENETIC ENGINEERING; HUMAN GENETICS; MOLECULAR BIOLOGY; NUCLEIC ACID.

F. Collins; E. Jordan

Bibliography. M. D. Adams et al., The genome sequence of *Drosophila melanogaster, Science*, 287: 2185–2195, 2000 (also related articles, 287:2181–2184, 2196–2224); C. elegans Sequencing Consortium, Genome sequence of the nematode *C. elegans*: A platform for investigating biology, *Science*, 282:2012–2018, 1998 (also related articles, 282:2018–2046); F. S. Collins, Medical and societal consequences of the Human Genome Project, *N. Engl. J. Med.*, 341:28–37, 1999; F. S. Collins et al., A vision for the future of genomics research: A blueprint for the genomic era, *Nature*, 422:835–847, 2003; F. S. Collins, M. Morgan, and A. Patrinos, The Human Genome Project: Lessons from large-scale biology, *Science*, 300:286–290, 2003; A. Goffeau, R. Aert, and M. L. Agostini-Carbone, The Yeast Genome Directory, *Nature*, supplement to vol. 387, May 29, 1997; E. D. Green, The Human Genome Project, in *Metabolic and Molecular Bases of Inherited Disease*, 8th ed., 2001; International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature*, 409:860–921, 2001; Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome, *Nature*, 420:520–562, 2002; National Research Council, *Mapping and Sequencing the Human Genome*, 1988; J. C. Venter et al., The sequence of the human genome, *Science*, 291:1304–1351, 2001.

Human-machine systems

Complex systems that comprise both humans and machines. Human-machine systems engineering is the analysis, modeling, and design of such systems. It is distinguished from the more general field of human factors and from the related fields of human-computer interaction, engineering psychology, and sociotechnical systems theory in three general ways. First, human-machine systems engineering focuses on large, complex, dynamic control systems that often are partially automated (such as flying an airplane, monitoring a nuclear power plant, or supervising a flexible manufacturing system). Second, human-machine systems engineers build quantitative or computational models of the human-machine interaction as tools for analysis and frameworks for design. Finally, human-machine systems engineers study human problem-solving in naturalistic settings or in high-fidelity simulation environments. *See* HUMAN-COMPUTER INTERACTION; HUMAN-FACTORS ENGINEERING.

Thus, human-machine systems engineering focuses on the unique challenges associated with designing joint technological and human systems. Historically it has grown out of work on cybernetics, control engineering, information and communication theory, and engineering psychology. Subsequently, researchers who focus on cognitive human-machine systems (in which human work is primarily cognitive rather than manual) have also referred to their specialization as cognitive engineering or cognitive systems engineering. *See* CYBERNETICS; INFORMATION THEORY.

The four major aspects of human-machine systems, in roughly historical order, are systems in which the human acts as a manual controller, systems in which the human acts as a supervisory controller, human interaction with artificial-intelligence systems, and human teams in complex systems. This general progression is related to advances in computer and automation technology. With the increasing sophistication and complexity of such technology, the human role has shifted from direct manual control to supervisory control of physical processes, to supervision of intelligent systems, and finally, with an increasing emphasis on the social and organizational aspects of complex systems, to teamwork in complex environments.

Aviation is an example of a human-machine system in which all of these developments have occurred. Early work in aircraft systems focused on manual control models of pilot performance. With increasing levels of automation, the pilot shifted to a more supervisory role in which tasks such as planning and programming the flight management computer became the predominant form of work. *See* AIRCRAFT INSTRUMENTATION; AUTOPILOT; FLIGHT CONTROLS.

Manual control. Manual control means that the individual exerts direct control on a dynamic system. Vehicle control tasks such as driving a car and flying an airplane, process control in chemical plants, and ship handling are examples of human work that have been modeled as manual control tasks, and in particular, often as closed-loop negative-feedback systems in which the system is continually trying to reduce error (defined as the difference between its goal and its current state) by getting feedback about its current state and then making control actions to compensate for this error. Such human-in-the-loop control involves a display, a human, a controlled system, and a goal (**Fig. 1**). Driving a car may be used as an example of manual control. In this case, the display is the view through the windshield, the human is the driver, the controlled system is the car, and the goal is to stay on the road. The process of driving successfully on the road can then be modeled as a closed-loop negative-feedback system in which the driver is continually getting information via the windshield (display) about the error, which is the difference between the way that the road curves (target state of the controlled system, the input) and the car's current position, velocity, and trajectory (actual state of the controlled system). The driver exerts a change in control, for example by turning the steering wheel. In response to this control input, as well as to any external disturbances such as wind gusts, the car's direction or output changes. This new system output is now fed back to the driver through the windshield, and the cycle continues. The ratio of output to input is called the transfer function, which describes the system dynamics mathematically in terms of a number of concepts such as system order, time lag, and gain. *See* CONTROL SYSTEMS; PROCESS CONTROL.

Classical control theory focuses on system stability in the frequency domain. The original crossover model is based on classical control theory, considers the human-machine system as the invariant, and describes how human manual controllers adjust gain to achieve stability. It provides an elegant mathematical characterization of human manual control: The human exerts control such that the combined human-machine system transfer function behaves like a first-order system with a time delay. Modern or optimal control theory focuses on prediction in the time domain and considers the human controller as a model-based controller that includes a Kalman filter, a model-based state estimator to keep running estimates of state variables, and an optimal controller to make control actions on the basis of predictions derived from these estimates (**Fig. 2**). Thus, the Kalman filter might be viewed as the mental model of the system. More elaborate versions of classical and optimal control models include components such as neuromuscular delays and additional predictors. *See* CONTROL SYSTEM STABILITY; ESTIMATION THEORY; OPTIMAL CONTROL THEORY.

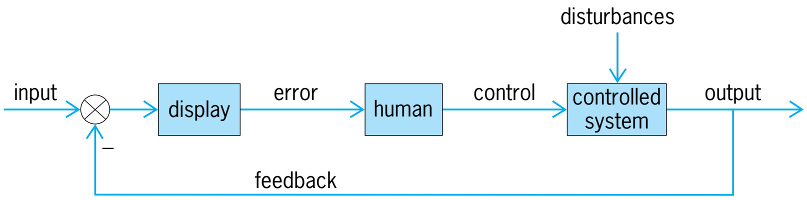


Fig. 1. Simple closed-loop negative-feedback system as a model of manual control.

Work based on control theory looks at the effects of different system dynamics on human performance in manual control (for example, varying gain, time lag, and system order) and different kinds of displays to provide better feedback for manual control. In particular, the use of pursuit, compensatory, and predictor displays is considered in manual tracking tasks (for example, in which the person seeks to minimize the error between a moving target object and the cursor on a computer display). Pursuit displays explicitly show the target and the cursor; compensatory displays show only the difference between the target and cursor. Predictive displays show both the current state of the system and a model-based estimate of its predicted future state. Quickening is a display technique in which higher-order system dynamics are taken into account when generating a predictive display. Teleoperation or remote manipulation, the real-time control of remotely located machines that act as eyes and hands of a person located elsewhere, has been used in undersea and lunar exploration and in microsurgery. *See* MICROMANIPULATION; REMOTE MANIPULATORS; UNDERWATER VEHICLES.

Supervisory control. With the advance of computer and automation technology, in many systems the human no longer acts as a direct manual controller but as a supervisor. This is seen in systems such as nuclear power plants and process-control plants, and is even true in some advanced aircraft systems in which the pilot programs the flight management computer rather than directly flying the airplane. In such systems, human supervisory controllers monitor and only intermittently control a highly automated process. Typical supervisory control activities include equipment configuration (for example, startup and shutdown of the plant), monitoring and fine-tuning system parameters, planning and scheduling control activities, predicting system state, and

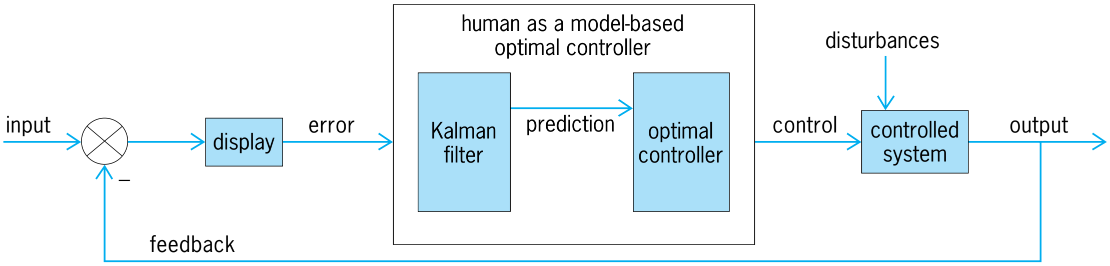


Fig. 2. Modern or optimal control paradigm.

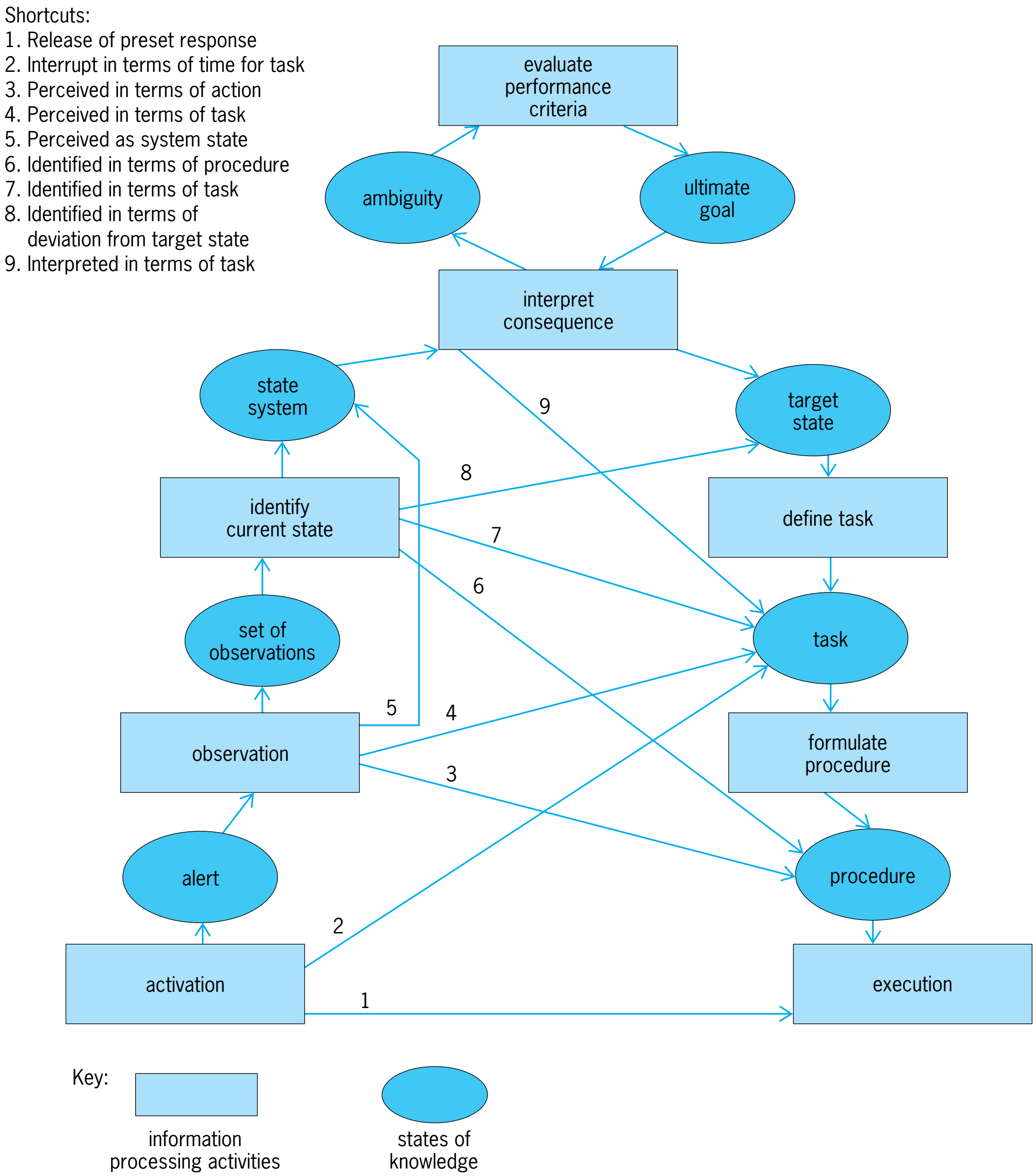

Shortcuts:
1. Release of preset response
2. Interrupt in terms of time for task
3. Perceived in terms of action
4. Perceived in terms of task
5. Perceived as system state
6. Identified in terms of procedure
7. Identified in terms of task
8. Identified in terms of deviation from target state
9. Interpreted in terms of task
evaluate performance criteria
ambiguity
ultimate goal
interpret consequence
state system
target state
identify current state
define task
set of observations
task
observation
formulate procedure
alert
procedure
activation
execution
1
2
3
4
5
6
7
8
9
Key:
information processing activities
states of knowledge

Fig. 3. **Decision ladder. Human performance can be described as skill-based (using shortcut 1), rule-based (using shortcuts 2–9), or knowledge-based (traversal of the entire ladder). (*After T. B. Sheridan and G. Johannsen, eds., Monitoring Behavior and Supervisory Control, Plenum Press, 1976*)**

compensating for system failures and abnormal conditions. Whereas earlier models of supervisory control tended to focus on monitoring and reactive, compensatory activities, more advanced research has emphasized the active, anticipatory nature of supervisory control. That is, supervisory controllers do not just monitor the system and react to perceived problems; rather, they use so-called mental simulation to predict the future state of the system, actively seek information, set goals, make plans, and perform control actions that anticipate future system state changes.

Human supervisory control differs in many important ways from human manual control. First, human supervisory control is primarily cognitive, that is, concerned with problem solving, planning, decision making, and other kinds of thinking rather than with manual skill and human physical constraints. Second, the systems that are supervised are frequently very complex and highly automated, requiring the human to manage multiple concurrent activities or, typically, requiring a team of human operators (as discussed below). Finally, the level of automation has important implications for the design of human-machine systems with respect to function allocation, human error and human reliability analysis, trust between humans and machines, and the overall integration of human problem-solving skills with potentially obscure workings on the part of the machine. In general, human supervisory controllers are at least one step removed from the actual process being controlled, and how they can maintain situation awareness and be ready to act quickly and appropriately is a very important and active area of research.

Models. Supervisory control has been modeled by using a wide variety of mathematical and analytic tools, including queueing theory, fuzzy set theory, network models of activity or state changes such as Petri nets, the discrete control model, the operator function model, and a variety of artificial-intelligence knowledge-representation schemes, including rule-based models and blackboard models that reflect activity and decision making in operational tasks. One influential model of human decision making in complex environments is the decision ladder that distinguishes between skill-, rule-, and knowledge-based behavior (**Fig. 3**). Alternative models emphasize the role of situational context and active information seeking, planning, and other anticipatory cognitive actions characteristic of supervisory control. *See* ARTIFICIAL INTELLIGENCE; DECISION ANALYSIS; FUZZY SETS AND SYSTEMS; QUEUEING THEORY.

System design. System design to support human supervisory control has followed three major lines: display design (or representation aiding), telerobotics, and intelligent support systems. Representation aiding considers how to represent and organize system information to support effective human problem solving. One approach is to use the abstraction hierarchy (**Fig. 4**) as a framework for supporting situation assessment. The abstraction hierarchy represents the functional properties of a system according to means-ends relationships; this describes, top-down, why the system works as it does and, bottom-up, how it works as it does. The basic idea of using the abstraction hierarchy as a basis for display design (an approach also called ecological interface design) is to directly show higher-order properties of the system (higher levels as described in the abstraction hierarchy) and thus in a sense change a complex reasoning task into a (presumably) more tractable perceptual one.

Telerobotics refers to the supervisory control of a sophisticated teleoperator that is sometimes called a telerobot. The human supervisor intermittently communicates goals, plans, and other information to a computer to guide the telerobot in carrying out its task. Through the computer intermediary, the human keeps track of the telerobot's progress.

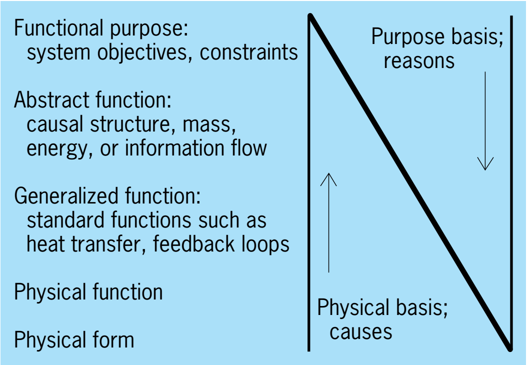


Fig. 4. **Abstraction hierarchy. (*After J. Rasmussen, Skills, rules, knowledge: Signals, signs, and symbols and other distinctions in human performance models, IEEE Trans. Sys. Man Cybernet., SMC-13, (3):257–267, 1983*)**

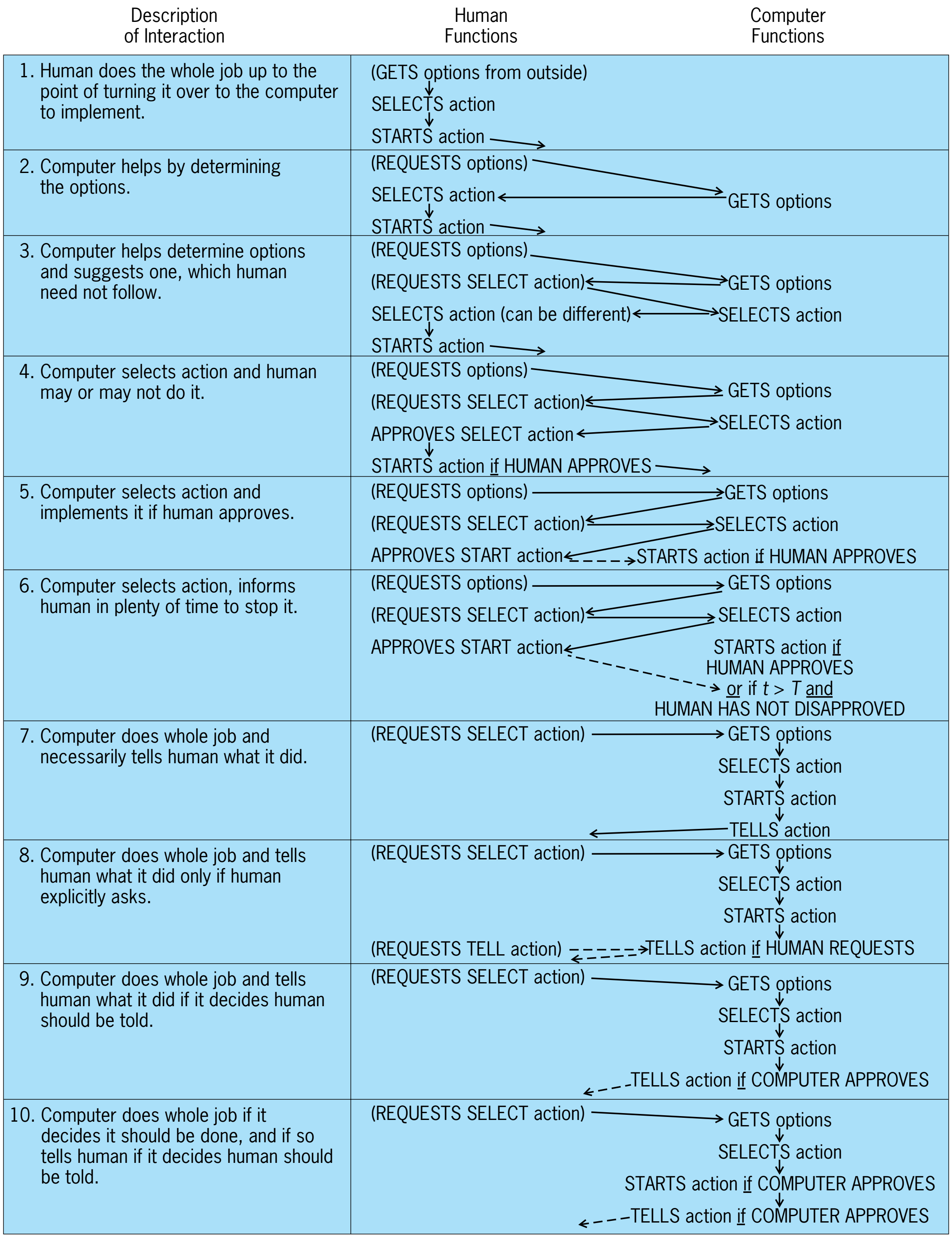
Intelligent support systems involve the use of artificial-intelligence programs that act as tutors, assistants, or both to human supervisory controllers. Artificial-intelligence applications in supervisory control systems include expert systems for automatic fault detection and diagnosis; intelligent associate systems that dynamically adapt their behavior to the human in recognition of the intent of operator actions; and intelligent tutoring systems that teach procedural knowledge and diagnostic skills. The concept of intelligent agents offers the prospect of personalized assistants that gradually provide automation in cooperation with the human user: The agent monitors the user's actions, recognizes recurring patterns, and eventually offers to invoke that sequence of actions automatically for the user; alternatively, users can program agents directly by themselves, sometimes via graphical programming languages. However, relatively few such systems have been deployed in actual supervisory control environments. *See* EXPERT SYSTEMS; FAULT ANALYSIS; INTELLIGENT MACHINE.

Human–intelligent-system interaction. In the design of intelligent systems for human use, many issues about trust, authority, and responsibility in decision making are important to consider. For example, in interacting with a typical expert system, the human user inputs data, waits for the expert system to generate a judgment or diagnosis by applying its reasoning algorithms, and then decides if that result makes any sense. This kind of interaction does not support human authority in the decision-making process; this is a particularly important issue in complex dynamic systems in which the costs of errors are potentially catastrophic and it is usually not possible to undo control actions. There are a number of different levels of intelligent automation in decision-making tasks (**Fig. 5**).

In general, emerging theories of how to design good intelligent systems for human use in complex dynamic environments advocate cooperative support for human authority rather than full automation of decision making and reasoning. For example, intelligent systems can support cooperative problem formulation, exploration of alternatives, and dynamic task allocation between human and machine (that is, adaptive automation, an issue that has received much attention).

Besides these general issues of knowledge representation, inference, and trust, control, and responsibility between humans and machines, intelligent systems for supervisory control have the difficulty of supporting problem solving under tight time constraints. A variety of quantitative and computational techniques have been used to provide reasoning under uncertainty in dynamic environments.

Teams in complex systems. Group work plays an important role in human-computer interaction, artificial intelligence, and human supervisory control.



Description of Interaction	Human Functions	Computer Functions
1. Human does the whole job up to the point of turning it over to the computer to implement.	(GETS options from outside) SELECTS action STARTS action	
2. Computer helps by determining the options.	(REQUESTS options) SELECTS action STARTS action	GETS options
3. Computer helps determine options and suggests one, which human need not follow.	(REQUESTS options) (REQUESTS SELECT action) SELECTS action (can be different) STARTS action	GETS options SELECTS action
4. Computer selects action and human may or may not do it.	(REQUESTS options) (REQUESTS SELECT action) APPROVES SELECT action STARTS action if HUMAN APPROVES	GETS options SELECTS action
5. Computer selects action and implements it if human approves.	(REQUESTS options) (REQUESTS SELECT action) APPROVES START action	GETS options SELECTS action STARTS action if HUMAN APPROVES
6. Computer selects action, informs human in plenty of time to stop it.	(REQUESTS options) (REQUESTS SELECT action) APPROVES START action	GETS options SELECTS action STARTS action if HUMAN APPROVES or if $t > T$ and HUMAN HAS NOT DISAPPROVED
7. Computer does whole job and necessarily tells human what it did.	(REQUESTS SELECT action)	GETS options SELECTS action STARTS action TELLS action
8. Computer does whole job and tells human what it did only if human explicitly asks.	(REQUESTS SELECT action) (REQUESTS TELL action)	GETS options SELECTS action STARTS action TELLS action if HUMAN REQUESTS
9. Computer does whole job and tells human what it did if it decides human should be told.	(REQUESTS SELECT action)	GETS options SELECTS action STARTS action TELLS action if COMPUTER APPROVES
10. Computer does whole job if it decides it should be done, and if so tells human if it decides human should be told.	(REQUESTS SELECT action)	GETS options SELECTS action STARTS action if COMPUTER APPROVES TELLS action if COMPUTER APPROVES

Fig. 5. Levels of automation in human-machine interaction. Other variations are possible. (*After T. B. Sheridan and W. L. Verplank, Human and Computer Control of Undersea Teleoperators, Tech. Rep., MIT Man-Machine Systems Laboratory, Cambridge, Massachusetts, 1978*)

The term distributed supervisory control refers to the situation in which a team of human operators works together in the supervisory control of a complex dynamic system. This situation is an example of computer-supported cooperative work, an interdisciplinary area that draws upon computer science, information systems, sociology, psychology, and anthropology to examine the mutual influences of technologies (such as electronic mail and video conferencing) on group work. Research areas include the organizational structures of the workplace, the methods by which the people coordinate and negotiate activity and responsibility, and the use of different technologies in the work environment. Supporting good cooperative work by system design relies on both an understanding of the formal requirements of work (which could be captured in engineering models) and the informal work that is done to coordinate activity (which cannot be modeled but must be possible). Included are analysis of social, cultural, organizational, and managerial practices as well as the more traditional engineering analyses associated with tasks and physical processes. *See* INFORMATION SYSTEMS ENGINEERING; SYSTEMS ENGINEERING. Patricia M. Jones

Bibliography. R. Baecker (ed.), *Readings in Groupware and Computer Supported Cooperative Work*, 1993; E. Hollnagel, G. Mancini, and D. Woods (eds.), *Cognitive Engineering in Complex Dynamic Worlds*, 1986; J. Rasmussen, A. Pejtersen, and L. Goodstein, *Cognitive Systems Engineering*, 1994; W. B. Rouse, *Systems Engineering Models of Human-Machine Interaction*, 1980; A. P. Sage (ed.), *System Design for Human Interaction*, 1987; T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, 1992.

Humidification

The process of increasing the water-vapor content (humidity) of a gas. This process and its reverse operation, dehumidification, are important steps in air conditioning for human comfort and in many industrial operations. For a discussion of the dehumidification of gases, with the exception of air, *see* DRYING

Humidity is normally expressed as the weight of water per unit weight of dry air. This value is also referred to as the absolute humidity or humidity ratio and has the units kg/kg, lb/lb, or grains/lb. The water content of air also is frequently indicated as a relative humidity, which is the ratio (expressed as a percentage) of the partial pressure of water vapor in the air to the vapor pressure of pure water at the same temperature. *See* HUMIDITY.

Air (or other gas) can be humidified by direct injection of water vapor (steam) or, more commonly, by the evaporation of liquid water in contact with the airstream. When evaporation occurs, heat is required to provide the latent heat of vaporization. If no external source of heat is provided, either the water or the air, or both, will be cooled. The cooling of water by this process is the basis of operation for industrial cooling towers, whereas evaporative air coolers often used in hot, dry climates depend upon the air-cooling effect. In both these types of apparatus, humidification of the air occurs, although it is not the prime objective of the operation. In units designed primarily for humidification, the incoming air is usually heated to provide the latent heat of evaporation and to permit the air to leave the unit at controlled levels of both temperature and humidity.

Psychrometric chart. The operation of humidification equipment can best be understood by reference to a psychrometric chart as shown in **Fig. 1**. Every point on this chart represents a specific air condition with regard to temperature (bottom scale) and humidity (right-hand scale). Several other characteristics of the air are indicated by the location of the points on the chart, including the relative humidity and the wet-bulb temperature. The latter is defined as the equilibrium temperature attained by a small surface of liquid evaporating into a large amount of

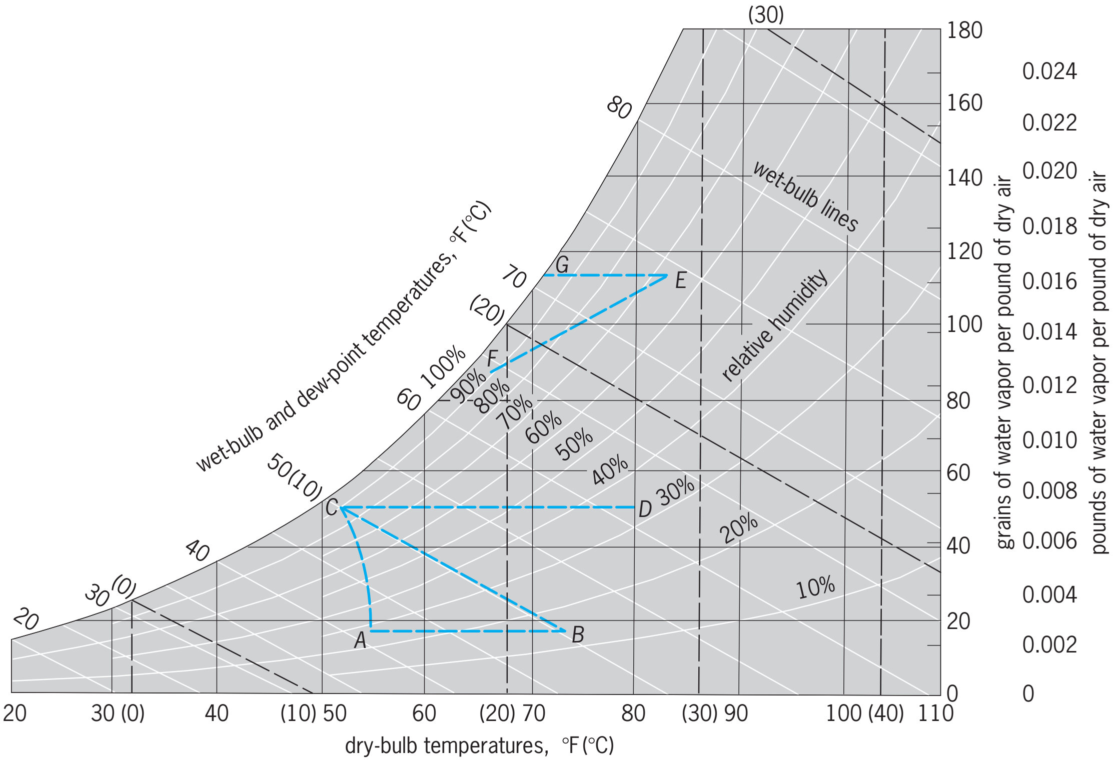


Fig. 1. An example of a psychrometric chart.

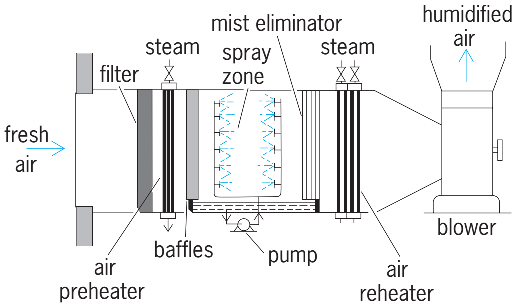


Fig. 2. Schematic of air conditioning humidifier.

unsaturated air. It is normally measured by moving air rapidly past a thermometer bulb covered with a wetted wick—hence the name. For the air-water system, the wet-bulb lines can also be used with sufficient accuracy to indicate the adiabatic saturation temperature, that is, the temperature which a gas will attain when saturated by adiabatic contact with water which is already at the adiabatic saturation temperature. *See* PSYCHROMETRICS.

Operation of an air conditioning humidifier such as shown in **Fig. 2** can be represented by the path *A-B-C-D* on the psychrometric chart. Air entering at the conditions indicated by point *A* is first heated to point *B* without change in humidity by contact with a heated coil. It then passes through a water-spray zone and is adiabatically cooled and humidified to the conditions represented by point *C*. The humidified air is then heated to the desired final outlet conditions of point *D*. Alternatively, the entering air can be contacted with heated water so that humidification occurs with little or no cooling. This path is represented by line *A-C*. *See* AIR CONDITIONING.

Cooling towers. Line *A-C* also represents a typical cooling-tower operation in which warm water is contacted with air for the purpose of cooling the water. Cooling towers are of considerable industrial importance in power plants, refineries, chemical plants, and large air conditioning or refrigeration installations in which considerable quantities of water are used in condensers and coolers to remove process heat. Two general types of water cooling towers are employed—natural circulation and mechanical draft. The former depend primarily upon wind and natural draft effects to provide air circulation, whereas the latter employ fans to move air through the tower. Cooling towers may be further subdivided into crossflow designs, in which the air moves horizontally while the water falls vertically, and counterflow designs (**Fig. 3**), in which the air moves upward countercurrent to the falling water. The crossflow design is useful for towers which must be kept to a minimum height; however, counterflow operation is theoretically more efficient, and this type of tower is capable of producing colder water. *See* COOLING TOWER.

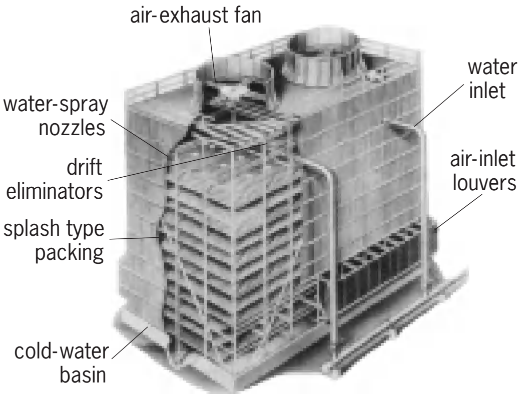


Fig. 3. Cutaway view of large industrial counterflow cooling tower. (*Fluor Products Co.*)

Dehumidification. The dehumidification of air is indicated by line *E-F* on the psychrometric chart. Dehumidification may be accomplished by contacting the air with cold water in a device similar to that of Fig. 2, except that the heater coils are not necessary, or by passing the air across banks of finned tubes through which cold water or refrigerant is passed. Because water condenses on the outside walls of the tubes, the two processes are equivalent in that the warm, moist air is in direct contact with cold water. Condensation results in the release of the latent heat of condensation, and this raises the temperature of the cooling liquid. It is not necessary for the entire gas stream to be cooled to the dew point (point *G*) for condensation to occur; however, no water can be condensed from the gas unless the cooling surface is below the dew point. *See* DEHUMIDIFIER.

Dehumidification can also be accomplished by the use of solid desiccants such as silica gel or alumina, or liquid absorbents such as triethylene glycol or lithium chloride solutions. Both types are commonly employed in regenerative systems in which the spent dehydration agent is stripped of water by applying heat. *See* DESICCANT; HEAT EXCHANGER; REFRIGERATION; STRIPPING (CHEMICAL ENGINEERING); UNIT OPERATIONS.

Arthur L. Kohl

Bibliography. American Society of Heating, Refrigerating, and Air Conditioning Engineers, *ASHRAE Handbook: Systems*, 1992, *Fundamentals*, 1994, *Applications*, 1993; G. W. Brundrett, *Criteria for Moisture Control*, 1990; J. M. Coulson and J. F. Richardson (eds.), *Chemical Engineering*, vol. 1: *Fluid Flow, Heat Transfer, and Mass Transfer*, 4th ed., 1990.

Humidistat

A controller that measures and controls relative humidity. A humidistat may be used to control either humidifying or dehumidifying equipment by the regulation of electric or pneumatic switches, valves, or dampers. Most methods for measuring humidity rely upon the swelling and shrinking of materials, such as human hair, silk, horn, gold-beater's skin, and wood, with increases and decreases in relative humidity.

Human hair is most commonly used because of its small diameter, which contributes to rapid absorption and dissipation of moisture. Strands of hair are

Fig. 1. Human-hair-element humidistat. (*Honeywell Inc.*)

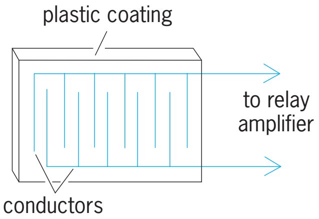


Fig. 2. Electronic humidistat. (*Honeywell Inc.*)

bunched and several such bunches are combined in a ribbonlike element (**Fig. 1**).

As the relative humidity of the air decreases, the strands of hair shorten; this movement is transmitted through a suitable lever mechanism to an electric switch or pneumatic valve, which is part of the humidistat.

An electronic humidistat includes a sensing element and a relay amplifier. The sensing element consists of alternate metal conductors on a small, flat plate with a plastic coating (**Fig. 2**).

An increase or decrease of the relative humidity causes a decrease or increase in the electrical resistance between the two sets of conductors; the change in resistance is measured by the relay amplifier. Small changes in relative humidity can be measured in this way for precise control. *See* HUMIDITY; HUMIDITY CONTROL; PSYCHROMETRICS.

John E. Haines; Richard L. Koral

Bibliography. I. Stepnich, *Humidity Control,* 1988.

Humidity

Atmospheric water-vapor content, expressed in any of several measures, especially relative humidity, absolute humidity, humidity mixing ratio, and specific humidity. Quantity of water vapor is also specified indirectly by dew point (or frost point), vapor pressure, and a combination of wet-bulb and dry-bulb (actual) temperatures. *See* DEW POINT; VAPOR PRESSURE.

Relative humidity is the ratio, in percent, of the moisture actually in the air to the moisture it would hold if it were saturated at the same temperature and pressure. It is a useful index of dryness or dampness for determining evaporation, or absorption of moisture. *See* PSYCHROMETRICS.

Human comfort is dependent on relative humidity on warm days, which are oppressive if relative humidity is high but may be tolerable if it is low. At other than high temperatures, comfort is not much affected by high relative humidity. *See* COMFORT TEMPERATURES.

However, very low relative humidity, which is common indoors during cold weather, can cause drying of skin or throat and adds to the discomfort of respiratory infections. The term indoor relative humidity is sometimes used to specify the relative humidity which outside air will have when heated to a given room temperature, such as 72°F (22°C), without addition of moisture. The indoor relative humidity always has a low value in cold weather and is then a better measure of the drying effect on skin than is outdoor relative humidity. This is even true outdoors because, when air is cold, skin temperature is much higher and may approximate normal room temperature. *See* BIOMETEOROLOGY.

Absolute humidity is the weight of water vapor in a unit volume of air expressed, for example, as grams per cubic meter or grains per cubic foot.

Humidity mixing ratio is the weight of water vapor mixed with unit mass of dry air, usually expressed as grams per kilogram. Specific humidity is the weight per unit mass of moist air and has nearly the same values as mixing ratio.

Dew point is the temperature at which air becomes saturated if cooled without addition of moisture or change of pressure; frost point is similar but with respect to saturation over ice. Vapor pressure is the partial pressure of water vapor in the air. Wet-bulb temperature is the lowest temperature obtainable by whirling or ventilating a thermometer whose bulb is covered with wet cloth. From readings of a psychrometer, an instrument composed of wet- and dry-bulb thermometers and a fan or other means of ventilation, values of all other measures of humidity may be determined from tables. *See* HYGROMETER; MOISTURE-CONTENT MEASUREMENT; PSYCHROMETER.

J. R. Fulks

Bibliography. D. Ahrens, *Meteorology Today: An Introduction to Weather, Climate, and the Environment*, 5th ed., 1994; R. A. Anthes, *Meteorology*, 7th ed., 1996; W. L. Donn, *Meteorology*, 4th ed., 1975; A. Miller and J. C. Thompson, *Elements of Meteorology*, 6th ed., 1999.

Humidity control

Regulation of the degree of saturation (relative humidity) or quantity (absolute humidity) of water vapor in a mixture of air and water vapor. Humidity is commonly mistaken as a quality of air. *See* HUMIDITY.

When the mixture of air and water vapor is heated at constant pressure, not in the presence of water or ice, the ratio of vapor pressure to saturation pressure decreases; that is, the relative humidity falls, but absolute humidity remains the same. If the warm mixture is brought in contact with water in an insulated system, adiabatic humidification takes place; the warm gases and the bulk of the water are cooled as heat is transferred to that portion of the water which evaporates, until the water vapor reaches its saturation pressure corresponding to the resultant water-air-vapor mixture temperature. Relative humidity is then 100% and absolute humidity has increased. Heating of the mixture and use of the heated mixture to evaporate water is typical of many industrial drying processes, as well as such common domestic applications as hair drying. This same sequence occurs when warm furnace air is passed over wetted, porous surfaces to humidify air for comfort conditioning. *See* AIR CONDITIONING.

To remove moisture from the air-vapor mixture, the mixture is commonly cooled to the required dew point temperature (corresponding to the absolute humidity to be achieved) by passage over refrigerated coils or through an air washer where the mixture is brought in contact with chilled water. The result is a nearly saturated mixture which can be reheated, if required, to achieve the desired relative humidity. *See* DEHUMIDIFIER.

Moisture is also removed without refrigeration by absorption, a process in which the mixture passes through a spray of liquid sorbent that undergoes physical or chemical change as it becomes more dilute. Typical sorbents include lithium and calcium chloride solutions and ethylene glycol. *See* ABSORPTION.

Another means of dehumidification, by adsorption, uses silica gel or activated bauxite which, through capillary action, reduces the vapor pressure on its surface so that the water vapor in its vicinity, being supersaturated, condenses. *See* ADSORPTION; PSYCHROMETRICS. Richard L. Koral

Bibliography. G. W. Brundrett, *Criteria for Moisture Control*, 1990; R. Havrella, *Heating, Ventilating and Air Conditioning Fundamentals*, 1995; Instrument Society of America, *Moisture and Humidity,* 1985; I. Stepnich, *Humidity Control,* 1988.

Humite

A homologous series of magnesium nesosilicate minerals having the general composition $Mg_{2n+1}(SiO_4)_n(F,OH)_2$. The known species include norbergite ($n = 1$), chondrodite ($n = 2$), humite ($n = 3$), and clinohumite ($n = 4$). They are structurally related to forsterite olivine, Mg_2SiO_4, and brucite, $Mg(OH)_2$. All are based on hexagonal close-packed oxygen and fluorine atoms, the Mg atoms occupying octahedral interstices and the Si atoms occupying tetrahedral interstices. The hexagonal close-packed repeat distance is approximately 0.47 nm in these minerals. Forsterite, norbergite, and humite are orthorhombic; chondrodite and clinohumite are monoclinic; brucite is trigonal. Manganese analogs of these minerals occur as pink grains in metamorphosed manganese ores derived from preexisting siliceous carbonates and sedimentary manganese oxides. Other cations which can occur as substituents are Fe^{2+}, Ca^{2+}, Al^{3+}, and Ti^{4+}. Titanoclinohumite, for example, a high-pressure phase, has been found in deep-seated rocks. It may be one of the storage minerals for water at depth.

The minerals of the humite group have similar physical properties. The luster is resinous, and the color usually light yellow, brown, orange, or red. The pure synthetic Mg end members are colorless. Hardness is 6–6$^1/_2$ on Mohs scale, specific gravity is 3.1–3.2. They are very difficult to distinguish visually, and x-ray diffraction, electron microprobe, or optical techniques are required. They are found in regionally crystallized marbles, usually the skarn minerals associated with iron ores. Several species may occur in zoned contact with each other. Typical sources are the Grenville-age marbles in New York and Ontario, marble ejecta from Mount Vesuvius, and marbles from central Sweden. *See* SILICATE MINERALS.

Paul B. Moore

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals: Orthosilicates*, 2d ed., 1997.

Humus

A group of substances that are natural products of earth surface environments. Probably the most widely distributed organic carbon–containing materials in terrestrial and aquatic environments, they are dark-colored, predominantly aromatic, acidic, hydrophilic, molecularly flexible polyelectrolytes. Humic substances constitute 70–80% of the organic matter in inorganic soils and are formed from the chemical and biological degradation of plant and animal residues and from synthetic activities of microorganisms.

Chemistry

Based on their solubility in alkali and acid, humic substances are partitioned into three main fractions: humic acid, which is soluble in dilute alkali but is coagulated by acidification of the alkaline extract; fulvic acid, which is the humic fraction that remains in solution when the alkaline extract is acidified, that is, it is soluble in both dilute alkali and dilute acid; and humin, which is the humic fraction that cannot be extracted from the soil by dilute base or acid.

Analytical characteristics. The elemental composition and functional group content of a typical humic acid and fulvic acid show that (1) the humic acid contains approximately 10% more carbon (C) but 10% less oxygen (O) than the fulvic acid; (2) there is relatively little difference between the two fractions in hydrogen (H), nitrogen (N), and sulfur (S) contents; (3) the total acidity and COOH group content of the fulvic acid are appreciably higher than those of the humic acid; (4) both materials contain per unit weight approximately the same concentrations of phenolic OH and total C═O and OCH_3 groups, but the fulvic acid is richer in alcoholic OH groups than is the humic acid; (5) about 78% of the oxygen in the humic acid is present in functional groups, but all of the oxygen in the fulvic acid is similarly distributed; (6) the E_4/E_6 ratio (the ratio of absorbance at 460 over that at 660 nanometers) of the fulvic acid is twice as high as that of the humic acid, which means that the fulvic acid has a lower particle or molecular weight than the humic acid; and (7) the free-radical content of the humic acid is higher than that of the fulvic acid but the other electron spin resonance

TABLE 1. Analytical characteristics of a typical humic acid and fulvic acid*

Characteristic	Humic acid	Fulvic acid
Element, %		
Carbon	56.2	45.7
Hydrogen	4.7	5.4
Nitrogen	3.2	2.1
Sulfur	0.8	1.9
Oxygen	35.5	44.9
Functional groups, meq/g		
Total acidity	6.7	11.3
COOH	3.6	8.2
Phenolic OH	3.1	3.1
Alcoholic OH	2.6	6.1
Quinonoid C═O	2.9	2.7
Ketonic C═O	2.9	2.7
OCH_3	0.6	0.8
E_4/E_6	4.8	9.6
Free radicals, spins/g × 10^{-17}	2.85	0.64
Line width, gauss	4.5	3.5
g-value†	2.0043	2.0043

*From M. Schnitzer and S. U. Khan (eds.), *Soil Organic Matter*, Elsevier, 1978.
†Spectroscopic splitting constant.

parameters are similar (**Table 1**). *See* ELECTRON SPIN; FREE RADICAL.

Analytical characteristics of humins are similar to those of humic acids. The insolubility of humin in aqueous solutions arises from its being strongly retained or complexed by hydrous oxides and clay minerals.

Analysis of chemical structure. Aside from elemental and functional group analyses, the methods which have been used over the past century for obtaining information on the chemical structure of humic substances can be grouped into nondegradative and degradative methods. Nondegradative methods include different types of spectrophotometry and spectroscopy, x-ray analysis, electron microscopy, colloid-chemical, electrochemical, and radiochemical methods. Among degradative methods employed are oxidative and reductive degradation, hydrolysis, various types of irradiations, thermal analysis, and biological degradation. Of the many methods used, ^{13}C nuclear magnetic resonance spectroscopy and the oxidative degradation of methylated humic preparations have been especially informative.

Carbon-13 nuclear magnetic resonance spectra of humic substances provide inventories of the different components that make up these materials. Especially interesting is the 0–40-ppm region of the spectrum, which indicates the presence of paraffinic materials. The latter may constitute up to 35% of the weights of humic and fulvic acids. Important information on the identities of the major paraffinic components of humic and fulvic acids has been obtained by pyrolysis-field ionization and field desorption mass-spectrometric analyses of organic extracts of these materials. The latter contain *n*-alkanes (C_{17}–C_{101}), *n*-fatty acids (C_{15}–C_{34}), *n*-diols (C_{16}, C_{24}, C_{31}, C_{32}), sterols (C_{28}, C_{29}), *n*-alkyl monoesters (C_{40}–C_{68}), and *n*-alkyl diesters (C_{56}–C_{66}). In addition, humic acid extracts contain *n*-alkyl triesters (C_{75}–C_{93}). The highest molecular weight identified has a mass of close to 1500. The composition of the aliphatics in the humic acid and fulvic acid extracts resembles that of natural waxes. This suggests that humic materials contain significant amounts of waxlike materials that could positively affect the structural stability and water-retaining capacity of soils. The aliphatics present in the humic and fulvic acids could have originated from algae, fungi, bacteria, insects, earthworms, small animals, and plants. Because of their hydrophobic properties, these components would be expected to resist biodegradation and have long residence times in soils. *See* NUCLEAR MAGNETIC RESONANCE (NMR).

The oxidative degradation of humic and fulvic acids and humins produces aliphatic carboxylic, phenolic, and benzene carboxylic acids. The most abundant aliphatic degradation products are monocarboxylic acids (acetic to *n*-caprylic acids); dicarboxylic acids (oxalic to azelaic acids); and other polycarboxylic acids (propanetricarboxylic acid, butanetetracarboxylic acid, and methylfurandicarboxylic acid). Major phenolic acids include those with one to three OH groups and one to five COOH groups per aromatic ring. Prominent benzenecarboxylic acids are the tri, tetra, penta, and hexa forms. Humic and fulvic acids have similar chemical structures; however, the oxidation of humic acids yields more benzenecarboxylic acids but fewer phenolic acids than that of fulvic acid. Major products resulting from the reductive degradation (by zinc-dust fusion) of humic and fulvic acids are methyl-substituted-naphthalene, -anthracene, -phenanthrene, -pyrene, and -perylene. *See* ANALYTICAL CHEMISTRY.

Molecular structure. The chemical structure for humic acid can be characterized as aromatic rings that are joined by alkyl chains of various lengths. Oxygen is present in the form of carboxyls, phenolic and alcoholic hydroxyls, ether and ketone groups, and nitrogen in heterocyclic forms and as nitriles. The oxidative degradation of this structure would produce benzenecarboxylic acids, and the reductive degradation would yield aromatic polycyclics. The structure contains voids of various dimensions that could trap and bind other organics such as carbohydrates, proteins, lipids, and biocides as well as inorganics such as clay minerals and hydrous oxides. It is assumed that most of the carbohydrates and proteinaceous materials that are usually found in humic substances are adsorbed on the surfaces and in the voids and are not integral structural components. *See* ADSORPTION; CHEMICAL BONDING.

Chemical reactions. One of the most striking characteristics of humic substances is their ability to interact with metal ions, oxides, hydroxides, minerals, and organics, including toxic pollutants, to form water-soluble and water-insoluble associations of widely differing chemical and biological stabilities. The following types of reactions have been observed to occur in terrestrial and aquatic systems: formation of water-soluble simple metal complexes; formation of water-soluble mixed-ligand complexes; adsorption on and desorption from water-soluble mixed-ligand complexes; adsorption on and desorption from water-insoluble humic acids and metal-humate complexes; dissolution of minerals; adsorption on external mineral surfaces; and adsorption in clay interlayers. Through the formation of water-soluble complexes, humic materials can dissolve, mobilize, and transport metals and organics in soils and waters, while the formation of water-insoluble complexes brings about their accumulation in certain soil horizons and in sediments. Adsorption on clay surfaces and in clay interlayers stabilizes humic materials and protects them against chemical and biological decomposition over long periods of time. Thus, the formation of metal-humate and metal-fulvate complexes is important in soil genesis, for the formation of a good soil structure and for the availability of nutrients, especially those present only at microconcentrations. Humic substances can also interact with herbicides (such as atrazine), insecticides (DDT), and plasticizers (dialkylphthalates) by making them soluble in water and thus modifying their mobility and reactivity in terrestrial and aquatic environments. *See* COORDINATION CHEMISTRY; COORDINATION COMPLEXES; SOIL; SOIL CHEMISTRY.

Functions and uses. Humic substances enhance plant growth directly through positive physiological effects and indirectly by affecting the physical, chemical, and biological properties of soils. They also have nutritional effects in that they serve as sources of nitrogen, phosphorus, and sulfur for plants and microorganisms; biological functions in that they affect the activities of microorganisms; and physical functions in that they promote good soil structure, thereby improving tilth, aeration, and retention of moisture. All of these effects increase agricultural productivity. *See* NITROGEN CYCLE.

Humic substances are good chelating agents, have relatively large surface areas, are excellent dispersants, and act as reducing agents. Most of the uses that have been proposed for humic materials take advantage of these properties. Applications in various fields include (1) agriculture: additives to fertilizers and sprays, treatment and coating of seeds, and nutrients in hydroponics; (2) industry: in drilling muds for oil well rigs, as boiler scale removers, pigment extenders, emulsifiers, dispersants, corrosion inhibitors, wood preservatives and flotation reagents; (3) environment: in waste management, deodorizing liquids and gases, stack scrubbing, and absorption of herbicides; and (4) medicine: antimicrobial, anti-inflammatory, antitumor agents, liver stimulants, as well as medications for treating gastric ulcers, bleeding, and skin burns. *See* CHELATION; HYDROPONICS. Morris Schnitzer

Biochemistry of Humus Formation

Humus is formed during the biodegradation or decay of plant, animal, and microbial organic residues in soils, swamps, and waters. It is a very important constituent of soils and has a number of beneficial properties. These include slow releasing of plant nutrient elements, especially nitrogen; improving soil physical properties; assisting trace-element nutrition of plants through chelation reactions; promoting solubilization of plant nutrients from insoluble minerals; having a high adsorptive or exchange capacity for nutrient cations; promoting growth via certain components; favoring heat absorption because of its dark color; increasing soil buffer capacity; supporting a greater and more varied soil population, which favors biological control; reducing the toxicity of both natural and anthropogenic toxic substances; and promoting increased soil water–holding capacity.

Although humus is a mixture of numerous organic substances, two types of polymers, humic acids and polysaccharides, constitute the major fractions. Humic acids appear to be complex polymers of hydroxyphenols, hydroxybenzoic acids, and other aromatic structures with linked peptides (proteins), amino sugar compounds, fatty acids, microbial cell wall and protoplasmic fractions, and possibly other constituents. The polysaccharide fraction is a complex mixture of polymers of a large number of sugar and sugar-type constituent units. *See* BIOPOLYMER.

The use of ^{14}C-labeled plant residues or specific residue constituents has made it possible to follow more precisely the residue decomposition and transformation processes and the stabilization in humic substances.

Plant residues. As soon as plant residues are returned to the earth and environmental factors such as temperature and moisture are favorable, microorganisms, insects, and worms begin to utilize them as food (carbon), nutrient, and energy sources. Some of the carbon is released as carbon dioxide (CO_2), some is used for synthesis of cell constituents (biomass) and products, and a portion is converted to relatively resistant humic acid–type polymers or to substances which are stabilized by complexing with inorganic metal ions or clays. Initially, about 40–60% of the carbon of readily available plant constituents will be converted to biomass and synthesized organic products. These are subject to decay as they are produced or when the cells die.

TABLE 2. Decomposition and stabilization in humus of organic residue carbons after 1 year in a fertile sandy loam top soil, in percent

Carbon	Glucose	Algal protein	Wheat straw	Polysaccharide from wheat straw	Cornstalk ring lignin C	Cornstalk OCH_3 lignin C	*Aspergillus glaucus* melanin
Added C evolved as CO_2	85	77	69	82	33	58	13
Residual C in biomass	16	11	7.4	10.8	0.5	1.2	0.2
Residual C in humic acid	20	22	34	24	66	58	71
Residual C lost upon 6 *N* HCl hydrolysis	82	78	75	73	9	14	5

Generally after 3 months, about 50–60% of the carbon of most crop residues and plant leaves returned to the soil will have evolved as CO_2. After a year, the carbon loss will have increased to about 55–70%. About 5–15% of the residual carbon will be present in soil biomass and 85–95% in the new humus. Residues having relatively high lignin contents degrade at a slower rate, and young succulent plant tissues and high-nitrogen residues such as legumes decompose a little faster. The bulk of the residual carbon will be present in components which are hydrolyzed in 6 *N* hydrochloric acid (HCl), primarily peptides or proteins, and polysaccharides. A smaller portion will be present in aromatic polymers. The aromatic units are derived primarily from the plant lignins and through microbial synthesis (**Table 2**).

Biodegradable substances. Readily available, small molecular substrates or plant constituents such as sugars, amino acids, aliphatic acids, and pyrimidines are metabolized within a few hours or days and carbon is rapidly evolved as CO_2. After a year, 80–90% of the original carbon will have evolved as CO_2. About 10–20% of the residual carbon will be in the form of soil biomass and 80–90% in new humus. With time, the percentages of residual substrate carbons in biomass will decline and the percentage in humus will increase. In most soils, the biomass constitutes about 2–4% of the organic carbon.

About 20% of the residual carbon from readily biodegradable substrates will be associated with the humic acid fraction of humus, and possibly 10–20% of this carbon is present in aromatic units. The bulk of the residual carbon, however, will be present in the form of peptides and polysaccharides, which are released as sugar or amino acid units upon acid hydrolysis. This would be expected as the major portion of the metabolized carbon not released as CO_2 would be transformed into microbial protoplasm, cell wall polymers, and polysaccharides that could be partially linked into humic molecules during autolysis and decay of the cells. *See* AMINO ACIDS; PEPTIDE; POLYSACCHARIDE.

Polysaccharides, proteins, and lipids. Sixty percent or more of most organic residues consist of cellulose and other polysaccharides. Some residues such as legumes and microbial tissues contain 6–65% protein. Specific plant and microbial tissues may also contain appreciable amounts of lipids. Most of these materials are highly biodegradable, but especially during the early stages of decomposition many decompose a little slower than the simple sugars and acids (Table 2). But after 6 months or a year, about 70–85% of the carbon will have evolved as CO_2 and about 6–16% of the residual carbon will be present in soil biomass and 84–94% in new humus. The small decrease in loss of carbon as CO_2 compared with the more simple substrates indicates that some of the original polymer carbons, or more likely, partially degraded units are stabilized by incorporation into the new humus. *See* CELLULOSE; LIPID.

Phenols. Simple phenolic substances and other aromatic compounds may be present in plant and microbial residues and are released during biodegradation of aromatic polymers such as lignins. Some soil fungi synthesize up to 30 or more phenolic compounds. Phenolic substances are degraded by numerous species of organisms; however, in soil, the carbon losses are not as great as would be expected. After 3 months losses vary from about 10 to 71%. This indicates that a portion of intact ring molecules are stabilized in humus. The more reactive the compound with respect to radical formation or oxidative polymerization reactions, the greater the stability. Also, the more reactive the phenol, the smaller the percentage of residual carbon incorporated into biomass. *See* PHENOL.

Lignins. This is the second most abundant polymer synthesized by plants. It is an important source of structural units for humus formation. During one year, about 15–30% of the ring and 2-side-chain carbons of model and cornstalk lignins are evolved as CO_2. This compares with about 30–50% loss of 1- and 3-side-chain and methoxyl carbons. During the second year, the losses are greatly reduced.

Whereas the bulk of the residual carbons from readily available substrates are associated with peptide and polysaccharide polymers, the major portion of the residual lignin carbons are associated with aromatic complexes. If most of the aromatic rings were cleaved by microbial enzymes, a substantial portion of the ring carbons would have been utilized by the microbes and would be present in the peptide and polysaccharide fractions of the humus. Biomass estimations also support this conclusion: after 6 months or 1 year, about 5–15% of the residual carbons from the readily available substrates are present in biomass compared to about 0.2–1% for the residual lignin carbons. *See* BIOMASS.

Lignin biodegradation and stabilization studies indicate that lignin is indeed an important substrate for humus formation, but it undergoes very profound changes during humification. In peat bogs and swamps, greater quantities of aromatic lignin–derived structures persist as most organisms cannot readily cleave the benzene ring under anaerobic conditions. *See* LIGNIN.

Melanins. Many organisms, including fungi, bacteria, and actinomycetes, synthesize dark polymers called melanins. The melanins from *Epicoccum nigrum, Aspergillus sydowi, Hendersonula toruloidea,* and *Eurotium echinulatum* are similar to soil humic acids with respect to elemental composition, reactive chemical groups, amino acids released upon acid hydrolysis, phenols released upon sodium-amalgam reductive degradation, resistance to microbial degradation, types of structures released upon oxidative degradation and pyrolysis, nuclear magnetic resonance spectra, and low polysaccharide content. It is therefore very likely that these and other fungal melanins contribute to humic acid formation.

Humic acid formation. The ability of numerous phenolic compounds to undergo enzymatic and autoxidative polymerization reactions is probably of major importance in the formation of humic acid molecules. During the early microbial metabolism of these compounds, transformations such as beta oxidation of side chains, decarboxylation, demethoxylation, and formation of additional OH groups form a large variety of phenolic substances. Many of these such as dihydroxy and trihydroxy phenols readily autoxidize to form polymers at pH values of 6 and above. Also these and less reactive phenols such as ferulic acid, coniferyl alcohol, vanillic acid, and orcinol are readily oxidized by microbial phenolases and peroxidases present in soil or synthesized by soil organisms. Less reactive phenols such as *p*-hydroxybenzoic and 2,5-dihydroxybenzoic acids are linked into the polymers if present at the reaction sites. The phenol oxidation process forms radicals that stabilize by linkage to form dimers, or on further oxidation forms quinones which link other phenols or substances with free amino groups such as peptides or amino sugar polysaccharides, and microbial cell wall structures through nucleophilic addition reactions. Condensed ring aromatic structures and parts of lignin and melanin molecules may be linked into the humic polymer molecules. Model humic acids prepared by the enzymatic polymerization of phenol mixtures and peptides are similar to natural humic acids. *See* AUTOXIDATION; FREE RADICAL.

Polysaccharide fraction. Inasmuch as most plant polysaccharides readily decompose, it is probable that the major portion of this humus fraction originates through microbial synthesis. Also polysaccharides contain amino sugar units which are synthesized by many microorganisms but are not generally present in plant polysaccharides. Microbial polysaccharides are also subject to decomposition, but the rate varies with specific polymers. Several factors may explain how polysaccharides are stabilized in the soil humus. The polysaccharide fraction of humus contains about 10% uronic acid units. These and units with *cis*-hydroxyls or phosphoric acid esters form salts or complexes with di- and trivalent metal ions which may greatly reduce their susceptibility to degradation. Also, polysaccharides may be complexed with clay minerals through metal ion-clay linkages which increase resistance to biodegradation. Complex polysaccharides containing amino acid or amino sugar units may be stabilized by linkage into humic acid polymers through nucleophilic addition to quinones, or strong hydrogen bonding to humic acids. Chitosan, which is a polymer of amino sugar units, and *Chromobacterium violaceum* polysaccharide, which contains glucosamine, are much more resistant to biodegradation than many other plant and microbial polysaccharides. Polysaccharides may further be protected inside dense aggregates, and highly branched polymers may be more resistant than straight-chain structures.

Hydrolysis procedures indicate that soil polysaccharide fractions normally contain 10 or more major sugar units and many others in smaller amounts. Attempts to isolate significant fractions with fewer structural units have been unsuccessful. Generally it has been concluded that it is extremely difficult to separate a complex mixture of polysaccharides into single types. Another possibility is that microbial and plant polysaccharides at all stages of decomposition could be recombined through the action of soil enzymes stabilized in humus or released during cell autolysis. Those structures which resist biodegradation or which readily form complexes with clays and metal ions would persist and contribute to the humus polysaccharide fraction. *See* SOIL MICROBIOLOGY.

James P. Martin

Bibliography. W. R. Jackson, *Humic, Fulvic, and Microbial Balance: Organic Soil Conditioning,* 1993; P. MacCarthy et al. (eds.), *Humic Substances in Soil and Crop Sciences: Selected Readings*, 1990; E. A. Paul and A. D. McLaren (eds.), *Soil Biochemistry*, vol. 4, 1975; F. J. Stevenson, *Humus Chemistry*, 2d ed., 1994; R. L. Tate III, *Soil Organic Matter: Biological and Ecological Effects*, 1987, reprint 1992.

Hunger

A term most commonly used to refer to the subjective feelings that accompany the need for food; however, the study of this topic has come to include consideration of the overall control of food intake. More specifically, experimental work on the problem of hunger has been concerned with the sensory cues that give rise to feelings of hunger, the physiological mechanisms that determine when and how much food will be ingested, and the mechanism governing the selection of the food to be eaten.

Cannon theory. The earliest experimental approach to these problems concentrated almost exclusively on the question of the sensations of hunger or, as they have come to be known, hunger pangs.

The early work of W. B. Cannon, which resulted in the so-called local theory, led to the conclusion that for both hunger and thirst the appropriate sensations arose peripherally in the body. According to theory, hunger sensations (pangs) arose from stomach contractions that stimulated local sensory nerves. Subsequent work substantiated the idea that increased stomach contractions do indeed accompany the state of hunger in ordinary circumstances. It seems unlikely, however, that these contractions contribute substantially to the detailed control of eating behavior. For example, when the sensory nerves from the stomach are cut or even when the stomach is altogether absent, eating behavior can go on in an essentially normal manner. Whatever influence stomach contractions may have on the ingestion of food, it is known that the motility of this organ can be controlled by both a neural and a hormonal route. The hormone involved may be one that is secreted by the stomach itself. *See* HORMONE; NERVOUS SYSTEM (VERTEBRATE).

Physiological mechanisms. Food consumption is basically controlled by the organism's nutritional status. Food deprivation leads to eating, and the ingestion of food materials terminates hunger sensations. The issues are to determine which physiological processes vary quantitatively with nutritional status, and to find out if these changes can be detected by the nervous system in a manner that would instigate and terminate food consumption. Attempts by researchers to find a simple humoral factor that might be involved linearly in this regulation have not been illuminating. In fact, not a single humoral factor has been identified that can be reliably linked to feeding onset as it occurs spontaneously.

Blood-sugar level. Blood-sugar level, which has received more attention than any other factor, can be used as a case in point. The concentration of blood sugar does indeed vary appropriately in a general way with the periodicity of the food cycle. Moreover, hyperglycemia (extremely high blood sugar) and hypoglycemia (extremely low blood sugar) have been observed to decrease and increase hunger, respectively. Detailed analyses of normal life variations of blood sugar, however, reveal that the relation between the concentration of blood sugar and hunger is not sufficiently close for this single humoral factor to be able to control hunger in any simple and direct manner. Moreover, feeding following injections of insulin occurs only after lethal levels of hypoglycemia have been reached. *See* CARBOHYDRATE METABOLISM.

Tissue utilization of food. The evaluation of more local tissue utilization of food has proved a more promising approach to this problem. There is now some evidence suggesting that the status of the liver is pivotal in the control of feeding. Depletion of liver glycogen stimulates feeding; its repletion terminates feeding in rats and rabbits. The control exerted by these hepatic signals may be mediated by the vagus nerve, as complete destruction of the hepatic vagus markedly decreases feeding. *See* LIVER.

Feeding termination. Many stimuli that terminate feeding have been identified. Eating in food-deprived animals is inhibited by the reduction of either cellular water or of plasma fluid. It is also reduced by gastric distension and by infusing nutrients into the intestine and into the systemic, especially venous hepatic, circulation. Satiation produced by nutrient absorption from the intestine may be mediated, in part, by the gut hormone cholecystokinin. Cholecystokinin's effects may be central because injections of this hormone (albeit in very high doses) into the brain ventricles reduce feeding. It is more likely, however, that the cholecystokinin is effective because it reduces the rate at which food passes through the stomach. Generally, inhibitory signals in the control of feeding are more effective when they are brought about in the context of the feeding act than when introduced outside this context.

Neural centers. The previously held notions of discrete neural centers for the onset and termination of feeding have been abandoned, as the complexity of the feeding act and its corresponding neural complexity have become more widely appreciated. Deficits in feeding produced by lesions to the lateral hypothalamus are now recognized to be part of a larger constellation of behavioral impairments. To be sure, destruction of the lateral hypothalamus results in a loss of feeding and drinking behaviors that only slowly recover. But these animals also suffer from a variety of other profound behavioral incapacities. They do not attend to sensory stimuli. They have severe motor impairments (for example, they do not even groom) and cannot be easily aroused. They are deficient in learning. Conversely, although electrical stimulation of the lateral hypothalamic area induces feeding and drinking, it often also leads to participation in other behaviors, such as nest building, carrying, and copulation—even in a given animal.

Other brain areas have been implicated in feeding control. Destruction of the ventromedial nuclei of the hypothalamus leads to overeating, but again the deficits are not specific. Such rats are vicious, hyposexual, and do not attend to their young. There are also metabolic consequences of the lesion that may cause overeating. In short, more recent work has shown that the idea of hypothalamic feeding and satiety centers is an oversimplification and that, seemingly, the entire brain is involved in feeding control. Even decerebrate rats (rats that have had their entire brain removed, from the hypothalamus forward) make very sensitive discriminations about the quality of solutions placed in their mouths. Like intact animals, they reject quinine or very salty solutions and avidly accept the sugars.

Neurochemical control of feeding. It appears that increasing the rate at which the transmitter norepinephrine is taken up by various tissues in the forebrain increases food intake. Conversely, increasing the transmission of serotonin decreases feeding. Again, the cautions raised about the brain lesion studies must be sounded here. The specificity of effects are not known. This is especially problematic because only a small fraction of the neurotransmitters that exist in the central nervous system have been identified.

Specific hungers. Deprivation of certain, specific food substances precipitates an increased appetite for the needed substance. This so-called specific hunger behavior has been demonstrated experimentally with many substances, such as salt, calcium, fats, proteins, and certain vitamins in children and in the lower animals studied. Theories to account for the control of eating by specific food deprivations have suggested the possibilities of learning, differentially lowered taste thresholds, and direct effects upon the brain.

It is now clear that only the hunger for salt in salt-deprived animals appears before the animal has learned about the beneficial consequences of salt ingestion. Specific hungers for other minerals, proteins, and vitamins appear only gradually and reflect the animal's learning that certain foods are no longer beneficial and, in fact, may be harmful. The evidence that animals learn that the "right" food is actually beneficial to them is meager. They appear to stop eating the food associated with the deficit and sample new foods, stopping when the one that corrects the deficit is eaten for a while. This behavior is in contrast with salt appetite, which increases immediately in sodium-deficient animals.

Feeding development. It may be tentatively concluded, in rats at least, that suckling in these animals, until about 2 weeks of age, is not under the control of internal stimuli that manage adult feeding. Suckling does, however, come under the control of these stimuli at about the time that rats start to eat from the environment. Portions of the feeding system appear to be present even in the newborn, however. When tested at very warm temperatures (90°F or 32°C), 3-day-old rats actually eat a liquid diet, and the amount eaten is proportional to the length of time that they have been food-deprived. Such animals, however, do not eat at room temperature. Feeding on a variety of diets at room temperature occurs at about the start of the third week after birth. *See* THIRST AND SODIUM APPETITE.

Elliott M. Blass

Huntington's disease

A hereditary disorder of the basal ganglia causing progressive motor incoordination, abnormal involuntary movements (chorea), and intellectual decline. The disease, which progresses gradually over 15–20 years, is invariably fatal.

Huntington's disease is a rare disorder that affects approximately five to ten of every 100,000 persons. Inherited as an autosomal dominant mendelian trait, it inevitably develops in those who carry the gene if they live long enough. Men and women are affected equally. The average age at onset is between 35 and 40 years, but the disease can begin as early as 2 years or as late as 80 years. The mutation rate is very low, and no confirmed new mutations have ever been reported. The disease appears to have originated in northern Europe, England, or Scandinavia. Pathologically, the nerve cells in the caudate nucleus and putamen degenerate, and although other regions of the brain are affected to a minor extent, these regions are the most seriously damaged.

The disease begins insidiously with restlessness, choreiform (dancelike) movements, and personality changes (irritability, depression, and occasionally psychosis). The symptoms gradually worsen, with severe involuntary, large-amplitude motions, grimacing, poor balance, and very poor fine-motor coordination. Abnormalities of eye movement also occur early in the disease and worsen with its progression; speech and swallowing are always affected. Death frequently results from secondary causes such as pneumonia. About 10% of those affected have a rigid form of the disease in which choreiform movements are less prominent, and stiffness and slowness of movement and abnormal posturing (dystonia) are the dominant features. This variety is more likely to occur when the disease begins before the age of 20 years. Therapy is merely supportive: no medications significantly affect the course of the disease or functional capacity of the sufferer. Depression or psychosis, however, can be temporarily alleviated by antidepressant and antipsychotic medications. A progressive intellectual decline is usually observed, but even in the late stages of the disease, patients can often recognize family members and interact with other people to a limited extent.

The gene for Huntington's disease has been localized at the end of the short arm of chromosome 4. The gene, termed IT15, contains a series of three nucleotides that are repeated from 11 to 30 times. On Huntington's disease chromosomes, this triplet repeat is expanded and occurs in greater numbers (37 to over 86 repeats). The IT15 gene itself is very large (10 kilobases) and is thought to encode a protein of approximately 3144 amino acids. The predicted protein, huntingtin, does not resemble any known protein. Mutations to Huntington's disease are rare but do occur. Offspring of individuals suffering from the disease have a 50% risk of developing it, and can be tested by recombinant genetic technology. *See* CHROMOSOME; MUTATION.

The diagnosis of symptomatic Huntington's disease is usually straightforward if a family history of the disease is confirmed through autopsy. The diagnosis can become more complicated, however, if affected relatives have died prematurely or have been misdiagnosed. In the absence of a clear family history, a tentative diagnosis can be made on the basis of the clinical signs and symptoms after other disorders with similar manifestations have been ruled out. *See* HUMAN GENETICS; NERVOUS SYSTEM DISORDERS.

Anne Young

Bibliography. S. E. Folstein, *Huntington's Disease: A Disorder of Families*, 1989; J. F. Gusella et al., A polymorphic DNA marker genetically linked to Huntington's disease, *Nature*, 306:234–238, 1983; Huntington's Disease Collaborative Research Group, A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes, *Cell*, 72:971–983, 1993; J. B. Martin and J. F. Gusella, Huntington's disease: Pathogenesis and management, *N. Engl. J. Med.*, 315:1267–1276, 1986; J. P. Vonsattel et al., Neuropathological classification of Huntington's disease, *J. Neuropathol. Exp. Neurol.*, 44:559–577, 1985; A. B. Young et al., Huntington's disease in Venezuela: Neurologic features and functional decline, *Neurology*, 36:244–249, 1986.

Hurricane

A tropical cyclone whose maximum sustained winds reach or exceed a threshold of 74 mi/h (119 km/h). In the western North Pacific Ocean it is known as a typhoon. Many tropical cyclones do not reach this wind strength. *See* CYCLONE.

Maximum surface winds in hurricanes range up to about 200 mi/h (320 km/h). However, much greater losses of life and property are attributable to inundation from hurricane tidal surges and riverine or flash flooding than from the direct impact of winds on structures.

Tropical cyclones of hurricane strength occur in low latitudes of all oceans except the South Atlantic and the eastern South Pacific, where combinations of cooler sea temperatures and prevailing winds whose velocities vary sharply with height prevent the establishment of a central warm core through a deep enough layer to sustain the hurricane wind system. *See* WIND.

Impact. Severe hurricanes are responsible for many of the world's greatest natural disasters. Hurricane Camille in 1969 devastated coastal communities of the central Gulf of Mexico; Hurricane Tracy on Christmas 1974 virtually destroyed the Australian city of Darwin; in 1989, Hurricane Hugo, the most destructive atmosphere event of all times, caused $7 billion damage in the Caribbean Sea and South Carolina; and the Bangladesh hurricane of 1970 killed more than 300,000 people. Lesser hurricanes often inflict years of social disruption and economic devastation in small tropical countries.

In the United States, hurricane damage continues to climb, primarily because of increasing exposure of property as the population continues to migrate to the seashores. At the same time, loss of life has decreased sharply because of more effective warnings which reflect the result of extensive research, more complete storm surveillance, and improved programs of public awareness.

On a global scale, the impact of hurricanes may change in the decades ahead as a consequence of increasing carbon dioxide in the atmosphere which may reduce radiative heat losses from the ground—the greenhouse effect. While the impact of global incidence of hurricanes remains controversial, the most likely change is an increase in the number of hurricanes at higher latitudes, owing to possible increases in sea surface temperatures. However, it will remain difficult to distinguish normal climatic aberrations from changes attributable to greenhouse warming for at least several decades. In some regions, the number of hurricanes could decrease. *See* GREENHOUSE EFFECT.

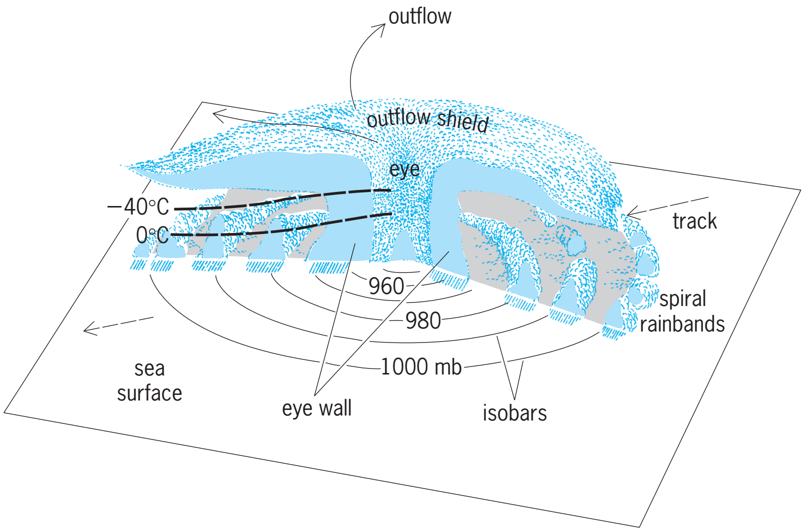


Fig. 1. Model of a hurricane circulation and cloud structure. °F = (°C × 1.8) + 32.

Structure. The characteristic signature of a hurricane (**Figs. 1** and **2**), as viewed by radar or satellite, is of a central, relatively cloud-free eye encircled by an eye wall of towering cumulonimbus clouds 15–50 mi (25–80 km) wide. A dense cloud deck covers the storm at upper levels, and spiral bands of

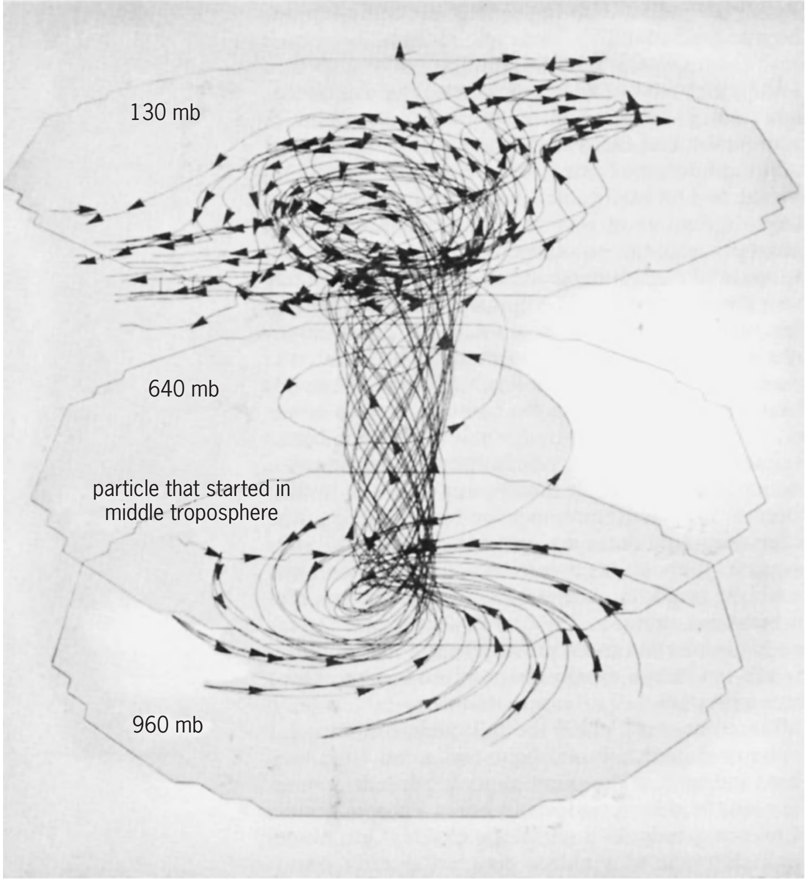


Fig. 2. Particle trajectories calculated over an 8-day period (90–282 h in 9-h intervals) in an experiment with a three-dimensional model hurricane. All particles start in the lower atmospheric boundary layer except one, which starts in the middle troposphere. 1 mb = 10^2 Pa. (*After R. A. Anthes, S. L. Rosenthal, and J. W. Trout, Preliminary results from an asymmetric model of the tropical cyclone, Mon. Weath. Rev., 99:744–758, 1971*)

cumulus clouds wind inward from the environment to the eye wall. *See* RADAR METEOROLOGY; SATELLITE METEOROLOGY.

Near the Earth's surface the hurricane appears as a nearly circular vortex of low pressure, typically 200–400 mi (320–640 km) in diameter. Its dynamic and thermodynamic properties, however, are asymmetrically distributed both at the surface and in upper layers. The cyclonic circulation (counterclockwise in the Northern Hemisphere) extends through most of the troposphere (to an atmospheric depth of about 9 mi or 15 km). In lower layers (up to 2 mi or 3 km), winds spiral inward and accelerate toward lower pressure, reaching peak speeds in a narrow annulus typically 10–20 mi (15–30 km) from the pressure center. Here there is a near balance between the pressure forces acting radially inward and the centrifugal and Coriolis forces acting outward, so that the air, no longer able to move radially, is forced upward, forming and maintaining the cloudy eye wall.

Momentum is transported upward from surface layers, so that wind speeds near the top of the eye wall are almost as strong as peak winds near the surface. Since pressure forces in the relatively warm eye wall must diminish with height, the consequent imbalance of forces causes the air in upper layers to spiral out of the vortex and join environmental circulations, carrying with it a canopy of cloud debris which may extend great distances into the environment. The raging winds in the eye wall, so closely adjacent to the relative calm within the eye, act as a centrifuge, dragging air from the eye which is replaced from above, the sinking motion therefrom gradually filling the eye with very dry, often cloud-free, warm air—warmer, in fact, than in the moist eye wall.

The spiral bands of cumulus clouds extending from the environment and entwining the eye wall result from complex processes. Some portions of these bands grow tall enough to augment the outflow from the eye wall. *See* CLOUD PHYSICS; CORIOLIS ACCELERATION; TROPOSPHERE.

Energy sources and processes. The hurricane draws energy from two main sources, its atmospheric environment and the ocean surface.

In the hurricane vortex, the circulation of mass—inward at low and midlevels, upward and outward aloft—at the rate of some 2.2 million tons (2 metric tons) per second, constitutes an atmospheric heat pump involving the import of angular momentum from the environment. The primary source of fuel for this pump is a combination of the latent heat released by the formation of towering cumulus clouds and heavy precipitation, and the much smaller but crucial flux of sensible heat from the ocean. Since low-level air, flowing from the environment into the storm core, is observed to maintain a nearly constant temperature, an important up-flux of heat energy from the warm ocean surface tends to compensate for the cooling that otherwise would occur during the transit toward lower pressure. Without this oceanic heat source, the pump would be unable to generate and sustain wind speeds of hurricane strength. The loss of this compensating sea-to-air heat flux is the primary reason for rapid decreases in strength usually observed when a hurricane moves over land. *See* ANGULAR MOMENTUM.

The process by which the tall cumuli in the eye wall maintain the warm, light air in the hurricane core, and in turn the pressure forces that determine the strength of hurricane winds, is not a simple matter of releasing latent heat within the clouds. Latent heat released by cumulus clouds does not directly warm the atmosphere, since most of this heat energy is converted to potential energy as air parcels rise. However, the direct role of the latent heat released is to maintain the buoyancy of the rising cloud air and thus to maintain the vertical circulation of the eye wall. The high temperatures at the storm center are created and maintained by adiabatic compression of descending air.

Since latent heat from atmospheric water vapor is the major fuel for the hurricane heat engine, the efficiency and available fuel supply is dependent upon both the atmospheric and oceanic environment. From a purely thermodynamic point of view, the overall hurricane efficiency is a function of the temperature difference between the ocean surface and the cold upper atmosphere where outflowing air, mixing with a colder environment, is further cooled by long-wave radiation to space. Thus warm tropical oceans provide the best medium for generating and sustaining the most intense hurricanes. *See* HEAT; PSYCHROMETRICS.

Hurricane intensity is also constrained by the amount of preexisting rotation in the environment. This rotation is normally quantified in terms of angular momentum, a combination of that from the Earth's rotation and that of environmental flow. If there is a store of cyclonic angular momentum and the necessary processes are triggered to set the development in motion, the low-level inflow described earlier carries this momentum toward the vortex center, generating hurricane-force winds in the same manner that rapid spin is achieved when ice skaters bring their arms close to their bodies. Some momentum is lost to friction at the Earth's surface, so that air flowing out of the eye wall returns to the environment with an anticyclonic motion. *See* ANGULAR MOMENTUM; VORTEX.

The degree of rotation determines the potential for releasing latent heat that can be used to accelerate the winds. In the undisturbed tropics, rotation is very low, so that the latent heat released by daily-observed cumulus clouds and thunderstorms diffuses away, unable to organize a warm core and the consequent pressure drops necessary to produce storm-scale wind systems. Similar amounts of latent heat are released inside hurricanes, but the high rotation ensures that much of this heating is retained locally to lower the pressure.

Formation. An average of about 70 tropical cyclones develops gale or hurricane force somewhere on Earth each year, a figure some 15–20% higher than climatological records indicated before the advent of weather surveillance satellites in the 1960s. The global distribution of tropical cyclone occurrence is shown in **Fig. 3**. *See* METEOROLOGICAL SATELLITES.

