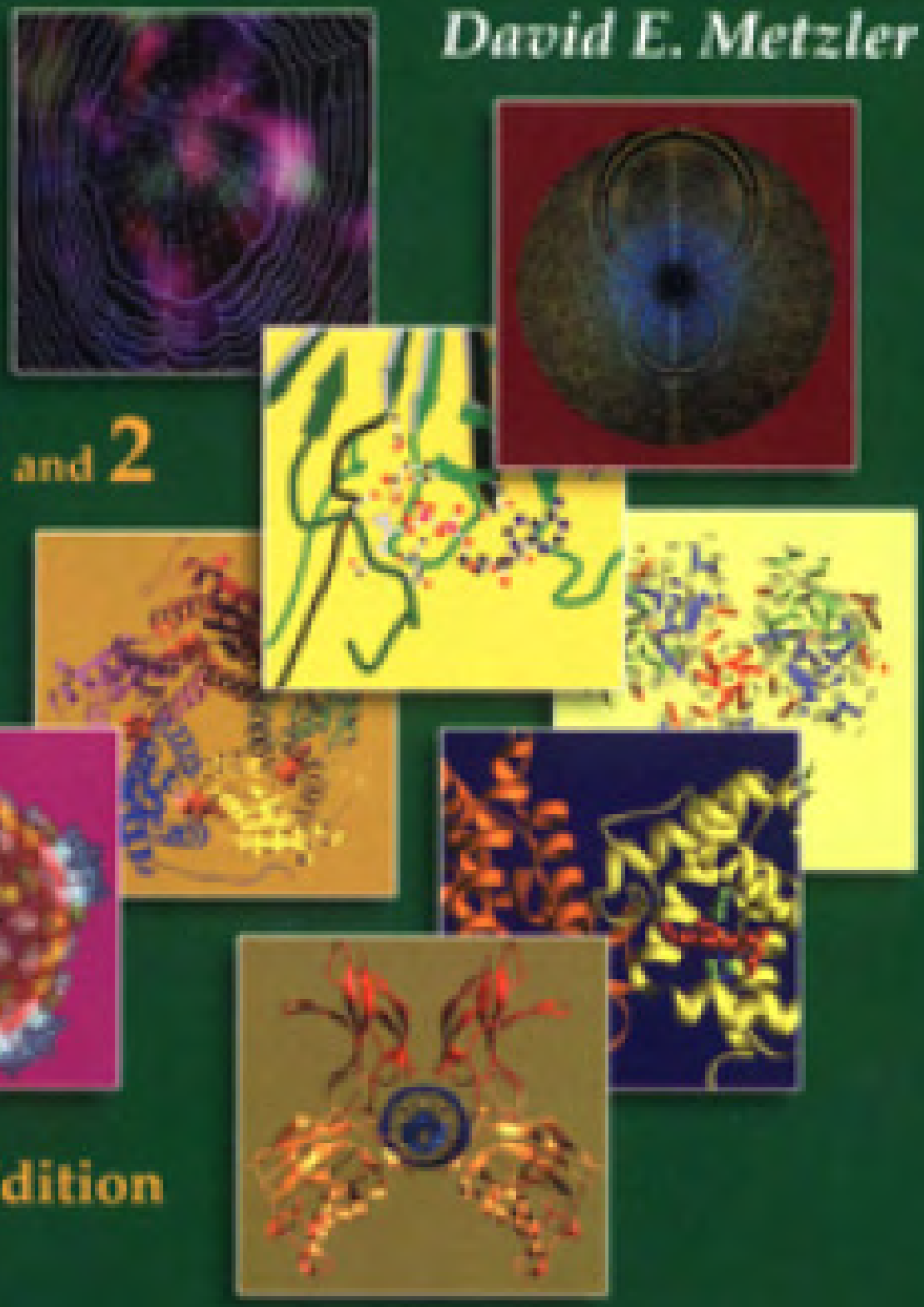
# BIOCHEMISTRY

## The Chemical Reactions of Living Cells

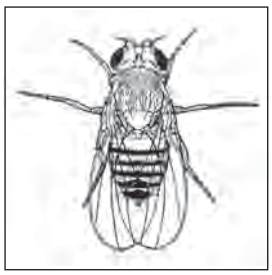
*David E. Metzler*



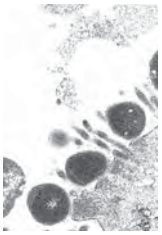
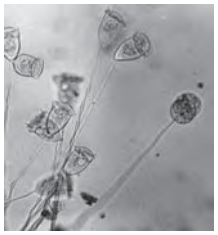
Volumes 1 and 2

Second Edition

# Table of Contents
## *Volume 1*



# Table of Contents
## *Volume 2*

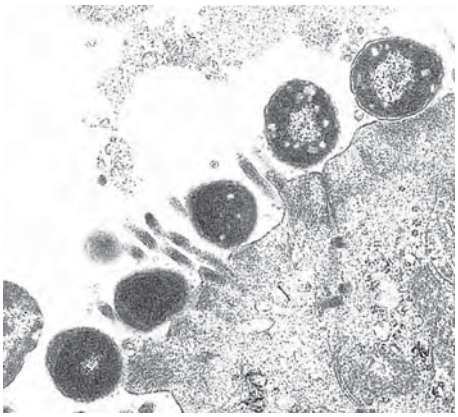




Left: Cells of the pathogenic O157:H7 strain of *Escherichia coli* attached to the surface epithelium of the cecum of a neonatal piglet. Electron-dense filaments (presumably polymerized actin) in the host cytoplasm can be seen subjacent to attached bacteria. The bacteria have effaced most micro-villi but some remain between the bacterial cells. Courtesy of Evelyn A. Dean-Nystrom, National Animal Disease Center, USDA, Agricultural Research Service, Ames, IA. Center: Many unicellular organisms such as these *Vorticella* inhabit wet and moist environments throughout the biosphere. Invertebrates have evolved as long as humans and have complex specializations such as the contractile stem of these protozoa. Courtesy of Ralph Buchsbaum. Right: Although 97% of animals are invertebrates, ~ 3% of the several million known species have backbones. Giraffe: © M. P. Kahl, Photo Researchers

# Contents

# The Scene of Action

# 1



This book is about the chemistry of living cells with special emphasis on the trillions of cells that make up your own body. Every aspect of life depends upon the chemical makeup of cells and on the chemical properties of the remarkable molecules found within the cells. The information presented here will give the reader a solid foundation for understanding not only the chemical basis of life but also the revolutionary developments in molecular biology, biochemical genetics, medicine, and agriculture which dominate today's scientific news and which will play an increasingly important role in our lives.

The first theme of the book is **biomolecular structure**. We'll look carefully at the complex structures of proteins, carbohydrates, RNA, DNA, and many other substances. We'll not only examine in-depth their molecular architecture but also study the chemical properties that make life possible.

A second theme is **metabolism**, the unceasing, complex network of thousands of chemical reactions by which cells grow and reproduce, take up foods and excrete wastes, move, and communicate with each other. Within cells we have a **steady state**, a condition in which the complex chemical constituents of cells are continuously being synthesized in one series of reactions and degraded in another. The result is a marvelous system of self-renewal or "turnover" of tissues. We'll examine the chemical reactions involved in these processes as well as the ways in which they are controlled. We will consider both the reaction sequences and the techniques such as cloning of genes, isotopic labeling, X-ray diffraction, and nuclear magnetic resonance spectroscopy, which are used today to study metabolism.

Human beings are surrounded by many other living creatures whose activities are important to us. Photosynthetic organisms obtain energy from sunlight and synthesize compounds that the human body requires but cannot make. Microorganisms cause decay of organic matter and convert it into forms usable by plants. This book deals with the chemical reactions occurring in all of these organisms. We'll look at strange and unusual reactions, along with those metabolic sequences common to most living things.

Each one of the thousands of chemical reactions of metabolism is catalyzed by an **enzyme**. Most of these enzymes are proteins, but others are made from **RNA** (ribonucleic acid). In both cases enzymes are very large molecules with precise three-dimensional structures. The study of the properties of enzymes and of **enzymatic catalysis** is a third theme of the book. Not only are the chemical mechanisms by which enzymes act of interest but also enzymes are often targets for useful drugs. Incorrectly formed enzymes can result in serious diseases.

The sequences of the amino acids in the chains from which proteins are constructed are encoded in the nucleotide sequences of **DNA** (deoxyribonucleic acid). The coding sequence for a protein in the DNA is found in the **structural gene** for that protein. The RNA enzymes are also encoded by DNA genes. A fourth major theme of the book deals with the nature of the **genetic code** used in DNA and with the processes by which cells read and interpret the code. It also includes study of the methods by which thousands of genes have been mapped to specific positions in chromosomes, isolated, cloned, and sequenced.