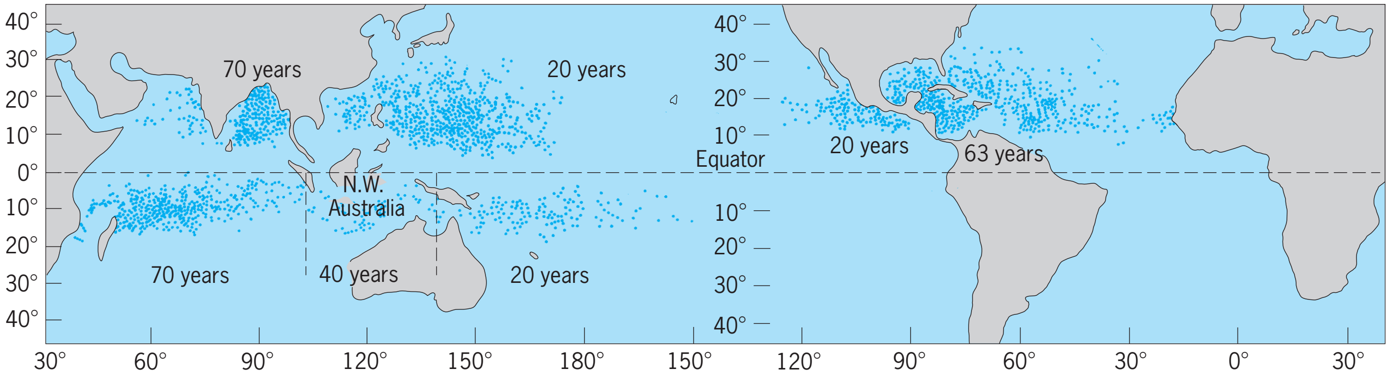


Fig. 3. **Points on the globe where tropical cyclones were first detected by weather observers. (*After W. Gray, Global view of the origin of tropical disturbance, Mon. Weath. Rev., 96(10):670, 1968*)**

Tropical disturbances, sometimes referred to as seedlings, that breed cyclones of hurricane strength originate mainly in tropical latitudes in the vicinity of the equatorial or monsoonal trough. Some have their origins over continental areas. Seedling disturbances, comprising an agglomeration of convective clouds 180–300 mi (300–500 km) in diameter, often move more than 1200 mi (2000 km) across tropical oceans as benign rainstorms before developing closed circulations and potentially dangerous winds. *See* MONSOON METEOROLOGY; TROPICAL METEOROLOGY.

A number of well-known factors may influence the development of seedlings. The atmosphere must be able to sustain and consolidate the growth of cumulus clouds to release latent heat and cause pressure falls. Also, ocean temperatures must be high enough to enhance the heat content of environmental air entering the system. It was established in the 1950s that this required an ocean temperature of greater than 79°F (26°C). Because of the need for background rotation, hurricanes rarely form or retain their identity equatorward of 5° latitude (Fig. 3).

The variation of prevailing winds with height also is important. If there is a sufficient change in wind velocity with height, the heat released by convection and the warming due to subsidence cannot be stored in vertical columns of sufficient height to produce the pressure drop required to sustain hurricane winds. Finally, conditions favorable to development in the low-level flow must be matched by an upper-layer environment that supports outflow from the convective system.

Nevertheless, even when these necessary conditions are satisfied, many seedlings fail to develop. In the North Atlantic, for example, 100 seedlings form each year (more than half emerging from Africa), yet only 9 develop into tropical cyclones. Knowledge of the complex interactions involved is not sufficient to explain this relatively low incidence of hurricanes.

Movement. The observed movement of hurricanes arises from two mechanisms, advection and propagation. Advection occurs when the environment directs an airflow over the hurricane that moves it in a fashion similar to a cork bobbing in a stream. Forecasters call this the steering current, and considerable effort is made to determine it accurately. Although there are many individual differences, the mean flow throughout the middle levels of the troposphere and over a radial band some 300–400 mi (500–650 km) from the center seems to provide the most consistent results.

Dynamically, the hurricane propagates by internal processes that are quite sensitive to the properties of the environment. A full understanding of many of these processes requires in-depth knowledge of fluid dynamics. One illustration, however, can provide an indication of the essential features. Any cyclone is made up of infinitesimally small rotating elements that are measured as vorticity. For a hurricane, maximum combinations of elements, and thus maximum cyclonic vorticity, is found near the center. Air that is stationary relative to the Earth also has cyclonic rotation and vorticity arising from the component of Earth rotation about the local vertical. The Earth's vorticity component increases from zero at the Equator to a cyclonic maximum at each pole. The flow around a hurricane thus brings increasing cyclonic vorticity to the west side, and decreasing cyclonic vorticity on the east side to produce a poleward flow over the tropical cyclone. This effect combined with associated internal rearrangements causes the hurricane to propagate westward and poleward. The magnitude of this propagation is defined by the outer structure of the hurricane; it is largely insensitive to the hurricane intensity.

The propagation velocity is generally about 3–5 mi/h (5–7 km/h) westward and poleward, but it may vary from near zero to over 10 mi/h (16 km/h) for a hurricane of large diameter. By comparison, most tropical cyclones reach hurricane intensity in the steady trade-wind belt where they are advected westward at a speed of 10–20 mi/h (15–30 km/h). As the hurricane approaches the western extremity of a subtropical ridge and begins moving poleward, the steering current first weakens, and then is replaced by westerlies whose speeds often exceed 40 mi/h (65 km/h). Thus cyclone propagation and advection tend to be in the same direction in the tropics but to act oppositely in midlatitudes.

Numerical simulations. Numerical simulation models have become a powerful tool for the study of hurricane dynamics and energetics. These models consist of finite-difference expressions for the

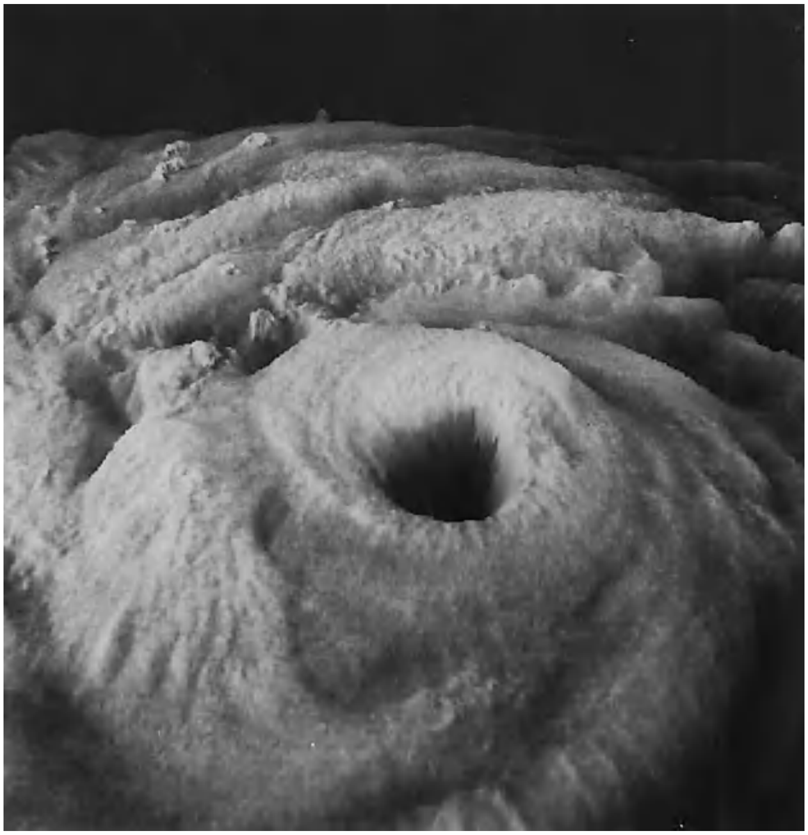


Fig. 4. Three-dimensional view of Hurricane Gilbert's eye and cloud system viewed from a weather satellite on September 12, 1988. (*F. H. Hasler, NASA Goddard Laboratory for Atmospheres*)

system of partial differential equations that govern the dynamics and energetics of the atmosphere. Pioneer experiments with computer models of hurricanes were carried out in the late 1950s. The limited computing capacity restricted hurricane models to be axisymmetric (assuming that hurricanes had no variations in horizontal azimuthal directions). These models were used to study the effects on hurricane intensity of sea-surface temperature, ambient atmospheric temperature and humidity stratification, and various hurricane modification strategies.

By the early 1970s, more advanced computers permitted development of fully three-dimensional models. Figure 2 shows the hurricane circulation obtained with the first of these models. By 1984, some research groups were using three-dimensional numerical models of the hurricane and its environment covering thousands of miles; they used telescoping, or nested, grids to provide fine resolution of mesoscale features near the eye wall, those with finest grid spacing moving with the hurricane. Topics of investigation ranged from the effects of vertical wind shear on hurricane development to the detailed changes of the hurricane's structure as it made landfall. Numerical models have been developed that cover the entire Earth with sufficient resolution to simulate many of the aspects of tropical cyclones. Some sophisticated models use nested grids which allow global integrations of the equations of motion and thermodynamics with coarse resolution while simulating the detailed circulations and cloud systems of tropical cyclones with very high resolution. An operational global model in use at the European Center for Medium Range Weather Forecasting (at Reading, England) uses powerful enough computers with sufficiently small grid spacings to show useful results in anticipating hurricane development and movement without using nested grids.

Prediction. Prediction of the path of the hurricane center as a function of time is the most important aspect of the hurricane forecast. If the track prediction contains major errors, all other aspects of the forecast (intensity, structure, rainfall, and so forth) are of no consequence. Because the high-energy (damage-producing) portion of the storm has small dimensions (60–90 mi or 100–150 km), the forecaster, in landfall situations, must balance the consequences of warning too small an area and missing the landfall, against overwarning a large area and thereby causing unnecessary expensive preparations by the public. The prediction problem is made especially difficult by the sparse coverage of the oceans by conventional surface and upper-air meteorological data. Meteorological satellites provide some data over these regions. Indeed, when storms are well away from landfall, satellites are often the only source of information concerning location of the storm. For the east and Gulf of Mexico coasts of the United States, center fixes are made several times a day by aircraft penetration when the storm is within 36 h of landfall. *See* STORM DETECTION.

An array of objective models is available to the hurricane forecaster for predicting the path of the hurricane. The National Oceanic and Atmospheric Administration's National Hurricane Center in Miami continues to make regular use of at least four models. However, these methods often yield predictive results that differ by a significant amount. The forecaster is then faced with the problem of deciding which is to be given greatest weight in the official forecast.

Forecast models generally fall into three categories. There are statistical models based solely on climatology and persistence. These produce a most probable hurricane track, the only current data required being the initial location of the storm center, past movement of the center, and the calendar date. A second class of statistical models adds, to climatology and persistence, information concerning the large-scale pressure field in which the storm is embedded. The third class of models is based entirely on atmospheric dynamics. These models are similar to, though less sophisticated than, the simulation models described earlier. They differ in that the initial conditions for simulation studies are generally a simplified idealized state of the atmosphere, whereas the dynamical prediction models have initial conditions based on actual observations of the current state of the atmosphere. Generally speaking, the more advanced dynamical models perform better than statistical models for longer-range forecasts (36–48 h), while the statistical models perform better for shorter-period predictions (12–24 h). *See* MODEL THEORY.

Averaged over the Atlantic Ocean, Caribbean Sea, and Gulf of Mexico, the magnitude of the prediction error (length of the line connecting the observed position of the storm center to the predicted position) for the official forecast by the National Hurricane Center is 109, 244, and 377 mi (202, 452, and 698 km), respectively, for forecasts of 24, 48, and 72 h. From the mid-1950s through the 1960s, a decrease of about 10% was observed in the average errors of official forecasts. Since then, however, the principal measure of improvement has been a reduction in standard deviations from the mean, as larger individual errors were significantly reduced.

The notable progress in technology and of data-processing capabilities for meteorological satellites offers great promise for hurricane research and for the design of new dynamically based predictions of both development and movement. **Figure** 4, an example of the observational capabilities of modern satellites, is a three-dimensional view of the cloud systems in Hurricane Gilbert, the most powerful Atlantic hurricane of record, displaying details of cloud structure and implications for both the details of circulation and the thermodynamic properties of the hurricane inner core. Under development is an array of microwave sounders to be installed on weather satellites that will "see through" the clouds to sense the three-dimensional thermal and circulation structure of hurricanes and its interacting environment, and for measuring rainfall distributions. Near the end of the century, satellites are planned which will be able not only to measure the winds throughout the storm core but also to observe the core's thermal structure. Such observations will have a significant impact on the understanding of energetic and dynamic processes which control the degree of development and severity of a hurricane; and they may well provide data for initializing more sophisticated prediction models incorporating far more comprehensive physics than was possible with the previous models. With the increasing archives of satellite data from existing sensors, there are already promising opportunities for prediction research, applying well-known methodologies for principal component analyses (statistical procedures for analyzing data sets) and pattern recognition techniques that could reduce prediction errors for forecasts of 24 h or more, the period most critical for issuance of coast warning. *See* REMOTE SENSING; UPPER-ATMOSPHERE DYNAMICS; WEATHER FORECASTING AND PREDICTION.

Greg Holland; Joanne Simpson; Robert Simpson

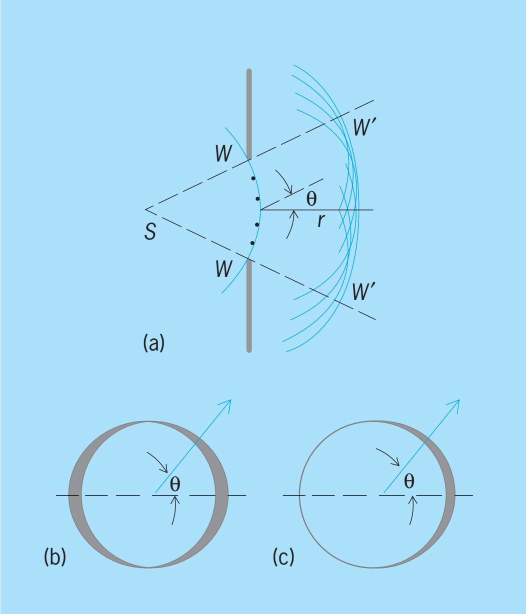
Bibliography. R. A. Anthes, *Tropical Cyclones: Their Evolution, Structure and Effects*, 1982; C. Neumann and J. M. Pelissier, An analysis of Atlantic tropical cyclone forecast errors, 1970–1979, *Mon. Weath. Rev.*, 109:1248–1266, 1981; C. Neumann and J. M. Pelissier, Models for the prediction of tropical cyclone motion over the North Atlantic: An operation evaluation, *Mon. Weath. Rev.*, 109:522–538, 1981; R. A. Pielke, *The Hurricane*, 1990; R. H. Simpson and H. Riehl, *The Hurricane and Its Impact*, 1981; H. E. Willoughby, J. Clos, and M. Shoreibah, Concentric eyewalls, secondary wind maxima, and the evolution of the hurricane vortex, *J. Atm. Sci.*, 39:395–411, 1982.

Huygens' principle

An assumption regarding the behavior of light waves, originally proposed by C. Huygens in the seventeenth century to explain the fact that light travels in straight lines and casts sharp shadows. Large-scale waves, such as sound waves or water waves, bend appreciably into the shadow. The special behavior of light may be explained by Huygens' principle, which states that "each point on a wavefront may be regarded as a source of secondary waves, and the position of the wavefront at a later time is determined by the envelope of these secondary waves at that time." Thus a wave *WW* originating at *S* is shown in **illus.** *a* at the instant it passes through an aperture. If a large number of circular secondary waves, originating at various points on *WW*, are drawn with the radius *r* representing the distance the wave would travel in time *t*, the envelope of these secondary waves is the heavily drawn circular arc *W′W′*. This represents the wave after *t*. If, as Huygens' principle requires, the disturbance is confined to the envelope, it will be 0 outside the limits indicated by points *W′*.

Careful observation shows that there is a small amount of light beyond these points, decreasing rapidly with distance into the geometrical shadow. This is called diffraction. *See* DIFFRACTION.

The Huygens-Fresnel principle, a modification of Huygens' original formulation, is capable of explaining diffraction. A. Fresnel in 1814 postulated that the amplitude of any secondary wave decreases in proportion to cos θ, when θ is the angle between the normal to the original wavefront and any point on the secondary wave (see illus. *b*, where the thickness of the arc indicates the amplitude). Fresnel then modified Huygens' requirement that the disturbance be confined to the envelope, by specifying that at any point the disturbance was the resultant of all



Huygens' principle. (*a*) The construction for a spherical wave. (*b*, *c*) Amplitude of the secondary wave according to Fresnel and Kirchhoff, respectively.

displacements due to secondary waves reaching that point. In this way Fresnel was able to explain the complicated diffraction patterns that are produced by sending light through small apertures. Subsequent theoretical investigations by G. Kirchhoff showed that the correct obliquity factor should be $1 + \cos\theta$ instead of $\cos\theta$ (illus. *c*). Approximations made by Fresnel had compensated for this. A discrepancy in the phase of the resultant wave of one-quarter period was also explained by Kirchhoff.

Francis A. Jenkins; William W. Watson

Bibliography. J. W. Goodman, *Introduction to Fourier Optics*, 3d ed., Roberts, 2005.

Hyades

A small cluster of stars that makes up the nearest well-defined open cluster (galactic cluster) to the Earth. With a total mass of about 300 suns and a population of 400–500 mostly low-mass stars, the Hyades is a typical example of the 2000 or so small star clusters in the Milky Way Galaxy. Most of its stars are located in a loose, roughly spherical system approximately 40 light-years (2.4×10^{14} mi or 3.8×10^{14} km) in diameter. It is located primarily in the constellation Taurus, the Bull, and forms the faint stellar background to the V-shaped asterism that defines the head of the Bull. *See* TAURUS.

Distance. For many years the Hyades was the backbone of the stellar distance scale in the Milky Way Galaxy and beyond. Because of its proximity, its motion through space can be detected from Earth by successive photographs taken years apart. Combining the motion in the sky of its individual stars with motions in the line of sight obtained with spectrographs, astronomers were able to determine a geometric distance to the Hyades. The technique is called the moving cluster method, and the Hyades has until recently been the only cluster close enough to use the method reliably. The galactic distance scale was established by comparing the properties of stars in other, more distant clusters with Hyades stars.

However, it was found in 1967 that certain unexpected errors had entered into the process. Evidence suggested that the distance to the Hyades (and, therefore, almost all other distances in the Milky Way Galaxy and beyond) had been underestimated by 20%. This conclusion was confirmed in 1996, when the European space telescope *Hipparcos* was able to make a more direct and more precise measure of the Hyades distance of 149 light-years (8.7×10^{14} mi or 1.4×10^{15} km). *See* ASTROMETRY; PARALLAX (ASTRONOMY).

Age. The age of the Hyades has been determined by directly comparing its stars, especially its brightest main-sequence stars (stable, long-lived stars), with detailed computer models of stars of different ages. This process must take into account the abundances of the various elements in Hyades stars which, spectroscopy indicates, are similar to the abundances in the Sun, but with all elements heavier that helium being slightly elevated. The age of the cluster is found to be approximately 6.5×10^8 years. Thus the Hyades were formed quite recently, after about 95% of the life of the Milky Way Galaxy had already occurred. *See* ASTRONOMICAL SPECTROSCOPY; STELLAR EVOLUTION.

Chemical composition. The age-measuring process must take into account the abundances of the various elements in Hyades stars. A 2006 study of a large number of individual stars using powerful new telescopes found that the stars of the Hyades are remarkably uniform in their chemical composition, all of them 45% higher than the Sun in their abundances of elements heavier than helium. *See* ASTRONOMICAL SPECTROSCOPY; TELESCOPE.

Stellar membership. The Hyades is made up mainly of normal main sequence stars, but it also contains many binary stars, a few cool giant stars, and many normal cool dwarf stars that are weak x-ray sources. The use of the *Hubble Space Telescope*, together with giant ground-based telescopes using adaptive optics, led to the detection of a population of a new kind of star, called brown dwarfs, in the Hyades. *See* ADAPTIVE OPTICS; BINARY STAR; BROWN DWARF; DWARF STAR; GIANT STAR; HERTZSPRUNG-RUSSELL DIAGRAM; HUBBLE SPACE TELESCOPE; MILKY WAY GALAXY; STAR; STAR CLUSTERS.

Paul Hodge

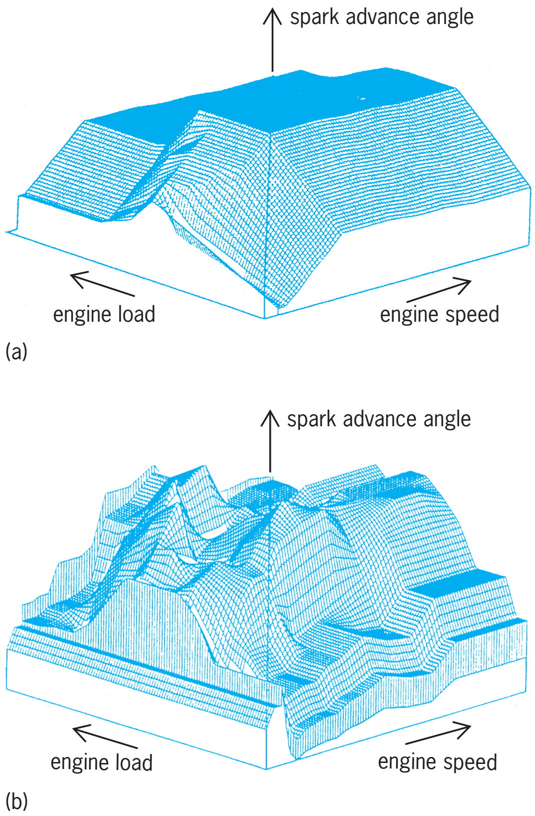
Bibliography. G. M. De Silva et al., Chemical homogeneity in the Hyades, *Astron J.*, 131:455–471, 2006; P. Hodge, How far are the Hyades?, *Sky Telesc.*, 75(2):138–140, February 1988.

Hybrid control

Control by means of systems that include both digital and analog devices working together to enhance the controller's functionality. As computers and their associated memory have become cheaper, faster, and easier to use, engineers have used them to expand the capabilities of control systems. Most controllers built today are, to some extent, hybrids because digital implementation is usually cheaper and more reliable than analog. The computer in the controller can then be used easily to provide additional capability, such as better response to internal failures. However, the potential benefits of control systems combining the symbol manipulation and decision-making capability of the computer with the precise tracking capability of analog controllers are much greater.

Automobile engine control unit. The current state of the art in hybrid control is well represented by a typical automobile engine control unit. Bosch's Motronic engine management system primarily controls ignition and fuel injection. The ignition map, basically a nonlinear function mapping engine speed and load into spark advance angle, is extremely complex in comparison to analog versions of the same controller. The complexity results from individually optimizing dozens of points on the map and then interpolating smoothly between them. This would not be feasible without the computer in the controller (see **illustration**). *See* AUTOMOTIVE ENGINE.

The rest of the Motronic engine management system includes several such complex, optimized



Ignition timing maps. (*a*) Mechanical advance system. (*b*) Electronically optimized system. (*After R. K. Jurgen, ed., Automotive Electronics Handbook, 2d ed., McGraw-Hill, 1999*)

functions. It also includes two other functions that depend on the computer. First, the controller is adaptive. The controller continually monitors certain controlled variables and changes the controller's parameters to maintain the overall system's performance despite changes in the engine's parameters. This also minimizes the need for adjustments during vehicle servicing. Second, the hybrid controller continuously monitors itself for component failures. When a failure is detected, it modifies its operation to protect other components that might be damaged by the failure, and adopts an emergency mode that allows the vehicle to still operate. The computer stores the diagnostic information and supplies it, on demand, to a repair person. Lastly, since the driver may be unaware of the problem, it provides a warning indication.

The engine management system has some decision-making capability. Another automotive example illustrates additional possibilities. The controller that deploys the airbags and seat belt pretensioners in a car is a decision-making feedback control system. The controller senses the vehicle's acceleration, either positive or negative. The analog signals from the sensors are then input to a digital system that must decide whether to deploy the seat belt pretensioners and the airbags and, if so, at what precise instants. The decision is, in principle, difficult. The vehicle is continually accelerated by the driver, by bumps in the road, by the wind, and occasionally by minor impacts. Deployment must not occur unless there is a real crash and, furthermore, it must occur at the right instant. Future systems are expected to have enhanced decision-making capability. They are expected to classify the occupant and adapt the airbag and seat belt pretensioner to the occupants and the crash severity. Current systems also self-test, a very common feature of hybrid control systems.

Incorporating humanlike decision-making into an automatic control system creates very exciting possibilities. For example, there is a great deal of current research on various types of autonomous vehicles. These include pilotless reconnaissance and combat aircraft, driverless earthmoving equipment, and various mobile autonomous robots. With the Global Positioning System (GPS) providing very accurate location sensing; the computer optimizing routing, adjusting speed, and choosing the best lane; and the analog controller maintaining speed and heading, driverless passenger vehicles are nearly feasible. The lack of a good obstacle sensor, cost, and people's willingness to trust their lives to a decision-making machine are all that stand in the way of a true "auto" mobile.

Design. The digital and analog parts of a hybrid control system have different requirements from those imposed on them in nonhybrid situations. Even though step inputs are common tests of control systems, it is generally undesirable for the inputs to a physical system to change suddenly by a large amount. For example, rapidly flooring the accelerator of a car will usually spin the wheels. Rapid changes in the parameters of an analog control system can have even worse effects and should be avoided.

Digital computers can respond almost instantly to a change in an input signal. Thus, they facilitate rapid changes. Hybrid control systems must be designed to prevent harmful jumps in signals and parameters. Alternatively, the controller can be designed to mitigate the effects of sudden changes. In either case the designer has a problem to solve.

A second design challenge results from the fact that digital computers normally act asynchronously. The exact instant at which the computer returns an answer does not matter. Analog systems operate in real time; they need a control signal at every instant. In a hybrid system the computer is constrained to respond no later than a rigid deadline imposed by the analog system. Because the computer is typically capable of many more operations per second than are needed for analog control, it is usually performing other operations, such as a self-test, in its spare time. The designer has to ensure that this does not interfere with the analog operation.

Ensuring that the controller works properly in all situations is an extremely important, and difficult, aspect of hybrid control system design. A software crash that freezes the computer is normally a nuisance; in a hybrid controller it can be a disaster. Testing and verification of the hybrid controller's operation is related to, but substantially more difficult than, the software verification problem.

Theory. Most of the theoretical work on hybrid control has been devoted to the problem of finding a single, unified description of an arbitrary hybrid system. Without such a framework it is impossible to answer many basic questions about control. For example, is the system stable? This single mathematical model must be a reasonably accurate description of an interesting class of real hybrid systems; otherwise it has only academic interest. The difficulty results from the completely different mathematics associated with the two main parts of any hybrid system.

Analog systems act continuously on continuously applied inputs; they are "time-driven." The driver must keep her foot on the accelerator pedal of her car; the car responds to any change in the pedal angle. The usual mathematical models of analog systems are differential equations or difference equations in which time is an independent variable. Digital systems are more naturally viewed as "event-driven;" they respond to discrete instantaneous inputs. The computer does nothing until the operator strikes the return key; it then executes a string of instructions, stops, and waits for the next command. The mathematical description of discrete-event (event-driven and discrete-state) systems is more like a language—if this event occurs then that event results. In many such systems and their models, time is simply ignored. When time is included, it is just an attribute of an event and not the driving independent variable.

Many mathematical models for hybrid systems have been proposed. There does not seem to be agreement on one such model. The most useful model seems to depend on the specific application. Two examples are given below. They were chosen because they are substantially less abstract that most.

One research project has concentrated on automobile engine control. Motivation includes the importance of the automobile to society and the importance of hybrid controllers to the automobile. An additional motivation is that the four-stroke single-cylinder internal combustion engine can be viewed as a discrete-event system in its own right, even without the hybrid controller. The onset of each stroke is an event. Torque generation, which depends on the stroke, is then a discrete-event system. The power train and air dynamics are described by differential equations; equivalently, they are analog systems. Naturally, the group of researchers involved in this project formulates a hybrid model for the internal combustion engine. They also pose several optimal control problems associated with this engine. A key to the solution of these problems is that they are able to "relax" the original hybrid model of the engine to a purely analog description. Finally, they map the relaxed solution back to the full hybrid engine model to obtain a suboptimal solution to the original problem.

Several aspects of the project illustrate the value of good mathematical modeling. The system specifications can be given at a high level of abstraction so that they are independent of implementation decisions. The rigorous mathematical framework ensures formal correctness of the control algorithms.

Another research project on hybrid control focuses on motion control systems, primarily in robotics. Controlling a movement in such a system requires both discrete event and analog elements. The discrete-event part is obvious—the robot moves until it touches the wall. The analog part is less obvious—it is necessary to describe and control the analog compliance that governs the interaction with the wall. A hybrid model is therefore created for movement description. This model leads to a motion description language, a basis for robot programming that facilitates specification of both the motion and its analog control. Theory such as this that simplifies the programming of hybrid controllers would greatly facilitate their development.

There are many current research projects on various aspects of hybrid control. Two large projects are HYCON (Hybrid Control), a European research consortium, and CHESS (Center for Hybrid and Embedded Software Systems), a cooperative project involving universities in the United States. At a more applied level, in 2006 the U.S. Defense Advanced Research Projects Agency (DARPA) issued a grand challenge to develop an autonomous vehicle that would successfully travel through a 60-mi (96-km) urban course at an average speed of at least 10 mi/h (16 km/h) without hitting any of the numerous obstacles in the environment. The run in which the qualifying vehicles would be expected to meet this challenge was scheduled for November 2007. The critical problem is not the vehicle but the hybrid control system. Another challenge is the RoboCup: to develop a soccer team of fully autonomous humanoid robots that can win against the human world soccer champion team by 2050. *See* CONTROL SYSTEMS; DIGITAL CONTROL; ROBOTICS. William S. Levine

Bibliography. R. J. Alur and G. Pappas (eds.), *Hybrid Systems: Computation and Control*, Springer-Verlag, Berlin, 2004; M. Egerstedt, Control of autonomous mobile robots, in D. Hristu-Varsakelis and W. S. Levine (eds.), *Handbook of Networked and Embedded Control Systems*, pp. 767–778, Birkhäuser Verlag, Basel, 2005; B. K. Mattes, Passenger safety and convenience, in R. K. Jurgen (ed.), *Automotive Electronics Handbook*, 2d ed., chap. 23, McGraw-Hill, 1999; Robert Bosch GmbH, *Automotive Handbook*, 6th ed., 2005.

Hybrid dysgenesis

A syndrome of abnormal traits that appears in the hybrids between certain strains of the fruit fly *Drosophila melanogaster*. The traits include partial sterility and greatly elevated rates of genetic mutations and chromosome rearrangements. Strains can be classified as P for paternally contributing or M for maternally contributing, so that only the hybrid sons and daughters of M females mated to P males show hybrid dysgenesis.

Cause. Hybrid dysgenesis is caused by the action of a family of transposable genetic elements, that is, segments of the genetic material (deoxyribonucleic acid, or DNA) with the special ability to move from one chromosomal site to another. Such elements range in size from a few hundred to a few thousand nucleotide pairs and typically occur in 10–100 genetic locations scattered throughout the chromosomes. Altogether, transposable elements are thought to compose 10–15% of the entire genetic complement of *Drosophila melanogaster* and are probably also common in most animal and plant species.

The family of transposable elements that causes most cases of hybrid dysgenesis is called the P family, and the individual elements are called P factors because they occur only in the paternally contributing strains. A typical P strain might have 30 to 50 factors in widely scattered chromosomal locations. The cross of P males to M females activates the transposition mechanisms of these elements, resulting often in mutations and chromosome rearrangements due to the chromosome breakage involved in the transposition process. Transposition occurs preferentially in the cell lines destined to form the gametes (the germ line), which probably accounts for the partial sterility seen in the hybrids. *See* TRANSPOSONS.

Behavior of P factors. P factors can be either complete or defective. The complete copies are 2907 nucleotide pairs long, with the first 31 bases exactly matching the last 31 bases in reverse order. This reverse repeat structure undoubtedly contributes to their ability to transpose. The defective elements differ from the complete ones by the deletion of interior sequences; the end repeats are intact in both types of factor.

The genetic information in the complete elements codes for an enzyme product (known as transposase) that gives the elements their mobility. The defective elements cannot produce their own transposase; however they can still move about if there are complete P factors present in the cell to provide the transposase.

When there is a large number of P factors present in the chromosomes of certain individuals, the transposition process stops and the number of elements does not increase further. This situation is known as the P cytotype. A likely interpretation is that the P factors produce a second gene product that regulates their own transpositional activity. These parental regulator molecules would be packaged in the egg cells of the P strains but not in the sperm cells, thus explaining why only hybrids from P males and M females have active P factors.

P factors in biotechnology. The ability of P factors to transpose at high frequencies and to be regulated by the P cytotype has made them especially useful as tools for manipulating recombinant DNA molecules. For example, P factors commonly cause mutations by inserting into genes and thus inactivating them. When this happens, the DNA of the inactivated gene can then be identified through its association with the P factor. P factors can also be used as vehicles for transporting a gene of interest from the test tube back into the organism, where its expression can be studied. *See* GENETIC ENGINEERING.

Importance of P factors in evolution. The evolutionary significance of P factors and other transposable elements is not known. One possibility is that they act as DNA parasites whose only adaptation is to maintain their presence in the chromosomes. Two arguments that strengthen this interpretation are (1) that the only known effects of P factors on the organism is the hybrid dysgenesis syndrome, which is clearly detrimental; and (2) M strains lack P factors, yet still have normal viability and fertility. However, even in the role of parasites, P factors and other transposable elements might make a positive contribution to the evolution of their hosts by additional mutations and chromosome rearrangements produced as by-products of transposition. Some of these might be favored by natural selection. In addition, the partial reproductive incompatibility between P and M strains could, in some conditions, aid in the process of splitting one species into two. Although no such instance has been observed, it is postulated that if two subspecies are already partly isolated, by geographical barriers for example, one subspecies and not the other could acquire transposable elements such as P factors. In that case, the resulting dysgenesis in some of the hybrids between the two subspecies would yield natural selection that would favor further reproductive isolation and perhaps eventual splitting into separate species. *See* GENE; MUTATION; ORGANIC EVOLUTION.

William R. Engels

Bibliography. W. R. Engels, The P family of transposable elements in *Drosophila, Annu. Rev. Genet.*, 17:315–344, 1983; M. G. Kidwell, J. F. Kidwell, and J. A. Sved, Hybrid dysgenesis in *Drosophila melanogaster*: A syndrome of aberrant traits including mutation, sterility, and male recombination, *Genetics*, 36:813–83, 1977; K. O'Hare and G. M. Rubin, Structures of P transposable elements of *Drosophila melanogaster* and their sites of insertion and excision, *Cell*, 34:25–35, 1983.

Hydatellales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Commelinidae of the class Liliopsida (monocotyledons). The order consists of a single family with five species native to Australia, New Zealand, and Tasmania. The plants are small, submersed or partly submersed aquatic annuals with greatly simplified internal anatomy. The leaves are tufted at the base of the stem, and the inflorescence is a terminal head with two to several bracts, each subtending one to several reduced, unisexual flowers. These plants have sometimes been included within the Restionales, but the structural details of the ovary and seed set them apart. They are of no economic significance. *See* COMMELINIDAE; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM; RESTIONALES. T. M. Barkley

Hydrate

A particular form of a solid compound which has water in the form of H_2O molecules associated with it. For example, anhydrous copper sulfate is a white solid with the formula $CuSO_4$. When crystallized from water, a blue crystalline solid which contains water molecules as part of the crystals is formed. Analysis shows that the water is present in a definite amount, and the hydrate may be given the formula $CuSO_4 \cdot 5H_2O$. Four of the water molecules are attached to the copper ion in the manner of coordination complexes, and the fifth water molecule is related to the sulfate and presumably held by hydrogen bonding. *See* HYDROGEN BOND.

Water can also be present in definite proportions in the crystal without being associated directly with the anion or cation. The water occupies a definite place in the crystal lattice. Alums, with their 12 molecules of water, are examples of this. *See* ALUM. Frank Wagner

Gas hydrates (gas clathrates) are crystalline compounds in which an isometric (cubic) ice (H_2O) lattice contains cages that incorporate small guest gas molecules. They are stable at moderate to high pressures and low temperatures, above and below the ice point. These ice lattices are stable only when the cages contain a gas molecule. The pressure and temperature constraints restrict them to oceanic continental margins in the uppermost few hundred meters of slope and rise sediments where water depths exceed 300–500 m, and to permafrost in polar regions. Under the ocean, the amount of gas hydrates is at least an order of magnitude higher than in permafrost.

Methane (CH_4) hydrate is the dominant natural gas hydrate on Earth, and is probably found on the planets Uranus and Neptune. One cubic meter of methane hydrate when dissociated can contain 165–180 m^3 of methane gas. The total amount of methane in gas hydrates is estimated to be very large; about 10^{19} g of methane carbon is stored in them, approximately twice that in fossil fuels. Geochemical studies of the methane, especially its carbon isotope ratios, indicate that the methane is largely biogenic in origin. But in some regions, for example, the Gulf of Mexico and Caspian Sea, its origin is thermogenic.

Gas hydrate is distributed inhomogeneously in ocean sediment (see **illus.**). It cements coarser sediments, is disseminated in silty and clay-rich sediments, and is abundant in faults or lithologic boundaries.

Methane hydrate core sample taken from the Costa Rica margin slope sediments about 200 m below the sea floor.

Recent interest in natural gas hydrates, most of which are methane hydrates, has resulted from the recognition that global warming may destabilize the enormous quantities of methane hydrate in shallow marine slope sediments and permafrost. The environmental impact of releasing large quantities of methane into the ocean and atmosphere could have important consequences. In the ocean, intense microbial oxidation of methane would reduce the amount released into the atmosphere, but would result in extensive oxygen consumption and somewhat reduce the capacity of the ocean to absorb atmospheric carbon dioxide.

Gas hydrates also affect sediment physical properties. During deglaciation or global warming, the gas hydrate may decompose at the sediment's base layer. The released gas will be trapped in the sediment, causing instability and possibly triggering giant landslides that may release methane to the ocean and atmosphere.

Methane gas is an important contributor to the atmospheric radiation balance. Any flux of methane into the atmosphere beyond the present annual rate of ~0.8% per year would enhance global warming. New evidence has been reported concerning the possible role of methane hydrate decomposition in rapid global warming, in the Santa Barbara Basin (southern California) over the past 50,000–60,000 years and in the deep ocean about 55.6 million years ago. Both episodes were accompanied by thermal maxima—times when the climate on Earth was unusually warm.

The fossil fuel resource potential of the enormous quantities of marine methane hydrates is being evaluated. *See* MARINE SEDIMENTS; METHANE; PERMAFROST. Miriam Kastner

Bibliography. B. E. Conway, *Ionic Hydration in Chemistry and Biophysics*, 1981; G. R. Dickens, M. M. Castillo, and J. G. Walker, A blast of gas in the latest Paleocene: Simulating first-order effects of massive dissociation of oceanic methane hydrate, *Geology*, 25:259–262, 1997; F. Franks (ed.), *Water: A Comprehensive Treatise*, vol. 2: *Water in Crystalline Hydrates*, 1973; V. Gornitz and I. Fung, Potential distribution of methane hydrates in the world's oceans, *Global Biogeochem. Cycles*, 8:335–347, 1999; H. Kleeberg (ed.), *Interactions of Water in Ionic and Nonionic Hydrates*, 1987; K. A. Kvenvolden, Methane hydrate—a major reservoir of carbon in the shall geosphere?, *Chem. Geol.*, 71:41–51, 1988; E. G. Nisbet and D. J. W. Piper, Giant submarine landslides, *Nature*, 392:329–330, 1998; D. D. Sloan, Jr., *Clathrate Hydrates of Natural Gases*, 2d ed., Marcel Dekker, New York, 1998.

Hydration

The incorporation of molecular water into a complex with the molecules or units of another species. The complex may be held together by relatively weak forces or may exist as a definite compound. Many salts form solid hydrates when exposed to water vapor under certain conditions of temperature and pressure. Copper sulfate, for example, forms a monohydrate ($CuSO_4 \cdot H_2O$) when exposed at 25°C (77°F) to water vapor at a pressure of 0.8 mm of mercury (1100 pascals). At higher pressures other hydrates are formed. Water is lost from these compounds when they are heated or when the water vapor pressure falls below a minimum value. Solids forming hydrates at low pressures are used as drying agents. *See* DELIQUESCENCE; DESICCANT; EFFLORESCENCE; HYDRATE; SOLUTION; SOLVATION.

Francis J. Johnston

Hydraulic accumulator

A pressure vessel which operates as a fluid source device or shock absorber. It is used to store fluid under pressure or to absorb excessive pressure increases. The hydraulic accumulator is an energy-efficient component, which allows the use of a smaller pump to achieve the same end results in terms of cylinder rod actuation speeds. In certain circuit designs, the accumulator will permit a pump motor to be completely shut down for an extended period of time while the accumulator supplies the necessary fluid to the circuit.

The operation of the hydraulic accumulator is induced by a pressurized gas (usually nitrogen), a spring, or a weighted plunger. The accumulator supplies fluid for actuator movement or to replace fluid lost by leakage. *See* SHOCK ABSORBER.

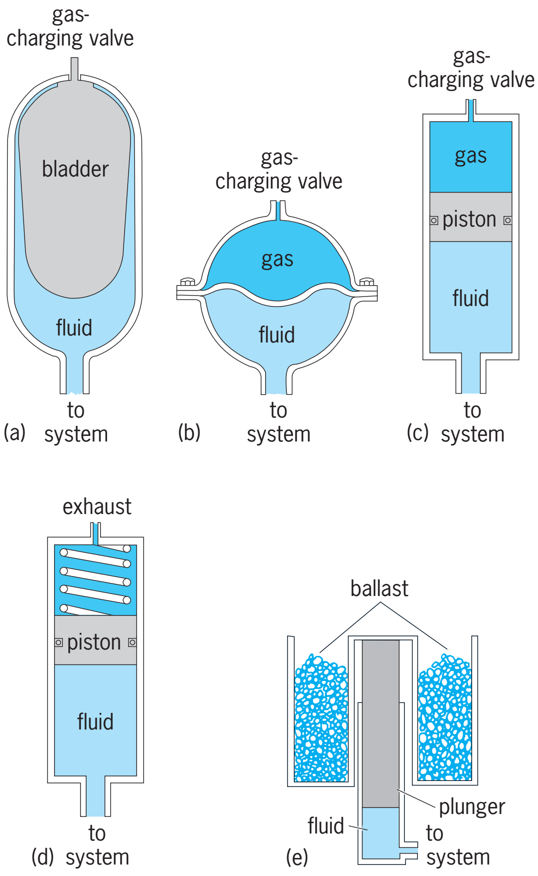
The gas-charged accumulator and the spring-type accumulator discharge their fluid into the system at pressures which are decreasing as the gas or spring expands. The weighted accumulator allows stored fluid to be discharged into the system at a constant pressure for the entirety of its downward stroke.

The gas-charged accumulator can be bladder type, diaphragm type, or piston type (see **illus.**) The bladder-type accumulator consists of a one-piece outer steel shell whose upper portion contains a bladder of elastomeric material compatible with the hydraulic fluid and charged with nitrogen gas to a pressure of 50–75% of the system pressure. The bottom of the accumulator housing is ported to allow free entry and exit of the fluid as required. When the system needs fluid from the accumulator, the nitrogen pressure expands the bladder and thus forces the fluid out into the system. As the bladder expands, its pressure is reduced, as is the system pressure.

The diaphragm-type accumulator consists of two hemispheres making up the outer shell, with the split being horizontal, and a compatible convoluted elastomeric diaphragm is bolted between the hemispheres. The upper half of the accumulator is charged with nitrogen gas, and the unit functioning is similar to that of the bag-type accumulator.

The piston-type accumulator is similar to a cylinder, except there is no piston rod. The position of the piston is determined by the input of fluid into the accumulator, or the withdrawal of fluid from it. The piston seal prevents gas from leaking into the fluid.

The spring-type accumulator is a piston-type accumulator that uses a powerful spring instead of a gas charge. *See* SPRING (MACHINES).



Diagrams showing the five types of hydraulic accumulators. (*a*) Bladder. (*b*) Diaphragm. (*c*) Piston. (*d*) Spring. (*e*) Weighted.

The weighted-type accumulator has a weight pan attached to the plunger's uppermost point. The pan carries an evenly distributed load of scrap metal whose weight is sufficient to induce the desired pressure.

The charging of the accumulator with fluid is accomplished during any period of the system cycle in which the total output of the pump is not needed for actuator function. *See* CONTROL SYSTEMS.

James E. Anders, Sr.

Bibliography. T. C. Frankenfield, *Using Industrial Hydraulics,* 1985; Parker-Hannifin Corp., *Design Engineers Handbook*, 1973; W. W. Vockroth, *Industrial Hydraulics*, 1994.

Hydraulic actuator

A cylinder or fluid motor that converts hydraulic power into useful mechanical work. The mechanical motion produced may be linear, rotary, or oscillatory. Operating pressures range from a few pounds per square inch gage (psig or lb/in.^{2} gage) to several thousand, but usually are about 500–5000 psig (3.4–34 megapascals). Sizes vary from 0.2-in.^{2} (1.3-cm^2) linear actuators capable of a few pounds of force to extremely large linear or rotary actuators capable of exerting hundreds of tons of force. Operation exhibits high force capability, high power per unit weight and volume, good mechanical stiffness, and high dynamic response. These features lead to wide use in precision control systems and in heavy-duty machine tool, mobile, marine, and aerospace applications. *See* CONTROL SYSTEMS.

Cylinder actuators. To provide a fixed length of straight-line motion, linear cylinder actuators usually consist of a tight-fitting piston moving in a closed cylinder. The piston is attached to a rod that extends from one end of the cylinder to provide the mechanical output. The double-acting cylinder (**Fig. 1**) has a port at each end of the cylinder to admit or return hydraulic fluid. A four-way directional valve functions to connect one cylinder port to the hydraulic supply and the other to the return, depending on the desired direction of the power stroke.

Some examples of the many possible types of linear hydraulic actuators are shown in **Fig. 2**. Single-acting types have a power stroke in one direction only, with the return stroke accomplished by some external means or effected by a spring. Telescopic cylinders are used to accomplish extremely long single-acting motions. A double-ended rod is employed when it is necessary to have equal forces and velocities in each direction of travel. Tandem cylinder actuators consist of two or more pistons and rods fastened together so that they operate as a unit. In this actuator a combination of low-force and high-speed piston travel can be obtained, followed by high-force and low-speed piston travel.

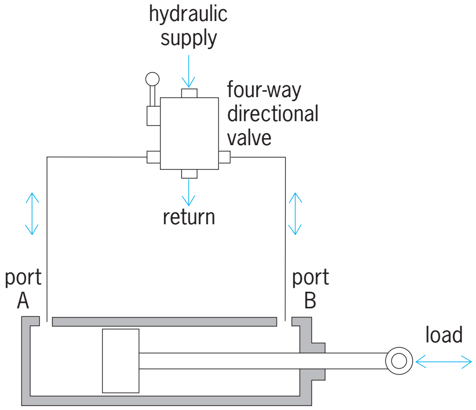


Fig. 1. Function of a hydraulic double-acting cylinder.

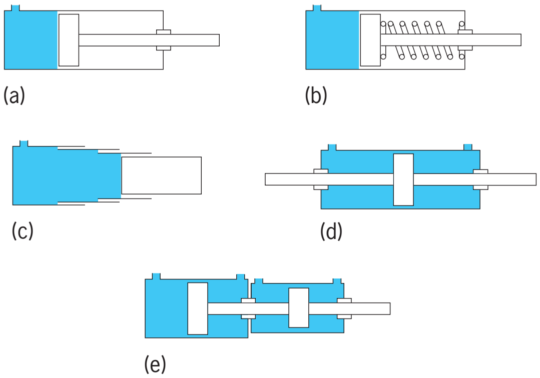


Fig. 2. Types of hydraulic cylinder (linear) actuators. (*a*) Single-acting, external return. (*b*) Single-acting, spring return. (*c*) Telescopic. (*d*) Double-acting double-ended rod. (*e*) Tandem cylinders.

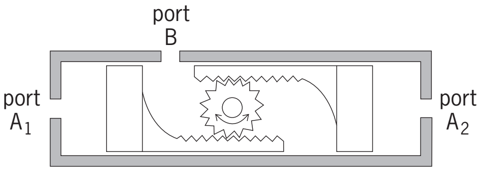


Fig. 3. Piston-rack type rotary actuator.

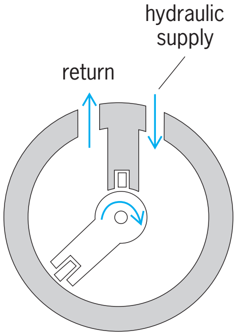


Fig. 4. Single-vane rotary actuator.

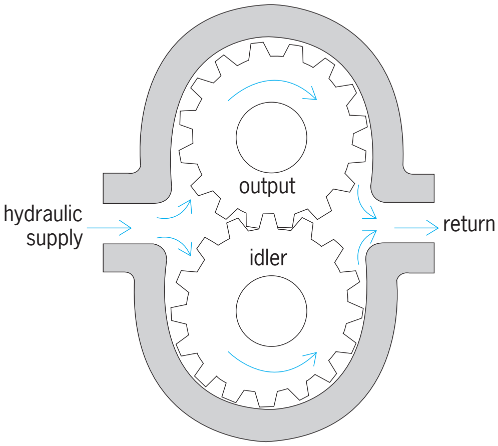


Fig. 5. Gear motor rotary actuator.

Limited-rotation actuators. For lifting, lowering, opening, closing, indexing, and transferring movements, limited-rotation actuators produce limited reciprocating rotary force and motion. Rotary actuators are compact and efficient, and produce high instantaneous torque in either direction.

In the piston-rack type of rotary actuator (**Fig. 3**), the pinion gear is attached to the load. Hydraulic fluid is applied to either the two end chambers or the central chamber to cause the two pistons to retract or extend simultaneously so that the racks rotate the pin ion gear. The single-vane type of actuator (**Fig. 4**) has a fixed stationary barrier and a rotating vane that forms two variable volume chambers. Hydraulic fluid is ported to one chamber and returned from the other to effect output shaft rotation of about 280° in either direction. A double-vane actuator has two stationary barriers and two rotating vanes, which limit rotation to about 100°. Its advantages are balanced radial loads on the output shaft

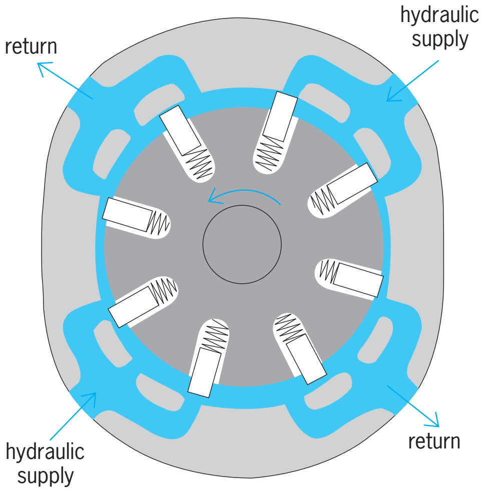


Fig. 6. Vane motor rotary actuator.

and a doubling of the torque output over that of a single-vane unit of comparable dimensions.

Rotary motor actuators. Coupled directly to a rotating load, rotary motor actuators provide excellent control for acceleration, operating speed, deceleration, smooth reversals, and positioning. They allow flexibility in design and eliminate much of the bulk and weight of mechanical and electrical power transmissions.

Motor actuators are generally reversible and are of the gear or vane type. The gear motor actuator (**Fig. 5**) has one gear connected to the output shaft. Supply fluid enters and flows around the chamber as shown, forcing the gears to rotate. The vane motor actuator (**Fig. 6**) consists of a rotor with several spring-loaded sliding vanes in an elliptical chamber. Hydraulic fluid enters the chamber and forces the vanes before it as it moves to the outlets.

Charles Mangion

Bibliography. P. Dransfield, *Hydraulic Control Systems: Design and Analysis of the Dynamics*, 1981; D. A. Pease, *Basic Fluid Power*, 2d ed., 1986; J. Prokes, *Hydraulic Mechanisms in Automation*, 1977; H. L. Stewart and J. M. Storer, *Fluid Power*, 3d ed., 1980; J. Stringer, *Hydraulic Systems Analysis*, 1976.

Hydraulic jump

An abrupt increase of depth in a free-surface liquid flow. A hydraulic jump is characterized by rapid flow and small depths on the upstream side, and by larger depths and smaller velocities on the downstream side. A jump can form only when the upstream flow is supercritical, that is, when the fluid velocity is greater than the propagation velocity c of a small, shallow-water gravity wave ($c = \sqrt{gh}$, where g is the acceleration of gravity and h is the depth). A considerable amount of energy is dissipated in the conversion from supercritical to subcritical flow. *See* OPEN CHANNEL; SURFACE WAVES.

Donald R. F. Harleman

Hydraulic press

A combination of a large and a small cylinder connected by a pipe and filled with a fluid so that the pressure created in the fluid by a small force acting on the piston in the small cylinder will result in a large force on the large piston. The operation depends upon Pascal's principle, which states that when a liquid is at rest the addition of a pressure (force per unit area) at one point results in an identical increase in pressure at all points. Therefore, in **Fig. 1**, the pressure due to the application of force F_1 is shown by Eq. (1), and the equilibrium force F_2 is shown by Eq. (2), where A_1 and A_2 are the areas

$$p = \frac{F_1}{A_1} \tag{1}$$

$$F_2 = pA_2 = F_1\frac{A_2}{A_1} \tag{2}$$

of pistons 1 and 2, respectively. The mechanical advantage is shown by Eq. (3).

$$\mathrm{MA} = \frac{F_2}{F_1} = \frac{A_2}{A_1} \tag{3}$$

The principle of the hydraulic press is used in lift jacks, earth-moving machines, and metal-forming presses (**Fig. 2**). A comparatively small supply pump creates pressure in the hydraulic fluid. The fluid then acts on a substantially larger piston to produce the action force. In this way forces greater than 15,000 tons (14,000 metric tons) are developed over the entire stroke of hydraulic presses. Heavy objects are accurately weighed on hydraulic scales in which

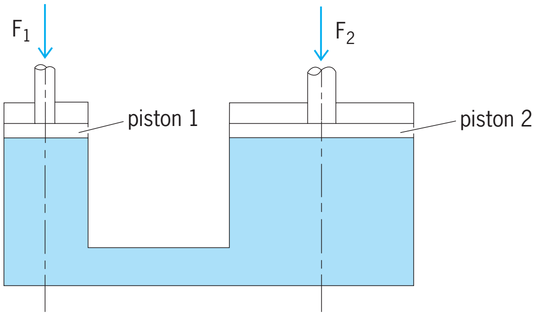


Fig. 1. Principle of hydraulic press.

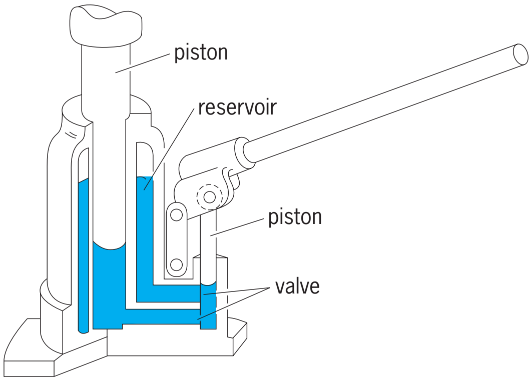


Fig. 2. Hydraulic jack.

precision ground pistons introduce negligible friction. *See* MECHANICAL ADVANTAGE; SIMPLE MACHINE.

Richard M. Phelan

Bibliography. M. Manohar and P. Krishnamachar, *Fluid Mechanics*, vol. 2: *Hydraulic Machinery and Advanced Hydraulics*, 1983; J. P. Pippenger and T. G. Hicks, *Industrial Hydraulics*, 3d ed., 1979.

Hydraulic turbine

A machine which converts the energy of an elevated water supply into mechanical energy of a rotating shaft. Most old-style waterwheels utilized the weight effect of the water directly, but all modern hydraulic turbines are a form of fluid dynamic machinery of the jet and vane type operating on the impulse or reaction principle and thus involving the conversion of pressure energy to kinetic energy. The shaft drives an electric generator, and speed must be of an acceptable synchronous value. *See* GENERATOR.

The impulse or Pelton unit has all available energy converted to the kinetic form in a few stationary nozzles and subsequent absorption by reversing buckets mounted on the rim of a wheel (**Fig. 1**). Reaction units of the Francis or the Kaplan types run full of water, submerged, with a draft tube and a continuous column of water from head race to tail race (**Figs. 2** and **3**). There is some fluid acceleration in a continuous ring of stationary nozzles with full peripheral admission to the moving nozzles of the runner in which there is further acceleration. The draft tube produces a negative pressure in the runner with the propeller or Kaplan units acting as suction runners; the Francis inward-flow units act as pressure runners. Mixed-flow units give intermediate degrees of rotor pressure drop and fluid acceleration. *See* IMPULSE TURBINE.

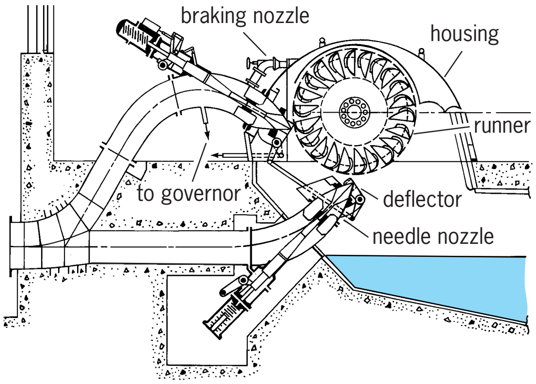


Fig. 1. Cross section of an impulse (Pelton) type of hydraulic-turbine installation.

For many years reaction turbines have generally used vertical shafts for better accommodation of the draft tube whereas Pelton units have favored the horizontal shaft since they cannot use a draft tube. Vertical-shaft Pelton units have found increasing acceptance in large sizes because of multiple jets (for example, 4–6) on a single wheel; these provide reduced runner windage and friction losses and, consequently, higher efficiency. Axial-flow (**Fig. 4**) and diagonal-flow reaction turbines offer improved hydraulic performance and economic powerhouse structures for large-capacity low-head units. Kaplan units employ adjustable propeller blades as well as adjustable stationary nozzles in the gate ring for higher sustained efficiency. Pelton units are preferred for high-head service (1000± ft or 300± m), Francis runners for medium heads (200± ft or

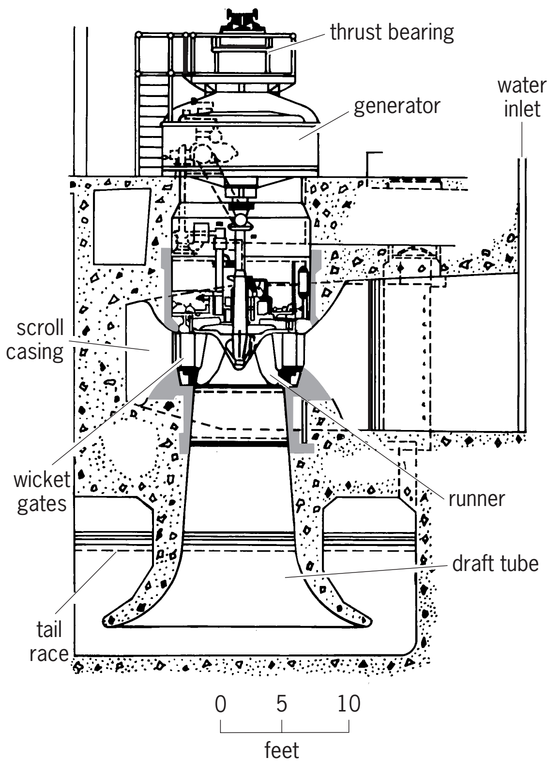


Fig. 2. Cross section of a reaction (Francis) type of hydraulic-turbine installation. 1 ft = 0.3 m.

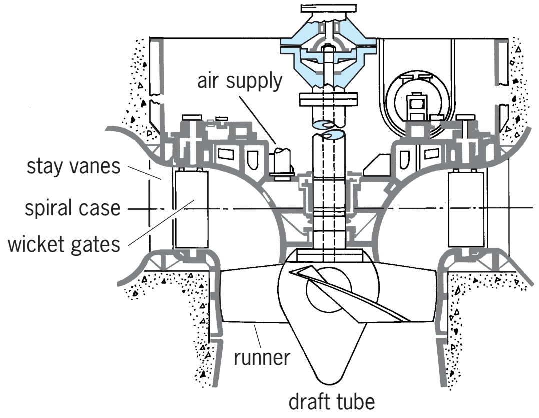


Fig. 3. Cross section of a propeller (Kaplan) type of hydraulic turbine installation.

60± m), and propeller or Kaplan units for low heads (50± ft or 15± m).

Hydraulic-turbine performance is rigorously defined by characteristic curves, such as the efficiency characteristic. The proper selection of unit type and size is a technical and economic problem. The data of **Fig. 5** are significant because they show synoptically the relationship between unit type and site head as the result of accumulated experience on some satisfactory turbine installations. Specific speed N_s is a criterion or coefficient which is uniquely applicable to a given turbine type and relates head, power, and speed, which are the basic performance data in the selection of any hydraulic turbine. Specific speed is defined in the equation below, where rpm is revolutions per

$$N_s = \frac{\text{rpm} \times (\text{shp})^{0.5}}{(\text{head})^{1.25}}$$

minute, shp is shaft horsepower (1 hp = 0.7 kW), and head is head on unit in feet (1 ft = 0.3 m). Specific speed is usually identified for a unit at the point of maximum efficiency. Cavitation must also be carefully scrutinized in any practical selection.

The draft tube (Fig. 2) is a closed conduit which (1) permits the runner to be set safely above tail water level, yet to utilize the full head on the site, and (2) is limited by the atmospheric water column to a height substantially less than 30 ft (9 m), and when made flaring in cross section will serve to recover velocity head and to utilize the full site head.

Efficiency of hydraulic turbine installations is always high, more than 85% after all allowances for hydraulic, shock, bearing, friction, generator, and mechanical losses. Material selection is not only a problem of machine design and stress loading from running speeds and hydraulic surges, but is also a matter of fabrication, maintenance, and resistance to erosion, corrosion, and cavitation pitting.

Governing problems are severe, primarily because of the large masses of water involved, their positive and negative acceleration without interruption of the fluid column continuity, and the consequent shock and water-hammer hazards. *See* PRIME MOVER.

Pumped-storage hydro plants have employed various types of equipment to pump water to an elevated storage reservoir during off-peak periods and to generate power during on-peak periods where the water flows from the elevated reservoir through hydraulic turbines. Although separate, single-purpose, motor-pump and turbine-generator sets give the best hydraulic performance, the economic burden of investment has led to the development of reversible pump-turbine units. Components of the conventional turbine are retained, but the modified pump runner gives optimum performance when operating as a turbine. Compromises in hydraulic performance, with some sacrifice in efficiency, are more than offset by the investment savings with the dual-purpose machines. *See* TURBINE; WATERPOWER.

Theodore Baumeister

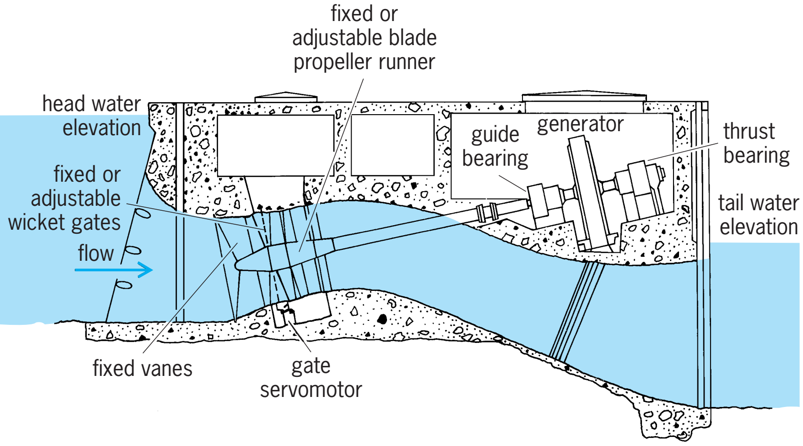


Fig. 4. **Axial-flow tube-type hydraulic-turbine installation.**

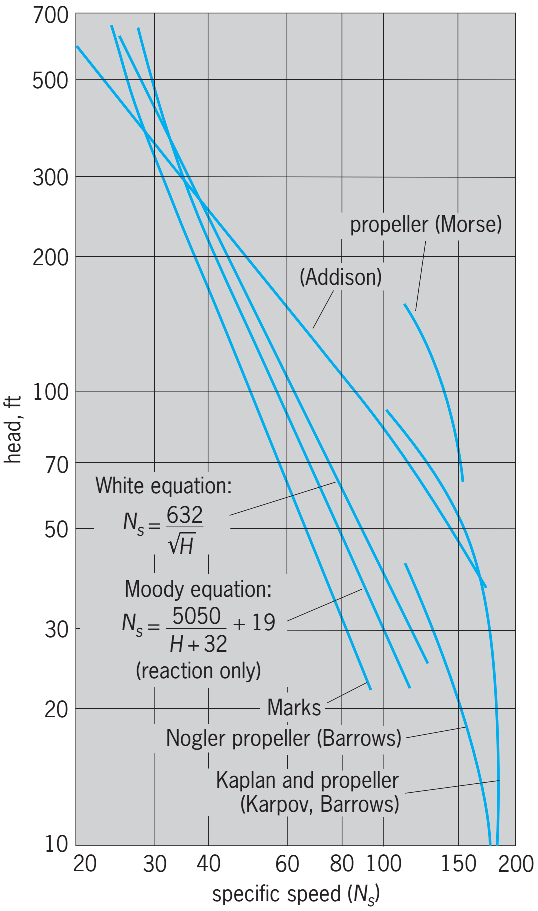


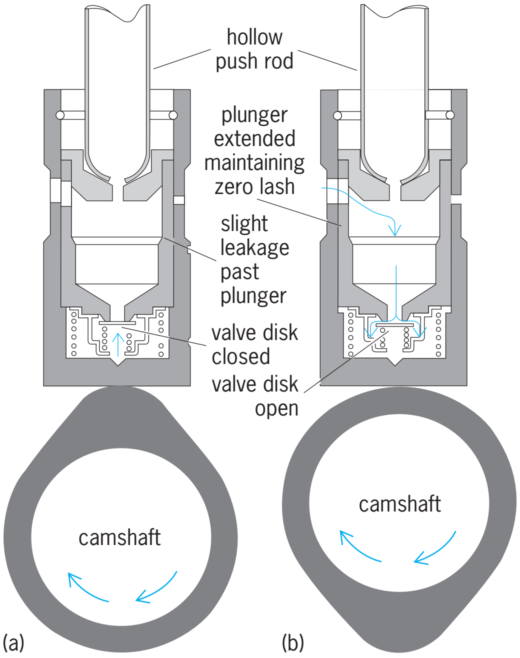
Fig. 5. **Hydraulic-turbine experience curves, showing specific speed versus head. 1 ft = 0.3 m.**

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Standard Handbook for Mechanical Engineers*, 10th ed., 1996; D. G. Fink and H. W. Beatty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999.

Hydraulic valve lifter

A device that eliminates the need for mechanical clearance in the valve train of internal combustion engines. Clearance is normally required to prevent the valve's being held open and destroyed as the valve train undergoes thermal expansion. However, clearance requires frequent adjustment and is responsible for much operating noise. The hydraulic lifter is a telescoping compression strut in the linkage between cam and valve, consisting of a piston and cylinder (see **illus.**). When no opening load exists, a weak spring moves the piston, extending the strut and eliminating any clearance. This action sucks oil into the cylinder past a check valve. The trapped oil transmits the valve-opening forces with little deflection. A slight leakage of oil during lift shortens the strut, assuring valve closure. The leakage oil is replaced as the spring again extends the strut at no load. *See* VALVE TRAIN.

Augustus R. Rogowski



Positions of the hydraulic valve lifter, with engine valve (*a*) open and (*b*) closed. (*After W. H. Crouse, Automotive Mechanics, 5th ed., McGraw-Hill, 1965*)

Hydraulics

The physical description of fluids in motion is based upon the conservation laws of mass, momentum, and energy. The mathematical statement of these laws may be approached from a highly theoretical point of view involving advanced mathematics, and the results obtained from this approach can be very. However, for practical engineering work these physical principles must be supplemented with empiricism and this semi-empirical approach is generally referred to as hydraulics. *See* CONSERVATION OF ENERGY; CONSERVATION OF MASS; CONSERVATION OF MOMENTUM.

Conservation laws. The mathematical form of the conservation laws may be obtained by considering the flow through an arbitrary closed surface drawn in the fluid, which may be either moving or fixed. This surface is called the control surface, and the volume inside it is called the control volume. If the control volume is of finite size (large), the equations governing the flow are obtained in integral form. If the control volume is of infinitesimal size (very small), the equations in differential form are obtained.

Equation of continuity. The continuity equation is the mathematical form of the law of conservation of mass. For a control surface, this law states that the net mass rate of flow out of the control volume through the control surface (mass flow rate out minus mass flow rate in) is equal to the rate at which the mass inside the control volume is decreasing. *See* EQUATION OF CONTINUITY.

For a one-dimensional steady flow (varying over space but not over time) through a fixed control volume with one inlet and one outlet, the equation of continuity becomes Eq. (1), where subscript 1 refers

$$\rho_1 A_1 V_1 = \rho_2 A_2 V_2 \tag{1}$$

to conditions at the entrance and subscript 2 to conditions at the outlet, A_1 is the area through which fluid is entering the control volume, A_2 is the area through which it is leaving, and ρ and V are the fluid density and velocity.

Momentum equation. The momentum equation is obtained by applying Newton's second law to the fluid passing through the control surface. In this case, the second law states that the sum of all the forces acting on the fluid inside the control volume must be equal to the time rate of change of the momentum of the fluid. The forces acting on the fluid include body forces such as gravity, forces that arise from pressure gradients, and viscous stresses. *See* MOMENTUM; NAVIER-STOKES EQUATION; NEWTON'S LAWS OF MOTION; VISCOSITY.

For a steady one-dimensional flow of an incompressible fluid, the momentum equation in integral form is Eq. (2), in which $\Sigma \boldsymbol{F}$ is the resultant force

$$\Sigma F = \rho Q(V_2 - V_1) \tag{2}$$

vector acting on the fluid inside the control volume, ρ is the fluid density, Q is the volume rate of flow, and V_1 and V_2 are the vector velocities of the fluid entering and leaving the control volume.

Energy equation. The energy conservation principle states that the rate at which work is done to the fluid inside the control volume is equal to the rate of change of the energy of the fluid. The energy consists of kinetic and potential forms. *See* ENERGY; THERMODYNAMIC PRINCIPLES.

For steady flow of an incompressible fluid, the integral form of the energy equation is Eq. (3), where

$$\frac{p_1}{\rho g} + \frac{V_1^2}{2g} + z_1 = \frac{p_2}{\rho g} + \frac{V_2^2}{2g} + z_2 + h_s + h_f \tag{3}$$

ρ is the fluid density, g is the acceleration of gravity, h_f is the energy-loss term due to internal fluid friction, the subscripts 1 and 2 refer to the entrance and exit ports of the control volume, p and V are the fluid pressure and velocity, and z is the vertical coordinate. The term h_s is the shaft work of a pump or turbine per unit weight leaving the control volume. This equation is frequently referred to as the modified Bernoulli equation. *See* BERNOULLI'S THEOREM; FLUID-FLOW PRINCIPLES; FLUID MECHANICS; GAS DYNAMICS.

Applications. Applications of hydrodynamics include the study of closed-conduit and open-channel flow, and the calculation of forces on submerged bodies.

Closed-conduit flow. Flow in closed conduits, or pipes, has been extensively studied both experimentally and theoretically. If the pipe Reynolds number, given

$$\mathrm{Re}_D = \frac{VD\rho}{\mu} \qquad (4)$$

by Eq. (4), where V is the average velocity and D is the pipe diameter, is less than about 2000, the flow in the pipe is laminar. In this case, the solution to the continuity, momentum, and energy equations is readily obtained, particularly in the case of steady flows. If Re_D is greater than about 4000, the flow in the pipe is turbulent, and the solution to the continuity, momentum, and energy equations can be obtained only by employing empirical correlations and other approximate modeling tools. The Re_D region between 2000 and 4000 is the transition region in which the flow is intermittently laminar and turbulent. *See* LAMINAR FLOW; REYNOLDS NUMBER; TURBULENT FLOW.

The loss term in Eq. (3) is given by Eq. (5), where

$$h_f = f\frac{L}{D}\frac{V^2}{2g} \qquad (5)$$

L is the length of pipe between stations 1 and 2, and f is the friction factor. For laminar flow, the friction factor can be calculated exactly, and is given by Eq. (6). For turbulent flow, f is a function of both Re_D and

$$f_{\mathrm{lam}} = \frac{64}{\mathrm{Re}_D} \qquad (6)$$

the relative roughness of the pipe wall, and values of f are obtained from the Colebrook formula, Eq. (7), or various other semiempirical correlations. Here,

$$\frac{1}{f^{1/2}} = -2.0 \log\left(\frac{\epsilon/D}{3.7} + \frac{2.51}{\mathrm{Re}_D f^{1/2}}\right) \qquad (7)$$

ϵ is the absolute internal surface roughness, whose values have been compiled for various materials. If the pipe has valves and other pipe fittings, the loss due to each of these is given by Eq. (8),where K is

$$h_f = K\frac{V^2}{2g} \qquad (8)$$

the loss coefficient for the particular fitting. Values of K are determined experimentally and have been compiled. The force exerted by the fluid on bends, elbows, and reducers is calculated by using the momentum equation. *See* PIPE FLOW.

Open-channel flow. Confined flows that have a liquid surface exposed to the atmosphere (a free surface) are called open-channel flows. Flows in rivers, canals, partially full pipes, and irrigation ditches are examples. The difficulty with these flows is that the shape of the free surface is one of the unknowns to be calculated.

A special case that is sometimes a good approximation is steady, uniform flow. For this type of flow, the free surface of the liquid is parallel to the channel bottom, and the channel discharge is given by Eq. (9), where Q is the volume rate of flow, A is the

$$Q = CA(R_h S_0)^{1/2} \qquad (9)$$

cross-sectional area of the flow, C is the Chézy coefficient, R_h is the hydraulic radius (sometimes called the hydraulic mean depth), and S_0 is the slope of the channel bottom. The hydraulic radius R_h is the ratio of the cross-sectional area of the flow to the wetted perimeter, which is the perimeter of the channel cross section in contact with the fluid. The Chézy coefficient is easily obtained from Manning's formula (10), in which n is the boundary roughness factor,

$$C = \frac{1.49}{n}[R_h(\mathrm{ft})]^{1/6} = \frac{1.0}{n}[R_h(\mathrm{m})]^{1/6} \qquad (10)$$

whose values have been compiled for various materials.

In most open-channel flows, however, the bottom slope and the water depth change with position, and the free surface is not parallel to the channel bottom. If the slopes are small and the changes are not too sudden, the flow is called a gradually varied flow. An energy balance between two sections of the channel yields a differential equation for the rate of change of the water depth with respect to the distance along the channel. The solution of this equation, which must be accomplished by using one of many available numerical techniques, gives the shape of the water surface.

Flow over spillways and weirs and flow through a hydraulic jump are examples of rapidly varying flows. In these cases, changes of water depth with distance along the channel are large. Here, because of large accelerations, the pressure distribution with depth may not be hydrostatic as it is in the cases of gradually varied and uniform flows. Solutions for rapidly varying flows are accomplished by using approximation techniques. *See* HYDRAULIC JUMP; OPEN CHANNEL.

Forces on submerged bodies. The force exerted by a fluid flowing past a submerged body is in principle calculated by integrating the pressure distribution over the surface of the body. This force is resolved into two components, the lift and the drag. The drag force is the component parallel to the velocity U of the undisturbed stream (flow far away from the body), and the lift force is the component perpendicular to the undisturbed stream. These force components are related to the undisturbed stream velocity by Eqs. (11), where C_L and $C_D$are the lift and

$$\begin{aligned} \text{Lift} &= C_L A \frac{\rho U^2}{2} \\ \text{Drag} &= C_D A \frac{\rho U^2}{2} \end{aligned} \qquad (11)$$

drag coefficients, respectively. the values of C_d and C_l have been measured and compiled for many shapes ranging from spheres to airfoil sections. For nonlifting bodies, the area A in the drag formula is the projected area of the body on a plane normal to the flow. For lifing bodies A is the planform area. C_L and C_D depend upon the Reynolds number, surface roughness, and the orientation of the object to the stream. *See* BOUNDARY-LAYER FLOW.

Warren M. Hagist

Bibliography. D. J. Acheson, *Elementary Fluid Dynamics*, 1990; V. T. Chow, *Open Channel Hydraulics*, 1959; L. M. Milne-Thomson, *Theoretical Hydrodynamics*, 5th ed., 1996; H. Schlichting, *Boundary Layer Theory*, 8th ed., 2000; F. M. White, *Fluid Mechanics*, 4th ed., 1998; F. M. White, *Viscous Fluid Flow*, 2d ed., 1991; E. B. Wylie, V. L. Streeter, and L. Suo, *Fluid Transients in Systems*, 1993.

Hydrazine

A colorless liquid, H_2NNH_2 (boiling point 114°C or 237°F), with a musty, ammonialike odor. Physically hydrazine is similar to water, but chemically it is reducing, decomposable, basic, and bifunctional. Its derivatives range from simple salts to ring compounds, polymers, and coordination complexes. Major uses of hydrazine include such diverse applications as rocket fuels (since combustion of hydrazine is highly exothermic), corrosion inhibition in boilers, and syntheses of biologically active materials.

Hydrazine is manufactured by two routes: the reaction of chloramine with ammonia and the reaction of sodium hypochlorite with urea. Both processes require the presence of glue or gelatin to inhibit catalytic decomposition of the product by unreacted oxidants. Because hydrazine forms an azeotrope containing 31% water (boiling point 121°C or 250°F), anhydrous hydrazine is isolated from aqueous process streams by extractive distillation with aniline.

Added to feedwater, hydrazine reduces rust in boilers to a hard film of magnetic iron oxide and reduces oxygen to water on catalytic metal surfaces. Metal ions such as Cu^{2+} and Ni^{2+} are reduced to free metals, and organic nitro compounds are reduced to amines by hydrazine. The energetic reaction of hydrazine with strong oxidants, such as nitric acid, is utilized in rocket propulsion. The thermal decomposition of hydrazine produces free radicals and gases, useful in rubber curing and foam-rubber production.

A slightly weaker base than ammonia, hydrazine forms most of the analogs of ammonia derivatives as well as distinctive hydrazine derivatives in which both nitrogen atoms are involved. For example, hydrazine forms two series of salts, such as $N_2H_4 \cdot HCl$ and $N_2H_4 \cdot 2HCl$, and forms not only hydrazones, $RCH{=}NNH_2$, but also azines, $RCH{=}NN{=}CHR$, by reaction with aldehydes. In a manner similar to ammonia, hydrazine attacks polar bonds, in one case displacing ammonia from urea to form semicarbazide, a reaction which can be reversed by an excess of ammonia.

Prominent uses of hydrazine derivatives include rocket fuels [1,1-dimethylhydrazine, $(CH_3)_2NNH_2$]; antituberculin drugs (isonicotinic hydrazide, $C_5H_4N \cdot CO \cdot NHNH_2$); plant-growth regulators (maleic hydrazide, $NH \cdot CO \cdot CH{=}CH \cdot CO \cdot NH$, and β-hydroxyethyl hydrazine, $HOCH_2CH_2NHNH_2$); dye and explosive intermediates [aminoguanidine, $NH_2 \cdot C(NH) \cdot NHNH_2$]; algacides and fungicides [copper dihydrazinium sulfate, $CuSO_4 \cdot (N_2H_5)_2SO_4$]; soldering fluxes (hydrazine hydrobromide, $N_2H_4 \cdot HBr$); blowing agents for foam rubber (azides $R \cdot CO \cdot N_3$, and sulfonyl hydrazides, $R \cdot SO_2 \cdot NHNH_2$); insecticides (1,4-diphenylsemicarbazide, $C_6H_5 \cdot NH \cdot CO \cdot NHNH \cdot C_6H_5$); heterocycle syntheses (thiosemicarbazide, $NH_2 \cdot CS \cdot NHNH_2$) and polymers (dihydrazide-formaldehyde resins). *See* AMMONIA; NITROGEN.

Theodore H. Dexter

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, vol. 12, 3d ed., 1980.

Hydride

The isolated atomic hydrogen anion, H^-. It consists of a singly charged positive nucleus and two electrons. The electron-electron repulsion almost overwhelms the nuclear-electron attraction. Thus, the "extra" electron is held weakly and is readily donated. Ionic salts containing this large and easily polarized ion are highly reactive, strongly basic, and powerfully reducing. This makes them important reagents despite the fact that they are readily destroyed by the presence of the relatively acidic compound water (H_2O) or by exposure to the relatively oxidizing dioxygen (O_2) as found in air. *See* ELECTRON CONFIGURATION.

The term hydride also refers to salts containing the H^- anion and a highly electropositive alkali or alkaline-earth metal as the cation. The salt names reflect this high ionic character, for example, sodium hydride (NaH). In such salts the ionic radius of H^- is comparable to that of Cl^-.

There are also complex metal hydrides that are formed from the formal reaction of H^- salts with some more covalent metal or metalloid hydrogen compound. Among the earliest to be investigated were lithium aluminum hydride ($LiAlH_4$) and sodium borohydride ($NaBH_4$). These two species also have numerous derivatives in which one, two, or three hydrogens have been replaced by other univalent groups. All of these species formally transfer the H^- anion to suitable organic hydride acceptors. In these and related reactions, all the components of H^- (that is, two electrons and one proton) are transferred, although the precise electronic details depend on the molecular details.

Ideally, the term hydride should be reserved for those species that contain H^- or that at least formally transfer this ion to another substance in a so-called hydride transfer reaction. Such reactions are found in the industrial synthesis of 2,2,4-trimethylpentane (isooctane) from isobutylene and isobutane, as well as in many of the classical organic chemistry named reactions, for example, the Cannizzaro reaction. Hydride transfer is also important in most of the biologically important oxidation-reduction reactions of the vitamin niacin (nicotinamide) as found in the forms of nicotinamide adenine dinucleotide/hydrogenated nicotinamide adenine dinucleotide (NAD^+/NADH) and nicotinamide adenine dinucleotide phosphate ($NADP^+$)/NADPH. *See* NIACIN; NICOTINAMIDE ADENINE DINUCLEOTIDE (NAD); NICOTINAMIDE ADENINE DINUCLEOTIDE PHOSPHATE (NADP).

There are compounds with metal-hydrogen bonds that are also referred to as hydrides. For example, hydridocarbonyl [$HCo(CO)_4$] and dihydridotetracarbonyl iron [$H_2Fe(CO)_4$] are often named cobalt tetracarbonyl hydride and iron tetracarbonyl dihydride by analogy to corresponding halides. However, both species are quite acidic. These carbonyl hydrides and derived anions are important for both the academic and industrial chemist. So are other metal-hydrogen complexes with so-called soft ligands. The presence of two or more hydrogens in a metal complex does not necessarily convey two metal-hydrogen bonds of whatever polarity, but may be a complex of molecular hydrogen instead. *See* COORDINATION COMPLEXES; HYDRIDO COMPLEXES; HYDROGEN; LIGAND; METAL CARBONYL; METAL HYDRIDES.

Joel F. Liebman

Bibliography. F. A. Cotton and G. Wilkinson, *Advanced Inorganic Chemistry*, 6th ed., 1999; N. N. Greenwood and A. Earnshaw, *Chemistry of the Elements*, 2d ed., 1997; M. Smith and J. March, *March's Advanced Organic Chemistry*, 5th ed., 2000.

Hydrido complexes

Complex hydrides containing a hydride ligand bonded to a central atom. The prefix hydro instead of hydrido is sometimes used. Sodium tetrahydridoborate, $NaBH_4$ (or sodium tetrahydroborate, originally called sodium borohydride), and lithium tetrahydridoaluminate, $LiAlH_4$ (originally called lithium aluminum hydride), are important reducing agents in synthetic and industrial reactions. $NaBH_4$ is employed in aqueous or alcoholic solutions, and $LiAlH_4$ is employed in ethers. Sodium cyanotrihydridoborate, $NaBH_3CN$, can be used in acidic medium. A family of aluminum-based reducing agents is now available commercially, including sodium diethyldihydridoaluminate, $NaAlH_2(C_2H_5)_2$; sodium tri-*tert*-butoxohydridoaluminate, $NaAlH(O\text{-}t\text{-}C_4H_9)_3$; and sodium bis(2-methoxyethoxo)-dihydridoaluminate, $NaAlH_2(OCH_2CH_2OCH_3)_2$. All are soluble in aromatic hydrocarbons. They are rather expensive, but their specific reducing powers make them attractive for synthesizing high-value products, such as pharmaceuticals, flavorings, fragrances, dyes, and insecticides.

Similar hydrides are C_6H_5MgH and $C_6H_5Mg_2H_3$. If zinc is considered a nontransition metal, then complex hydrides such as Li_3ZnH_5, Li_2ZnH_4, $LiZnH_3$, $NaZn_2H_5$, $LiZnH(CH_3)_2$, and $LiZn(CH_3)_2AlH_4$ may be grouped with the above compounds. *See* HYDROBORATION; METAL HYDRIDES.