A large number of proteins present in the outer surfaces of cells serve as **receptors** that receive chemical messages and other signals from outside the cell. The receptors, which are sometimes enzymes, respond by generating internal signals that control metabolism and cell growth. Such **molecular signaling** is another major area of contemporary biochemistry.

Biologists have described over a million species, and several millions of others probably exist.[1] Many of these organisms have very specialized ways of life. However, they all have much chemistry in common. The same 20 amino acids can be isolated from proteins of plants, animals, and microorganisms. Formation of lactic acid in both bacteria and human muscle requires the same enzymes. Except for some small variations, the genetic code is universal—the same for all organisms. Thus, there is a unity of life and we can study metabolism as the entirety of chemical transformations going on in all living things. However, the differences among species are also impressive. Each species has its own gene for almost every protein.

When the enzyme that catalyzes a particular metabolic reaction is isolated from a number of different organisms, it is usually found to have similar properties and a similar mechanism of catalysis, regardless of the source. However, the *exact* sequence of amino acids in the enzyme will be almost unique to the organism that produced it. When the three-dimensional structures are compared it is found that differences between species often affect only the peripheral parts of an enzyme molecule. The interior structure of the protein, including the catalytic machinery, is highly conserved. However, the surface regions, which often interact with other macromolecules, vary greatly. Such interactions help to control metabolism and may account for many differences in the metabolism among living beings.

Variations in protein structures are not limited to differences between species. Individuals differ from one another. Serious genetic diseases sometimes result from the replacement of a single amino acid unit in a protein by a different amino acid. Genetic deviations from the "normal" structure of a protein result from **mutations**. Many mutations, whether they occurred initially in our own cells or in those of our ancestors, are detrimental.

However, such mutations also account for variation among individuals of a species and allow for evolution. The chemical nature and consequences of mutations and their significance to health, medicine, and agriculture are dealt with throughout the book. We now have reliable methods for inducing in the laboratory mutations at any specific place in a protein sequence and also for synthesizing new DNA sequences. These techniques of **genetic engineering** have given biochemists the ability to modify protein structures freely, to create entirely new proteins, and to provide a basis for the rapidly developing field of **genetic therapy**.

It should be clear from this introduction that **biochemistry** deals with virtually every aspect of life. The distinguishing feature of the science is that it approaches biological questions in terms of the underlying chemistry. The term **molecular biology** is often regarded as synonymous with biochemistry.

However, some scientists use it to imply a more biological approach. These molecular biologists also emphasize structure and function but may have a goal of understanding biological relationships more than chemical details. **Biophysics**, a closely related science, encompasses the application of physical and mathematical tools to the study of life.

## A. The Simplest Living Things

The simplest organisms are the **bacteria.**[2–5] Their cells are called **prokaryotic** (or procaryotic) because no membrane-enclosed nucleus is present. Cells of all other organisms contain nuclei separated from the cytoplasm by membranes. They are called **eukaryotic**. While viruses (Chapter 5) are sometimes regarded as living beings, these amazing parasitic objects are not complete organisms and have little or no metabolism of their own. The smallest bacteria are the **mycoplasmas**.[6–8] They do not have the rigid cell wall characteristic of most bacteria. For this reason they are easily deformed and often pass through filters designed to stop bacteria. They are nutritionally fussy and are usually, if not always, parasitic. Some live harmlessly in mucous membranes of people, but others cause diseases.

### BOX 1-A ABOUT MEASUREMENTS

In 1960 the International General Conference on Weights and Measures adopted an improved form of the metric system, **The International System of Units** (SI). The units of mass, length, and time are the kilogram (kg), meter (m), and second (s). The following prefixes are used for fractions and multiples:

| | | |
|---|---|---|
| $10^{-18}$, *atto* (a) | $10^{-6}$, *micro* (μ) | $10^{9}$, *giga* (G) |
| $10^{-15}$, *femto* (f) | $10^{-3}$, *milli* (m) | $10^{12}$, *tera* (T) |
| $10^{-12}$, *pico* (p) | $10^{3}$, *kilo* (k) | $10^{15}$, *peta* (P) |
| $10^{-9}$, *nano* (n) | $10^{6}$, *mega* (M) | $10^{18}$, *exa* (E) |

There is an inconsistency in that the prefixes are applied to the gram (g) rather than to the basic unit, the kilogram.

SI units have been used throughout the book whenever possible. There are no feet, microns, miles, or tons. Molecular dimensions are given uniformly in nanometers rather than in angstrom units (Å; 1Å = 0.1 nm). Likewise the calorie and kilocalorie have been replaced by the SI unit of energy, the **joule** (J; 1 calorie = 4.184 J).

Throughout the book frequent use is made of the following symbols:

, "appr oximately equal to"

~, "approximately" or "about"

For example, M*ycoplasma pneumoniae* is responsible for primary atypical pneumonia.

Cells of mycoplasmas sometimes grow as filaments but are often spherical and as small as 0.3 micrometer (μm) in diameter. Their outer surface consists of a thin **cell membrane** about 8 nanometers (nm) thick. This membrane encloses the **cytoplasm**, a fluid material containing many dissolved substances as well as submicroscopic particles. At the center of each cell is a single, highly folded molecule of DNA, which constitutes the bacterial chromosome. Besides the DNA there may be, in a small spherical mycoplasma, about 1000 particles ~20 nm in diameter, the **ribosomes**. These ribosomes are the centers of protein synthesis. Included in the cytoplasm are many different kinds of proteins, but there is room for a total of only about 50,000 protein molecules. Several types of RNA as well as many smaller molecules are also present. Although we don't know what minimum quantities of proteins, DNA, and other materials are needed to make a living cell, it is clear that they must all fit into the tiny cell of the mycoplasma.

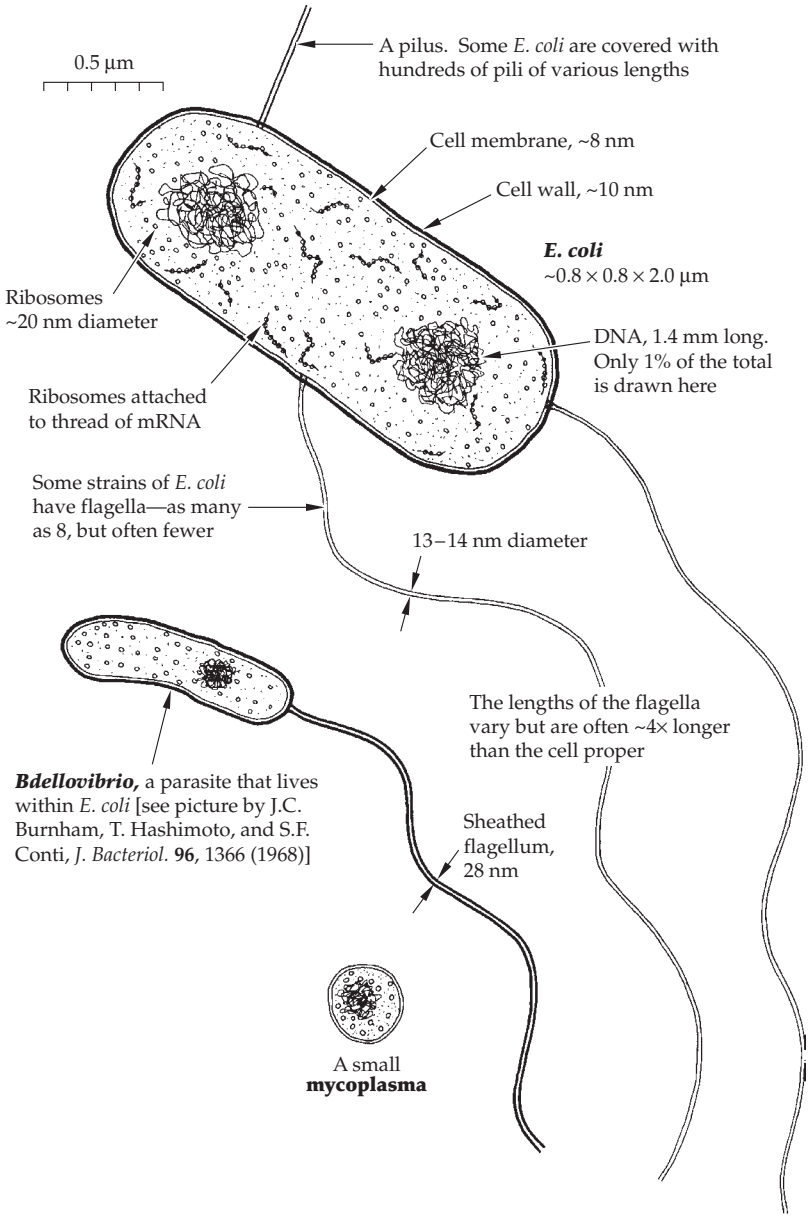



**Figure 1-1** *Escherichia coli* and some smaller bacteria.

## 1. *Escherichia coli*

The biochemist's best friend is *Escherichia coli*, an ordinarily harmless inhabitant of our intestinal tract. This bacterium is easy to grow in the laboratory and has become the best understood organism at the molecular level.[4,9] It may be regarded as a typical true bacterium or **eubacterium**. The cell of *E. coli* (Figs. 1-1, 1-2) is a rod ~2 μm long and 0.8 μm in diameter with a volume of ~1 $\mu m^3$ and a density of ~1.1 $g/cm^3$. The mass is ~1 x $10^{-12}$ g, i.e., 1 picogram (pg) or ~0.7 x $10^{12}$ daltons (Da) (see Box 1-B).[4] It is about 100 times bigger than the smallest mycoplasma but the internal structure, as revealed by the electron microscope, resembles that of a mycoplasma.