Transition metals. More than a thousand hydrido complexes of transition metals have been prepared. Excepting Sc, Y, and La, hydrido compounds of all three transition series metals are known. In only two cases do the complex anions contain the central atom bonded to hydride ions and no other ligand; these are K_2ReH_9 and K_2TcH_9 (see below). The rest contain H-M-L linkages, where hydrogen is bonded to a transition metal M which is also bonded to one or more pi-bonded ligands L [for example, CO, CN^-, PF_3, or $P(C_6H_5)_3$]. The metal-hydrogen bonds are stabilized by the pi bonding between the metal and other ligands. The M-H bond lengths are generally 0.16–0.17 nanometer. These compounds are exceedingly diverse in nature and undergo a multitude of reactions. A few industrial processes involve hydrido complexes. Some chemists suspect that the enzymes which convert atmospheric nitrogen to ammonia (fixation) depend on Fe-H or Mo-H functions. *See* COOR-DINATION CHEMISTRY; COORDINATION COMPLEXES.

Enneahydridorhenate ion. A compound of formula K_2ReH_9 was isolated in 1960. Brackets are frequently employed to identify complex ions or molecules, for example, $K_2[ReH_9]$. The structure of the enneahydridorhenate ion, $ReH_9{}^{2-}$, is that of a triangular prism with three extra hydrogen atoms bonded through three faces, as shown in **Fig. 1**. One of these facial hydrogen atoms can be replaced by a phosphine molecule, the compound $K[ReH_8PH_3]$ being formed. The technetium compound analogous to $K_2[ReH_9]$, namely $K_2[TcH_9]$, is the second transition metal hydrido complex known which has only hydride ligands. Both compounds are fairly stable in alkaline aqueous solution.

Carbonyl hydrides. Hydrido complexes of metals also bonded to carbonyl ligands were the first such compounds discovered in 1931. The iron

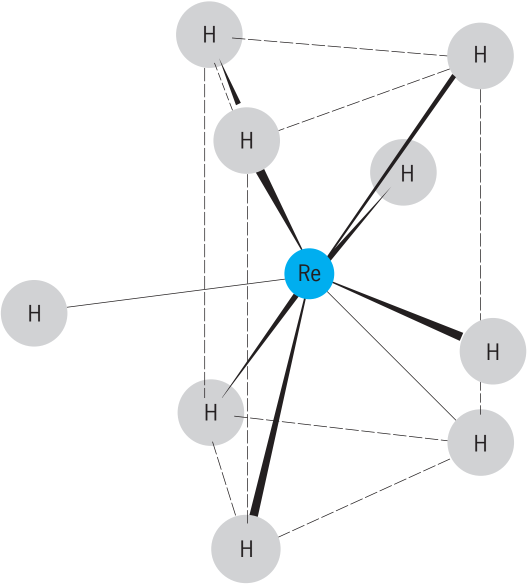


Fig. 1. Structure of $ReH_9{}^{2-}$.

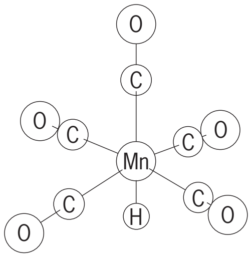


Fig. 2. Structure of $[HMn(CO)_5]$.

compounds were prepared by the action of an alkali on iron pentacarbonyl as shown in reaction (1).

$$[Fe(CO)_5] + 3OH^- \longrightarrow [HFe(CO)_4]^- + CO_3{}^{2-} + H_2O \quad (1)$$

Acidification of the product from (1) yields dihydridotetracarbonyliron, $[H_2Fe(CO)_4]$. The corresponding cobalt compound is prepared by hydrogenation of dicobalt octacarbonyl as shown in reaction (2).

$$[Co_2(CO)_8] + H_2 \rightleftharpoons 2[HCo(CO)_4] \quad (2)$$

Some other hydridocarbonyls are $[HMn(CO)_5]$ (**Fig. 2**), $[HRe(CO)_5]$, and $[HV(CO)_6]$. The effective atomic number rule is valid in these cases. These compounds are acidic, and many derivatives are known, such as $[(CO)_4Co—Hg—Co(CO)_4]$, a substance containing metal-metal bonds. The hydridocarbonyls $[HCo(CO)_4]$ and $[HV(CO)_6]$ are strong acids, while $[H_2Fe(CO)_4]$, with $pK_1 = 4.4$, is about as acidic as acetic acid. All of these hydridocarbonyls are volatile, toxic compounds with obnoxious odors. Some hydridocarbonyls are listed below.

$[HZrCl(C_5H_5)_2]$
$[H_2Zr(C_5\{CH_3\}_5)_2]$
$[H_3Ta(C_5H_5)_2]$
$[HCr(C_5H_5)(CO)_3]$
$[H_2Mo(C_5H_5)_2]$
$[H_3Mo(C_5H_5)_2]^+$
$[H_2W(C_5H_5)_2]$
$[HW(C_5H_5)(CO)_3]$
$[HMn(CO)_5]$
$[HMn(PF_3)_5]$
$[HRe(C_5H_5)_2]$
$[H_2Re(C_5H_5)_2]^+$
$[HFe(CO)_4]^-$
$[HFe(C_5H_5)(CO)_2]^+$
cis-$[H_2Fe(PF_3)_4]$
$[HCo(CO)_4]$
$[HCo(PF_3)_4]$
$[HCo(CO)_3P(C_6H_5)_3]$
$[H_2Co(P\{C_6H_5\}_3)]_3]$
$K_3[HCo(CN)_5]$
$[HRu(C_5H_5)_2]^+$
$[HRu(C_5H_5)(CO)_2]$
$[HRuCl(P\{C_6H_5\}_3)_3]$
$[HPtCl(P\{C_2H_5\}_3)_2$

Knowledge of the hydrogen-metal bond in these complexes derives primarily from neutron and x-ray diffraction studies, nuclear magnetic resonance studies, and infrared absorption spectra. Proton (hydride) nuclear magnetic resonance shifts are exceptionally large, and to high field. *See* INFRARED SPECTROSCOPY; NEUTRON DIFFRACTION; NUCLEAR MAGNETIC RESONANCE (NMR); X-RAY DIFFRACTION.

The hydridocarbonyls are chemically reactive substances, as illustrated by olefin and carbon monoxide insertion reactions of the cobalt compound, reaction (3). The last compound reacts with hydrogen to regenerate $[HCo(CO)_4]$, forming an aldehyde, as shown in reaction (4). Thus hydridotetracarbonylcobalt functions as a catalyst. Further hydrogenation of the aldehyde leads to an alcohol. The above sequence of reactions is known as hydroformylation or the oxo reaction, and is employed industrially to produce long-chain alcohols. *See* HYDROFORMYLATION; METAL CARBONYL.

$$[HCo(CO)_4] \xrightarrow{H_2C=CH_2} [CH_3CH_2Co(CO)_4] \xrightarrow{CO} [CH_3CH_2COCo(CO)_4] \quad (3)$$

$$[CH_3CH_2COCo(CO)_4] \xrightarrow{H_2} [HCo(CO)_4] + CH_3CH_2CHO \quad (4)$$

Hydridocyclopentadienyl complexes. The hydrocarbon cyclopentadiene, C_5H_6, readily loses one hydrogen ion to form the cyclopentadienide ion, $C_5H_5{}^-$. The iron(II) derivative, $[Fe(C_5H_5)_2]$, is bis(cyclopentadienyl)iron, the famous sandwich compound, ferrocene.

In 1955 hydrido-bis(cyclopentadienyl)rhenium, $[HRe(C_5H_5)_2]$, was synthesized, and nuclear magnetic resonance studies proved that the hydrogen atom is bonded to the metal (**Fig. 3**).The compound is a Brönsted base, and neutralized dilute acids to form $[H_2Re(C_5H_5)_2]^+$. In 60% dioxane the value of pK_b for $[HRe(C_5H_5)_2]$ is 8.5; pK_b for ammonia under the same conditions is 8.85. It is believed that in $[HRe(C_5H_5)_2]$ and similar compounds there are three orbitals directed from the metal atom; electrons in these orbitals cause the two C_5H_5 rings to be angular rather than parallel. In $[H_2Re(C_5H_5)_2]^+$, hydrogen atoms are bonded through two of these orbitals, and all three are occupied in $[H_3Ta(C_5H_5)_2]$ (which is not a base) and in $[H_3W(C_5H_5)_2]\ ^+$. Mixed ligand complexes such as $[HCr(C_5H_5)(CO)_3]$ were also discovered in 1955.

When attempts were made to prepare $[HM(C_5H_5)_2]$, where M is Co, Rh, or Ir, it was found that the hydrogen atom bonds, not to the metal atom, but to one of the five-carbon rings, resulting in $[M(C_5H_5)(C_5H_6)]$, which is not a hydrido complex.

Hydridophosphine and hydridocyano complexes. Trifluorophosphine (PF_3), triethylphosphine $[P(C_2H_5)_3]$ and triphenylphosphine $[P(C_6H_5)_3]$ are important ligands in hydrido-transition metal complexes. Examples of the first type are the colorless liquids $[HMn(PF_3)_5]$, *cis*-$[H_2Fe(PF_3)_4]$, and $[HCo(PF_3)_4]$, which resemble the corresponding carbonyls. The structures of $[HRuCl(P\{C_6H_5\}_3)_3]$ and of $[H_3Ir(P\{C_6H_5\}_3)_3]$ are sketched in **Fig. 4**. Within a given periodic family the stability of a given type of hydride increases as shown by the following trend: $[HNiCl(P\{C_2H_5\}_3)_2]$ can be detected but not isolated.

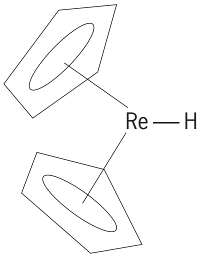


Fig. 3. Structure of $[HRe(C_5H_5)_2]$. The aromatic C_5H_5 groups are represented by pentagons.

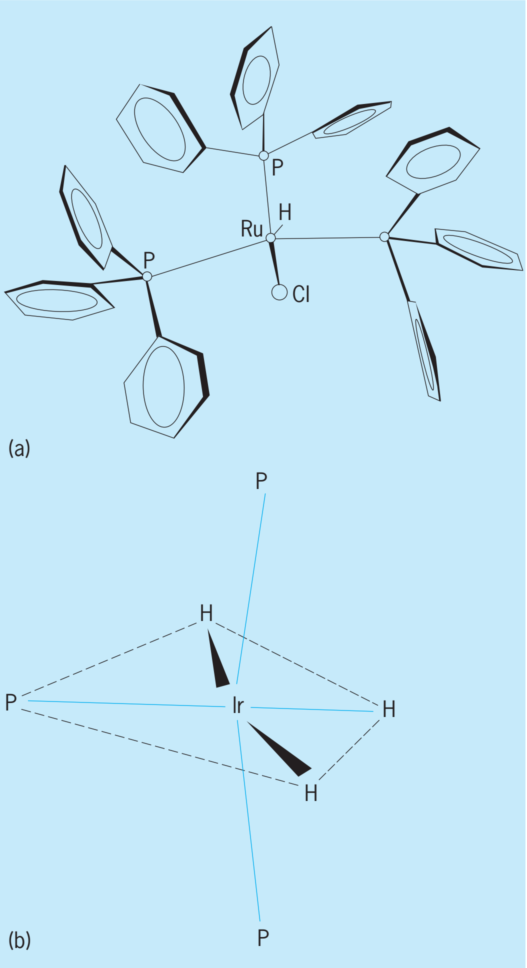
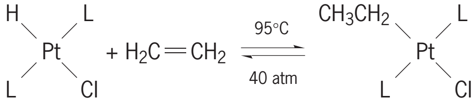


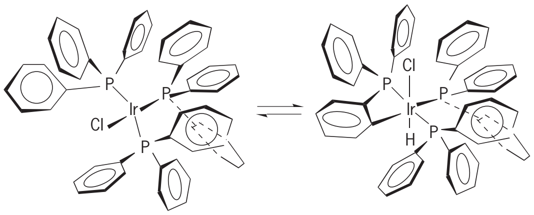
Fig. 4. Hydridophosphine structures, with the C_6H_5 groups not shown. (*a*) $[HRuCl(P\{C_6H_5\}_3)_3]$. (*b*) $[H_3Ir(P\{C_6H_5\}_3)_3]$.

$[HPdCl(P\{C_2H_5\}_3)_2]$ can be isolated, but is unstable. $[HPtCl(P\{C_2H_5\}_3)_2]$ is stable in air even at 100°C (212°F), and can be vacuum-distilled. Certain hydrido complexes of the platinum metals function as hydrogenation catalysts at room temperature and 1 atm (100 kilopascals) pressure. The catalysis is attributed to the ability of the metal atom to coordinate, via a vacant site, the reactant molecules, thereby orienting them and lowering the activation energy for bond making and breaking. The transition metal atom can supply or accept electrons as necessary. The hydrogen atom is facile, and under these circumstances is inserted into the substrate molecule. One well-known compound of this type is chloro-tris(triphenylphosphine)rhodium(I), Wilkinson's catalyst. Its role in catalyzing the hydrogenation of an olefin (RCH═CH$_2$) is shown in **Fig. 5**.

The mobility of the hydrido ligand is further shown by the reversible insertion of ethylene into an H-Pt bond. This is shown in reaction (5), where the ligand L is $P(C_6H_5)_3$. The conversion of orange chloro-tris (triphenylphosphine)iridium(I) to its colorless hydrido isomer is shown inreaction (6). Hydrido(tri-*n*-butylphosphine) copper(I), $[HCuP(C_4H_9)_3]$, which might be a polymer, is a mild, selected reducing agent employed in organic syntheses. It and similar complexes are very unstable and must be employed below −20°C (−4°F).



(5)



(6)

Reversible hydrogenation of the cyanocobalt(II) complex yields the hydridocyanocobalt(III) derivative, shown in reaction (7). Hydrido complexes of

$$2[H_2OCo(CN)_5]^{3-} + H_2 \rightleftharpoons [HCo(CN)_5]^{3-} + 2H_2O \qquad (7)$$

this type are catalysts in selectively reducing diene hydrocarbons such as butadiene, C_4H_6, to the monoene stage, C_4H_8.

Polynuclear hydrido complexes. Reduction of Cr-(CO)$_6$ with NaBH$_4$ yields the yellow binuclear

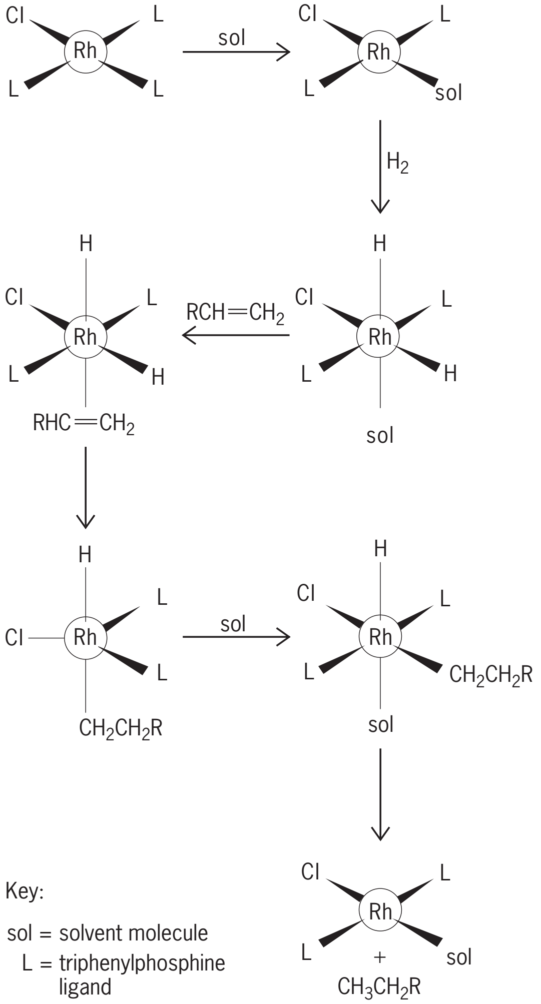


Fig. 5. Reaction scheme for hydrogenation of an olefin using Wilkinson's catalyst.

(two-chromium) anion, $[HCr_2(CO)_{10}]^-$. The earliest structural studies of this ion indicated that the hydrogen atom lies on the Cr-Cr axis, bonding the two $Cr(CO)_5$ units together. More recently it was demonstrated that the Cr-H-Cr linkage is inherently bent, as indicated in **Fig. 6**. The hydrogen atom is about 0.03 nm from the center of the Cr-Cr bond. In $[HW_2(CO)_{10}]^-$ the hydrogen atom is about 0.07 nm off-axis. In these compounds there is an electron-deficient three-center linkage. Only one electron pair is involved, as in the case of diborane. When $[C_5H_5Mo(CO)_3]_2$ is dissolved in concentrated sulfuric acid, it is protonated to $[C_5H_5Mo(CO)_3]_2H^+$, which also has a hydrogen bridge bond. In $[H_2W_2(CO)_8]^{2-}$ there are two such bridging atoms, as sketched in **Fig. 7**. The molecule $[H_8Re_2(P\{C_2H_5\}_2C_6H_5)_4]$ has an Re-Re bond and no less than four bridging H atoms, as well as four terminal H atoms. The core of this structure is shown in **Fig. 8**. Each hydrogen atom in $[H_3Ni_4(C_5H_5)_4]$ is bounded to three nickel atoms. The hydrogen atom of $[HCo_6(CO)_{15}]^-$, a cluster ion, is in the center of the octahedron formed by the six cobalt atoms. A recently discovered hydrido complex of thorium is the dimer of $[H_2Th(C_5\{CH_3\}\ _5)_2]$. The organic ligands are pentamethylcyclopentadienyl groups. As shown in **Fig. 9**, two hydrogen atoms are bridging and two are terminal. Some polynuclear hydrido complexes are listed below.

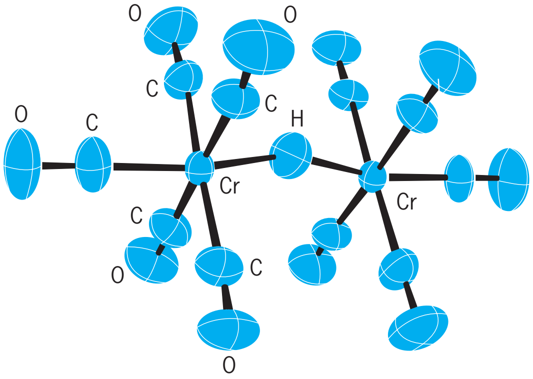


Fig. 6. Structure of $[HCr_2(CO)_{10}]^-$.

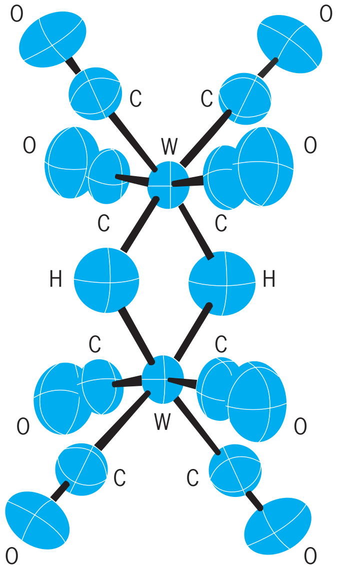


Fig. 7. Structure of $[H_2W_2(CO)_8]^{2-}$.

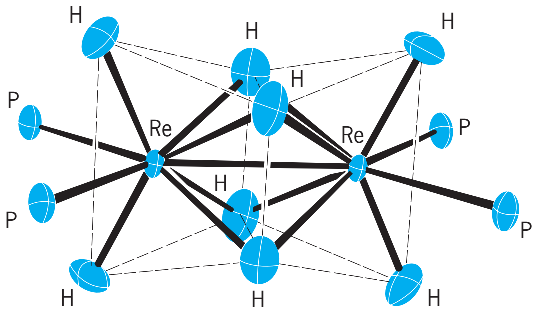


Fig. 8. Core of the $[H_8Re_2(P\{C_2H_5\}_2C_6H_5)_4]$ molecule. All organic groups are omitted for clarity.

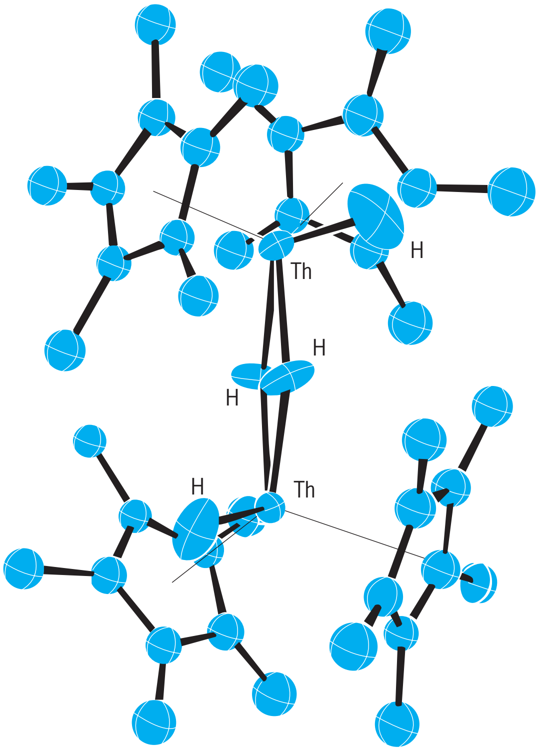


Fig. 9. Structure of $[H_2Th(C_5\{CH\ _3\}_5)_2]_2$. The hydrogen atoms of the methyl groups are omitted for clarity.

$[HNb_6I_{11}]$
$[HCr_2(CO)_{10}]^-$
$[C_5H_5Mo(CO)_3]_2H^+$
$[HW_2(CO)_{10}]-$
$[H_2W_2(CO)_8]^{2-}$
$[H_2Mn_2(CO)_9]$
$[H_8Re_2(PR_3)_4]$,
(R = ethylorphenyl)

$[H_2Re_3(CO)_{12}]^-$
$[HFe_3(CO)_{11}]$
$[HCo_6(CO)_{15}]^-$
$[H_2Ni_2(CO)_6]$
$[H_3Ni_4(C_5H_5)_4]$
$[HNi_{12}(CO)_{21}]^{3-}$
$[HCuP(C_6H_5)_3]_6$

James C. Warf

Bibliography. A. Dedieu, *Transition Metal Hydrides*, 1992; A. P. Ginsberg, *Transition Metal Chemistry*, vol. 1, 1965; E. L. Muetterties, *Transition Metal Hydrides*, 1971.

Hydroboration

The process of producing organoboranes by the addition of diborane to unsaturated organic compounds. In ether solvents the addition of diborane to such molecules is exceedingly rapid and essentially quantitative. This reaction therefore makes the organoboranes readily available. Such organoboranes find application as intermediates for organic synthesis.

Procedures. Diborane is highly soluble in tetrahydrofuran, where it exists as the addition compound tetrahydrofuran-borane. Such solutions are often used for hydroboration, and merely involve bringing the two reactants together as indicated by reaction (1).

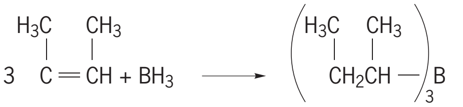
$$3RCH{=}CH_2 + C_4H_8O{:}BH_3 \longrightarrow (RCH_2CH_2)_3B + C_4H_8O \quad (1)$$

Alternatively, sodium borohydride may be utilized to achieve hydroboration by the addition of boron trifluoride etherate. This is shown by reaction (2).

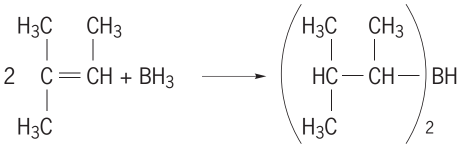
$$12RCH{=}CH_2 + 3NaBH_4 + 4(C_2H_5)_2O{:}BF_3 \longrightarrow 4(RCH_2CH_2)_3B + 3NaBF_4 + 4(C_2H_5)_2O \quad (2)$$

Usually the organoborane is not isolated but is utilized in place, similar to applications of the Grignard reagent in synthesis.

Scope and stoichiometry. Essentially all molecules containing one or more double or triple bonds undergo rapid conversion to organoboranes by this procedure. In general, disubstituted olefins react to give trialkylboranes, trisubstituted olefins react rapidly to give dialkylboranes and only slowly beyond, and tetrasubstituted olefins react rapidly to the monoalkylborane stage. The respective reactions are indicated by (3), (4), and (5).

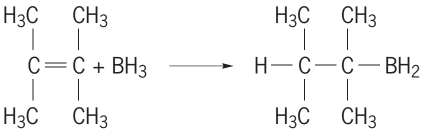


(3)



(4)

Disiamylborane

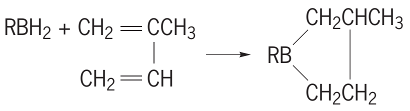


(5)

t-Hexylborane

Such substituted organoboranes are often very useful for controlled hydroborations, as indicated by reactions (6) and (7).

$$R_2BH + H_2C{=}CHCH_2Cl \longrightarrow R_2BCH_2CH_2CH_2Cl \quad (6)$$

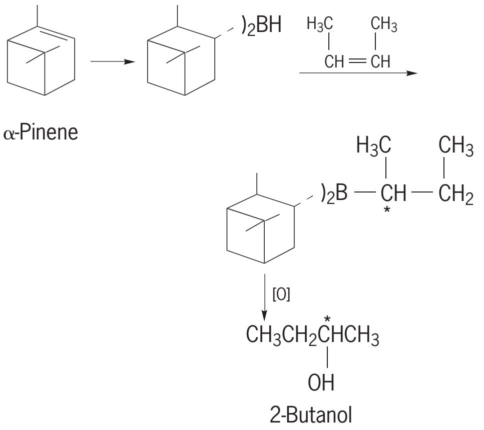


(7)

The hydroboration of unsaturated molecules containing functional groups is usually easily accomplished. Reaction (8) illustrates this technique. In this way organoboranes containing such functional groups become readily available for organic synthesis.

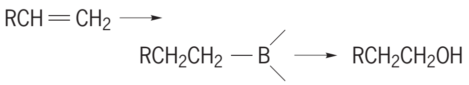
$$R_2BH + CH_2{=}CHCH_2CO_2C_2H_5 \longrightarrow R_2BCH_2CH_2CH_2CO_2C_2H_5 \quad (8)$$

Asymmetric syntheses. The hydroboration of optically active α-pinene yields an optically active dialkylborane. This can be utilized to introduce an asymmetric center, as illustrated in reaction (9), the synthesis of optically active 2-butanol in high optical purity. In reaction (9) the asymmetric center is indicated by an asterisk and [O] indicates an oxidation step.



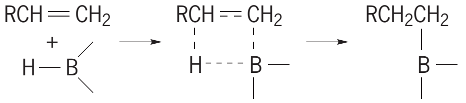
(9)

Directive effects. Hydroboration generally proceeds to place the boron atom at the less substituted of the two carbon atoms of a double bond (anti-Markovnikov addition). Since the boron atom may readily be replaced by many functional groups, such as hydroxyl and amino, this makes possible the anti-Markovnikov hydration and amination of double bonds, as shown by reaction (10).



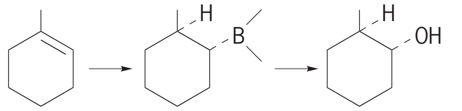
(10)

Cis addition. The hydroboration reaction appears to involve a simple four-center cis addition of the hydrogen-boron bond to the carbon-carbon double bond. Reaction (11) indicates this addition.



(11)

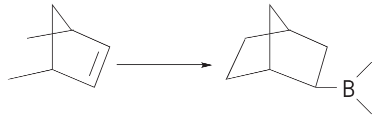
This oxidation proceeds with retention of configuration. Thus the hydroboration-oxidation of cyclic olefins, such as of 1-methylcyclohexene in reaction (12), provides the pure trans alcohol.



(12)

1-Methyl cyclohexene

Steric effects. The hydroboration reaction is quite sensitive to steric influences. Thus it hydroborates bicyclic olefins, such as norbornene in reaction (13),

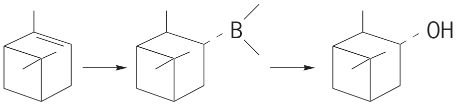


(13)

Norbornene 99.5% *exo-*

predominantly from the less hindered side. Moreover, the reaction does not cause rearrangements even in labile systems.

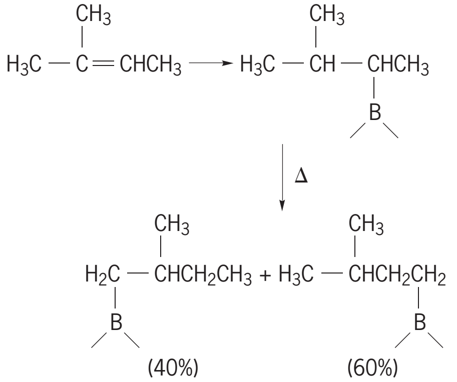
Stereospecific syntheses. These unusual characteristics—(1) anti-Markovnikoff addition, (2) freedom from rearrangement, (3) cis addition, and (4) high degree of steric control—give the hydroboration reaction major importance for achieving stereospecific syntheses. This is indicated in the reaction sequence (14), which is the synthesis of isopinocampheol from α-pinene.



(14)

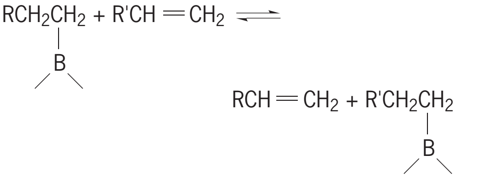
α-Pinene Isopinocampheol

Isomerization. Although the hydroboration reaction is remarkably free of rearrangements of the carbon structure, even in labile systems, the boron atom is capable of facile migration around the carbon skeleton at temperatures of 100–150°C (212–300°F). The reaction achieves a thermodynamic equilibrium among the possible organoboranes with that particular carbon skeleton. The preferred isomer is the one in which the boron atom is in the least crowded position. Reaction (15) illustrates this isomerization.



(15)

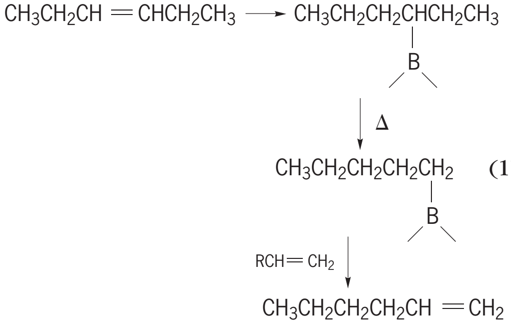
Displacement reaction. Heating an organoborane with another olefin transfers the boron atom to the new olefin, liberating the original alkyl group as an olefin, as in reaction (16). The reaction can be made



(16)

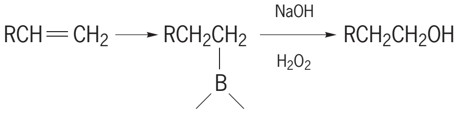
to proceed to essential completion by taking advantage of differences in volatility of the olefins, by using an excess of the displacing olefin, or by using an olefin, such as ethylene, which forms a very stable organoborane.

Contrathermodynamic isomerization. A combination of isomerization of the organoborane and the displacement reaction makes it possible to move double bonds from the more stable position at an alkyl branch or internal position to the less stable, unbranched terminal position. This is shown by the sequence in reaction (17).



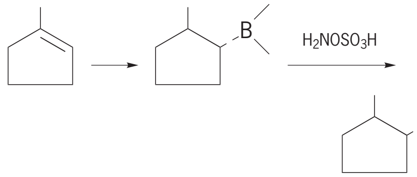
(17)

Oxidation. Organoboranes are readily oxidized by oxygen. Consequently, reactions of organoboranes are generally carried out under an inert atmosphere, such as nitrogen. Organoboranes are oxidized exceedingly readily by alkaline hydrogen peroxide, and this is the reagent of choice for the synthesis of alcohols via hydroboration, as shown by reaction (18).



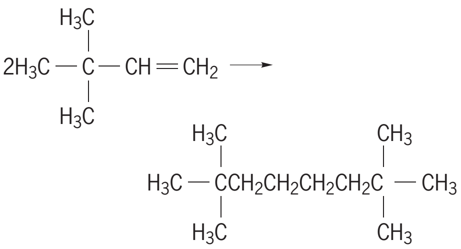
(18)

Amination. Chloramine and *o*-hydroxylamine-sulfonic acid convert organoboranes into the corresponding amine. Amination with *o*-hydroxylamine-sulfonic acid is shown by reaction (19).

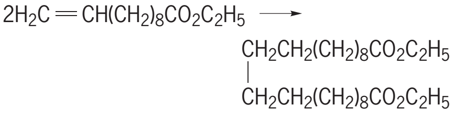


(19)

Coupling. Treatment of an olefin after hydroboration with alkaline silver nitrate produces a coupled product, reactions (20) and (21), that is a



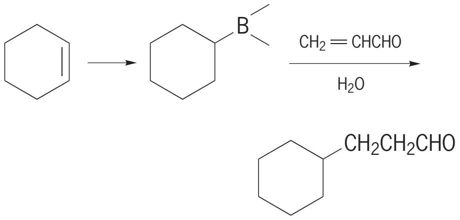
(20)



(21)

new compound composed of two molecules of the original.

1,4-Additions. Organoboranes react very rapidly with certain α,β-unsaturated aldehydes and ketones to give saturated products. This is illustrated by reaction (22).



(22)

Condensations. Organoboranes react with α-halo-substituted carbanions to transfer alkyl groups from boron to carbon. This is shown by reactions (23) and (24), where the symbol KO*t*-Bu represents potassium tertbutoxide.

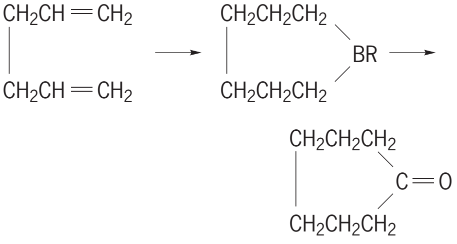
$$R_3B + CH_2BrCO_2C_2H_5 \xrightarrow{KOt\text{-}Bu} RCH_2CO_2C_2H_5 \quad (23)$$

$$R_3B + CH_2BrCO_2C_2H_5 \xrightarrow{KOt\text{-}Bu} RCHBrCO_2C_2H_5 \quad (24)$$

Carbonylation. Organoboranes react with carbon monoxide to produce intermediates which can be converted to tertiary alcohols, as in reaction (25),

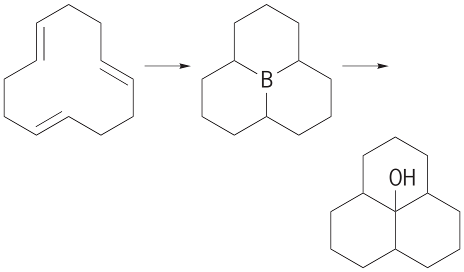
$$R_3B + CO \longrightarrow R_3CBO \xrightarrow{[O]} \underset{\substack{\text{Tertiary}\\\text{alcohol}}}{R_3COH} \quad (25)$$

secondary alcohols, ketones, aldehydes, methylol derivatives, ring ketones, as in reaction (26), and polycyclics, as in reaction (27).



(26)

Ring ketone



(27)

Polycyclic compound

Future developments. It is already apparent that the organoboranes are among the most versatile synthetic intermediates that are available to the organic chemist. The hydroboration reaction has made these intermediates readily available. *See* BORANE; CARBORANE; METAL HYDRIDES; ORGANIC SYNTHESIS; ORGANOMETALLIC COMPOUND.

Herbert C. Brown

Bibliography. H. C. Brown, *Hydroboration*, 1962; H. C. Brown, *Organic Syntheses via Boranes*, 1975; G. E. Coates and K. Wade, *Organometallic Compounds*, vol. 1, pt. 1: *Groups I–III*, 4th ed., 1999; G. Cragg, *Organoboranes in Organic Synthesis*, 1973; T. Onak, *Organoborane Chemistry*, 1975; K. Smith, *Organometallic Compounds of Boron*, 1985.

Hydrocharitales

An order of aquatic flowering plants, division Magnoliophyta Angiospermae), in the subclass Alismatidae of the class Liliopsida (monocotyledons). The order consists of the single family Hydrocharitaceae, with about 100 species. Within the subclass the order is marked by its inferior, compound ovary with basically laminar placentation. The ovules are scattered over the walls of the individual carpels, which are weakly connate to form partial partitions (intruded placentas) in the ovary. The aquarium plant *Vallisneria spiralis*, called tape grass or eelgrass, is not a grass but belongs to the Hydrocharitaceae. *See* ALISMATIDAE; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM.

Arthur Cronquist

Hydrocracking

A catalytic, high-pressure process flexible enough to produce either of the two major light fuels—high octane gasoline or aviation jet fuel. It proceeds by two main reactions: adding hydrogen to molecules too massive and complex for gasoline and then cracking them to the required fuels. The process is carried out by passing oil feed together with hydrogen at high pressure (1000–2500 lb/in.^{2} gage or 7–17 megapascals) and moderate temperatures (500–750°F or 260–400°C) into contact with a bifunctional catalyst, comprising an acidic solid and a hydrogenating metal component. Gasoline of high octane number is produced, both directly and through a subsequent step such as catalytic reforming; jet fuels may also be manufactured simply by changing conditions with the same catalysts. The process is characterized by a long catalyst life (2–4 years), though a slow decline in activity occurs, caused by the deposition of carbonaceous material on the catalyst. Regeneration at intervals by burning off these deposits restores the activity, but eventually the catalyst porosity is destroyed and it must be replaced.

Generally, the process is used as an adjunct to catalytic cracking. Oils, which are difficult to convert in the catalytic process because they are highly aromatic and cause rapid catalyst decline, can be easily handled in hydrocracking, because of the low cracking temperature and the high hydrogen pressure, which decreases catalyst fouling. Usually, these oils boil at 400–1000°F (200–540°C), but it is possible to process even higher-boiling feeds if very high hydrogen pressures are used. However, the most important components in any feed are the nitrogen-containing compounds, because these are severe poisons for hydrocracking catalysts and must be removed to a very low level.

Hydrocracking was carried out on a practical scale in Germany and England starting in the 1930s. In this early work, a common hydrocracking catalyst was tungsten disulfide on acid-treated clay; thus, both hydrogenation and acidic components were present. Generally, a light oil from coal or coking products was vaporized and passed over the catalyst at high pressure. After separation of gasoline from the products, the unconverted material was returned to the reactor with a fresh portion of feed. Because this catalyst was not very active, the process had to be carried out at very high pressures and temperatures (4000 lb/in.^{2} gage or 28 MPa; 750°F or 400°C). It was costly and the products were not of high quality.

Research in the United States concentrated on the development of much more active catalysts, a different mode of operation, and the use of heavier oil feeds. As a result, the reaction is carried out in two separate, consecutive stages; in each, oil and hydrogen at high pressure flow downward over fixed beds of catalyst pellets placed in large vertical cylindrical vessels.

First stage. In the first, or pretreating, stage the main purpose is conversion of nitrogen compounds in the feed to hydrocarbons and to ammonia by hydrogenation and mild hydrocracking. Typical conditions are 650–740°F (340–390°C), 150–2500 lb/in.^{2} gage (1–17 MPa), and a catalyst contact time of 0.5–1.5 h; up to 1.5 wt % hydrogen is absorbed, partly by conversion of the nitrogen compounds, but chiefly by aromatic compounds which are hydrogenated. It is most important to reduce the nitrogen content of the product oil to less than 0.001 wt% (10 parts per million). This stage is usually carried out with a bifunctional catalyst containing hydrogenation promotors, for example, nickel and tungsten or molybdenum sulfides, on an acidic support, such as silica-alumina. The metal sulfides hydrogenate aromatics and nitrogen compounds and prevent deposition of carbonaceous deposits; the acidic support accelerates nitrogen removal as ammonia by breaking carbon-nitrogen bonds. The catalyst is generally used as $^1/_8 \times {}^1/_8$ in. (0.32 × 0.32 cm) or $^1/_{\sqrt{16}} \times {}^1/_8$ in. (0.16 × 0.32 cm) pellets, formed by extrusion.

Second stage. Most of the hydrocracking is accomplished in the second stage, which resembles the first but uses a different catalyst. Ammonia and some gasoline are usually removed from the first-stage product, and then the remaining oil, which is low in nitrogen compounds, is passed over the second-stage catalyst. Again, typical conditions are 600–700°F (300–370°C), 1500–2500 lb/in.^{2} gage (10–17 MPa) hydrogen pressure, and 0.5–1.5 h contact time; 1–1.5 wt % hydrogen may be absorbed. Conversion to gasoline or jet fuel is seldom complete in one contact with the catalyst, so the lighter oils are removed by distillation of the products and the heavier, high-boiling product combined with fresh feed and recycled over the catalyst until it is completely converted.

The catalyst for the second stage is also a bifunctional catalyst containing hydrogenating and acidic components. Metals such as nickel, molybdenum, tungsten, or palladium are used in various combinations, dispersed on solid acidic supports such as synthetic amorphous or crystalline silica-aluminas, such as zeolites. These supports contain strongly acidic sites and sometimes are enhanced by the incorporation of a small amount of fluorine. A long period (for example, 1 year) between regenerations is desirable; this is achieved by keeping a low nitrogen content in the feed and avoiding high temperatures, which lead to excess cracking with consequent deposition of coke on the catalyst. When activity of the catalyst does decrease, it can be restored by carefully controlled burning of the coke.

The catalyst is the key to the success of the hydrocracking process as now practiced, particularly the second-stage catalyst. Its two functions must be most carefully balanced for the product desired; that is, too much hydrogenation gives a poor gasoline but a good jet fuel. The oil feeds are composed of paraffins, other saturates, and aromatics—all complex molecules boiling well above the required gasoline or jet-fuel product. The catalyst starts the breakdown of these components by forming from them carbonium ions, that is, positively charged molecular fragments, via the protons (H^+) in the acidic function. These ions are so reactive that they change their internal molecular structure spontaneously and break down to smaller fragments having excellent gasoline qualities. The hydrogenating function aids in maintaining and controlling the ion reactions and protects the acid function by hydrogenating coke precursors off the catalyst surface, thus maintaining catalyst activity. Any olefins formed in the carbonium ion decomposition are also hydrogenated.

Products. The products from hydrocracking are composed of either saturated or aromatic compounds; no olefins are found. In making gasoline, the lower paraffins formed have high octane numbers; for examble, the 5- and 6-carbon number fractions have leaded research octane numbers of 99–100. The remaining gasoline has excellent properties as a feed to catalytic reforming, producing a highly aromatic gasoline which, with added lead, easily attains 100 octane number. Both gasolines are suitable for premium-grade motor gasoline. Another attractive feature of hydrocracking is the low yield of gaseous components, such as methane, ethane, and propane, which are less desirable than gasoline. When making jet fuel, more hydrogenation activity of the catalysts is used, since jet fuel contains more saturates than gasoline. *See* GASOLINE.

The hydrocracking process is being applied in other areas, notably, to produce lubricating oils and to convert very asphaltic and high-boiling residues to lower-boiling fuels. Its use will certainly increase greatly in the future, since it accomplishes two needed functions in the petroleum-fuel economy: Large, unwieldy molecules are cracked, and the needed hydrogen is added to produce useful, high-quality fuels. *See* AROMATIZATION; CRACKING; HYDROGENATION; ISOMERIZATION; REFORMING PROCESSES. Charles P. Brewer

Bibliography. L. W. Hall, *Petroleum Production Operations*, 1986; J. McKetta, *Petroleum Processing Handbook*, 1992.

Hydroelectric generator

A low-speed generator driven by water turbines. Hydrogenerators may have a horizontal or vertical shaft. The horizontal units are usually small with speeds of 300–1200 revolutions per minute (rpm). The vertical units are usually larger and more easily adapted to small hydraulic heads. The rotor diameters range from 2 to 62 ft (0.6 to 19 m) and capacities from 50 to 900,000 kVA. The generators are rated in kVA (kilovolts times amperes). The kilowatt output is the product of kVA and power factor. The normal power-factor rating of small synchronous generators is between 0.8 and 1.0 with 0.9 being common. For large generators a rating of 0.9–0.95 is common with the machines able to operate up to 1.0 when the load requires. The generators may also supply reactive power. *See* ALTERNATING CURRENT; ELECTRIC POWER MEASUREMENT; VOLT-AMPERE.

Structure. The turbine shown in the **illustration** has an adjustable blade propeller, typical of large, low-head units that are common on large river power plants. The water enters the turbine spiral scroll casing, falls down through the turbine, causing rotation, and empties into the river. The shaft transmits the rotation to the generator spider or hub and thence to the rotor rim and poles. The magnetic field of the rotor poles transmits the torque to the stator and changes the mechanical power to electrical power. *See* HYDRAULIC TURBINE; TURBINE.

The poles are spaced around the rotor rim and are magnetized by direct current flowing in the turns of the field coil around each pole. The magnetic field, or flux, crosses the air gap between rotor and stator, flows radially through the stator teeth and thence to the area one pole pitch away, and back to the adjacent pole on the rotor. The magnetic flux is stationary with respect to the rotor poles but sweeps around the stator at the peripheral rotor speed. Coils are installed in the stator slots between the teeth. Thus there is an ever- changing flux linking stator coils, which causes an induced electromotive force in the coils according to Faraday's law. The alternating voltage induced by the changing flux is given by

$$E = 4.4fN\phi$$

where E is the root-mean-square (rms) value of the induced voltage in volts, f is the system frequency in hertz, N is the number of turns in the coil, and ϕ is the maximum flux linkage in the coil in webers (= flux in lines × 10^{-8}). *See* ALTERNATING-CURRENT GENERATOR.

Almost all machines have three groups of coils per pole, making a three-phase generator. The windings are connected; thus the sum of the coil voltages in each phase is equal to the rated phase voltage. The phase voltages have a time distribution: sequential phases reach their positive maximum at intervals of one-third of each cycle. This provides the best practical combination for transmission and a rotating field for driving motors.

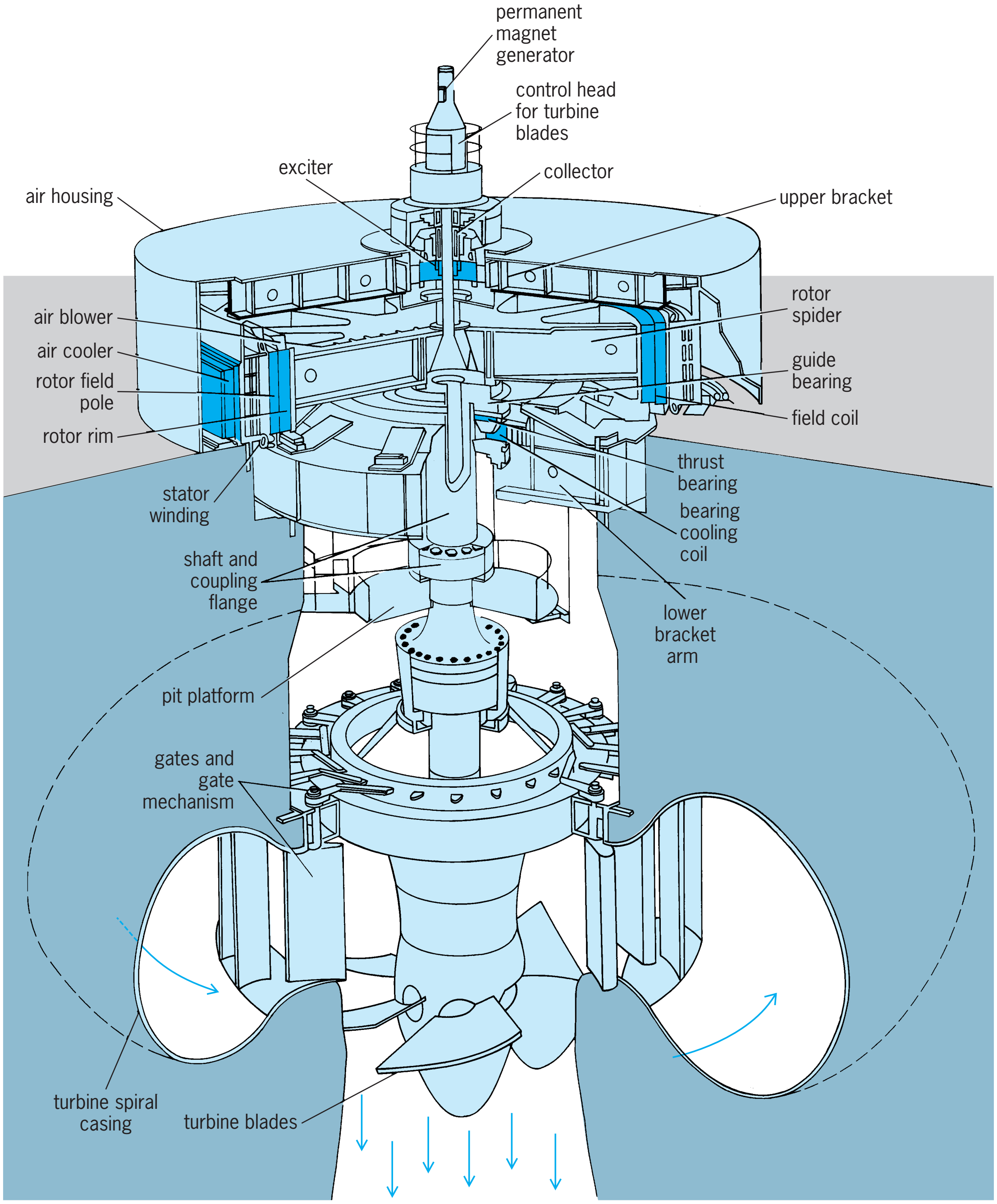
The direct current required to magnetize the field poles is supplied by a small exciter generator mounted on the rotor or by an electronically controlled rectifier. The current is fed to the field through brushes on collector rings on the rotor.

The rotating parts of both generator and turbine are supported by the generator thrust bearing, which also supports the hydraulic thrust of the turbine. This thrust may be as much as the total weight of the rotating parts. Guide bearings on the turbine and generator keep the rotating parts on center. These bearings are usually of the shoe type, operating in a bath of oil.

All hydromachines are required to withstand the highest attainable speed (runaway speed) with the governor inoperative after tripping from the line. For good governing, the flywheel effect may have to be larger than inherent in the electrical parts. This may affect the diameter and weight of the unit. *See* GOVERNOR.

Large machines need a housing for reasons of safety, noise reduction and cleanliness. The air is circulated by the fan action of the rotor between the poles and through the radial ducts in the stator core, cooling both the stator iron and the coils. The air then passes through coolers and recirculates.

Standards and characteristics. Most modern machines are built to the 1977 ANSI standard with class F insulation on the windings but temperature limits for class B insulation. The limit for this class of machines is a stator-winding temperature rise of 80°C (144°F), above an ambient temperature of 40°C (104°F), as measured by a resistance detector. The temperature rise observed by the detector is frequently less than the 80°C (144°F) specified as the practical insulation limit. The temperature-rise limit on the rotor field windings is also 80°C (144°F), but as measured by change in field resistance.



Large hydroelectric generator. (*Westinghouse Electric Corp*.)

In 1977 the standard method of rating hydrogenerators was changed to allow the 80°C (144°F) temperature rise but with no overload rating. The rated kVA of the generator should nearly match the maximum expected turbine capability and should be operated below the name-plate rating most of the time for best winding life. Before this change, the generator name-plate kVA and power factor were based on a 60°C (108°F) rise and about matched the turbine operating point of highest efficiency. With 15% overload and a higher temperature rise, the generator could usually handle the turbine capability.

Hydrogenerators of normal characteristics usually serve for application in the United States with the plant near a large power concentration or substation. However, when long lines are required or special problems exist, subnormal reactances and other characteristics with larger underexcited capability may be required for stability and voltage regulation. These features may be purchased with the equipment.

Induction generator. An induction generator is the same as an induction motor, except that it is driven by a turbine at a few percent above a synchronous speed instead of the motor driving a load a few percent below a synchronous speed. Induction generators are used locally to recover energy at small dams. They receive excitation or magnetizing kVA from the transmission line; thus their kVA ratings are always higher than those equivalent synchronous machines with the same turbine kilowatt output. They often have step-up gears to raise the generator speed to 720–1800 rpm, and must withstand the higher speed required by the turbine. The power system capability at the machine terminals must be large compared to the machine capacity. The power company may require a penalty to supply the necessary reactive kVA. Some installations use capacitors connected to the line to correct this power factor. This can be dangerous if the machine is tripped free of the system with the capacitors still connected; the machine may overexcite by resonance and fail because of the overvoltage generated. Induction machines are satisfactory for many applications and are a little less expensive than synchronous machines because of the elimination of a voltage regulator and excitation supply. However, induction machines are seldom satisfactory below 400–500 rpm because of their low power factor (0.6–0.8) and low efficiency (2–7% less). In addition, the mechanical deflection of parts with the variable hydraulic thrust from the turbine causes gear and generator misalignment and noise. *See* ELECTRIC POWER GENERATION; ELECTRIC ROTATING MACHINERY; GENERATOR; INDUCTION MOTOR; WATERPOWER. Eugene C. Whitney

Bibliography. American National Standards Institute, *Rotating Electrical Machinery-Synchronous Machines*, ANSI Stand. C50.10-1990, 1990; E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; Institute of Electrical and Electronics Engineers, *IEEE Guide for Operation and Maintenance of Hydro-Generators*, IEEE Stand. 492–1999, 1999; Institute of Electrical and Electronics Engineers, *IEEE Guide: Test Procedures for Synchronous Machines*, IEEE Stand. 115–1995, 1995; Institute of Electrical and Electronics Engineers, *IEEE Standard for Salient-Pole 50 Hz and 60 Hz Synchronous Generators and Generator-Motors for Hydraulic Turbine Applications Rated 5 MVA and Above*, IEEE Stand. C50.12–2005, 2005.

Hydrofoil craft

A form of high-speed ship that supports its weight by means of wings (properly called hydrofoils, or simply foils) beneath the surface of the water. The hydrofoils generate lift by movement in the same manner as an airplane wing. The hydrofoil was conceived in order to produce faster ships. The most effective means of developing a faster ship is to find a way to lift the hull clear of the water. This greatly reduces the drag on the hull, in turn greatly reducing the power required to drive the ship. The hydrofoil ship is one means to this end. *See* AIRFOIL; SHIP POWERING, MANEUVERING, AND SEAKEEPING.

Many inventors, including Alexander Graham Bell, pursued the hydrofoil concept. A patent for a hydrofoil was granted in the United States in the late 1880s. The earliest record of a successful hydrofoil flight was in 1894, when the Meacham brothers demonstrated a 14-ft (4.3-m) test craft in Chicago. This use of a wing to support a boat precedes by nearly 10 years the Wright brothers' first airplane flight in 1903.

TABLE 1. Specifications of U.S. Navy PHM Class hydrofoil craft

Weight	243 metric tons	239 long tons
Length	44 m	145 ft
Speed	48 knots	89 km/h
Power	13.4 MW	18,000 hp
Year first built	1997	
Number built	6	

The hydrofoil concept was not successfully exploited until the development of lightweight engines and structural materials. These developments—many of which came from the aircraft industry—made possible commercial and military hydrofoils.

There are two basic types of hydrofoils: fully submerged and partially submerged.

Fully submerged hydrofoils. The most efficient hydrofoil craft, from a powering standpoint, are those with fully submerged foils; that is, the wings are completely below the surface of the water. They are not subject to surface interactions such as air drawing, or to the danger of broaching through small waves. They are also relatively unaffected by wave action, resulting in excellent ship ride comfort. However, because they are fully submerged, the ship must be actively controlled. A conventional surface ship will stay on the surface, regulating its height by the simple action of floating. A fully submerged hydrofoil, by contrast, has no passive or inherent means to maintain its ride height and attitude. The ship may slowly (or rapidly) rise or fall, subject to small perturbations in the foil angles. This behavior could be wild and chaotic in an uncontrolled hydrofoil. All existing fully- submerged-type hydrofoils overcome this through the use of an active control system which senses the craft's height and attitude and adjusts the

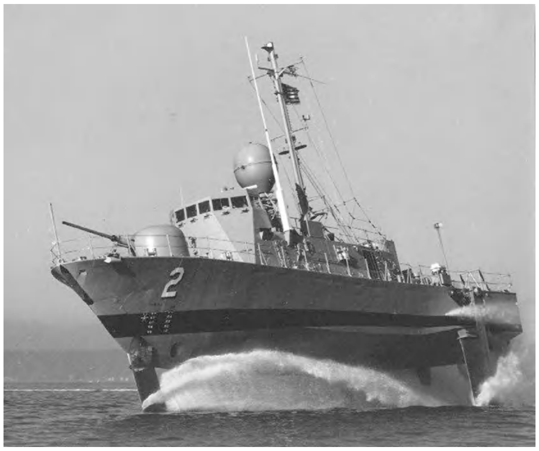


Fig. 1. U.S. Navy Patrol Hydrofoil, Missile (PHM) Class ship. (*U.S. Navy*)

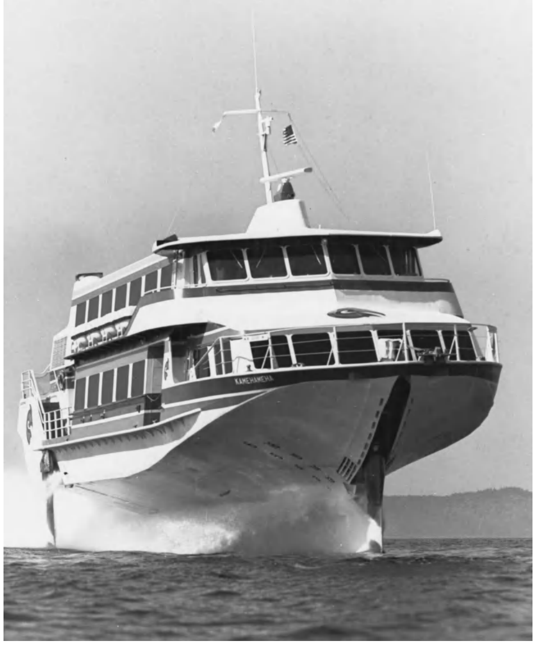


Fig. 2. JetFoil *Kamehameha*. (*Boeing Airplane Co.*)

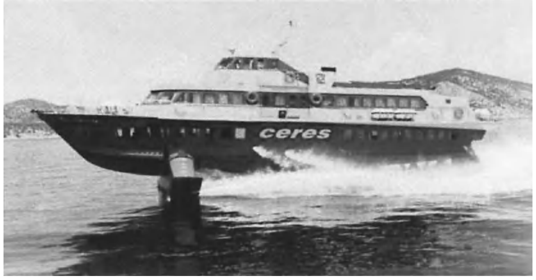


Fig. 3. Rodriquez hydrofoil in the RHS-160 series.

wings as necessary to maintain the desired condition. *See* CONTROL SYSTEMS.

Partially submerged hydrofoils. The need for an active control system in a fully submerged hydrofoil led to the development of the partially submerged hydrofoil. In a partially submerged hydrofoil, the foils are generally V-shaped when viewed head on, with the apex of the V below the water and the tips of the wings above the water. In this case the amount of lift generated will increase if the foils are more deeply submerged, and it will decrease if the foils are less submerged. (The increase and decrease in lift are due to the change in the wetted area of the foils.)

The result of this behavior is that if the craft rises, the foils will lose some lift, resulting in its settling downward. If the craft settles beyond equilibrium, excess lift will be generated and the craft will rise. In actual designs, this behavior results in a steady ride attitude at an equilibrium balance of lift. A disadvantage of partially submerged foils is that the ride is not as smooth, especially in rough water, since the foils will react to each wave they encounter.

Limitations. Two significant limitations constrain the application of hydrofoil craft in any given mission. The first is a consequence of size. The weight of a ship increases approximately as the cube of a change in ship dimensions. (Double the length, width, and height of the ship, and the weight will increase approximately eightfold.) The lift of a hydrofoil, however, increases only as the square of a dimension change. (Double the span and chord of the foil, and the lift will increase approximately fourfold.) This cube/square relationship is increasingly difficult to overcome as the respective values increase. The variation in size between, say, a 30-m (100-ft) craft and a 40-m (130-ft) craft can be overcome with detailed engineering. This is not easily accomplished, however, for a 100-m-long (330-ft) craft. As a consequence, there are no hydrofoils larger than about 1000 tons.

The second challenge is that the hydrofoil is a one-speed craft. The lift on a foil varies as the square of the speed. Increasing the speed by 20% will lead to a 44% increase in lift. It is very difficult to accommodate a lift variation this large by any system of foil control. As a consequence, there is generally a rather narrow band of speeds in which the boat is best operated. This one-speed characteristic presents challenges for certain missions, such as military operations, which may involve extended periods of time at low or moderate speeds during which the hydrofoil sits down off its foils on the basic hull structure. The hydrofoil is then supported by buoyancy and offers no advantage over a conventional hulled craft.

Significant modern hydrofoils. The following ships are representative of modern hydrofoils.

US Navy PHM Class. The U.S. Navy Patrol Hydrofoil, Missile (PHM) Class ships were developed as the result of a NATO cooperative program. Similar hydrofoils were put in service by Italy as part of this program, and related craft were built for Israel and France. These craft use fully submerged foils, the main one located aft and a secondary one located forward. The U.S. Navy craft (**Table 1**; **Fig. 1**) were retired from service in 1996–1997. *See* NAVAL SURFACE SHIP.

Boeing JetFoil. The USN PHM Class ships were built by the Boeing Company, which used the same technology to produce a line of commercial hydro-

TABLE 2. Characteristics of five types of Rodriquez hydrofoils

	RHS-70	RHS-150	RHS-160	RHS-200	MEC-1
Length, m (ft)	22 (72)	29 (95)	31 (102)	36 (118)	25 (82)
Breadth, m (ft)	7.85 (25.75)	5.85 (19.2)	6.2 (20.3)	7.0 (23.0)	8.4 (27.6)
Weight, metric tons (long tons)	31.5 (31.0)	65.6 (64.6)	85 (84)	120 (118)	47.7 (46.8)
Passengers	69	180	205	300	146
Cruise speed, knots (km/h)	32.5 (60.2)	32.5 (60.2)	38 (70)	35 (65)	36 (67)

TABLE 3. Characteristics of Russian Katran Class hydrofoils

Length	34.5 m	113.2 ft
Breadth	10.3 m	33.8 ft
Weight	72 metric tons	71 long tons
Passengers	147	
Cruise speed	35 knots	65 km/h

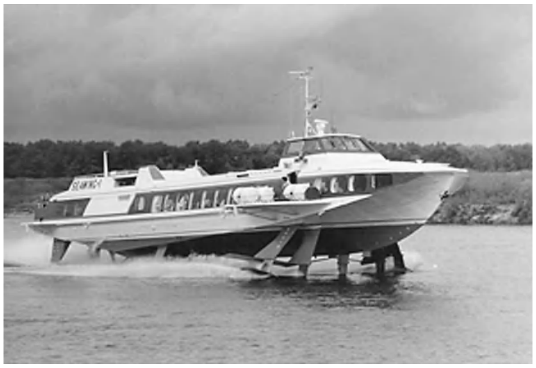


Fig. 4. Russian hydrofoil in Katran Class. (*Volga Shipyard*)

foil ferries. The Boeing JetFoil ferry (**Fig. 2**) was produced in a small variety of sizes and capacities, and several units were sold commercially. The majority of them are in service as ferries in Hong Kong. In the late 1980s, Boeing sold the rights and technology for the JetFoil. JetFoils are now available from Kawasaki Jetfoil. *See* FERRY.

Rodriquez ferries. The Boeing JetFoil and the USN PHM use fully submerged foils. Since the 1950s, the Italian shipbuilder Rodriquez Cantieri Navali SpA has been building commercial hydrofoils of the surface-piercing type. The earliest Rodriquez ferries were the PT-20 type, which have all but disappeared from service. These were followed by a range of craft up to 40 m (130 ft) in length (**Table 2**; **Fig. 3**). Rodriquez continues to make incremental improvement and developments in their hydrofoil product line.

Russian hydrofoils. The former Soviet Union sponsored development of several hydrofoils for civilian river transport. These craft were all of the surface-piercing type. A few have been successfully exported for commercial operation, such as the Katran Class (**Table 3**; **Fig. 4**).

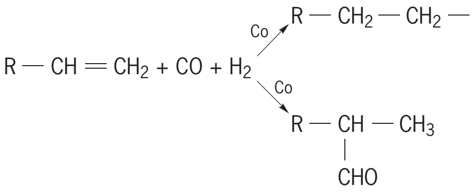
Buoyant hydrofoils. Since the mid-1990s, there has been some research on buoyant hydrofoils, also called lifting bodies. These foils are very thick and are intended to provide substantial hydrostatic lift in addition to their hydrodynamic lift. The purpose of these foils is to marry the speed potential of the hydrofoil with the seakindliness of the SWATH (small waterplane area, twin hull) ship concept. These concepts, however, are still developmental. Milestones in this development include the MIDFOIL craft and the HDV-100, both developed by Navatek Ships in Honolulu, and the QUEST HYSWAS developed by Maritime Applied Physics Corporation.

Chris B. McKesson

Bibliography. *Fast Ferry International* (formerly *Hovering Craft and Hydrofoil*), Fast Ferry Information Ltd., Kent, U.K.; *Ferry Technology*, Riviera Maritime Media Ltd, Enfield, Middlesex, U.K.; Capt. R. Johnson (ed.), Hydrofoils, *Naval Eng. J.*, pp. 142–199, February 1985; E. V. Lewis (ed.), *Principles of Naval Architecture*, 2d ed., 3 vols., Society of Naval Architects and Marine Engineers, Jersey City, NJ, 1988.

Hydroformylation

An aldehyde synthesis process that falls under the general classification of a Fischer-Tropsch reaction but is distinguished by the addition of an olefin feed along with the characteristic carbon monoxide and hydrogen. In the oxo process for alcohol manufacture, hydroformylation of olefins to aldehydes is the first step. The second step is the hydrogenation of the aldehydes to alcohols. At times the term "oxo process" is used in reference to the hydroformylation step alone. In the hydroformylation step, olefin, carbon monoxide, and hydrogen are reacted over a cobalt catalyst to produce an aldehyde which has one more carbon atom than the feed olefin. The olefin conversion takes place by the addition of a formyl group (CHO) and a hydrogen atom across the double bond. This is represented by reaction (1).



(1)

See FISCHER-TROPSCH PROCESS.

The aldehyde is then treated with hydrogen to form the alcohol. In commercial operations, the hydrogenation step is usually performed immediately after the hydroformylation step in an integrated system.

A wide range of carbon number olefins, C_2–C_{16}, have been used as feeds. Propylene, heptene, and nonene are frequently used as feedstocks to produce normal and isobutyl alcohol, isooctyl alcohol, and primary decyl alcohol, respectively. Feed streams to oxo units may be single-carbon-number or mixed-carbon-number olefins.

The lower-carbon-number alcohols such as butanols are used primarily as solvents, while the higher-carbon-number alcohols go into the manufacture of plasticizers, detergents (surfactants), and lubricants.

Reactants. Reactants are CO, H_2, and olefins. The H_2 and CO are usually fed in a 1-to-1 ratio, as synthesis gas from methanol conversion. The hydroformylation takes place in the liquid phase. If the olefins are normally gaseous at the reactor conditions, a heavier liquid solvent is used as a suspension medium. The synthesis gas is usually fed in considerable excess of the required stoichiometric amount.

The reaction mass is contacted with the cobalt catalyst either by flowing the liquid reaction mass over a

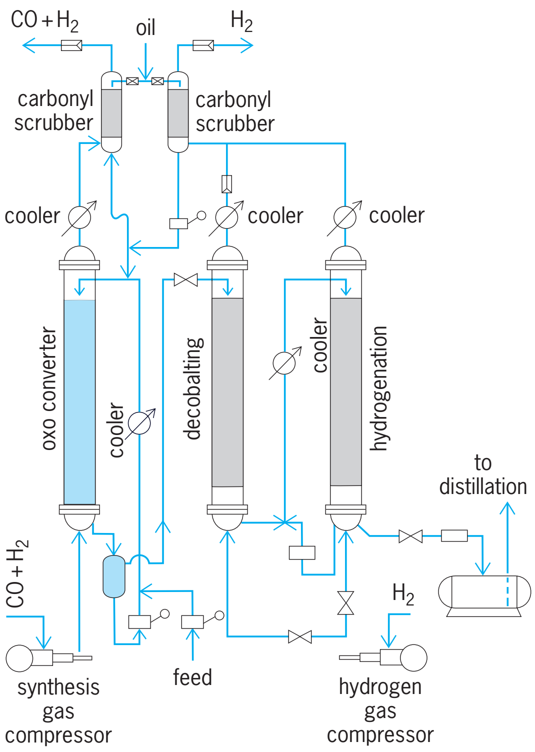


Fig. 1. Oxo process with fixed catalyst bed.

fixed bed of supported cobalt catalyst or by pumping a cobalt catalyst slurry into the liquid reaction mass in a continuous stirred tank or plug flow reactor. In the latter case, the catalyst is carried along with the reaction mass in what is frequently called the slurry system.

The olefin reactants are fed in blocked operation as either single-carbon-number C_7, C_8, or C_9 or in multiple-carbon-number feeds, such as C_{11}–C_{13}. For straight-chain olefins the double bond can be either in the terminal position or internal position. The product distribution is approximately the same in either case, roughly 60–40% normal alcohol and 40–60% alpha branched alcohol. Although the product distribution is nearly the same, the reaction rate when the double bond is in the terminal position is almost 3.5 times faster than the rate with double bond in an internal position.

The reaction rate has been found to be directly proportional to the olefin concentration, the cobalt concentration, and the hydrogen pressure, and inversely proportional to carbon monoxide pressure. The proposed mechanism for the reaction begins with the formation of an olefin-carbonyl complex and carbon monoxide from the reaction of dicobalt octacarbonyl with olefin, as shown in reaction (2). The complex

$$Co_2(CO)_8 + RCH{=}CH_2 \longrightarrow Co_2(CO)_7RCH{=}CH_2 + CO \quad (2)$$

decomposes through a reaction with a hydrogen or cobalt hydrocarbonyl to form the aldehyde and a precursor of dicobalt octacarbonyl. The heat of reaction of the hydroformylation step is about 30 kcal/mole (125,500 joules/mole). The capability of removing the heat of reaction is often the limiting factor in the reactor capacity.

Commercial operations. Two general types of liquid-phase process have been used. The first process employs a fixed bed of cobalt catalyst with the liquid reactant mixture flowing past the catalyst. The second type of process, the slurry type, carries the catalyst along with the liquid reactant. The slurry type of reactors can be either the mechanically agitated, stirred- tank type or the gas-sparged type. A stirred-tank type with mechanical agitation can be used when the operating pressure is not too high. However, many operations are run at high pressure, 100–300 atm (1–3 × 10^6 pascals). For these processes, a gas-sparged slurry-type reactor can be designed to use the gas rising through the liquid to provide mixing and thus eliminate the mechanical agitator and seals.

The fixed-bed process is shown schematically in **Fig. 1**. In this process, soluble cobalt salts of fatty acids or naphthenates are pumped with the olefin to the top of the first reactor and flow countercurrent to the synthesis gas. One type of fixed-bed catalyst consists of 2% metallic cobalt on a pumice carrier. Part of the cobalt is converted to carbonyl, leaves the reactor with the overhead product, and is replaced by cobalt salts in the feed. Unreacted synthesis gas leaving the top of the reactor is cooled, passed through a packed tower countercurrent to the olefin feed to remove cobalt carbonyl, and recycled to the reactor.

The second vessel is a decobalting converter in which cobalt carbonyl, dissolved in the product from the first reactor, is decomposed by treatment with hydrogen at about 200–220 atm (2–2.2 × 10^6 Pa) and 257–300°F (125–150°C). The liquid enters at the top and flows countercurrent to the hydrogen. Metallic cobalt is deposited on the packing. Gas flows from the top of the decobalting unit to the carbonyl scrubber, and the liquid leaving the bottom is sent to the hydrogenation reactor for conversion to alcohol.

The slurry type of process shown schematically in **Fig. 2** begins with cobalt oxide being mixed with

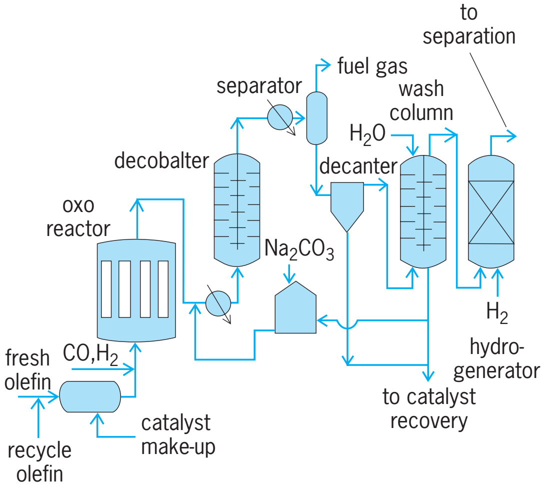


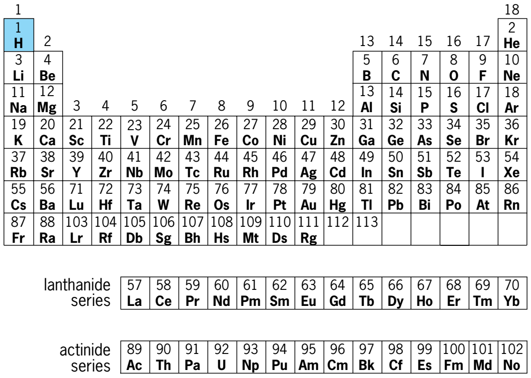
Fig. 2. Slurry process, in which the catalyst is carried with the liquid reactant.

recycle olefins. This slurry is then combined with the main olefin feed stream to give a 3–5 wt % cobalt slurry. This reaction slurry is fed into the bottom of the first reactor along with the synthesis gas. This process may employ up to five reactors in series. In the reactors, the catalyst reacts with the synthesis gas to form cobalt hydrocarbonyl, which is a gas at the reactor conditions of 356°F (180°C) and 220 atm (2.2 × 10^6 Pa). The liquid-gas mixture exits the last reactor of the series, and an aqueous solution of sodium carbonate is injected into this stream to form the water-soluble sodium cobalt carbonylate. The aqueous phase is separated from the organic stream in the decanter, and the organic stream is then water-washed to remove the last traces of catalyst. The washed aldehyde stream is then fed to the hydrogenator from which the product alcohol goes to the separation train downstream. The catalyst is recovered from the wash water and recycled to the process. D. L. Holt

Hydrogen

The first chemical element in the periodic system. Under ordinary conditions it is a colorless, odorless, tasteless gas composed of diatomic molecules, H_2. The hydrogen atom, symbol H, consists of a nucleus of unit positive charge and a single electron. It has atomic number 1 and an atomic weight of 1.00797. The element is a major constituent of water and all organic matter, and is widely distributed not only on the Earth but throughout the universe. There are three isotopes of hydrogen: protium, mass 1, makes up 99.98% of the natural element; deuterium, mass 2, makes up about 0.02%; and tritium, mass 3, occurs in extremely small amounts in nature but may be produced artificially by various nuclear reactions. *See* DEUTERIUM; ISOTOPE; PERIODIC TABLE; TRITIUM.

Uses. The largest single use of hydrogen is in the synthesis of ammonia. A rapidly expanding use for hydrogen is in petroleum-refining operations, such as hydrocracking and hydrogen treatment for removal of sulfur. Large quantities of hydrogen are consumed in the catalytic hydrogenation of unsaturated liquid vegetable oils to make solid fats. Hydrogenation is used in the manufacture of organic chemicals. Large quantities of hydrogen are used as a rocket fuel, in conjunction with oxygen or fluorine, and as a propellent for nuclear-powered rockets.


1 18
1 H 2 13 14 15 16 17 2 He
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

Properties of hydrogen

Property	Value
Melting point	−259.2°C
Boiling point at 1 atm	−252.8°C
Density of solid at −259.2°C	0.0866 g/cm^3
Density of liquid at −252.8°C	0.0708 g/cm^3
Critical temperature	−240.0°C
Critical pressure	13.0 atm
Critical density	0.0301 g/cm^3

Properties. Ordinary hydrogen has a molecular weight of 2.01594. The gas has a density at 0°C and 1 atm of 0.08987 g/liter. Its specific gravity, compared to air, is 0.0695. Hydrogen is the lightest substance known. Some additional properties of hydrogen are listed in the **table**.

Hydrogen is somewhat more soluble in organic solvents than in water. Many metals adsorb hydrogen. The adsorption of hydrogen in steel may cause “hydrogen embrittlement,” which sometimes leads to the failure of chemical processing equipment.

At ordinary temperatures hydrogen is a comparatively unre-active substance unless it has been activated in some manner, for example, by a suitable catalyst. At elevated temperatures it is highly reactive.

Although ordinarily diatomic, molecular hydrogen dissociates at high temperatures into free atoms. Atomic hydrogen is a powerful reducing agent, even at room temperature. It reacts with the oxides and chlorides of many metals, including silver, copper, lead, bismuth, and mercury, to produce the free metals. It reduces some salts, such as nitrates, nitrites, and cyanides of sodium and potassium, to the metallic state. It reacts with a number of elements, both metals and nonmetals, to yield hydrides such as NaH, KH, H_2S, and PH_3. With oxygen atomic hydrogen yields hydrogen peroxide, H_2O_2. With organic compounds atomic hydrogen reacts to produce a complex mixture of products. With ethylene, C_2H_4, for example, the products include ethane, C_2H_6, and butane, C_4H_{10}. The heat liberated when hydrogen atoms recombine to form hydrogen molecules is used to obtain very high temperatures in atomic hydrogen welding.

Hydrogen reacts with oxygen to form water. At room temperature this reaction is immeasurably slow, but is accelerated by catalysts, such as platinum, or by an electric spark, and then may take place with explosive violence.

With nitrogen, hydrogen undergoes an important reaction to give ammonia. Hydrogen reacts at elevated temperatures with a number of metals to give hydrides. The oxides of many metals are reduced by hydrogen at elevated temperatures either to the free metal or to lower oxides. Hydrogen reacts at room temperature with the salts of the less electropositive metals and reduces them to the metallic state. In the presence of a suitable catalyst hydrogen reacts with unsaturated organic compounds and adds to the double bond. *See* HYDROGENATION.

Principal compounds. Hydrogen is a constituent of a very large number of compounds containing one or more other elements. Such compounds include water, acids, bases, most organic compounds, and many minerals. Compounds in which hydrogen is combined with a single other element are commonly referred to as hydrides. For additional details on the compounds of hydrogen *see* ACID AND BASE; HYDRAZINE; HYDRIDE; HYDRIDO COMPLEXES; HYDROGEN FLUORIDE; HYDROGEN PEROXIDE; WATER.

Preparation. A large number of methods may be used to prepare hydrogen gas. The choice of method is determined by such factors as the quantity of hydrogen desired, the purity required, and the availability and cost of raw materials. Among the processes frequently used are the reactions of metals with water or acids, the electrolysis of water, the reaction of steam with hydrocarbons or other organic materials, and the thermal decomposition of hydrocarbons. The principal raw materials for hydrogen production are hydrocarbons, such as natural gas, oil refinery gas, gasoline, fuel oil, and crude oil.

Louis Kaplan

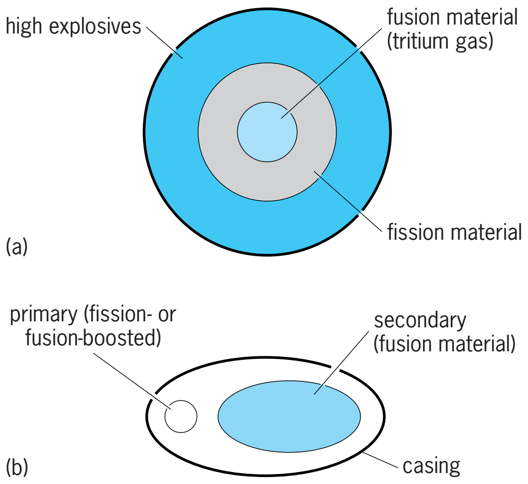
Bibliography. G. F. Bassani, M. Inguscio, and T. W. Hansch (eds.), *The Hydrogen Atom*, 1989; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; J. S. Rigden, *Hydrogen: The Essential Element*, 2003.

Hydrogen bomb

A device in which an uncontrolled, self-sustaining, thermonuclear fusion reaction is carried out in heavy hydrogen (deuterium or tritium) to produce an explosion. In a fusion reaction, the collision of two energy-rich nuclei results in a mutual rearrangement of their protons and neutrons to produce two or more reaction products, together with a release of energy of amount E given by A. Einstein's formula $E = mc^2$, where m is the mass difference between the original and produced nuclei, and c is the velocity of light. *See* NUCLEAR FUSION; THERMONUCLEAR REACTION.

For the hydrogen bomb reaction to become self-sustaining, a so-called critical temperature of about 3.5×10^7 K ($6.3 \times 10^{7\circ}$F) must be attained with the aid of the enormous temperature created by a fission explosive. Once this temperature is achieved, the energy released in the initial reaction maintains the temperature, and the chain proceeds either until the supply of fusionable material is exhausted or until sufficient expansion has taken place that the material is cooled below the critical temperature. When the isotopes fuse, the result is a release of energy and radiation. Some of the radiation consists of fast neutrons, whose role is discussed below. *See* ATOMIC BOMB.

Fusion is a more difficult process than fission to start and propagate because of the high temperatures required—temperatures comparable to those at the center of stars (or in nuclear fission explosions). Also, deuterium and tritium can be produced only in highly specialized facilities. Although a fission explosion requires special isotopes carefully arranged in a weapons configuration, a fusion explosive must have separated isotopes of different light elements, and is much more complex to design and build. There are two ways to use fusion: boosting of fusion explosive yields or generating multistage thermonuclear reactions.


high explosives
fusion material (tritium gas)
fission material
(a)
primary (fission- or fusion-boosted)
secondary (fusion material)
casing
(b)

Design configuration of two types of fusion weapons: (*a*) fusion-boosted atomic bomb, (*b*) multistage hydrogen bomb. (*After A. DeVolpi et al.*, *Born Secret*: *The H-Bomb*, *the Progressive Case and National Security*, *Pergamon*, *1981*)

Fusion-boosted weapons. In a fusion-boosted warhead (**illus.** *a*), when the sphere of fissile materials is compressed (imploded) by the chemical explosion, an uncontrolled fission chain reaction begins. The fissionable material rapidly (in tenths of a millionth of a second) gets as hot as the center of the Sun. If there is fusionable material inside the device, thermonuclear reactions will boost the fission yield. This type of weapon is called fusion- boosted because the fusion reactions do not directly contribute very much to the explosive energy, but instead enhance the fission rate, due to the release of a large number of additional neutrons. Fusion-boosted weapons are militarily more desirable than pure-fission weapons because they are generally lighter, more efficient, and more powerful. *See* ATOMIC BOMB.

Multistage weapons. Multistage thermonuclear weapons are conceptually quite different from fission and fusion-boosted devices. They contain three essential components, which are physically separated from each other (illus. *b*). One component is a small fission or fusion-boosted explosive called a primary or trigger. Separated from the primary is an assembly of lithium-deuteride fusion material called the secondary. Surrounding the primary and secondary is the third major component, a massive casing. Before thermonuclear ignition, neutrons from the exploding primary convert some of the lithium deuteride in the secondary to a fusionable mixture of deuterium and tritium.

Multistage thermonuclear explosive detonation is basically generated in three major phases: ignition caused by the fission-explosive primary stage, coupling of x-radiation to the physically separated secondary fusion stage, and secondary-stage implosion induced by fission x-rays. The radiation-induced implosion of the secondary compresses the lithium-deuteride.

As neutrons from the primary traverse the compressed lithium-deuteride compound, they are most likely to be absorbed in enriched lithium-6, which immediately decomposes into tritium and an alpha particle. If the tritium thus created fuses with deuterium, an energetic alpha particle and a 14-MeV neutron are released. Alpha particles heat the material when they lose kinetic energy.

The fission triggers are themselves highly sophisticated, efficient nuclear explosive devices. The type, chemical form, isotopic content, density, and geometrical arrangement of the fissile materials are important. The exploding trigger emits a great deal of energy as photons, in blackbody radiation. At the temperatures present (about 5×10^7 K or 9×10^7°F), much of that energy is in the soft x-ray region. The photons fill the inside of the casing, traveling hundreds of times faster than the material portions of the exploding primary and other parts of the bomb. The casing behaves at first like a bottle, for a time keeping the energy confined.

There is so much energy released that very large compressive forces are exerted on the secondary, offset by expansive forces on the casing, caused by the shock wave from the primary while it disassembles. Because the casing is massive relative to the fusion package, it moves slowly compared with the rate of compression of the fusion materials. Soon those materials reach densities and temperatures where thermonuclear ignition occurs, liberating many times more energy than that which came originally from the trigger. *See* HEAT RADIATION.

Neutrons from the thermonuclear reactions escape in large numbers from the fusing materials and strike the nuclei in the casing. If the massive casing is made mostly of uranium-238 (natural or depleted uranium), the fusion neutrons will cause the uranium nuclei to undergo fission, giving off still more energy. A device of this sort can be regarded as a three-stage fission-fusion-fission bomb.

Yield. The yield, or total energy, of a hydrogen bomb is expressed in megatons (1 megaton equals 10^{15} calories or 4.18×10^{15} joules). Typical fusion-boosted weapons yield hundreds of kilotons (tenths of megatons), and typical multistage weapons yield megatons. The Soviet Union exploded a bomb with a yield of over 50 megatons in 1961; it is now known that this bomb was designed for a total of 150 megatons. Because a larger proportion of its yield is in radiation than in blast, the neutron-warhead type of thermonuclear weapon can be deployed for use with field artillery. *See* NUCLEAR EXPLOSION.

Complexity. Multistage hydrogen (fusion) bombs are far more complex than the fission weapon or the booster type of fusion weapon. Even with massive national resources, it has required an average of 5 years to develop such weapons after testing a fission explosive.