Each cell of *E. coli* contains from one to four identical DNA molecules, depending upon how fast the cell is growing, and ~15,000–30,000 ribosomes. Other particles that are sometimes seen within bacteria include food stores such as fat droplets and granules (Fig. 1-3). The granules often consist of **poly-β-hydroxybutyric acid**[10] accounting for up to 25% of the weight of *Bacillus megaterium*. **Polymetaphosphate**, a highly polymerized phosphoric acid, is sometimes stored in "metachromatic granules." In addition, there may be droplets of a separate aqueous phase, known as **vacuoles**.

## 2. The Bacterial Genome

The genetic instructions for a cell are found in the **DNA molecules**. All DNA is derived from four different kinds of monomers, which we call **nucleotides**. DNA molecules are double-stranded: two polymer chains are coiled together, their nucleotide units being associated as **nucleotide pairs** (see Fig. 5-7). The genetic messages in the DNA are in the form of

sequences of nucleotides. These sequences usually consist of a series of code "words" or **codons**. Each codon is composed of three successive nucleotides and specifies which one of the 20 different kinds of amino acids will be used at a particular location in a protein. The sequence of codons in the DNA tells a cell how to order the amino acids for construction of its many different proteins.

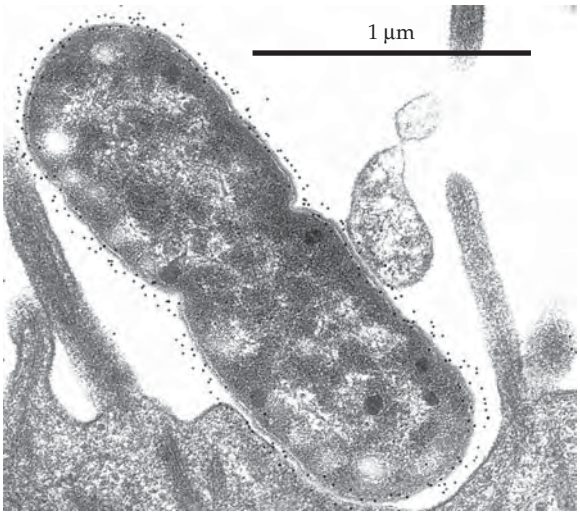



**Figure 1-2** Transmission electron micrograph of a dividing cell of *Escherichia coli* O157:H7 attached to the intestinal epithelium of a neonatal calf. These bacteria, which are able to efface the intestinal microvilli, form characteristic attachments, and secrete shiga toxins, are responsible for ~73,000 illnesses and 60 deaths per year in the U. S.[10a] After embedding, the glutaraldehyde-fixed tissue section was immunostained with goat anti-O157 IgG followed by protein A conjugated to 10-nm gold particles. These are seen around the periphery of the cell bound to the O-antigen (see Fig. 8-28). Notice the two microvilli of the epithelium. Courtesy of Evelyn A. Dean-Nystrom, National Animal Disease Center, USDA, Agricultural Research Service, Ames, IA.

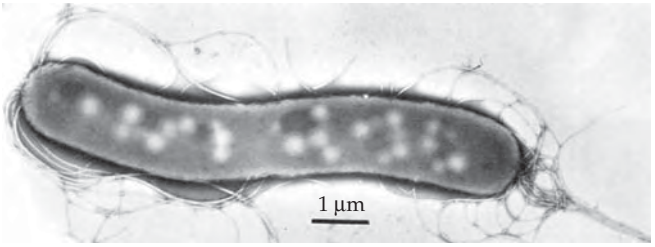



**Figure 1-3** A cell of a *Spirillum* negatively stained with phosphotungstic acid. Note the tufts of flagella at the ends, the rough appearance of the outer surface, the dark granules of poly-β-hydroxybutyric acid and the light-colored granules of unknown nature. Courtesy of F. D. Williams, Gail E. VanderMolen, and C. F. Amstein.

Assume that a typical protein molecule consists of a folded chain of 400 amino acids. Its structural gene will therefore be a sequence of 1200 nucleotide pairs. Allowing a few more nucleotides to form spacer regions between genes we can take ~1300 as the number of nucleotide pairs in a typical bacterial gene. However, some genes may be longer and some may be much shorter. The **genome** is the quantity of DNA that carries a complete set of genetic instructions for an organism. In bacteria, the genome is a single chromosome consisting of one double-stranded DNA molecule. *Mycoplasma genitalium* is the smallest organism for which the DNA sequence is known.[11] Its genome is a double-helical DNA circle of 580,070 nucleotide pairs and appears to contain about 480 genes (an average of ~1200 nucleotides per gene).

The average mass of a nucleotide pair (as the disodium salt) is 664 Da. It follows that the DNA of *M. genitalium* has a mass of ~385 x $10^6$ Da. The relative molecular mass ($M_r$) is 0.385 x $10^9$ (See Box 1-B for definitions of dalton and $M_r$). The DNA of *E. coli* is about seven times larger with a mass of ~2.7 x $10^9$ Da. It contains ~4.2 x $10^6$ nucleotide pairs and encodes over 4000 different proteins (see Table 1-3).

Each nucleotide pair contributes 0.34 nm to the length of the DNA molecule; thus, the total length of DNA of an *E. coli* chromosome is 1.4 mm. This is about 700 times the length of the cell which contains it. Clearly, the molecules of DNA are highly folded, a fact that accounts for their appearance in the electron microscope as dense aggregates called **nucleoids**, which occupy about one-fifth of the cell volume (Fig. 1-4).

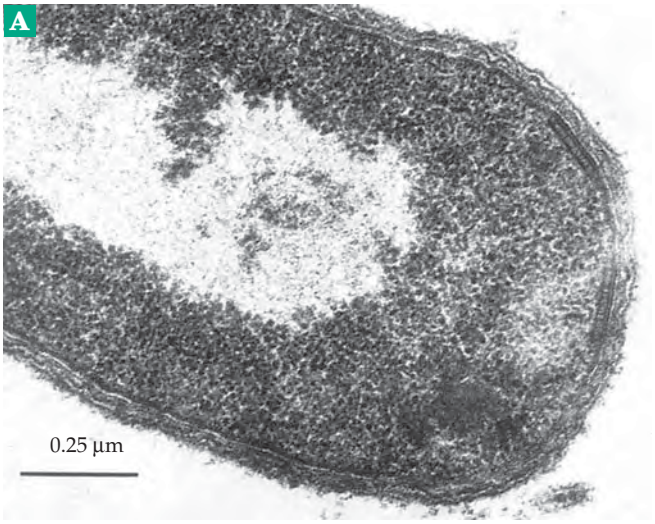
## BOX 1-B RELATIVE MOLECULAR MASS, $M_r$, AND DALTONS

Atomic and molecular masses are assigned relative to the mass of the carbon isotope, $^{12}C$, whose atomic weight is defined as exactly 12. The actual mass of a single atom of $^{12}C$ is defined as 12 **daltons**, one dalton being 1.661 x $10^{-24}$ g. The mass of a molecule can be given in daltons **(Da)** or kilodaltons **(kDa)**. This molecular mass in daltons is numerically equivalent to the relative molecular mass **($M_r$)** or molecular weight **(MW)**[a] and also to the molar mass **(g/mol)**. However, it is not correct to use the dalton for the unitless quantity $M_r$. Masses of structures such as chromosomes, ribosomes, mitochondria, viruses, and whole cells as well as macromolecules can be given in daltons.[b]

[a] The Union of Pure and Applied Chemistry renamed molecular weight as **relative molecular mass** with the symbol $M_r$; $M_r$ = MW.

[b] J. T. Edsall (1970) *Nature* (*London*) **228**, 888.