A. DeVolpi

Bibliography. A. DeVolpi et al., *Born Secret: The H-Bomb, the Progressive Case and National Security*, 1981; C. Hansen, *U.S. Nuclear Weapons: The Secret History*, 1988; H. York, *The Advisors: Oppenheimer, Teller, and the Superbomb*, 1976, reprint 1989.

Hydrogen bond

The interaction which occurs when a hydrogen atom, covalently bonded to an electronegative atom (as in A—H), interacts with another atom to form the aggregate A—H· · ·Y. The shortest and strongest bond is indicated as A—H, while the secondary and weaker interaction is written as H· · ·Y. Thus A—H is a proton donor, while (Y) is a proton acceptor which often contains lone pair electrons and can act as a base. The strongest hydrogen bonds are formed between the most electronegative (A) atoms such as fluorine, nitrogen, and oxygen which interact with (Y) atoms having electronegativity greater than that of hydrogen (C, N, O, S, Se, F, Cl, Br, I). The weakest of hydrogen bonds are formed by acidic protons of C—H groups, as in chloroform and acetylene, and by olefinic and aromatic π-electrons acting as (Y).

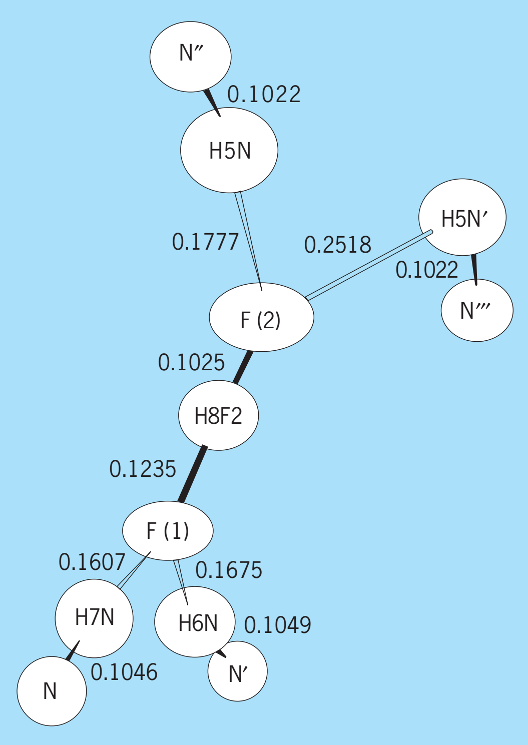
Bond energies. The majority of hydrogen bonds have energies in the range 4–6 kcal/mole (17–25 kilojoules/mole) and involve those between O—H functional groups (as in water, alcohols, or acids) or N—H groups (as in amides or amines) and oxygen atoms (as in water, alcohols, carbonyls, or esters). The strongest hydrogen bond known is that found in the hydrogen difluoride ion, $(F—H—F)^-$, which has been variously estimated at 37–55 kcal/mole (155–230 kJ/mole). Therefore, the average hydrogen bond is of much lower energy than a normal chemical bond (>100 kcal/mole or 418 kJ/mole). Although hydrogen bonding gives rise to a specific interaction between atoms, resulting in a complex with characteristic A—H· · ·Y distances and angles, especially in the solid state, it is difficult to establish a lower limit for the H-bond enthalpy because experimental methods of detection are becoming increasingly more sensitive and accurate.

The weaker the hydrogen bond, the shorter the lifetime of the complex it forms. The detection of weak hydrogen bonds often amounts to measuring shorter and shorter lifetimes of rapidly associating and dissociating species in equilibrium. This is a difficult problem because proton transfer along hydrogen bonds belongs to the fastest known chemical reactions, and in most experimental studies only mean values and "average" structures are determined.

An important aspect of weak hydrogen bond formation is that the different molecular aggregates which do form can be easily and reversibly transformed. Thus the small energy changes resulting in the rapid making and breaking of hydrogen bonds in biological systems are of great importance; for example, hydrogen bonding determines the configuration of the famous α-helix, and the structures of most proteins, thereby serving an important function in determining the nature of all living things. *See* DEOXYRIBONUCLEIC ACID (DNA); PROTEIN.

Spectroscopy. Even though slight energy changes are usually involved in hydrogen bond formation, the aggregate once formed changes almost every measureable physical property of the original species. When investigating hydrogen bonding, the most frequently used techniques are infrared and nuclear magnetic resonance spectroscopy. In the infrared method the A—H stretching frequency is shifted to lower values and is accompanied by band broadening and increased intensity. Such changes are usually easily discernible, but in the case of very strong hydrogen bonds, such as $(F—H—F)^-$, the shift and broadening is so drastic that it is difficult to assign the new frequencies correctly. In the case of nuclear magnetic resonance, hydrogen bonding usually shifts the proton resonance to lower fields. *See* INFRARED SPECTROSCOPY; NUCLEAR MAGNETIC RESONANCE (NMR).

Neutron diffraction. While infrared and nuclear magnetic resonance techniques can yield considerable information about hydrogen bond formation, far greater information content is derived from diffraction studies of crystalline solids. The method of choice is neutron diffraction (versus x-ray diffraction), because in the former case hydrogen atoms scatter almost as well as any other atom while their scattering is often swamped out in the x-ray case. Neutron diffraction crystal structure analysis has become the best probe available for the study of the geometry of A—H···Y bonds. Using this technique, it is generally observed that the A—H separation is ~0.10 ± 0.01 nanometer and is much less than the H···Y separation; that is, hydrogen bonds are usually asymmetric. In the extreme case the hydrogen atom may be equally bonded to both atoms, as in certain O—H—O and $(F—H—F)^-$ containing systems where the atomic environment around A and Y is identical and symmetric. In such systems strong hydrogen bond formation is indicated when the A···Y separation is ~0.02–0.03 nm less than the sum of the van der Waals radii. However, contrary to prior belief, even the shortest and strongest hydrogen bond known, $(F—H—F)^-$, may be asymmetric. This was demonstrated in a neutron diffraction study of *p*-toluidinium hydrogen difluoride in which the two terminal F atoms exist in vastly different F···H—N hydrogen bonding environments (see **illus.**). The F—H distances are unequal, and the $(F—H—F)^-$ ion is asymmetric, because of the very different N—H···F hydrogen bonding environments around the F atoms. Thus it seems that the H atom of a strong $(X—H—X)^-$ bond is a probe of the X atom environment. Indeed, the hydrogen dichloride ion, $(ClH_2)^-$, with a bond energy of about 12 kcal/mole (49 kJ/mole) appears to be symmetric (centered H atom) in some salts and asymmetric in others. Other types of hydrogen atom interactions such as those of M—H···M and C—H···M, where M is a metal, are becoming increasingly important in catalysis. *See* NEUTRON DIFFRACTION.



Hydrogen difluoride ion, $(F—H—F)^-$, geometry. Distances are in nanometers, and atoms are represented as ellipsoids of 50% probability. Numbers following symbols for the elements signify that like atoms are not equivalent structurally. The primes on N atoms indicate that they are structurally equivalent and related by symmetry operations.

Theory. Developments in theory have made it possible to better define certain contributions to hydrogen bond energies. The relative importance of forces of different origin (dispersion, polarization, exchange, coulomb, and so on) have become possible to estimate by using both molecular orbital methods and perturbation theory. In general, it appears that quantum theory gives reliable descriptions of isolated imers and trimers, but fails when dealing with large clusters of the type found in condensed phases. Jack M. Williams

Bibliography. P. Schuster, G. Zundel, and C. Sardorfy (eds.), *The Hydrogen Bond*: *Recent Developments in Theory and Experiments*, vols. 1–3, 1976; M. J. Winter, *Chemical Bonding*, 1994.

Hydrogen fluoride

The hydride of fluorine and the first member of the family of halogen acids. Anhydrous hydrogen fluoride is a mobile, colorless liquid that fumes strongly in air. It has the empirical formula HF, melts at −83°C (−117°F) and boils at 19.8°C (67.6°F). The vapor is highly aggregated, and gaseous hydrogen fluoride deviates from perfect gas behavior to a greater extent than any other gaseous substance known. Aggregate formation in both the vapor and liquid phase arises from unusually strong hydrogen-bond interactions. Hydrogen fluoride is prepared on the large industrial scale by treating fluorspar (calcium fluoride, CaF_2) with concentrated sulfuric acid. The crude product is purified by fractional distillation to yield a product containing more than 99.5% HF; the remaining impurities are principally water and small amounts of sulfur dioxide, silicon tetrafluoride, and boron trifluoride. Very dry hydrogen fluoride can be obtained by electrolysis or by treatment with reagents such as fluorine or cobaltic fluoride that react with water. *See* HYDROGEN BOND.

Properties. Anhydrous hydrogen fluoride is an extremely powerful acid, exceeded in this respect only by 100% sulfuric acid. Like water, hydrogen fluoride is a liquid of high dielectric constant that undergoes self-ionization and forms conducting solutions with many solutes. Because anhydrous hydrogen fluoride is a superacid, many organic solutes dissolve in it to form stable carbonium ions. Alkali metal fluorides and silver fluoride dissolve readily in hydrogen fluoride to form conducting solutions. The alkali metal fluorides are bases in the hydrogen fluoride system and correspond to solutions of alkali metal hydroxides in water. Conversely, antimony pentafluoride and boron trifluoride act as acids in hydrogen fluoride and accentuate the already strong acid properties of the solvent. *See* SUPERACID.

Anhydrous hydrogen fluoride dissolves a wide variety of organic compounds. Oxygen-, nitrogen-, and sulfur-containing compounds usually have high solubility in liquid hydrogen fluoride, generally higher than that found in water. Aromatic hydrocarbons are moderately soluble, and even saturated aliphatic hydrocarbons show appreciable solubility. Despite the fact that hydrogen fluoride is a strong dehydrating agent, many organic solutes can be recovered unchanged from hydrogen fluoride solution. Surprisingly, the enzymes trypsin and lysozyme survive dissolution and recovery from solution in anhydrous liquid hydrogen fluoride with full retention of biological activity.

Uses. Hydrogen fluoride is a widely used industrial chemical. It was formerly used in the petroleum refining industry for the isomerization of aliphatic hydrocarbons to form more desirable automotive fuels, but this application has been superseded by other methods. The largest industrial use of hydrogen fluoride is in making fluorine-containing refrigerants (Freons, Genetrons).

Another important use of hydrogen fluoride is in the preparation of organic fluorocarbon compounds by the Simons electrochemical process. In this procedure, an organic compound is dissolved in hydrogen fluoride, and an electric current is passed through the solution, whereupon the hydrogen atoms in the organic solute are replaced by fluorine. Hydrogen fluoride is employed in the electrochemical preparation of fluorine and for the preparation of inorganic fluorides. Thus, hydrogen fluoride is used for the conversion of uranium dioxide to uranium tetrafluoride, an intermediate in the preparation of uranium metal and uranium hexafluoride. With the great increase in nuclear energy–produced electricity, this represents an important use of hydrogen fluoride.

Important organic reactions may in some cases be performed to advantage in hydrogen fluoride solution, and nitration, sulfonation, diazotization, cyclization, and polymerization reactions have been carried out in this medium. Anhydrous hydrogen fluoride can be used for the carboxylation of olefins with carbon monoxide, for the alkylation of isoparaffins with olefins, for the esterification of fatty acids, and for the removal of protective groups in peptide preparation by solid-phase reactions.

Aqueous solutions of hydrogen fluoride (hydrofluoric acid) are relatively weakly acidic as compared to hydrochloric acid. Fluoride salts are formed by reaction of hydrofluoric acid with metal oxides and carbonates. Of particular importance is the rapid reaction of hydrofluoric acid or anhydrous hydrogen fluoride with silica, which leads to the application of these substances as etching agents for glass.

Both hydrogen fluoride and hydrofluoric acid cause unusually severe burns; appropriate precautions must be taken to prevent any contact of the skin or eyes with either the liquid or the vapor. *See* FLUORINE; HALOGENATED HYDROCARBON; ISOMERIZATION.

Joseph J. Katz

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; P. Henderson, *Inorganic Chemistry*, 1982; A. G. Sharpe, *Inorganic Chemistry*, 2d ed., 1987; H. Tsunoda and M. H. Yu (eds.), *Fluoride Research*, 1985.

Hydrogen ion

A proton combined with a number of water molecules. It is often written as H_3O^+ and called the hydronium ion. However, this species is best considered as an excess proton on a tetrahedral group of four water molecules and so would be designated as $H_9O_4{}^+$. For simplicity, it is most commonly written as H^+ (aq). *See* PROTON.

Chemical activity. Since it is formed by the self-ionization of water according to reaction (1), the hydrogen ion is present in all aqueous solutions. This

$$H_2O \rightleftharpoons H^+(aq) + OH^-(aq) \qquad (1)$$

formation also means that H^+(aq) is always found in the company of the hydroxide ion, OH^-(aq). The relationship between the concentrations of these two species is a very important property of water and at 25°C (77°F) is given by Eq. (2). The equilibrium

$$K_w = [H^+(aq)][OH^-(aq)] = 10^{-14} \qquad (2)$$

constant K_w increases rapidly with increasing temperature, and for more precision the concentrations in Eq. (2) should be replaced by the activities of the ions. *See* IONIC EQUILIBRIUM.

Reaction (1) and Eq. (2) indicate that the H^+(aq) and OH^-(aq) concentrations in pure water are equal to each other with a value of 10^{-7} mole/liter. Any aqueous solution with this concentration of $H^+(aq)$ is called a neutral solution. If the $H^+(aq)$ concentration is greater than 10^{-7} mole/liter, the solution is called acidic. Basic solutions are those in which the $H^+(aq)$ concentration is less than 10^{-7} mole/liter. It is clear from Eq. (2) that as the $H^+(aq)$ concentration increases the $OH^-(aq)$ concentration must decrease, and vice versa. In the most straightforward system, acids are substances that can donate an $H^+(aq)$, and bases are substances that can accept one. Therefore, the action of common acids can be written, for example, as reaction (3) or (4), and the action of common bases can be exemplified by reaction (5).

$$HCl(aq) \xrightarrow{H_2O} H^+(aq) + Cl^-(aq) \qquad (3)$$

$$H_2SO_4(aq) \xrightarrow{H_2O} H^+(aq) + HSO_4{}^-(aq) \qquad (4)$$

$$NH_3(aq) + H^+(aq) \longrightarrow NH_4{}^+(aq) \qquad (5)$$

Metal cations can also act as acids, particularly when they are small ions with high positive charges. An example of such a hydrolysis reaction is shown in reaction (6). The reaction of an acid and a base is called

$$Al^{3+}(aq) \rightleftharpoons AlOH^{2+}(aq) + H^+(aq) \qquad (6)$$

a neutralization reaction and has great importance. A typical example of a neutralization is reaction (7).

$$HCl(aq) + NaOH(aq) \longrightarrow NaCl(aq) + H_2O \qquad (7)$$

The products of such a neutralization are always a salt, in this case NaCl, and water. *See* ACID AND BASE.

The reaction of $H^+(aq)$ with bicarbonates, carbonates, sulfites, bisulfites, and sulfides produces the volatile gases carbon dioxide, sulfur dioxide, and hydrogen sulfide, respectively. The hydrogen ion also reacts with metals above hydrogen in the electromotive force series to produce hydrogen gas and the cation of the metal [reaction (8)].

$$Zn(s) + 2H^+(aq) \longrightarrow H_2(g) + Zn^{2+}(aq) \qquad (8)$$

See ELECTROCHEMICAL SERIES.

Concentration. Hydrogen ion concentration determines the course of many chemical reactions that occur in living organisms and in the chemical industry. The control of hydrogen ion concentration is achieved in living organisms and in the laboratory by buffer systems. These are chemical mixtures designed to resist change in hydrogen ion concentration. Careful control of acid concentration is crucial in industries such as brewing, pharmaceutical manufacturing, electroplating, and textiles. Water sanitation, whether in a swimming pool or a water treatment plant, demands close attention to acidity. *See* BUFFERS (CHEMISTRY).

Electrical conductivity. Another property of the $H^+(aq)$ ion is important in both theoretical and practical ways. $H^+(aq)$ is the best conductor of electricity of any ion in aqueous solution. Its conductance at 25°C (77°F) is almost five times as large as the next-most-conducting ion. This abnormally large conductance is due to a unique mechanism of electricity conduction. Ordinary ions must travel physically through the solution to conduct electricity. The $H^+(aq)$ conducts electricity by a chain mechanism whereby an excess proton can be attached to one side of a water molecule cluster, and another excess proton lost from the opposite side of the same cluster. This high conductivity is of great practical importance in batteries such as the common lead-acid cell, where the sulfuric acid electrolyte keeps the internal cell resistance low. *See* ELECTROLYTIC CONDUCTANCE.

Relation between ion concentrations and pH in water

$[H^+(aq)]$	$[OH^-(aq)]$	pH	$[H^+(aq)]$	$[OH^-(aq)]$	pH
1	1×10^{-14}	0	1×10^{-7}	1×10^{-7}	7
1×10^{-1}	1×10^{-13}	1	1×10^{-8}	1×10^{-6}	8
1×10^{-2}	1×10^{-12}	2	1×10^{-9}	1×10^{-5}	9
1×10^{-3}	1×10^{-11}	3	1×10^{-10}	1×10^{-4}	10
1×10^{-4}	1×10^{-10}	4	1×10^{-11}	1×10^{-3}	11
1×10^{-5}	1×10^{-9}	5	1×10^{-12}	1×10^{-2}	12
1×10^{-6}	1×10^{-8}	6	1×10^{-14}	1	14

pH. The hydrogen ion concentration can vary over fourteen powers of 10. To avoid dealing with such exponentials, the concept of pH was advanced by S. P. L. Sørensen in 1909. It is defined by Eq. (9). In

$$pH \equiv -\log[H^+(aq)] \qquad (9)$$

precise work the activity of the hydrogen ion would be used to replace the concentration in Eq. (9). Since all aqueous solutions contain both hydrogen ion and hydroxide ion, it is possible to define all degrees of acidity and basicity on the pH scale. The **table**, which gives hydrogen ion concentrations, hydroxide ion concentrations, and pH for aqueous solutions at 25°C (77°F), illustrates certain characteristics of the pH scale. Neutral solutions have a pH of 7, while acid solutions have a pH less than 7, and basic solutions have a pH greater than 7. A change of one pH unit corresponds to a tenfold change in acidity. The pH is also useful in pointing out the differences between strong acids such as hydrochloric acid and weak acids such as acetic acid, the main component of vinegar. For a strong acid, reaction (10) is almost 100% complete.

$$HCl(aq) + H_2O \longrightarrow H^+(aq) + Cl^-(aq) \qquad (10)$$

Therefore, a 0.1 mole/liter solution of HCl has a pH of 1.0. For the weak acid, reaction (11) occurs to

$$CH_3COOH(aq) + H_2O \rightleftharpoons H^+(aq) + CH_3CO\overline{O}(aq) \qquad (11)$$

a very limited extent. A 0.10 mole/liter solution of acetic acid has a pH of 2.9 and thus is almost 100 times less acidic than the hydrochloric acid solution. *See* PH.

Determining concentrations. Two general methods are used for the determination of hydrogen ion concentrations. For relatively crude work, colorimetric methods are commonly used. These methods depend on the fact that certain natural and synthetic dyes have colors that depend on the hydrogen ion concentration. At times, paper is impregnated with such an indicator. By dipping the paper into the aqueous solution, the color of the paper changes due to the acidity of the solution. The paper can then be compared to a standard printed color series, and a rough idea of the pH obtained. In another application, a small amount of such an indicator dye can be added to a solution whose acidity is being determined. Small increments of a base solution are then added in a titration. If the indicator is chosen correctly, the solution will change color when the exact amount of base has been added to completely neutralize the original acid. A simple calculation will then give the amount of acid initially present. *See* ACID-BASE INDICATOR.

In most precise work, a potentiometric method is used for the determination of hydrogen ion concentration. This method depends on an electrode whose potential is sensitive to hydrogen ion concentration. The only electrode commonly in use for practical pH measurements is the glass electrode. A glass electrode, together with a reference electrode, is placed in the solution of unknown acidity. The potential of the electrochemical cell formed is then measured by a high-input-impedance voltmeter.

Many good commercial pH meters are available. Because it is difficult to relate the measured voltage directly to a hydrogen ion concentration, the pH meter–electrode combination is calibrated with buffer solutions of well-defined pH values. The calibrated instrument can then be used to measure the pH of unknown solutions. The importance of pH measurement led to the creation of very precisely defined pH standards at the National Bureau of Standards. It is now possible to make meaningful pH measurements, even in extreme conditions such as very high pressures (1000 bars or 100 megapascals) or very high temperatures (200°C or 390°F). The technology of pH measurement has advanced to the point where many industrial processes are automatically controlled through pH measurements on process streams. *See* ION-SELECTIVE MEMBRANES AND ELECTRODES; TITRATION.

G. Atkinson

Bibliography. J. O'M. Bockris and A. K. N. Reddy, *Modern Electrochemistry*, vol. 1, 1970; S. A. Borman (ed.), *Instrumentation in Analytical Chemistry*, 1982; G. Eisenman (ed.), *Glass Electrodes for Hydrogen and Other Cations*, 1967; G. K. McMillan, *pH Measurement and Control*, 2d ed., 1993; J. E. Ricci, *Hydrogen Ion Concentration*, 1952; L. Wilson, *Quantitative Analysis: Gravimetric, Volumetric and Instrumental Analysis,* 2d ed., 1989.

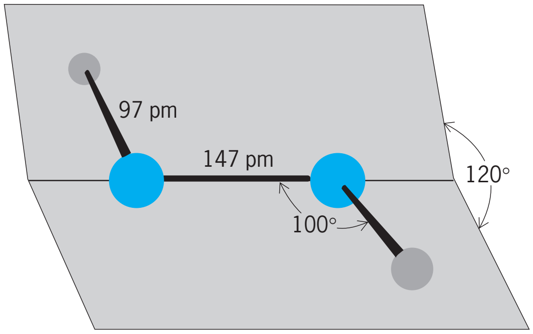
Hydrogen peroxide

A binary compound of hydrogen and oxygen, empirical formula H_2O_2, used mostly in dilute aqueous solutions as an oxidizing agent. It was discovered in 1818 by the French chemist Louis-Jacques Thenard, who named it *eau oxygénée*. Its most remarkable feature is its tendency to decompose readily into water and oxygen, the first observed instance of contact catalysis.

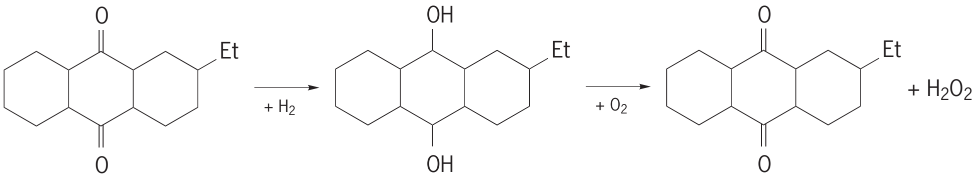
Properties. Anhydrous hydrogen peroxide is a clear, colorless liquid, of nearly the same viscosity and dielectric constant as water, but of greater density, 1.442 g · cm^{-3} at 25°C (77°F). Like water, it is strongly associated through hydrogen bonds. It boils at 150°C (302°F) with violent, sometimes explosive decomposition. With considerable supercooling, it can be frozen into needle-shaped crystals which melt at −0.41°C (31°F). The decomposition, strongly exothermic (690 cal · g^{-1} or 2.89 kJ · g^{-1}), is almost always catalytic, becoming homogeneous only in the vapor above 420°C (790°F). The rate varies considerably with the nature of the catalyst, the temperature, and the surface-volume ratio of the sample. Decomposition by light begins only in the near ultraviolet. As a solvent, hydrogen peroxide resembles water, except that acids and bases show much lower electrical conductivity. Although a fairly strong oxidant, it can act as a mild reducing agent, for example, with permanganates and perchromates.

The H_2O_2 molecule has the skew chain configuration shown in the **illustration**, the simplest example of internal rotation or torsion about a single bond. This is a semirigid structure, because the torsion is hindered by a rather low potential barrier (1 kcal · mol^{-1} or 4 kJ · mol^{-1} for the trans versus 7 kcal · mol^{-1} or 29 kJ · mol^{-1} for the cis configuration).

Formation. Hydrogen peroxide occurs only in traces in nature, mostly in rain and snow. It has not yet been detected in interstellar space. Its formation, usually in low concentration, has been studied in a variety of systems: (1) from the elements in the products of the oxyhydrogen flame quickly quenched in liquid air, or in explosion of hydrogen-rich mixtures near 550°C (1020°F); (2) from water, liquid or vapor, irradiated by ultraviolet light of a wavelength shorter than 185 nm, with and without sensitizer, and by ionizing radiation; and (3) from a silent electrical discharge through water vapor at atmospheric pressure, or under reduced pressure (below 10^{-3} atm or 10^{2} pascals) in electrodeless discharges through a fast-flowing system, followed by quick chilling in a liquid-nitrogen trap. Under optimum conditions, concentrations up to 50% can be reached. This last



Structural parameters of the H_2O_2 molecule.



(**1**) (**2**) (**1**)

system, developed for the synthesis of deuterium peroxide, D_2O_2, from heavy water, has led to identification of the long-postulated higher oxides, H_2O_3 and H_2O_4, as metastable intermediates.

Manufacture. The original barium peroxide–sulfuric acid method was superseded long ago by the electrochemical process, which involves oxidation of sulfuric acid, or ammonium, or potassium sulfate in concentrated solutions at high current density on a platinum anode. This also is now largely replaced by the anthraquinone process based on the cyclic oxidation-reduction of some substituted anthraquinone as shown in the reaction above. The anthraquinone derivative (**1**) dissolved in an appropriate solvent is first reduced catalytically by hydrogen at atmospheric pressure. The resulting hydroquinone (**2**) is then oxidized by air, and the hydrogen peroxide (5 to 15 g per liter) is removed by countercurrent extraction with water, thereby regenerating the working substance. The efficiency of the latter step and the avoidance of side reactions are essential for economical operation.

Uses. Hydrogen peroxide is used for bleaching cotton and other fibers, natural or synthetic, and in the pulp and paper industry. Because of its milder action on fibers and the fact that it leaves no undesirable residue, it is preferable to chlorine and its compounds. Its cosmetic use as hair bleach consumes relatively little of the commercial 10% (30 volume) solution. In medicine it is useful for cleansing wounds and cuts, although its antiseptic action is rather slow. A limited but important use of the concentrated peroxide is for energy production in rockets, submarines (during submersion), airplanes (at takeoff), and the steering of space vessels. Fast decomposition is achieved by sudden addition of a catalyst or by blowing the vapor through a porous bed of catalyst-impregnated ceramics. The jet from a 90% solution can reach temperatures of about 750°C (1380°F). *See* ANTISEPTIC; BLEACHING; CHEMICAL FUEL; STERILIZATION.

Handling. Hydrogen peroxide, especially when concentrated, requires great care in handling and storing. When dropped on paper or wood, it can start a fire. Contact with the skin causes blotches that can be painful, but they disappear after a few hours without leaving traces. Appropriate containers are made of Pyrex glass, poly(tetrafluoroethylene), or polyethylene; for industrial storage and shipment, tanks or carboys are made of electropolished stainless steel or pure aluminum. All containers must be fitted with a vent for the escape of oxygen, and be of such design as to prevent spilling.

Slight decomposition may be overcome by distillation, or by addition of some stabilizer (such as zinc salts, phosphates, or ascorbic acid). For special purposes, a 90% product, stabilizer-free, is sometimes available. It can be concentrated to over 99% by fractional distillation in an all-Pyrex-glass still. The last traces of water are easily removed by fractional crystallization, provided that the solid is frozen very slowly into a large single crystal. *See* HYDROGEN; OXYGEN; PEROXIDE.

Paul A. Giguère

Bibliography. W. C. Schumb, C. N. Satterfield, and R. L. Wentworth, *Hydrogen Peroxide*, ACS Monogr. 128, 1955; G. Strykul (ed.), *Catalytic Oxidations with Hydrogen Peroxide as an Oxidant*, 1993.

Hydrogenation

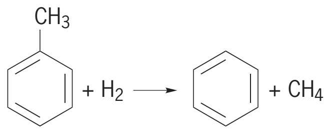
The chemical reaction of hydrogen with another substance, generally an unsaturated organic compound, and usually under the influence of temperature, pressure, and catalysts. There are several types of hydrogenation reactions. They include (1) the addition of hydrogen to reactive molecules; (2) the incorporation of hydrogen accompanied by cleavage of the starting molecules (hydrogenolysis); and (3) reactions in which isomerization, cyclization, and so on, result. Other reactions that involve molecular hydrogen and catalysts are reductive amination (hydroammonolysis) and hydroformylation (oxo reaction).

Hydrogenation is synonymous with reduction in which oxygen or some other element (most commonly nitrogen, sulfur, carbon, or halogen) is withdrawn from, or hydrogen is added to, a molecule. When hydrogenation is capable of producing the desired reduction product, it is generally the simplest and most efficient procedure.

Hydrogenation is used extensively in industrial processes. Important examples are the synthesis of methanol, liquid fuels, hydrogenated vegetable oils, fatty alcohols from the corresponding carboxylic acids, alcohols from aldehydes prepared by the aldol reaction, cyclohexanol and cyclohexane from phenol and benzene, respectively, and hexamethylenediamine for the synthesis of nylon from adiponitrile.

Hydrogenolysis. This term refers particularly to cleavages in a molecule associated with the addition of hydrogen. Hydrogenolysis is analogous to hydrolysis and ammonolysis, which involve cleavage of a bond induced by the action of water and ammonia, respectively. Chemical bonds which are broken by hydrogenolysis reactions include carbon-carbon, carbon-oxygen, carbon-sulfur, and carbon-nitrogen.

Two examples are hydrodealkylation of toluene to form benzene and methane, reaction (1), and reduction of methyl laurate to form lauryl alcohol and methanol, reaction (2).



(1)

$$CH_3\,(CH_2)_{10}\,COOCH_3 + H_2 \longrightarrow CH_3\,(CH_2)_{10}\,CH_2OH + CH_3OH \quad (2)$$

Catalytic hydrogenation. A variety of organic compounds can be hydrogenated easily in the presence of a catalyst. Acetylenes readily add two moles of hydrogen giving the saturated derivatives, as shown in reaction (3), where R and R′ are aliphatic, aromatic, or certain other groups. Under proper conditions the hydrogenation can be stopped at the intermediate olefin stage.

$$\begin{aligned} RC \equiv CR' + 2H_2 &\longrightarrow RCH = CHR' + H_2 \\ &\longrightarrow RCH_2CH_2R' \end{aligned} \quad (3)$$

Catalytic hydrogenation of olefins can be carried out either in gas or in liquid phase, depending on their molecular weights. A nickel-containing catalyst and sometimes platinum or palladium catalysts are employed.

Aromatic compounds may be reduced either in the vapor phase at atmospheric pressure or in the liquid phase at hydrogen pressures up to 200 atm (2×10^4 kilopascals). In the latter case, aromatics, such as benzene, toluene, and *p*-cymene, can be hydrogenated readily in the presence of a nickel catalyst. In the case of naphthalene or substituted naphthalenes, the product may be the tetra- or decahydronaphthalene derivative.

The reduction of carbonyl compounds, such as aldehydes and ketones to the corresponding alcohols is represented by reaction (4), where R is an aliphatic or aromatic group, and R′ may be the same group or a hydrogen atom. Frequently when R is an aromatic group, it is difficult to stop the reduction at the alcohol. Instead, it proceeds further to yield a hydrocarbon, RCH_2R'. In general, aldehydes are reduced more rapidly than ketones, although there are numerous examples in which both undergo reduction at room temperature and only a few atmospheres of hydrogen pressure. Often a small amount of water (1–10%) is added to the feed to the hydrogenator to suppress ether formation.

$$RCOR' + H_2 \rightarrow RCHOHR' \quad (4)$$

Other processes. Hardening of various animal fats and vegetable oils (such as soybean, cottonseed, fish, whale, and peanut) is carried out on a large scale by partial hydrogenation. The resultant plastic fats have a consistency and other properties suitable for the manufacture of shortenings, margarine, soaps, and a variety of other edible and industrial products. Chemically, the process involves the conversion of glycerides of unsaturated fatty acids (for example, oleic and linoleic) to saturated ones. Mild conditions (such as 212–1380°F or 100–750°C, 1–14 atm or $1\text{–}14 \times 10^2$ kPa and Ni catalysts) are employed to avoid the hydrogenolysis of the ester linkage. The conversion of olein to stearin may be expressed as reaction (5).

$$(C_{17}H_{33}COO)_3C_3H_5 + 3H_2 \longrightarrow (C_{17}H_{35}COO)_3C_3H_5 \quad (5)$$

Unlike the hardening of fats, which involves only the hydrogenation of ethylene linkages, the hydrogenolysis of the carboxyl group of acids and esters takes place with the formation of alcohols, as shown by reaction (6). The olefinic bonds in the fatty chain may or may not be reduced. A reduced copper–ammonium chromate catalyst is used at 660–750°F (350–400°C).

$$RCOOH + 2H_2 \longrightarrow RCH_2OH + H_2O \quad (6)$$

The synthesis of methanol from carbon monoxide and hydrogen is carried out at high pressures (3000–5000 lb/in.^{2} or 20,600–34,500 kPa), because the reaction (7) involves a decrease in volume. The practical reaction temperature range is small. Below 570°F (300°C) the rate is slow; above 750°F (400°C) the equilibrium becomes unfavorable. Mixed catalysts consisting of oxides of zinc, chromium, manganese, or aluminum, such as zinc oxide with 10% chromium oxide, are utilized. Carbon monoxide can also be hydrogenated to give various hydrocarbons and higher alcohols.

$$CO + 2H_2 \rightarrow CH_3OH \quad (7)$$

Petroleum, tar, and coal are hydrogenated to (1) improve existing products; (2) convert low-grade materials such as heavy oils into valuable fuels; and (3) transform solid fuels such as lignites and coal into liquid fuels. Several of these processes are hydrodesulfurization, hydrocracking, and hydrodealkylation. By a proper selection of catalysts and operating conditions, such hydrogenations can be directed to give desired end products and at the same time cause impurities that are common catalytic poisons, such as sulfur, nitrogen, and oxygen, to be detached from their molecular linkages and to be removed as hydrogen sulfide, ammonia, and water. *See* HYDROCRACKING.

Catalytic hydrogenation may continue to increase in importance, size, and the variety of processes. Examples are possible commercialization of coal-hydrogenation processes and the conversion of petroleum residues and shale oils to lighter fractions. *See* COAL GASIFICATION; COAL LIQUEFACTION.

Thermodynamics. Hydrogenation reactions are generally reversible. Catalysts affect the rate or speed of reaction, but have nothing to do with the inherent tendency of the reaction to proceed. To know whether or not the reaction is feasible, the free-energy change of a reaction, ΔG, can be determined. For example, for the reaction represented by (8), the change in free-energy content per

$$CO(g) + 3H_2(g) \rightleftharpoons CH_4(g) + H_2O(g) \qquad (8)$$

atom of carbon in each molecule at 750°F (400°C) is shown in Eqs. (9) and (10).

$$-40,500 + 0 + \Delta G = -\,3600\; -\; 50,000 \qquad (9)$$

$$G = -13,100 \text{ cal at } 400^\circ\text{C } (1 \text{ cal} = 4184 \text{ J}) \qquad (10)$$

The decrease in free energy means that, at this temperature, reduction of carbon monoxide to methane is possible. At 1830°F (1000°C), the reverse reaction for the production of hydrogen and carbon monoxide from natural gas and steam is feasible. It so happens that both the forward and the reverse reactions are industrially important. *See* FREE ENERGY.

Hydrogenation reactions are exothermic, that is, heat is released during the reaction. Typically, the heat release per gram mole of hydrogenated material formed is about 28–30 kcal (117–126 kilojoules) for the hydrogenation of alkenes to alkanes and is 50 kcal (209 kJ) for the hydrogenation of benzene to cyclohexane. The heat of reaction must be removed by heat exchangers in the reactor or is utilized to heat the feeds to the reaction temperature.

Effect of temperature. The reaction temperature affects the rate and the extent of hydrogenation as it does any chemical reaction. Practically every hydrogenation reaction can be reversed by increasing temperature. High temperatures often lead to loss of selectivity and, therefore, yield of desired product, if a second functional group is present. As a practical measure, hydrogenation is carried out at as low a temperature as possible compatible with a satisfactory reaction rate. Although the optimum temperature depends on the catalyst type and age, the temperatures for hydrogenation reactions are generally below 930°F (500°C).

Effect of pressure. Hydrogenation rates are generally increased by increasing the hydrogen pressure. Pressure also increases the equilibrium yield in hydrogenations where there is a decrease in volume as the reaction proceeds. For economic reasons, many industrial hydrogenation processes are carried out under an imposed pressure but seldom above 300 atm (3×10^2 kPa).

Catalysts. For industrial applications, hydrogenation catalysts are generally solids consisting of metals, metal oxides, and some salts. These catalysts may be classified in accordance with their customary use. Vigorous catalysts suitable for the hydrogenation of alkyne and alkene linkages, aldehydes, and ketones include nickel and cobalt, and molybdenum and tungsten oxides or sulfides. Mild catalysts, useful for stepwise hydrogenations of aldehydes and ketones include oxides of copper, zinc, and chromium, and metallic platinum and palladium. Molybdenum sulfide and especially tungsten disulfide are active catalysts for operations at 3000 lb/in.^{2} (20,600 kPa). These catalysts are useful for the hydrogenation of unsaturates and to effect the cleavage of C-C, C-O, and C-N bonds. *See* CATALYSIS.

A wide range of metal ions and complexes has been found to catalyze hydrogenation reactions homogeneously in solution. These ions and complexes have been derived from a variety of metals, including platinum, cobalt, rhodium, and copper. Homogeneous catalytic systems are inherently simpler chemically and kinetically, and are often more selective. Judging from the patent activity in this area, the use of homogeneous catalysts to effect hydrogenation shows considerable promise. *See* HOMOGENEOUS CATALYSIS.

Equipment. There are two common types of reaction vessels. The first is used with liquids or solids, as in the hydrogenation of oils and viscous hydrocarbons. Internal agitators bring about an intimate mixing of organic compound, catalyst, and hydrogen. Alternatively, hydrogen is kept dispersed in the oil by recirculating the gas extracted from the head space of the reactor to the bottom by means of a blower. Usually, these are batch processes, although continuous mode of operation is becoming more attractive. The second type of reactor resembles a column or tube containing a fixed bed of catalyst, and is used where the organic compound has sufficient vapor pressure at the reaction temperature, as in the synthesis of methanol from carbon monoxide, to permit gas-phase, continuous operations.

The design and construction of process equipment which can withstand hydrogen gas at high temperature or high pressure, or both, are complicated. Alloy steels are the most common materials of construction. *See* DEHYDROGENATION; FISCHER-TROPSCH PROCESS; HIGH-PRESSURE PROCESSES; HYDROFORMYLATION; HYDROGEN; ORGANIC SYNTHESIS; OXIDATION-REDUCTION.

Roberto Lee

Bibliography. M. Freifelder, *Catalytic Hydrogenation in Organic Synthesis: Procedures and Commentary*, 1978; B. R. James, *Homogeneous Hydrogenation*, 1973; H. P. Patterson, *Hydrogenation of Fats and Oils*, 1983; P. N. Rylander, *Hydrogenation Methods,* 1985; M. J. Satriana (ed.), *Hydroprocessing Catalysts for Heavy Oil and Coal*, 1982.

Hydrography

The science of measuring and describing the physical features and conditions of navigable waters and adjoining coastal areas. Hydrography is an applied science involving the study of marine areas, including oceans, rivers, and lakes. It involves geodesy, physical oceanography, marine geology, geophysics, photogrammetry (in coastal areas), remote sensing, and marine cartography. Basic parameters observed during a hydrographic survey are time, geographic position, depth of water, and bottom type. However, observation, analysis, and prediction of tides and currents area are also normally included in order to reduce depth measurements to a common vertical datum. *See* GEODESY; PHOTOGRAMMETRY.

A principal objective of hydrography is to provide for safe navigation and protection of the marine environment through the production of up-to-date nautical charts and related publications. In addition, hydrographic data are essential to a multitude of other activities such as global studies, for example, shoreline erosion and sediment transport studies; coastal construction; delimitation of maritime boundaries; environmental protection and pollution control; exploration and exploitation of marine resources, both living and nonliving; and development of marine geographic information systems (GIS). *See* GEOGRAPHIC INFORMATION SYSTEMS; NAVIGATION.

The International Hydrographic Organization (IHO), which is intergovernmental, coordinates the activities of the national hydrographic offices, develops sciences in the field of hydrography and physical oceanography, and achieves the greatest possible uniformity in nautical charts and documents.

Electronic position-fixing systems provide continuous availability of the necessary data, which previously were limited to line-of-sight observations. The electronic systems are usually classified in terms of their range capability, that is, long, medium, or short range. The long-range (Loran) systems have accuracies far offshore to a fraction of a nautical mile. Medium-range systems are typically comparison systems, extending 100–200 mi (160–320 km) offshore, and their accuracies are in the range of several tens of meters. Short-range systems are typically high-frequency, with the data converted to the elapsed time for a pulse to travel between two units into a measured distance; accuracy is in the range of meters. Distances measured from two known stations on land to the mobile unit (ship or launch) then provide the necessary position. Laser measurements provide highly accurate but short-range distance measurements (centimeters). *See* LORAN.

Satellite navigation. The U.S. Department of Defense satellites of the NAVSTAR GPS (Navigation System with Time and Ranging Global Positioning System) provide worldwide, continuous, and precise three-dimensional real-time positioning. Another global satellite system, GLONASS, is operated by Russia. Many satellite navigation receivers are compatible with both GPS and GLONASS. By using the differential technique, wherein a shore-based fixed unit is used to monitor variations of the satellite signal that are broadcast to the mobile unit to refine the ship%s position, positions can be obtained that are accurate to within less than 33 ft (10 m). Submeter accuracies are obtainable by monitoring the phase of the carrier signal from the satellites. This advance has revolutionized hydrography, making it possible to survey the world%s oceans with great accuracy. An added advantage is that satellite systems provide positions referenced with a common horizon datum rather than the regional datum of the past. *See* AIR NAVIGATION; MILITARY SATELLITES; SATELLITE NAVIGATION SYSTEMS.

Sonar. Modern depth information is achieved with sonar measurements. Dual-frequency echo sounders are used, with a high-frequency, narrow beam to measure the depth below the vessel, and a lower-frequency, wider beam to obtain larger coverage of the terrain. Ships used for offshore surveys and boats used for nearshore shallow-water surveys usually make observations along parallel sounding lines that run perpendicular to the depth contours, that is, usually perpendicular to the beach, which provides for better information for interpreting the bathymetry. The spacing of the lines depends mainly on the scale of the survey and the sea bottom morphology. The principal deficiency of this method is the possibility that a hazard may be undetected between the sounding lines.

Side-scan sonar, an instrument that transmits acoustic signals obliquely through the water, is normally towed behind the survey vessel and displays the returning echoes via an onboard graphic recorder. Although this technique does not allow exact determination of position and depth (both can be approximated), it provides excellent resolution with a depiction with what lies to either side of the vessel. As towing must be done at slower-than-normal sounding speeds and the area of coverage varies according to height of the towed sensing element (known as towfish) above the sea floor and the sonar frequency, this technique is limited in its speed of survey. Multibeam side-scan sonars provide for increased numbers of sonar returns, thereby allowing the vessel speed to be increased while still detecting small objects and providing improved depiction of the sea-floor texture.

Multibeam hydrographic survey systems consist of hull-mounted arrays such that a fan-shaped array of sound beams is transmitted perpendicular to the direction of the ship%s track. This provides for the possibility of 100% coverage of the sea floor. The resolution of the system depends on the size of the sonar beam (beam angle) and depth of the water. Since beams are transmitted at an angle through the water column, refraction becomes a problem. While all sonar systems need information about the structure of the water column so as to properly account for the speed of sound in seawater (affecting depth), with multibeam systems both depth and position of the depth measurement are affected. Also, with multibeam systems it is important to measure the attitude of the vessel (heave, roll-pitch, and yaw). Such systems cover a swath width of typically two to six times the water depth. *See* ECHO SOUNDER; SONAR.

Airborne systems. Laser airborne systems mounted in fixed-wing aircraft or helicopters are also available for hydrographic surveys. The system emits a two-color laser beam, usually green and red, such that a return is received from the surface of the water by the red laser and from the bottom by the lower-frequency green laser, allowing the depth to be determined from the time difference. They can be operated in depths down to 165 ft (50 m), but more normally to 66 ft (20 m), depending on water clarity. High turbidity can preclude use of this method. Resolution depends on the scattering of the laser beam as it passes through the water column, and it may not detect the shoal point of a hazard. The great advantage of this system is its speed of survey in coastal waters. It can survey an area at a rate of perhaps

60 times that of a traditional vessel and may be quickly deployed to different areas. *See* LASER.

Hydrographers use tide-coordinated aerial photography to delineate the high and low water lines for charting, which in turn is used for base-line determination of offshore boundaries. Satellite positioning of the aircraft using the Global Positioning System with carrier phase measurement and postprocessing of the data provides for determination of the position of the aircraft of the decimeter level, significantly reducing the need for a ground crew and geodetic control. Thus, metric information can be obtained in areas to which geodetic control has not or cannot (because of ice covering) be extended. *See* AERIAL PHOTOGRAPHY.

Digital data. Hydrographers provide suitable data to feed the Electronic Chart Display and Information System (ECDIS), which may be used instead of paper charts. The resolutions of the International Maritime Organization (IMO), taken from the approved standards of the International Hydrographic Organization, regulate the carriage requirements of electronic charts and their use. *See* MARINE GEOLOGY; OCEANOGRAPHY.

Christian Andreasen; Giuseppe Angrisano

Bibliography. Hydrographic Society, *Hydrographic Journal*, quarterly; A. E. Ingham, *Hydrography for the Surveyor and Engineer*, 3d ed., 1993; IMO Assembly resolutions A.817(19) and MSC 64(67), MSC 86(70), International Hydrographic Bureau, *International Hydrographic Bulletin*, monthly.

Hydroida

An order of the cnidarian which includes the fresh-water hydras, the attached and usually colonial hydroids, and many of the smaller jellyfish. It is the largest order of the class Hydrozoa.

Taxonomy. The order Hydroida includes two principal suborders, Gymnoblastea and Calyptoblastea; the descriptions in this article apply mainly to them. The Gymnoblastea are those hydroids which lack protective cups around the hydranths and gonozoids. Jellyfish produced by these athecate hydroids are Anthomedusae. The Calyptoblastea include the hydroids with protective cups around the hydranths (hydrothecae) and around the gonozoids (gonothecae). Jellyfish of these thecate hydroids are called Leptomedusae.

Two minor suborders are Limnomedusae and Chondrophora. Limnomedusae include a few species of fresh-water jellyfish (such as *Craspedacusta*) which have small and simple polyp stages in their life history. The marine genus *Gonionemus* also belongs to this suborder. Chondrophora are animals of the open sea; the best known is *Velella*, the by-the-wind sailor or purple sail. Because they possess floats, they have sometimes been included with the Siphonophora. A direct comparison, however, may be made with the structure of a symnoblastic polyp whose stem region is used for attachment, but for the chondrophoran the stem region is modified to become the float. The medusae which are produced by budding resemble anthomedusae.

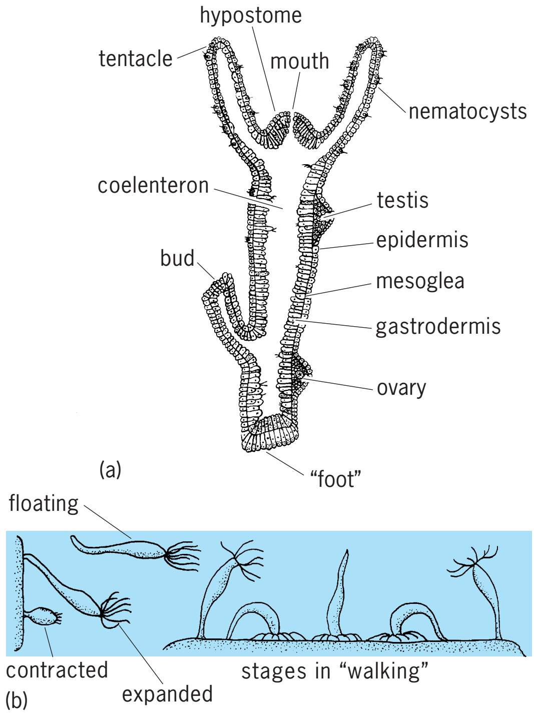


Fig. 1. Hydra. (*a*) Longitudinal section. (*b*) Movements. (*After T. I. Storer and R. L. Usinger, General Zoology, 4th ed., McGraw-Hill, 1965*)

Morphology. Anthomedusae are typically ovoid jellyfish, often with eyespots. Leptomedusae are usually flattened or saucer-shaped, have statocysts (sense organs of balance), and lack eyespots. The gonads of Anthomedusae are generally on the wall of the stomach just above the mouth, and those of Leptomedusae below the radial canals.

Young hydranths of gymnoblastic hydroids are small with five or more tentacles, but subsequently grow much larger and add more tentacles. Calyptoblastic hydranths, in contrast, emerge from a bud

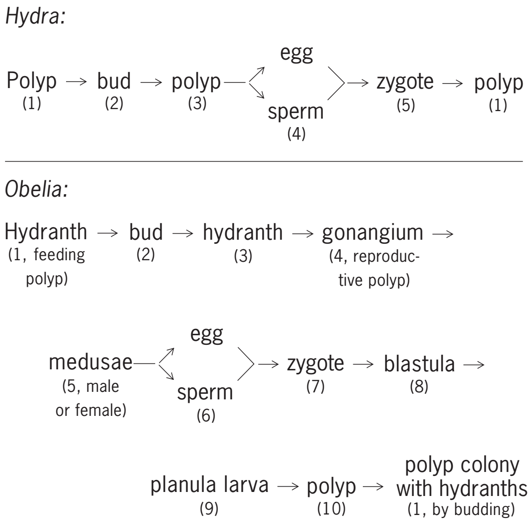


Fig. 2. Life cycle stages of *Hydra* and *Obelia*.

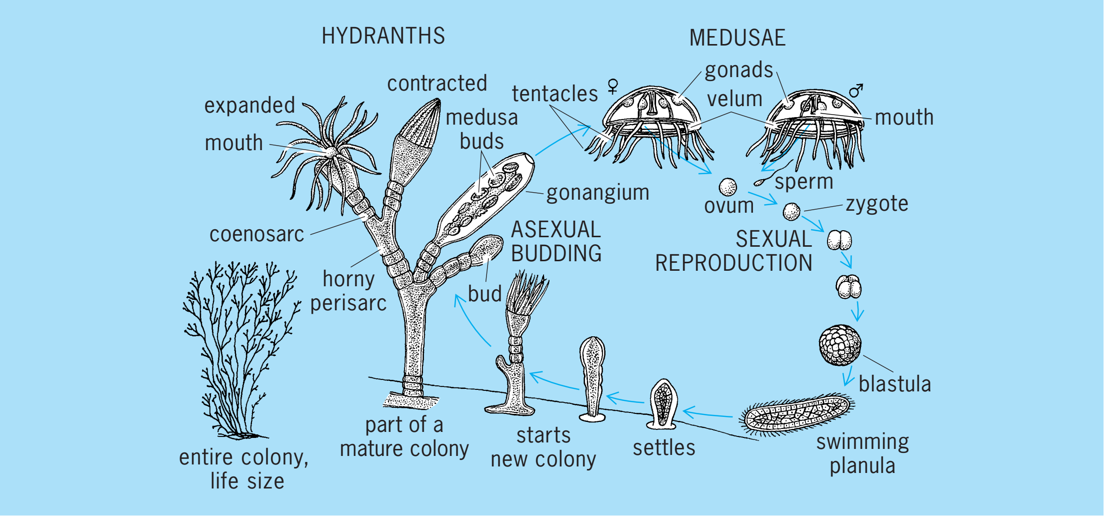


Fig. 3. **Diagram illustrating the life cycle of *Obelia*, a polymorphic colony of polyps which has a medusoid state. (*After T. I. Storer and R. L. Usinger, General Zoology, 4th ed., McGraw-Hill, 1965*)**

with a full complement of parts; they do not grow, live for only about a week, undergo regression and absorption by the colony, and are then replaced by new hydranths.

The fresh-water hydras are simple, motile polyps which do not produce colonies (**Fig. 1**). Buds separate from the parent and become individual polyps. Simple gonads develop on the body and there is no medusa. Hydras are sometimes included in the Gymnoblastea and sometimes placed in a suborder by themselves, the Hydrida.

Life cycle. The fertilized egg usually develops into a free-swimming ciliated larva, the planula, which soon attaches to some support and develops a mouth surrounded by tentacles at its free end. This attached stage is called a polyp and produces stolons, stems, and further polyps which remain connected to make up a hydroid or hydroid colony. Such colonies may show elaborate patterns of branching and may consist of hundreds of polyps, also called hydranths. Buds develop on the colony and are liberated as jellyfish or medusae which produce the sperm and eggs. *See* INVERTEBRATE EMBRYOLOGY.

Although the life cycle described for *Obelia geniculata* is regarded as typical, many species show reduction of either the polyp or the medusa stage. A comparison showing the differences in the life cycles of *Hydra* and *Obelia* is shown in **Fig. 2**. *Hydra*, a solitary polyp, has no medusoid stage, while *Obelia*, a polymorphic colony of small polyps, has a medusoid state (**Fig. 3**).

The colony. In a hydroid colony there is sometimes only one kind of polyp and the medusae are developed on the hydranth or stem. Often, however, there are special reproductive polyps, with partial or complete reduction of mouth and tentacles. Such reproductive polyps are called gonozooids or gonangia and may produce small medusae which are liberated to grow and reproduce sexually. Often the medusa is incompletely developed except for its gonads, and the gametes ripen while the meduosoid body remains attached to the parent hydroid.

In addition to ordinary nutritive hydranths and reproductive ones, some hydroids have other specialized types which are sensory, defensive, or aid in capturing food.

The stems of a colony are covered by a stiff secreted substance, the perisarc or periderm, which supports and protects it. The periderm may form cups around the hydranths and gonozooids.

Ecology. Hydroids are species in which the polyp stage is usually dominant. Most hydroids are found near the shore attached to various supports such as rocks, wharves, boats, mussels, barnacles, worm tubes, crab and snail shells, and seaweeds. Few occur on mud or sand. Sometimes the medusa stage is well developed and the hydroid stage may be lacking. Medusae are abundant both in coastal waters and in the open sea.

Some hydroids live in direct association with other animals. A few are parasitic on fish; others live inside clams and on snails, hermit crabs, and worm tubes. The larger animal is valuable to the hydroid because its activity produces a movement of water past the hydroid and thus suspended food is made available.

Use in research. Hydras and hydroids have been used extensively in research on problems of growth, development, and regeneration. They have a high capacity for reorganization. Missing parts are quickly replaced, a bit of stem can produce a new hydranth, and completely disorganized masses of cells can reconstitute a new polyp. *See* REGENERATIVE BIOLOGY.

Dried *Sertularia*, a large and delicately branched hydroid, is sold as decorative material known as white weed (air fern when dyed green). *See* HYDROZOA.

Sears Crowell

Bibliography. R. D. Barnes, *Invertebrate Zoology*, 6th ed., 1994; J. G. Engemann, *Invertebrate Zoology*, 3d ed., 1981; C. M. Fraser, *Hydroids of the Atlantic Coast of North America*, 1944; C. M. Fraser, *Hydroids of the Pacific Coast of Canada and the*

United States, 1937; L. H. Hyman, *The Invertebrates*, vol. 1, 1940; F. S. Russell, *The Medusae of the British Isles*, 1953.

Hydrology

The science dealing with all aspects of the waters of the Earth: their occurrence, circulation, and distribution; their chemical and physical properties; and their reaction with the environment, including their relation to living things.

Hydrologic cycle. Water in liquid and solid form covers most of the crust of the Earth. By a complex process powered by gravity and the action of solar energy, an endless exchange of water, in vapor, liquid, and solid forms, takes place between the atmosphere, the oceans, and the crust. Water circulates in the air and in the oceans, as well as over and below the surface of landmasses (**Figs. 1** and **2**). The distribution of water in the planet is uneven. General patterns of circulation are present in the atmosphere, the oceans, and the landmasses, but regional features are very irregular and seemingly random in detail. Therefore, while causal relations underlie the overall process, it is believed that important elements of chance affect local hydrological events. *See* ATMOSPHERIC GENERAL CIRCULATION.

Hydrologic studies. Water is essential for all living things. It also participates in the physical and geochemical evolution of most nonliving matter on Earth. Its adequate supply is a key factor for urban, agricultural, and industrial development. Water can also be the recipient of pollutants that degrade its quality for all uses; and it may be a destructive agent when it inundates valleys and causes death and great damage during floods. The rising of ground-water levels in agricultural lands may cause deterioration of the soils and loss of fertility by waterlogging and increased salinity. Erosion of soils by flowing waters, and the ultimate deposition of the sediment in lakes, reservoirs, stream channels, and harbors, are also serious problems. Thus, the means by which natural waters may be captured and controlled are of utmost importance for the development of human economy. The study of hydrology provides the information necessary for determining those means. *See* BIOSPHERE; EROSION; WATER POLLUTION.

Whereas the global linkages of the hydrologic cycle are recognized, the science of hydrology has traditionally confined its direct concern to the detailed study of the portion of the cycle limited by the physical boundaries of the land; thus, it has generally excluded specialized investigations of the ocean (which is the subject of the science of oceanography) and the atmosphere (which is the subject of the science of meteorology). The heightened interest in anthropogenically induced environmental impacts has, however, underlined the critical role of the hydrologic cycle in the global transport and budgeting of mass, heat, and energy. Hydrology has become recognized as a science concerned with processes at the local, regional, and global scales. This enhanced status has strengthened its links to meteorology, climatology, and oceanography. *See* CLIMATOLOGY; METEOROLOGY; OCEANOGRAPHY.

A number of field measurements are performed for hydrologic studies. Among them are the amount and intensity of precipitation; the quantities of water stored as snow and ice, and their changes in time; discharge of streams; rates and quantities of infiltration into the soil, and movement of soil moisture; rates of production from wells and changes in their water levels as indicators of ground-water storage;

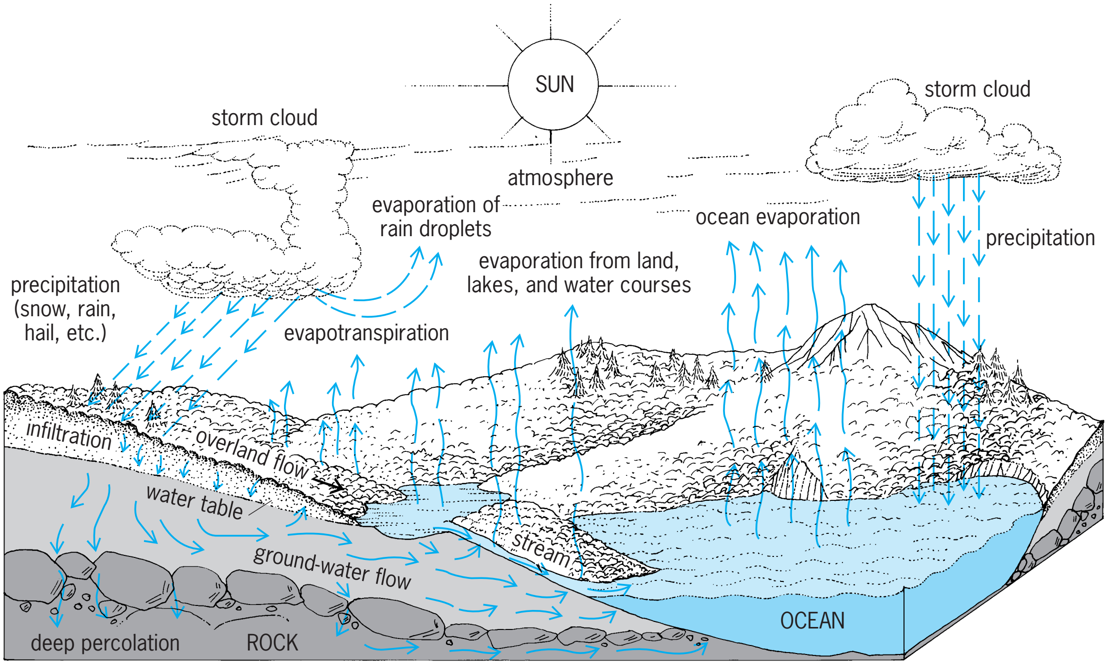
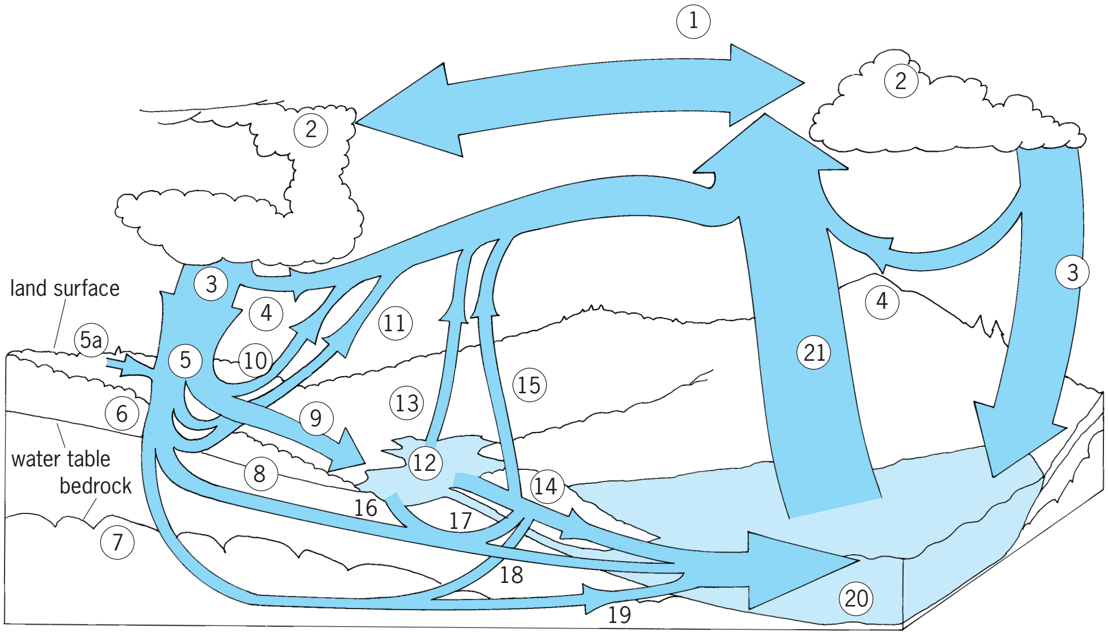


Fig. 1. Diagram of the hydrologic cycle.



Key:

1. Atmospheric water in circulation
2. Formation of strom clouds
3. Precipitation (snow, rain, hail, etc.)
4. Evaporation of precipitation in transit and moisture diffusion
5. Net precipitation on Earth surface
5a. Snow storage and melt
6. Infiltration
7. Deep percolation
8. Ground-water flow and storage
9. Overland flow and depression storage
10. Evporation of intercepted and surface water
11. Evapotranspiration
12. Surface storage in lakes and reservoirs
13. Evaporation from lake surfaces
14. Surface streams
15. Evaporation from streams
16. Ground-water exchange from channel storage (base flow and return)
17. Ground-water exchange from channel storage (base flow and return)
18. Ground-water flow to ocean
19. Ground-water flow to ocean
20. Storage in the ocean
21. Evaporation from ocean surfaces

Fig. 2. Water flow scheme of the hydrologic cycle.

concentration of chemical elements, compounds, and biological constituents in surface and ground waters; amounts of water transferred by evaporation and evapotranspiration to the atmosphere from snow, lakes, streams, soils, and vegetation; and sediment lost from the land and transported by streams.

In addition to the making of these measurements with specially designed instruments and devices, hydrology is concerned with research on the phenomena and mechanisms involved in all physical and biological components of the hydrologic cycle, with the purpose of understanding them sufficiently to permit quantitative predictions and forecasting. The field investigations and measurements not only provide the data whereby the behavior of each component may be evaluated in detail, permitting formulation in quantitative terms, but also give a record of the historical performance of the entire system. Thus, two principal vehicles for hydrological forecasting and prediction become available: a set of elemental processes, whose operations are expressible in mathematical terms, linked to form deterministic models that permit the prediction of hydrologic events for given conditions; and a group of records or time series of measured hydrologic variables, such as precipitation or runoff, which can be analyzed by statistical methods to formulate stochastic models that permit inferences to be made on the future likelihood of hydrologic events.

Surface water. Hydrology, when specifically applied to the solution of problems, such as the estimation of water supplies from streams, flood control and protection, dam and reservoir design and operation, urban drainage, and changes in flow regimes due to modifications introduced by humans, traditionally has been designated as surface-water hydrology. This does not imply that flow barriers exist between surface water and ground water; however, because ground-water flows usually occur at much slower rates, the processes can be decoupled to some degree when the primary objective is to evaluate surface flows. Both deterministic and stochastic models are employed.

A variety of deterministic models have been proposed; they differ from each other mainly in the degree of simplification and schematization used to describe the components of the hydrologic cycle and their linkages. The common underlying concept is that the overall process is composed of distinct elements of water storage and translation, each following specific rules but all obeying the hydrologic balance equation, expressed, for any finite period of time, as inflow = outflow + change in water storage. Typically, this water accounting is carried out at short intervals of time, both internally for each element and as a whole for the entire catchment, to yield the surface outflows to be expected as a result of changing precipitation rates. Data on relevant physical and vegetative characteristics of the catchment, as well as on factors affecting evaporation and evapotranspiration, are needed for the calculations, which are generally performed on computers, except in the simplest models for rough approximations. *See* STREAM GAGING; WATER SUPPLY ENGINEERING.

Predictions on the fate of chemical and biological species contained naturally in surface waters or introduced as pollutants due to urbanization and industrial and agricultural developments can be made by appropriate coupling of the equations of chemical and biochemical reactions with the equations representing the dynamics of water movement. Solutions to these coupled equations are typically obtained by using computers.

Stochastic process models are used for purposes such as drawing inferences on the probabilities of occurrence of future extreme events such as floods and droughts, or simulating possible future flow sequences similar in likelihood to those of the historical record. These inferences and synthetic time series can be used in the planning and design of water resource systems. Research on climatic change is an area of study that has a profound influence on the interpretation of historical hydrologic data. All conclusions drawn from recorded data need to be consistent with the possibility that local hydrologies may or may not remain as observed in the recent past. *See* CLIMATE HISTORY; STOCHASTIC PROCESS.

Ground water. Waters contained in porous formations below the ground are extremely valuable sources of supply. They depend for their long-term availability on replenishment (or recharge) from the surface. Water-bearing layers suitable for economical extraction by means of wells are called aquifers; they operate simultaneously as reservoirs and as flow media. The ability of an aquifer material to store water depends primarily on its porosity; and the ability to permit fluid motion, on its permeability or hydraulic conductivity. Because the properties of aquifers vary in space, the flow patterns vary from point to point. These patterns are further altered by the presence of wells, which act as localized sink points. For the rational management of ground-water basins that may contain several nonuniform and nonhomogeneous aquifers, special computer models, based on the equations of flow through porous media, have been developed. These specialized models can be deterministic in nature, where the aquifer properties are assumed to be completely known, or stochastic in nature, where aquifer properties are assumed to vary randomly over an area. Physical, chemical, and biological processes can be incorporated into either type of computer model to predict the quality of the water in the aquifer. Studies can also be undertaken by means of these models for the planning of conjunctive use operations of surface and ground waters, whereby controlled aquifer recharges and discharges can be combined with regulated surface flows and reservoirs for optimizing the management of water resources. *See* AQUIFER; FLUID-FLOW PRINCIPLES; GROUND-WATER HYDROLOGY; RESERVOIR; WATER TABLE; WELL.

Snow. In many parts of the world, snow is an important component of the total water supply. Snow fields constitute natural reservoirs that accumulate water during the winter and release it during the melting season. Snow hydrology is concerned with the study of the conditions of accumulation of snow, the properties of snowpacks as porous media, and the processes of heat and vapor exchange between the atmosphere and the snow as the snowpack evolves and melts under gradually varying conditions of solar radiation. These studies, together with field measurements and remote sensing of snow covered by satellites, permit the formulation and solution of mathematical (computer) models for the prediction of snowmelt as a function of time. Accordingly, it becomes possible to plan the operation of surface reservoir systems in combination with expected snowmelt yields for optimal water resource management. *See* REMOTE SENSING; SNOW; SNOW SURVEYING; SNOWFIELD AND NÉVÉ.

Global hydrology. The role of water in its liquid, gaseous (water vapor), and solid (ice) phases as a critical agent and regulator of biogeochemical and energy transport and balance in the planet Earth has been recognized since the beginning of the twentieth century. The strong coupling of land, atmosphere, and oceans in determining the rates of evapotranspiration, precipitation, ice melting, and biogeochemical cycles, and the importance of water as a greenhouse gas, have resulted in the emergence of hydrology as a scientific discipline with strong emphasis in basic research. Modern hydrologic curricula include training in meteorology, climatology, physical geography, and the basic natural sciences (biology, chemistry, physics), in addition to the traditional instruction in engineering, physical hydrology, mathematics, fluid mechanics, and geology. *See* BIOGEOCHEMISTRY; GREENHOUSE EFFECT; HYDROSPHERE.

Miguel A. Mariño

Bibliography. P. B. Bedient and W. C. Huber, *Hydrology and Floodplain Analysis*, 2d ed., Addison-Wesley, Reading, MA, 1992; S. L. Dingman, *Physical Hydrology*, Macmillan, New York, 1994; P. S. Eagleson (ed.), *Opportunities in the Hydrologic Sciences*, National Research Council, National Academy Press, Washington, DC, 1991; A. D. Feldman (ed.), *Engineering Hydrology: Symposium Proceedings*, Hydraulics Division, American Society of Civil Engineers, New York, 1987; R. A. Freeze and J. A. Cherry, *Groundwater*, Prentice Hall, Englewood Cliffs, NJ, 1979; R. K. Linsley, Jr., et al., *Hydrology for Engineers*, 3d ed., McGraw-Hill, New York, 1982; D. R. Maidment (ed.), *Handbook of Hydrology*, McGraw-Hill, New York, 1993; M. A. Mariño and J. N. Luthin, *Seepage and Groundwater*, Elsevier Scientific, Amsterdam, 1982; W. Stumm and J. J. Morgan, *Aquatic Chemistry*, John Wiley, New York, 1996.

Hydrolysis

A chemical reaction in which splitting of a molecule by water occurs.

Hydrolysis as applied to organic molecules can be considered a reversal of such reactions as esterification and amide formation. The hydrolysis of esters [reaction (1)] and of amides [reaction (2)]

$$\underset{\text{Ethyl acetate}}{CH_3COOC_2H_5} + H_2O \rightarrow \underset{\text{Acetic acid}}{CH_3COOH} + \underset{\text{Ethanol}}{C_2H_5OH} \quad (1)$$

$$\underset{\text{Acetamide}}{CH_3CONH_2} + H_2O \rightarrow \underset{\text{Acetic acid}}{CH_3COOH} + \underset{\text{Ammonia}}{NH_3} \quad (2)$$

is shown. Other classes of organic compounds that are subject to hydrolysis include acetals, acyl and alkyl halides, ketals, and peptides. While the overall hydrolysis reaction [as in reactions (1) and (2)] appears to involve the addition of the water molecule (H_2O), the reaction is in fact more complicated. There are several reaction steps, such as the formation of a complex with either a proton [H^+; acid-catalyzed hydrolysis] or a hydroxyl ion (OH^-; base-catalyzed hydrolysis), followed by elimination of these ions to give the overall equation. The hydrolysis reaction is frequently encountered in biological systems. The kinetics of these reactions are greatly enhanced by the action of enzymes (biological catalysts) such as the esterases and the peptidases. *See* ENZYME; HYDROGEN ION.

In inorganic chemistry, hydrolysis, also called aquation, represents a class of reactions involving metal coordination complexes in which one of the coordinated ligands is displaced by either H_2O or OH^-. Hydrolysis is a special case of the class of reactions termed ligand displacement reactions [reaction (3), where M is a metal, L and X are ligands,

$$L_nMX + Y \rightarrow L_nMY + X \quad (3)$$

and Y is the displacing ligand (H_2O in hydrolytic reactions)]. *See* COORDINATION COMPLEXES; LIGAND.

The phenomena attributed to hydrolysis are commonly explained in terms of Brönsted acid-base theory. An acid is a species capable of donating a proton; a base, one that can accept a proton. A conjugate acid-base pair comprises two species that differ by a single transferable proton. From this, it follows that a solution of a weak acid (HB) will contain some of the conjugate base to an extent governed by the acid dissociation constant, K_a, and its concentration, C_{HB}. As the K_a of a weak acid gets smaller, the extent of dissociation of its conjugate base increases. Not all acids or bases are neutral and uncharged. Positively charged ammonium ion (NH_4^+) is the conjugate acid to ammonia (NH_3), and negatively charged acetate ion is the conjugate base of acetic acid. A solution of sodium acetate will be basic; its pH will be greater than 7. Since the proton must be donated by water, the extent of this reaction is determined by the competition between OH^-, the conjugate base of water, and acetate ion; since acetate ion is a weaker base than OH^-, the extent is small. Similarly, ammonium salts will have pH values somewhat lower than 7. Hence, to say that acetates and ammonium salts undergo hydrolysis is no longer necessary in view of the Brönsted theory. *See* ACID AND BASE; HYDROLYTIC PROCESSES; ION; PH.

Henry Freiser

Bibliography. H. Freiser, *Concepts and Calculations in Analytical Chemistry*, 1992; I. M. Kolthoff et al., *Quantitative Chemical Analysis*, 4th ed., 1969.

Hydrolytic processes

Reactions of both organic and inorganic chemistry wherein water effects a double decomposition with another compound, hydrogen going to one component, hydroxyl to another, as in reactions (1)–(3).

$$XY + H_2O \rightarrow HY + XOH \quad (1)$$

$$KCN + H_2O \rightarrow HCN + KOH \quad (2)$$

$$C_5H_{11}Cl + H_2O \rightarrow HCl + C_5H_{11}OH \quad (3)$$

Although the word hydrolysis means decomposition by water, cases in which water brings about effective hydrolysis unaided are rare, and high temperatures and pressures are usually necessary. *See* HYDROLYSIS.

In the field of organic chemistry, the term "hydrolysis" has been extended to cover the numerous reactions in which alkali or acid is added to water. An example of an alkaline-condition hydrolytic process is the hydrolysis of esters, reaction (4), to produce

$$CH_3COOC_2H_5 + NaOH \rightarrow CH_3COONa + C_2H_5OH \quad (4)$$

alcohol. An example of an acidic-condition process is the hydrolysis of olefin to alcohol in the presence of phospheric acid, reaction (5). The addition of acids

$$C_2H_4 + H_2O \xrightarrow{H_2PO_4} C_2H_5OH \quad (5)$$

of alkalies hastens such reactions even if it does not initiate the reaction.

Hydrolysis reactions may be classified as follows: (1) hydrolysis with water alone; (2) hydrolysis with dilute or concentrated acid; (3) hydrolysis with dilute or concentrated alkali; (4) hydrolysis with fused alkali with little or no water at high temperature.

Alcohol processes. The direct synthesis of ethanol from ethylene in the presence of phosphoric acid has already been mentioned. On a commercial scale this reaction is conducted at high temperature (300°C or 572°F) and high pressure (1000 lb/in.^{2} or 6895 kilopascals) with the reactants in the vapor state and the phosphoric acid supported on pelleted diatomaceous earth. During operation a small stream of phosphoric acid is injected at the inlet of the reactor to replace acid that is carried out by the reactants. Conversion of ethylene per pass is low, about 4%. Unreacted ethylene is recycled, and the overall process yield is 95–98%.

An alternative process for ethanol production is a liquid-phase process which uses sulfuric acid as the catalyst. In this process, the ethylene is absorbed into concentrated sulfuric acid and forms monoethyl and diethyl sulfate. The esters are transferred to a hydrolyzer vessel along with water to form the alcohol. The stepwise reaction is shown in (6). The sulfuric-acid-catalyzed process is also used to produce isopropyl alcohol from normal propylene, secondary butanol from normal butylene, tertiary butanol from isobutylene, and secondary and tertiary amyl alcohol from pentenes. The process produces secondary and tertiary alcohols almost exclusively and very little normal alcohols. *See* ALCOHOL.

$$C_2H_4 + H_2SO_4 \rightarrow C_2H_5OSO_3 \xrightarrow{H_2O} C_2H_5OH + H_2SO_4 \quad (6)$$

Phenol processes. The commercial production of phenol may be accomplished by several hydrolysis processes. The processes discussed here represent earlier technology and, although valid, are no longer economically attractive for new plants. Some existing units are still operated. One of the older processes is an example of an alkali fusion process. Benzene is the starting material for the lengthy process. The reactions involved are summarized in (7)–(10).

$$C_6H_6 + H_2SO_4 \rightarrow C_5H_5SO_3H + H_2O \quad (7)$$

$$2C_6H_5SO_3H + Na_2SO_3 \rightarrow 2C_6H_5SO_3Na + H_2SO_3 \quad (8)$$

$$C_6H_5SO_3Na + 2NaOH \rightarrow C_6H_5ONa + Na_2SO_3 + H_2O \quad (9)$$

$$2C_6H_5ONa + H_2SO_3 \rightarrow 2C_6H_5OH + Na_2SO_3 \quad (10)$$

Reactions (7) and (8) represent the formation of benzenesulfonic acid from benzene and sulfuric acid and the subsequent neutralization with sodium sulfite to form sodium benzenesulfonate. The fusion of sodium hydroxide and sodium benzenesulfonate in reaction (9) produces sodium phenoxide, and in reaction (10), sodium phenoxide is neutralized to yield phenol.

A second process is based on the hydrolysis of chlorobenzene in caustic soda solution at 350°C (662°F) and 4000 lb/in.^{2} (27,580 kPa). The reaction is summarized in (11)–(13). In a commercial operation the reaction is carried out in a very long, high-pressure tubular reactor. The overall reaction is exothermic, and external heating is needed only during start-up.

$$C_6H_5Cl + 2NaOH \rightarrow C_6H_5ONa + NaCl + H_2O \quad (11)$$

$$C_6H_5ONa + C_6H_5Cl \rightarrow C_6H_5\text{—}O\text{—}C_6H_5 + NaCl \quad (12)$$

$$C_6H_5\text{—}O\text{—}C_6H_5 + H_2O \rightarrow 2C_6H_5OH \quad (13)$$

A third phenol process is a regenerative process that takes place in the vapor phase. The two steps in the process are shown in reaction (14) and (15).

$$C_6H_6 + HCl + O \rightarrow C_6H_5Cl + H_2O \quad (14)$$

$$C_6H_5Cl + H_2O \rightarrow 2C_6H_5OH \quad (15)$$

In the first step benzene, hydrogen chloride, and atmospheric oxygen are consumed, and chlorobenzene and water are products. The chlorobenzene and water are consumed in the second-step reaction to yield the desired phenol product and the first-step reactant, hydrogen chloride. The reaction is carried out in two catalyzed vapor-phase reactions. Although the overall reaction is simple and straightforward, in practice the per pass conversion in the first step is only about 12% while the per pass conversion in the second step is only 15%. The separation and recycle equipment required to conserve the reactants complicates the commercial unit considerably. *See* HALOGENATED HYDROCARBON.

Soap manufacture. Perhaps some of the oldest and largest-volume hydrolysis technology is involved in soap manufacture. A two-step process for saponification is shown in reactions (16) and (17). In the first step, glyceryl stearate acid, a fat, is hydrolyzed with water to yield stearic acid and glycerin. In the second step, the stearic acid is neutralized with caustic soda to give sodium stearate, the soap, and water. Originally performed as a batch operation, the process is now automated to a continuous process. *See* FAT AND OIL.

$$(C_{17}H_{35}COO)_3C_3H_5 + 3H_2O \rightarrow 3C_{17}H_{35}COOH + C_3H_5(OH)_3 \quad (16)$$

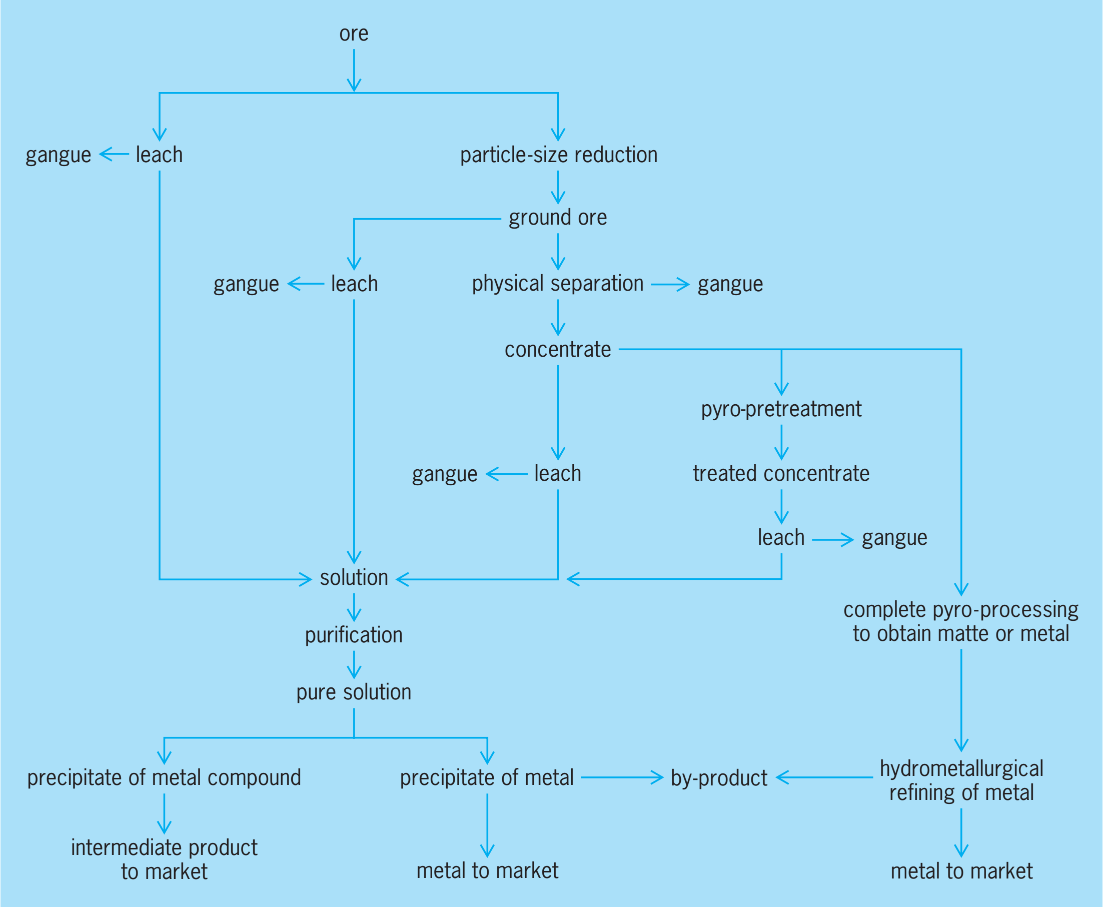
$$C_{17}H_{35}COOH + NaOH \rightarrow C_{17}H_{35}COONa + H_2O \quad (17)$$

Products. Most hydrolytic processes take place in acid or alkaline conditions which in some cases are quite severe. The acids used are primarily sulfuric, hydrochloric, and phosphoric. Much of the equipment in hydrolytic processes must be of special construction such as glass-lined steel or special metal alloy.

Hydrolytic processes account for a huge product volume. Conversion of starch such as cornstarch into maltose and glucose (sugar syrups) by treatment with hydrochloric acid is a major industry. Similarly, the production of furfural from pentosans of oat hulls or other cereal by-products such as corn cobs, rice hulls, or cottonseed bran is another commercial hydrolytic process. The production of alcohols, soaps, and industrial intermediate chemicals have already been mentioned. The volume and diversity of the products serve to illustrate the fundamental nature and importance of hydrolytic processes. D. L. Holt

Hydrometallurgy

The extraction and recovery of metals from their ores by processes in which aqueous solutions play a predominant role. Two distinct processes are involved in hydrometallurgy: putting the metal values in the ore into solution via the operation known as leaching; and recovering the metal values from solution, usually after a suitable solution purification or concentration step, or both. The scope of hydrometallurgy is quite broad and extends beyond the processing of ores to the treatment of metal concentrates, metal



Generalized metallurgical flow sheet.

scrap and revert materials, and intermediate products in metallurgical processes. Hydrometallurgy enters into the production of practically all nonferrous metals and of metalloids, such as selenium and tellurium.

A generalized metallurgical flow sheet (see **illus.**) provides an indication of the nature and extent of the role of hydrometallurgy in metal production. It also shows the manner in which hydrometallurgical and pyrometallurgical processes complement each other. *See* PYROMETALLURGY.

The advantages of hydrometallurgy include applicability to low-grade ores (copper, uranium, gold, silver) and complex sulfide ores, amenability to the treatment of materials of quite different compositions and concentrations, adaptability to separation of highly similar materials (hafnium from zirconium), and flexibility in terms of the scale of operations. Hydrometallurgical operations are amenable to effective control leading to automation and continuous operation.

There are some important disadvantages in hydrometallurgical processes. The processes are generally energy-intensive and can involve the handling of large volumes of dilute solutions that may be corrosive or hazardous. In addition, the processes yield residues and effluents that must be disposed of in an environmentally acceptable manner.