Each bacterial nucleoid contains a complete set of genetic "blueprints" and functions independently. Each nucleoid is **haploid**, meaning that it contains only a single complete set of genes. In addition to their chromosome, bacteria often contain smaller DNA molecules known as **plasmids**. These plasmids also carry genetic information that may be useful to bacteria. For example, they often encode proteins that confer resistance to antibiotics. The ability to acquire new genes from plasmids is one mechanism that allows bacteria to adapt readily to new environments.[12] Plasmids are also used in the laboratory in the cloning of genes and in genetic engineering (Chapter 26).



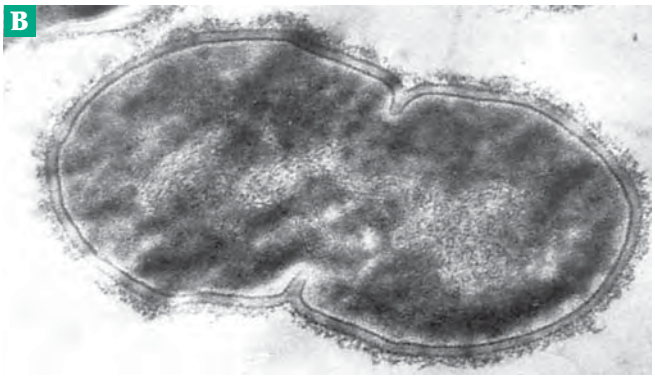




**Figure 1-4** (A) Thin (~60 nm) section of an aquatic gram-negative bacterium, *Aquaspirillum fasciculus*. Note the light-colored DNA, the dark ribosomes, the double membrane characteristics of gram-negative bacteria (Chapter 8, Section E), and the cell wall. In addition, an internal "polar membrane" is seen at the end. It may be involved in some way in the action of the flagella. (B) A thin section of dividing cell of *Streptococcus*, a gram-positive organism. Note the DNA (light-stranded material). A portion of a mesosome is seen in the center and septum can be seen forming between the cells. Micrographs courtesy of F. D. Williams, Gail E. VanderMolen, and C. F. Amstein.

## 3. Ribonucleic Acids (RNA) and the Transcription and Translation of Genetic Information

The genetic information in the DNA is not utilized directly by the protein-synthesizing machinery of cells. Instead, molecules of ribonucleic acid (RNA) are synthesized according to the instructions encoded in the DNA, a process called **transcription**. Although they differ from DNA significantly in their structure, these RNA molecules carry the same coded information as is found in a length of DNA that contains one or a few genes. If DNA is regarded as the "master blueprint" of the cell, molecules of RNA are "secondary blueprints." This concept is embodied in the name **messenger RNA** (mRNA) which is applied to a small, short-lived fraction of RNA that carries information specifying amino acid sequences of proteins. Each molecule of mRNA carries the genetic message from one or more genes to the ribosomes where the proteins are made.

Ribosomes are extraordinarily complex little protein-synthesizing machines. Each ribosome of *E. coli* has a mass of 2.7 x $10^6$ Da and contains 65% of a stable **ribosomal RNA** and ~35% protein. About 50 different kinds of protein molecules are present as parts of the ribosomal structure. Working together with a variety of **transfer RNA** molecules and enzymes, the ribosomes are able to read the genetic messages from mRNA and to accurately assemble any kind of protein molecule that a gene may specify. This process is called **translation** of the genetic message.

## 4. Membranes and Cell Walls

Like the mycoplasma, the *E. coli* cell is bounded by an 8-nm membrane which consists of ~50% protein and 50% lipid. When "stained" (e.g., with permanganate) for electron microscopy, this single membrane appears as two very thin (2.0 nm) dark lines separated by an unstained center band (~3.5 nm) (Fig. 1-4; see also Fig. 8-4). Single membranes of approximately the same thickness and staining behavior occur in all cells, both of bacteria and of eukaryotes.

A cell membrane is much more than just a sack. It serves to control the passage of small molecules into and out of the cell. Its outer surface carries receptors for recognition of various materials. The inside surface of bacterial membranes contains enzymes that catalyze most of the oxidative metabolism of the cells. Bacterial cell membranes are sometimes folded inward to form internal structures involved in photosynthesis or other specialized reactions of metabolism such as oxidation of ammonia to nitrate.[2] In *E. coli* replication of DNA seems to occur on certain parts of the membrane surface, probably under the control of membrane-bound enzymes. The formation of the new membrane which

divides multiplying cells proceeds synchronously with the synthesis of DNA.

A characteristic of true bacteria (**eubacteria**) is a rigid **cell wall** which surrounds the cell membrane. The 40-nm-thick wall of *E. coli* is a complex, layered structure five times thicker than the cell membrane. Its chemical makeup is considered in Chapter 8. One of the layers is often referred to as the **outer membrane**. In some bacteria the wall may be as much as 80 nm thick and may be further surrounded by a thick **capsule** or **glycocalyx** (slime layer).[13] The main function of the wall seems to be to prevent osmotic swelling and bursting of the bacterial cell when the surrounding medium is hypotonic.

If the osmotic pressure of the medium is not too low, bacterial cell walls can sometimes be dissolved, leaving living cells bounded only by their membranes. Such **protoplasts** can be produced by action of the enzyme lysozyme on gram-positive bacteria such as *Bacillus megaterium*. Treatment of cells of gram-negative bacteria with penicillin (Box 20-G) produces **spheroplasts**, cells with partially disrupted walls. Spheroplasts and protoplasts are useful in biochemical studies because substances enter cells more readily when the cell wall is absent. Strains of bacteria lacking rigid walls are known as **L forms**.

## 5. Flagella and Pili

Many bacteria swim at speeds of 20–60 μm/s, ten or more body lengths per second! Very thin thread-like **flagella** of diameter 13–20 nm coiled into a helical form are rotated by the world's smallest "electric motors" to provide the motion.[14] While some bacteria have a single flagellum, the corkscrew-like *Spirillum* (Fig. 1-3) synchronously moves tufts of flagella at both ends. Some strains of *E. coli* have no flagella, but others contain as many as eight flagella per cell distributed over the surface. The flagella stream out behind in a bundle when the bacterium swims. The flagella of the helical **spirochetes** are located inside the outer membrane.[15,16]

In addition to flagella, extremely thin, long, straight filaments known as **pili** or **fimbriae** (Fig. 1-2) project from the surfaces of many bacteria.[14] The "sex pili" (F pili and I pili) of *E. coli* have a specific role in sexual conjugation. The similar but more numerous common pili or fimbriae range in thickness from 3 to 25 nm and in length from 0.2 to 2 μm. Pili are involved in adhesion of bacteria to surrounding materials or to other bacteria and facilitate bacterial infections.[17–19] A typical *E. coli* cell has 100–300 pili.[5]

## 6. Classification and Evolution of Bacteria

Bacteria vary greatly in their chemistry and metabolism, and it is difficult to classify them in a rational way. In higher organisms species are often defined as forms that cannot interbreed and produce fertile offspring, but such a criterion is meaningless for bacteria whose reproduction is largely asexual and which are able readily to accept "visiting genes" from other bacteria.[12] The classification into species and genera is therefore somewhat arbitrary. A currently used scheme (Table 1-1)[20] classifies the prokaryotes into 35 groups on the basis of many characteristics including shape, staining behavior, and chemical activities. Table 1-1 also includes genus names of most of the bacteria discussed in this book.

Bacteria may have the shape of spheres or straight or curved rods. Some, such as the **actinomycetes**, grow in a branching filamentous form. Words used to describe bacteria often refer to these shapes: a **coccus** is a sphere, a **bacillus** a rod, and a **vibrio** a curved rod with a flagellum at one end. A **spirillum** is screw-shaped. These same words are frequently used to name particular genera or families. Other names are derived from some chemical activity of the bacterium being described.

The **gram stain** provides an important criterion of classification that depends upon differences in the structure of the cell wall (see Chapter 20). Bacterial cells are described as **gram-positive** or **gram-negative** according to their ability to retain the basic dye crystal violet as an iodine complex. This difference distinguishes two of four large categories of bacteria.[20] Most actinomycetes, the spore-forming bacilli, and most cocci are gram-positive, while *E. coli*, other enterobacteria, and pseudomonads are gram-negative. A third category consists of eubacteria that lack cell walls, e.g. the mycoplasma.