The first commercial hydrometallurgical operation was the application of cyanidation to the treatment of gold ores in 1889, more than 40 years after the discovery that gold can be dissolved in dilute alkaline solutions of sodium cyanide, as shown in reaction (1).

$$4Au + 8NaCN + O_2 + 2H_2O \longrightarrow 4NaAu(CN)_2 + 4NaOH \quad (1)$$

Leaching. A major unit process in hydrometallurgy is the leaching of ores and concentrates. A particularly noteworthy and time-honored dissolution process is the caustic pressure leaching of bauxite ore to produce pure alumina. This process, which was developed in the nineteenth century, is still in use as the principal method of producing alumina. The ever-increasing demand for aluminum and the rising cost of bauxite stimulated research on the development of processes for leaching nonbauxite alumina materials such as a alunite, anorthosite, clays, and coal shales. *See* ALUMINUM; LEACHING.

In the case of sulfide concentrates, leaching in acid or ammoniacal solution under oxygen pressure provides a number of advantages, including accelerated leaching and the recovery of sulfur in a suitable form, such as ammonium sulfate in the ammoniacal leaching of nickel concentrates in the Sherritt-Gordon process or as elemental sulfur in the acid pressure leaching of zinc concentrates.

The economic recovery of metals from low-grade ores became increasingly important as mining costs rose and as higher-grade materials became more difficult to find. In many operations, what was at one time waste is now considered as ore to be treated by never leaching methods such as heap, dump, and in situ leaching. In the United States, copper in excess of 160,000 tons (145,000 metric tons) is extracted annually by such processes. The same technology is applied in the leaching of uranium ores and in the treatment of gold and silver ores.

Biotechnology has come to play a more significant role in mineral leaching and metal separation processes. Of particular importance is the microbiological leaching of low-grade sulfide copper ores and of uranium ores in the presence of pyrite utilizing the action of the bacterium *Thiobacillus ferrooxidans*. The same organism provides an attractive option in the treatment of refractory gold ores. The bioleaching of minerals such as pyrite and arsenopyrite, in which gold is entrapped, facilitates the subsequent recovery of the gold by cyanide leaching. *See* BIOLEACHING.

The establishment of more stringent regulations in many countries concerning sulfur dioxide emissions from smelters has led not only to the adoption of new and improved smelting processes but also to a great deal of research on hydrometallurgical alternatives to smelting, especially of copper concentrates. In the case of copper concentrates, the principal difficulty has been the realization of a hydrometallurgical process that has sufficiently well-defined economic advantages. From an environmental standpoint, there are definite advantages which are afforded by hydrometallurgy, principally the elimination of sulfur dioxide emissions and the conversion of sulfur to elemental sulfur or a sulfate. The overall economics of copper hydrometallurgy can be improved by the recovery of copper in final form by a modified electrowinning process or by an alternative process such as the hydrogen reduction of cuprous chloride.

Environmental and other considerations led to the cessation of the production of zinc by the carbothermic reduction of zinc calcine at 2200–2400°F (1200–1300°C) in horizontal retorts by reaction (2), where

$$\mathrm{ZnO}(s) + \mathrm{C}(s) \rightarrow \mathrm{Zn}(g) + \mathrm{CO}(g) \qquad (2)$$

(s) and (g) denote solid and gas. This process has been replaced by the sulfuric acid leaching of zinc calcine, purification of the resulting leach solution, and the recovery of zinc by electrowinning. An important alternative to the production of zinc calcine by fluid-bed roasting is the acid pressure leaching of zinc concentrates, in which the sulfide is oxidized to yield elemental sulfur.

Recovery. The recovery of metals from leach solutions is often preceded by the purification or concentration of the metal to be isolated. Two processes that can be used to achieve this objective are ion exchange and solvent extraction.

Ion exchange. Ion exchange involves the selective adsorption of metal ions onto a synthetic organic ion-exchange resin (commonly a styrene-divinylbenzene copolymer containing suitable functional groups for ion exchange). The process is stoichiometric and reversible, so that the adsorbed ions are subsequently desorbed or eluted from the resin by a solution known as the eluant. The resulting solution or eluate is treated for metal recovery or for production of a metal compound, such as ammonium diuranate, $(NH_4)_2U_2O_7$, in uranium extraction. Ion exchange, which is widely used in water purification, is employed extensively in uranium recovery. Here continuous ion-exchange processes are in commercial use, with resultant increases in efficiency and productivity. Chelating resins should find increasing use in the separation of metals such as copper, nickel, cobalt, and zinc. *See* CHELATION; ION EXCHANGE.

Solvent extraction. As applied in hydrometallurgy, solvent extraction relates to the selective transfer of metal species from an aqueous solution to an immiscible organic solvent with which it is in contact. The solvent itself can be the active reagent in the process, as in the extraction of uranium by tributylphosphate. On the other hand, the solvent can contain an organic extractant which forms a complex or chelation compound with the metal ion from the aqueous phase. The extraction process is reversible, so that the extracted metal can be stripped from the organic solution and returned to the aqueous phase.

With the development of suitable chelating compounds, solvent extraction became established as an excellent technique for the upgrading of copper leach solutions. The application of solvent extraction is especially suited to the treatment of the dilute solutions resulting from the heap and dump leaching of low-grade copper ores. The selectivity of the organic reagents, notably the hydroxyoximes, is such as to permit an excellent separation from iron. The stripping of the organic phase by spent electrowinning electrolyte leads to a strong solution, from which copper can be electrowon without difficulty.

Solvent extraction is used commercially in the separation of many metals, including copper, nickel, cobalt, zinc, uranium, thorium, zirconium, hafnium, molybdenum, tungsten, niobium, tantalum, and beryllium. *See* SOLVENT EXTRACTION.

Recovery from aqueous solution. Metal recovery from aqueous solution can be effected by a number of reduction processes, the most prominent being electrowinning and gaseous reduction using hydrogen. The latter process yields metal powder, whereas the former results in cathodes which are melted and cast to produce various marketable shapes.

Importance. Hydrometallurgy occupies an important role in the production of aluminum, copper, nickel, cobalt, zinc, gold, silver, platinum metals, selenium, tellurium, tungsten, molybdenum, uranium, zirconium, and other metals. Considering the versatility of hydrometallurgy and the need to process more complex ores, as well as lower-grade ores and secondary materials, and to produce high-purity advanced materials, hydrometallurgical processes are expected to play an even greater part in the production of metals and materials in the future. *See* ELECTROMETALLURGY; METALLURGY.

W. Charles Cooper

Bibliography. J. H. Canterford, Hydrometallurgy: Winning metals with water, *Chem. Eng.*, pp. 41–48, October 28, 1985; W. C. Cooper, G. E. Lagos, and G. Ugarte (eds.), Hydrometallurgy and electrometallurgy of copper, *Copper 87*, vol. 3, University of Chile, 1988; G. A. Davies (ed.), *Separation Processes in Hydrometallurgy*, Society of the Chemical Industry, U.K., 1987; H. L. Ehrlich and C. L. Brierley, *Microbial Mineral Recovery*, 1990; J. B. Hiskey and C. W. Warren (eds.), *Hydrometallurgy: Fundamentals, Technology, and Innovations*, 1993; K. Osseo-Asare and J. D. Miller (eds.), *Hydrometallurgy*: *Research*, *Development and Plant Practice*, AIME, 1983; G. Rossi, *Biohydrometallurgy*, 1990.

Hydrometeorology

The study of the occurrence, movement, and changes in the state of water in the atmosphere. The term is also used in a more restricted sense, especially by hydrologists, to mean the study of the exchange of water between the atmosphere and continental surfaces. This includes the processes of precipitation and direct condensation, and of evaporation and transpiration from natural surfaces. Considerable emphasis is placed on the statistics of precipitation as a function of area and time for given locations or geographic regions.

Water occurs in the atmosphere primarily in vapor or gaseous form. The average amount of vapor present tends to decrease with increasing elevation and latitude and also varies strongly with season and type of surface. Precipitable water, the mass of vapor per unit area contained in a column of air extending from the surface of the Earth to the outer extremity of the atmosphere, varies from almost zero in continental arctic air to about 1.4 oz/in.^{2} (6 g/cm^2) in very humid, tropical air. Its average value over the Northern Hemisphere varies from around 0.46 oz/in.^{2} (2.0 g/cm^2) in January and February to around 0.84 oz/in.^{2} (3.7 g/cm^2) in July. Its average value is around 0.63 oz/in.^{2} (2.8 g/cm^2), an amount equivalent to a column of liquid water slightly greater than 1 in. (2.5 cm) in depth. Close to 50% of this water vapor is contained in the atmosphere's first mile, and about 80% is to be found in the lowest 2 mi (3 km).

Atmospheric water cycle. Although a trivial proportion of the water of the globe is found in the atmosphere at any one instant, the rate of exchange of water between the atmosphere and the continents and oceans is high. The average water molecule remains in the atmosphere only about 10 days, but because of the extreme mobility of the atmosphere it is usually precipitated many hundreds or even thousands of miles from the place at which it entered the atmosphere.

Evaporation from the ocean surface and evaporation and transpiration from the land are the sources of water vapor for the atmosphere. Water vapor is removed from the atmosphere by condensation and subsequent precipitation in the form of rain, snow, sleet, and so on. The amount of water vapor removed by direct condensation at the Earth's surface (dew) is relatively small.

A major feature of the atmospheric water cycle is the meridional net flux of water vapor. The average precipitation exceeds evaporation in a narrow band extending approximately from 10°S to 15°N lat. To balance this, the atmosphere carries water vapor equatorward in the tropics, primarily in the quasi-steady trade winds which have a component of motion equatorward in the moist layers near the Earth's surface. Precipitation also exceeds evaporation in the temperate and polar regions of the two hemispheres, poleward of about 40° latitude. In the middle and higher latitudes, therefore, the atmosphere carries vapor poleward. Here the exchange occurs through the action of cyclones and anticyclones, large-scale eddies of air with axes of spin normal to the Earth's surface.

For the globe as a whole the average amount of evaporation must balance the precipitation (**Table 1**). This exchange is related to the characteristics of the general circulation of the atmosphere. It seems likely that a similar cycle would be observed even if the Earth were entirely covered by ocean, although details of the cycle, such as the flux across the Equator, would undoubtedly be different.

Complications in the global pattern arise from the existence of land surfaces. Over the continents the only source of water is from precipitation; therefore, the average evapotranspiration (the sum of evaporation and transpiration) cannot exceed precipitation. The flux of vapor from the oceans to the continents through the atmosphere, and its ultimate return to atmosphere or ocean by evaporation, transpiration, or runoff is known as the hydrologic cycle. Its atmospheric phase is closely related to the air mass cycle. In middle latitudes of the Northern Hemisphere, for example, precipitation occurs primarily from maritime air masses moving northward and eastward across the continents. Statistically, precipitation from

TABLE 1. Meridional flux of water vapor in the atmosphere

Latitude	Northward flux, 10^{10} g/s
90°N	0
70°N	4
40°N	70
10°N	−61
Equator	45
10°S	71
40°S	−75
70°S	1
90°S	0

these air masses substantially exceeds evapotranspiration into them. Conversely, cold and dry air masses tend to move southward and eastward from the interior of the continents out over the oceans. Evapotranspiration into these continental air masses strongly exceeds precipitation, especially during winter months. These facts, together with the extreme mobility of the atmosphere and its associated water vapor, make it likely that only a small percentage of the water evaporated or transpired from a continental surface is reprecipitated over the same continent. *See* ATMOSPHERIC GENERAL CIRCULATION; HYDROLOGY.

Precipitation. Hydrometeorology is particularly concerned with the measurement and analysis of precipitation data. Radar plays an important role in estimating precipitation. By relating the intensity of radar echo to rate of precipitation, it has been possible to obtain a vast amount of detailed information concerning the structure and areal distribution of storms. *See* METEOROLOGICAL INSTRUMENTATION; RADAR METEOROLOGY; STORM DETECTION.

Deficiencies in the observational networks over the oceans and over the more sparsely inhabited land areas of the Earth are now being bridged through the use of meteorological satellite observations. Progress toward the development of methods for estimating rainfall amounts from satellite observations of cloud type and distribution is of particular significance to hydrometeorology. *See* METEOROLOGICAL SATELLITES.

Precipitation occurs when the air is cooled to saturation. The ascent of air toward lower pressure is the most effective process for causing rapid cooling and condensation. Precipitation may therefore be classified according to the atmospheric process which leads to the required upward motion. Accordingly, there are three basic types of precipitation: (1) Orographic precipitation occurs when a topographic barrier forces air to ascend. The presence of significant relief often leads to large variations in precipitation over relatively short distances. (2) Extratropical cyclonic precipitation is associated with the traveling regions of low pressure of the middle and high latitudes. These storms, which transport sinking cold dry air southward and rising warm moist air northward, account for a major portion of the precipitation of the middle and high latitudes. (3) Air mass precipitation results from disturbances occurring in an essentially homogeneous air mass. This is a common precipitation type over the continents in mid-latitudes during summer. It is the major mechanism for precipitation in the tropics, where disturbances may range from areas of scattered showers to intense hurricanes. In most cases there is evidence for organized lifting of air associated with areas of cyclonic vorticity, that is, areas over which the circulation is counterclockwise in the Northern Hemisphere or clockwise in the Southern Hemisphere.

The availability of data from geosynchronous meteorological satellites, together with surface and upper-air data acquired as part of the Global Atmospheric Research Program (GARP), is leading to significant advances in the understanding of the character and distribution of tropical precipitation. *See* HURRICANE; STORM; TROPICAL METEOROLOGY.

Precipitation may, of course, be in liquid or solid form. In addition to rain and snow there are other forms which often occur, such as hail, snow pellets, sleet, and drizzle. If upward motion occurs uniformly over a wide area measured in tens or hundreds of miles, the associated precipitation is usually of light or moderate intensity and may continue for a considerable period of time. Vertical velocities accompanying such stable precipitation are usually of the order of several centimeters per second. Under other types of meteorological conditions, particularly when the density of the ascending air is less than that of the environment, upward velocities may locally be very large (of the order of several meters per second) and may be accompanied by compensating downdrafts. Such convective precipitation is best illustrated by the thunderstorm. Intensity of precipitation may be extremely high, but areal extent and local duration are comparatively limited. Storms are sometimes observed in which local convective regions are embedded in a matrix of stable precipitation. *See* PRECIPITATION (METEOROLOGY).

Analysis of precipitation data. Precipitation is essentially a process which occurs over an area. However, despite the use of radar, most observations are taken at individual stations. Analyses of such "point" precipitation data are most often concerned with the frequency of intense storms. These data (such as in **Table 2**) are of particular importance in evaluating

TABLE 2. Record observed point rainfalls*

Duration	Depth, in. (cm)	Station	Date
1 min	1.23 (3.12)	Unionville, Maryland	July 4, 1956
8 min	4.96 (12.6)	Fuüssen, Bavaria	May 25, 1920
15 min	7.80 (19.8)	Plumb Point, Jamaica	May 12, 1916
42 min	12.00 (30.5)	Holt, Missouri	June 22, 1947
2 h 45 min	22.00 (55.9)	Near D'Hanis, Texas	May 31, 1935
24 h	73.62 (187)	Cilaos, La Reunion (Indian Ocean)	March 15–16, 1952
1 month	366.14 (930)	Cherrapunji, India	July 1861
12 months	1041.78 (2646.1)	Cherrapunji, India	August 1860 to July 1861

*From D. R. Maidment (ed.), *Handbook of Meteorology*, 1993.

local flood hazard, and may be used in such diverse fields as the design of local hydraulic structures, such as culverts or storm sewers, or the analysis of soil erosion. Intense local precipitation of short duration (up to 1 h) is usually associated with thunderstorms. Precipitation may be extremely heavy for a short period, but tends to decrease in intensity as longer intervals are considered.

A typical hydrometeorological problem might involve estimating the likelihood of occurrence of a storm of given intensity and duration over a specified watershed to determine the required spillway capacity of a dam. Such estimates can only be obtained from a careful meteorological and statistical examination of large numbers of storms selected from climatological records.

Evaporation and transpiration. In evaluating the water balance of the atmosphere, the hydrometeorologist must also examine the processes of evaporation and transpiration from various types of natural surfaces, such as open water, snow and ice fields, and land surfaces with and without vegetation. From the point of view of the meteorologist, the problem is one of transfer in the turbulent boundary layer. It is complicated by topographic effects when the natural surface is not homogeneous. In addition the simultaneous heating or cooling of the atmosphere from below has the effect of enhancing or inhibiting the transfer process. Although the problem has been attacked from the theoretical side, empirical relationships are at present of greatest practical utility. *See* METEOROLOGY; MICROMETEOROLOGY.

Eugene M. Rasmusson

Bibliography. R. K. Linsley et al., *Hydrology for Engineers*, 3d ed., 1982; F. K. Lutgens and E. J. Tarbuck, *Atmosphere: An Introduction to Meteorology*, 8th ed., 2000; R. R. Maidment (ed.), *Handbook of Meteorology*, 1993; W. D. Sellers, *Physical Climatology*, 1965.

Hydrometer

A direct-reading instrument for indicating the density, specific gravity, or some similar characteristic of liquids. Almost all hydrometers are made of a high-grade glass tubing. The main body is the float section in the bottom of which ballast, such as small shot, is secured. A small-diameter tube, the stem, extends from the upper end of the float section. Inside the stem is the scale, printed on heavy-grade paper, and well-secured within the stem so its position will not change. When the hydrometer is placed in a liquid, the stem extends vertically above the surface for a portion of its length.

The regular, or plain, hydrometer is illustrated in **Fig. 1***a*, and the thermohydrometer, which has a thermometer enclosed in the float section, in Fig. 1*b*.

Hydrometers may be classified according to the indication provided by graduations of the scale as follows: (1) density hydrometers, to indicate densities at a particular temperature, and usually for a particular liquid; (2) specific gravity hydrometers to indicate specific gravity of a liquid, with reference to water, at a particular temperature; (3) percentage hydrometers to indicate, at a particular temperature, the percentage of a substance such as salt, sugar, or alcohol dissolved in water; (4) arbitrary scale hydrometers, indicating the density, specific gravity, or concentration of a liquid in terms of an arbitrarily defined scale, at a defined temperature.

Alcoholometers, which belong in group 3, indicate the percentage of ethyl alcohol, either by volume or by weight, in a water-alcohol mixture, or the scale may be percent "proof spirit." (Percent proof spirit is twice the percentage of ethyl alcohol by volume at 60°F or 15.6°C.) The Brix and Balling (Plato) saccharimeters indicate the percentage by weight of pure sucrose solutions.

Possibly the most widely used arbitrary scale hydrometers are the API (American Petroleum Institute) and the Baumé (Bé). The relation between the

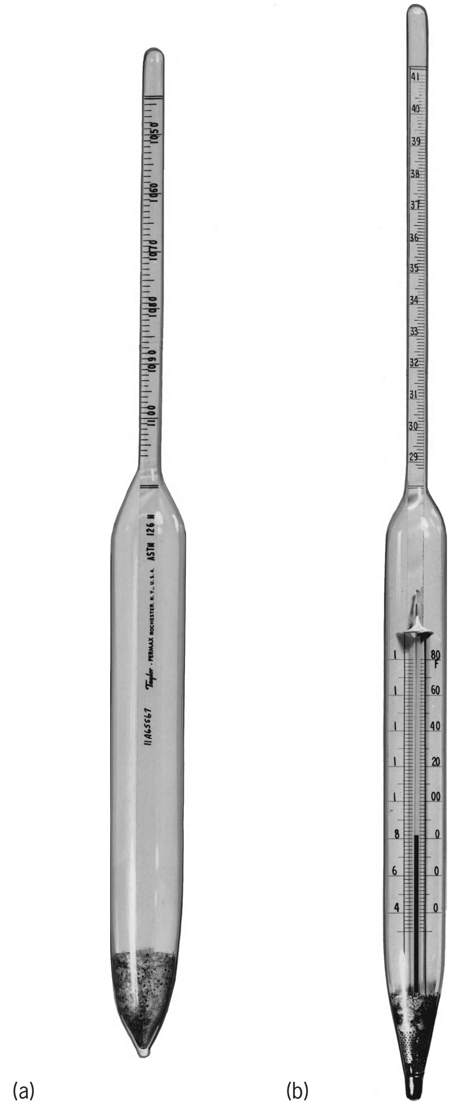


Fig. 1. Types of hydrometer. (*a*) Plain hydrometer. (*b*) Thermohydrometer. (*Taylor Instrument Co.*)

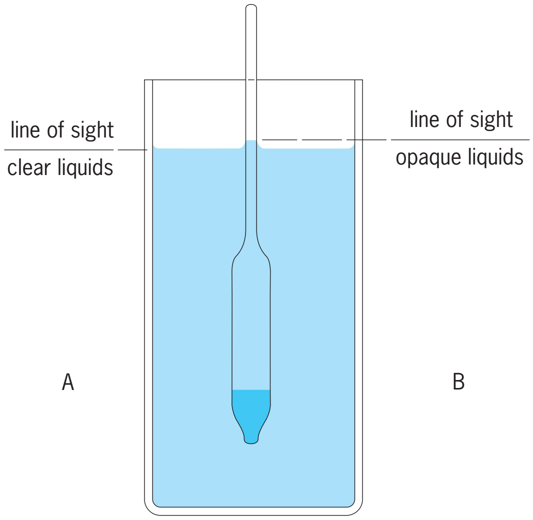


Fig. 2. Illustration of method of reading hydrometers.

API scale and specific gravity is given by Eq. (1). The

$$\text{Degrees API} = \frac{141.5}{\text{sp gr } 60/60 \text{ F}} - 131.5 \qquad (1)$$

denominator, sp gr 60/60 F, means the specific gravity of the liquid (an oil) when the temperatures of both the liquid and the water to which it is referred are 60°F (15.6°C).

The relation between the Baumé scale and specific gravity for liquids lighter than water is given by Eq. (2); for liquids heavier than water, by Eq. (3).

$$\text{Degrees Bé} = \frac{140}{\text{sp gr } 60/60 \text{ F}} - 130 \qquad (2)$$

$$\text{Degrees Bé} = 145 - \frac{145}{\text{sp gr } 60/60 \text{ F}} \qquad (3)$$

Other special scales are the Quevenne, used in lactometers for testing milk, and the barkometer and Twaddle scales used by the leather-tanning industry in testing the strength of tanning extracts.

The best way to read a hydrometer in clear liquids is to start with the eyes slightly below the plane of the liquid surface and slowly to raise the eyes until the surface of the liquid appears as a straight line (**Fig. 2**, side A). The place where the line crosses the scale is the reading. With opaque liquids, such as milk and many oils, it is necessary to read the hydrometer at the top of the meniscus (Fig. 2, side B).

For accurate readings the hydrometer, especially the stem, must be absolutely clean. Also, the surface of the liquid must be clean and free of dust. With a precision grade hydrometer, with a long small-diameter stem, specific gravity values may be read to 0.0001. In the general use of hydrometers, however, the uncertainty of readings will probably be about ±0.001 specific gravity or the equivalent. *See* SPECIFIC GRAVITY.

Howard S. Bean

Bibliography. D. M. Considine, *Process/Industrial Instruments and Controls Handbook*, 4th ed., 1993; R. W. Herschy, *Hydrometry; Principles and Practices*, 1978; J. C. Hughes, *Testing of Hydrometers*, U.S. Department of Commerce, Nat. Bur. Stand. Circ. 555, October 1954.

Hydrophone

The underwater equivalent of a microphone, which generates an electrical signal as a response to the pressure component of an acoustic signal. Modern hydrophones are usually composed of piezoelectric ceramics such as barium titanate and lead zirconate titanate. The piezoelectric effect is the ability of materials to generate a voltage in response to applied mechanical stress. This effect is reciprocal in that an applied voltage will induce a small deformation in a piezoelectric material. Thus, some hydrophones can also be used as acoustic sources or projectors. In general, the capacitance of a hydrophone is relatively small, so a preamplifier is required to be placed near the piezoelectric material. The preamplifier boosts the electrical signal to allow the use of long cables without reducing the sensitivity of the hydrophone due to the added capacitance. When a preamplifier is installed in a hydrophone, the hydrophone is no longer reciprocal and may not be used as a projector. Hydrophone design requirements differ from microphones in that they must be electrically insulated from water and may need to operate at high hydrostatic pressure (deep depths). *See* ELECTRET TRANSDUCER; MICROPHONE; PIEZOELECTRICITY; PREAMPLIFIER.

Acoustic signals involve very small displacements of particles (water molecules, for instance) that propagate with relatively high speeds (1500 m/s or 4900 ft/s for water). These displacements result in local condensations (higher mean density) and rarefactions (lower mean density). Associated with the density fluctuations are changes in pressure, a scalar quantity containing no directional information. In order to obtain directional information, it is necessary to use arrays of hydrophones distributed along a line, over a surface, or throughout a volume. The acoustic signal arriving from a distant source will reach each hydrophone at a different time. This time difference, along with the known location of each hydrophone, provides information on the arrival direction. Modern underwater sensor systems also use sensors to detect the vector component (particle displacement, velocity, or acceleration) of acoustic signals to provide directional information in a more compact array than is achievable in a hydrophone array. *See* SOUND; SOUND PRESSURE.

Hydrophones are currently used to detect and determine the location of submarines, with hydrophone arrays mounted on or towed behind surface ships or submarines, mounted on some underwater weapons, or suspended from air-deployed sonobuoys. Submarines are also detected with stationary bottom-mounted hydrophones, such as the Sound Surveillance System (SOSUS) arrays located in the Atlantic and North Pacific oceans. Submarine detection systems may be active or passive. In a passive system the hydrophone detects acoustic signals made by the submarine. In an active system the hydrophone detects acoustic signals generated by the system sources which have been reflected by the submarine. Hydrophones are also used to detect and track marine mammals, to study global climate change by detecting and mapping temperature change in the world's oceans by means of the change in sound speed caused by the temperature change, to find schools of fish, to assess ocean biomass, and to prospect for offshore seismic oil. *See* ANTISUBMARINE WARFARE; OIL AND GAS, OFFSHORE; SONOBUOY; UNDERWATER SOUND. Peter H. Rogers; David H. Trivett

Bibliography. R. J. Bobber, *Underwater Electroacoustic Measurements*, Government Printing Office, Washington, DC, 1970, reprint, Peninsula Publishing 1990; J. W. Horton, *Fundamentals of Sonar*, United States Naval Institute Press, Annapolis, 1957; L. E. Kinsler et al., *Fundamentals of Acoustics*, 4th ed., Wiley, New York, 2000.

Hydroponics

Techniques for supplying nutrients and water directly to the roots of plants, without soil or other media. Methods that utilize an inert medium such as sand, gravel, peat, or vermiculite to provide the root environment, with water and nutrients added in solution, are soilless culture but are not hydroponic in the strict sense.

Hydroponic systems range in complexity from a single plant supported above an aerated jar of nutrient solution to thousands of plants supported above a large area of flowing solution in which pH, temperature, and nutrient concentrations are controlled by using a sophisticated computer system and automated chemical analysis. In hydroponic culture, precise control of the pH and the concentrations of elements in the solution is critical; all essential elements must be provided and in the correct ratios for plant growth. Elements known to be essential are nitrogen, phosphorus, potassium, calcium, magnesium, iron, manganese, boron, copper, molybdenum, and zinc. Oxygen in the root zone is also necessary. *See* PLANT MINERAL NUTRITION.

Hydroponic systems. Hydroponic systems offer a number of advantages when compared to soil culture. They reduce water, pH, and nutrient stress; yield clean roots and leaves; and facilitate rapid crop turnaround and automation. The disadvantages are that disease may spread more rapidly, pH and nutrient control are required, and initial expenses are higher. In theory, the growth-limiting factors in hydroponic systems are the availability of photosynthetic light and carbon dioxide. *See* PHOTOSYNTHESIS.

Among typical hydroponic systems are aerated standing culture, intermittent-flow culture, and continuous-flow culture. These techniques require careful preparation of the nutrient solution, continued monitoring, and adjustment or periodic replacement of the solution.

The aerated standing-culture technique (**Fig. 1***a*) is widely used in research and was the first of the hydroponic methods. In this system, plants are supported above an opaque container, with the roots submerged in a nutrient solution that is aerated by bubbling. Light must be excluded from the solution to avoid the growth of algae. Either the pH and nutrients must be adjusted frequently or the solution must be replaced daily to weekly, depending upon the volume per plant. Intermittent- and continuous-flow cultures use similar techniques for delivering the solution and differ only in the length of time that the roots are immersed. Such systems can be open (single-pass) or closed (recirculating), depending on the requirements for reuse of the solution.

The most widely used active solution-delivery techniques are flowing-solution culture with standing solution, the nutrient film technique, and aeroponics. The flowing-standing culture employs intermittent or continuous replacement of the standing solution in the root zone by using a tub or trough with an overflow drain pipe. The solution is pumped from a reservoir to the plant roots, overflows through the drain, and is either discarded or returned to the reservoir (Fig. 1*b*) for aeration, pH control, and replenishment of nutrients. The nutrient film technique is similar, except that the root mass lies at the bottom of the tray or trough (Fig. 1*c*) and is exposed to a thin film of solution that is introduced to the upper end of a slightly tilted tray. The solution flows across the bottom of the tray and leaves at the lower end, to be either discarded or returned to the reservoir. In early applications of this technique, a capillary matting was used to distribute nutrients to the roots. In aeroponics, the plant is suspended, and the nutrient is sprayed on the roots and allowed to drain off to be discarded or recycled (Fig. 1*d*). The last two techniques require less water in the root zone, and with adequate control, the total water requirements can be much reduced. However, there is a greater danger of crop loss in case of pump failure. Regardless of the technique, recirculation increases the danger of spreading root disease.

Research and applications. Because it makes accurate control of the root zone possible, hydroponic culture is widely used in research on plant nutrition and on the effect of temperature and pH on roots. Of specific interest is the role of different mineral nutrients. Hydroponics can also be used to study the effect of microbes on plant health.

Commercial systems have been improved by the use of plastics and computer control, but still tend to be limited to certain crops for which appearance is important, such as lettuce, spinach, and tomatoes. Such growing is typically done in greenhouses or similar controlled-environment buildings.

Hydroponic systems also have obvious value in the field of education and for the amateur horticulturist, who can grow flowers or vegetables in a

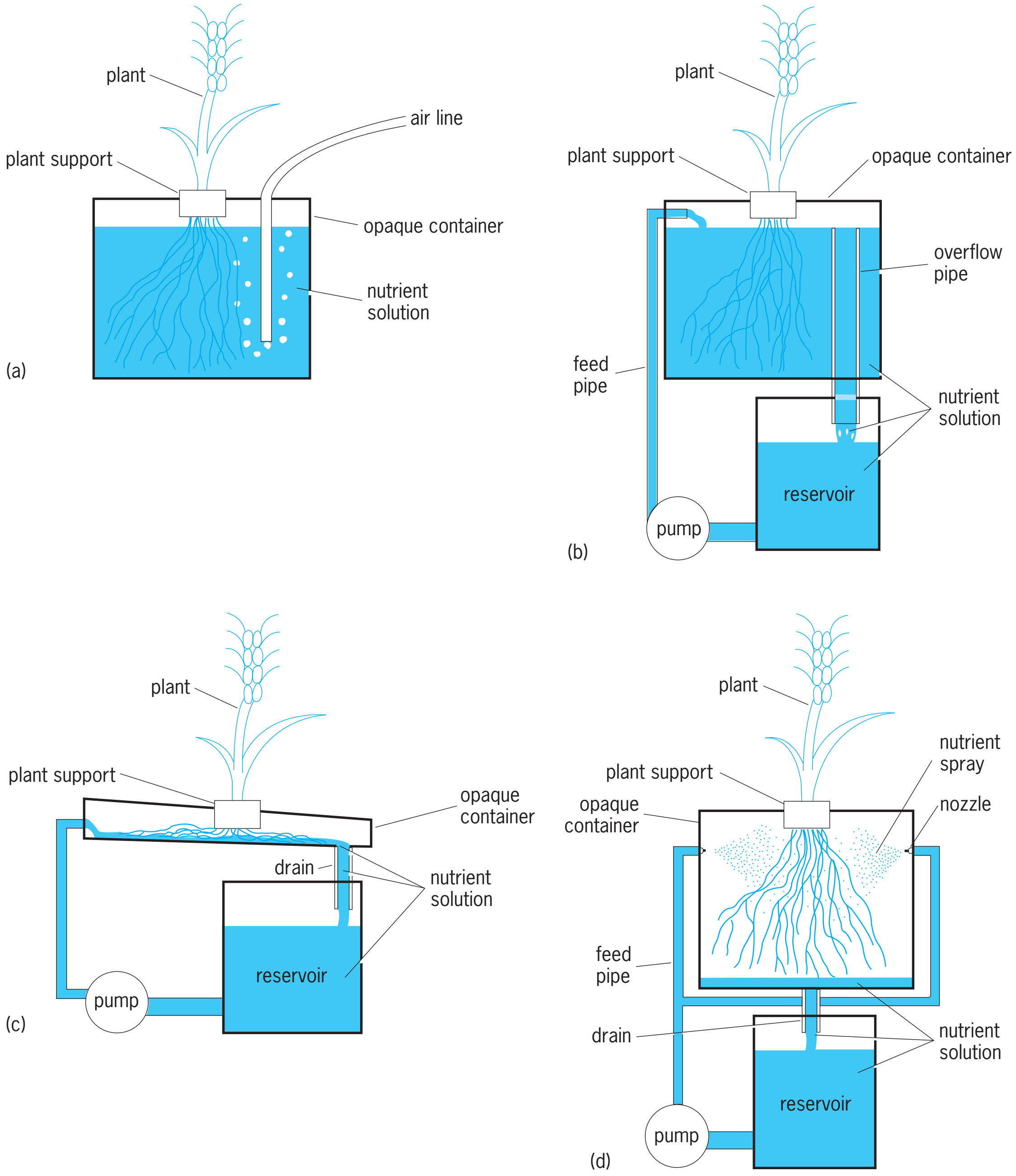


Fig. 1. Hydroponic techniques: (*a*) aerated standing culture; (*b*) flowing-solution culture with standing solution; (*c*) nutrient film technique (NFT); (*d*) aeroponics.

confined space with an indoor hydroponic garden. Hydroponic systems that have been developed for the NASA Controlled Ecological Life Support Systems (CELSS) program are used for crops not typically in hydroponic culture such as wheat, soybeans, white potatoes, and sweet potatoes. In a sealed, controlled-environment growth chamber at the Kennedy Space Center, crops are produced in a 120-ft^2 (16-m^2) growing space (**Fig. 2**), simulating a food production module for space. It is believed that conventional hydroponic systems on Mars or the Moon would produce oxygen and food and remove carbon dioxide, but that such systems could not be used in space because of the absence of the gravity needed for solution flow. Among the alternatives that have been proposed, the most promising system feeds nutrient solution to roots through a hydrophilic, microporous material. The roots grow on the surface of the material but do not penetrate it, and an opaque covering excludes light and retains moisture. A system using this technique at the Kennedy Space Center utilizes a tube of porous material contained within an opaque solid tube that shades and holds the roots, which surround the porous tube (**Fig. 3**). The nutrient solution is fed from a reservoir (bladder) through the porous tube under slight suction. Some solution moves by capillary action through the porous material to the roots; the rest is pumped back to the

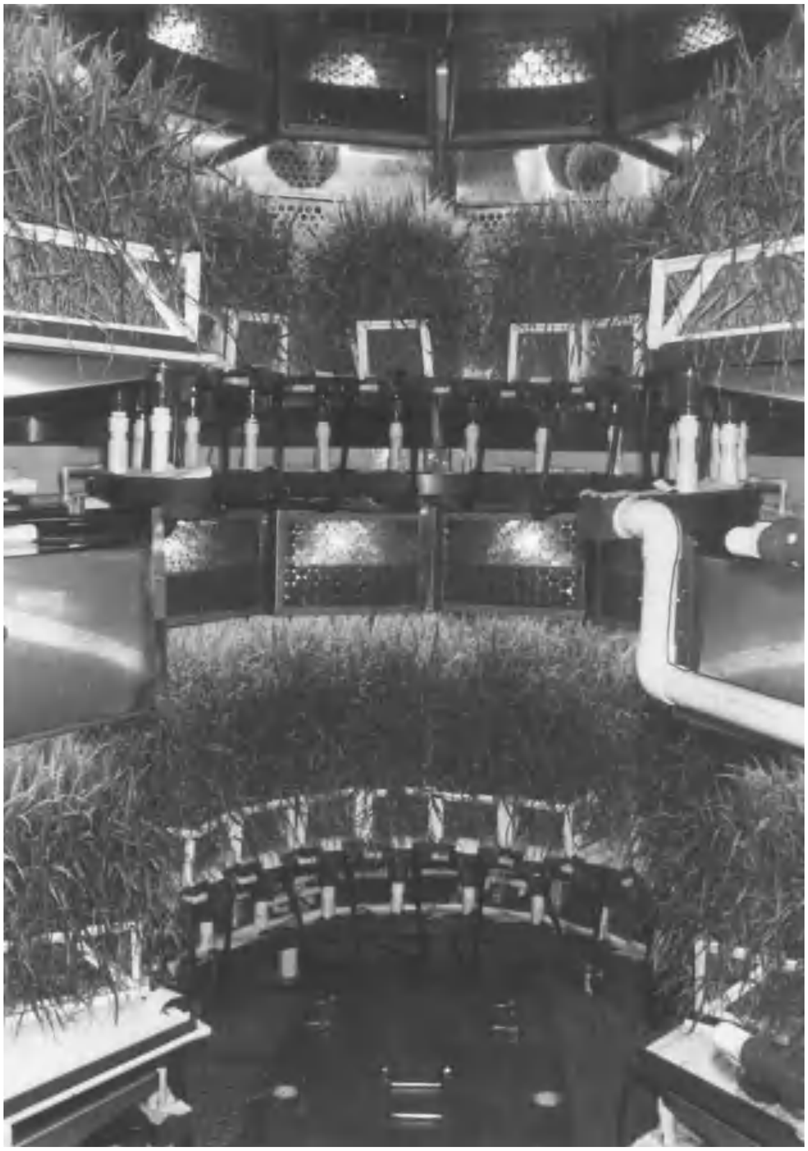


Fig. 2. Wheat grown under artificial lighting by using the nutrient film technique in the Biomass Production Chamber at the Kennedy Space Center. (*NASA*).

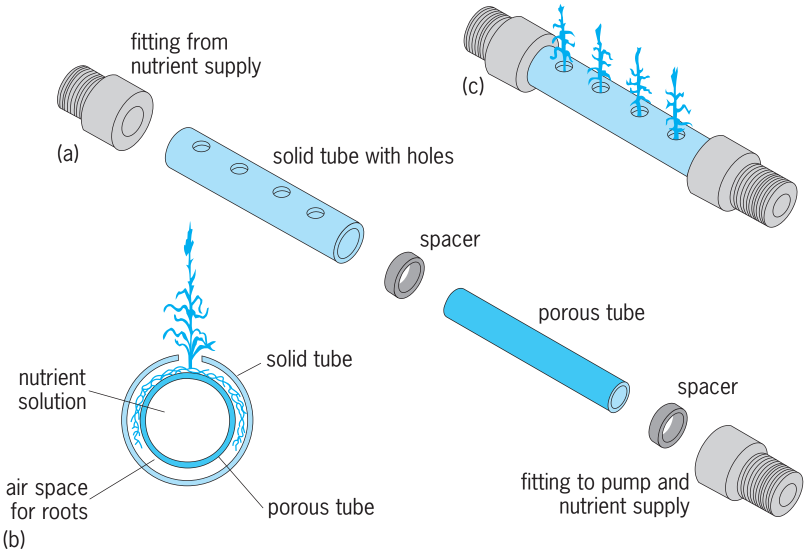


Fig. 3. Porous-tube plant-growth unit shown in (*a*) exploded view, (*b*) cross section, and (*c*) exterior view.

reservoir for control of pH and for replenishment of water and nutrients. Wheat, lettuce, beans, soybeans, and potatoes have been grown by using the system, with varying success. *See* PLANT GROWTH.

T. W. Dreschel

Bibliography. A. Cooper, *The ABC of NFT: Nutrient Film Technique*, 1979; T. W. Dreschel, *The Results of Porous Tube Plant Growth Unit Experiment T6B*, NASA Tech. Mem. 100988, 1988; D. R. Hoagland and D. I. Arnon, *The Water Culture Method for Growing Plants Without Soil*, Calif. Exp. Sta. Circ. 347, 1950; J. B. Jones, *A Guide for the Hydroponic and Soilless Culture Grower*, 1983.

Hydrosphere

The water portion of the Earth as distinguished from the solid part and from the gaseous outer envelope (atmosphere). Approximately 74% of the Earth's surface is covered by water, in either the liquid or solid state. These waters, combined with minor contributions from ground waters, constitute the hydrosphere.

The oceans account for about 97% of the weight of the hydrosphere, while the amount of ice reflects the Earth's climate, being higher during periods of glaciation. There is a considerable amount of water vapor in the atmosphere. The circulation of the waters of the hydrosphere results in the weathering of the landmasses. The annual evaporation from the world oceans and from land areas results in an annual precipitation of 320,000 km^3 (76,000 mi^3) on the world oceans and 100,000 km^3 (24,000 mi^3) on land areas. The rainwater falling on the continents, partly taken up by the ground and partly by the streams, acts as an erosive agent before returning to the seas. *See* GROUND-WATER HYDROLOGY; HYDROLOGY; LAKE.

The unique chemical properties of water make it an effective solvent for many gases, salts, and organic compounds. Circulation of water and the dissolved material it contains is a highly dynamic process driven by energy from the Sun and the interior of the Earth. Each component has its own geochemical cycle or pathway through the hydrosphere, reflecting the component's relative abundance, chemical properties, and utilization by organisms. The introduction of materials by humans has significantly altered the composition and environmental properties of many natural waters.

Rainwater. Rainwater contains small but measurable concentrations of many elements derived from the dissolution of airborne particulate matter and produced by equilibration of rainwater with atmospheric gases.

Total dissolved solids in rainwater range from over 10 parts per million (ppm) in rain formed in marine air masses to less than 1 ppm in rain precipitated over continental interiors. The major dissolved constituents of rainwater are chloride, sodium, potassium, magnesium, and sulfate (**Table 1**). These salts are derived over oceans and coastal areas from the dissolution of aerosol particles formed during the evaporation of sea spray. *See* AEROSOL.

TABLE 1. Chemical composition of waters, in parts per million

Constituent	Average rainwater	Average river water	Seawater
Sodium (Na)	1.98	6.3	10,500
Potassium (K)	0.3	2.3	380
Magnesium (Mg)	0.27	4.1	1,300
Calcium (Ca)	0.09	15.0	400
Chlorine (Cl)	3.79	7.8	19,000
Sulfate (SO_4)	0.58	11.2	2,650
Bicarbonate (HCO_3)	0.12	58.4	140
Silica (SiO_2)	—	13.1	6

A significant portion of the dissolved sodium, potassium, calcium, and sulfate in rain formed over continental areas is introduced by reaction with land-derived dust particles. Additional sulfate comes from the oxidation of sulfur dioxide, produced by the oxidation of hydrogen sulfide and by the burning of fossil fuels and smelting of sulfide ores [reaction (1)].

$$2SO_2 + O_2 + 2H_2O \rightarrow 4H^+ + 2SO_4^{2-} \tag{1}$$

A map of the average sodium content of rain in the continental United States (**Fig. 1*a***) shows sodium contours subparallel to the coastlines, reflecting mixing of continental air masses with salt-rich marine air. The distribution of sulfate (Fig. 1*b*) is more complex, and reflects significant continental input from dust storms and industrial activity.

Rainwater contains in dissolved state each of the gases present in the lower atmosphere. Carbon dioxide (CO_2) is derived from both biological respiration and the burning of fossil fuels. It reacts with rain to form carbonic acid (H_2CO_3) [reaction (2)].

$$CO_2 + H_2O \rightarrow H_2CO_3 \tag{2}$$

The presence of free oxygen and carbon dioxide makes rain both a natural oxidizing agent and an acid. Rain equilibrated with normal atmosphere has a pH of 5.7. More highly acidic rains form in areas where the industrial discharge of carbon dioxide or sulfur dioxide is intense. *See* ACID RAIN; PH.

Rain contains variable trace concentrations of many additional elements and compounds. Some of these, such as heavy metals and radionuclides, are derived from industrial pollution and nuclear testing, respectively. Precipitation of rain is the primary process by which many of these materials are transported from the atmosphere to the continents and oceans. *See* PRECIPITATION (METEOROLOGY).

Soil waters. As rainwater percolates downward and laterally through the soils and surface rocks of the continents, a complex group of reactions occurs.

The release of carbon dioxide and organic acids by bacterial processes increases the chemical reactivity of waters passing through the upper part of the soil zone. Weathering of most carbonate or silicate minerals generally involves acid attack, with carbonic acid dominating [reactions (3) and (4)].

$$\underset{\text{Calcite}}{CaCO_3} + H_2CO_3 \rightarrow Ca^{2+} + 2HCO_3 \qquad (3)$$

$$\underset{\text{Illite}}{2KAl_5Si_7O_{20}(OH)_4} + 2H_2CO_3 + 13H_2O \longrightarrow \underset{\text{Kaolinite}}{5Al_2Si_2O_5(OH)_4} + 4H_4SiO_4 + 2K^+ + 2HCO_3^- \qquad (4)$$

The weathering of carbonates and silicates consumes acid and produces a soil water enriched in cations, bicarbonate, and dissolved silica (H_4SiO_4). Dissolved sulfate is derived from dissolution of sulfate minerals and by reaction between sulfide minerals and dissolved oxygen in soil waters, as in reaction (5). Chloride is derived from the weathering of

$$\underset{\text{Pyrite}}{4FeS_2} + 15O_2 + 8H_2O \longrightarrow \underset{\text{Hematite}}{2Fe_2O_3} + 8SO_4^{2-} + 16H^+ \qquad (5)$$

fluid inclusions in silicate minerals and dissolution of halite (NaCl). *See* HALITE; SILICATE MINERALS; WEATHERING PROCESSES.

The ease with which a particular element can be accommodated in soil waters depends in part on the ionic radius *r* and charge *Z* of the cation that it forms (**Fig. 2**). The ratio of *Z* to *r* is known as the ionic potential. Large cations with a small charge, such as potassium (K^+) and calcium (Ca^{2+}), are usually readily accommodated in aqueous solution. In oxidizing environments, elements that form small, highly charged cations, such as the sulfur cation S^{6+}, combine with oxygen to form highly soluble and stable anionic complexes (for example, SO_4^{2-}). Cations of intermediate size and charge, however, including aluminum (Al^{3+}) and iron^{III} (Fe^{3+}), are only sparingly soluble. These elements are usually incorporated in the solid products of weathering, for example, aluminum in kaolinite in reaction (4) and iron in hematite in reaction (5). In soil waters depleted in free oxygen, the more highly soluble, reduced form of iron, Fe^{2+}, may go into solution. Other elements may also be solubilized in the absence of oxygen or in the presence of suitable complexing agents.

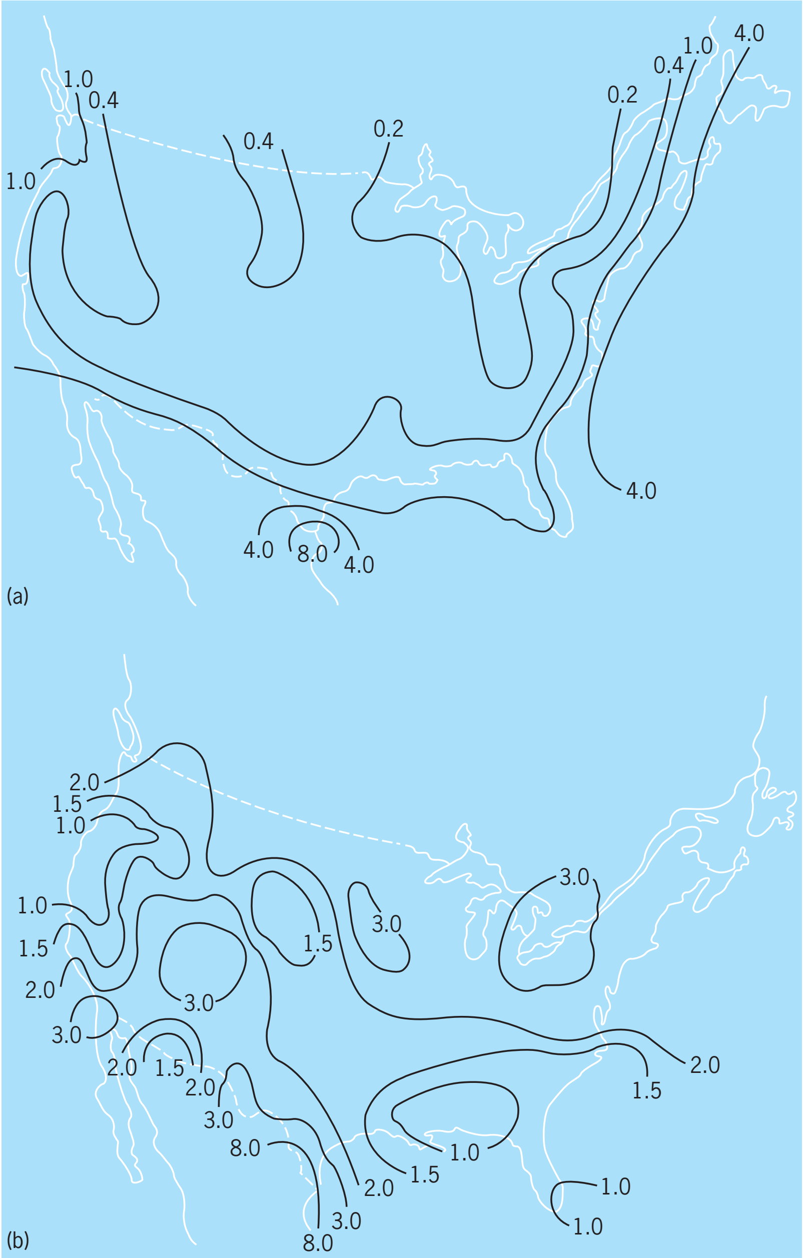


Fig. 1. Dissolved (*a*) sodium and (*b*) sulfate in rain (in parts per million) over the continental United States (*After R. M. Garrels and F. T. Mackenzie, Evolution of Sedimentary Rocks, W. W. Norton, 1971*)

River water. River water represents a variable mix of subsurface waters, which enter the river at the ground-water table, and surface runoff. Some of the material in river water is derived from the dissolved sea salts and dust present in rainwater, but most has been introduced through weathering reactions. River waters are higher in bicarbonate and dissolved silica, and the relative abundance of the cations they contain reflects the lithology of the drainage basins from which they are derived (Table 1). Waters draining carbonate terranes are typically enriched in calcium and magnesium [reaction (3)].

Shale terranes, in contrast, will produce waters preferentially enriched in potassium, which is released during weathering of illite [reaction (4)]. The salinity of river waters varies from less than 40 ppm for the Amazon River, which drains a region of exceptionally high rainfall, to over 800 ppm for the Rio Grande, which drains a region of low rainfall and high evaporation. Dissolved organic material is high in tropical streams and rivers, where rates of organic production and decay are high. Organics in many rivers draining the southeastern United States exceed the concentration of dissolved inorganic salts. The composition salinity of a given river may vary seasonally.

Rivers and streams have been used by humans since earliest history as a source of potable water, a place to discard wastes, and a vehicle for the transportation of goods. The concentrations of many metals and organic compounds deliberately or

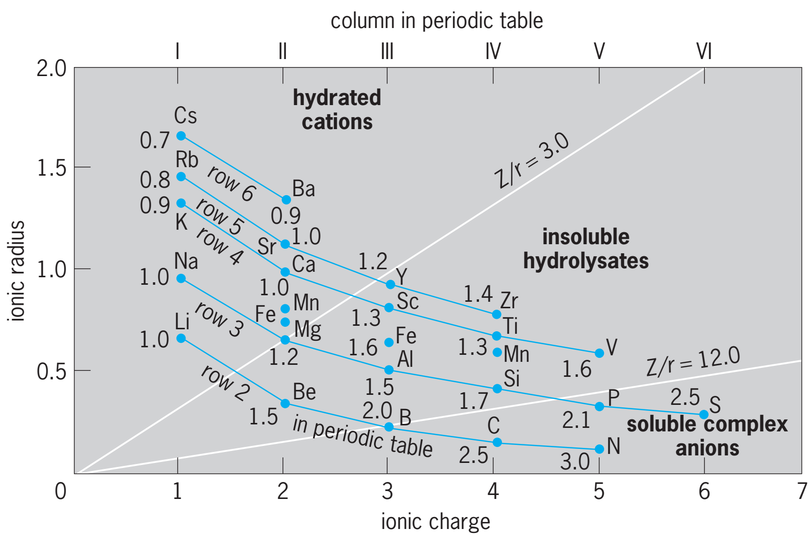


Fig. 2. **Accommodation of cations in aqueous solution. (*After H. Blatt, G. Middleton, and R. Murray, Origin of Sedimentary Rocks, Prentice-Hall, 1972*)**

accidentally introduced as wastes now often exceed natural river levels of these materials. Humans have also introduced compounds such as chlorinated hydrocarbons, which were unknown in the natural environment.

Most river water eventually mixes with marine waters in coastal and estuarine areas. The concentrations of most of the major cations and anions in these zones of mixing are not affected by processes other than the physical mixing of fresh and marine waters. Such constituents are said to behave conservatively. Many minor and trace constituents, however, behave nonconservatively, and are preferentially introduced into or removed from aqueous solution by chemical or biological processes occurring in the zone of mixing. Significant quantities of barium, for example, are desorbed from river clays when these particles are transported into marine waters. Humic-metal colloids present in river waters are flocculated as they mix with marine waters. The removal of dissolved iron by this process has been extensively documented. Field studies have shown that silica is removed from solution in some river estuaries. However, the question of whether this removal is due to biological uptake, reaction with suspended mineral particles, or both, has not been resolved. *See* RIVER.

Seawater. The dissolved salt content of open ocean water varies between 32,000 and 37,000 ppm. This range reflects dilution of seawater by rain and concentration by evaporation. Chloride, sodium, sulfate, magnesium, calcium, and potassium ions dominate sea salt (Table 1) and, with the exception of calcium, are present in remarkably constant proportions throughout the oceans. Other elements, such as boron, bromine, and fluorine, also show a constant ratio with chloride, but the chloride ratios of many elements vary significantly. The average pH of seawater is 8.2.

Most variations in the composition of seawater arise from the removal of elements by organisms that are living in surface seawater and the later release of these elements by the destruction of biologically produced particles which have sunk downward into deeper waters. Exceptions to this general rule are dissolved gases, whose solubility and concentration in surface seawater increase with decreasing temperatures.

Marine plants can live only in surface seawater, where sufficient light is available for photosynthesis. These organisms give off oxygen and extract carbon dioxide and nearly all of the dissolved nitrate and phosphate from seawater to produce organic matter. Some plants, in addition, secrete solid particles of calcium carbonate ($CaCO_3$) or opaline silica ($SiO_2 \cdot nH_2O$). Marine plants are consumed by animals, some of which also extract dissolved calcium, bicarbonate, and silica to make carbonate or opaline shells or tests. During the downward rain of particles produced by plants and animals in surface waters, destruction of organic matter by bacteria and animals releases dissolved nitrogen, phosphorus, and carbon dioxide back into the water column at depth and consumes dissolved oxygen (**Fig. 3**). Ocean waters are undersaturated with respect to opaline silica, and these particles begin to dissolve, releasing dissolved silica, after the death of their parent organism. Some particles, however, reach the sea floor to accumulate as siliceous oozes. Carbonate particles are stable in surface waters, which are supersaturated, and accumulate readily in shallower areas of the sea floor. Deeper waters are undersaturated because of increased pressure, and below depths of 4000 m (2.5 mi) the degree of undersaturation is such that carbonate dissolves very rapidly.

Thus, in response to biological processes, nitrogen, phosphorus, and silicon are almost totally depleted in surface waters, and marine plant life can flourish only where upwelling currents renew surface water in these biolimiting elements. Elements that show some lowering in concentration in surface waters are carbon, copper, nickel, and cadmium and the alkaline earths calcium, strontium, barium, and radium. The behavior of strontium, barium, and radium may reflect in part their coprecipitation with calcium in carbonate. Dissolved oxygen is unique in that it is produced at the surface and consumed at depth (Fig. 3). Analytical data for many elements are not precise enough to establish patterns of variation in their concentration. *See* UPWELLING.

In closed basins on the sea floor, stagnant bottom waters can become totally depleted in dissolved oxygen. In anoxic waters, anaerobic respiration reduces sulfate and forms hydrogen sulfide. Iron and manganese become more soluble and may increase in concentration, while other metals, such as copper, precipitate out as sulfides. *See* ANOXIC ZONES; SEAWATER.

Surface brines. Evaporation of fresh waters flowing into closed basins on the continents typically produces alkaline brines (**Table 2**; Soap Lake). Calcium and magnesium precipitate out as insoluble carbonates or hydroxysilicates. Sodium and potassium concentrate continuously, and total carbonate and pH increase. Chemical evolution of marine waters during evaporation follows a different course. Gypsum ($CaSO_4 \cdot 2H_2O$) is the first mineral to precipitate out during continued evaporation, followed by halite (NaCl). Reaction with carbonates to form dolomite [$CaMg(CO_3)_2$] may remove magnesium (Table 2; Dead Sea). *See* SALINE EVAPORITES.

Subsurface waters. Marine waters trapped in the pore spaces of sediments during deposition react with the mineral and organic particles surrounding them and undergo significant changes in composition. Pore waters in organic-rich sediments are quickly depleted in dissolved oxygen, and anaerobic reduction of sulfur destroys dissolved sulfate and produces hydrogen sulfide. Anaerobic reduction of carbon dioxide in the absence of sulfate produces methane. Changes in the relative proportions of dissolved cations occur as a result of diagenetic reactions with silicate minerals. The vertical variation in concentration of dissolved species in sediment pore waters provides valuable information on the nature and rates of diagenetic reactions in fine-grained marine sediments.

TABLE 2. Chemical composition of some saline waters, in parts per million

Constituent	Soap Lake, Washington	Dead Sea, southwest Asia	Subsurface brine, Louisiana
Sodium (Na)	12,500	39,700	63,900
Potassium (K)	12,500	7,590	869
Magnesium (Mg)	23	42,430	1,070
Calcium (Ca)	4	17,180	9,210
Chlorine (Cl)	4,680	219,250	124,000
Sulfate (SO_4)	6,020	420	153
Bicarbonate (HCO_3)	11,270	220	115
Carbonate (CO_3)	5,130	—	—
Silica (SiO_2)	101	—	16

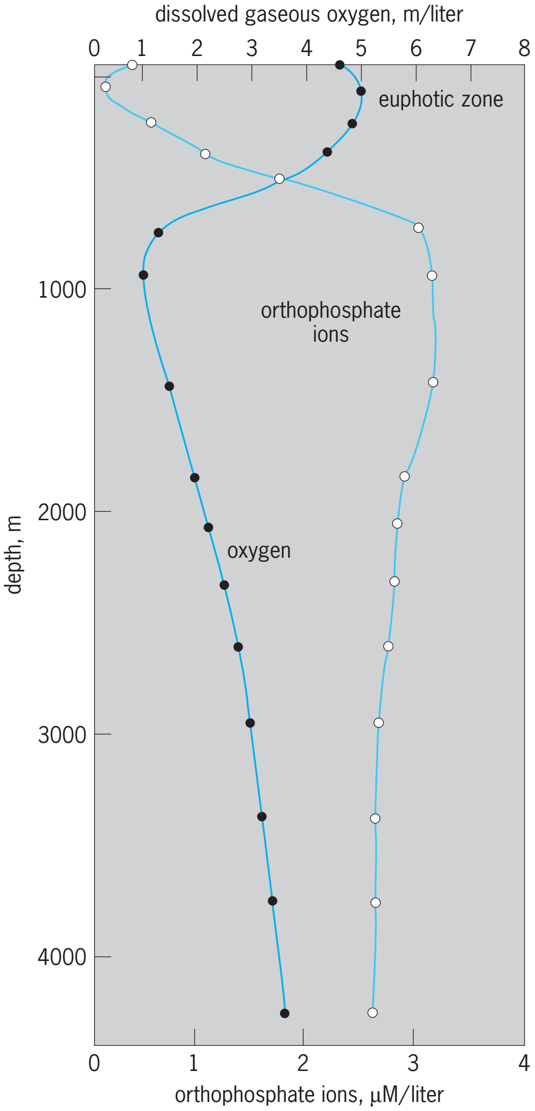


Fig. 3. Distribution of dissolved gaseous oxygen (low values) and nutrient species (high concentrations) of orthophosphate ions at 26°22′.4 N and 168°57′.5 W in the Pacific Ocean. 1 m = 3.3 ft.

Subsurface waters in deep sedimentary basins are often far more saline than the connate pore waters trapped at the time of sediment deposition. Salinities of 100,000–150,000 ppm are common but can exceed 400,000 ppm. Basinal brines are typically dominated by sodium and chloride, but calcium becomes the most abundant cation at extreme salinities. Subaerial evaporation of marine and continental waters and the subsurface dissolution of halite can produce the range of salinities and dissolved chloride concentrations observed for most subsurface basinal brines, but not their major cation compositions (Table 2). The observed systematic increases in individual dissolved major cations and decreases in pH and alkalinity with increasing salinity observed in deep sedimentary waters worldwide support the hypothesis that the approach toward thermodynamic buffering by silicate-carbonate mineral assemblages is the first-order control on subsurface fluid compositions, even at temperatures well below 100°C (212°F).

Dissolved organic acid anions, such as acetate, are associated primarily with low-salinity basinal waters; and dissolved ore-forming metals, such as copper, lead, and zinc, are preferentially found in basinal brines having salinities in excess of 200,000 ppm. The high chloride concentration and low pH of these saline waters may enhance solubilization of metals through chloride complexing. Long-distance migration of brines and other subsurface fluids appears to be a common aspect of the geological evolution of sedimentary basins. *See* BASIN; DIAGENESIS.

Ice. Ice is a nearly pure solid, and in contrast to the solvent power of liquid water, few foreign ions can be accommodated in its lattice. Ice does contain particulate matter, however, and the change in the composition of these particles with time, as recorded in the successive layers of ice that have accumulated in polar regions, has provided much information on the progressive input of lead and other materials into the environment by humans. *See* SEA ICE.

Jeffrey S. Hanor

Bibliography. E. K. Berner and R. A. Berner, *The Global Water Cycle: Geochemistry and Environment*, 1987; J. I. Drever, *The Geochemistry of*

Natural Waters, 3d ed., 1997; J. S. Hanor, *Origin of Saline Fluids in Sedimentary Basins*, 1994.

Hydrostatics

The study of fluids at rest; that is, at every point in the fluid, the velocity and acceleration are zero. The only force acting at a point on the surface of a small volume of fluid is due to the pressure at that point, and the force is perpendicular to the surface. Application of Newton's second law shows that the pressure is independent of the orientation of the surface on which it is acting, and that the pressure increases with depth.

The pressure difference between any two points in the fluid is given by Eq. (1), where p_1 and p_2 are the

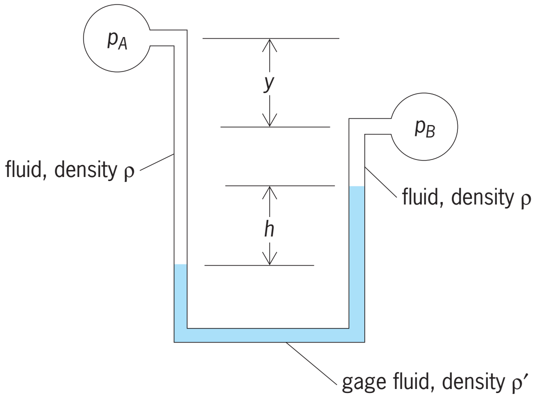
$$p_2 - p_1 = \int_{z_1}^{z_2} \rho g \, dz \qquad (1)$$

pressures at the two points, z_1 and z_2 are the vertical coordinates (z positive downward) of the two points, ρ is the density of the fluid, and g is the acceleration due to gravity. The integration on the right-hand side can be carried out only if a relationship between density and elevation is known. Variations of g with elevation are negligible in earth-bound engineering problems. If the density of the fluid is constant, Eq. (1) becomes Eq. (2), where $h = z_2 - z_1$ is the

$$p_2 - p_1 = \rho g(z_2 - z_1) = \rho g h \qquad (2)$$

vertical distance between the two points. The pressures p_2 and p_1 may be either absolute pressures or gage pressures. The absolute pressure is the sum of the gage pressure and atmospheric pressure. *See* PRESSURE.

Manometers. Equation (2) shows that a pressure difference may be expressed in terms of the height h of a column of fluid of density ρ. A manometer is a device that makes use of this equivalence. An example of a U-tube manometer with its attendant equation is shown in **Fig. 1**. Manometers are available in a wide variety of configurations. Each configuration has its own attendant equation, but all of the equations are derivable from Eq. (2). *See* MANOMETER; PRESSURE MEASUREMENT.



$$p_A - p_B = gh(\rho' - \rho) - \rho g y$$

Fig. 1. U-tube manometer and its attendant equation for equation for determining pressure difference $p_A - p_b$.

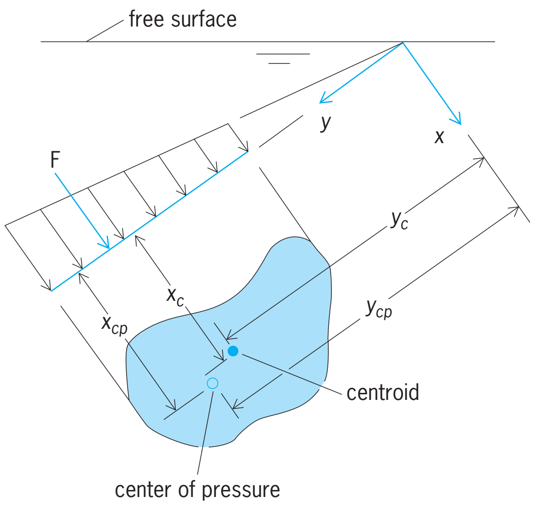


Fig. 2. Fluid force on a plane surface. Symbols are explained in text.

Forces on plane submerged surfaces. The design of tanks, dams, and other devices for containing fluids at rest requires the calculation of the force exerted by the fluid on the walls of the container. The force is calculated by integrating the pressure distribution over the submerged area. **Figure 2** illustrates the pressure distribution on a plane submerged area and shows the notation used. The coordinates are chosen so that the plane of the area is the xy plane, and the intersection of this plane with the free surface is the x axis.

For a fluid of constant density, the magnitude of the force exerted by the fluid is given by Eq. (3), where

$$F = p_c A \qquad (3)$$

p_c is the pressure at the centroid of the submerged area and A is the submerged area. The force **F** is distributed over the entire submerged area, but it is convenient to replace it with a concentrated equivalent force. The point of application of the concentrated equivalent force is called the center of pressure. The location of the center of pressure is given by Eqs. (4),

$$\begin{aligned} y_{cp} &= y_c + \frac{I_{yc}}{y_c A} \\ x_{cp} &= x_c + \frac{I_{xyc}}{y_c A} \end{aligned} \qquad (4)$$

where x_c and y_c are the coordinates of the centroid of the submerged area, I_{yc} is the moment of inertia of the submerged area with respect to an axis parallel to the x axis through its centroid, and I_{xyc} is the product of inertia with respect to centroidal axes. *See* CENTROIDS (MATHEMATICS); MOMENT OF INERTIA; PRODUCT OF INERTIA; STATICS.

Forces on curved submerged surfaces. The easiest way to calculate the force exerted on a curved submerged surface (**Fig. 3**) is to calculate the horizontal and vertical components of the force and then

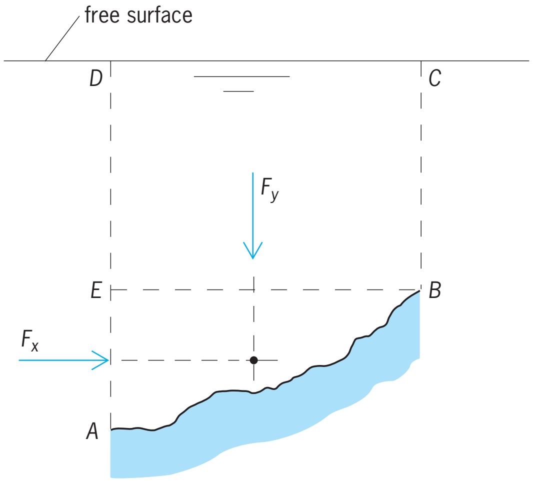


Fig. 3. Fluid force on a curved surface. Symbols are explained in text.

combine them into a single force, if necessary. The horizontal component F_x is determined by calculating the force on a projected area which is equal to the projection of the curved area on a vertical plane (area *EA* in Fig. 3). The horizontal force component is calculated by using Eqs. (3) and (4) applied to this projected area.

If the fluid is above the curved surface, the vertical force component F_y is equal to the weight of the fluid above the surface (*DEABCD* in Fig. 3), and its line of action passes through the center of gravity of this weight of fluid. If the fluid is below the curved surface, the vertical force component acts upward and is equal to the weight of the fluid displaced by the curved surface. The line of action of the upward force passes through the center of gravity of the displaced fluid.

Submerged and floating bodies. The pressure distribution on a floating or submerged body gives rise to a net upward force which is called the buoyant force F_B. This buoyant force is equal to the weight of the fluid displaced by the body. This equality is known as Archimedes' principle. If the weight of the body *W* is greater than the buoyant force, the body will sink to the bottom of the fluid. If the weight of the body is equal to the buoyant force, the body will float. The line of action of the buoyant force is through the center of gravity of the displaced fluid, which, in this case, is called the center of buoyancy. *See* ARCHIMEDES' PRINCIPLE; BUOYANCY.

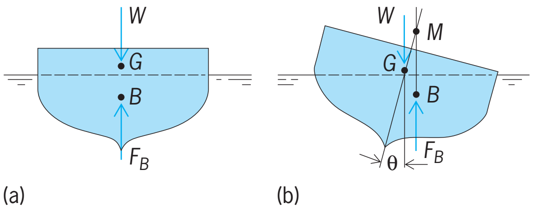


Fig. 4. Stability of a floating body (*a*) at equilibrium position and (*b*) at angular displacement θ from equilibrium position. Symbols are explained in text.

Stability of floating bodies. The stability of a floating body depends on the position of the center of buoyancy *B* relative to the center of gravity *G* of the body (**Fig. 4**). The position of the center of gravity of a body depends on its shape and on the distribution of mass within that shape. The position of the center of buoyancy, since it is the center of gravity of the fluid displaced by the body, depends only on the shape of that part of the body which is submerged.

For a particular body in Fig. 4, an angular displacement θ from the equilibrium position produces a couple that tends to return the body to its equilibrium position. This body has a positive righting moment, and the equilibrium is stable. The point *M* located at the intersection of the originally vertical line through *G* and the vertical line through the center of buoyancy *B* is called the metacenter, and the distance *GM*, the metacentric height. As the angle θ is increased, the shape of the displaced volume of fluid changes, and as a result the position of the center of buoyancy changes. If the angle θ is made large enough, the point *M* will lie below the point *G* and the righting moment will become negative. In this position the equilibrium is unstable and the body will capsize. *See* SHIP DESIGN. Warren M. Hagist

Bibliography. R. L. Daugherty, J. B. Franzini, and E. J. Finnemore, *Fluid Mechanics*, 9th ed., 1997; R. W. Fox and A. T. McDonald, *Introduction to Fluid Mechanics*, 5th ed., 1998; V. L. Streeter and E. S. Wylie, *Fluid Mechanics*, 9th ed., 1998; F. M. White, *Fluid Mechanics*, 4th ed., 1998.

Hydrothermal ore deposits

The predominant sources of scarce metals, such as copper, zinc, lead, tin, molybdenum, antimony, tungsten, mercury, bismuth, uranium, silver, and gold. Fascinating because of their variety, they are the subject of intense research aimed at understanding them better and finding them efficiently. *See* ORE AND MINERAL DEPOSITS.

Formation. Hydrothermal ore deposits are inferred to have formed by mineral precipitation from hot hydrous fluids. This deposition occurred within the rocks through which these fluids traveled, or as the fluids entered bodies of water (ocean or lakes), or as they reached the atmosphere at the Earth's surface. These inferences are based on the following observations: (1) Most hydrothermal ores occur in places that favor the passage of fluids: fracture zones (veins), pipelike conduits (chimneys, pipes) in easily soluble rocks (such as limestone), sedimentary beds that are porous or susceptible to chemical replacement (mantos), fractured caps over intrusive cupolas (porphyries), and porous calcsilicates (skarns) at contacts of magmatic intrusives with limestone. Other hydrothermal ores occur where fracture zones reached lake or ocean waters (as along mid-ocean ridges) or the atmosphere (as in hot springs above presently active geothermal reservoirs). (2) Hydrothermal minerals often contain small fluid inclusions consisting mostly of water with some carbon dioxide; this water has significant concentrations of common salt and of several of the metals commonly found in hydrothermal deposits. These inclusions are interpreted as trapped remnants of the mineralizing fluids. Such fluid inclusions tend to homogenize between 50 and 500°C (120 and 930°F), suggesting that their host minerals formed at temperatures above the homogenization temperatures from one-phase fluids. (3) The original composition of hydrothermal ore host rocks was generally modified by the formation of water-bearing minerals (hydrothermal alteration). (4) Many hydrothermal deposits are in the vicinity of magmatic intrusive rocks, implying that they formed either from fluids emanating from a crystallizing magma or from fluids put in motion by magmatic heat. *See* LIMESTONE; MID-OCEANIC RIDGE; PORPHYRY; SEDIMENTARY ROCKS; SKARN.

Hydrothermal fluids may be generated in different ways: (1) as water expelled from a crystallizing magma; (2) as meteoric water that becomes ground water and is mobilized by gravity, tectonic activity (for example, tilting or folding), or magmatic heat and reacts with continental crustal rocks; (3) as seawater that infiltrates near mid-ocean ridges, is mobilized by ocean ridge magmatism, and reacts with oceanic crust; or (4) as water trapped in sediments (connate water) and released upon compaction and heating (diagenesis and metamorphism). *See* DIAGENESIS; MAGMA; METAMORPHISM.

Composition of ore and host rocks. The ore minerals in hydrothermal deposits are mostly sulfides of copper, zinc, lead, tin, molybdenum, antimony, mercury, bismuth, and silver. However, some metals occur as oxides (tin and uranium), tellurides (gold), tungstates (tungsten), or native (gold). The accompanying nonvaluable gangue minerals consist mostly of sulfides (pyrite, pyrrhotite, orpiment, realgar), oxides (quartz), carbonates (calcite), sulfates (barite), fluorides (fluorite), and silicates (rhodonite). *See* MINERAL; SULFIDE AND ARSENIDE MINERALS.

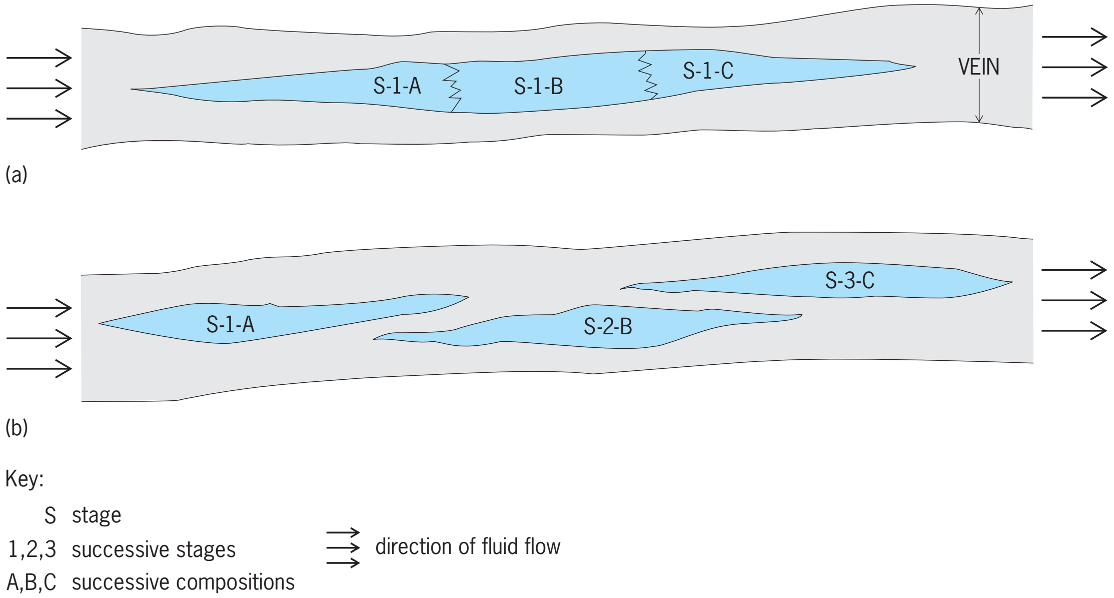
Hydrothermal alteration. Hydrothermal alteration of the surrounding host rocks due to contact with the hot hydrous fluids may extend from a few meters to a few kilometers from an ore. The alteration generally consists of addition of silica and conversion of anhydrous silicates (orthoclase, plagioclase, olivine) to hydrous silicates (muscovite, sericite, kaolinite, biotite, epidote, chlorite) by a low-pH fluid. Where present, the sulfate mineral alunite indicates that the fluid was oxidized and acidic. Where magma intruded limestone, hydrothermal fluids promoted the formation of porous silicates (garnet, wollastonite)—the skarns that host hydrothermal ores of copper, zinc, gold, or tungsten.

Mineralization pulses and stages. Hydrothermal ores often contain specific minerals or mineral groups (paragenetic associations or assemblages) that either form successive layers or cut prior minerals, thus defining a paragenetic sequence of mineralization stages. Thin sections of transparent minerals and etching or electron microprobe analyses of opaque minerals reveal that each mineral consists of many hundreds of growth bands, proving that an ore deposit is made by hundreds or thousands of mineralization pulses.

Zoning. The chemical composition of hydrothermal ores often varies spatially within a mining district or within a given vein, manto, pipe, or disseminated deposit. Over a century ago, this led to a zonal theory, according to which certain minerals or elements were considered to be characteristic of the upper or peripheral parts (chalcedony, barite, fluorite, mercury, antimony, gold-silver), others of the central part (silver, lead, zinc, copper), and still others of the inner or lower part (pyrrhotite, arsenopyrite, tourmaline, cassiterite, gold, arsenic, bismuth, tungsten, tin). This was only a generalized scheme, because every mineral deposit and district has its own zoning sequence.

Hydrothermal zoning has traditionally been considered to result from progressive changes in the temperature, pressure, or composition of the hydrothermal fluids along their paths. However, field observations have so far failed to provide convincing evidence that a specific paragenetic stage (S-1) evolves chemically (from composition A to compositions B and C) from one part of a deposit to another (**illus.** *a*). It seems more likely that successive stages (S-1, S-2, and S-3) are deposited by different fluids (A, B, C) at various points along the hydrothermal fluid paths, thus generating a zoned deposit (illus. *b*).

Oxidation, leaching, and supergene enrichment. The near-surface part of most hydrothermal ore deposits contains hydroxides and oxides of iron, copper, and manganese (limonite, hematite, cuprite, tenorite, pyrolusite, manganite), sulfate of lead (anglesite), carbonates of copper, lead, and zinc (malachite, azurite, brochantite, cerussite, smithsonite), and chloride of silver (cerargyrite). This oxide zone passes in depth into the underlying hypogene sulfides. Copper was often leached from the oxide zone, but precipitated again as a sulfide (chalcocite, covellite) between the sulfide and oxide zones to form a supergene enrichment zone or blanket. This sequence results from the near-surface atmospheric oxidation and dissolution of most of the original hypogene sulfides, which leads to the formation of acid ground waters. The downward-percolating ground waters are reduced at depth, just below the water table, leading to the precipitation of the supergene copper sulfides. Other factors influencing near-surface processes are the reactivity of the host rock to the acid ground water (for example, limestone is much more effective in neutralizing acid ground water than igneous, metamorphic, or other sedimentary rocks) and changes in the water table due to climatic or sea-level variations. Supergene enrichment of copper played an important role in making it economic to mine disseminated (porphyry) copper deposits in the southwestern United States, Mexico, and northern Chile. Residual enrichment in the oxide zone of relatively insoluble gold, and lead or zinc carbonates, may have played a role in the discovery of some deposits. *See* CARBONATE MINERALS; OXIDE AND HYDROXIDE MINERALS.



Alternative explanations of zoning in hydrothermal ore deposits. (*a*) A single-stage fluid (S-1) evolves along its path, depositing successively different mineral compositions or assemblages (A, B, and C). (*b*) Three different fluids flow through the vein at various times or stages (S-1, S-2, and S-3), depositing various mineral compositions or assemblages (A, B, and C).

Ore deposition factors. Perhaps more important than the sources of the metals and of the hydrothermal fluids are the factors determining the uptake and transport of ore constituents in hydrothermal fluids, as well as the causes of hydrothermal ore deposition. It has long been known that elevated temperatures and acidity (low pH) promote increased metal solubility. Experimental studies also highlighted the importance of chloride, sulfide, organic, and other complexing agents in metal solubility. Thus, magmas contaminated by major evaporite sequences (which contain chloride in halite, and sulfide in gypsum or anhydrite) would produce more "fertile" hydrothermal fluids than uncontaminated magmas. Similarly, heated sedimentary basin brines or infiltrating seawater would be more likely to dissolve and transport significant amounts of metals than dilute ground waters. In contrast, hydrothermal ore deposition would require cooling or neutralization (increasing pH) of the metal-bearing fluids.

Temperature. In the early 1900s, the temperature of the hydrothermal fluids was considered to largely determine the mineralogical composition of an ore deposit. This led to a temperature classification of hydrothermal ore deposits (epithermal, leptothermal, mesothermal, hypothermal) based mainly on minerals deemed to be indicative of the temperature of ore deposition. However, subsequent laboratory studies of pertinent chemical systems demonstrated that supposedly "diagnostic" minerals are stable over wide ranges of temperature and are, therefore, generally unreliable as geothermobarometers. For example, pyrrhotite (FeS_{1-x}) had been thought to deposit at a higher temperature than pyrite (FeS_2), but the experimental studies demonstrated that both are stable over the general range of hydrothermal ore deposition (and that the deposition of one or the other depends more on the concentrations of sulfur and iron in the hydrothermal fluid). Unfortunately, reliable geothermometers, such as temperature ranges of different polymorphs of a mineral (for example, argentite and acanthite) or incompatible mineral assemblages (pyrite-magnetite versus pyrrhotite-hematite), only indicate large temperature ranges, are generally scarce, or are restricted to unusual ore deposition temperatures. Studies of fluid inclusion homogenization temperatures and of light-isotope fractionations have shown that hydrothermal ore deposition occurs at temperatures from several hundred to a few tens of degrees Celsius. In addition, they showed that in some ore deposits there was a fast initial increase and a subsequent gradual decrease in temperature, as evidenced by the depositional temperatures of successive paragenetic mineralization stages. These studies have so far failed to clearly demonstrate thermal gradients along hydrothermal solution paths for a single depositional stage. At present, the temperature gradient factor is supported mainly by pointing out the temperature gradients observed in geothermal fields and by noting mineral precipitation by cooling solutions in laboratory experiments.

Acidity. During the mid-1900s, it became evident that the hydrothermal alteration of rocks adjoining veins and disseminated ore results largely from attack by acid fluids. This has been reinforced by abundant laboratory evidence, and a pH increase due to the interaction of hydrothermal fluids with host rocks is now seen as a more general cause of hydrothermal ore deposition, especially for lead-zinc ores in limestone. In some geothermal fields (for example, Yellowstone), hot springs of substantially different pH occur in proximity, suggesting the possibility that fluid mixing may increase the pH of an acid metal-bearing fluid and cause mineral precipitation. Fluid mixing also may be responsible for mineral precipitation by bringing together ions that form insoluble compounds.

Pressure. Decreasing pressure gradients have been invoked as causes for mineral deposition because hydrothermal fluids tend to flow toward low-pressure regions. Fluid inclusion studies have shown that some minerals precipitated from boiling solutions (because different inclusions in a given mineral trapped varying proportions of liquid to water, thus giving an apparent range of homogenization temperatures). However, this appears to be an occasional rather than a predominant factor.

Most chemical equilibria vary with both temperature and pressure (that is, they are really geothermobarometers), so that reliable pressure gradients can be determined only if a pressure-independent geothermometer is available (such as a light-isotope fractionation).

Oxidation and reduction. Reduction of oxidized hydrothermal fluids is generally credited for the deposition of uranium, vanadium, copper, or silver ores in redbeds containing organic matter. Conversely, oxidation of hydrothermal fluids is also considered to be occasionally important in ore precipitation. *See* REDBEDS.