Comparisons of amino acid sequences of proteins and the nucleotide sequences of DNA and RNA have provided a new approach to classification of bacteria.[21–28] Although the origins of life are obscure, we can easily observe that the genome changes with time through mutation and through the enzyme-catalyzed process of **genetic recombination**. The latter gives rise to the deletion of some nucleotides and the insertion of others into a DNA chain. When we examine sequences of closely related species, such as *E. coli* and *Salmonella typhimurium*, we find that the sequences are very similar. However, they differ greatly from those of many other bacteria. Consider the 23S ribosomal RNA, a molecule found in the ribosomes of all bacteria. It contains ~3300 nucleotides in a single highly folded chain. The basic structure is highly conserved but between any two species of bacteria there are many nucleotide substitutions caused by mutations as well as deletions and insertions. By asking what is the minimum number of

mutations that could have converted one 23S RNA into another and by assuming a more or less constant rate of mutation over millions of years it is possible to construct a **phylogenetic tree** such as that shown in Fig. 1-5.

One conclusion from these comparisons is that the methane-producing bacteria, the **methanogens**,[24] are only distantly related to most other bacteria. Methanogens together with the cell wall-less *Thermoplasma*,[28] some salt-loving **halobacteria**, and some **thermophilic** (heat-loving) sulfur bacteria form a fourth major category. They are often regarded as a separate kingdom, the **archaeobacteria**,[25] which together with the kingdom of the eubacteria form the superkingdom prokaryota. Certain archaeobacteria have biochemical characteristics resembling those of eukaryotes and

**TABLE 1-1**
**A Systematic Classification Scheme for Bacteria[a,b]**

**Kingdom Procaryotae**

The bacteria are classified according to the following 35 groups. Within these groups many genera are classified into subgroups or families. A few genera, most of which are mentioned elsewhere in this book, are listed here by name. Members of a single subgroup are placed together and are separated by semicolons from members of other subgroups.

1. The spirochetes (long bacteria, up to 500 μm, that are propelled by the action of filaments wrapped around the cell between the membrane and wall). *Borrelia* (*B. burgdorferi*, Lyme disease), *Leptospira*, *Treponema* (*T. pallidum*, syphilis)
2. Aerobic spiral and curved motile gram-negative bacteria. *Bdellovibrio*, *Campylobacter* (*C. jejuni*, diarrhea), *Helicobacter* (*H. pylori*, gastric ulcers), *Spirillum*
3. Nonmotile gram-negative curved bacteria
4. Gram-negative, aerobic rods and cocci. *Acetobacter, Agrobacterium, Azotobacter, Brucella* (*B. abortus*, brucellosis), *Flavobacterium, Gluconobacter, Legionella* (*L. pneumophila*, Legionnaire's disease), *Methylomonas*, *Neisseria* (*N. gonorrhea*, gonorrhea), *Pseudomonas*, *Rhizobium*, *Thermus, Xanthomonas, Rochalimaea* (*R. henselae*, cat scratch disease)
5. Gram-negative, facultatively anaerobic rods. *Enterobacter*,[c] *Proteus*, *Yersinia* (*Y. pestis*, plague), *Escherichia, Klebsiella, Salmonella* (*S. typhi*, typhoid fever), *Serratia, Shigella* (*S. dysenteriae*, bacterial dysentery), *Haemophilus; Vibrio* (*V. cholerae*, Asiatic cholera); *Zymomonas*
6. Gram-negative, anaerobic bacteria. *Butyrivibrio*
7. Dissimilatory sulfate- or sulfur-reducing bacteria. *Desulfovibrio*
8. Anaerobic gram-negative cocci. *Veillonella*
9. The rickettsias (parasitic bacteria with exacting nutritional requirements and small genome sizes) and chlamydias. *Chlamydia* (*C. trachomatis*, trachoma), *Rickettsia* (*R. rickettsii*, Rocky Mountain spotted fever)
10. Anoxygenic photosynthetic bacteria. Green sulfur bacteria. *Chlorobium, Prosthecochloris;* purple nonsulfur bacteria: *Rhodopseudomonas, Rhodospirillum;* purple sulfur bacteria: *Chromatium, Thiospirillum*
11. Oxygenic photosynthetic bacteria. Cyanobacteria (blue-green algae): *Synechocystis; Anabaena, Nostoc; Oscillatoria*
12. Aerobic, chemolithotrophic bacteria. Colorless sulfur bacteria: *Thiobacillus;* iron or manganese-oxidizing bacteria, magnetotactic bacteria; nitrifying bacteria: *Nitrobacter, Nitrosomonas*
13. Budding and/or appendaged bacteria. *Caulobacter*
14. Sheathed bacteria
15. Nonphotosynthetic, nonfruiting gliding bacteria. *Beggiatoa* (a filamentous bacterium containing sulfur granules)
16. Fruiting, gliding bacteria. *Myxococcus*
17. Gram-positive cocci. *Leuconostoc, Micrococcus, Peptococcus, Staphylococcus* (*S. aureus,* boils, infections), *Streptococcus* (*S. pyogenes,* scarlet fever, throat infections, *S. pneumoniae,* pneumonia)
18. Endospore-forming gram-positive rods and cocci. Aerobic: *Bacillus* (*B. anthracis,* anthrax), anaerobic: *Clostridium* (*C. tetani,* tetanus; *C. botulinum,* botulism)
19. Regular nonsporing gram-positive rods. *Lactobacillus*
20. Irregular nonsporing gram-positive rods. *Actinomyces, Bifidobacterium, Corynebacterium* (*C. diphtheriae,* diphtheria), *Propionibacterium*
21. Mycobacteria. *Mycobacterium* (*M. tuberculosis,* tuberculosis; *M. leprae,* leprosy)

22-29. Actinomycetes

30. Mycoplasmas. *Acholeplasma, Mycoplasma*
31. Methanogens. *Methanobacterium; Methanosarcina; Methanospirillum*
32. Archaeal sulfate reducers
33. Halobacteria. *Halobacterium*
34. Cell wall-less archaeobacteria. *Thermoplasma*
35. Very thermophilic $S^0$-Metabolizers. *Sulfolobus; Thermococcus*

[a] From Bergey's Manual of Systematic Bacteriology, 9th ed. J. G. Holt, N. R. Krieg, P. H. A. Sneath, J. T. Staley and S. T. Williams, Eds. (1994) Williams and Wilkins, Baltimore, Maryland. For another recent list see http://www.ncbi.nlm.nih.gov/

[b] The human diseases caused by some species are also listed.

[c] Formerly *Aerobacter*.

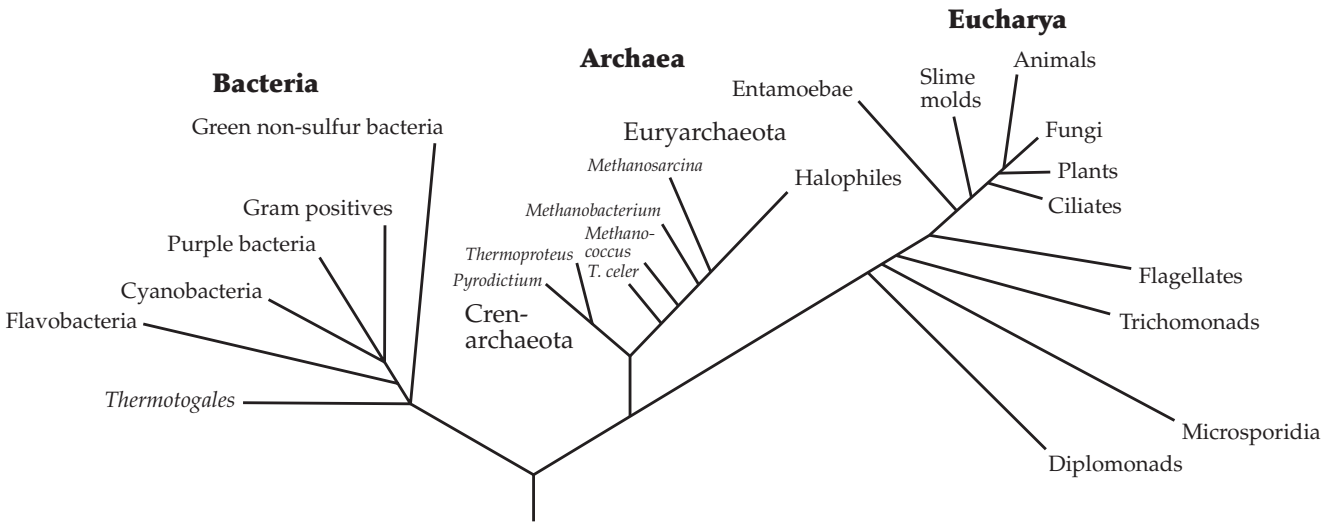



**Figure 1-5** Universal phylogenetic tree. From Wheelis *et al.*[29]

some biologists therefore classify them as **archaea** and rank their kingdom as equal to that of the bacteria and the eukaryotes (Fig. 1-5).[27,29,30,30a,30b] Others disagree.[31] In Table 1-1, the archaeobacteria are found in groups 31–35. Most bacteria are very small in size but there are species large enough to be confused with eukaryotic protozoa. The record for bacteria seems to be held by *Epulopiscium fishelsoni*, a parasite of the surgeonfish intestinal tract. A single cell measured > 600 μm by 80 μm diameter, over $10^6$ times larger in volume than a cell of *E. coli*.[32] The organism is a gram-positive bacterium as judged by analysis of its cloned ribosomal RNA genes.

## 7. Nutrition and Growth of Bacteria

**Autotrophic** (self-nourishing) bacteria can synthesize all of their organic cell constituents from carbon dioxide, water, and inorganic forms of nitrogen and sulfur. The **photoautotrophs** extract their energy from sunlight, while the **chemoautotrophs** obtain energy from inorganic chemical reactions. For example, the hydrogen bacteria oxidize $H_2$ to $H_2O$ and sulfur bacteria oxidize $H_2S$ to $H_2SO_4$. Like the fungi and animals, most bacteria are **chemoheterotrophic**; they obtain energy from the breakdown of organic compounds. Some of these heterotrophic bacteria are **anaerobes** which live without $O_2$. Many of them metabolize complex organic substances such as sugars in the absence of oxygen, a process called **fermentation**. Others oxidize organic compounds with an inorganic oxidant such as nitrite or sulfate. Members of the genus *Clostridium* are poisoned by oxygen and are known as **obligate anaerobes**. Others, including *E. coli*, are **facultative anaerobes**, able to grow either in the presence or in the absence of oxygen. **Obligate aerobes** depend for energy upon combustion of organic compounds with oxygen.

One of the largest groups of strictly aerobic heterotrophic bacteria, the pseudomonads (*Pseudomonas* and related genera), are of interest to biochemists because of their ability to oxidize organic compounds, such as alkanes, aromatic hydrocarbons, and steroids, which are not attacked by most other bacteria. Often, the number of oxidative reactions used by any one species of bacteria is limited. For example, the acetic acid bacteria that live in wine and beer obtain all of their energy by oxidation of ethanol to acetic acid:

$$CH_3CH_2OH + O_2 \rightarrow CH_3COOH + H_2O$$

Bacteria can grow incredibly fast. Under some conditions, it takes a bacterial cell only 10–20 min to double its size and to divide to form two cells.[4] An animal cell may take 24 h for the same process. Equally impressive are the rates at which bacteria transform their foods into other materials. One factor contributing to the high rate of bacterial metabolism may be the large surface to volume ratio. For a small spherical bacterium (coccus) of diameter 0.5 μm, the ratio of the surface area to the volume is $12 \times 10^6$ $m^{-1}$, while for an ameba of diameter 150 μm the ratio is only $4 \times 10^4$ $m^{-1}$ (the ameba can increase this by sticking out some pseudopods). Thimann[33] estimated that for a 90-kg human, the ratio is only 30 $m^{-1}$.

When food is limited, some bacteria such as the *Bacillus* form **spores**. These are compact little cells that form inside the vegetative cell and are therefore called **endospores**. They sometimes have only 1/10 the volume of the parent cell. Their water content is very low, their metabolic rate is near zero, and they are

extremely resistant to heat and further desiccation. Under suitable conditions, the spores can "germinate" and renew their vegetative growth. Spore formation is one of several examples of the development of specialized cells or **differentiation** among prokaryotes.

## 8. Photosynthetic and Nitrogen-Fixing Prokaryotes

It is likely that the earth was once a completely anaerobic place containing water, ammonia, methane, formaldehyde, and more complicated organic compounds. Perhaps the first forms of life, which may have originated about $3.5 \times 10^9$ years ago, resembled present-day anaerobic bacteria. The purple and green photosynthetic bacteria may be related to organisms that developed at a second stage of evolution: those able to capture energy from sunlight. Most of these gram-negative photosynthetic bacteria are strict anaerobes. None can make oxygen as do higher plants. Rather, the hydrogen needed to carry out the reduction of carbon dioxide in the photosynthetic process is obtained by the splitting of inorganic compounds, such as $H_2S$, thiosulfate, or $H_2$, or is taken from organic compounds. Today, photosynthetic bacteria are found principally in sulfur springs and in deep lakes, but at one time they were probably far more abundant and the only photosynthetic organisms on earth.

Before organisms could produce oxygen a second complete photosynthetic system, which could cleave $H_2O$ to $O_2$, had to be developed. The simplest oxygen-

## BOX 1-C IN THE BEGINNING

No one knows how life began. Theories ranging from the biblical accounts to recent ideas about the role of RNA are plentiful but largely unsatisfying. In the 1800s the great physical chemist Arrhenius was among scientists that preferred the idea held by some scientists today that a "seed" came from outer space. Until recently the only concrete data came from fossils. Making use of a variety of isotopic dating methods it can be concluded that cyanobacteria were present $2.2 \times 10^9$ years ago and eukaryotes $1.4 \times 10^9$ years ago. About $0.5 \times 10^9$ years ago the "Cambrian explosion" led to the appearance of virtually all known animal phyla. Many of these then became extinct about $0.2 \times 10^9$ years ago.

New insights published in 1859[a] were provided by Charles Darwin. However, his ideas were only put into a context of biochemical data after 1950 when sequencing of proteins and later nucleic acids began. From an astonishingly large library of sequence data available now we can draw one firm conclusion: *Evolution can be observed;*[b] *it does involve mutation of DNA*. Comparisons of sequences among many species allow evolutionary relationships to be proposed.[c-e] In general these are very similar to those deduced from the fossil record. They support the idea that evolution occurs by natural selection and that duplication of genes and movements of large pieces of DNA within the genome have occurred often. As many as 900 "ancient conserved regions" of DNA in the *E. coli* genome corresponding to those in human, nematode, and yeast DNA are thought to date back perhaps $3.5 \times 10^9$ years.[f] However, nobody has explained how life evolved before there was DNA.

One of the first scientists to devote his career to biochemical evolution was I. V. Oparin,[g] who published a book on the "origin of life" in 1924. Oparin and J. B. S. Haldane, independently, proposed that early life was anaerobic and that energy was provided by fermentation. In 1951 Stanley Miller built an apparatus that circulated $CH_4$, $NH_3$, $H_2O$, and $H_2$, compounds thought to be present in a primitive atmosphere, past an electric discharge. He found glycine, alanine, β-alanine, and other amino acids among the products formed.[h] Schrödinger pointed out that a flux of energy through a system will tend to organize the system. The solar energy passing through the biosphere induces atmospheric circulation and patterns of weather and ocean currents.[i,j] Perhaps in the primordeal oceans organic compounds arose from the action of light and lightning discharges. These compounds became catalysts for other reactions which eventually evolved into a rudimentary cell-less metabolism. It is a large jump from this to a cell! Among other problems is the lack of any explanation for the development of individual cells or of their genomes. However, because it helps to correlate much information *we will always take an evolutionary approach in this book* and will discuss the "beginnings" a little more in later chapters.