Fugacity. The decrease in emphasis on depositional temperature was supplanted by an increased awareness of the roles played by the fugacities of sulfur, oxygen, and carbon dioxide in hydrothermal solutions. (Fugacity is the partial pressure of a component in a perfect gas that is in equilibrium with a liquid or solid containing this gas in solution.) Eventually, this led to the current tendency to classify ore deposits as low-sulfidation, high-sulfidation, or acid-sulfate. But given that such expressions provide little information on the nature of a specific ore deposit, and given the difficulty involved in pinpointing the exact causes of ore deposition, there has been a parallel development of a hodge-podge terminology of ore deposit types, such as "Cordilleran vein, manto, pipe or breccia," "five-element (Ag-Ni-Co-Fe-As) vein," "disseminated or porphyry," "skarn," "volcanogenic massive sulfide," "redbed or sediment-hosted stratiform Cu-Ag," and "igneous-metamorphic." *See* FUGACITY. Ulrich Petersen

Bibliography. H. L. Barnes, *Geochemistry of Hydrothermal Ore Deposits*, 3d ed., Wiley, 1997; J. M. Guilbert, Linkages among hydrothermal ore deposit types, *Pro-Explo 2001*, Lima, Peru, 2001; J. M. Guilbert and C. F. Park, Jr., *The Geology of Ore Deposits*, W. H. Freeman, 1986; H. D. Holland, Some applications of thermochemical data to problems of ore deposits, I. Stability relations among the oxides, sulfides, sulfates and carbonates of ore and gangue minerals, *Econ. Geol.*, 54:184–233, 1959; H. D. Holland, Some applications of thermochemical data to problems of ore deposits, II. Mineral assemblages and the composition of ore-forming fluids, *Econ. Geol.*, 60:1101–1166, 1965.

Hydrothermal vent

A hot spring on the ocean floor, where heated fluids exit from cracks in the Earth's crust. Most hydrothermal vents occur along the central axes of mid-oceanic ridges, which are underwater mountain ranges that wind through all of the deep oceans. The best-studied vents are at tectonic spreading centers on the East Pacific Rise and at the Mid-Atlantic Ridge. However, vents are also found over hot spots such as the Hawaiian Islands and Iceland, in back-arc basins such as those in the western Pacific, in shallow geothermal systems such as those off the Kamchatka Peninsula, and on the flanks of some underwater volcanoes and seamounts. Major parts of the mid-oceanic ridges, including those of the Indian Ocean and Southern Hemisphere, are poorly explored, but available evidence suggests that these areas should also have extensive hydrothermal venting. Hydrothermal vent sites, or closely grouped clusters of vent deposits and exit ports, may cover areas from hundreds to thousands of square feet (tens to hundreds of square meters). Individual vent sites may be separated along mid-ocean ridges by more than 1000 mi (1600 km). *See* MID-OCEANIC RIDGE; SEAMOUNT AND GUYOT.

Hydrothermal fluid formation. All of the hydrothermal vent sites occur in areas where quantities of magma exist below the sea floor (**Fig. 1**). Cold seawater is drawn down into the oceanic crust toward the heat source. As the seawater is heated and reacts with surrounding rock, its composition changes. Sulfate and magnesium are major components of seawater lost during the reactions; sulfide, metals, and gases such as helium and methane are major components gained. This modified seawater is known as hydrothermal fluid. Buoyant, hot hydrothermal fluid rises toward the sea floor in a concentrated zone of upflow to exit from the sea floor at temperatures ranging from 50°F (10°C) to greater than 750°F (400°C), depending on the degree of cooling and of mixing with seawater during the ascent. If the sea floor is shallow enough and the fluid hot enough, the solution may boil; but it usually does not because of the pressure of overlying seawater at depths of 4900–9800 ft (1500–3000 m) or more. *See* MAGMA.

Fluid venting. Hydrothermal fluid that mixes extensively with seawater below the sea floor surface may reach the sea floor as warm springs, with temperatures of 50–86°F (10–30°C). This outflow is usually detectable as cloudy or milky water, but the flow is slow and no mineral deposits accumulate except for some hydrothermal staining or oxidation of sea floor basalts. When hotter, relatively undiluted hydrothermal fluid reaches the sea floor, it is still buoyant with respect to seawater, so that the hot solution rises out of cracks in the sea floor at velocities up to about 6 ft (2 m) per second, mixing turbulently with seawater as it rises. Mixing of hydrothermal fluid with seawater leads to precipitation of minerals from solution, forming mineral deposits at the exit from the sea floor and so-called smoke, tiny mineral particles suspended in the rising plume of fluid. Black smoker vents are distinguished by the presence of such large quantities of minute mineral particles that the plumes become virtually opaque (**Fig. 2**).

The roughly cone-shaped plume of hydrothermal fluid continues to rise while mixing with seawater until the density of the mixture is equal to that of the surrounding seawater. It then spreads horizontally as a layer within the water column that is detectable by chemical analysis for trace components such as helium-3, an isotope originating in the Earth's interior, or methane (CH_4). Detection of this hydrothermal layer, which can be traced back to the general area of its source, is one of the techniques for detecting large areas of hydrothermal venting.

Formation and outflow of hydrothermal fluid makes a major contribution to the concentration and balance of elements in the oceans by changing the composition of seawater. The quantities of elements added or removed from the oceans by hydrothermal venting around the world are comparable to quantities contributed by the worldwide flow of rivers into the oceans. Hydrothermal venting also represents a major flow of heat from the Earth's crust and a major mechanism for cooling of new oceanic lithosphere. *See* LITHOSPHERE; SEAWATER.

Mineral deposits. Mineral deposits form at hydrothermal vents as the result of mixing between hot hydrothermal fluids and cold seawater, which leads to mineral precipitation. These deposits provide an opportunity to analyze a major ore-forming process in progress, although the exact relationship between hydrothermal-vent deposits and ore deposits on land is not always clear. The hydrothermal deposits consist largely of iron, copper, and zinc sulfides; calcium and barium sulfates; and silica. Two major structural types of mineral deposits, chimneys and mounds, are found at high-temperature vents.

Chimneys are roughly tubular, vertical structures with clearly defined central channelways through which hydrothermal plumes exit (Fig. 2). They consist of concentric layers of minerals, usually with a monomineralic channel lining of zinc sulfide [wurtzite, (Zn,Fe)S], copper-iron sulfide [chalcopyrite ($CuFeS_2$) or cubanite ($CuFe_2S_3$)] and outer layers dominated by iron sulfides [pyrite (FeS_2) and pyrrhotite (FeS_{1-x})], zinc sulfides [wurtzite or sphalerite, both (Zn,Fe)S], and anhydrite [$CaSO_4$].

Mounds vary from low-lying features about 16 ft (5 m) high and about 50 ft (15 m) in diameter to massive edifices up to 56 ft (17 m) high and 115 ft (35 m) in diameter. The larger edifices often have horizontally oriented outgrowths called flanges, with pools of high-temperature fluid trapped underneath. The flanges consist of iron sulfides [pyrrhotite, pyrite, and marcasite (also FeS_2)], oxyhydroxides, zinc sulfides, rare copper-iron sulfides, anhydrite, barite ($BaSO_4$), and silica. The flanges grow from deposition within the underlying pools and from the overflow of hot fluid from the pools. The large edifices themselves appear to have formed from the coalescence of flanges cemented by late-depositing silica.

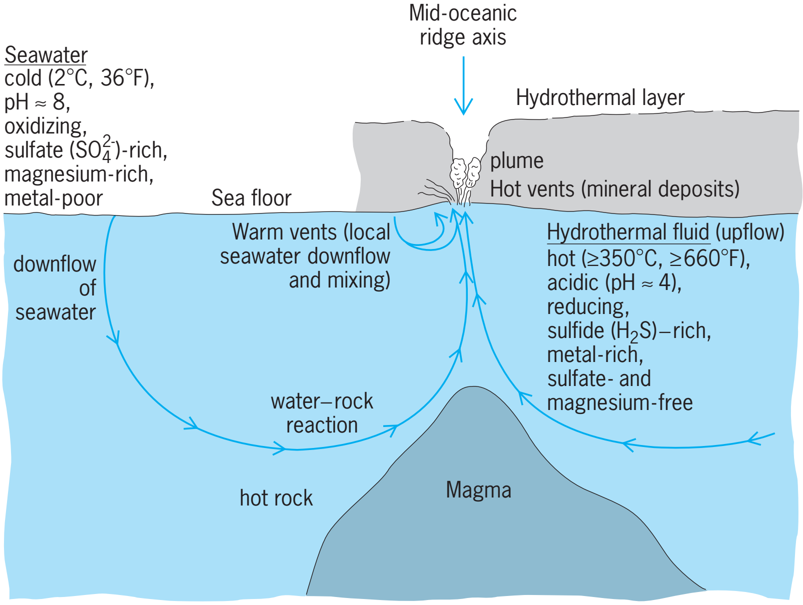


Fig. 1. Schematic view of the seawater circulation system creating hydrothermal vents along a mid-oceanic ridge.

Warm springs rarely show much mineralization at the sea floor, but minerals may be deposited below the surface where the initially hot fluids mix with seawater. *See* ORE AND MINERAL DEPOSITS.

Vent biology. Perhaps the most striking feature of sea-floor hydrothermal vents is their dense biologic communities, easily detectable in sea-floor photographs as indicators of the presence of even merely warm vent fluid. Biologists have compared the discovery of the vents to landing on a new planet.

About 450 animal species have now been identified at hydrothermal vents. About 95% of these have been species new to science, and about 75%

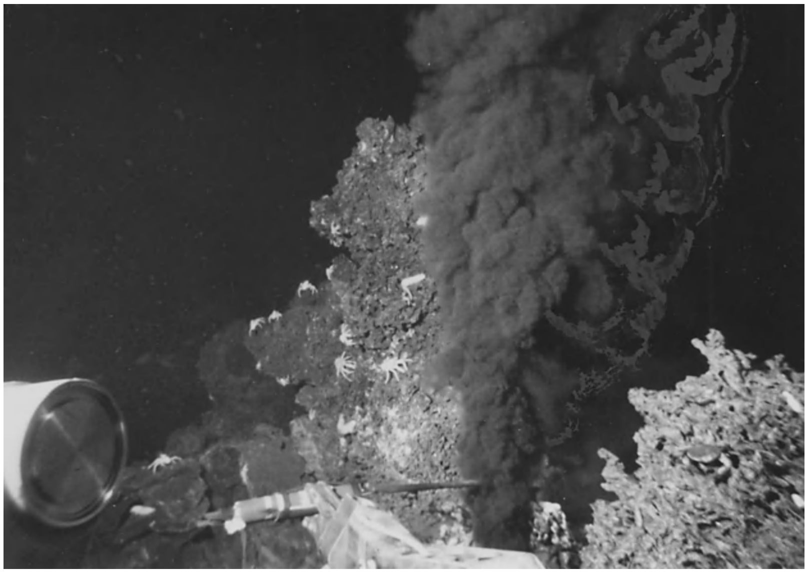


Fig. 2. Black smoker plume at the hydrothermal vent site at 21°N on the East Pacific Rise, off the tip of Baja California. Behind the plume is another hydrothermal chimney; visible in the foreground is the mechanical arm of the submersible *ALVIN* holding a temperature probe into the plume. Chimneys at this site reach a maximum of about 16 ft (5 m); the section in the background is about 6 ft (2 m) high.

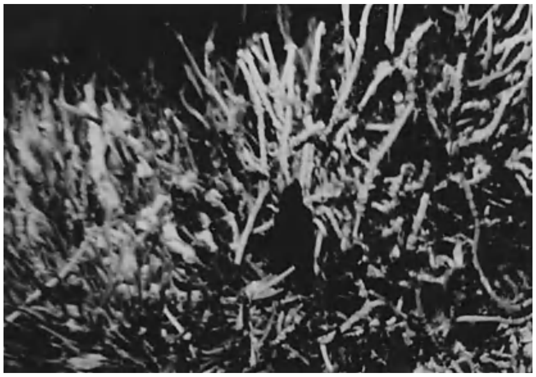


Fig. 3. Giant *Riftia* tubeworms from warm vents at the Galápagos Rift vent site. These tubeworms reach lengths of at least 6 ft (2 m).

are known from only a single vent site. Vent faunas tend to be dominated by mollusks, annelids, and crustaceans, whereas faunas on nonvent hard-bottom habitats consist predominantly of cnidarians, sponges, and echinoderms. Large mussels or clams are found in nearly all vent systems, as are a diverse assortment of tiny limpets and other gastropods. Gutless vestimentiferan annelids (tube worms), including the enormous *Riftia pachyptila* (**Fig. 3**) with its brilliant red plume and white tube up to 2 m (6 ft) in length, are characteristic of Eastern Pacific vents, but are completely absent from vents on the Mid-Atlantic Ridge. The latter sites are dominated by several species of caridean shrimp which swarm in massive numbers near areas of active venting.

Biologically, vents are among the most productive ecosystems on Earth—an interesting paradox, since they occur in the deep ocean where food is normally very scarce. Sulfide from hydrothermal fluids provides the energy to drive these productive systems. Whereas most animal life depends on food of photosynthetic origin (inorganic carbon converted to useful sugars by plants using energy from the Sun), the animals at hydrothermal vents obtain most or all of their food by a process of chemosynthesis. Chemosynthesis is accomplished by specialized bacteria residing in hydrothermal fluids, in mats on the sea floor, or in symbiotic relationships with other organisms. The bacteria convert inorganic carbon to sugars by mediating the oxidation of hydrogen sulfide, thereby exploiting the energy stored in chemical bonds. A few vent animals are also known to use methane gas as a source of energy and carbon.

The distribution of organisms at individual vents and at different vent sites seems to be closely related to the details of the distribution of temperature and fluid composition around the vents. *See* DEEP-SEA FAUNA.

The physical and chemical conditions at hydrothermal vents would be lethal to most marine animals, but vent species have adapted to the conditions there. For example, small pompeii worms (*Alvinella pompeiana*) occupy tubes in the sides of chimneys where water temperatures are very high; these worms may experience extreme thermal gradients between their anterior and posterior ends. Vestimentiferan tube worms absorb toxic sulfide into their bloodstreams for use by symbiotic bacteria, which reside in a large internal organ known as the trophosome. The hemoglobin in their blood binds the sulfide, thereby protecting the tissues of the worm, and permitting maximum delivery of sulfide to the symbionts. One of the most remarkable adaptations is found in *Rimicaris exoculata*, a very abundant shrimp at deep vent sites on the Mid-Atlantic Ridge. This shrimp has an eyelike structure on the back of its head capable of detecting the infrared radiation that is emitted from the hot water vents. This is apparently a mechanism for locating black smokers, where there is sufficient sulfide for the chemosynthetic bacteria upon which the shrimp depends for food.

Because vent fields tend to be dynamic and ephemeral, and the life of an individual vent is measured in years to decades, scientists have long wondered how new vents are colonized and how sulfide-dependent organisms survive local extinction when old vents shut down. Most vent species produce planktonic larvae capable of dispersing for long distances without the benefit of food. Eggs of many vent species are buoyant, causing them to move rapidly upward, where they can disperse with the ocean currents. The larvae of many vent species have been captured in the buoyant hydrothermal plumes far above vent fields in the Eastern Pacific, and the larvae of the mid-Atlantic vent shrimps have been taken in plankton samples hundreds of meters above the bottom. In those systems that have been studied, the prevailing currents run along the axes of the mid-ocean ridges, a fortuitous pattern that probably increases the likelihood that larvae will locate another vent. It has also been suggested that vent species may use whale carcasses and other sulfide-rich deep-sea habitats as “stepping stones” between distant vent sites. Some scientists speculate whether life itself could have originated in environments similar to those at hydrothermal vents. Many ubiquitous biochemical mechanisms involving the use of essential metals, heat-shock proteins, and certain enzymes appear to have originated in hydrothermal systems. It has even been suggested that the biochemical pathways of photosynthesis may have begun as part of an infrared heat-detection system such as that used by vent shrimp. *See* PREBIOTIC ORGANIC SYNTHESIS.

In a remarkable discovery, it was shown that chemosynthetic microbes known as Archaea are flushed from cavities deep within the Earth’s crust by hydrothermal and volcanic activity. These microbes are hyperthermophilic (hot-water-loving) and thrive at temperatures exceeding 90°C (194°F). It is now suspected that an entire community of such microbes inhabits the rocks deep within the water-saturated portions of the Earth’s crust. This discovery has led to the speculation that similar microbial communities could exist on Europa, a moon of Jupiter that may have a liquid ocean with deep-sea volcanism similar to that found on Earth.

The mineral deposits at hydrothermal vents have great commercial potential, and plans to mine vent deposits are being formulated in some parts of the world. There is increasing concern, however, that these systems are not well enough known for exploitation, and many scientists are therefore opposed to the commercial mining of vents. The genetic heritage of vent organisms may be an equally valuable resource; indeed, unique biochemical features of vent organisms are already having an impact in biotechnology. A heat-tolerant enzyme discovered in hydrothermal vent microbes is now an essential component of the polymerase chain reaction (PCR), a recent technological breakthrough that permits deoxyribonucleic acid (DNA) sequencing and fingerprinting from minuscule biological samples. This vent-dependent technology has become critically important in fields as disparate as medicine, criminology, and archeology. Marjorie Goldfarb; Craig M. Young

Bibliography. T. J. Barrett and J. L. Jambor (eds.), Seafloor hydrothermal mineralization, *Can. Mineralog.*, 26:429–888, September 1988; M. L. Jones (ed.), Hydrothermal Vents of the Eastern Pacific: An Overview, *Bull. Biol. Soc. Wash.*, no. 6, 1985; V. Robigou et al., Large massive sulfide deposits in a newly discovered active hydrothermal system, the High-Rise Field, Endeavour Segment, Juan de Fuca Ridge, *Geophys. Res. Lett.*, 20(17):1887–1890, September 3, 1993; A. P. Rona et al. (eds.), *Hydrothermal Processes at Seafloor Spreading Centers*, NATO Conference Series IV: Marine Sciences, vol. 12, 1983; The Southern Juan de Fuca Ridge: Hydrothermal fluids, sulfides, and geophysical studies, *J. Geophys. Res.*, 92:11,281–11,433, 1987.

Hydroxide

A compound containing the hydroxide ion (OH^-) and having the general formula $M(OH)_n$, where M represents a metal. Hydroxides are a subset of compounds containing the hydroxyl group (—OH) and range in chemical character from strongly basic, to amphoteric (having both acidic and basic characteristics), to essentially acidic. The hydroxide ion has a closed-shell electronic structure with a singlet ground state ($^1\Sigma^+$).

In the Lewis acid-base scheme, where a base is defined as an electron pair donor and an acid as an electron pair acceptor, a typical hydroxide decreases in base strength as attraction for electrons of the cation increases. For example, the hydroxides of electropositive elements such as the alkali metals and alkaline earths tend to be bases. When these ionic compounds are dissolved in water, they form metal ions and hydroxide ions, as inreaction (1), where

$$M^+(OH^-)_n(S) + nH_2O \longrightarrow M^{n+}(aq) + nOH^-(aq) \qquad (1)$$

s represents a solid and aq represents a water solution. The hydroxides of nonmetals (X) such as boron hydroxide [$B(OH)_3$], where the X-O bonds are covalent, are generally acidic, as shown in reaction (2),

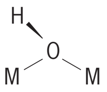
$$X(OH)_n + nH_2O \longrightarrow XO_n^{n-}(aq) + nH_3O^+(aq) \qquad (2)$$

where H_3O^+ is the hydronium ion. The amphoteric hydroxides may dissociate by either mechanism, depending on the presence of strong acids or bases. *See* ACID AND BASE; HYDROGEN ION.

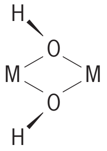
The alkali metal hydroxides such as sodium hydroxide (NaOH) are extremely important as reagents in metallurgy and photography and in the manufacture of soaps and detergents. Calcium hydroxide [$Ca(OH)_2$], known as slaked lime, is used in the preparation of mortar for brick laying. Minerals such as brucite [$Mg(OH)_2$] and pyrochroite [$Mn(OH)_2$] are naturally occurring hydroxides.

The lanthanides (Ln) also form hydroxides [$Ln(OH)_3$]; they are not simply hydrated oxides. They are prepared by precipitation of Ln(III) salts in aqueous NaOH and have hexagonal structures with tricapped prismatic coordination (a classification of ligand geometry around the lanthanide). Lanthanides exhibit the lanthanide contraction (that is, a decrease in atomic and ionic size with a corresponding increase in atomic number). Therefore, the hydroxides of lanthanides exhibit a reduction in basic character from lanthanum hydroxide [$La(OH)_3$] to lutetium hydroxide [$Lu(OH)_3$]. *See* HYDRATE; LANTHANIDE CONTRACTION.

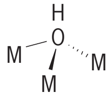
The hydroxide ion may behave as a ligand in transition-metal complexes; examples are the hydroxides of cobalt {$[Co(OH)_4]^{2-}$}, nickel [$Ni(OH)_2$], and zinc [$Zn(OH)_2$]. Ferrous hydroxide [$Fe(OH)_2$] is oxidized to red-brown ferrous oxide (Fe_2O_3), commonly known as rust. In addition, hydroxo bridges are well documented; single (**1**), double (**2**), or even triply (**3**) bridged structures are known. The double-bridged species (**2**) are the most common. *See* LIGAND. Thomas J. Meade



(**1**)



(**2**)



(**3**)

Hydroxyl

A chemical group in which oxygen and hydrogen are bonded and act as a single entity. In inorganic chemistry the hydroxyl group is known as the hydroxide ion (OH^-), and it is frequently bonded to metal cations, for example, sodium hydroxide (NaOH). In organic chemistry it frequently acts as a functional group, for example, in an alcohol (ROH, where R represents an alkyl group). *See* ACID AND BASE.

Oxygen toxicity. Many of the intermediate redox forms of dioxygen are toxic and damage important biomolecules. Much of this toxicity is thought to involve the generation and reactivity of hydroxyl (·OH), which is sometimes called the hydroxy radical. Radicals are chemical species with unpaired electrons (a dot represents an unpaired electron). Radicals can be formed via the addition of a single electron to a molecule, so superoxide, which is generated by the addition of a single electron to dioxygen, is also a radical and is sometimes written as $O_2^{\cdot\,-}$. Radicals can also be generated by removal of a single electron or by breaking in half the bonds that are formed by sharing two electrons. *See* FREE RADICAL; OXYGEN TOXICITY.

Production. The most common means for producing hydroxyl is the reaction of a reducing agent with hydrogen peroxide (H_2O_2). Other biologically available radicals, such as nitric oxide or semiquinones, may act as the reducing agent, but it is believed that transition-metal ions, such as ferrous ion (Fe^{2+}), are the most common reducing agents for generating $\cdot OH$, according to the Fenton reaction (1). Most of

$$Fe^{2+} + H_2O_2 \longrightarrow Fe^{3+} + OH^- + \cdot OH \qquad (1)$$

the free iron available in biological systems exists as Fe^{3+}, but the superoxide anion is capable of reducing Fe^{3+} to Fe^{2+} [reaction (2)]. The net reaction is

$$O_2^{\cdot -} + Fe^{3+} \longrightarrow O_2 + Fe^{2+} \qquad (2)$$

therefore the reaction of O_2^- with H_2O_2, which is called the Haber-Weiss reaction (3), and is catalyzed

$$O_2^{\cdot -} + H_2O_2 \longrightarrow O_2 + \cdot OH + OH^- \qquad (3)$$

by redox-active metal ions such as Fe^{3+}. Since many transition metals exist in oxidation states that differ by a single electron, many reactions that form radicals use transition metals as an electron source or sink.

Homolytic bond cleavages are reactions in which the two electrons from a bond are divided equally between the two atoms in the bond, so homolytic cleavage always produces two radicals. Electrons are more stable when shared as a pair by two atoms; thus, homolytic cleavage requires the input of energy, often in the form of radiation. Another common means for generation of hydroxyl comes from the action of ionizing radiation on water, which leads to the homolytic cleavage of one of the O-H bonds in water to produce hydroxyl and a free hydrogen atom [reaction (4)]. The action of ultraviolet light on hydrogen peroxide also produces $\cdot OH$ via homolytic bond cleavage [reaction (5)]. The O-H bond in water

$$HO{-}H \xrightarrow[\text{radiation}]{\text{ionizing}} H\cdot + \cdot OH \qquad (4)$$

$$HO{-}OH \xrightarrow[\text{light}]{\text{ultraviolet}} HO\cdot + \cdot OH \qquad (5)$$

is stronger than the O-O bond in hydrogen peroxide, so radiation of higher energy is required for homolytic cleavage of water.

Chemical behavior. Once generated, hydroxyl is a potent one-electron oxidant that forms the very stable OH^- ion [reaction (6)], and it abstracts hydrogen atoms from organic molecules that contain C-H bonds to form the stronger O-H bond in water [reaction (7), where R represents an organic

$$\cdot OH + e^- \longrightarrow OH^- \qquad (6)$$

$$\cdot OH + R{-}H \longrightarrow R\cdot + H_2O \qquad (7)$$

molecule that contains C-H bonds and R· is an organic radical with an unpaired electron on a carbon atom]. This reaction demonstrates an important property of radicals: the reaction of a radical, which contains an uneven number of electrons, with a molecule, which contains an even number of electrons paired in bonds, must generate a radical, because the number of electrons cannot change during the reaction. Thus, most reactions of radicals generate new radicals in processes called radical chain reactions. Reactions of radicals with molecules will continue to produce new radicals until other odd-electron species (such as transition-metal ions or other radicals) react with the radicals to produce even-electron molecules via termination reactions.

In biological systems, reaction of hydroxyl with lipids, which are organic molecules that make up cell membranes, leads to chain reactions involving dioxygen in a process called lipid peroxidation. This process is very destructive to cells, and it is a primary pathway of hydroxyl (and hence dioxygen) toxicity. Other deleterious effects of hydroxyl involve damage to nucleic acids [deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)], which also contain many C-H bonds that react with hydroxyl to produce new radicals. *See* CATALYSIS; CHAIN REACTION (CHEMISTRY); DEOXYRIBONUCLEIC ACID (DNA); ENZYME; PROTEIN; RIBONUCLEIC ACID (RNA); TRANSITION ELEMENTS.

H. Holden Thorp

Bibliography. I. Bertini et al. (eds.), *Bioinorganic Chemistry*, 1994; S. J. Lippard and J. M. Berg, *Principles of Bioinorganic Chemistry*, 1994; M. Roberfroid and P. Buc Calderon, *Free Radicals and Oxidation Phenomena in Biological Systems*, 1994.

Hydrozoa

A class of the phylum Cnidaria which includes the fresh-water hydras, the marine hydroids, many of the smaller jellyfish, a few special corals, and the Portuguese man-of-war. The Hydrozoa may be divided into seven orders: the Hydroida, Milleporina, Stylasterina, Trachylina, Chondrophora, Siphonophora, and Spongiomorphida. See separate article on each order.

Hydrozoa differ from the Scyphozoa, which are mostly the large jellyfish, and from the Anthozoa, to which the sea anemones and most corals belong, in the following features: The digestive space is not divided by longitudinal partitions; it lacks nematocyst-bearing structures, and has no stomodeum. The medusa has a velum but lacks the highly specialized sense organs characteristic of the Scyphozoa. *See* ANTHOZOA; SCYPHOZOA.

The form of the body varies greatly among the hydrozoans. This diversity is due in part to the existence of two body types, the polyp and the medusa. A specimen may be a polyp, a medusa, a colony of polyps, or even a composite of the first two. Polyps are somewhat cylindrical, attached at one end, and have a mouth surrounded by tentacles at the free end. Medusae are free-swimming jellyfish with tentacles around the margin of the discoidal body.

In a representative life cycle, the fertilized egg develops into a swimming larva which soon attaches itself and transforms into a polyp. The polyp develops stolons (which fasten to substrates), stems, and other polyps to make up a colony of interconnected polyps. Medusae are produced by budding and liberated to feed, grow, and produce eggs and sperm.

Not all hydroids conform to the foregoing description. Variation among them is due largely to differences in the pattern of growth and the extent to which the medusa stage is developed. In *Hydra* and a few hydroids, the medusa stage is absent, and new polyps, produced by budding, separate from the parent. In these cases each polyp is solitary. Corallike forms are similar to hydroids with the addition of a calcareous skeleton. In some cases the polyp form is greatly reduced or absent and the medusa stage predominates as in the Trachylina. The greatest complexity occurs in the Siphonophora, whose bodies have several different types of both medusoid and polypoid components.

Most hydrozoans are carnivorous and capture animals which come in contact with their tentacles. The prey is immobilized by poison injected by stinging capsules, the nematocysts. Most animals of appropriate size can be captured, but small crustaceans are probably the most common food. *See* CNIDARIA; FEEDING MECHANISMS (INVERTEBRATE).

Sears Crowell

Bibliography. L. H. Hyman, *The Invertebrates*, vol. 1, 1940; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Hyena

An African carnivore represented by three species of the family Hyaenidae; this family also includes the aardwolf (*Proteles cristatus*). Hyena adults are over 4 ft (1.2 m) high and almost 5 ft (1.5 m) in length, and may weigh up to 165 lb (75 kg). While they resemble dogs superficially, they are more closely related to the felids (cats). All three species are adapted for feeding on carrion, having well-developed foreparts, reduced hindquarters, a rounded head, and short strong jaws. The 34 teeth (dental formula I 3/3 C 1/1 Pm 4/3 M 1/1) are extremely strong and together with the powerful jaw muscles can crush bone.

Hyenas are four-toed digitigrade (walking on the toes) animals with blunt, nonretractile claws that are used for digging and disinterring bodies. The gestation period is about 13 weeks, with two to four young comprising the yearly litter. The maximum life-span of this animal is about 25 years.

Spotted hyena (*Crocuta crocuta*). (*Photo by H. Vannoy Davis; © 2001 California Academy of Sciences*)

The spotted hyena (*Crocuta crocuta*; see **illus.**) ranges south of the Sahara, while the brown hyena (*Hyaena brunnea*) is restricted to southern Africa. The striped hyena (*H. hyaena*) is found in northeastern Africa and down into southern Asia. The spotted hyena is more active diurnally than the striped hyena, which is nocturnal and solitary in its activities. *Crocuta crocuta* is interesting in that it is difficult to distinguish between the sexes, since the female closely resembles the male in the external appearance of the reproductive organs. *See* CARNIVORA.

Charles B. Curtin

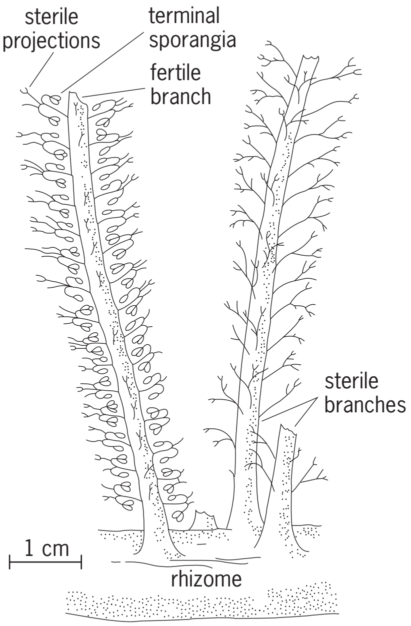
Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Hyeniales

An order of Devonian plants considered by some to be related to the spenophytes. The small dichotomously forked leaves tend to be borne in whorls. Some leaves bear terminal sporangia, but these appendages are neither aggregated into a tight cone nor separated by bracts.

The Middle Devonian genus *Hyenia* (see **illus.**) has aerial axes borne on stout rhizomes. Its leaves are two to four times dichotomized (branched repeatedly into two branches) and their arrangement approaches the whorled condition. Fertile leaves are characterized by recurved segments which bear paired, terminal sporangia, and often by erect, sterile segments as well. The anatomy of *Hyenia* is essentially unknown. *Protohyenia,* believed to be a precursor of *Hyenia*, has since proved to be a *Pseudosporochnus*.

Calamophyton, also from the Middle Devonian, consists of a stout axis that divides digitate-fashion into three or more lesser branches which may dichotomize. The axes frequently appear to be jointed, and whorled leaves are borne at the nodes. The leaves dichotomize two to four times in three dimensions. The best-known species, *C. bicephalum*, bears whorls of fertile leaves, each of whose two main subdivisions bear three recurved appendages. Each appendage bears a pair of terminal sporangia. Sporangia are fusiform and dehisce longitudinally along one edge. Erect, sterile portions of the fertile



***Hyenia*, showing stout rhizome with fertile and sterile branches.**

appendages project beyond the position of the sporangia. Spores assignable to the genus *Dibolisporites gibberosus* have been found in some sporangia. They range from 86 to 166 micrometers in diameter and are ornamented by coni and spinae up to 4.5 micrometers long. The vascular strand of *Calamophyton* has been regarded as siphonostelic (xylem surrounding a pith). The anatomy of *C. bicephalum* has been shown to consist of a number of separate strands of xylem, thus resembling *Pseudosporochnus*. *See* PALEOBOTANY; PLANT KINGDOM. Harlan P. Banks

Hygrometer

In the most general sense, a hygrometer is any device used to measure humidity in air or other gases. The more modern definition is that hygrometers are only devices that measure humidity via changes in material properties. Humidity may also be measured by other methods, chiefly psychrometry and dewpoint, both of which rely on temperature changes: due to evaporation for psychrometers, and due to condensation for dewpoint. In contrast, hygrometers are generally isothermal. Primitive hygrometers relied on mechanical changes, such as the slackening of hair with increasing humidity. However, the vast majority of modern hygrometers are based on changes in the electrical resistance or capacitance of a material with changes in humidity. Most hygrometers are designed and calibrated to indicate relative humidity, which is the percentage of water in air (or another gas) relative to the saturation condition. When air is saturated with water, further water content appears as mist or frost on exposed surfaces. Many commercial hygrometers are combined with thermometers, sometimes in the same electrical circuit, and these are known as thermohygrometers. *See* DEWPOINT; HUMIDISTAT; HUMIDITY; MOISTURE-CONTENT MEASUREMENT; PSYCHROMETER; PSYCHROMETRICS.

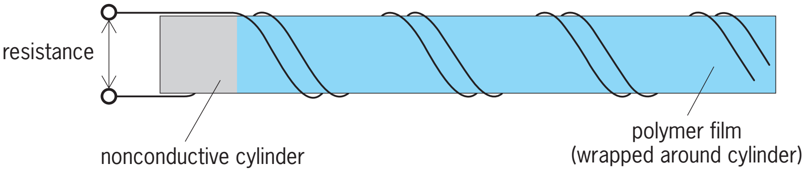


Fig. 1. Simplified resistive hygrometer circuit in which the electrical resistance of the polymer film depends on the ambient humidity. The electrical resistance of the polymer is measured across the electrical leads as shown. The ends of the two electrical leads are not joined but are connected to the film.

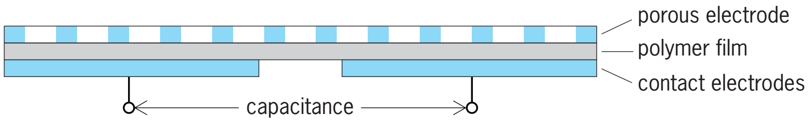


Fig. 2. Simplified capacitive hygrometer circuit in which the electrical capacitance of the contact electrodes across the polymer film depends on the ambient humidity.

There are three common classes of electronic hygrometer: resistive, capacitive, and piezoelectric. The first two types may also be combined to operate as an inductive circuit. Resistive and capacitive hygrometers are commonly based on changes in resistance or capacitance of a thin polymer film. Sulfonated polystyrene, created by treating polystyrene with sulfuric acid, is an example of a suitable resistive polymer. It is important that the polymer film be well exposed to air. The rate of response of a hygrometer is dependent on the ability of air to reach the sensor. Some hygrometers even incorporate fans to accelerate response time. *See* POLYSTYRENE RESIN.

Figures 1 and **2** illustrate common resistive and capacitive circuits incorporating a thin film. In both circuits, the film is open to air, and humidity can be measured due to changes in the resistance or capacitance of the film. In the capacitive circuit (Fig. 2), the film is shielded by the porous electrode, while in the resistive circuit (Fig. 1) a porous sheath would be required to protect the sensor from damage. Polymer films are preferred over many other materials because they can be chosen to respond to relative humidity. Other materials, including ceramics such as aluminum oxide, tend to respond to absolute humidity (the amount of absolute water content in air, rather than the amount relative to saturation). *See* CAPACITANCE MEASUREMENT; RESISTANCE MEASUREMENT.

Piezoelectric crystals such as quartz can also be used to measure absolute humidity. The mass of the oscillating crystal changes directly with the amount of water absorbed, which is typically proportional to absolute humidity. This effect, in turn, changes the frequency of oscillation, which can be measured extremely precisely. *See* PIEZOELECTRICITY.

Commercial thin-film hygrometers tend to be precise to within an error of ±2% for relative humidity. This accuracy can be refined by calibration against a reference humidity close to the temperature of interest. The simplest calibration references are saturated

Relative humidities for saturated salt soluitions

Salt	Relative humidity*, %
Lithium chloride	11.31
Potassium acetate	23.11
Potassium carbonate	43.16
Sodium chloride	75.47
Potassium sulfate	97.59

*Saturated solution, 20°C (68°F).

salt solutions. These are created by dissolving an excess of a given salt to form a slushy solution. The hygrometer is then sealed in a container with the solution. Relative humidities for saturated salt solutions are shown in the **table** for 20°C (68°F). The salts shown give humidities that are almost constant with temperature. For example, sodium chloride (table salt) produces relative humidities of 75.51% at 0°C (32°F), 75.47% at 20°C (68°F), 74.41% at 55°C (176°F), and 76.29% at 80°C (176°F). For careful calibration, the solution should be allowed to return to the reference temperature after dissolving the salt, as the heat of dissolution can change the temperature significantly. Brian J. Lowry

Bibliography. J. J. Carr, *Sensors and Circuits*, Prentice Hall, 1993; W. C. Dunn, *Fundamentals of Industrial Instrumentation and Process Control*, McGraw-Hill, 2005; H. N. Norton, *Hand-book of Transducers*, Prentice Hall, 1989.

Hymenomycetes

An artificial class of fungi in the phylum Basidiomycota. It was traditionally divided into two subclasses: Holobasidiomycetidae, delimited by nonseptate basidia and the absence of a yeast phase; and Phragmobasidiomycetidae, frequently with septate basidia and often forming a yeast phase. A typical hymenomycete produces a fruit body or basidiome with spore-bearing basidia organized in a membranelike layer called the hymenium. The fruit bodies are designed to allow for basidiospore discharge into air currents either directly off basidia or after falling from elaborate fertile surfaces. Falling spores, indicative of active discharge, accumulate in powdery masses called spore prints. The shape of the hymenium varies from lamellate (gilled as in mushrooms), poroid (as in conk or bracket fungi), toothed (in hedge hog fungi), coralloid (coral fungi), labyrinthoid (daedaleoid fungi), wrinkled (merulioid fungi), or smooth to diffuse (corticioid fungi). Exceptional hymenomycetes may be aquatic, lack a mycelial phase, or lack a fruit body.

Antibiotics have been isolated from many species. Commercially grown edible species include the button mushroom (*Agaricus bisporus*), Shiitake (*Lentinula edodes*), Paddy Straw mushroom (*Volvariella volvacea*), and Wood Ear (*Auricularia polytricha*). Wild harvested species include the Matsutake (*Tricholoma matsutake* and *T. magnivelare*), chanterelles (*Cantharellus cibarius* and allies), and the King Bolete (*Boletus edulis*).

Most genera are either saprophytic (for example, *Agaricus* and *Polyporus*), or mycorrhizal with trees (*Albatrellus, Cortinarius, Ramaria*, and *Thelephora*). Others are parasites. *Heterobasidion* causes destructive tree diseases; *Rhizoctonia* and *Typhula* (snow molds) cause field crop losses; *Mycena citricolor* blights coffee leaves; and *Exobasidium*, an obligate plant pathogen, induces the formation of galls and leaf curls. Other notable pathogens include *Hohenbuehelia* and *Pleurotus*, which capture nematodes; *Serpula*, a major dry-rot agent; *Dictyonema*, a basidiolichen; and *Septobasidium*, which harnesses living scale insects. *See* BASIDIOMYCOTA; EUMYCOTA; FUNGI; PLANT PATHOLOGY. Scott A. Redhead

Bibliography. D. S. Hibbett et al., *Proc. Natl. Acad. Sci.*, 94:12002-12006, 1997; K. K. Nakasone, *Cultural Studies and Identification of Wood-Inhabiting Corticiaceae and Selected Hymenomycetes from North America*, 1990; P. Stamets, *Growing Gourmet and Medicinal Mushrooms*, 1993; E. C. Swann and J. W. Taylor, *Mycologia*, 85:923–936, 1993.

Hymenoptera

The third largest order of insects, containing the sawflies, ants, wasps, bees, and related forms. There are some 15,700 described species, subspecies, and varieties from America north of Mexico. Conservative estimates suggest that the world fauna may comprise well over 100,000 described species of this order, with many thousands still to be described.

This order is of great importance to humans. Some members such as the sawflies, certain chalcidoids, and most cynipoids, feed during the larval stage on foliage or other plant tissues. Many species, such as the ichneumon flies, most chalcid flies, and wasps, are parasites or predators of other insects or spiders during their larval stage. Bees are indispensable in the pollination of many fruits, vegetables, and forage crops. *See* POLLINATION.

Hymenoptera occur in all major faunal zones but are more abundant and have greater diversity of species in the tropical and temperate zones. Representation in the northern parts of the boreal zone is limited to a few sawflies, some Parasitica, and very few Aculeata, of which the bumblebees are the most conspicuous representatives. *See* INSECTA.

Adult Hymenoptera usually may be recognized by having two pairs of membranous wings with reduced venation, the hind pair smaller than the front pair, and by mouthparts formed for biting and often for lapping or sucking. In the higher forms, the abdomen is constricted basally, its first segment fused with the hind part of the thorax. Females always have an ovipositor modified for sawing, piercing, or stinging. Metamorphosis is complete, and sawfly larvae resemble caterpillars except for the abdominal prolegs which lack a series of hooklets. Larvae of higher forms are legless and maggotlike but have a well-developed head. Pupae have the appendages free

TABLE 1. Families of Hymenoptera

Classification	Common name	No. of species
Suborder Symphyta	Sawflies	1009
Superfamily Megalodontoidea		120
Xyelidae		33
Pamphiliidae	Web-spinning sawflies	87
Superfamily Tenthredinoidea		849
Pergidae		13
Argidae		32
Cimbicidae	Cimbicid sawflies	12
Diprionidae	Conifer sawflies	35
Tenthredinidae	Sawflies	757
Superfamily Siricoidea		28
Syntexidae		1
Siricidae	Horntails	15
Xiphydriidae		6
Orussidae		6
Superfamily		12
Cephoidea		12
Cephidae	Stem sawflies	12
Suborder Apocrita		13,346
Superfamily Ichneumonoidea		3814
Stephanidae		7
Braconidae	Braconid wasps	1239
Ichneumonidae	Ichneumon flies	2568
Superfamily Chalcidoidea		2032
Mymaridae	Fairy flies	110
Trichogrammatidae	Minute egg parasites	39
Eulophidae		544
Elasmidae		17
Thysanidae		18
Eutrichosomatidae		2
Tanaostigmatidae		4
Encyrtidae		320
Eupelmidae		89
Eucharitidae		27
Perilampidae		31
Agaontidae	Fig insects	2
Torymidae	Torymids	181
Ormyridae		17
Pteromalidae		321
Eurytomidae	Seed and stem chalcids	203
Chalcididae	Chalcids	101
Leucospidae		6
Superfamily Cynipoidea	Gall wasps	877
Ibaliidae		6
Liopteridae		2
Figitidae		58
Cynipidae	Cynipids of gall wasps	811
Superfamily Proctotrupoidea		985
Evaniidae	Ensign flies	11
Gasteruptiidae		50
Pelecinidae	Pelecinid wasps	1
Vanhorniidae		1
Roproniidae		3
Heloridae		1
Proctotrupidae		54
Ceraphronidae		101
Diapriidae		304
Scelionidae	Scelionid wasps	272
Platygasteridae		182
Trigonalidae		5
Superfamily Bethyloidea		345
Chrysididae	Cuckoo wasps	124
Bethylidae		100
Sclerogibbidae		1
Dryinidae		120
Superfamily Scolioidea		643
Tiphiidae	Tiphiid wasps	185
Sierolomorphidae		2
Mutillidae	Velvet ants	409
Rhopalosomatidae		2
Scoliidae		26
Sapygidae		19
Superfamily Formicoidea		786
Formicidae	Ants	786
Superfamily Vespoidea		368
Vespidae	Hornets, yellow jackets, potter wasps	368
Superfamily Pompiloidea		279
Pompilidae	Spider wasps	279
Superfamily Sphecoidea	Fossorial wasps	1215
Ampulicidae		3
Sphecidae		1212
Superfamily Apoidea	Bees	3310
Colletidae	Colletid bees	149
Andrenidae	Andrenid bees	852
Halictidae	Halictid and sweat bees	472
Melittidae		31
Megachilidae	Leafcutting bees	730
Apidae	Honeybees, bumblebees, and carpenter bees	1076

from the body except for certain chalcidoids.

The first four superfamilies of the Apocrita—the Ichneumonoidea, Chalcidoidea, Cynipoidea, and Proctotrupoidea—are commonly called the Parasitica, and the remaining superfamilies are known as the Aculeata. The Aculeata are stinging forms and the Parasitica are parasites of other insects. It is impossible to demarcate these two groups sharply because some Aculeata are parasites and some Parasitica are phytophagous. However, except for the phytophagous species of Parasitica, these insects lay their eggs in or on an insect or spider host while the Aculeata place theirs in nests with a provision of food. Some connecting forms apparently occur in the Procotrupoidea and Bethyloidea. Some of the Parasitica are more highly specialized, morphologically, than any other Hymenoptera, but the Aculeata show the greatest specialization in behavior. **Table 1** presents the major classification of the order as recognized in North America. The number of described species, subspecies, and varieties in America north of Mexico is noted for each family, and common names are given.

Morphology

The adult hymenopteron has a clearly differentiated head, thorax, and abdomen. Wings, when present, and legs are attached to the thorax (**Fig. 1**).

Head. Typically the head is so oriented that mouthparts are directed downward; however, all variations occur, and in some species the mouthparts are directed forward. The large compound eyes occupy much of the sides of the head, though they are reduced in size in many ants and some Parasitica. Three ocelli are typically present on the top of the head, but they may be reduced or absent in wingless forms. The paired antennae arise from the face between the eyes, and they may be close to the mouthparts or removed from them. Primitively, the antennae are long, slender, and multisegmented, but in the more highly specialized members of all major groups there has been a tendency toward reduction in the number of segments.

The mouthparts consist of paired mandibles, and a labiomaxillary complex formed by membranous connections between the maxillae and labium (**Fig. 2**). Most sawflies have simple biting mouthparts, but in the higher forms the labiomaxillary complex is variously modified to permit lapping or sucking types of feeding. In many bees and a few wasps, the labiomaxillary complex is greatly elongated to permit these insects to suck nectar from flowers which have deeply seated nectaries. The mandibles, in forms other than sawflies, do not function primarily in feeding but are used in manipulation of parasites or prey, construction of nests, and to escape from the cocoon or host body.

Thorax. The thorax consists of three segments, tightly fused together. Each segment bears a pair of legs, and each of the last two segments bears a pair of wings. In the sawflies, the thorax is broadly joined to the abdomen, but in the Apocrita the true first abdominal segment, the propodeum, has fused firmly with the thorax and is separated from the remainder of the abdomen by a constriction.

Wings. Most species have two pairs of wings, of which the posterior pair is smaller. In flight, the fore- and hindwings are joined by a row of tiny hooks, the hamuli, along the fore margin of the posterior wing, which fit into the downfolded hind margin of the anterior wing. The most complete venation is shown by the sawflies, and extensive reduction of venation in varying degrees has occurred in all higher forms (**Fig. 3**). Flightless species with shortened, nonfunctional wings, or no wings at all occur in most major groups, except the sawflies and bees. The females usually exhibit a greater reduction than do the males.

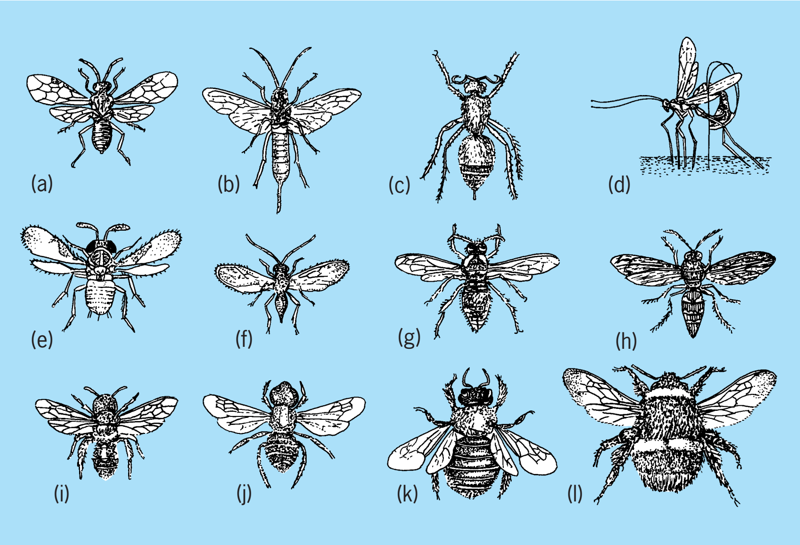


Fig. 1. **Representatives of 12 families of Hymenoptera. (*a*) Tenthredinidae. (*b*) Siricidae. (*c*) Mutillidae. (*d*) Ichneumonidae. (*e*) Chalcididae. (*f*) Cynipidae. (*g*) Vespidae. (*h*) Sphecidae. (*i*) Andrenidae. (*j*) Megachilidae. (*k*) Xylocopidae. (*l*) Bombidae. (*After T. I. Storer and R. L. Usinger, General Zoology, 3d ed., McGraw-Hill, 1957*)**

Legs. Each leg consists of a coxa which articulates with the corresponding thoracic segment, a trochanter, femur, tibia, and segmented tarsus. The tarsus usually has five segments, but these may be reduced to four or three in many Chalcidoidea and males of *Orussus* (Orussidae). The last tarsal segment typically bears a pair of claws. In most Parasitica, there is a constriction near the base of the femur which causes the trochanter to appear two-segmented.

The legs are frequently modified to serve varied specialized uses. Many species which dig through the soil, like the Scolioidea, have thickened femora and tibiae bearing numerous stout spines; while other digging wasps have a comb of bristles on each fore tarsus which aids in raking the soil out of the burrow. Some bees, such as *Andrea*, have groups of curled, branched hairs on the hindlegs to collect

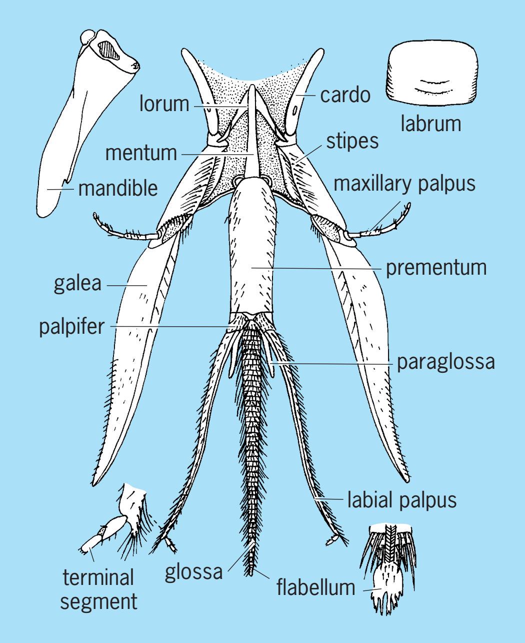


Fig. 2. **Mouthparts of *Anthophora edwardsii*. (*After E. O. Essig, College Entomology, MacMillan, 1942*)**

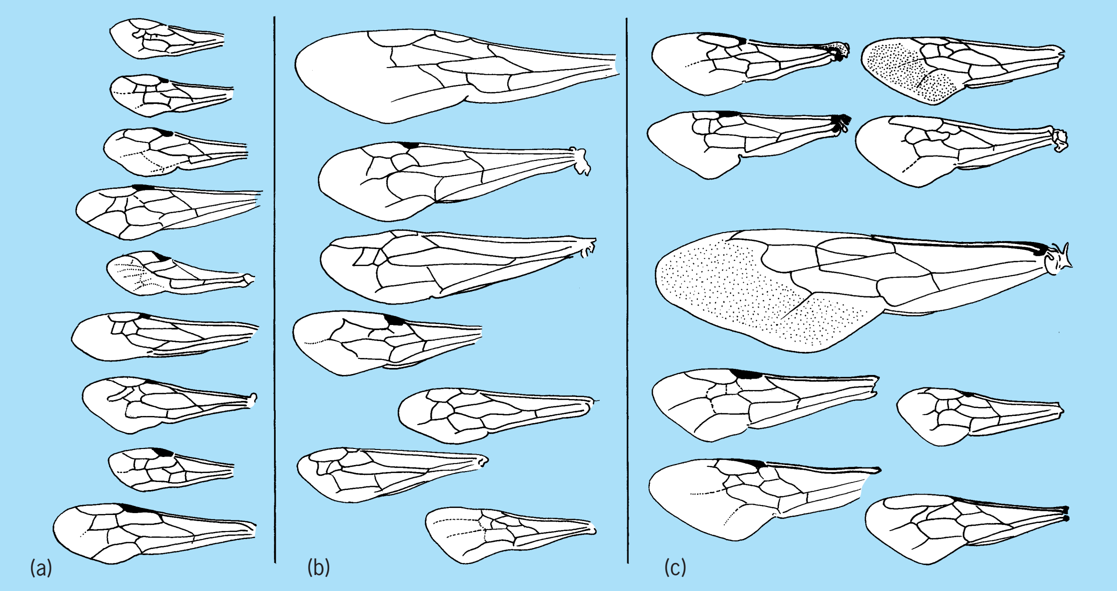


Fig. 3. Wings of Hymenoptera. Most species have two pairs, of which the posterior pair is smaller. (*a*) Sphecoid wasps. (*b*) Vespoid wasps. (*c*) Bees. (*After E. O. Essig, College Entomology, Macmillan, 1942*)

pollen, while others, like the honeybee and bumblebees, have the hind tibia expanded, flattened, and fringed by long curved hairs to form a pollen basket, or corbicula. Females of most Dryinidae have the fore tarsus modified into a chela to grasp the prey.

Abdomen. The abdomen primitively consists of 10 segments, though the number appears to be less because of modification or loss in the higher forms. In the sawflies, 10 terga and 9 sterna may be recognized, the apical parts being modified in connection with the male genitalia or female ovipositor. In the Apocrita, the first abdominal segment has fused with the thorax, and 6 other segments are normally visible in the female, 7 in the male. An exception to this is the Chalcidoidea, where there are 7 terga. Frequently there are fewer visible segments, either because of fusion, as in some Braconidae and Proctotrupidae, or because of retraction of the posterior 2 or 3 segments, as in the Chrysididae.

The female ovipositor, or sting, is formed from processes of the eighth and ninth sterna (**Fig. 4**). In the sawflies, parts of the ovipositor have ridges which terminate ventrally in the teeth of the saw. Vestiges of such ridges in the Apocrita constitute the barbs, which, if well-developed as in the honeybee, cause the sting to remain in the wound. The eggs of both Symphyta and Parasitica pass through the ovipositor during oviposition. In the Aculeata, the ovipositor is purely a stinging mechanism, and the egg issues from an opening at its base. The sting is reduced or lacking in some ants and in the stingless honeybees of the tropics.

Venom. In the Apocrita there is a pair of acid glands opening into a poison sac connected with the ovipositor. The secretion of these glands produces either a temporary paralysis when injected into their hosts by some Parasitica, or, usually, permanent paralysis when injected into their prey by aculeate wasps. Bees use their stings purely for defense. The venom of the Aculeata is a complex substance consisting of a protein and certain enzymes, as well as other constituents. Apparently the composition varies slightly with each species, which complicates the preparation of a desensitizing agent. When a human is stung, the enzymes react with his tissues to release histamine. Death may occasionally result from anaphylactic shock, or from mechanical suffocation due to swelling of the lymphatic system. Medical assistance should be sought if severe swelling occurs following a sting, especially one on the face or throat.

Biology

Practically all hymenopterous adults are terrestrial forms, living in, on, or near the Earth's surface. A few species are secondarily aquatic, the adults swimming or walking under water to search out and parasitize aquatic or subaquatic hosts.

Most adults feed on plant nectar or honeydew secretions of various insects. A few sawflies prey on other insects. Some species of Parasitica and Aculeata imbibe body juices of the host or prey which they attack primarily for oviposition. Not too much is known of the nutritional requirements of newly emerged adults, but it is likely that many of them require some nutrient materials in order for the eggs to mature.

Reproduction. Mating takes place in a variety of situations, but it is always of rather short duration. Customarily, the males emerge from one to several days earlier than the females. Among the aculeates, the males of ground-nesting species may indulge in prenuptial flights over the nesting site while awaiting the emergence of females. Usually they pounce on the females and mating takes place on the ground. Males of some wood-nesting species may hover in front of the burrows harboring the females, and in this instance mating usually takes place during flight. Other species meet and mate on flowers. Among the ants and social wasps, large numbers of reproductive females and males emerge simultaneously from several colonies and form mating swarms in the air. The honeybee has a similar nuptial flight, but it is composed of only one virgin queen and a number of drones.

Most species are represented by both males and females. Males are usually produced from unfertilized eggs and have half the normal number of chromosomes, while females are produced from fertilized eggs and have the normal number of chromosomes. However, both facultative and obligate parthenogenesis are more common than in any other order of insects. In those species in which facultative parthenogenesis occurs, apparently only males are produced from virgin females. In those species exhibiting obligate parthenogenesis, females are produced from unfertilized eggs. Certain Cynipoidea have alternating generations, in which one kind of female produces both sexes parthenogenetically, and the fertilized females of that generation produce only females.

In addition to males and fertile females found in most Hymenoptera, the social species have a third form called the worker caste. These are females which are sterile except under unusual circumstances. The workers of social wasps and bees are normally very similar in appearance to the queens, differing principally in their smaller size and occasionally in color pattern. The honeybee queen is quite different from the workers because she lacks wax glands and pollen-collecting apparatus, and she has a smooth rather than a barbed sting. Worker ants ordinarily differ more from their queens than do other species of social Hymenoptera—wings are absent, and the thorax is more modified. In some ants, the workers may be dimorphic, some of them being modified into a soldier form with huge head and mandibles for defense of the colony.

Life history. Hymenoptera exhibit complete metamorphosis during development and pass through an egg, larval, and pupal state (**Fig. 5**). So far as is known, Hymenoptera always lay eggs. These are deposited in a protected situation on or near the supply of larval food. The hymenopterous egg is usually ovoid or sausage-shaped, and many species in some groups, like the Cynipoidea and some of the other Parasitica, have stalked eggs. The surface is usually unsculptured and very delicate. The eggs hatch after an incubation of variable length, and the resulting larvae begin a growth period consisting of three or four instars.

The larvae of sawflies are very similar in gross appearance to lepidopterous caterpillars. There is a well-developed head bearing powerful biting mouthparts. There are three pairs of segmented thoracic legs and, in all but the forms that bore in stems or wood, there are from six to eight pairs of unsegmented abdominal prolegs which do not bear tiny hooks, or crochets, as in the Lepidoptera.

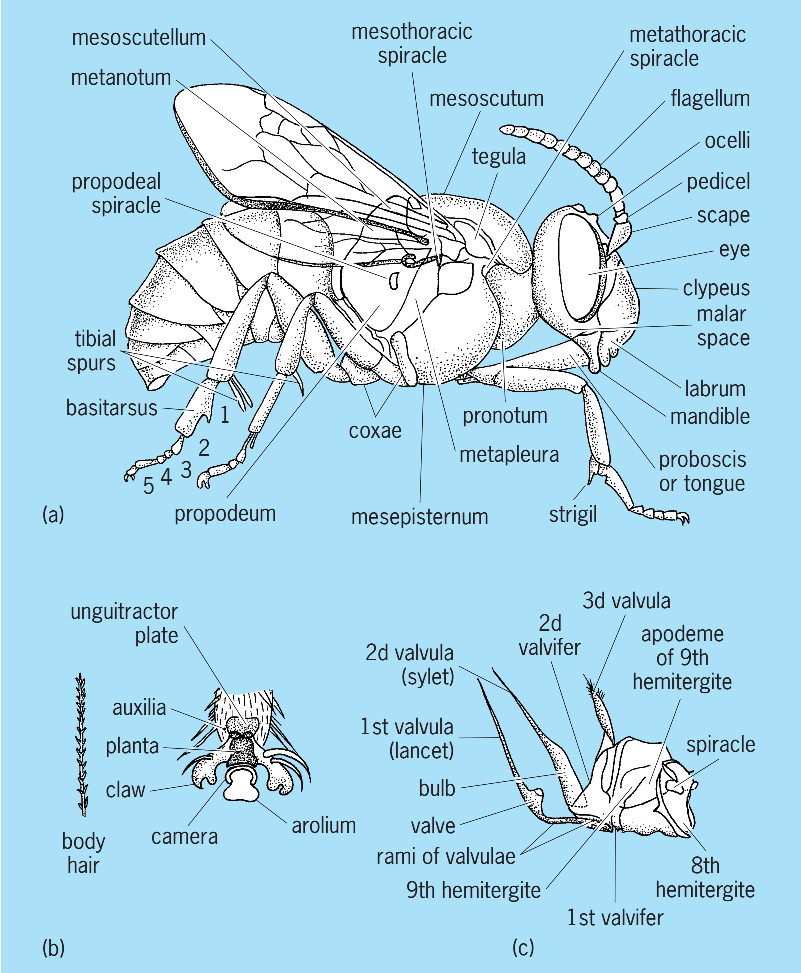


Fig. 4. The bee *Anthophora edwardsii*. (*a*) Adult male. (*b*) Sting. (*c*) Accessory organs of the female showing the important characters used in classifying bees. (*After E. O. Essig, College Entomology, Macmillan, 1942*)

The larvae of Apocrita are legless, maggotlike forms in the later instars with a well-developed head and biting mouthparts. Occasionally, fleshy processes are present which aid in locomotion. The first instar larvae of some Apocrita, particularly those of the Parasitica and of the social parasites among the Aculeata, are very different from the larvae of later instars, a condition known as hypermetamorphosis.

The insect then spins a cocoon, if characteristic of the particular species, and enters the prepupal stage which may be of short or lengthy duration. The pupal stage usually lasts only two or three weeks, and the adult then ecloses, or emerges from the pupal case. The hymenopterous pupa has the appendages free from the body, except in certain Chalcidoidea. It is usually enclosed in a cocoon either constructed of silk alone or with other interwoven substances. The cocoon may be well developed, or it may be just a thin silken lining of the larval cell. It is absent in almost all Chalcidoidea and Cynipoidea, and, occasionally, in all other superfamilies. The adult usually remains in the cocoon for one to several days until the integument is thoroughly hardened,

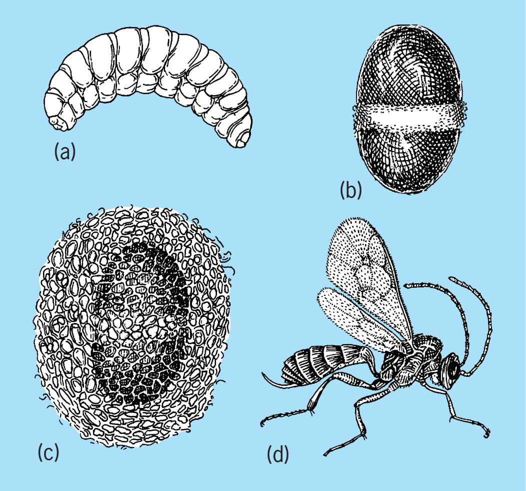


Fig. 5. Life cycle of *Bathyplectes curculionis*, a parasite of alfalfa weevil. (*a*) Larva. (*b*) Pupa. (*c*) Pupa within cocoon of weevil. (*d*) Adult female. (*After E. O. Essig, College Entomology, Macmillan, 1942*)

though occasionally the adult, as in some species of *Osmia* in the Megachilidae, may remain in the cocoon for 10 months before beginning its active adult existence.

Some species have only one generation a year in temperate zones, others have two, and many breed continually during the warmer months. Hymenoptera usually overwinter as prepupae but occasionally ants, social wasps, and some bees overwinter as adults or as larvae, as in some Parasitica. So far as is known, none passes the winter in the egg or pupal stage.

Dimorphism. Sexual dimorphism is often very marked. The two sexes of some species are so dissimilar that earlier students placed them in different genera or families. Even today there are many puzzles, and sexes have not been associated for many of the species having wingless females and winged males. Ordinarily the males are somewhat smaller than females, though the reverse is true in most species having wingless females. *See* SEXUAL DIMORPHISM.

In addition to sexual dimorphism, there are other kinds of variations in adult Hymenoptera. Species of Parasitica or of the social parasites among the Aculeata sometimes vary a great deal in size, depending on the size of the host or on the number of parasites per host. Also, there may be variation in form and size between the generations of species having more than one generation a year.

Behavior. Behavior of the Parasitica is much less complex than in the Aculeata. The primitive pattern is for the female to hunt for her host and then oviposit in or on it, without causing it any great disturbance. The host continues to live a normal life until it is killed by the developing parasite larva. This pattern is typical of many Ichneumonidae. A more complicated pattern is developed by stinging of the host to produce a temporary or permanent paralysis, or even death, before oviposition. This method is typical of many Braconidae and Chalcidoidea. Some species of Parasitica are hyperparasites or secondary parasites; that is, they are parasitic on parasites of the host. Some species may be either primary or secondary parasites, while tertiary parasitism, that is, species parasitic on secondary parasites, also has been reported in a few chalcidoids.

Nests and nesting. The higher Aculeata are remarkable for the amount of care taken in construction of nests for their progeny. The nest may consist of a burrow in the ground or a boring in wood. The nest may also be constructed of clay, paper formed from masticated wood fibers, gums, resins, masticated leaf pulp, pieces of leaves, or wax made in the body of the insect. There may be only one cell per nest, a series of a dozen or so cells arranged in a clump or in a linear series, or there may be a multicellular nest, as in the social wasps and bees.

There are several basic types of nesting behavior among the aculeates. The most primitive pattern is for the wasp to search out its prey, paralyze it by stinging, and then to lay one or several eggs on it, leaving the prey where it was found. This is typical of most Bethylidae and many Scolioidea. A somewhat more complicated pattern, typical of most Pompilidae, is for the wasp to capture and paralyze its prey before constructing a nest for it in a sheltered situation. Another pattern is for the cell to be constructed and an egg deposited in it before the first specimen of prey is brought in. This occurs typically among the Vespidae and a few of the Sphecidae which practice progressive provisioning. A fourth pattern is a variation of the third and consists of preparation of the cell, storing the cell with prey or a pollen-nectar mixture, and then oviposition. This method is typical of most Sphecidae and the Apoidea.

The solitary wasps store varied amounts of prey per cell. The Scolioidea, Pompilidae, and a few Sphecidae use only one specimen of prey per cell. Others, like the Vespidae and most Sphecidae, store from two to more than two dozen specimens of prey per cell depending upon the size of the prey used. Some solitary wasps are quite restrictive in their prey preferences and store only one species. The majority are not so selective and prey upon different species. However, most of these latter prey on species belonging to a single family or order, or to species occurring in a very restricted ecological niche. Thus, in the Sphecidae, species of *Pemphredon* prey on a number of species of aphids (Aphidae), and species of *Ectemnius* on a wide variety of Diptera, but *Symmorphus canadensis* in the Vespidae preys on leafmining larvae belonging to either the Coleoptera or Lepidoptera.

Sociality. Social parasites, also known as cuckoo wasps and bees, have arisen independently in several different families of wasps and bees. Except in the Chrysididae, ordinarily each is closely related, taxonomically, to the particular host wasp or bee which it parasitizes. These parasites either oviposit on the prey of the host wasp as the latter transports it to a nesting site, or they keep the nest under observation and slip into it at some time during the provisioning cycle to deposit the egg while the host female is absent from the nest.

Social life has arisen independently in several stocks of aculeate Hymenoptera, such as the ants, vespid wasps, halictid bees, bumblebees, stingless honeybees, and true honeybees. In the more primitive social insects, the colony dies out at the end of the year except for the recently emerged, fertilized queens which hibernate. These colonies are necessarily less populous than in the more advanced social forms, and are exemplified by the hornets and allies, the halictid bees, and bumblebees. In the higher social forms consisting of the ants, social wasps of the tropics, stingless honeybees, and true honeybees, the colonies are persistent and frequently number several thousands of individuals. The colonies of these wasps and bees divide by swarming, a process by which a queen and a number of workers leave the parent nest and begin a new colony. *See* SOCIAL INSECTS.

Communication. One of the most remarkable facts that has been learned about the honeybee is that the bees employ a sign language to tell other members of the colony that they have located flowers with a copious flow of nectar or abundance of pollen, and approximately how far and in what direction this source is located. If the source is more than 100 yd (90 m) from the hive, the worker performs a tail-wagging dance, usually on the vertical surface of a comb in the darkness. This dance consists of running over a figure-eight pattern with the part between the two loops of the 8 straight. The bee wags her abdomen from side to side when on this straight section. The number of runs over the straight section indicates the distance from the food source, and varies from 9 to 10 runs within 15 s for a distance of 130 yd (120 m) to 3 within 15 s for a distance of 830 yd (750 m). The direction of the source is indicated by the angle of the straight section of the dance. If straight up, the source is toward the Sun; if straight downward, the source is directly away from the Sun; if at an angle to one side or the other of the vertical, then the source is at a corresponding angle and direction away from the Sun. If the food source is less than 100 yd (90 m) from the hive, the worker performs a rapid round dance. The other bees do not receive any clue as to the direction or distance but only an indication that there is a copious food source nearby. *See* ANIMAL COMMUNICATION.

Reproduction. Sawflies deposit their eggs on or in foliage, fruits, stems, or woody tissue, in accordance with the feeding requirements of the larva. Many Parasitica and a few Aculeata deposit more than one egg per host. Some species of Parasitica are polyembryonic, that is, from two to a thousand or more embryos may develop from a single egg. Usually the resulting progeny are all of one sex but exceptions occur. Polyembryony occurs in a few Braconidae, some species of *Macrocentrus*, some species of several genera of Encrytidae, a few Platygasteridae (some *Platygaster*), and one species of Dryinidae, *Aphelopus theliae*.

The size and number of eggs laid is subject to great variation within the order. Most Symphyta and Chalcidoidea deposit a rather small number, from 10 to 50. Most of the Ichneumonoidea deposit larger numbers of eggs, from several hundred to more than a thousand. Some subsocial wasps which practice progressive provisioning may deposit only 6 eggs. Most aculeates lay from 10 to 75 eggs, but apparently some bethylids can lay 150 or more. Honeybees are probably the most fecund of the Aculeata. A queen may lay as many as 1500 eggs per day and 200,000 a year for at least 3 years.

Eggs of the Parasitica are deposited more or less at random on or in the host, but in the Aculeata each species deposits its egg at a particular location on the prey or in the nest.

Sawfly larvae are phytophagous and live on or in foliage, in stems, or in woody tissue. Larvae of the Parasitica may be either internal or external parasites, depending on whether the development takes place inside or outside the host. They can also be classified as solitary if there is only one larva per host, as in most Ichneumonidae, or gregarious, if there are two or more per host as in many Braconidae and Chalcidoidea. Larvae of most wasps are predaceous, feeding on paralyzed insects or spiders stored for them by the mother. Bee larvae feed on pollen-nectar mixture stored by the mother, except for the social parasites, which are predaceous in their first instar on the host egg or young larva and feed on the pollen-nectar mass in the later instars. The larvae of most Aculeata have a blind stomach during most of the larval life so that waste matter is not excreted until the end of the larval life when the connection between stomach and intestine is opened. The larvae of a few wasps, such as Pemphredoninae in the Sphecidae, and of the megachilid bees have this connection opened early in the larval life and excrete small meconial pellets during the last three instars.

Many larvae of the Parasitica and those of the social parasites among the Aculeata are hypermetamorphic, that is, the first instar larva is very different in appearance and behavior from the succeeding instars, which assume the typically maggotlike form of mature larvae of the Apocrita. There are at least 10 different primary larval forms among the species of Apocrita exhibiting hypermetamorphosis. Rarely, the second instar larva differs in appearance from either the first or third.

There are few reliable data on the number of larval instars. Rather meager information indicates that there are three or four instars in various groups of Aculeata.

Silken cocoons are of general occurrence throughout the order, although they are absent in most Chalcidoidea and Cynipoidea, some Formicoidea, and absent sporadically throughout the rest of the Aculeata. The cocoons may be of two layers, as in a few sawflies and wasps, but normally there is only one layer of silk, or the cocoon may consist of only a silken cap. Foreign material may be frequently incorporated in the cocoon. The larva may impregnate the silk with secretions from the gut which cause the cocoon wall to become brittle and varnished. Sand, mud, or prey remnants may be incorporated in the silk of the cocoon. Almost all Hymenoptera

TABLE 2. Economically important Hymenoptera

Family	Scientific name	Common name	Economic importance
Pamphiliidae	*Neurotoma inconspicua*	Plum, web-spinning sawfly	Larvae spin webs and eat foliage of plums
	Pamphilius persicus	Peach sawfly	Larvae roll peach leaves and feed on them
Cimbicidae	*Cimbex americana*	Elm sawfly	Larvae eat foliage of elm, willow, poplar, and maple
Diprionidae	*Diprion hercyniae*	European spruce sawfly	Larvae defoliate spruce; introduced from Europe
	Neodiprion lecontei	Red-headed pine sawfly	Larvae defoliate pines
Tenthredinidae	*Caliroa cerasi*	Pear slug	Larvae skeletonize foliage of pear, cherry, plum; probably introduced from Europe
	Fenusa ulmi	Elm leaf miner	Larvae mine leaves of elm; probably introduced from Europe
	Cladius isomerus	Bristly rose slug	Larvae skeletonize rose leaves
	Hoplocampa cookei	Cherry fruit sawfly	Larvae feed inside cherries
	Pteronidea ribesii	Imported currant worm	Larvae feed on leaves of currants, gooseberries; introduced from Europe
Siricidae	*Tremex columba*	Pigeon tremex	Larvae bore in trunks of weakened or dead maple, oak, elm, other deciduous trees
Cephidae	*Hartigia trimaculata*		Larvae bore in stems of roses, blackberries
	Cephus cinctus	Wheat stem sawfly	Larvae bore in stems of wheat, rye, timothy, and wild grasses
Braconidae	*Aphidius* sp.		Larvae are internal parasites of aphids
	Meteorus sp.		Larvae are internal parasites of lepidopterous and coleopterous larvae
	Macrocentrus sp.		Larvae are internal parasites of lepidopterous larvae
	Apanteles sp.		Larvae are internal parasites of lepidopterous larvae
	Opis sp.		Larvae are dipterous parasites.
	Spathius sp.		Larvae are external parasites of coleopterous larvae
Ichneumonidae	*Scambus* sp.		Larvae are internal parasites of small lepidopterous larvae in leaf mines, leaf rolls, galls
	Polysphincta sp.		Larvae are external parasites of spiders
	Ephialtes sp.		Larvae are internal parasites of lepidopterous pupae
	Megarhyssa macrurus		Larvae are external parasites of wood-boring sawfly larvae
	Gelis sp.		Larvae are parasites in cocoons of other Ichneumonoidea and in egg sacs of spiders
	Acroricnus aequatus		Larvae are parasites in nests of mud dauber wasps
	Diplazon laetatorius		Eggs are laid in syrphid (Diptera) eggs or young larvae, and adults emerge from host puparia
Mymaridae	*Anagrus armatus*		Larvae are parasites in eggs of Hemiptera
Trichogrammatidae	*Trichogramma minutum*		Larvae are parasites in eggs of other insects
Eulophidae	*Sympiesis* sp.		Larvae are parasites of leaf-mining coleopterous and lepidopterous larvae
	Tetrastichus sp.		Larvae are parasites of a wide variety of insect eggs and larvae including those of other Chalcidoidea
Thysanidae	*Thysanus* sp.		Larvae are parasites of Homoptera or of other Chalcidoidea which parasitize Homoptera
Encyrtidae	*Copidosome gelechiae*		Polyembryonic parasites of gall-making lepidopterous larvae
	Aphycus sp.		Larvae are parasites of scale insects
	Ooencyrtus sp.		Larvae are parasites in eggs of other insects
Agaontidae	*Blastophagus psenes*	Fig wasp	Lives within figs and fertilizes them; introduced from Europe
Torymidae	*Megastigmus* sp.		Larvae live in seeds
Eurytomidae	*Harmolita tritici*	Wheat jointworm	Larvae bore in stems of wheat
	Harmolita grandis	Wheat strawworm	Larvae bore in stems of wheat
	Bruchophagus gibbus	Clover seed chalcid	Larvae develop in seeds of clover, alfalfa; possibly an introduced species
Cynipidae	*Diplolepis rosae*	Mossy rose gall	Larvae develop in galls on rose stems; introduced from Europe
	Acraspis erinacei	Oak hedgehog gall	Agamic generation (no males); develops in hedgehog gall on oak leaves; sexual generation develops in soft galls in oak buds
	Amphibolips confluenta	Large oak-apple gall	Larvae develop in large globular galls on oak leaves
Evaniidae	*Evania appendigaster*		Parasitic in egg cases of cockroaches; introduced
Scelionidae	*Telenomus* sp.		Parasitic in eggs of many orders of insects
Chrysididae	*Chrysis* sp.	Cuckoo wasps	Social parasites of other wasps and bees
Bethylidae	*Cephalonomia tarsalis*		Parasites of beetle larvae infesting stored grains
	Scleroderma sp.		Parasites of old-house borer (Coleoptera); have very painful sting
	Goniozus sp.		Parasites of lepidopterous leaf rollers
Tiphiidae	*Tiphia vernalis*	Spring tiphia	Parasites of Japanese beetle larvae; introduced from Japan, Korea, China
Multillidae	*Dasymutilla occidentalis*	Cow killer	Parasite of bumblebee pupae
Formicidae	*Eciton* sp.	Legionary ants	Predacious on other insects
	Pheidole sp.	Harvesting ants	Nest in ground and feed on seeds
	Monomorium pharaonis	Pharaoh ant	A pest ant which nests in buildings; introduced

TABLE 2. Economically important Hymenoptera (*cont.*)

Family	Scientific name	Common name	Economic importance
Formicidae (cont.)	*Solenopsis saevissima var. richteri*	Imported fire ant	Nests in soil; very aggressive; introduced from South America
	Tetramorium caespitum	Pavement ant	Infests houses; introduced
	Atta sp.	Fungus ants	Nest in soil and feed upon fungi which they cultivate on beds of masticated leaves
	Tapinoma sessile	Odorous house ant	Infests houses
	Camponotus sp.	Carpenter ants	Many species nest in wood
	Lasius interjectus	Larger yellow ant	Nests around building foundations and beneath cellar floors
	Myrmecocystus sp.	Honey ants	Nest in soil in Southwest; some workers store large quantities of honey until abdomen is greatly distended
	Formica exsectoides	Allegheny mound ant	Nests in ground beneath large mounds of excavated soil
	Polyergus sp.	Slave-making ants	Enslave workers of other ants to build their nests and feed their young
Vespidae	*Vespa crabro germana*	Giant hornet	Social wasp nesting in sidings of buildings and hollow trees; introduced from Europe
	Vespula (Vespula) sp.	Yellow jacket	Social wasps nesting in ground
	Vespula (Dolichovespula) maculata	Bald-faced hornet	Social wasp which builds large paper nests in trees and shrubs
	Polistes sp.	Paper wasps	Social wasps which build a single umbrella-shaped comb, frequently under eaves or porch roofs
	Eumenes sp.	Jug wasp	Solitary wasps which build small clay jugs in which to rear their young; prey on caterpillars
Pompilidae	*Pepsis* sp.	Tarantula hawks	Prey on tarantulas
Sphecidae	*Trypoxylon politum*		Builds the familiar clay "pipe-organ" nests in which it stores many small spiders
	Chlorion ichneumoneum	Great golden digger wasp	Nests in soil and preys on long-horned grasshoppers
	Sceliphron caementarium	Black and yellow mud dauber	Builds clay cells in which it stores many small spiders
	Sphecius speciosus	Cicada killer	Nests in soil and preys on adult cicadas
	Stictia carolina	Horse guard	Nests in ground and preys on a variety of flies
Colletidae	*Colletes* sp.		Solitary bees which nest in ground, frequently gregariously, and collect pollen and nectar from a variety of flowers
Andrenidae	*Andrena* sp.		As in Colletidae
Halictidae	*Halictus* sp.		As in Colletidae, except that some species are subsocial or social
Megachilidae	*Megachile* sp.	Leaf-cutter bees	Solitary bees which nest in cavities in wood or in ground, and make cells from pieces cut from leaves
Apidae	*Anthophora* sp.	Mining bees	Solitary ground-nesting bees
	Xylocopa virginica	Carpenter bee	Solitary bee which excavates tunnels in solid wood in which it rears its brood
	Bombus sp.	Bumblebees	Social bees which next on or in ground, frequently in old mouse nests
	Trigona sp.	Stingless honeybees	The honeybee of the tropics; does not occur in North America
	Apis mellifera	Honeybee	The beekeeper's bee; escaped bees nest in hollow trees or house sidings; introduced from Europe

that spin cocoons produce the silk from labial glands through the mouthparts, but those Chalcidoidea that construct cocoons produce the silk in the Malpighian tubules and spin it from the anus.;

Economically important species. Many of the Hymenoptera are of economic importance and include both useful and destructive forms. The species listed in **Table 2** are North American except where noted to the contrary. Common names are given, and a brief note is included explaining the importance of each species. *See* ENTOMOLOGY, ECONOMIC.

Phylogeny

The earliest known fossil Hymenoptera are from the Middle Jurassic. Only Symphyta and Parasitica are known from these strata. The absence of Aculeata perhaps substantiates the claim that this section evolved from the Parasitica; or its absence may be only apparent and due to the meager fossil record. Unfortunately, the wing venation of these fossil sawflies is quite specialized and offers no clues to the derivation of the order. The consensus is that the Hymenoptera probably arose from an ancestral type that also gave rise to the Neuroptera and other related orders. The Baltic amber, which dates from the Tertiary, preserved many ants and a few bees and other Hymenoptera. The bees are all members of genera which are now extinct, but some of the ants are thought to be the same as species living now.

Karl V. Krombein

Bibliography. R. H. Arnett and R. L. Jacques, *Insect Life: A Field Entomology Manual for the Amateur Naturalist*, 1985; R. J. Elzinga, *Fundamentals of Entomology*, 5th ed., 1999; H. E. Evans et al., *Insect Biology: A Textbook of Entomology*, 1984; R. L. Jeanne, *Interindividual Behavior Variability in Social Insects*, 1988; B. Klausnitzer, *Insects: Their Biology and Cultural History*, 1987; R. D. Shenfelt, *Hymenopterorum Catalogus*, 1980; F. Taylor and R. Karban (eds.), *The Evolution of Insect Life Cycles*, 1986; A. Wooten, *Insects of the World*, 1985.

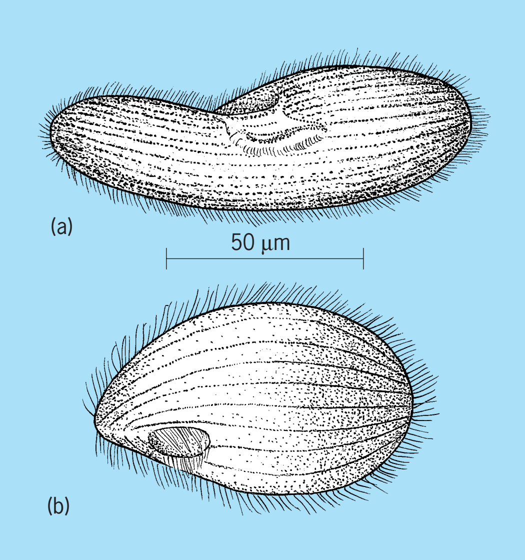
Hymenostomatida

An order of the Holotrichia which contains many species that often are of small size and fairly uniform ciliation. Primarily, these protozoa are of importance as the first possessors of a definite, though inconspicuous, buccal ciliature. This ciliature consists of an undulating membrane on the right side of the buccal cavity and an adoral zone of membranelles that is primitively composed of three membranelles on the left side. This tetrahymenal, or four-part, buccal ciliary apparatus is considered the fundamental condition from which the oral ciliature of many subsequent higher groups evolved. *See* CILIOPHORA.

The majority of hymenostomes are free-living fresh-water forms. On the basis of differences in their patterns of stomatogenesis (new-mouth formation), one group of hymenostomes has been authoritatively considered as representing a new and different ordinal group of the Holotrichia; but the better-known, if more conservative, scheme of classification is espoused here.

The parasitic genus *Ichthyophthirius* is often mistakenly considered to be a gymnostome. *Paramecium* (**illus.** *a*) is the best-known genus of ciliates in the entire subphylum. It is a good-sized widely distributed ciliate. Being easy to recognize and culture, it is a much-studied form. It was at one time classified as a trichostome. As an experimental animal, *Paramecium* has played a major role in the advance of protozoan genetics. *Tetrahymena* (illus. *b*), beginning to rival *Paramecium* as a favorite ciliate in much experimental work, owes its scientific popularity primarily to its ability to grow axenically, that is, free from all other organisms, in a chemically defined medium. Its species are the first animal organisms, excluding the green plantlike flagellate protozoans, to be so grown. The achievement of axenic growth represents a great forward step in experimental studies, particularly those of a biochemical nature, in cancer research and in other important fields of direct, immediate interest to humans. *See* AXENIC CULTURE; GYMNOSTOMATIDA; HOLOTRICHIA; PROTOZOA.

John O. Corliss



Hymenostomatida. (*a*) *Paramecium.* (*b*)*Tetrahymena*.

Hyperbaric oxygen chamber

A specially equipped pressure vessel used in medicine and physiological research to administer oxygen at elevated pressures.