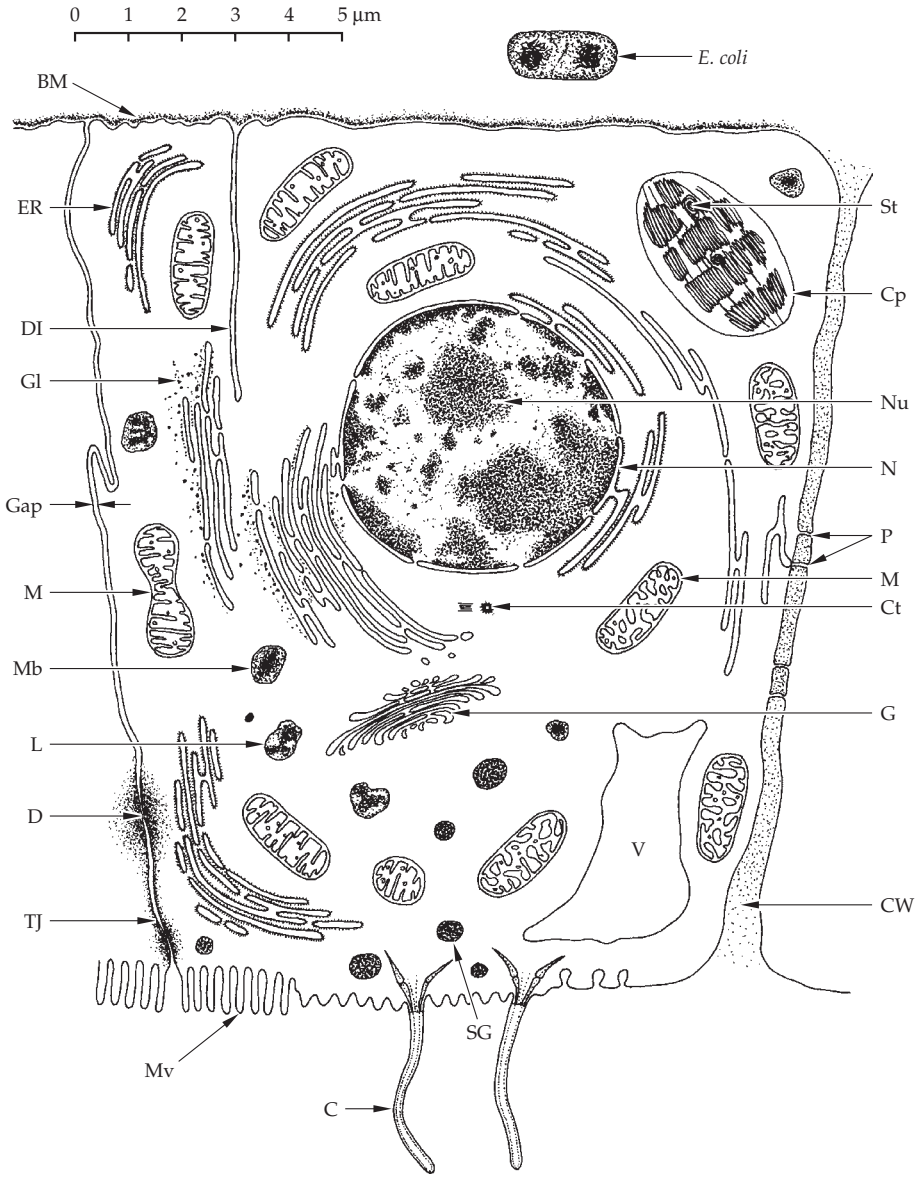
---

[a] Maynard-Smith, J. (1982) *Nature (London)* **296**, 599–601
[b] Lenski, R. E., and Travisano, M. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6808–6814
[c] Wilson, A. C. (1985) *Sci. Am.* **253**(Oct), 164–173
[d] Eigen, M., Gardiner, W., Schuster, P., and Winkler-Oswatitsch, R. (1981) *Sci. Am.* **244**(Apr), 88–118
[e] Doolittle, R. F. (1992) *Protein Sci.* **1**, 191–200
[f] Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., and Claverie, J.-M. (1993) *Science* **259**, 1711–1716
[g] Broda, E. (1980) *Trends Biochem. Sci.* **5**, IV–V
[h] Miller, S. L. (1953) *Science* **117**, 528–529
[i] Mason, S. (1993) *Trends Biochem. Sci.* **18**, 230–231
[j] Welch, G. R. (1996) *Trends Biochem. Sci.* **21**, 452

producing creatures existing today are the **cyanobacteria**,[34] also known as blue-green algae. Many cyanobacteria are unicellular, but others such as *Oscillatoria*, a slimy "plant" that often coats the inside walls of household aquaria, consist of long filaments about 6 μm in diameter (see Fig. 1-11). All cyanobacteria contain two groups of pigments not found in other prokaryotes: **chlorophyll a** and **β-carotene**, pigments that are also found in the chloroplasts of true algae and in higher plants. A recently discovered group of bacteria, the **prochlorophytes**, are even closer to chloroplasts in their pigment composition.[35]

In addition to pigmented cells, some cyanobacteria contain paler cells known as **heterocysts**. They have a specialized function of fixing molecular nitrogen. The development of the ability to convert $N_2$ into organic nitrogen compounds represents another important evolutionary step. Because they can both fix nitrogen and carry out photosynthesis, the blue-green algae have the simplest nutritional requirements of any organisms. They need only $N_2$, $CO_2$, water, light, and minerals for growth.

Evolution of the photosynthetic cleavage of water to oxygen was doubtless a major event with far-reaching consequences. Biologists generally believe that as oxygen accumulated in the earth's atmosphere, the obligate anaerobes, which are poisoned by oxygen, became limited to strictly anaerobic environments. Meanwhile, a new group of bacteria, the **aerobes**, appeared with mechanisms for detoxifying oxygen




Abbreviations:

**BM**, basement membrane
**ER**, rough endoplasmic reticulum (with ribosomes attached; smooth ER is depicted nearer the nucleus and on the right side of the cell.)
**DI**, deep indentation of plasma membrane
**Gl**, glycogen granules
**Gap**, space ~10-20 nm thick between adjacent cells
**M**, mitochondrion
**Mb**, microbody
**L**, lysosome
**D**, desmosome
**TJ**, tight junction
**Mv**, microvilli
**C**, cillium
**SG**, secretion granule
**V**, vacuole
**Nu**, nucleolus
**G**, Golgi apparatus
**CW**, cell wall (of a plant)
**Ct**, centrioles
**P**, plasmodesmata
**N**, nucleus
**Cp**, chloroplast
**St**, starch granule

**Figure 1-6** The "average" eukaryotic cell. This composite drawing shows the principal organelles of both animal and plant cells approximately to the correct scale. (Adapted from a drawing by Michael Metzler.)

and for using oxygen to oxidize complex organic compounds to obtain energy.

## B. Eukaryotic Cells

Cells of the **eukaryotes** contain true nuclei and are much larger and more complex internally than are those of prokaryotes. The nucleus of a cell contains most of its DNA and is separated from the cytoplasm by membranes. Within the cytoplasm are various **organelles** with characteristic structures. These include **mitochondria**, **lysosomes**, **peroxisomes,** and **centrioles**. Eukaryotic cells come in so many sizes and shapes and with so many specialized features that it is impossible to say what is typical. Nevertheless, Fig. 1-6 is an attempt to portray some sort of "average" cell, partly plant and partly animal.

As can be seen from Table 1-2, which lists the diameters and volumes of several roughly spherical cells, there is a great variation in size. However, a diameter of 10–20 µm may be regarded as typical for both plants and animals. For growth of a large cell such as the ovum, many adjacent cells assist in synthesis of foodstuffs which are transferred to the developing egg cell. Plant cells are often large but usually 90% or more of the cell is filled with a **vacuole** or **tonoplast**,[36] which is drawn unrealistically small in Fig. 1-6. The metabolically active protoplasm of plant cells often lies in a thin layer at their peripheries.

Many cells are far from spherical; for example, human red blood cells are discs 8 x 8 x 1 to 2 µm with a volume of 80 $\mu m^3$. Plant fiber cells may be several millimeters in length. Nerve cells of animals have long extensions, the **axons**, which in the human sometimes attain a length of a meter. Muscle cells fuse to give very long multinucleate fibers.

**TABLE 1-2**
**Approximate Sizes of Some Cells**

| Cell | Diameter (µm) | Approximate volume ($\mu m^3$) |
|---|---|---|
| *E. coli* | 1 | 1.0 |
| Small thymus cell | 6 | 120 |
| Liver cell | 20 | 4,000 |
| Human ovum (mature) | 120 | 500,000 |
| Hen's egg (white excluded) | 20,000 | $4 \times 10^{12}$ |
| Yeast cell | 10 | 500 |
| Onion root (meristematic cell) | 17 | 2,600 |
| Parenchyma cell of a fruit | 1,000 | $1 \times 10^{8}$ |

### 1. The Nucleus

In a typical animal cell the nucleus has a diameter of ~5 µm and a volume of 65 $\mu m^3$. Except at the time of cell division, it is densely and almost uniformly packed with DNA. The amount of DNA present is larger than that in bacteria as is indicated in Table 1-3. Yeast contains about three times as much genetic matter as *E. coli* and a human being or a mouse about 700 times as much. However, genes are sometimes duplicated in higher organisms and large amounts of **repetitive DNA** of uncertain significance are often present. Some amphibians have 25 times *more* DNA per cell than do humans. The fruit fly *Drosophila* contains about 13,600 functioning genes and a human being perhaps 50,000.[37]

Because of its acidic character, DNA is stained by basic dyes. Long before the days of modern biochemistry, the name **chromatin** was given to the material in the nucleus that was colored by basic dyes. At the time of cell division, the chromatin is consolidated into distinct **chromosomes** which contain, in addition to 15% DNA, about 10% RNA and 75% protein.

Nearly all of the RNA of the cell is synthesized (transcribed) in the nucleus, according to the instructions encoded in the DNA. Some of the RNA then moves out of the nucleus into the cytoplasm where it functions in protein synthesis and in some other ways. Many eukaryotic genes consist of several sequences that may be separated in the DNA of a chromosome by **intervening sequences** of hundreds or thousands of base pairs. The long RNA transcripts made from these **split genes** must be cut and **spliced** in the nucleus to form the correct messenger RNA molecules which are then sent out to the ribosomes in the cytoplasm.