Basic principle. Under normal conditions the red blood cells provide the main transport mechanism for distributing oxygen through the bodies of warm-blooded animals. In humans less than 5% of the oxygen in the body is dissolved in body fluids. The transport capacity of red blood cells permits warm-blooded animals to maintain high body temperatures even in cold climates, and to supply the heavy oxygen demand of a large active brain. However, circulation of the red cells through the blood vessels requires a great amount of work by the heart and can be reduced or stopped by damage or blockage of the blood vessels.

The amount of oxygen dissolved in the body fluids is related to the pressure of oxygen in the lungs (Henry's law). When a person breathes pure oxygen, the amount of oxygen dissolved in body fluids is about six times that when breathing air. This is still too low to supply the needs of the human body. However, breathing pure oxygen at three times normal air pressure causes the amount of oxygen dissolved in body fluids to be equal to that normally carried by the red cells. Research animals have been kept alive in a hyperbaric chamber for some time with all the red cells removed from their blood. At the end of the experimental period the red cells were returned and the animals subsequently led perfectly normal lives.

A principal advantage of dissolved oxygen is that it can be transported throughout the body wherever there is fluid of any sort. It is not limited to circulation through blood vessels. A second advantage is that a given volume of blood contains twice the normal amount of oxygen. The high level of oxygen in the body aids the patient in the following ways:

(1) Oxygen can be carried past an obstruction in the circulatory system, thus relieving oxygen-starved tissues. (2) The work load on the heart can be reduced, since one-half the normal blood flow will provide the normal amount of oxygen required by the body. (3) Poisons such as carbon monoxide can be eliminated. (4) Anaerobic bacteria such as tetanus can be destroyed. (5) The effectiveness of radiation treatment of cancer is increased. *See* RADIATION BIOLOGY; TETANUS.

Equipment. For physiological studies and for some types of patient treatment chambers just large enough for one person have been built. These chambers are relatively inexpensive to build but have limited usefulness because the patient cannot be treated or cared for while sealed up inside the oxygen-filled pressure chamber.

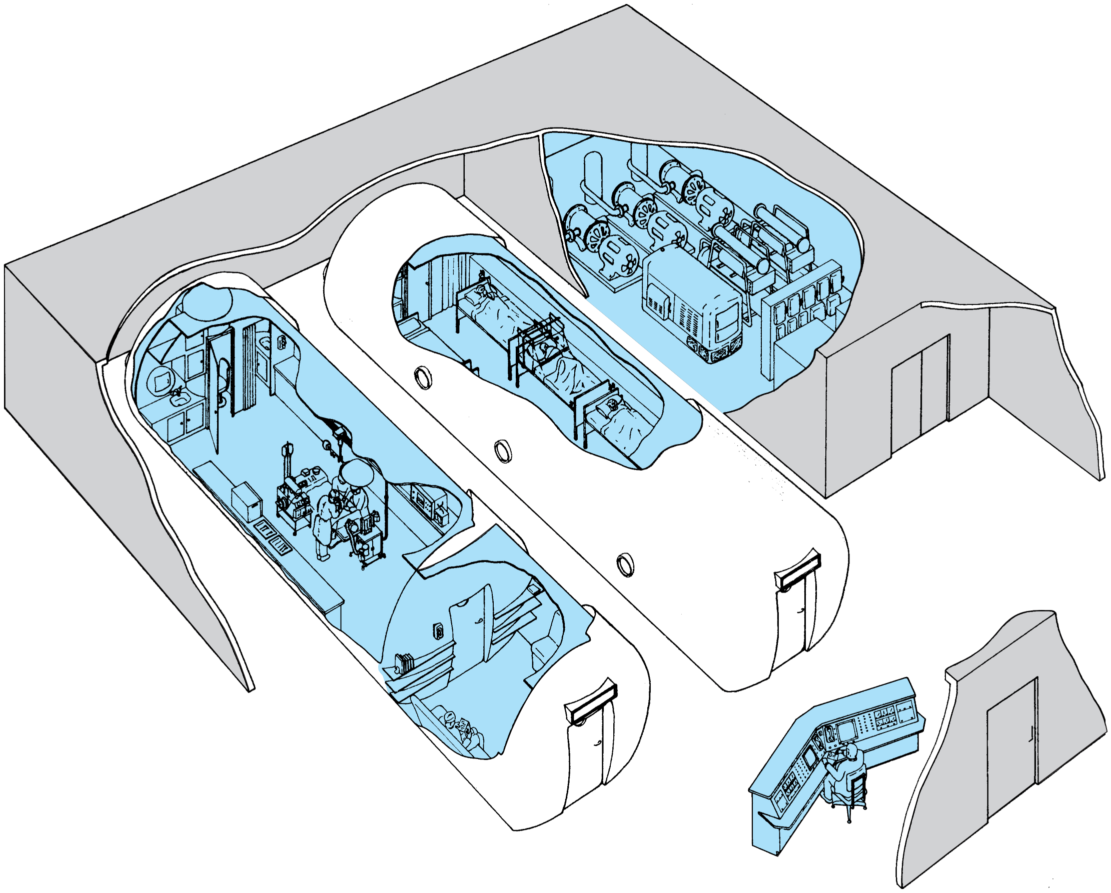
In order to give the patient adequate care or to perform surgery, it is necessary to use much larger equipment. The **illustration** shows a hyperbaric facility consisting of two large chambers. One chamber contains a fully equipped operating room, while the other contains a small medical ward. These large chambers are filled with ordinary air that is compressed to the appropriate pressure. The patients breathe pure oxygen through masks or tents; the attending doctors and nurses breathe the air that fills the chamber.

Each chamber consists of a steel pressure vessel which is roughly 12 ft (4 m) in diameter and about 45 ft (14 m) long. Larger vessels have been employed, but they are extremely difficult to install and require excessively large compressors to provide a fresh air supply.

Inside, the vessel is divided into three areas. In the front area is a large entry lock in which patients on stretchers can be compressed or decompressed while the main section remains at elevated pressure. The main area of the chamber is a room about 30 ft (9 m) wide because of the curvature of the walls. In one of the chambers this room contains a fully equipped surgical operating table. In the other chamber, beds for six patients can be accommodated. Beyond the main section there is a small lavatory and another small lock through which attending personnel can enter or leave quickly during an operation. There is also an instrument lock through which small equipment and supplies can be passed.

In a nearby room outside the chambers air compressors and air conditioning equipment are located. From a central console the operator can monitor all that is happening in each of the chamber areas and can activate any of the mechanical equipment by remote control.

Proper conditioning of the air under a wide variety



Hyperbaric oxygen treatment facility.

of operating pressures is necessary for the comfort and safety of the people in the chamber. When air is compressed, that part of the total pressure represented by each component increases to the same extent (Dalton's law). For example, on a comfortable day relative humidity is about 50–75%. This means that the water vapor is 50–75% of the saturation pressure. If this air is compressed to three times its normal pressure, the pressure of each component including the water vapor will be increased threefold. On the other hand, the water vapor saturation pressure will remain constant as long as the temperature remains the same. Therefore, the humidity will rise to 150–225% which means that some of the water vapor must condense as fog or even rain. Without any special provision to remove moisture the hyperbaric chamber would always be foggy and damp, and the atmosphere would have a very unpleasant effect on the senses. To keep the atmosphere clean and comfortable, the air in the chamber is changed completely every 20 min. The fresh air is dried and cooled after being compressed.

Providing normal facilities in a hyperbaric chamber presents novel problems. For example, normal water pressure may not be sufficient to make the water flow out of a faucet into the high-pressure environment in the chamber. Booster pumps raise the pressure enough to ensure normal flow. Draining wastewater out of the chamber under 3 atm (300 kilopascals) pressure could have rather spectacular results if the drain lines led directly to the sewers. Special waste-receiving tanks are provided which permit the high pressure to bleed off before releasing the wastewater to the normal drainage system.

Changes in pressure must take place rather gradually for the comfort and safety of the people in the chamber. The ear-popping problems that commonly occur in express elevators and airplanes are considerably magnified because the total change in pressure is considerably greater. Normally a period of 5–10 min is required to pressure up or down.

A special problem exists for the attending personnel who breathe the compressed air. This is the well-recognized problem of the elimination of dissolved nitrogen from the body. For working periods in the chamber up to about 1 h no problem is involved, because nitrogen dissolves slowly. However, for periods of more than 1 h enough nitrogen will accumulate to cause formation of gas bubbles in the bloodstream if decompression takes place too fast. These gas bubbles cause the condition known as caisson disease, or the bends. Tables governing the safe decompression schedules for various working times and pressures have been prepared by the U.S. Navy.

For example, if a surgical team conducts an open-heart operation lasting 3 h in a hyperbaric chamber at three times normal pressure, the Navy tables indicate that decompression should include a 19-min hold at $1^2/_3$ times normal pressure and another hold of 79 min at $1^1/_3$ times normal pressure. Altogether this means that decompression requires almost 2 h after a 3-h working period. This would not be required for the patient, since the pure oxygen administered during the surgery does not present any hazard. *See* DECOMPRESSION ILLNESS.

Because of the potentially long periods of time which may be required to decompress, the people in the chamber are relatively isolated from the outside world. They are highly dependent on the chamber operator, who remains at the control console continually as long as any person is in the chamber. Communication is maintained by closed-circuit television with the main section and each of the personnel locks. The operator continually observes the pressure, temperature, and gas composition in the chamber.

While loss of electric power would not of itself be dangerous, the chamber would soon become rather uncomfortable without the constant flow of fresh air, and it would soon become necessary to decompress and leave the chamber. This could mean stopping an operation before completion, which could pose serious problems. Emergency power supplies are therefore provided, and clean dry air is stored in high-pressure cylinders so that no interruption disrupts the work of the physicians. During the great power failure that blacked out the Northeast on November 9–10, 1965, the hyperbaric facility at Mount Sinai Hospital in New York City was an island of light in otherwise complete darkness.

History. Physicians have long been interested in the medical applications of elevated pressures. Orval Cunningham observed beneficial results from such conditions when treating influenza victims during World War I. However, treatment with oxygen at atmospheric pressure produced the same benefits with much less difficulty. It was not until the 1950s that the combination of elevated pressure and pure oxygen was suggested, especially for open-heart surgery. In 1965 a large chamber was built at the University of Amsterdam in the Netherlands. Since that time facilities have been constructed at a number of medical schools and large hospitals in Great Britain, Canada, and the United States. *See* OXYGEN; RESPIRATION.

Arthur W. Francis

Bibliography. B. Fischer et al., *Handbook of Hyperbaric Oxygen Therapy*, 1988; G. S. Innes (ed.), *The Production and Hazards of a Hyperbaric Oxygen Environment*, 1970.

Hyperbola

A curve cut from a cone or revolution by a plane that intersects both nappes of the cone and does not contain the apex (**Fig. 1**). In analytic geometry it is shown, as shown in **Fig. 2**, that a hyperbola is the locus of points P in a plane, such that $PF = \epsilon \cdot PD$, where PF and PD denote the distances of P from a fixed point F (focus) and a fixed line (directrix) of the plane, respectively, and ϵ is a constant, greater than 1. It is also the locus of points P, the difference of whose distances from two fixed points F, F' (foci)

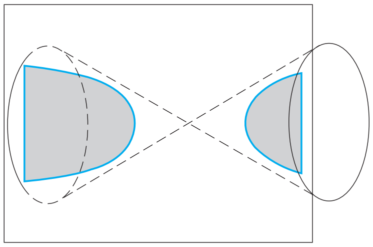


Fig. 1. Hyperbola as a conic section.

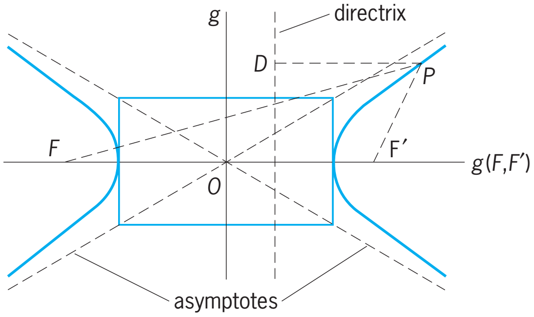


Fig. 2. Hyperbola as a locus of points.

$PF - PF'$ is a constant $2a$ that is less than the distance $2c$ between the foci. The curve is symmetric to the line $g(F, F')$ determined by F, F' and to O, their midpoint. It consists of two branches that are images of each other in the line g through O, perpendicular to $g(F, F')$. There are two lines through O, making equal angles with $g(F, F')$, and to each of which points on each branch get indefinitely close; that is, if point P traverses either branch of the hyperbola, its distance from these lines approaches zero.

The lines are called asymptotes of the hyperbola. If the asymptotes are mutually perpendicular, the hyperbola is called rectangular or equilateral, since then its transverse axis $AA' = 2a$, where A, A' are on $g(F, F')$ with $OA = OA'$, equals its conjugate axis $BB' = 2b = 2(c^2 - a^2)^{1/2}$ (B, B' on g, $OB = OB'$). The eccentricity $\epsilon = c/a$ of every equilateral hyperbola is $\sqrt{2}$. The asymptotes of each hyperbola are the diagonals of a rectangle with center at O, having two sides of length $2a$ parallel to $g(F, F')$, and the other two parallel sides have length $2b$. If an ellipse and a hyperbola have the same foci F, F', they are called confocal. Through each point of the plane there is exactly one ellipse and one hyperbola with the same given foci. They intersect each other at right angles at each of the four points they have in common.

The area of the triangle formed with the asymptotes of a hyperbola by a variable tangent is equal to ab. Conversely, if a line cuts the asymptotes of a hyperbola in two points that lie on the same side of the conjugate axis and forms with them a triangle of area ab, then the line is tangent to the hyperbola. The tangent and normal at a point P bisect the angles formed by the lines joining P to the foci F, F'. *See* ANALYTIC GEOMETRY; CONIC SECTION. Leonard M. Blumenthal

Hyperbolic function

The hyperbolic sine and cosine of a real or complex variable z are defined by Eqs. (1). Both sinh z and cosh z have a period $2\pi i$ of e^z. From $De^z = de^z/dz = e^z$, Eqs. (2) are obtained.

$$\sinh z = \frac{e^z - e^{-z}}{z} \qquad \cosh z = \frac{e^z + e^{-z}}{z} \tag{1}$$

$$D \sinh z = \cosh z \qquad D \cosh z = \sinh z \tag{2}$$

Since

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

by definition, Eqs. (3) hold and the series converge for all z and yield relations (4). Thus relations between the circular functions become hyperbolic (and vice versa) when z is replaced by iz. In particular $\cos^2 z + \sin^2 z = 1$ becomes $\cosh^2 z - \sinh^2 z = 1$; and addition theorems (5) become Eqs. (6). Moreover Eqs. (1) yield eulerian equations (7).

$$\begin{aligned} \sinh z &= \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} \\ \cosh z &= \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} \end{aligned} \tag{3}$$

$$\begin{aligned} \sinh iz &= i \sin z \qquad \cosh iz = \cos z \\ \sin iz &= i \sinh z \qquad \cos iz = \cosh z \end{aligned} \tag{4}$$

$$\begin{aligned} \sin(z + \zeta) &= \sin z \cos \zeta + \cos z \sin \zeta \\ \cos(z + \zeta) &= \cos z \cos \zeta - \sin z \sin \zeta \end{aligned} \tag{5}$$

$$\begin{aligned} \sinh\left(z + \zeta\right) &= \sinh z \cosh \zeta + \cosh z \sinh \zeta \\ \cosh\left(z + \zeta\right) &= \cosh z \cosh \zeta + \sinh z \sinh \zeta \end{aligned} \tag{6}$$

$$i \sin\ z = \frac{e^{iz} - e^{-iz}}{2} \qquad \cos z = \frac{e^{iz} + e^{iz}}{2}\text{''} \tag{7}$$

With $z = x + iy$, the addition theorems give Eqs. (8). Hence Eqs. (9) hold and all zeros of sinh z and cosh z lie on the axis $x = 0$ at the points $y = 2n\pi i$ and $y = (2n + {}^1\!/_2)\pi i$, respectively. All zeros are simple.

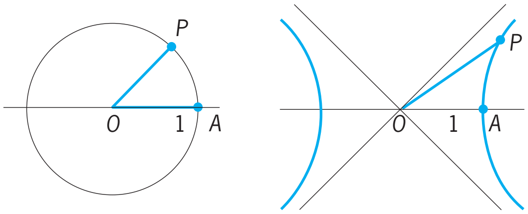
$$\begin{aligned} \sinh z &= \sinh x \cos y + i \cosh x \sin y \\ \cosh z &= \cosh x \cos y + i \sinh x \sin y \end{aligned} \tag{8}$$

$$\begin{aligned} |\sinh z|^2 &= \sinh^2 x + \sin^2 y \\ |\cosh z|^2 &= \sinh^2 x + \cos^2 y \end{aligned} \tag{9}$$

The functions e^z, sin z, cos z, sinh z, cosh z are analytic throughout the complex plane; all have an essential singularity at $z = \infty$; e^z has no zeros, sin z and cos z vanish only on the axis of reals, sinh z and cosh z only on the axis of imaginaries.

The hyperbolic tangent and cotangent are defined by Eqs. (10). They have the period πi and have simple poles at the zeros of cosh z and sinh z,

$$\tanh z = \frac{\sinh z}{\cosh z} \qquad \coth z = \frac{\cosh z}{\sinh z} \tag{10}$$



Circle, $x^2 + y^2 = 1$; hyperbola; $x^2 - y^2 = 1$.

respectively. Moreover Eqs. (11) hold true. By definition Eqs. (12) hold.

$$D\tanh z = \frac{1}{\cosh^2 z} \qquad D\coth z = -\frac{1}{\sinh^2 z} \qquad (11)$$

$$\operatorname{sech} z = \frac{1}{\cosh z} \qquad \operatorname{csch} z = \frac{1}{\sinh z} \qquad (12)$$

The hyperbolic functions are related to the equilateral hyperbola in much the same way that the circular functions are related to the circle (equilateral ellipse). For the hyperbola of Eqs. (13), and for the circle of the area of the sector *OAP* shown in the **illustration** the circuit integral is given by Eq. (15). In both cases of Eqs. (14), parameter $t = \sigma$.

$$x = \cosh t \qquad y = \sinh t \qquad (13)$$

$$x = \cos t \qquad y = \sin t \qquad (14)$$

$$\sigma = \frac{1}{2}\oint_{OAP} (x\,dy - y\,dx) = \frac{1}{2}\oint_0^t dt = \frac{t}{2} \qquad (15)$$

See HYPERBOLA; TRIGONOMETRY. Louis Brand

Hyperbolic navigation system

A navigation system that produces hyperbolic lines of position (LOPs) through the measurement of the difference in times of reception (or phase difference) of radio signals from two or more synchronized transmitters at fixed points. Such systems require the use of a receiver which measures the time difference (or phase difference) between arriving radio signals. Assuming the velocity of signal propagation is relatively constant across a given coverage area, the difference in the times of arrival (or phase) is constant on a hyperbola having the two transmitting stations as foci (**Fig. 1**). Therefore, the receiver measuring time or phase difference between arriving signals must be located somewhere along the hyperbolic line of position corresponding to that time or phase difference. If a third transmitting station is available, the receiver can measure a second time or phase difference and obtain another hyperbolic line of position. The intersection of the lines of position provides a two-dimensional navigational fix (Fig. 1). User receivers typically convert this navigational fix to latitude and longitude for operator convenience.

The choice of frequency and locations of transmitters determines both the utility and the accuracy of hyperbolic navigation systems. In general, longer baselines (Fig. 1) enhance accuracy; however, transmitter power limitations may constrain a system to shorter baselines in order to maintain accurate synchronism.

Accuracy is a primary concern for any navigation system. In hyperbolic systems (as with all navigation systems) there are fundamental limits to accuracy. Important geometric concepts in describing accuracy limits for hyperbolic navigation systems are LOP crossing angles, gradient, and horizontal dilution of precision (HDOP).

LOP crossing angle. This is the angle between lines of tangency at the intersection of the lines of position. The first station's signal may be denoted as a master, and all other time differences can be measured relative to the master signal. Each line tangent to a hyperbolic line of position at the user's position *P* (**Fig. 2**) bisects one of the angles, α or β, formed at *P* between one of the two secondary stations and the master station. As a result, the LOP crossing angle, γ, equals $\alpha/2 + \beta/2$, or one-half the angle between the two secondary stations as observed at the user's position. LOP crossing angle is purely a function of the bearings to the two secondary stations being used, and is independent of the bearing to master. For a two-LOP crossing, angles of 120° are considered to be optimal due to correlation in time differences; however, such cannot be achieved throughout the coverage area. As crossing angles diminish from 90° to about 30°, the fix-error contour increases in area and elongates to form an ellipse. Crossing angles of less than 30° are generally considered unacceptable.

Gradient. This is a measure of the change in calculated geodetic position caused by a change in the measured time or phase difference; that is, gradient can be considered an expression of geodetic error sensitivity. Gradient varies depending on user location within the coverage area. For hyperbolic systems that measure difference in times of reception of two radio signals propagating at a nominal velocity of 983.24 ft/μs (299.69 m/μs), the gradient Δ

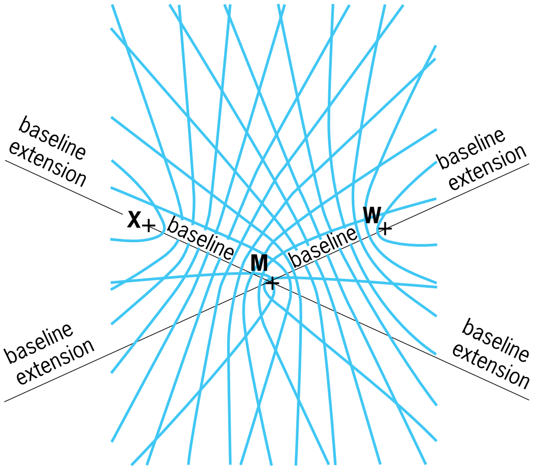


Fig. 1. Typical grid of hyperbolic lines of position formed by a master (M) and two secondary stations (W and X).

can be expressed by the equation below, where ψ

$$\Delta = \frac{491.62}{\sin\,(\psi/2)}\,\text{ft}/\mu\text{s} = \frac{149.85}{\sin\,(\psi/2)}\ \text{m}/\mu\text{s}$$

is the angular difference in bearing between the two stations (that is, one master and one secondary) as observed at the user's position. Gradient is purely a function of the difference in bearing to master and bearing to secondary. The minimum gradient occurs for positions located on the baseline, that is, directly between the two stations (Fig. 1), where the angle ψ is 180°. Gradient tends toward infinity for positions located on the baseline extension (Fig. 1), where the angle ψ tends toward zero. For a single pair of transmitting stations, smaller gradient implies improved accuracy, so system planners typically specify that hyperbolic navigation system baselines occur over areas where improved accuracy is needed. For a single pair of transmitting stations, no reliable position information can be obtained by measuring time or phase differences in the vicinity of the baseline extension (Fig. 1), since gradient tends to infinity at these locations.

Horizontal dilution of precision. This is a dimensionless multiplier which takes into account the geometric effects of both gradient and crossing angle, and shows how accuracy is degraded as a function of bearing to the master and two secondary stations. **Figure 3** shows a master and two secondary transmitters along with curves of constant HDOP. Small values of HDOP (as seen within the triangle *MXY*) suggest areas of improved accuracy. HDOP is multiplied by ranging error (σ) to obtain root mean square position error (drms), and by twice ranging error (2σ) to obtain 2 drms. For example, if gaussian signal stability statistics are assumed, then the receiver's actual position will be outside a circle of radius 2 drms, centered at the receiver's computed fix position, between 1.8 and 4.6% of the time. In other words, if a receiver is allowed to record positions over a period of time, between 95.4 and 98.2% of the calculated positions will be within a circle of radius 2 drms centered at the actual position. *See* AIR NAVIGATION; DECCA; LORAN; MARINE NAVIGATION; STATISTICS. Benjamin B. Peterson; Richard J. Hartnett

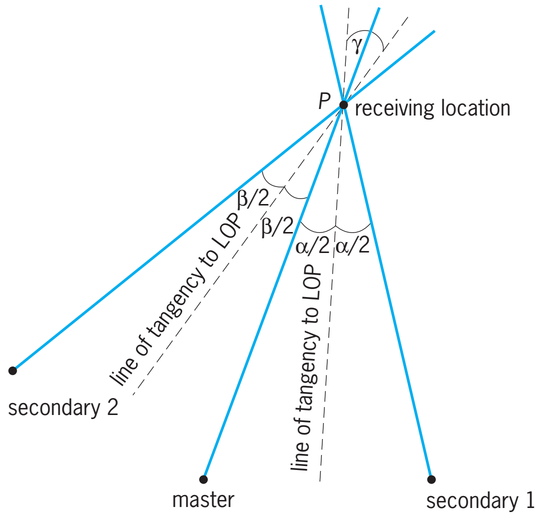


Fig. 2. Individual lines of position from a master and two secondaries at location *P*. γ = LOP crossing angle.

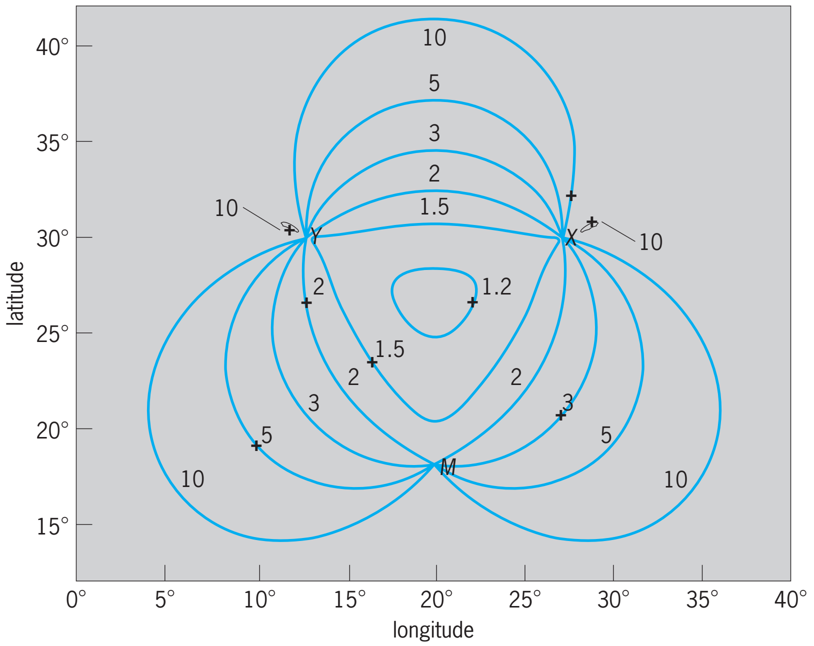


Fig. 3. Typical curves of constant horizontal dilution of precision (HDOP) for a master (*X*) and two secondary stations (*X* and *Y*). Numbers give HDOP values of contours.

Bibliography. Special issue on global navigation systems, *Proc. IEEE*, vol. 71, no. 10, 1983; L. Tetley and D. Calcutt, *Electronic Aids to Navigation*, 2d ed., 1992; U.S. Department of Transportation, *U.S. Coast Guard Loran-C User Handbook*, COMDTPUBP16562.6, 1992; J. C. Whitaker (ed.), *The Electronics Handbook*, CRC Press, 1996.

Hypercharge

A quantized attribute, analogous to electric charge, introduced in the classification of a subset of elementary particles—the so-called baryons—including the proton and neutron as its lightest members. As far as is known, electric charge is absolutely conserved in all physical processes. Hypercharge was introduced to formalize the observation that certain decay modes of baryons expected to proceed by means of the strong nuclear force simply were not observed. *See* ELECTRIC CHARGE.

Unlike electric charge, however, the postulated hypercharge was found not to be conserved absolutely; the weak nuclear interactions do not conserve hypercharge—and indeed can change hypercharge by ± 1 or 0 units. These observations establish certain relationships but do not define the hypercharge. The simplicity of the relationships suggested a hypercharge (symbol *Y*) as follows: for the proton and neutron, $Y = +1$; for the pion, $Y = 0$.

For example, consider the heavier Λ^0 baryon, which in principle is energetically unstable against decay into a proton and a pion in a characteristic time of about 10^{-23}s. But it is observed that the Λ^0

lives about 10^{13} times longer than this characteristic time, suggesting that, for this baryon, $Y = -1$ and thus it can decay into a proton and a pion only by means of the weak nuclear interaction. Systematic observations of this kind have permitted assignment of hypercharge values to all the baryons and antibaryons.

When the known baryons are classified according to their electric charge and their hypercharge, they naturally group into octets in the scheme first proposed by M. Gell-Mann and K. Nishijima. The quarks, hypothesized as the fundamental building blocks of matter, must have fractional hypercharge as well as electrical charge; the simplest quark model suggests values of 1/3 and 2/3, respectively. *See* BARYON; ELEMENTARY PARTICLE; QUANTUM MECHANICS; QUARKS; SYMMETRY LAWS (PHYSICS); UNITARY SYMMETRY.

D. Allan Bromley

Hyperfine structure

A closely spaced structure of the spectrum lines forming a multiplet component in the spectrum of an atom or molecule, or of a liquid or solid. In the emission spectrum for an atom, when a multiplet component is examined at the highest resolution, this component may be seen to be resolved, or split, into a group of spectrum lines which are extremely close together. This hyperfine structure may be due to a nuclear isotope effect, to effects related to nuclear spin, or to both.

Isotope effect. The element zinc, for example, has three relatively abundant naturally occurring nuclear isotopes, ^{64}Zn, ^{66}Zn, and ^{68}Zn. The radius of a nucleus increases with the nuclear mass and, for a given element, the Coulomb interaction of the nucleus with the atomic *s*-electrons will be slightly weaker when the nuclear size is larger. This nuclear size effect causes a slight shift of certain of the spectrum lines, and this shift will be different for each isotope. For a mixture of ^{64}Zn, ^{66}Zn, and ^{68}Zn, certain of the multiplet components will thus consist of three closely spaced lines, one line for each isotope. A study of the isotope effect for an element leads, for example, to information about the dependence of the nuclear size on isotope mass, that is, on the number of neutrons in the isotope. *See* ISOTOPE SHIFT.

Structure due to nuclear spin. For the zinc isotopes discussed above, the nuclear spin $I = 0$, and these nuclei will be nonmagnetic and, in effect, have a spherical shape. If $I \neq 0$, however, two new nuclear properties may be observed. The nucleus may have a magnetic moment, and the shape of the nucleus may not be spherical but rather may be that of a prolate or oblate spheroid; that is, it may have a quadrupole moment. *See* SPIN (QUANTUM MECHANICS).

Atoms and molecules. If the electrons in an atom or a molecule have an angular momentum, the electron system may likewise have a magnetic moment. An electron quadrupole moment may also exist. The magnetic moment of the nucleus may interact with the magnetic moment of the electrons to produce a magnetic hyperfine structure. The quadrupole moments of the nucleus and of the electrons may couple to give an electric quadrupole hyperfine structure. In a simple example, the magnetic and the electric quadrupole hyperfine structure may be described by an energy operator H_{hfs} which has the form below,

$$H_{\text{hfs}} = A\bar{I} \cdot \bar{S} + P\left[\bar{I}_z^{\,2} - 3I\,(I+1)\right]$$

where $\bar{S}$ is an operator describing electron spin, $\bar{I}$ and $\bar{I}_z$ are operators describing the nuclear spin and its z component, and I gives the magnitude of the nuclear spin. A and P are coupling constants which may take positive or negative values, and may range in magnitude from zero to a few hundred meters^{-1}. The term in A describes a magnetic, and the term in P a quadrupole, hyperfine structure.

The measurement of a hyperfine structure spectrum for a gaseous atomic or molecular system can lead to information about the values for A and P. These values may be interpreted to obtain information about the nuclear magnetic and quadrupole moments, and about the atomic or molecular electron configuration.

Important methods for the measurement of hyperfine structure for gaseous systems may employ an interferometer, or use atomic beams, electron spin resonance, or nuclear spin resonance. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; INTERFEROMETRY; MAGNETIC RESONANCE; MOLECULAR BEAMS; NUCLEAR MAGNETIC RESONANCE (NMR).

Liquid and solid systems. Hyperfine structure coupling may also occur and may be measured for liquid and solid systems. For liquids and solids, measurements are often made by electron spin or nuclear spin resonance methods. For solids, and for radioactive nuclei, one may, for example, also employ the Mössbauer effect or the angular correlation of nuclear gamma rays. *See* GAMMA RAYS; MÖSSBAUER EFFECT.

For a diamagnetic solid, $A = 0$ in the equation above, and if the crystalline environment of an atom is cubic, $P = 0$ also. If this environment is not cubic, P may have a finite measurable value. *See* DIAMAGNETISM.

If the solid is paramagnetic, ferromagnetic, or antiferromagnetic, A may be finite and measureable, and again P may or may not be zero depending on whether the atomic environment is cubic or not. *See* ANTIFERROMAGNETISM; FERROMAGNETISM; PARAMAGNETISM.

One may gain information about the nuclear moments and about electron bonding and magnetic structure from measurements of hyperfine structure for liquids and solids. Such measurements are extensively used, for example, in atomic and condensed matter physics, chemistry, and biology. *See* ATOMIC STRUCTURE AND SPECTRA; NUCLEAR MOMENTS.

Louis D. Roberts

Bibliography. A. Abragam, *The Principles of Nuclear Magnetism*, 1961, reprint 1983; D. M. Brink and G. R. Satchler, *Angular Momentum*, 3d ed., 1994; R. S. Raghavan and D. E. Murnick (eds.), *Hyperfine Interactions IV*, 1978.

Hypergeometric functions

The analytic continuation of the function defined by the series in Eq. (1), where the shifted factorial $(a)_n$ is defined by Eq. (2).

$$ {}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n n!} z^n \qquad |z| < 1 \tag{1}$$

It satisfies Eq. (3), a linear, homogeneous differential equation whose only singular points in the full complex plane are regular singular points at 0, 1, and ∞.

$$ (a)_n = a(a+1) \cdots (a+n-1) \qquad n = 1, 2, \ldots, \qquad (a)_0 = 1 \tag{2}$$

$$ z(1-z)y'' + [c - (a+b+1)z]y' - aby = 0 \tag{3}$$

For $|z| < 1$, Re $c >$ Re $b > 0$, it is given by the integral representation of Eq. (4), due to L. Euler, where Γ represents the gamma function. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES; GAMMA FUNCTION; SERIES.

$$ {}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \cdot \int_0^1 (1-zt)^{-a} t^{b-1} (1-t)^{c-b-1} dt \tag{4}$$

The interest in hypergeometric functions comes from the many important functions which are special cases of the general hypergeometric function, the rich theory which has been developed for the general hypergeometric function, and the many times they occur in applications. Classically hypergeometric functions have arisen in science as solutions to differential equations. As discrete, rather than continuous, models of physical phenomena have become increasingly useful, hypergeometric functions have continued to arise as solutions to the equations governing these models. *See* DIFFERENTIAL EQUATION.

Elementary functions. Among the special cases are the elementary functions in Eqs. (5).

$$\begin{aligned} \log(1-z) &= z\,{}_2F_1(1, 1; 2; z) \\ \sin^{-1} z &= z\,{}_2F_1\left({}^1/_2, {}^1/_2; {}^3/_2; z^2\right) \\ \tan^{-1} z &= z\,{}_2F_1\left({}^1/_2, 1; {}^3/_2; -z^2\right) \\ (1-z)^{-a} &= {}_2F_1(a, b; b; z) \end{aligned} \tag{5}$$

Continued fraction expansions. An example of a result which is useful, far from obvious, and yet is a simple consequence of the general theory of hypergeometric functions is the continued fraction expansion of $\tan^{-1}z$ in Eq. (6). Since $\tan^{-1} 1 = \pi/4$, the continued fraction in Eq. (6) gives an explicit expression for π.

$$ \tan^{-1} z = \cfrac{z}{1 + \cfrac{\frac{1.1}{1.3} z^2}{1 + \cfrac{\frac{2.2}{3.5} z^2}{1 + \cfrac{\frac{3.3}{5.7} z^2}{1 + \cdots}}}} \tag{6}$$

See CIRCLE.

C. F. Gauss found a continued fraction for the ratio ${}_2F_1(a,\ b+1;\ c+1;\ z)/{}_2F_1(a,\ b;\ c;\ z)$, which reduces to Eq. (6) in the special case $a = {}^1/_2$, $b = 0$, $c = {}^1/_2$. To obtain this continued fraction, Gauss derived three-term recurrence relations which connect three hypergeometric functions, two of which are contiguous if the three parameters, "a," "b," and "c," are equal and the third differs by one. These contiguous relations can be thought of as three-term difference equations which are a discrete analog of differential equation (3). Most of the occurrences of hypergeometric functions arise because they satisfy a second-order differential or difference equation. *See* INTERPOLATION.

Transformations. A simple change of variables ($t = 1 - s$) in the integral of Eq. (4) gives linear (fractional) transformation (7).

$$ {}_2F_1(a, b; c; z) = (1-z)^{-a}\,{}_2F_1[a, c-b; c; z/(z-1)] \tag{7}$$

Iterating this transformation, after using the symmetry in "a" and "b," gives Eq. (8).

$$ {}_2F_1(a, b; c; z) = (1-z)^{c-a-b}\,{}_2F_1(c-a, c-b; c; z) \tag{8}$$

See CONFORMAL MAPPING.

There is a very important subclass of hypergeometric functions, depending on two, rather than three, parameters, which has a "quadratic transformation." The two basic ones are Eqs. (9).

$$\begin{aligned} &{}_2F_1\left(2a, 2b; a+b++{}^1/_2; z\right) \\ &\quad = {}_2F_1\left[a, b; a+b+{}^1/_2; 4z(1-z)\right] \\ &{}_2F_1(2a, a; 2a; z) \\ &\quad = (1-z)^{-b}\,{}_2F_1\left[b, a-b; a+{}^1/_2; z^2/(4z-4)\right] \end{aligned} \tag{9}$$

This class is important because each function in it can be multiplied by an algebraic function to give a Legendre function, and all Legendre functions arise in this way. *See* LEGENDRE FUNCTIONS.

Confluent hypergeometric functions. Differential equation (3) has regular singular points at the points 0, 1, and ∞. These can be moved to arbitrary points z_1, z_2, z_3 by a linear fractional transformation, and the resulting equation exhibits the symmetries given in linear and quadratic transformations (7)–(9), in a more transparent fashion. Also the resulting singular

Orthogonal polynomials expressed as generalized hypergeometric functions

Polynomial	Hypergeometric representation	Distribution	
Jacobi $P_n^{(\alpha,\beta)}(x)$	$\dfrac{(\alpha+1)_n}{n!}\ {}_2F_1[-n,\quad n+\alpha+\beta+1;\alpha+1;(1-x)/2]$	$(1-x)^\alpha(1+x)^\beta$	$-1 < x < 1$
Laguerre $L_n^\alpha(x)$	$\dfrac{(\alpha+1)_n}{n!}\ {}_1F_1(-n;\alpha+1;x)$	$x^\alpha e^{-x}$	$x > 0$
Hermite $H_n(x)$	$(2x)^n_2F_0\left(-\dfrac{n}{2},\dfrac{1-n}{2};-;-\dfrac{1}{x^2}\right)$	e^{-x^2}	$-\infty < x < \infty$
Hahn $Q_n(x;\alpha,\beta,N)$	${}_3F_2\left(\begin{matrix}-x,-n,n+\alpha+\beta+1\\ -N,\alpha+1\end{matrix};1\right)$	$\dfrac{(\alpha+1)_x}{x!}\ \dfrac{(\beta+1)_{N-x}}{(N-x)!}$	$x = 0, 1, \ldots, N$
Meixner $M_n(x;\beta,c)$	${}_2F_1(-n,-x;\beta;1-c^{-1})$	$\dfrac{(\beta)_x c^x}{x!}$	$x = 0, 1, \ldots$
Krawtchouk $K_n(x;p,N)$	${}_2F_1(-n,-x;-N;p^{-1})$	$\binom{N}{x}p^x(1-p)^{N-x}$	$x = 0, 1, \ldots, N$
Charlier $C_n(x;a)$	${}_2F_0(-n;-x;-;-a^{-1})$	$a^x/x!$	$x = 0, 1, \ldots$

points can be made to coalesce. For the ordinary hypergeometric function this procedure is called confluence. The function ${}_2F_1(a,b;c;z/b)$ satisfies a differential equation whose regular singular points are at 0, b, and ∞, and if b is allowed to become large, the resulting function is called a confluent hypergeometric function. Explicitly, it is given by Eq. (10). Transformation formula (8) becomes Eq. (11). Quadratic transformation (9) implies Eq. (12). The function on the right-hand side is a simple multiple of a Bessel function of imaginary argument, or $I_{a-(1/2)}(z)$. Using this limit, differential equation (3) becomes Eq. (13).

$$ {}_1F_1(a;c;z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(c)_n}\frac{z^n}{n!} \qquad (10)$$

$$ {}_1F_1(a;c;z) = e^z_1F_1(c-a;c;-z) \qquad (11)$$

$$ e^{-z}{}_1F_1(a;2a;2z) = {}_0F_1\left(-;a+{}^1/_2;z^2/4\right) \qquad (12)$$

$$ zy'' + (c-z)y' - ay = 0 \qquad (13)$$

It has a regular singular point at 0 and an irregular singular point at infinity, and one of its solutions is ${}_1F_1(a;c;z)$. *See* BESSEL FUNCTIONS.

Generalized hypergeometric functions. Since a number of sums, such as ${}_2F_1$, ${}_1F_1$, and ${}_0F_1$, arise in a very natural way, it is useful to consider the more general hypergeometric function defined by Eq. (14), where

$$ {}_pF_q(a_p;b_q;z) = \sum_{n=0}^{\infty} \frac{(a_1)_n\cdots(a_p)_n z^n}{(b_1)_n\cdots(b_q)_n n!} \qquad (14)$$

$a_p = a_1, \ldots, a_p$; $b_q = b_1, \ldots, b_q$. All of these generalized hypergeometric functions satisfy differential equations of order max $(p, q+1)$, where if $p > q + 1$ one of the a's is assumed to be a negative integer, since the series does not converge except when $z = 0$ without this assumption.

Generalized hypergeometric functions in one variable can also be thought of as coming from the series (15), where the term ratio a_{n+1}/a_n is a rational function of n. Since the 1980s, significant progress has been made in studying many variable hypergeometric functions through integral representations and differential equations, including the discovery of new beta-function-type integrals.

$$ \sum_{n=0}^{\infty} a_n \qquad (15)$$

Orthogonal polynomials. These functions also satisfy difference equations in the parameters a_i and b_j. Some of these difference equations give rise to polynomials which are orthogonal with respect to many of the important distribution functions in statistics (see **table**).

These polynominals are useful in studying statistical and probabilistic problems. Meixner and Laguerre polynomials play an essential role in birth and death processes when the birth and death parameters are linear functions of the size of the population. Jacobi polynomials for certain special values of the parameters α and β are the zonal spherical harmonics on spheres and projective spaces. For the case $\alpha = \beta = 0$, the polynomials reduce to Legendre polynomials and the sphere is the unit sphere in three dimensions. This point of view leads to some of the deepest formulas known about hypergeometric functions, especially those known as addition formulas. For example, Jacobi polynomials satisfy formula (16) where notation (17) applies,and $c(m,k,\alpha,\beta,n)$ is a product of shifted factorials. *See* PROBABILITY; SPHERICAL HARMONICS; STATISTICS.

$$ \begin{aligned} P_n^{(\alpha,\beta)}\Bigg[\frac{(1+x)(1+y)}{2} &+ \frac{r^2(1-x)(1-y)}{2} \\ &+ (1-x^2)^{1/2}(1-y^2)^{1/2} r\cos\phi - 1\Bigg] \\ = \sum_{k=0}^{n}\sum_{m=0}^{k} & c(m,k,\alpha,\beta,n) f_{n,m,k}(x) f_{n,m,k}(y) \\ &\cdot P_m^{(\alpha-\beta-1,\beta+k-m)}(2r^2-1) r^{k-m} P_{k-m}^{(\beta-1/2,\beta-1/2)}(\cos\phi) \end{aligned} \qquad (16)$$

$$ f_{n,m,k}(x) = (1-x)^{(k+m)/2}(1+x)^{(k-m)/2} \cdot P_{n-k}^{(\alpha+k+m,\beta+k-m)}(x) \qquad (17)$$

Krawtchouk polynomials when $p = {}^1/_2$ are the zonal spherical harmonics on the space consisting of the vertices of the unit cube in N-space and there is a corresponding addition formula. The vertices of the unit cube can be considered as a message of zeros and ones, and Krawtchouk polynomials have been shown to play an important role in coding theory. *See* INFORMATION THEORY; ORTHOGONAL POLYNOMIALS.

Special values of the argument. A number of hypergeometric functions can be evaluated as quotients of gamma functions when the argument z takes on special values. Gauss' sum, Eq. (18), occurs often

$$ {}_2F_1(a,b;c;1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \qquad c > a+b \tag{18} $$

in applications. For the generalized hypergeometric series, conditions have to be placed on the parameters before the series can be explicitly summed. Two general classes of sums often arise. One, called well-poised, occurs when $p = q + 1$ and the parameters can be paired so that $a_1 + 1 = a_2 + b_1 = \cdots = a_{q+1} + b_q$. The general well-poised ${}_3F_2$ has been summed at $z = 1$. A series is called very well-poised if it is well poised and $a_2 = b_1 + 1$. In the second type, called k-balanced, $p = q + 1$, $a_1 + \cdots + a_{q+1} + k = b_1 + \cdots + b_q$ for some integer k, and one of the a_i's is a negative integer The l-balanced ${}_3F_2$ has been summed at $z = 1$. The most complicated sum of this type which has been discovered is the sum of the very well-poised, 2-balanced ${}_7F_6$ at $z = 1$.

These sums, and related transformation formulas, are fundamental and occur in many applications. Among these are complex spectra, symmetries of the Clebsch-Gordan coefficients, and the fluctuation theory of random walks. *See* ANGULAR MOMENTUM; STOCHASTIC PROCESS.

Generalizations. There are many generalizations of hypergeometric functions which are also very useful.

Basic hypergeometric functions.. The study of these functions started with Euler in the 1740s. From a power-series point of view, these are series (15) with the term ratio a_{n+1}/a_n a rational function of q^n, where q is a fixed parameter. This is equivalent to replacing the shifted factorial in Eq. (1) or Eq. (14), by the quantities in Eq. (19), and introducing appropriate

$$ (q^a)_{q,n} = (1-q^a)(1-q^{a+1})\cdots(1-q^{a+n-1}) \tag{19} $$

powers of q as multiplicative factors. The resulting series are connected with theta functions, and with a pair of formulas which are important in number theory and combinatorial analysis. One of these is Eq. (20)

$$ \sum_{n=0}^{\infty} \frac{qn^2}{(q)_{q,n}} = \frac{1}{\prod\limits_{n=0}^{\infty}(1-q^{5n+1})(1-q^{5n+4})} \tag{20} $$

$$ \text{where} \qquad \prod_{n=0}^{\infty} a_n = a_0 a_1 \cdots a_n \cdots $$

is an infinite product. In terms of partitions of positive integers as the sum of positive integers, this says that the number of ways of writing a positive integer as the sum of positive integers so that the difference between any two of the parts is at least two is equinumerous with the number of ways of writing the same integer as the sum of positive integers so that each of the parts is either one more or one less than an integer divisible by five. An infinite number of sums similar to Eq. (19) have been found, but the sums are now multiple sums. *See* COMBINATORIAL THEORY; ELLIPTIC FUNCTION AND INTEGRAL.

The coefficients $[(q)_{q,n}]/[(q)_{q,k}(q)_{q,n-k}]$ arise when counting the subspaces of dimension k of an n-dimensional vector space over a field of q elements. This property was used to obtain some identities for q-hypergeometric functions. The results were previously known, but these observations opened up new fields for applications of this generalization of hypergeometric functions.

Since the 1970s there have been notable developments in the study of basic hypergeometric functions. These functions have arisen in the study of the hard hexagon model in statistical mechanics. They are used to study and develop codes, in a discrete setting corresponding to the occurrence of Legendre functions on the surface of a sphere. They also occur in the study of quantum groups. These are Hopf algebras rather than groups, but some examples have enough structure so that detailed calculations can be done in the same way as for Lie groups. *See* LIE GROUP.

One example of the type of formula which occurs is a noncommutative extension of the binomial theorem. The binomial theorem is Eq. (21), where the

$$ (x+y)^n = \sum_{k=0}^{n} = \binom{n}{k} x^{n-k} y^k \tag{21} $$

binomial coefficient is given by Eq. (22). A noncom-

$$ \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{22} $$

mutative extension of Eq. (21) assumes that $yx = qxy$, $qx = xq$, and $qy = yq$. Then the expansion (21) becomes Eq. (23), with the q-binomial coefficient a

$$ (x+y)^n = \sum_{k=0}^{n} \binom{n}{k}_q x^{n-k} y^k \tag{23} $$

polynomial in q with integer coefficients. These coefficients are given by Eq. (24).

$$ \binom{n}{k}_q = \frac{(q)_{q,n}}{(q)_{q,k}(q)_{q,n-k}} \tag{24} $$

The coefficients in the polynomial

$$ \binom{n}{k}_q $$

count the number of inversions needed to take a given pattern to a fixed one. For example, there are 6 ways to arrange two red and two black balls in four spots. These are rrbb, rbrb, rbbr, brrb, brbr, and bbrr. If the first is taken as the fixed pattern, then there are 2 inversions needed to take each of brrb and brbr to rrbb. The others require 1, 3, and 4 inversions respectively. Then Eq. (25) is valid.

$$\binom{4}{2}_q = \frac{(1-q^4)(1-q^3)}{(1-q)(1-q^2)}$$
$$= 1 + q + 2q^2 + q^3 + q^4 \qquad (25)$$

Functions of multiple variables and matrices. In 1880 P. Appell introduced four hypergeometric functions of two variables. These functions arise very naturally from a group theoretic point of view, and thus have possible applications in the study of elementary particles by methods employing group theory. *See* ELEMENTARY PARTICLE; GROUP THEORY.

A final generalization is to hypergeometric functions of matrices. These functions have arisen in statistical multivariate analysis and in number theory. *See* MATRIX THEORY; NUMBER THEORY.

Richard Askey

Bibliography. G. Andrews et al. (eds.), *Ramanujan Revisited*, 1988; B. Dwork, *Generalized Hypergeometric Functions*, 1990; A. Erdélyi et al., *Higher Transcendental Functions*, 3 vols., 1953–1955, reprint 1981; G. Gasper and M. Rahman, *Basic Hypergeometric Series*, 1990; J. B. Seaborn, *Hypergeometric Functions and Their Applications*, 1991; N. Ja. Vilenkin, *Special Functions and the Theory of Group Representations*, 1968, reprint 1988.

Hyperiidea

A suborder of amphipod crustaceans. Most hyperiids can be recognized by the large eyes which cover nearly the entire surface of the head. The first maxillae and especially the maxillipeds are greatly reduced in comparison to the suborder Gammaridea. In the prehensile pereiopods, the claw is formed by the fifth and sixth segments, the carpus and propodus, rather than by the sixth and seventh segments or the propodus and dactyl, as in the Gammaridea. The second and third somites of the urosome are always fused, a condition rarely found in the Gammaridea. *See* AMPHIPODA.

Biology. The Hyperiidea are exclusively pelagic and marine. They are found in all the oceans, from the surface to great depths. Most are characteristic of oceanic rather than neritic waters, although some species frequent inshore waters in the tropics. After the Copepoda and Euphausiacea, the Hyperiidea are the most abundant planktonic crustaceans.

Little is known about the habits of the hyperiids. Some species are frequently found in association with other animals. Members of the genus *Vibilia* deposit their larvae in salps and feed on the salp until they are able to swim freely. Some species of *Hyperia* and *Hyperoche* are commonly found on Scyphomedusae, clinging to the surface of the bell. The young of *Hyperia galba* have been found in the subgenital pits. *Eupronoe* also sometimes occurs with medusae.

The adult females of *Phronima sedentaria* are found in deep water in gelatinous barrel-shaped cases open at both ends. These cases are fashioned from the tests of pyrosomes. The young hatch in the cases, and they pass through several molts before leaving them to move nearer to the surface.

Other species are truly free-living. Except in tropical and subtropical seas, species of *Parathemisto*, formerly *Themisto*, are the most abundant hyperiids. *Parathemisto libellula* occurs in large numbers in Arctic waters and is an important food for ringed seals. Japanese sardines feed on *Parathemisto japonica. Phronisa semilunata*, *Phromina sedentaria*, *Streetsia chellengeri*, and *Brachyscelus crusculum* are eaten by albacore.

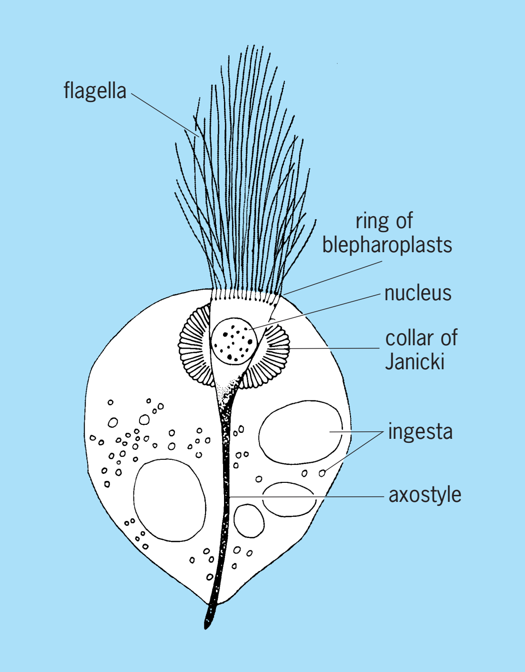
Some hyperiids are luminescent, like *Scina* and *Parapronoe crustulum*. The whole body glows with a greenish-yellow light. *See* BIOLUMINESCENCE.

Taxonomy. The suborder Hyperiidea is divided into two superfamilies, Physosomata and Genuina. Members of the Physosomata, except for a number of species of *Scina* and one species of *Lanceola*, are bathypelagic. The eyes are small or rarely absent, and the inner plates of the maxillipeds are free at the apex. The Genuina include the more familiar hyperiids with large eyes and completely fused maxillipedal inner plates, and are divided into three groups according to the structure of the male antennae. In the Recticornia, containing the genera *Vibilia*, *Paraphronima*, and *Cystisoma*, the first antennae are straight. They arise from the anterior margin of the head, and have few flagellar segments. The first antennae of the Filicornia, in such genera as *Hyperia*, *Dairella*, *Primno*, and *Phronima*, are also inserted anteriorly, and have many-segmented flagella. The antennae of the Curvicornia originate from the ventral margin of the head. This may be observed in *Eupronoe*, *Lycaea*, *Oxycephalus*, and *Platyscelus*, in which the first segment is large and curved, with the remaining segments few in number. The flagellar segments of the second antennae are folded on themselves. The Curvicornia comprise globular forms with greatly enlarged basal segments on the fifth and sixth legs as well as slender, elongate species, culminating in the rod-shaped *Rhabdosoma*. *See* CRUSTACEA.

Thomas E. Bowman

Hypermastigida

An order of the Protozoa in the class Zoomastigophorea comprising the most complex flagellates, both structurally and in modes of division. All inhabit the alimentary canal of termites, cockroaches, and woodroaches. These organisms are multiflagellate, often with complicated blepharoplast-parabasal-axostylar structures. The nucleus is single and the organisms are plastic and slow-moving, generally ovoid to elongate. Flagella occur in spiral rows, in tufts, or over the entire body. These flagellates vary from 15 to 350 micrometers in size. They may be holozoic or saprozoic. Ingestion is usually pseudopodial, in the posterior region. A high degree of adaptiveness exists, and species peculiar to one host species are not viable in another. Sexual processes



***Lophomonas blattarum*, zylophagous (wood-eating parasite) symbiont of cockroach**

and multiple fission are known, but trinary fission is most common. *Holomastigotoides* has spirochetes attached to its trophic body zone.

Lophomonas inhabits the cockroach intestine and is apparently cosmopolitan. The body is round to pyriform with a tuft of apical flagella arising from a blepharoplast ring which forms a calyx enclosing a nucleus (see **illus.**). Six to eight chromosomes form on a spindle of which the poles are centrioles. Laboratory culture is easy. *See* CILIA AND FLAGELLA; PROTOZOA; SARCOMASTIGOPHORA; ZOOMASTIGOPHOREA. James B. Lackey

Hypernuclei

Nuclei that consist of protons, neutrons, and one or more strange particles such as lambda particles. The lambda particle is the lightest strange baryon (hyperon); its lifetime is 2.6 × 10^{-10} s. Because strangeness is conserved in strong interactions, the lifetime of the lambda particle remains essentially unchanged in the nucleus also. Lambda hypernuclei live long enough to permit detailed study of their properties. *See* BARYON; ELEMENTARY PARTICLE; HYPERON; STRONG NUCLEAR INTERACTIONS.

There is no bound lambda-nucleon system, demonstrating that the lambda-nucleon force is weaker than the force between nucleons which can bind two nucleons into deuterium. The lightest bound Λ hypernucleus is ${}^{3}_{\Lambda}$H—lambda hypertriton—which is composed of a proton, a neutron, and a lambda particle. Lambda hypernuclei up to ${}^{16}_{\Lambda}$O have been identified experimentally.

Two cases of the double lambda hypernuclei have been found so far; one is ${}^{6}_{\Lambda\Lambda}$He, which is composed of two lambda particles coupled to the ^{4}He nucleus. The bound double-lambda system is still being sought. In some theoretical models of the elementary particles a strong binding for the two lambda particles is predicted.

The lambda particle in the nucleus experiences an attractive potential, the strength of which is about two-thirds that for the nucleon. The spin-orbit force, which is strong in the case of the nucleon as it causes splitting comparable to the energy differences between the nucleon shells, is negligibly small in the case of the lambda particle. Theoretically, the difference between the lambda-nucleus and nucleon-nucleus interaction is explained as reflecting the differences in the internal quark structure of the two baryons.

Sigma, xi, and omega particles, all with a lifetime of about 10^{-10} s as free particles, convert, through the strong interaction, into lambda particles in the nucleus. Nevertheless, sigma hypernuclei have been experimentally observed, the lifetime being of the order of 10^{-21} s. This lifetime is long enough to permit determination of the most important parameters of the sigma-nucleus interaction. There is a good chance that xi and omega hypernuclei could also be investigated experimentally. *See* NUCLEAR STRUCTURE. Bogdan Povh

Bibliography. D. Wilkinson (ed.), *Progress in Particle and Nuclear Physics*, vol. 5, 1981.

Hyperon

A collective name for any baryon with nonzero strangeness number *s*. The name hyperon has generally been limited to particles which are semistable, that is, which have long lifetimes relative to 10^{-22} s and which decay by photon emission or through weaker decay interactions. Hyperonic particles which are unstable (that is, with lifetimes shorter than 10^{-22} s) are referred to as excited hyperons. The known hyperons with spin $^1/_2\hbar$ (where $\hbar$ is Planck's constant divided by 2π) are Λ, Σ^-, Σ^0, and Σ^+, with $s = -1$, and Ξ^- and Ξ^0, with $s = -2$, together with the Ω^- particle, which has spin $^3/_2\hbar$ and $s = -3$. The corresponding antihyperons have baryon number $B = -1$, and opposite values of strangeness *s* and charge *Q*; they are all known empirically.

The first excited hyperon was reported in 1960. The state $\Sigma(1385)^+$ was observed as a $\pi\Lambda$ resonance in the final state of reaction (1). The symbol $\Sigma(m)$ or

$$K^- + p \rightarrow \pi^- + \Sigma(1385)^+ \rightarrow \pi^- + \pi^+ + \Lambda \qquad (1)$$

$\Lambda(m)$ indicates that the strangeness is $s = -1$ and that the isospin is $I = 1$ or 0, respectively. The superscript gives the charge and *m* is the mass in MeV; if no *m* is given, the symbol refers to the ground state, for example, Λ means $\Lambda(1115.5)$. *See* I-SPIN.

Reaction (1) is an example of an excited hyperon production reaction. Formation reactions are also possible for most excited hyperons with $s = -1$. For example, the properties of $\Lambda(1520)$ are particularly well known from observations on its formation and decay, for K^- mesons incident on hydrogen, reaction (2). $\Sigma(1385)$ cannot be formed in this way, because

$$K^-p \rightarrow \Lambda(1520) \rightarrow \begin{cases} K^- + p \\ \pi^\pm + \Sigma^\pm \\ \pi^+ + \pi^- + \Lambda, \text{etc.} \end{cases} \tag{2}$$

its mass lies below the K^-p threshold energy, $m_K + m_p \simeq 1432$ MeV.

The name hyperon may be extended to include any baryon for which one or more of the quantum members s, c, and b are nonzero, where $c =$ charm and $b =$ bottom. For example, the hyperon Λ_c^+,which has the quark structure $c(ud - du)/\sqrt{2}$, has zero strangeness but nonzero charm. The heaviest $c = 1$ hyperon in Ω_c^0, which has structure ssc and mass 2719 ± 8 MeV; it has spin-parity $\frac{1}{2}^+$ and is semistable, with decay mode (3). The heaviest $b =$ 1 hyperon is Λ_b^0, with structure $b(ud - du)/\sqrt{2}$ and mass 5640 ± 50 MeV; it has spin-parity $\frac{1}{2}^+$, like Λ and Λ_c, and is semistable, with decay mode $\Lambda_b^0 \rightarrow \Lambda_c^+\pi^-$.

$$\Omega_c^0 \rightarrow \Xi^- K^- \pi^+ \pi^+ \tag{3}$$

There is no deep distinction between hyperons and excited hyperons, beyond the phenomenological definition above. Indeed, the hyperon $\Omega(1672)^-$ and the excited hyperons $\Xi(1530)$ and $\Sigma(1385)$, together with the unstable nucleonic states $\Delta(1236)$ are known to form a unitary decuplet of states with spin $\frac{3}{2}\hbar$. *See* BARYON; ELEMENTARY PARTICLE; QUARKS; SYMMETRY LAWS (PHYSICS); UNITARY SYMMETRY.

Richard H. Dalitz

Hypersonic flight

Flight at speeds well above the local velocity of sound. In the span of a few years speeds of aircraft, missiles, and spacecraft increased tremendously, from far less than the speed of sound (subsonic) to speeds required for orbiting or escape from the Earth. By convention, hypersonic flight starts at about Mach 5 (five times the speed of sound) and extends upward in speed indefinitely. A more precise definition, attributed to T. von Kármán, states that hypersonic flight starts when the cross-flow Mach number (or Mach-number component perpendicular to the longitudinal axis) becomes 1 or greater. In any case, hypersonic flight refers to atmospheric flight at hypersonic speed when the characteristics of the flow field about the body are sensibly unchanged as the speed of the body increases further.

Conditions at high speed. Below about 350,000 ft (107 km) altitude, air is sufficiently dense to be considered a continuum and to form viscous laminar and turbulent boundary layers on body surfaces. Below this altitude, the speed of flight divided by the speed of sound—that is, the Mach number—is a characteristic aerodynamic parameter. In the realm of superaerodynamics, air density above about 350,000 ft (107 km) is so low that the mean free path between molecular collisions is large compared to body dimensions, and a new parameter known as the molecular speed ratio (ratio of flight speed to most probable speed of an incident molecule) is used.

Subsonic, transonic, and supersonic vehicles fly at speeds less than, equal to, and greater than the local speed of sound, respectively. However, when the Mach number is high, the flow field around an object exhibits a special behavior, which is worth studying separately from supersonic flight. This behavior is characteristic of hypersonic flight. For long slender bodies and for thin airfoils at small angles of attack, transition from supersonic to hypersonic flight may require a Mach number approximately equal to or greater than 5 (a flight speed of approximately 5000 ft/s or 1500 m/s). On Earth the fastest-moving natural objects seen by observers are meteorites, which travel at 25,000–250,000 ft/s (7500–75,000 m/s), or from Mach 25 to 250.

A body entering the Earth's atmosphere from space (for example, a meteorite, a ballistic reentry vehicle, or a spacecraft) has high velocity and hence large kinetic and potential energy. The kinetic energy KE is shown in Eq. (1) and potential energy PE in Eq. (2), where W is the weight, V is velocity, g is acceleration of gravity, and Z is altitude. At hypersonic speeds the potential energy of a body is small compared to its kinetic energy. For instance, the potential energy of a body at an altitude of 200,000 ft (60 km) about equals the kinetic energy of the body if it is moving at 3600 ft/s (1080 m/s).

$$\mathrm{KE} = \frac{WV^2}{2g} \tag{1}$$

$$\mathrm{PE} = WZ \tag{2}$$

During reentry, drag forces act upon a reentry vehicle or spacecraft and cause it to decelerate, thus dissipating its kinetic and potential energy. The energy lost by the reentry vehicle is then transferred to the air within the flow field around the reentry vehicle. The flow field around the forward portion of a blunt-nosed vehicle (body of revolution or leading edge of a wing) generally exhibits (1) a distinct bow shock wave, (2) a shock layer of highly compressed hot gas, and (3) a highly sheared boundary layer over the surface (**Fig. 1**). The flow field is defined as the region of disturbed air between the body surface and the shock wave. The temperature of the shock layer is so high that molecules begin to dissociate during collisions at about 2500 K, or at Mach 7. At about Mach 12, gas in the stagnation region reaches a temperature of 4000 K and becomes ionized.

In general, at high hypersonic flight speed the characteristic temperature in the shock layer of a blunted body and in the boundary layer of a slender body is proportional to the square of the Mach number. (Thus the working fluid in high hypersonic flight cannot be considered to behave as a perfect gas.) Heat energy is then transferred from the hot gases to the vehicle surface by conduction and diffusion of chemical species in the boundary layer and by radiation from the shock layer near the nose. Heat

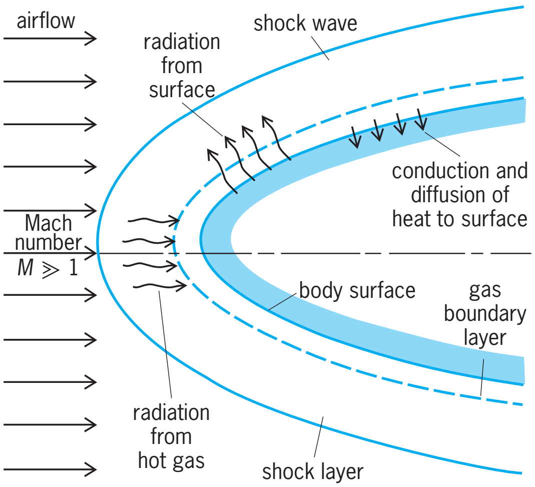


Fig. 1. Heat transfer in stagnation region of a blunted nose cap of a reentry vehicle or spacecraft.

energy is also radiated from the vehicle surface to space or to adjacent objects. One important problem confronting the designer of reentry vehicles or spacecraft is therefore to design a minimum-weight vehicle able to withstand large heat loads from adjacent hot-gas layers during reentry while retaining the ability to carry a given useful payload. *See* ATMOSPHERIC ENTRY; SPACECRAFT STRUCTURE.

Vehicular trajectories. A hypersonic vehicle in sustained powered flight maintains a constant speed (**Fig. 2**). It needs only to be initially accelerated to its operating velocity. An air-breathing engine or a rocket motor may provide the thrust to overcome aerodynamic drag, but the propulsive work done to sustain the flight speed must be dissipated as heat to the surrounding air or radiated to space. A hypersonic vehicle is in accelerating flight when the propulsive thrust is used to accelerate the vehicle as well as to overcome atmospheric drag.

There are three basic trajectories for the unpowered portion of the flight of a reentry vehicle or spacecraft reentering the Earth's atmosphere: ballistic, skip, and glide.

Ballistic trajectory. A trajectory such as that of an IRBM or ICBM is actually a segment of one of Kepler's planetary ellipses (Fig. 2). The range of the ballistic missile is shorter than that of other reentry vehicles for given initial conditions.

The short transit time of the ballistic missile in the Earth's atmosphere causes kinetic energy to be dissipated at a high rate as the reentry vehicle experiences rapid deceleration due to atmospheric drag. The effect of peak aerodynamic heating on the vehicle can be reduced if a blunt shape is employed to increase the vehicle's drag. This high-drag shape causes the vehicle to expend a large portion of its kinetic energy in pressure drag rather than in frictional forces. The effect of aerodynamic heating on the vehicle can be further reduced by the use of ablative, transpiration, shock generation and by magnetohydrodynamic cooling techniques. *See* BALLISTIC MISSILE; BALLISTICS.

Skip trajectory. If the trajectory is made up of ballistic phases alternating with skipping phases, the flight path resembles that of a stone skipping over a pond (Fig. 2). In the ballistic phase the vehicle experiences only gravitational and inertial forces; accordingly, each phase is simply a segment of one of Kepler's ellipses. During the skipping phase the vehicle experiences large aerodynamic forces, from which it develops lift. The forces tend to be so large when the reentry angle is steep that, for most practical purposes, it is permissible to neglect the gravitational force and to treat such a skipping phase as an impact. The heating effect is also more severe during the skipping process, even if ablation cooling is employed. Because of these effects a skip vehicle on a maximum-range trajectory must have a strong structure. The skip vehicle with a hypersonic lift-to-drag ratio between 1 and 4 appears to be more efficient than both the ballistic and glide types in converting high velocity into long range.

Glide trajectory. In a glide trajectory the vehicle descends and decelerates slowly through the atmosphere in such a manner that the aerodynamic lift plus inertial forces acting on the vehicle just counterbalance the weight and drag in the glide phase. For hypersonic lift-to-drag ratios that are greater than about 1.5, the range of the glide trajectory exceeds the maximum ballistic range for the same payload-to-takeoff mass ratio—that is, for the same velocity at burnout. For lift-to-drag ratios in the neighborhood of 4 and greater, the glide vehicle is comparable to the skip vehicle in its ability to convert velocity into range.

To obtain a relatively high lift-to-drag ratio, a slender shape is necessary. This low-drag configuration will in turn increase the total convective heat transfer to the vehicle because of the long flight time, but the effect is not so severe as for the skip vehicle. For high hypersonic lift-to-drag vehicles, the combination of radiative cooling, structural heat-sink techniques, and glide-trajectory control (gradient descent and deceleration) appears able to reduce skin temperatures to a level feasible for present materials. For glide vehicles with a hypersonic lift-to-drag ratio of about 2 or lower, ablation cooling is more atractive

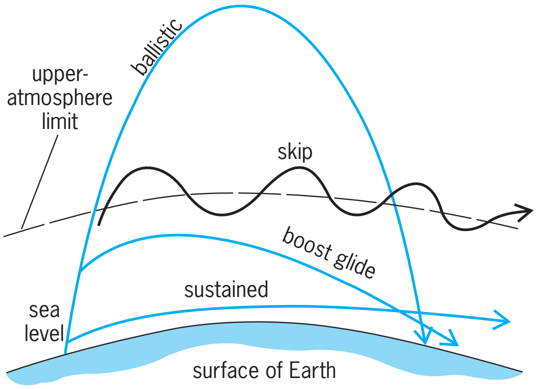


Fig. 2. Typical trajectories for hypersonic flight.

as a thermal protection technique on surface areas with high heating rates. For flight ranges on the order of the Earth's radius and greater, the performance of the hypersonic vehicle compares favorable with that of the supersonic airplane. *See* AERODYNAMICS; SUPERCRITICAL WING; TRAJECTORY. Shih-Yuan Chen

Bibliography. J. D. Anderson, *Hypersonic and High Temperature Gas Dynamics*, 2d ed., 2006; J. J. Bertin, *Hypersonic Aerothermodynamics*, 1994; W. H. Heiser and D. T. Pratt, *Hypersonic Airbreathing Propulsion*, 1994; C. Park, *Nonequilibrium Hypersonic Aerothermodynamics*, 1990.

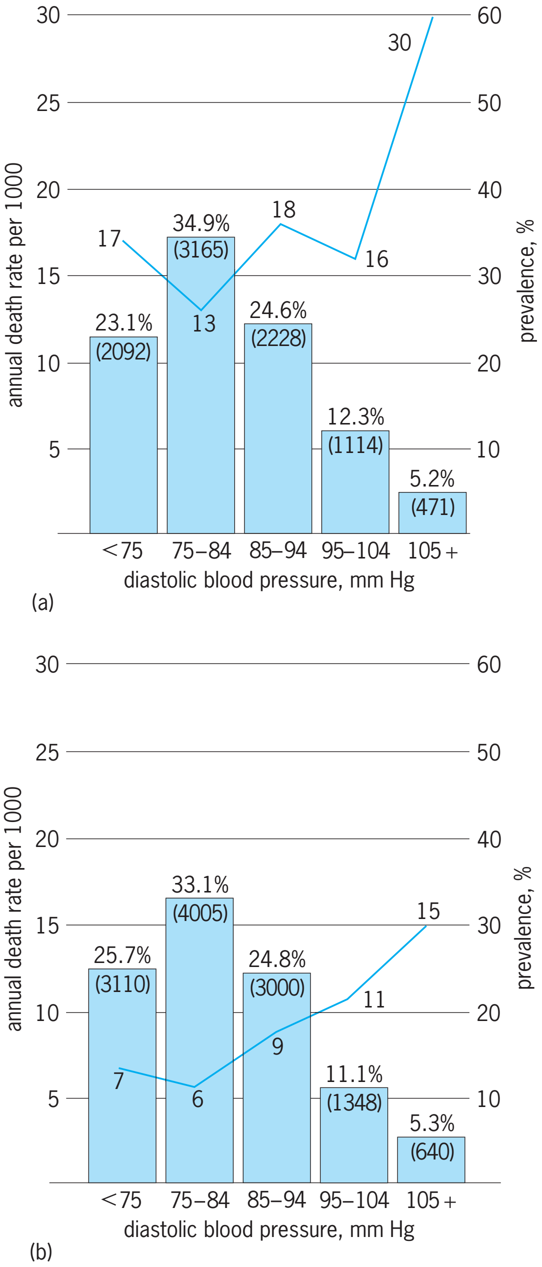
Hypertension

High blood pressure. The prevalence of this clinical disease in acculturated societies is sufficiently high to warrant its being referred to as a public health problem. Blood pressure is expressed in two numbers: the higher number is the systolic blood pressure, which is the pressure exerted by the blood against the walls of the blood vessels while the heart is contracting. The lower number is the diastolic blood pressure, which is the residual pressure that exists between heart contractions, or while the heart is relaxing. Normal blood pressure provides sufficient blood flow to the vital organs, including the brain, heart, kidneys, intestine, and skeletal muscle.

It is not entirely accurate to think of high blood pressure as a distinct disease. Epidemiological and actuarial data suggest that blood pressure is distributed unimodally within populations and is directly related to mortality (see **illus**.). Simply stated, high blood pressure appears to be both a disease and a risk factor for other diseases. At the highest end of the blood pressure distribution, there is an increased probability of premature death secondary to stroke, heart disease, or kidney failure. Lower on the distribution curve (for example, diastolic blood pressure of 90–104 mmHg, which is referred to as mild hypertension), the absolute risk of premature mortality is lower and continues to decline with further decreases in blood pressure. High blood pressure is thus a disease when its value is very high and a risk factor throughout its distribution. For diagnostic purposes, blood pressure is considered high when persistently above 140/90 mmHg.

Some cases of very high blood pressure are due to specific causes that may be surgically remediable. Most hypertension, however, results from the combination of a genetic predisposition and an environmental factor such as excessive sodium intake, sedentary habits, and stress. *See* ARTERIOSCLEROSIS; STRESS (PSYCHOLOGY).

High blood pressure can be controlled. Mild cases are treated by losing excess weight and reducing the intake of sodium and alcohol. More serious cases are treated with drugs such as diuretics, beta blockers, calcium antagonists, angiotensin-converting enzyme inhibitors, alpha blockers, and centrally acting compounds that affect regulatory centers in the brain. Treatment can usually assure a normal life. *See* CIRCULATION; HEART DISORDERS. Michael Horan



Mortality curves derived from the Framingham Heart Study show that in (*a*) men and (*b*) women aged 55–64 years, there is a correlation between diastolic blood pressure and mortality. The number of subjects examined is indicated in parentheses.

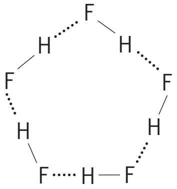
Bibliography. P. R. Hebert et al., The community-based randomized trials of pharmacologic treatment of mild-to-moderate hypertension, *Amer. J. Epidemiol.*, 127:581–590, 1988; W. B. Kannel, Role of blood pressure in cardiovascular morbidity and mortality, *Prog. Cardiovasc. Dis.*, 17:5–24, 1974; The 1988 Report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure, *Arch. Internal Med.*, 148:1023–1038, 1988; Society of the Association of Life Insurance Medical Directors and the Society of Actuaries, *Blood Pressure Study 1979*, 1980.

Hypervalent compounds

Group 1, 2, and 13–18 compounds which contain a number (*N*) of formally assignable electrons of more than eight (octet) in a valence shell directly associated with the central atom (X) in direct bonding with a number of ligands (*L*). The designation *N*-X-*L* is conveniently used to describe hypervalent molecules. *See* ELECTRON CONFIGURATION; LIGAND; VALENCE.

Compounds of main group elements in the second row (such as carbon, nitrogen, and oxygen) have eight valence electrons. As such, the fundamental shapes of their atoms are linear (such as acetylene, sp orbital), triangular, (such as ethylene, sp^2 orbital), and tetrahedral (such as methane, sp^3 orbital). In contrast, main group elements in the third row of the periodic table (such as silicon, phosphorus, and sulfur) may contain more than eight electrons in a valence shell. These are called hypervalent compounds. Fundamental shapes of *10*-X-*5* molecules (including *10*-X-*4* molecules bearing a pair of unshared electrons) are trigonal bipyramid (TBP) or square pyramid (SP), and those of *12*-X-*6* (including *12*-X-*5* bearing a pair of unshared electrons) are octahedral. Hence, there is an apparent similarity in shape between hypervalent compounds and organotransition metal compounds. *See* MOLECULAR ORBITAL THEORY.

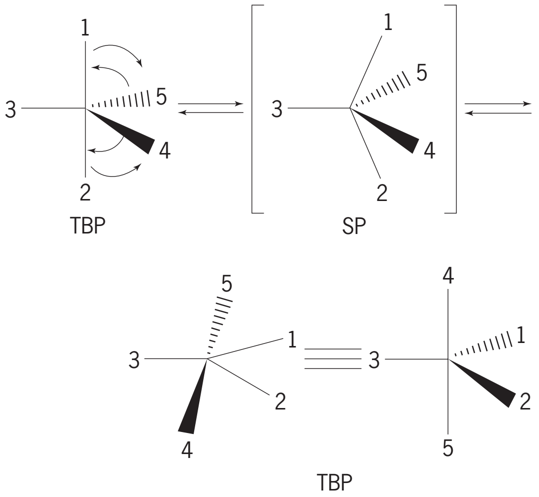
The hydrogen bond is one of the best examples of a hypervalent bond. The covalent nature of the hydrogen bond of the [N—H···N] system has recently been established experimentally and theoretically. The structure of $(HF)_5$ is pentagonal, with a bond length for F—H of 1.0 Å and H···F of 1.5 Å. The hydrogen atom shifts rapidly between the two fluorines in the range of 0.5 Å. However, $[F—H—F]^-K^+$, (*4*-H-*2*), is linear, and the two F—H bond lengths are elongated to 1.13 Å, which is typical for a hypervalent bond. *See* CHEMICAL BONDING; HYDROGEN BOND.



In order to accept extra electrons in a valence shell, an electron-rich and polarized sigma bond constitutes an apical bond (three-center, four-electron bond, which is defined as a hypervalent bond by molecular orbital theory) on the central atom of *10*-X-*5*. One of the most unstable hypervalent molecules is $[F—F—F]^-K^+$, (*10*-F-*2*). It is linear and the F—F bond is calculated to be 1.701 Å, which is elongated from that of F—F (1.412 Å) by accepting a fluoride ion. This is essentially the same as that of $[F—H—F]^-Li^+$, (*4*-H-*2*).

Novel and fundamental characteristics of hypervalent compounds are summarized as follows:

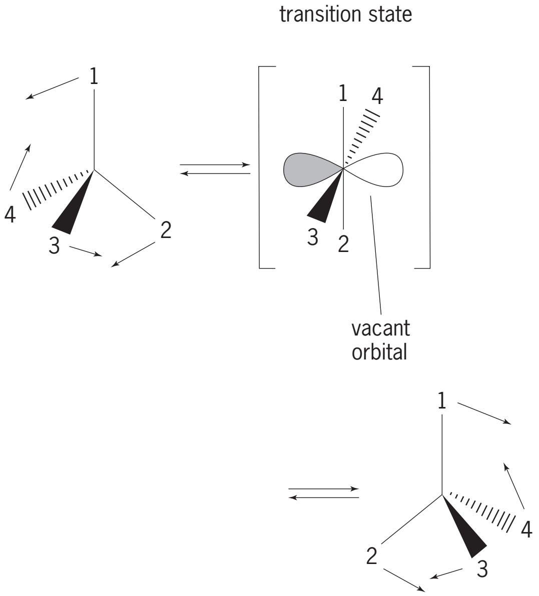
1. For TBP molecules of *10*-X-*5*, intramolecular positional isomerization between apical bonds and equatorial bonds (sp^2 orbital) is very rapid. This is demonstrated by a Berry pseudorotation, Substituents (ligands) with larger electron-withdrawing ability preferably occupy apical positions (apicophilicity).



2. Ligand exchange reactions (LER) and ligand coupling reactions (LCR) take place for hypervalent compounds of *10*-X-*5* and *10*-X-*4*. These are formally similar to LER and reductive elimination of organometallic compounds.

3. The stereochemistry of nucleophilic substitution at 4-coordinate main group compounds of higher rows is complex, unlike inversion at a carbon atom by a nucleophilic substitution reaction. Both retention and inversion occur due to the stability of intermediate *10*-X-*5* species, where rapid pseudorotation takes place to change the relative positions of substituents. *See* STEREOCHEMISTRY.

4. Edge inversion can be the mechanism of inversion for main-group element compounds in the higher-numbered rows,



This is applicable for *10*-X-*3* compounds bearing lone pair electrons. This mechanism involves the appearance of a vacant *p*-type orbital at the transition state which can be coordinated by two nucleophiles to help stabilize it. This is an essentially different mechanism from that of ammonia, which has a vertex inversion.

5. The stereochemistry of *12*-X-*6* compounds is stable in contrast to *10*-X-*5* compounds. This is composed of three hypervalent bonds. Kin-ya Akiba

Bibliography. K.-y. Akiba (ed.), *Chemistry of Hypervalent Compounds*, 1999; S. Borman, Hydrogen bonds revealed by NMR, *C&EN*, pp. 36–38, May 10, 1999.

Hyphochytriomycota

A phylum of microscopic, fungus-like organisms that reproduce with anteriorly uniflagellate zoospores (independently motile spores). In Hyphochytriomycota, two opposite rows of hairlike structures (mastigonemes) cover the flagellum. The mastigonemes are hollow and split into three parts at their tips. The zoospore swims with the flagellum directed straightforward and the tip moving back and forth, propelling the oval to pyriform zoospore forward. Functioning in dispersal, the zoospore uses stored food reserves until it reaches a substrate that it can use. After a period of swimming, between 30 min to 2 h, the zoospore rounds up, absorbs its flagellum, secretes a wall around itself, and develops into a thallus. The thallus can be a branching filament with spindle-shaped swellings and sporangia (spore-containing structures) connected by empty isthmuses (polycentric, eucarpic, **Fig. 1**); a sporangium with rootlike rhizoids (monocentric, eucarpic, **Fig. 2**); or simply an endobiotic (living in the cells or tissues of a host) sporangium (monocentric, holocarpic). [Polycentric means having a number of centers of growth and development and with more than one reproductive organ; eucarpic, using only part of the thallus for the fruit body; monocentric, having one center of growth and development; holocarpic, using the entire thallus for the fruit body.] The thallus is the growth and heterotrophic phase, acquiring nutrients by absorption. As the thallus grows, the sporangium becomes multinucleate. When mature, the sporangial cytoplasm cleaves out uninucleate, uniflagellate zoospores. Zoospore formation can occur within the sporangium, but in some genera, such as *Rhizidiomyces*, zoospores are formed externally. A long discharge tube grows from the sporangium (Fig. 2) as the sporangial protoplast extrudes over about a 15-min period and comes to lie as a pulsating mass at the tip of the exit tube. Zoospores are then cleaved, swarm as a mass for about 5 min, and eventually swim away. Hyphochytrids also produce resistant sporangia that aid in surviving adverse conditions. Most resistant sporangia are produced asexually, and there have been only a few reports of sexual reproduction in this group. *See* EUMYCOTA; FUNGI; MASTIGOMYCOTINA; REPRODUCTION (PLANT).

Taxonomy and phylogenetic relationships. Three families are generally recognized and are based on thallus type: Hyphochytriaceae (polycentric, eucarpic), Rhizidiomycetaceae (monocentric, eucarpic), and Anisolpidiaceae (monocentric, holocarpic). Taxonomists debate the classification of

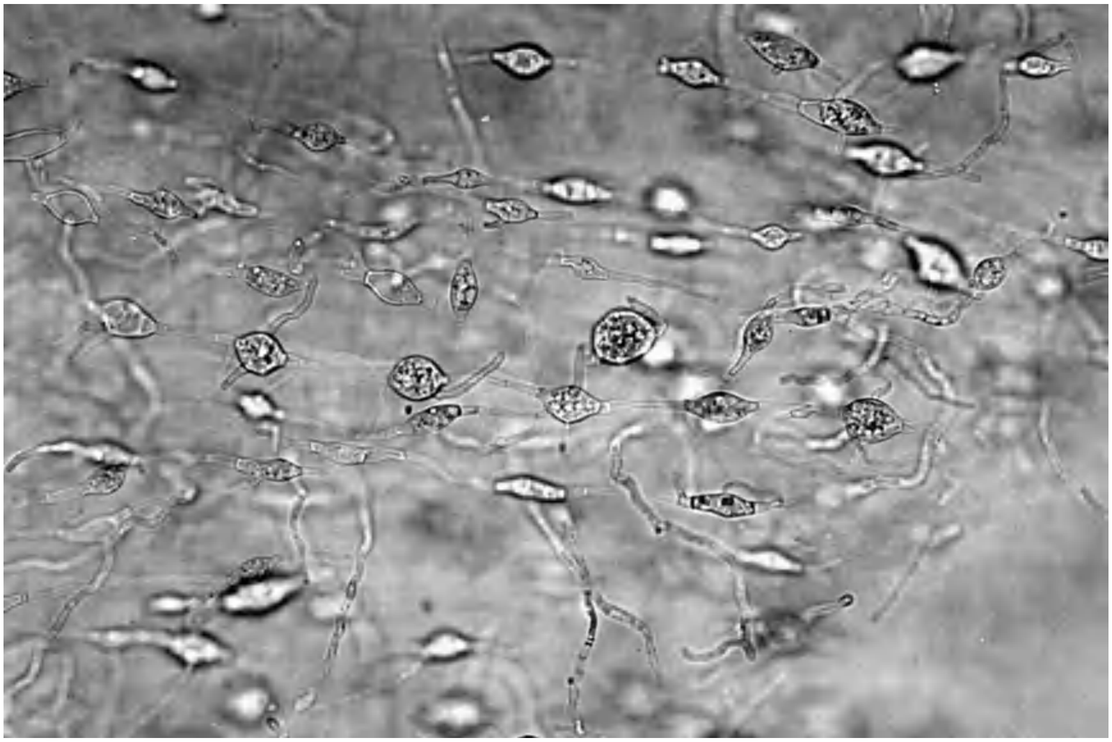


Fig. 1. Polycentric, filamentous thallus of *Hyphochytrium catenoides*. Fusiform swellings, separated by empty isthmus cells, will round up and become sporangia that later produce anteriorly uniflagellate zoospores.

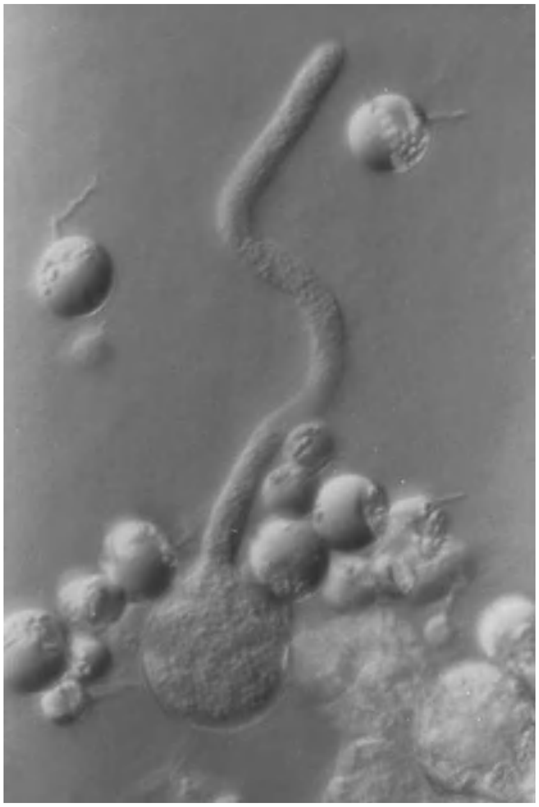


Fig. 2. Eucarpic, monocentric thallus of *Rhizidiomyces apophysatus* with a long, meandering exit tube through which the sporangial protoplast will extrude and cleave externally into zoospores. Recently encysted zoospores with threadlike rhizoids surround the mature thallus.

genera in this phylum, which encompasses approximately 29 described species in five to seven genera. There have been only limited molecular phylogenetic analyses of genes in this group, and it is not known whether classification based on thallus structure results in natural groupings or whether all taxa included in the phylum are actually related. Biochemical and molecular phylogenetic studies, however, do show that the hyphochytrids, *Hyphochytrium* and *Rhizidiomyces*, are closely related.

Hyphochytrid thallus forms parallel to those of the zoosporic fungi, Chytridiomycota, but hyphochytrids are actually more closely related to photosynthetic (phototrophic) and nonphotosynthetic (heterotrophic) organisms classified in the kingdom Straminopila or the monophyletic supergroup Chromalveolata. Hyphochytrids are clearly related to the water molds (Oomycetes, Peronosporomycetes) and to heterokont (having unequal flagella) algae, forming a monophyletic group in analyses of ribosomal sequences. Hyphochytrids share a number of biochemical characteristics with their biflagellate relative, the Oomycetes. Their cells walls contain cellulose, their mitochondrial cristae are tubular, and they produce lysine by the diaminopimelate (DAP) pathway. However, hyphochytrids are still distinctive. Some contain chitin as well as cellulose in their cell walls. Unlike Oomycetes, nuclear division in hyphochytrids is not totally closed, but fenestrae form at the poles of the nuclear envelope during mitosis and intranuclear vesicles coalesce around condensed separating chromatin, reconstituting new nuclear envelopes. *See* CHYTRIDIOMYCOTA; OOMYCOTA.

In phylogenetic analyses of ribosomal genes, the sequence of divergences among groups in the stramenopiles is controversial. However, most analyses reveal that the heterotrophic members are basal to the more derived phototrophic stramenopiles (such as diatoms, brown algae, chrysophytes, and yellow-green algae). The hyphochytrids and oomycetes are sister to each other (also known as pseudofungi) and cluster with other basal heterotrophic stramenopiles, including *Developayella elegans*, bicosoecids, thraustochytrids, labyrinthulids, and opalinids.

Zoospores. The most distinctive feature of the hyphochytrids is the ultrastructure of its zoospore (**Fig. 3**). Although its relatives have two flagella, one smooth and the other tinsel (covered with lateral filaments), hyphochytrid zoospores bear a single anterior tinsel flagellum (with tubular tripartite mastigonemes). The transition region of the flagellum resembles that of the oomycete *Saprolegnia*, consisting of a set of five to seven rings and a zone of struts perpendicular to the flagellar microtubules. The flagellum extents from a kinetosome (basal body), and a secondary centriole is positioned at a 150° angle. The backward projection of the secondary centriole suggests loss of the posteriorly directed flagellum and an earlier divergence from a

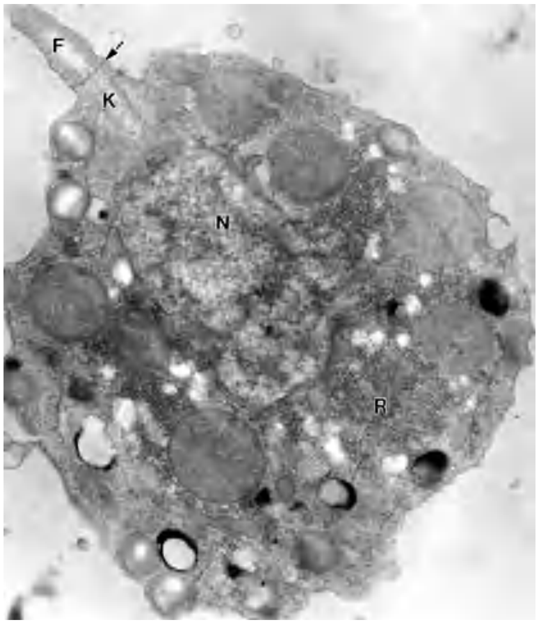


Fig. 3. Transmission electron microscopy image of a longitudinal section of a zoospore of *Rhizidiomyces apophysatus*. The flagellum (F) extends from the kinetosome (K). A cluster of coarsely granular ribosomes (R) surround the posterior portion of the single nucleus (N). The arrow points to the basal plate, with the transition zone located more anterior, with stacked rings (transitional helix) visible in longitudinal section as two stacks of dots.

biflagellate ancestor shared with oomycetes. With the loss of the flagellum, there was a loss of a root system. In the hyphochytrid zoospore, three of the four systems of microtubular roots found in oomycetes extend from the region around the kinetosome. The zoospore contains a single nucleus with ribosomes clustered over the posterior portion.

Importance. Despite being a relatively small group in terms of number of named species, hyphochytrids are ecologically diverse. Hyphochytrids can be found in soils and in freshwater and marine habitats as decay-causing organisms that break down plant and insect parts or as parasites of algae and fungi. As common parasites of algae in freshwater and marine environments, up to 50% of a population has been reported to be infected with hyphochytrids. *Hypochytrium catenoides* is weakly parasitic on root hairs of corn, and in soil it may be a component of suppressive soils inhibitory to the root rot oomycete, *Phytophthora cinnamomi*. Because hyphochytrids can parasitize spores of oomycetes and endomycorrhizae, they may have a role in natural control of microbial populations. Recent analyses of molecular sequences from marine environmental samples have revealed novel lineages closely related to hyphochytrids, and the prevalence of cryptic related taxa suggests hyphochytrids may be more diverse and more vital in microbial food webs than previously recognized.

Martha J. Powell

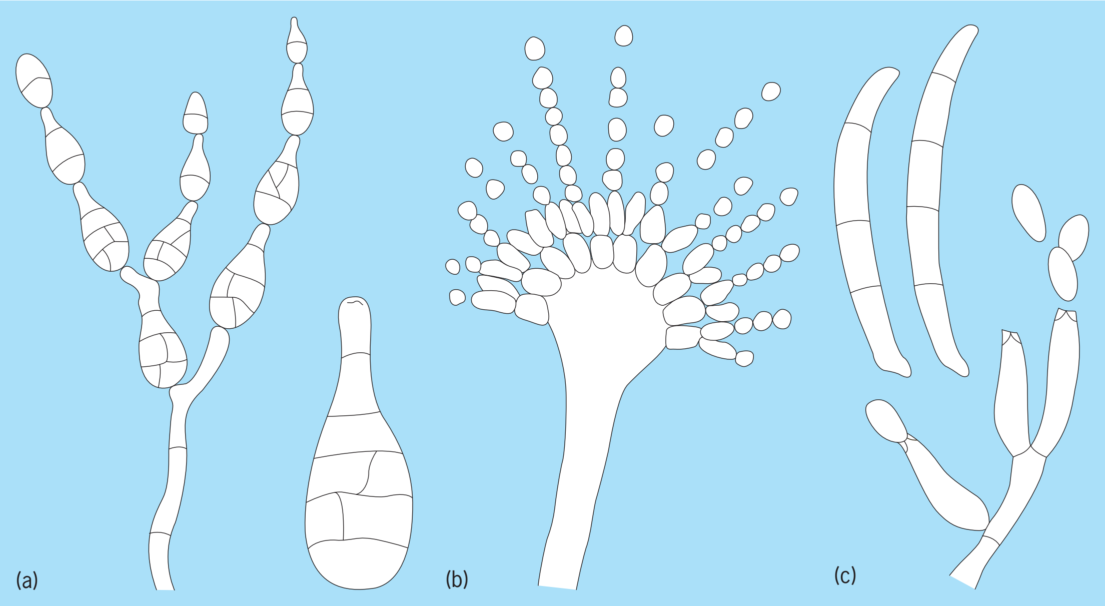
Bibliography. C. J. Alexopoulos, C.W. Mims, and M. Blackwell, *Introductory Mycology*, 4th ed., 1996; M. W. Dick, *Straminipilous Fungi*, 2001; M. S. Fuller, Phylum Hyphochytriomycota, in L. Marqulis et al. (eds.), *Handbook of Protoctista*, pp. 380–387, 1990; J. S. Karling, *Chytridiomycetarum Iconographia*, Lubrecht & Cramer, Monticello, NY, 1977; F. K. Sparrow, *Aquatic Phycomycetes*, 2d rev. ed., University of Michigan Press, Ann Arbor, 1960.

Hyphomycetes

Fungi that form spores asexually (via mitosis) and that produce their spores externally on threadlike hyphae (as opposed to bearing them inside a fruiting body). In traditional taxonomy, Hyphomycetes was a formal class, but because this group is nonphylogenetic (it is composed of more than one evolutionary lineage), many mycologists now either attach a prefix, as in "form-class Hyphomycetes," or use the term colloquially, as in "hyphomycetes." The mitotic spores of hyphomycetes (as well as those of coelomycetes) are called conidia. *See* AGONOMYCETES; COELOMYCETES; DEUTEROMYCOTINA; FUNGI.

Taxonomy and identification. Most hyphomycetes are life stages of ascomycetes, but some are related to basidiomycetes; for example, *Ingoldiella* is a relative of some of the jelly fungi, and *Nematoctonus* is closely related to a genus of mushrooms. Both these genera have clamp connections, buckle-shaped structures common in the hyphae of many basidiomycetes. However, most hyphomycetes are phylogenetically ascomycetes. For example, some *Penicillium* species are anamorphic (asexual) stages in the teleomorphic (sexual) ascomycete genus *Talaromyces*. Other species of *Penicillium* have sexual stages in a different ascomycete genus, *Eupenicillium*, or have no known sexual stage at all. Some hyphomycetes may appear to lack a sexual state simply because the teleomorph is as yet undiscovered, but in other instances the sexual state may no longer exist.

Hyphomycetes can be further divided on the basis of gross morphology, color and shape of conidia, and characteristics of their conidiogenous (conidium-producing) cells. Hyphomycetes which produce conidiophores (conidium-bearing hyphae) aggregated into cushion-shaped structures called sporodochia are in the form-order Tuberculariales. Some hyphomycetes have conidiophores aggregated into stalk- or club-shaped structures (synnemata), or head-shaped structures (coremia). These latter hyphomycetes belong to the form-order Stilbellales. The hyphomycetes whose conidiophores occur singly, or at least are not aggregated in the above manners, are in the form-order Moniliales. None of these form-orders is phylogenetic, but all are useful for identification. Moniliales has been further divided into dark- (dematiaceous) and light-spored groups, which themselves can be subdivided on the basis of conidial shape and number of septa (cell-dividing walls) in the conidia. All of these categories represent classifications of convenience, and some fungi display characters intermediate between these categories. For this reason, it has been very useful to utilize attributes of the conidiogenous cells and formation of conidia. Some conidia are produced through the tips of bottle-shaped cells called phialides (enteroblastic conidiogenesis), whereas others are produced from swellings of the conidiophore wall (holoblastic conidiogenesis). Sometimes conidia are produced through holes in a conidiophore wall (tretic conidiogenesis) or by formation and dissolution of many cross-walls at the apex of a conidiophore (thallic conidiogenesis). There are many variations of the above, including the direction and pattern of conidiogenesis (acropetal, basipetal, sympodial, etc.). This system of identification and classification by conidiogenesis, initially introduced by the mycologist S. J. Hughes, is the basis of modern identification of hyphomycetes. Identification efficiency is increased when characters from conidiogenesis are combined with information on spore color, septation, and ornamentation (spines, warts, grooves, etc.). Increasingly, identification is enhanced by use of molecular methods such as random amplified polymorphic DNAs (RAPDs), restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), DNA sequence determination (often obtained via polymerase chain reaction, PCR), or metabolic profiles.



Three hyphomycete genera. (*a*) *Alternaria*. (*b*) *Aspergillus*. (*c*) *Fusarium*.

See ASCOMYCOTA; BASIDIOMYCOTA; FUNGAL GENETICS; GENE AMPLIFICATION.

Cultivation and preservation. Most hyphomycetes can be cultivated on artificial media in the laboratory, a distinct asset in diagnosis and identification. In fact, cultivation under standardized conditions is a prerequisite for accurate identification of species in several genera such as *Alternaria*, *Aspergillus*, *Fusarium*, *Penicillium*, and others (see **illustration**). Production of conidia, a prerequisite for successful identification, is usually enhanced on media of low sugar content and sometimes by exposure to periodic fluorescent or near-ultraviolet light. Occasionally, it is necessary to experiment with plant materials, sterilized in an autoclave, by radiation, or by exposure to propylene oxide, and used as an amendment to the medium. For example, radiation-sterilized carnation leaves are used to promote formation of sporodochia by *Fusarium* isolates. Hyphomycetes, like other fungi, are best preserved in an inert (nongrowing) condition. One popular method is mixing conidia into vials with a dilute solution of glycerol and placing the vials into liquid nitrogen vapor or into an ultralow-temperature freezer [approximately −80°C (−112°F)]. Lyophilization (freeze-drying) in hermetically sealed vials can be used to preserve some species. Less technologically intensive approaches include putting conidia in silica gel or sterile soil, or drying them on filter paper, followed by storage in a conventional freezer [approximately −20°C (−4°F)]. Some laboratories prefer to excise several small (1–2 mm^3) cubes from an actively growing culture and store them at room temperature in vials of sterile water. Some fungi will keep well when grown on "slants" (test tubes, filled partly with medium and placed at an angle so that the medium hardens on a slant), then covered with sterile mineral oil and stored in a refrigerator [approximately 4°C (39°F)]. Whichever method is adopted, it is advisable to assess viability at periodic intervals by reculturing the fungi onto fresh artificial media. Although many fungi (including hyphomycetes, whose spores are often easily dispersed in air) are innocuous, some are pathogenic or allergenic to humans, so it is advised that manipulation of fungi take place in a biological safety cabinet or in a transfer box. *See* CRYOBIOLOGY.

Biology, economic importance, control, and utilization. In general, sexual stages of fungi provide for recombination of genes (via meiosis) and often serve to initiate new growth cycles in the spring. The subsequent asexual stages produce more numerous spores; hence they function strongly in the spread to new hosts, substrates, or environments. For example, Fusarium head blight, a disease of small grains (predominantly wheat and barley), is often initiated by ascospores from the sexual stage *Gibberella zeae*. Growth resulting from ascospores can produce conidia of the hyphomycetous stage, *Fusarium graminearum*, further fueling the epidemic. Species lacking a sexual cycle (such as *F. oxysporum*) may overwinter as resistant structures (chlamydospores, sclerotia) or by inhabiting buried debris. Some dark-spored hyphomycetes (the dematiaceous hyphomycetes) have conidia with thick melanized walls and contain ample nutrient reserves. In such cases, the conidia themselves are effective survival structures, as well as functioning in dissemination.

Numerous genera of hyphomycetes are well known as plant pathogens. Prominent examples include *Alternaria* (on many plants but especially fruits and vegetables), *Botrytis* (a serious problem on various fruits, vegetables, and cut flowers), *Curvularia* (often on grasses), *Cylindrocarpon* (a relative of *Fusarium*, and causing leaf spots, and shoot, stem, and root diseases of miscellaneous plants), *Fusarium* (on many plants but especially cereal grains), *Penicillium* (on grains, fruits, and vegetables), *Ramularia* (and its relatives *Cercospora*, *Cercosporella*, and *Pseudocercosporella* on miscellaneous hosts), *Stemphylium* (several species on forage legumes), and *Verticillium* (on various plants, with one species attacking cultivated mushrooms). Some of these pathogens are very host specific. For instance, *Fusarium oxysporum* f. sp. *cepae* is specialized to attack host plants in the genus *Allium* (onion and garlic). Others are opportunists with a broad host range, such as *Cladosporium herbarum*, generally a saprophyte (an organism living on dead or decaying organic matter) but also attacking ripe or overripe fruit. Some genera have very few pathogenic members. For example, *Trichoderma*, a generally saprophytic genus, is often used in biological control of plant diseases, but one species causes disease in cultivated mushrooms. Disease management for these and other fungi is most often exerted by combining agronomic practices with use of fungicides. *See* AGRONOMY; PLANT PATHOLOGY.

Some hyphomycetes cause disease in humans or animals. Especially well known are *Epidermophyton*, *Microsporum*, and *Trichophyton*; the first is known as the cause of athlete's foot, and the latter as agents of ringworm. *Blastomyces*, *Coccidioides*, *Histoplasma*, and *Paracoccidioides* are hyphomycetes causing serious diseases in humans. Several dematiaceous hyphomycetes are the cause of chromoblastomycosis (a skin disease characterized by warty nodules that may ulcerate); mycetoma (a chronic infection, usually of the feet, resulting in swelling), and other diseases in both normal and immunocompromised hosts. In fact, many other hyphomycetes (for example, *Acremonium*, *Chrysosporium*, *Fusarium*, and *Paecilomyces*) are well documented in clinical situations. Some species of *Aspergillus*, especially *A. fumigatus*, can cause respiratory diseases. Various species in *Alternaria*, *Aspergillus*, *Cladosporium*, and *Penicillium* are documented as allergens affecting a high proportion of humans, but the hyphomycete which has most captured the public imagination is *Stachybotrys*, one of the agents of sick building syndrome. Several of these same fungi, especially species in *Aspergillus*, *Fusarium*, and *Penicillium*, are well documented as producing mycotoxins in human or animal foods. Animals can also be affected by allergenic or toxigenic fungi in forage, moldy hay, or silage (fodder). Facial eczema in sheep is caused by a species of *Pithomyces*. Mycotic abortion in cows and heaves in horses are attributable to adverse effects of toxigenic hyphomycetes, especially *Aspergillus fumigatus*. Control of invasive fungal diseases is most often with antifungal compounds, sometimes combined with surgery. Indoor allergy problems are best resolved by structural improvements which deprive fungi of moisture. Quality control for production of food, feeds, and fodder is necessary to protect humans and animals from adverse effects of fungal contamination. *See* ALLERGY; MEDICAL MYCOLOGY; MYCOTOXIN.

Hyphomycetes also play a significant role in degradation or discoloration of manufactured products such as paper, textiles, building materials (including wood, stone, rubber, and plastic), and fossil fuels. Notable examples are *Hormoconis resinae* in diesel fuel, *Aureobasidium pullulans* on some plastics, *Phialophora* species in pulp and paper, and *Epicoccum* on stone. Modern building materials can be impregnated with a variety of fungi-resistant compounds, but restoration of deteriorating monuments and artifacts presents multiple challenges to conservators.

Although some hyphomycetes are undesirable, many species undoubtedly play beneficial roles as saprophytes in decomposition of plant residues and in nutrient recycling. Fungal mycelia are also an important component in soil structure. Many hyphomycetes are opportunistic phytopathogens (causing disease only on weakened or senescent plants) and are otherwise beneficial saprophytes. Other species cause disease on weeds (for example, *Myrothecium* on kudzu vine) or arthropod pests (for example, *Beauveria* on whiteflies and *Metarhizium* on locusts), or are used to counteract phytopathogenic fungi (for example, *Trichoderma* against *Rhizoctonia*, *Pythium*, and wood-rotting fungi). *Arthrobotrys* species are predators of nematodes. Not only are hyphomycetes increasingly useful in biological control, but several hyphomycetes play major roles in biotechnology. *Aspergillus niger* is used commercially to produce cellulases and citric acid, and *Aureobasidium pullulans* for the production of pullulan, a food additive. *Aspergillus oryzae* produces proteases used in the manufacture of breads and cheeses. Lipases for food manufacture are derived from some *Penicillium* species, and some *Penicillium* species have played crucial roles in the production of penicillin antibiotics. Perhaps the most famous example of fungi in food biotechnology is the manufacture of soy sauce with the hyphomycete *Aspergillus sojae*. *See* FUNGAL BIOTECHNOLOGY; FUNGAL ECOLOGY.

Frank M. Dugan

Bibliography. H. L. Barnett and B. B. Hunter, *Illustrated Genera of Imperfect Fungi*, 1998; J. W. Carmichael et al., *Genera of Hyphomycetes*, 1980; G. S. de Hoog et al., *Atlas of Clinical Fungi*, 2d ed., 2000; F. M. Dugan, *The Identification of Fungi*, 2006; E. Kiffer and M. Morelet, *The Deuteromycetes*, 2000; J. R. von Arx, *The Genera of Fungi Sporulating in Pure Culture*, 1981.

Hypnales

An order of the true mosses (subclass Bryidae). Also known as the Hypnobryales, this order consists of 14 families and some 135 genera, primarily put together because they share a hypnoid peristome (as described below). The families show considerable sporophytic unity; it is gametophyte structure that determines family membership.

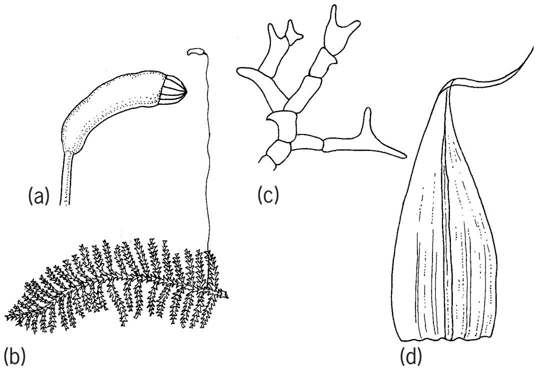


Fig. 1. *Thuidium recognitum*. (*a*) Urn and peristome. (*b*) Portion of plant. (*c*) Paraphyllium. (*d*) Perichaetial leaf. (*After W. H. Welch, Mosses of Indiana, Indiana Department of Conservation, 1957*)

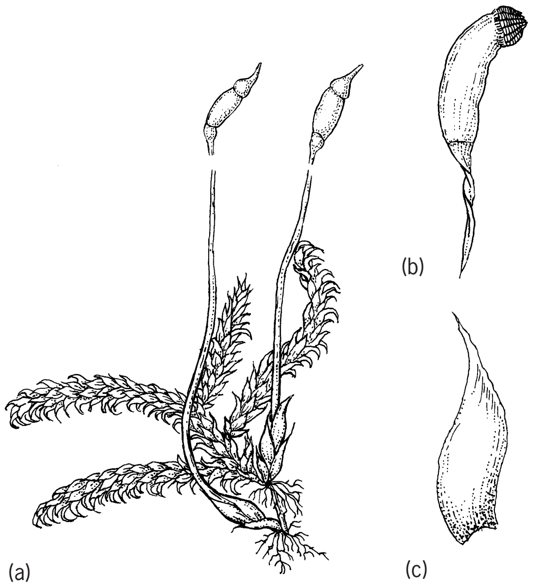


Fig. 2. *Hypnum reptile*. (*a*) Portion of plant, with stems shortened. (*b*) Urn and peristome. (*c*) Faintly bicostate leaf. (*After W. H. Welch, Mosses of Indiana, Indiana Department of Conservation, 1957*)

Hypnales are often known as feather mosses owing to their freely branched stems, with branches regularly or irregularly arranged in two rows. The plants form mats, especially in woodlands, with pseudoparaphyllia and paraphyllia often present. The leaves are most commonly acute or acuminate, and the costa is generally short and double or nearly lacking (**Figs. 1** and **2**). The upper cells are mostly elongate and smooth, but they may be variously papillose, and the alar cells are often differentiated. The sporophytes are lateral with elongate, smooth setae, and the capsules are generally inclined and asymmetric with well-developed double peristomes. The peristome generally has exostome teeth abruptly tapered, cross-striate below, bordered, and trabeculate, and endostome segments keeled, alternating with cilia, and rising from a high basal membrane. The calyptrae are cucullate and usually naked. The chromosome numbers range from 5 to 22, often in polyploid series. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA.

Howard Crum

Hypnosis

A presumed altered state of consciousness in which the hypnotized individual is usually more susceptible to suggestion than in his or her normal state. In this context, a suggestion is understood to be an idea or a communication carrying an idea that elicits a covert or overt response not mediated by the higher critical faculties (that is, the volitional apparatus).

Effects. Hypnosis cannot be physiologically distinguished from the normal awake state of an individual, and for this reason its existence has been questioned by some investigators. There are few phenomena observed in association with hypnosis, if any, that are specific to the hypnotic state. Most are directly or indirectly produced by suggestions. It is usual to assign a depth or degree to hypnosis that is proportional to the extent to which the presumably hypnotized person is found to be suggestible. Through suggestions given to hypnotized individuals, it is possible to induce alterations in memory, perception, sensation, emotions, feelings, attitudes, beliefs, and muscular state. Such changes can be, and usually are, incorporated into the complex behavior of the individual, resulting in amnesias and paramnesias, fuguelike conditions, paralysis, loss of sensory functions, changes in attention, personality alterations, hallucinatory and delusional behavior, and even physiological changes. Enhanced recall is sometimes possible. Although sometimes remarkable, the effects produced through hypnosis with the majority of individuals are much less spectacular than popularly believed.

Many of the phenomena seen in association with hypnosis and suggestion can be observed to occur in situations that are not obviously ones involving hypnosis or suggestion. Some investigators have viewed this as indicating that there is no hypnosis or suggestion, whereas others have interpreted this as indicating that all these situations effectively involve hypnosis or suggestion. However, there is just as much reason to conclude that neither is the case, and that the effects in question are simply not unique to hypnosis and suggestion. Hypnosis and suggestion are then simply convenient means for producing these effects.

Theories. The theories of hypnosis attempt to explain its essential nature. They all are inadequate in some respects. These theories, arranged approximately in order of their chronological development, adduce that hypnosis is: a form of subtle emanation between the hypnotist and subject; a modified form of sleep; a form of dissociation; a conditioned response resulting in ideomotor action (that is, nonvoluntary movement resulting from some idea); a form of role playing or goal-directed activity; a form of regression to a level representing the relation between parent and child; a form of regression in which the subject assumes the submissive role similar to that assumed by the female in the sexual role; a form of interpersonal relationship between the subject and hypnotist, but not necessarily of the kinds represented in the theories of regression. Finally, one theory, which is an extension of the theory of role playing, takes the position that there is no state of hypnosis, and that all hypnotic phenomena can be accounted for without this concept; this view is held by only a small number of investigators. Although the foregoing statements represent an oversimplification, they do give some idea concerning the nature of the theories.

Applications. The applications of hypnosis to the field of medicine and psychology are exceedingly wide in scope.

Diagnostic tool. Hypnosis has been employed successfully in distinguishing functional from organic disorders. For example, functional deafness may be distinguished from organic deafness; hysterical conversion paralysis may be differentiated from organic paralysis; psychogenic convulsions may be contrasted with organic convulsions. Hypnosis may also be used in determining a normal baseline for the functioning of certain endocrine glands. Anxiety may bring about dysfunction of the thyroid; consequently a metabolic or other test would not distinguish between a pathological and a functional disorder. The anxiety can be reduced by hypnotic relaxation, and a comparative relationship can be established between the output produced by an anxiety component and possible pathology.

Hypnosis can also be used in teasing out the part that a functional overlay plays in what appears to be an organic condition. For example, in a limb in which there is a partial paralysis due to any organic cause, there may be superimposed a further dysfunction due to psychological causes that can be demonstrated by hypnotic suggestion.

Aid in examination. Hypnosis can be used in controlling pain in certain types of examination, as in cystoscopy, and leaves the person able to assist the physician by giving information concerning the location of the instrument. It has an advantage over analgesics or anesthetics in such situations. Hypnosis may be useful in nerve block infiltrations and in making spinal punctures. It can also be employed successfully to reduce gagging when stomach tubes are inserted, and has been utilized in the control of bleeding, salivation, and muscle spasms.

Anesthesiology and surgery. Hypnosis has been used effectively for the control of pain. In cases where an analgesic or anesthetic is inadvisable, it may be used in place of, or in conjunction with, chemoagents. Hypnosis may also be used where tolerance becomes a factor in long-term cases. Pain control in all kinds of surgical operations is dependent on the susceptibility of the hypnosis subject. Tooth extractions, cranial operations, amputations, cardiac and lung surgery, as well as abdominal and genitourinary operations, have been performed without any kind of anesthesia except hypnosis. *See* ANALGESIC; ANESTHESIA; PAIN.

Symptom control. Intractable pain, such as that involved in trigeminal neuralgia, phantom limb, carcinoma, and burns, has been allayed. Hypnosis has been used in the treatment of hiccoughs, neurodermatitis, the dumping syndrome (weakness, nausea, vertigo, and palpitation occurring immediately after partial or complete removal of the stomach), coughing spasms, headaches, frigidity, impotence, dysmenorrhea, discomfort from contact lenses, stuttering, tics, and the pain of childbirth. Many organic symptoms, as well as most functional symptoms, can be suppressed in appropriate cases. The suppression may be complete or only partially controlled, so that the individual is still able to report dangerous developments that may be occurring. The above symptoms are only a partial list of those in which hypnosis has been applied.

Motivational tool. Hypnosis has been utilized to motivate speech retraining in asphasics, to increase the desire for psychotherapy, to enhance or decrease movement of limbs that are involved in surgical procedures, to increase motivation for early ambulation following surgery, and to enhance feeling tone in some very depressed individuals. Hypnosis has also been used to alter attitudes, likes, and dislikes, as in the case of obesity, for various foods and eating; in this connection, it has been used as an uncovering technique, with the person hypnotized only part of the time. *See* MOTIVATION.

Conjunction with analysis. Hypnosis is used in a number of ways with psychiatric analysis. Analysis may proceed entirely in the hypnotic state, or frequently hypnosis is used only as an aid in uncovering repressed material, and has been particularly useful in overcoming blocks in free association and in producing dreams. Hypnosis has been used to create artificial neurotic behavior that permits the individual to better understand his or her own neurotic symptoms and control them. Another application is to create illusory situations through which the individual works, thus enabling the therapist to gain a better understanding of the person's ability to cope with certain problems. Hypnosis has been utilized to enhance the feelings of schizophrenics who may develop some intellectual understanding of their problem but who otherwise remain unresponsive. *See* PSYCHOTHERAPY.

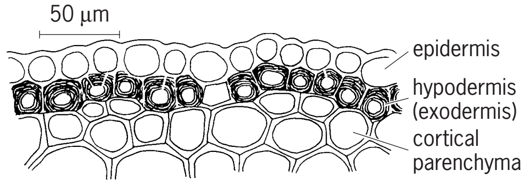
Conjunction with behavior therapy. One of the earliest uses of hypnosis was in the production of aversive reactions in the treatment of smoking and alcoholism. Hypnosis is an excellent tool for producing the state of relaxation which is called for by certain behavior-modification techniques, and it is also an effective way of creating various experiences which can be used for purposes of desensitization and of reinforcement.

Pharmacological substitute. Through the effects that may be produced in its presence, hypnosis may be used to provide substitutes for various pharmacological agents or may be employed with them to increase their effects. Hypnotic suggestions have thus been particularly used in place of or with soporifics, hypnotics, anesthetics, analgesics, and tranquilizers. Some lesser uses have included those of a stimulant, mood elevator, laxative, appetite reducer or enhancer, antiallergenic agent, lactation inhibitor or stimulant, and agent for control of menstrual discharge and cramps. In all of these instances, the advantages of hypnosis lie in the absence of tolerance problems, side effects, and other complications. The main disadvantage lies in the fact that such uses of hypnosis are probably limited to 20% of the population. Also, psychobiological problems can arise in unselected individuals. *See* CONSCIOUSNESS; PSYCHOPHARMACOLOGY. André M. Weitzenhoffer

Bibliography. S. J. Lynn and J. W. Rhue (eds.), *Theories of Hypnosis: Current Models and Perspectives*,1991; R. H. Rhodes, *Hypnosis: Theory, Practice, and Applications*, 1989, reprint 1998; A. Weitzenhoffer, *Practice of Hypnotism*, 2 vols., 2d ed., 1999; W. C. Wester, II and H. Alexander, Jr., *Clinical Hypnosis: A Multidisciplinary Approach*, 1991.

Hypodermis

The outermost cell layer of the cortex of plants. It forms a prominent layer immediately under the epidermis in many but not all plants (see **illus.**). In shoots, the hypodermis may be composed of parenchyma, collenchyma, or sclerenchyma and be from one to several cells thick. In roots, the hypodermis is often called the exodermis; it resembles the endodermis, and it develops Casparian strips, suberin deposits, and cellulose deposits impregnated with phenolic or quinoidal substances. Thus the root hypodermis is similar to the endodermis in cell wall anatomy and in its reaction to histochemical tests. The hypodermis is the mirror image of the endodermis in appearance; in the endodermis, wall deposits develop from the inner tangential wall outward; in the hypodermis, from the outer tangential wall inward. *See* CELL WALLS (PLANT).



Transection from the outer part of a *Smilax* root, illustrating a thick-walled hypodermis (exodermis) beneath the epidermis. One hypodermal cell is not thickened. (*After K. Esau*, *Plant Anatomy*, *2d ed.*, *John Wiley and Sons*, *1965*)

The hypodermis may produce substances that act as a barrier to the entry of pathogens, and in some plants it may function in the absorption of water and the selection of ions that enter the plant. Young cells of the hypodermis in the growing tip of the root contain cytochrome and peroxidase, enzymes associated with respiration and centers of metabolic activity. There is a large amount of protein and nucleic acid in the cells of the hypodermis, with unusually large nuclei, as in the endodermis. Similarity in substance content of the hypodermis and endodermis, with a similarity in anatomical form, suggests a common function for both tissues, but the function of the hypodermis has yet to be determined. *See* CORTEX (PLANT); ENDODERMIS. D. S. Van Fleet

Hypohalous acid

An oxyacid of a halogen [fluorine (F), chlorine (Cl), bromine (Br), iodine (I), or astatine (At)] possessing the general chemical formula HOX, where X is the halogen atom. The chemical behavior of hypofluorous acid (HOF) is dramatically different from the heavier hypohalous acids which, as a group, exhibit similar properties. These differences are attributed primarily to the high electronegativity and small size of the fluorine atom, which cause HOF to be an extremely strong oxidant with an anomalously weak O-F bond. Thus, the molecule is highly reactive and relatively unstable. (Gaseous HOF decomposes to HF and O_2 at room temperature with a half-life of about 1 h, and the liquid has a tendency to explode.) Because the most electronegative element in HOF is fluorine, whereas the other halogen atoms are less electronegative than oxygen, the O-X bond polarities are reversed in HOF and the heavier congeners. HOF therefore acts primarily as an oxygenating and hydroxylating agent, whereas the other hypohalous acids are electrophilic halogenating agents. For example, HOF hydroxylates aromatic compounds to form phenols and reacts instantaneously with water to give hydrogen peroxide [reaction (1)], whereas

$$\mathrm{HOF} + \mathrm{H_2O} \longrightarrow \mathrm{HF} + \mathrm{H_2O_2} \qquad (1)$$

hypochlorous acid (HOCl) chlorinates aromatic compounds and is unreactive toward water. *See* ASTATINE; HALOGEN ELEMENTS.

Hypochlorous, hypobromous, and hypoiodous acid solutions are formed by disproportionation of the corresponding halogen [reaction (2)]. In aqueous solutions, the equilibrium position lies increasingly to the left as the halogen size increases through the series Cl < Br < I. The reaction can be driven to completion by adding base or removing the halide ion by ion exchange or with precipitants such as silver or mercury salts. HOF is prepared at low temperature from F_2 and ice by the same reaction. Chloride-free solutions of HOCl can be prepared by hydrolysis of its anhydride, Cl_2O, which is in turn obtained by reaction of Cl_2 with mercuric oxide. Hypoiodous acid (HOI) can also been generated by oxidation of iodide ion or by hydrolysis of the polyhalide anion, ICl_2^-; these methods allow preparation of relatively high concentrations of HOI, which is unstable in aqueous solutions. Astatine is a radioactive compound formed during the decay of uranium and thorium. Its short lifetime permits isolation in only vanishingly small quantities, precluding quantitative investigation of its properties. Qualitatively, it appears to behave similarly to iodine. Evidence for formation of hypoastatous acid (HOAt) is that it reacts with weak oxidants and coprecipitates with iodinium dipyridine

$$\mathrm{X_2} + \mathrm{H_2O} \rightleftharpoons \mathrm{H^+} + \mathrm{X^-} + \mathrm{HOX} \qquad (2)$$

$[I(C_6H_5N)_2^+]$ salts, implying an At(I) oxidation state. *See* CHEMICAL EQUILIBRIUM.

Hypohalous acids are weak acids, that is, $HOX \rightleftharpoons H^+ + OX^-$, whose dissociation constants increase in the order HOI < HOBr < HOCl. Both the acids and their conjugate bases are thermodynamically unstable with respect to disproportionation [reactions (3) and (4)]. The decomposition mechanisms are

$$5HOX \longrightarrow 2X_2 + XO_3^- + H^+ + 2H_2O \qquad (3)$$

$$3OX^- \longrightarrow 2X^- + XO_3^- \qquad (4)$$

complex with rates that vary widely, increasing in the order $X = Cl^- < Br^- < I^-$. Hypochlorite solutions (for example, commercial bleach) are stable at room temperature and below for months but decompose at elevated temperatures; hypobromite solutions are stable only at reduced temperatures; and hypoiodite generally decays within minutes of its formation. Decomposition reactions of hypobromous acid (HOBr) and HOI are second-order in HOX and strongly buffer-catalyzed. Textbooks often incorrectly describe OI^- as so unstable that it appears only as a reaction transient with fleeting existence; under favorable medium conditions, it will decay at room temperature with reaction half times exceeding several minutes. *See* BLEACHING.

Reactivity. HOCl and HOBr and their anions are powerful oxidants that react rapidly with a wide range of organic and inorganic reductants. In addition, both acids halogenate aromatic compounds and form halohydrins with unsaturated organic compounds, and N-chloro compound with nitrogen bases. In general, HOCl reacts much more rapidly than OCl^-. Rate constants with a series of inorganic ions have been shown to increase proportionately with their nucleophilicity and, qualitatively, molecules that are good nucleophiles are oxidized much more rapidly than molecules that contain no nucleophilic centers. This high reaction selectivity is also observed in reaction environments as complex as living cells, where electron-rich compounds such as iron-sulfur clusters, metalloporphyrins, nitrogen heterocycles such as the adenosine nucleotides, conjugated polyenes such as carotenes, and biological amines and thiol-containing compounds are preferentially oxidized. This overall reaction behavior is taken to indicate that HOCl and, by inference, the other hypohalous acids react by electrophilic attack of the electron-deficient halogen atom at the nucleophilic center, forming an incipient halogen-nucleophile bond. For example, the reaction between HOCl and SO_3^{2-} is thought to occur by a stepwise mechanism [reactions (5) and (6)], involving transfer of Cl^+ from HOCl to the sulfite sulfur atom as the initial step. This general mechanism is supported in favorable cases by detection of chlorine-nucleophile intermediates. *See* HALOGENATION; OXIDATION-REDUCTION.

$$HOCl + SO_3^{2-} \longrightarrow ClSO_3^- + OH^- \qquad (5)$$

$$ClSO_3^- + H_2O \longrightarrow SO_4^{2-} + Cl^- + 2H^+ \qquad (6)$$

Cytotoxicity. HOCl and HOBr are potent cytotoxins. Hypochlorite was first used as a disinfectant around the beginning of the nineteenth century by the French pharmacist A. G. Labarraque, and has subsequently been widely applied to problems in public sanitation. Boric acid buffered solutions were prominent pharmaceutical agents in the early 1900s but have now been almost completely replaced by antibiotics. Recent studies with bacteria indicate that death is accompanied by disruption of metabolic functions associated with the plasma membrane, including inactivation of the adenosine triphosphate synthase and proteins involved with active transport of metabolites, and (in aerobes) inhibition of respiration. This pattern of reactivity suggests that toxicity arises from disruption of the energy-transducing capabilities of the cell. Remarkably, white blood cells also appear to use hypohalous acids to fight infection. Specifically, neutrophils and monocytes contain the enzyme myeloperoxidase, which catalyzes the oxidation of chloride ion to HOCl by respiration-generated hydrogen peroxide, and eosinophils contain a similar peroxidase which oxidizes bromide ion to HOBr. Many mammalian secretory fluids also contain peroxidases capable of oxidizing halides to their +1 oxidation state; prominent among these enzymes is lactoperoxidase, which in physiological environments is thought to oxidize the pseudohalide thiocyanate (SCN^-) to the hypohalite analog hypothiocyanite (HOSCN). Other chloroperoxidases and bromoperoxidases that catalyze organic halogenations by oxidation of chloride and bromide to Cl(I) and Br(I) are found in fungi and marine organisms. Some of these are being evaluated with respect to potential applications in medicine and as industrial biocatalysts.

James K. Hurst

Bibliography. J. C. Bailar et al. (eds.), *Comprehensive Inorganic Chemistry*, vol. 2, Pergamon Press, Oxford, 1973; J. Everse, K. E. Everse, and M. B. Grisham (eds.), *Peroxidases in Chemistry and Biology*, CRC Press, Boca Raton, FL, 1990; D. W. Johnson and D. W. Margerum, Non-metal redox kinetics: A reexamination of the mechanism of reaction between hypochlorite and nitrite ions, *Inorg. Chem.*, 30:4845–4851, 1991.

Hypothermia

A condition in which the internal temperature of the body is at least 3.6°F (2.0°C) below an internal temperature of 98.6°F (37°C).

Heat Regulation

Hypothermia represents a continuum of effects that vary with the severity of cold on physiological systems. Even though the human body can do without food for a number of weeks or water for a number of days, it needs a specific internal temperature that is regulated on a minute-by-minute basis to maintain all normal body functions. The many physiological and behavioral processes involved in maintaining the internal temperature constant are called thermoregulation. *See* THERMOREGULATION.

Core and peripheral temperatures. Core and peripheral temperatures are the two major categories of body temperature that are sensed by the brain and spinal cord in order to coordinate various body processes and maintain a constant internal, or core, temperature. The internal temperature is that of the brain, spinal cord, heart, and lungs. It must be maintained within strict limits; otherwise the heart and brain will be compromised, eventually leading to death. This temperature is usually monitored clinically by recording the temperature of the rectum, but a more accurate site would be the esophagus, because the esophagus is anatomically close to the heart, and the temperature of the heart is considered to be the overall average body temperature. Another area that can be used to monitor internal temperature is the eardrum, which is anatomically close to the brain and has the same temperature as the brain.

The peripheral, or shell, temperature is that of the skin and the muscles. It can vary over a much wider range (32–95°F or 0–35°C) than the core temperature. That is possible because the cells of the skin, muscles, and blood vessels are not as sensitive to decreased temperatures as are those of the vital organs. Skin temperatures, such as those recorded in the axilla, are not indicative of an internal temperature.

Physiological control of temperature. Temperature regulation involves the brain and spinal cord, which monitor the difference between the internal and the peripheral temperature, with subsequent psychological and physiological adjustments to maintain a constant internal temperature. The brain records the various body temperatures by means of specialized nerves throughout the body that respond specifically to temperature. These specialized nerve endings, called thermal receptors, are dedicated to responding either to cold or to heat. One theory as to how the system operates holds that the hypothalamus is programmed to maintain a specific temperature (the set point) in the internal parts of the body. If the brain records that the internal temperature is falling relative to the set point, certain automatic reactions occur to minimize or reverse the fall in internal temperature.

Striving for some form of thermal comfort is instinctive. For a human to be comfortable in a thermal environment, a relationship (known as the thermal comfort zone) must be maintained between the core–shell temperature profile and the temperature of the environment. The thermal comfort zone is about 68–72°F (20–23°C). Even on a relatively warm day of 70°F (22°C), the body must keep producing enough heat to minimize the heat transfer that occurs as a result of the difference between the internal temperature and the environment. This process explains how hypothermia can occur even in relatively moderate climates; cases of hypothermia have been reported in the tropics among malnourished individuals. *See* COMFORT TEMPERATURES.

Behavioral and physiological responses to cold stress. The initial response to a cold stress is a behavioral one: Leave the cold environment, button a jacket, or try to decrease body size by curling up into a ball to minimize heat loss. Unlike animals (which grow more fur, deposit fat in anticipation of the winter, or hibernate), humans must maintain a warm internal temperature by means of clothing and by behavior modification. Human response to cold stress is highly variable, however.

Cold stress can be either acute or chronic. People who live in a cold climate usually do not respond to a cold stress as vigorously as those who live in warmer areas. Tolerance to cold temperatures can be developed by constant exposure to the cold, but even in cold environments an internal temperature of 98.6°F (37°C) must be maintained: falling below a critical temperature has dire consequences. Hypothermic individuals as cold as 53.6°F (12°C) have been reported to have survived after medical treatment, however. In contrast with modern, urban populations, some societies whose members have been constantly exposed to the cold, such as the aborigines of Australia, are able to drop their internal temperature a number of degrees to minimize the metabolic stresses required to shiver and stay warm. These adaptations seem to become blunted or lost following a transition to a more contemporary, or urban, existence.

Physical processes influencing heat transfer. In general, urban populations take advantage of certain physical processes of heat transfer (radiation, conduction, convection, and evaporation) in their behavior, selection of clothing materials, and design of cold-weather gear in order to maintain the internal temperature in various climates. These same physical processes explain the body's physiological responses to the cold. In addition, food and exercise are also important in maintaining warmth.

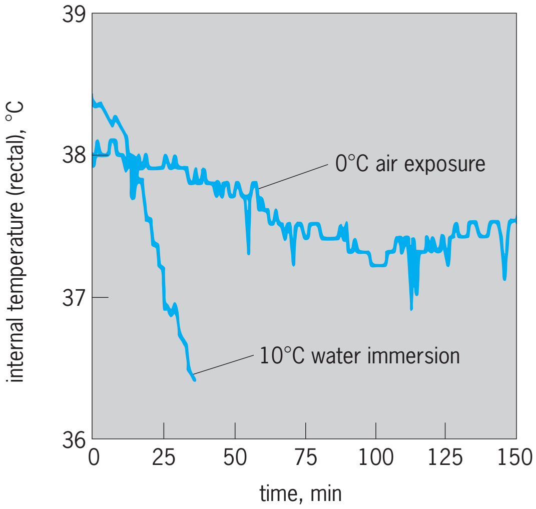
Radiation. For general considerations, the body radiates approximately as much heat as a 70-watt light bulb, and that heat transfer occurs between the skin surface and the environment. Whether heat is exchanged by radiation is determined by the temperature difference between the radiating surfaces, their emissivity, and their surface characteristics. Thus, if a number of people are in a room, the room temperature increases because the people are radiating heat. Conversely, if the surrounding walls are colder than the human skin, the people lose heat. That principle can be demonstrated at night by standing in a room near a large window surface. Even though the room air is warm, the colder surface of the window causes a significant heat transfer away from the body.

Heat loss can be minimized by using the concept of radiation. Usually clothing decreases the temperature difference between the skin and the environment and thus minimizes heat transfer. Thin blankets made of aluminum also minimize radiant heat loss. Such coverings are used by runners after a long race and patients who have undergone heart surgery in cold operating rooms. A decrease in the surface area of the radiating objects minimizes heat loss, and so curling up into a ball or assuming the fetal position minimizes heat transfer. *See* HEAT RADIATION.

Conduction. Conduction is heat transfer between two thermal objects in physical contact. All substances have thermal conductivity values. If a substance is a poor thermal conductor, it is a good insulator and thus minimizes heat transfer. Fat is an example of a poor conductor: The more fat that is accumulated, the less heat is transferred from the core—and the smaller the chance of becoming hypothermic. In addition to its being a poor thermal conductor and a good insulator, fat is a major energy depot for animals. Hibernating forms therefore gorge themselves before winter arrives to protect against the ravages of the oncoming cold and to accumulate a built-in supply of food. Studies suggest that under certain conditions of cold stress, humans also accumulate fat deposits. *See* ADIPOSE TISSUE; HIBERNATION AND ESTIVATION.

Materials that act as poor conductors, such as wool, feathers, and synthetic fibers, minimize heat transfer. The challenge is to find the right kind of material (conductor) so that the person is simultaneously comfortable and not too hot or cold while exercising or relaxing. An example is the wet suit, which is worn by those who swim in cold water or who work in cold environments and might accidentally fall into cold water. The suits allow a certain amount of water to enter, and it is continually warmed by the heat generated by the activity of the person and the body's metabolism. The layer of warm water between the skin and the suit provides an additional layer of insulation between the skin and the cold water, and thus minimizes heat transfer, and increases dramatically the time a person can stay in cold water.

Because water has a very high rate of conduction, heat transfer occurs faster in cold water than in cold air (see **illus.**). A person surrounded by air at a temperature of 50°F (10°C) experiences a drop in internal temperature that takes place more slowly than a person in 50°F (10°C) water. A nude person in 50°F (10°C) air becomes hypothermic in 24 h, whereas the same person in water of the identical temperature becomes hypothermic in 2 h. *See* CONDUCTION (HEAT).



Body heat lost with exposure to 0°C air and 10°C water. °C = (°F − 32)/1.8.

Convection. Convection is a process of heat gain or loss that takes place between two objects in physical contact when one object is in motion. As air moves over a thermal object, heat is transferred to the molecules of the air or fluid. These molecules then move out of contact and are replaced by cooler molecules, which again absorb heat and then move out of contact. This process takes place when riding a motorcycle or jogging, for example. Wind streams across the body and causes a major heat transfer from the body, which cools quickly.

This physical phenomenon is the basis for the term wind chill. Cold, dry wind has severe chilling effects because it increases the rate at which heat and moisture are drawn off the skin. The wind chill index is an attempt to quantify the combined effects of air speed and temperature on heat transfer. Because of the efficiency of convection as a heat transfer process, the cold wind becomes a continual heat drain. *See* CONVECTION (HEAT).

Evaporation. The process of evaporative heat loss can also be important in terms of facilitating heat transfer and predisposing a person toward hypothermia. Evaporation is effective as a heat dissipator since heat from the body is used to evaporate the sweat that is secreted from the body with exercise. If the sweat does not evaporate, it is ineffective in cooling the body and so clothing that does not allow the sweat to evaporate causes the sweat to cling to the body. If this sweat is cooled by cold wind, it cools the body and sets up the conditions that can lead to hypothermia. In addition to heat loss, evaporation promotes a net water loss from the body. Water is critical for maintaining normal body function, especially in the cold.

Although overt sweating occurs during exercise, insensible evaporation takes place when the body is at rest. It serves as a fine control of the body's temperature but can lead to hypothermia since in certain situations, when functioning in conjunction with other body processes, insensible sweating may promote a significant amount of body water and heat loss.

A significant amount of water is lost by way of the lungs, which humidify and heat the inspired air. When cold air is inspired, it cools the blood in the lungs and simultaneously removes some fluid. Considering that the lungs have the surface area of a tennis court, breathing in a cold environment can be a significant factor in inducing hypothermia. *See* EVAPORATION.

Metabolism. In addition to the physical processes that determine heat transfer, the body generates heat as a result of cellular reactions that produce energy. This metabolic heat, which is expressed in terms of internal temperature and which maximizes the chemical reactions of the body, is ultimately derived from the food that is consumed. The caloric value of food is a quantification of the amount of energy contained in that food. When food or caloric intake is insufficient, the body gets its energy from its own tissues. In extreme cases, the body uses all of its fat and then begins to digest its protein. In a state of malnourishment (for example, in cases of starvation or anorexia nervosa), the body feels cold since it is not taking in enough energy to stay warm. *See* ANOREXIA NERVOSA.

Metabolism is controlled by various hormones, among which are the thyroid hormone and epinephrine. The thyroid hormone increases overall body metabolism. In extreme cases, if the thyroid is not functioning, the body's metabolism is slowed and consequently the body temperature decreases. Epinephrine increases cell metabolism and thereby creates a feeling of warmth. Emotions such as fright or drugs such as amphetamines cause an increase in this hormone and thus increase the body's metabolism and the internal temperature. *See* EPINEPHRINE; THYROID HORMONES.

Exercise also raises internal temperature by increasing metabolism and generating more heat. Shivering, which generates heat by way of muscle contractions that are similar to exercise, represents an effort on the part of the nervous system to keep the body warm by causing involuntary muscle contractions. *See* METABOLISM.

Temperature clock. The physiological factors that affect overall heat transfer are under the control of the nervous system. The brain has one or more internal clocks, which regulate many of the heat-generating processes of the body over a 24-h period. During sleep, the internal temperature drops 0.9°F (0.5°C) and then rises and falls over the sleep period in a programmed manner. The relationship between the temperature clock and personality is a subject of interest in psychiatry, specifically with respect to depression, because a decrease in sleep time influences the clock.

Types of Hypothermia

Various environmental situations predispose humans to hypothermia, which can occur even in the absence of cold. In fact, hypothermia is more common in temperate regions than in the colder climates. Because of the uniqueness of the situations in which hypothermia can occur, various kinds of hypothermia have been classified, all of which can prove fatal.

Primary hypothermia. Primary hypothermia is a decrease in internal temperature that is caused by environmental factors in which the body's physiological processes are normal but thermoregulation capability is overwhelmed by environmental stress.

Air hypothermia. Air (formerly exposure) hypothermia is thought to be the most common form. A person exposed to cold air experiences the same processes as a person in cold water, but air hypothermia occurs more slowly. The induction of air hypothermia is more subtle and therefore more dangerous since it can occur over a number of weeks. People who exercise in cold air and who sweat excessively cool themselves faster than those who do not sweat, because water removes the body's heat. Cold air is a special concern to those who are confined and are unable to walk or exercise. A room with a temperature of 60°F (15.6°C) can become lethal over time for an individual who cannot exercise. On the other hand, mountain climbers, joggers, runners, back-packers, hikers, and victims of natural disasters in which they are exposed to cold are also likely victims of air hypothermia. Mountain climbers represent a subgroup since they have three major stressors: the cold environment, the decrease in oxygen at high elevations, and dehydration.

The degree to which a person reacts to a cold air stress is dependent on such factors as age, physical stamina, the intensity of the cold stress, and the responsiveness of the thermoregulatory system. One of the most convenient ways to determine whether someone is suffering from hypothermia is a noted change in personality: Complaints of fatigue, sluggish speech, and confusion are common, and in some cases the behavior resembles that of intoxication.

Initially, skin temperature falls rapidly, blood vessels to the skin constrict, and shivering begins. After 5–10 min, shivering ceases for about 10–15 min, but this is followed by uncontrollable shivering. The cardiovascular system, specifically the blood vessels to the extremities, circulates blood to satisfy the oxygen demands of the body organs and to distribute body heat. In a cold situation, the nervous system causes the blood to be redistributed away from the skin as the blood vessels of the skin close down to minimize heat transfer to the cold environment. The decrease in skin temperature coupled with vasoconstriction makes the person feel cold, and sometimes the fingers and toes can become painful. Internally, there is an increase in the levels of hormones that control metabolism, and blood is shunted primarily to the lungs, heart, and brain. The person becomes dehydrated as the inspired air is warmed and humidified. If the tense and shivering muscles do not generate enough heat, the hypothermic process begins and progresses for at least 3–5 h. As hypothermia continues, the arms become rigid, and the person loses the ability to make fine movements.

During this period of time the heart rate initially increases, then stabilizes and as the person's internal temperature becomes progressively colder, the heart rate and respiration slow. In severely hypothermic persons, it is very difficult to detect a slow heart rate or determine if the person is breathing. A temperature of 95°F (35°C) is only the beginning of mild hypothermia and shivering can continue for hours, depending on the muscle and fat supplies available. Eventually, the environment becomes overwhelming. At 86°F (30°C), the person loses consciousness and shivering ceases. Death does not occur until the internal temperature drops further: Death results at 68–77°F (20–35°C) because of cardiac standstill.

Immersion hypothermia. When a person falls into cold water, a gasping response is triggered by the thermal receptors on the skin. For some individuals, the cold stress may trigger a heart attack. Although as much of the body as possible should be kept out of the water, many victims of immersion hypothermia stay in the cold water because they cannot tell how cold they are. Shivering becomes generalized and, unlike its effect in cold air, may in fact cause a faster drop in internal temperature since the water layer closest to the body is stirred and convective heat loss is promoted. As a rule, any movement in cold water promotes heat loss due to heat convection. A person who is unable to come out of the water experiences a rapid drop in internal temperature as a result of movement in the water, the amount of clothing worn, the amount of body fat, and the temperature and state of the water.

The misconception persists that a person can die in 10 min in 50°F (10°C) water. Although the greater conductive property of water relative to air is a major heat sink, physiological and behavioral responses act to minimize the heat loss. Survival in 50°F (10°C) water is possible for several hours at most if the person is dressed in street clothes and a life jacket. Death within minutes may be attributed to drowning, not hypothermia.

In divers' hypothermia, a variation of immersion hypothermia, the person is properly suited and carries an air supply. Even with a wet suit to minimize the cold, cooling of the periphery of the body takes place because the cold water is a heat sink for body heat. In addition, breathing a cold, dry air mixture cools the core of the body, and movement in cold water increases heat loss. Gas mixtures such as helium–oxygen combinations promote an even greater loss of body heat due to their convective heat transfer characteristics. Thus these divers are cooling themselves both internally and externally. By breathing various gas mixtures that cool and dehydrate the core of the body, a drop in internal temperature can precede a drop in peripheral temperature, thus overriding the normal warning signs of approaching hypothermia. *See* DIVING.

Submersion hypothermia. The cooling of the body in this form of hypothermia (actually a situation in which a person drowns in cold water) allows the brain and heart to withstand approximately 45 min of oxygen debt. This is most operative for young children, and people making rescue attempts should always be aware of this. A child can survive for an extended period of time while completely submerged because of the fact that the body is undergoing both internal and external cooling. As the child is drowning, cold water is swallowed and enters the lungs, which cools the core. At the same time, the cold water that bathes the skin rapidly cools the periphery. The multiple effects of the internal and external cooling decrease the metabolic rate and give the child a window of safety of approximately 45 min. In warm water, survival is possible for only 5–7 min.

In addition to the physical factors associated with water cooling, a reflex may play a significant factor in a child's chance for survival. This reflex, called dive reflex, is triggered when cold water bathes the face. This stimulus causes the peripheral blood vessels to constrict, the heart rate to decrease, and the blood to move to the core, giving vital organs additional oxygen. This reflex is quite pronounced in children and, because of its oxygen-conserving results, plays a role in improving the submersed hypothermic victim's chance for survival.

Secondary hypothermia. A decrease in core temperature caused by an underlying pathology that prevents the body from generating enough core heat is referred to as secondary hypothermia. If any of the thermoregulatory systems are altered, the body's ability to generate heat subsequently decreases and hypothermia can then develop without warning. Insufficient muscle mass to generate heat, medications that interfere with metabolism, an underlying systemic infection, decreased thyroid hormone production, and paralysis predispose to hypothermia. Premature infants with low body fat and a large surface-to-volume ratio lose heat rapidly and are at risk for becoming hypothermic. The elderly are perhaps the most susceptible to secondary hypothermia. However, whether the process of aging with no associated debility also alters the thermoregulatory system in the elderly remains to be determined.

Ethanol is often imbibed to ward off the effects of the cold, and this leads to the mistaken belief that it is effective in warming victims of hypothermia. Actually, ethanol causes decreased sensitivity to the cold and results in depressed sensory input from the thermal receptors. Consequently, there is no vasoconstriction or shivering. Many drugs and chemicals affect the thermoregulatory system. Sedatives and antidepressants can decrease a person's ability to sense cold and respond to it, indirectly predisposing the user to hypothermia. Some antidepressants also cause peripheral vasodilation, giving the user a sense of warmth even though the person is losing heat to the environment.

Clinical hypothermia. Some cardiac surgical procedures require clinically induced cooling to stop the heart from beating. Induced hypothermia lowers the oxygen demand of the body tissues, so that oxygenated blood need not circulate. In the case of coronary bypass surgery, the entire body is cooled, enabling the surgeons to work for an extended period of time on the cold heart. During such procedures, the heart is actually the coldest part of the body. Also, cooling the heart cools the blood and thus minimizes hemorrhage.

Frostbite. In hypothermia, the body's internal temperature decreases, but no solid freezing takes place. In frostbite, which is freezing of the digits or the limbs, there is actual formation of ice crystals. Basically the digits go through various stages of cooling. Initially, in the prefreeze phase, the finger temperature is 37.4–50°F (3–10°C). Next, at 24.8°F (−4°C) ice crystals form outside the cells of the digits, circulation is limited, and cell death takes place if the process is allowed to continue. The cells of the digits and limbs can tolerate low temperatures that would be lethal to brain or nerve cells. However, once they are rewarmed and thawed, they develop an increased sensitivity to the cold and become more susceptible to frostbit Any part of the body can become frostbitten, but the fingers, toes, ears, nose, and cheeks are most often affected.

Cold-induced vasodilation. Associated with cold exposure is a phenomenon that occurs when fingers are placed in freezing water. When the fingers are first put into cold water, the blood vessels constrict and, therefore, the temperature of the fingers falls. After falling to 39.2–42.8°F (4–6°C), the blood vessels reopen and the finger temperature rises again. This sequence of vasoconstriction followed by vasodilation continues as long as the finger remains in cold water. This same phenomenon can also localize in areas such as the toes, fingers, earlobes, and buttocks. The process, which is thought to be protective in that vasodilation does not allow the finger to get too cold, is augmented among those who are often exposed to the cold for long periods of time so that the individuals can spend more time working in the cold.

Rewarming strategies. Even when the body becomes so cold that all function stops, revival with proper rewarming techniques is still possible. In some cases, physiologically dead patients have been revived as long as 45 min after any sign of life could be detected, because hypothermia stops the cells' metabolism and consequently diminishes any cellular demand for oxygen or glucose. Thus, sensitive cells in the brain and heart can survive cold temperatures for a period of time. This period, known as the metabolic icebox, is the primary reason that victims of hypothermia can survive. When a person becomes hypothermic, safe rewarming is the major concern. Since the brain and other vital organs are in a hypothermic state and their oxygen and glucose demands are small, the most important parameters to monitor in hypothermic individuals are heart rate, blood pressure, respiration, and fluid and electrolyte levels. Although rewarming would appear to be the first consideration, monitoring the heart and minimizing any mechanical or chemical damage to it are, in fact, the most critical.

Careful monitoring is important during rewarming. When a person is being rewarmed, the body temperature continues to fall for some time. This phenomenon, called after-drop, may contribute to further manifestations of hypothermia. Because of the effects of hypothermia, cells may lose their physiological or structural integrity, and when the body is warmed too fast, lethal concentrations of some ions, such as potassium, may stop the heart from beating.

For mild hypothermia, physical exercise is recommended because it effectively warms the core. Conversely, a campfire increases heat transfer to the skin but does not substantially warm the core. Similarly body-to-body rewarming warms only the periphery, not the core.

When severe hypothermia renders a person unconscious but with a detectable pulse and respiration, transfer to an emergency center or hospital for warming under the supervision of a physician is recommended. Rewarming in the field is not recommended because of the associated cardiovascular problems. *See* HOMEOSTASIS.

Robert S. Pozos; L. E. Wittmers; James Hodgdon

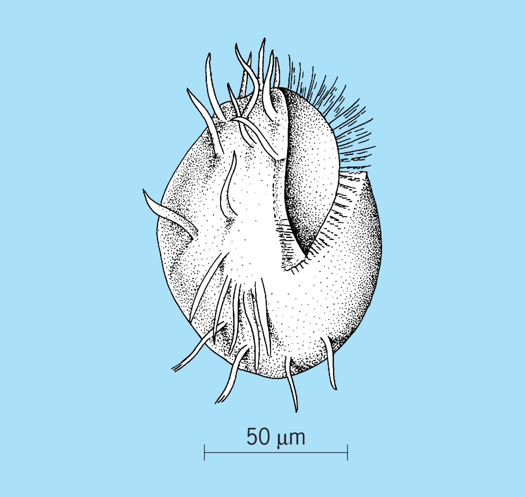
Bibliography. K. Bowler and B. J. Fuller (eds.), *Temperature and Animal Cells*, 1987; E. Kano (ed.), *Current Researches in Hyperthermia Oncology*, 1988; R. S. Pozos and L. E. Wittmers (eds.), *The Nature and Treatment of Hypothermia*, 1983; J. R. Sutton (ed.), *Hypoxia and Cold*, 1987; J. A. Wilkerson et al. (eds.), *Hypothermia*, *Frostbite and Other Cold Injuries*: *Prevention*, *Recognition*, *Pre-Hospital Treatment*, 1986.

Hypothesis

A tentative supposition with regard to an unknown state of affairs, the truth of which is thereupon subject to investigation by any available method, either by logical deduction of consequences which may be checked against what is known, or by direct experimental investigation or discovery of facts not hitherto known and suggested by the hypothesis. A classical example is the nebular hypothesis with regard to the origin of the solar system. The "truth" of this hypothesis may be checked by its ability to explain the properties of the solar system as now observed. Percy W. Bridgman; Henry Margenau

Hypotrichida

An order of Spirotrichia. These protozoa are commonly considered to represent the pinnacle of specialized development in the evolution of all ciliates. Somatic ciliature of the ventral surface has been replaced by cirri; on the dorsal surface it is absent or represented by inconspicuous sensory bristles. The adoral zone of membranelles is very prominently developed, and the buccal area may occupy



***Euplotes*, an example of Hypotrichida.**

a large part of the ventral surface of the body. The whole body is generally rigid and is pronouncedly flattened dorsoventrally. Hypotrichs occur ubiquitously in fresh- and salt-water habitats and have been studied quite extensively by experimental biologists. Common examples are *Euplotes* (see **illus.**) and *Diophrys*, as well as *Oxytricha* and *Stylonychia*. *See* CILIOPHORA; PROTOZOA; SPIROTRICHIA.

John O. Corliss

Bibliography. S. Anderson and J. K. Jones, Jr., *Recent Mammals of the World*, 1967; W. N. McFarland, *Vertebrate Life*, 1979; A. S. Romer, *Vertebrate Paleontology*, 1966.

Hypoxia

Failure of an adequate amount of oxygen to gain access to, or to be utilized by, the body. When the lack of oxygen is extreme, the term anoxia is used.

Oxygen deprivation may result from interference with inspiration of oxygen, passage of oxygen from the lungs into the blood, transport of oxygen to tissues, and use of oxygen by cells. Interference at any of these stages can produce damage that leads to further hypoxia. This is seen most dramatically when the brain is deprived of the necessary oxygen for more than a few minutes. Nerve cell degeneration begins quickly, and although the original cause of anoxia is removed, damage to the respiratory centers prevents resumption of breathing. *See* RESPIRATORY SYSTEM.

Oxygen may fail to enter the lungs in adequate amounts because of lowered concentrations in the atmosphere or because of obstructions in the airways. High altitudes, drowning, strangulation, aspiration of a foreign body, and lung disease are extrinsic causes of hypoxia. *See* RESPIRATION.

Inadequacy or failure of the respiratory mechanism may result from trauma, poisoning, progressive hypoxia, and other causes.

The passage of oxygen from the lung alveoli to the adjacent blood capillaries may be prevented or decreased by such conditions as chronic lung disease, infections, presence of foreign materials, and developmental defects.

A great number of conditions may interfere with the blood transport of oxygen, most of which is accomplished by the red blood cells. Examples include various forms of anemia, heart disease, trauma, hemorrhage, and circulatory diseases. In brief, anything that decreases blood supply or oxygenation of the blood cells may produce hypoxia. *See* CIRCULATION DISORDERS; RESPIRATORY SYSTEM DISORDERS.

Oxygen transfer from the blood into tissue cells may be prevented by various disturbances. Cellular damage that interferes with oxygen use may result from a host of circumstances, including previous anoxia or hypoxia. The most notable causes of such damage are various microorganisms, noxious chemicals, and physical injury. Failure of cellular respiration is sometimes termed histiotoxic anoxia.

Edward G. Stuart; N. Karle Mottet

Hyracoidea

An order of mammals closely related to elephants. Hyracoids were a diverse and successful group of mammals in Africa during the Eocene and Oligocene epochs (early part of the Age of Mammals, 55 to 34 million years ago), and they are still represented by a few living species. The early hyracoids ranged from animals as small as rabbits to ones as large as modern Sumatran rhinoceroses. The fossil skeletons of the early hyracoids indicate that some species were active runners and leapers, while others were heavy, piglike quadrupeds. Their teeth suggest herbivorous diets, ranging from fibrous leaves in some species to pulpy fruits and roots in others. Hyracoids originated in Africa but later extended their range into Europe and Asia, eventually attaining a distribution encompassed by China, Spain, and South Africa. In contrast to the early diversity of the order, the only living species are a few small-bodied animals (1.5–5.5 kg or 3.3–12 lb) that inhabit forests, scrubby brushlands, and rocky deserts in Africa and the Middle East. Hyracoidea represents a classic case of a spectacular adaptive radiation on an isolated continent, now reduced to a few remnant living taxa. *See* EOCENE; OLIGOCENE.

Classification and phylogeny. The modern hyracoids include about 7 to 12 species classified in three genera, *Procavia, Heterohyrax*, and *Dendrohyrax*. All are members of the family Procaviidae. *Gigantohyrax* is a fourth, extinct genus of Procaviidae known from Plio-Pleistocene caves in southern and eastern Africa. All of the remaining fossil hyracoids, representing at least 18 genera, are often classified in a separate, very diverse family, Pliohyracidae, pending better understanding of the phylogenetic relationships among forms. Important subfamilies include the Geniohyinae, a group of primitive, bunodont (lumpy toothed), piglike forms; the Saghatheriinae, the most abundant and diverse hyracoids of the Eocene and Oligocene; the Titanohyracinae, larger cursorial and graviportal (weight-bearing) hyracoids with teeth resembling those of early perissodactyls; and the Pliohyracinae, an assemblage of robust hyracoids, at least some of which were partially aquatic.

Morphological and molecular lines of evidence indicate that hyracoids are close relatives of elephants (order Proboscidea) and manatees and dugongs (order Sirenia). Hyracoids and elephants share several peculiar specializations, including continuously growing upper incisors, or tusks; mammary glands that are located on the chest rather than low on the abdomen; testes that remain in the body cavity rather than descending during development into a scrotum; and a peculiar linear arrangement of the wrist and foot bones (taxeopody). Molecular studies of blood and eye lens proteins further support a phylogenetic relationship among these three orders. Therefore, the orders Hyracoidea, Proboscidea, and Sirenia are best classified together in the superorder Paenungulata. *See* PROBOSCIDEA; SIRENIA.

Fossil record. Among the earliest hyracoids are *Seggeurius* and *Microhyrax*, both known only by jaws from the middle Eocene of Algeria (about 50 million years old). *Seggeurius* is a small geniohyine that may retain the most primitive dental morphology known within the order. *Microhyrax* is an even smaller hyracoid with crestier teeth than *Seggeurius*. Both are poorly known. Much more complete fossil material, including numerous jaws, crania, and limb bones, are known for three Eocene saghatheriines, *Saghatherium, Thyrohyrax*, and *Megalohyrax*. These genera differ from modern hyraxes in having long snouts, complete eutherian dentitions, and distinctive patterns of sexual dimorphism. Small *Thyrohyrax* and large *Megalohyrax* had a hollow chamber inside the mandible of one sex (perhaps the females), while the other sex did not. The hollow chamber is not understood but may have been a resonating device for loud vocalizations. Both sexes of *Saghatherium* lacked a chamber, but the males were much larger in size than the females. These patterns of sexual dimorphism hint that fossil hyracoids must have had complex social systems and behaviors. Another well known genus is *Titanohyrax*, a very large cursorial (running) hyracoid with teeth that indicate a leafy diet.

The best fossil record of early hyracoids comes from the Fayum, Egypt, where eight genera have been found in upper Eocene and lower Oligocene sediments (36–32 million years ago). After the early Oligocene, the diversity and importance of hyracoids began to decline, possibly due to competition with the newly arriving rhinos, bovids, and other groups migrating into Africa from the northern continents. By the early Miocene, only a few species are known from East Africa, but two evolutionary radiations of hyracoids were still to come. The distinct subfamily Pliohyracinae differentiated in Africa by the middle Miocene (*Parapliohyrax* and *Prohyrax*) and later spread to Eurasia. Some of the heavily built, aquatic pliohyracines survived until the Plio-Pleistocene in Europe (*Pliohyrax*) and China (*Postschizotherium*). These large pliohyracines had extremely hypsodont (high-crowned) teeth, indicating a very tough, fibrous plant diet. The modern family Procaviidae, characterized by relatively small size, reduced dentition (loss of canines and sometimes other teeth), and the development of a large gap between incisors and premolars, first appeared in the late Miocene of Namibia (*Heterohyrax auricampensis*). An abundance of Pleistocene procaviids is found throughout southern and eastern Africa, in habitats that are reconstructed by paleontologists as being similar to modern arid savannas and scrublands.

Habits and distribution. *Heterohyrax brucei* (bush hyrax) and species of *Procavia* (rock hyraxes) (see **illustration**) inhabit scrubby arid country in eastern Africa, from low desert to mountainous, alpine habitats. In addition, the range of *Procavia* extends into the deserts of northern and southern Africa, and the Middle East. *Procavia* and *Heterohyrax* prefer rocky outcrops, or kopjes, where they can hide in cracks and crevices, or bask on sunlit boulders. The elastic soles of their feet and the linear alignment of their wrist bones allow them to scramble and leap nimbly among the rocks. They live in colonies numbering from just a few to up to 40 individuals. A colony's territorial male, who is larger than the females, defends a rocky range inhabited by several females and immatures. Warning vocalizations, made loud with the help of an expanded chamber in the eustachian tube (the connection between throat and ear), alert the colony to predators (hyraxes are among the common prey of several carnivores and large birds). Female hyracoids produce few offspring (one to four per annual litter) which develop slowly for mammals of their small size, taking more than a year to reach sexual maturity. *Procavia* is a grazer that consumes a wide range of grasses and other coarse plant foods, an adaptation reflected by its high-crowned molar teeth, which are reminiscent of miniature rhino molars. Hyraxes do not ruminate, but they do have complex guts and utilize the help of gut microbes in digesting their plant food. As a consequence, they are able to digest some plants that are toxic to other mammals. *Heterohyrax* differs from *Procavia* in having a more general diet of softer leaves, grasses, and other plant parts. In contrast to its open country relatives, *Dendrohyrax* is typically an arboreal climber in equatorial tropical forests. It does not live in concentrated colonies but is more evenly distributed throughout the forest, communicating socially by loud calls at night. Because of the destruction of African forests, *Dendrohyrax* is the most threatened of the living hyraxes. *See* DENTITION; EUTHERIA; MAMMALIA. D. Tab Rasmussen

Cape hyrax (*Procavia capensis*). (*Photo by Lloyd Glenn Ingles; © 1999 California Academy of Sciences*)

Bibliography. J. Kingdon, *East African Mammals: An Atlas of Evolution in Africa*, University of Chicago Press, 1974; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 1984; R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; D. T. Rasmussen, The evolution of Hyracoidea: A review of the fossil evidence, in D. R. Prothero and R. M. Schoch (eds.), *The Evolution of Perissodactyls*, Oxford University Press, 1989.

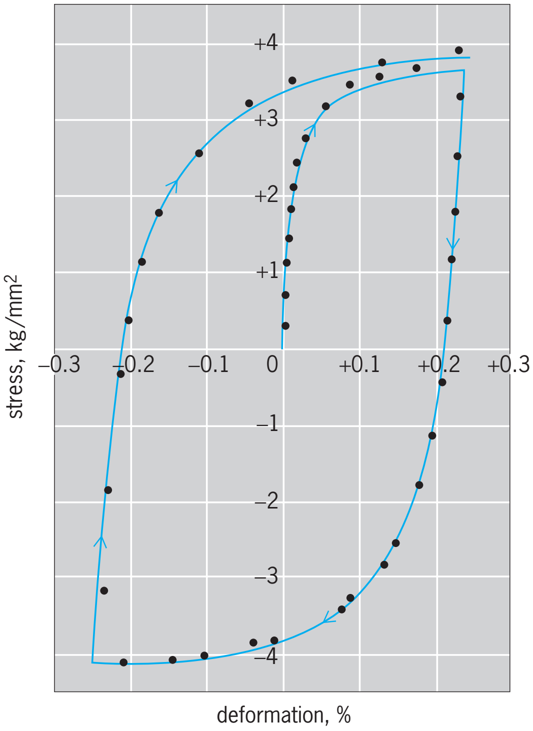
Hysteresis

A phenomenon wherein two (or more) physical quantities bear a relationship which depends on prior history. More specifically, the response Y takes on different values for an increasing input *X* than for a decreasing *X*.

If one cycles *X* over an appropriate range, the plot of *Y* versus *X* gives a closed curve which is referred to as the hysteresis loop. The response *Y* appears to be lagging the input *X*.

Hysteresis occurs in many fields of science. Perhaps the primary example is of magnetic materials where the input variable *H* (magnetic field) and response variable *B* (magnetic induction) are traditionally chosen. For such a choice of conjugate variables, the area of the hysteresis loop takes on a special significance, namely the conversion of energy per unit volume to heat per cycle. In a similar way the dielectric energy loss per unit volume per cycle can be obtained by using the variables *E* (electric field) and *D* (electric displacement). For mechanical hysteresis, it is customary to take the variables stress and strain (see **illus.**), where the energy density loss per cycle is related to the internal friction. In the framework of particle rather than continuum mechanics, the energy loss per cycle is given by the hysteresis loop for the applied force and the particle displacement. Thermal hysteresis is characteristic of many systems, particularly those involving phase changes, but here the hysteresis loops are not usually related to energy loss. *See* FERROELECTRICS; STRESS AND STRAIN; THERMAL HYSTERESIS.

Hysteresis is intimately associated with microscopic irreversibility. In the electric and magnetic situations the motion of domain walls in ordered structures is impeded by microscopic inhomogeneities in the matrix which catch and then let go under increasing stress. Such effects are clearly irreversible and in the magnetic case are evident as the Barkhausen noise. On the other hand, any domain rotation is largely reversible and does not contribute to the hysteresis loss. In the mechanical situation, the pinning and unpinning of dislocations gives rise to the hysteresis. *See* BARKHAUSEN EFFECT; CRYSTAL DEFECTS; DOMAIN (ELECTRICITY AND MAGNETISM).



Plot of stress versus strain (deformation) for a single crystal of brass. The positive and negative strains correspond to elongation and compression, respectively. (*After G. Sachs and H. Shoji, Zug-Druckversuche an Messing-Kristallen (Bauschinger effect), Z. Physik, 45:776–796, 1927*)

Time-dependent effects, where the final equilibrium value of the history-independent response *Y* occurs some time after the input *X* is applied, are complex and are not usually referred to as hysteresis effects. However, cycled experiments (resonance and relaxation) whose frequencies are comparable with the natural frequencies of the system will display an *X*-versus-*Y* dependence of the hysteretic form. The largest values of the hysteresis area will occur when the applied frequency coincides closely with natural frequencies of resonance or relaxation. H. B. Huntington; R. K. MacCrone

Hysteresis motor

A type of synchronous motor in which the rotor consists of a central nonmagnetic core upon which are mounted rings of magnetically hard material. The rings form a thin cylindrical shell of material with a high degree of magnetic hysteresis. The cylindrical stator structure is identical to that of conventional induction or synchronous motors and is fitted with a three-phase or a single-phase winding, with an auxiliary winding and series capacitor for single-phase operation. *See* INDUCTION MOTOR; SYNCHRONOUS MOTOR.

When the motor is running at synchronous speed, the hysteresis material is in a constant state of magnetization and acts as a permanent magnet. Full-speed performance is therefore exactly the same as in a permanent-magnet synchronous motor.

The outstanding special feature of a hysteresis motor is the production of nearly constant, ripple-free torque during starting. At this time the stator current is large, and the rotor hysteresis material is continuously cycled around its hysteresis loop. The resulting rotor flux is delayed in time (and hence in space) behind the rotating magnetomotive force (mmf) produced by the stator currents. This spatial angle between the stator mmf and rotor flux determines the starting torque. Because the angle depends only on the hysteresis loop shape and not on the rate at which the loop is traversed, the starting torque is nearly independent of rotor speed. When synchronous speed is attained, the rotor flux is synchronous with the stator mmf and the hysteresis material retains its last operating point as a constant state of magnetization, resulting in operation as a permanent-magnet synchronous motor. The stator current falls abruptly to its normal operating levels as the machine reaches synchronous speed.

Hysteresis motors are widely used in synchronous motor applications where very smooth starting is required, such as in clocks and other timing devices and record-player turntables, where smooth starting torque reduces record slippage. Hysteresis motors are limited to small size by the difficulty of controlling rotor losses caused by imperfections in the stator mmf wave. *See* ALTERNATING-CURRENT MOTOR; MOTOR. Donald W. Novotny

Bibliography. P. L. Cochran, *Polyphase Induction Motors: Analysis, Design, and Application*, 1989; S. A. Nasar, *Permanent Magnet, Reluctance and Self-Synchronous Motors*, 1993; S. A. Nasar and L. E. Unnewehr, *Electromechanics and Electric Machines*, 1979.

Hysteriales (lichenized)

An order of the Ascolichenes shared by the Ascomycetes. This order formerly included all lichens with linear elongate ascocarps called hysterothecia. But species with true paraphyses are now classified in the family Graphidaceae of the Lecanorales, leaving in the lichenized Hysteriales only those species with a so-called ascolocular structure. *See* LECANORALES.

The hymenium consists of densely interwoven, branched pseudoparaphyses and irregularly scattered locules in which the asci are located. At times it is difficult to distinguish branched paraphyses from pseudoparaphyses, and a few genera with hysterothecia are transitional between the Hysteriales and the Graphidaceae.

The growth form of the Hysteriales is either crustose or fruticose. There are five or six families now included here. Arthoniaceae is the largest family, with two genera, *Arthonia* and *Arthothelium*, both widespread on bark, rarely on leaves, in temperate and tropical regions. The ascocarps are exceedingly small and irregularly branched. The Opegraphaceae includes four major genera with larger elongate ascocarps, crustose on bark and rocks. The Roccellaceae is a family of conspicuous fruticose species that grow profusely on trees and rocks along the coastlines of Portugal, California, Baja California, and parts of western South America. The ascocarps are long and linear. These plants were collected by the ton in the Middle Ages for use as dyestuffs. These same lichens, on a smaller scale, now yield the dye used in litmus paper. *See* ASCOMYCOTA.

Mason E. Hale