Each cell nucleus contains one or more dense **nucleoli**, regions that are rich in RNA and may contain 10–20% of the total RNA of cells. Nucleoli are sites of synthesis and of temporary storage of ribosomal RNA, which is needed for assembly of ribosomes. The **nuclear envelope** is a pair of membranes, usually a few tens of nanometers apart, that surround the nucleus. The two membranes of the pair separate off a thin **perinuclear space** (Fig. 1-7). The membranes contain "pores" ~130 nm in diameter with a complex structure (see Fig. 27-8).[38,39] There is a central channel ~42 nm in diameter, which provides a route for controlled passage of RNA and other large molecules from the nucleus into the cytoplasm and also from the cytoplasm to the nucleus. Smaller ~10 nm channels allow passive diffusion of ions and small molecules.

**TABLE 1-3**
**Haploid Genome Sizes for Several Organisms**

| Organism; see footnotes for sequence information | Millions of nucleotide base pairs (Mb) | Number of chromosomes (haploid) | Estimated number of genes |
|---|---|---|---|
| *Mycoplasma genitalium*[a–c] | 0.580 | 1 | 482 |
| *Rickettsia prowazekii*[d] | 1.11 | 1 | 834 |
| *Haemophilus influenzae*[c,e–g] | 1.83 | 1 | 1,709 |
| *Methanococcus janaschii* (an archaeon)[h] | 1.66 | 1 | 1,738 |
| *Bacillus subtilis*[i] | 4.16 | 1 | |
| *Escherichia coli*[j,k] | 4.64 | 1 | 4,288 |
| *Myxococcus xanthis*[f] | 9.2 | 1 | |
| *Synechocystis* sp. (a cyanobacterium)[l] | 3.57 | 1 | 3,169 |
| *Saccharomyces cerevisiae* (a yeast)[k,m-p,t] | 13.5 | 17 | 6,241 |
| *Giardia lamblia* (a protozoan)[q] | 12 | | |
| *Plasmodium falciparum* (malaria parasite)[v] | 25-30 | 14 | |
| *Dictyostelium discoideum* (a slime mold)[r] | 34 | | |
| *Caenorhabditis elegans* (a nematode)[s,t] | 97 | 6 | 18,424 |
| Sea Urchin | 900 | | |
| *Drosophila melanogaster* (fruit fly)[u,t] | 180 | 4 | 13,601 |
| *Danio rerio* (zebrafish)[w] | 1,700 | 2.5 | |
| *Fugu rubripes* (pufferfish)[x] | 400 | | 60,000 |
| S. African lungfish[y] | 102,000 | 19 | |
| *Mus musculus* (mouse)[z] | ~3,000 | 20 | 80,000 |
| *Bos* (cow)[aa] | ~3,000 | 30 | 80,000 |
| *Homo sapiens* (human)[bb,cc] | ~3,000 | 23 | 50,000–150,000 |
| *Arabidopsis thaliana* (green plant)[dd,ee] | 115.4 | 5 | 25,498 |
| Rice[ff-hh] | 450 | 12 | |
| Maize[ii] or Wheat[ff] | ~2,700 | | |
| Lily[jj] | >100,000 | | |

## 2. The Plasma Membrane

The thin (8 nm) outer cell membrane or "plasma-lemma" (Fig. 1-7) controls the flow of materials into and out of cells, conducts impulses in nerve cells and along muscle fibrils, and participates in chemical communication with other cells. Deep infoldings of the outer membrane sometimes run into the cytoplasm. An example, is the "T system" of tubules which functions in excitation of muscle contraction (Figs. 19-7, 19-21). Surfaces of cells designated to secrete materials or to absorb substances from the surrounding fluid, such as the cells lining kidney tubules and pancreatic secretory cells, are often covered with very fine projections or **microvilli** which greatly increase the surface area. In other cases projections from one cell interdigitate with those of an adjacent cell to give more intimate contact.

## 3. Vacuoles, Endocytosis, and Lysosomes

Cells often contain vacuoles or smaller vesicles that are separated from the cytosol by a *single* membrane. Their content is often quite


[a] Fraser, C. M., and 28 other authors (1995) *Science* **270**, 397–403
[b] Goffeau, A. (1995) *Science* **270**, 445–446
[c] Brosius, J., Robison, K., Gilbert, W., Church, G. M., and Venter, J. C. (1996) *Science* **271**, 1302–1304
[d] Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C. M., Podowski, R. M., Näslund, A. K., Eriksson, A.-S., Winkler, H. H., and Kurland, C. G. (1998) *Nature (London)* **396**, 133–140
[e] Fleischmann, R. D., and 39 other authors (1995) *Science* **269**, 496–512
[f] He, Q., Chen, H., Kupsa, A., Cheng, Y., Kaiser, D., and Shimkets, L. J. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 9584–9587
[g] Mrázek, J., and Karlin, S. (1996) *Trends Biochem. Sci.* **21**, 201–202
[h] Bult, C. J., and 39 other authors; corresponding author Venter, J. C. (1996) *Science* **273**, 1058–1073
[i] Azevedo, V., Alvarez, E., Zumstein, E., Damiani, G., Sgaramella, V., Ehrlich, S. D., and Serror, P. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 6047–6051
[j] Blattner, F. R. and 16 other authors (1997) *Science* **277**, 1453–1462
[k] Winzeler, E. A., and 51 other authors (1999) *Science* **285**, 901–906
[l] Kaneko, T., and 23 other authors (1996) *DNA Res.* **3**, 109–136. See also http://www.kazusa.or.jp/cyano
[m] Dujon, B., and 107 other authors (1994) *Nature (London)* **369**, 371–378
[n] Taguchi, T., Seko, A., Kitajima, K., Muto, Y., Inoue, S., Khoo, K.-H., Morris, H. R., Dell, A., and Inoue, Y. (1994) *J. Biol. Chem.* **269**, 8762–8771
[o] Johnston, M., and 34 other authors. (1994) *Science* **265**, 2077–2082
[p] Williams, N. (1995) *Science* **268**, 1560–1561
[q] Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A., and Peattie, D. A. (1989) *Science* **243**, 75–77
[r] Loomis, W. F., and Insall, R. H. (1999) *Nature (London)* **401**, 440–441
[s] The *C. elegans* Sequencing Consortium (1998) *Science* **282**, 2012–2018 (See this article for list of authors.)
[t] Rubin, G. M., and 54 other authors. (2000) *Science* **287**, 2204–2215
[u] Adams, M. D., and 194 other authors. (2000) *Science* **287**, 2185–2195
[v] Su, X.-z, Ferdig, M. T., Huang, Y., Huynh, C. Q., Liu, A., You, J., Wootton, J. C., and Wellems, T. E. (1999) *Science* **286**, 1351–1353
[w] Postlethwait, J. H., Johnson, S. L., Midson, C. N., Talbot, W. S., Gates, M., Ballinger, E. W., Africa, D., Andrews, R., Carl, T., Eisen, J. S., Horne, S., Kimmel, C. B., Hutchinson, M., Johnson, M., and Rodriguez, A. (1994) *Science* **264**, 699–703
[x] Fishman, M. C. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10554–10556
[y] Kornberg, A., and Baker, T. A. (1992) *DNA Replication*, 2nd ed., pp. 19–21, Freeman, New York
[z] Dietrich, W. F., Copeland, N. G., Gilbert, D. J., Miller, J. C., Jenkins, N. A., and Lander, E. S. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10849–10853
[aa] Anonymous (1994) *Nature (London)* **368**, 167
[bb] Schuler, G. D. and 102 other authors (1996) *Science* **274**, 540–546
[cc] Koonin, S. E. (1998) *Science* **279**, 36–37
[dd] Olson, M. V. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4338–4344
[ee] The Arabidopsis Genome Initiative, (2000) *Nature* **408**, 796–815
[ff] Stevens, J. E. (1994) *Science* **266**, 1186–1187
[gg] Shimamoto, K. (1995) *Science* **270**, 1772–1773
[hh] Singh, K., Ishii, T., Parco, A., Huang, N., Brar, D. S., and Khush, G. S. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6163–6168
[ii] Carels, N., Barakat, A., and Gernardi, G. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 11057–11060
[jj] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994) *Molecular Biology of the Cell*, 3rd ed., Garland, New York

acidic.[40] Small vesicles sometimes bud inward from the plasma membrane in a process called **endocytosis**. In this manner the cell may engulf particles (**phagocytosis**) or droplets of the external medium (**pinocytosis**). The resulting endocytotic vesicles or **endosomes** often fuse with **lysosomes**, which are small acidified vesicles containing a battery of enzymes powerful enough to digest almost anything in the cell. In cells that engulf bits of food (e.g., ameba) lysosomes provide the digestive enzymes. Lysosomes also take up and digest denatured or damaged proteins and may digest "worn out" or excess cell parts including mitochondria. Lysosomes are vital components of cells,[41] and several serious human diseases result from a lack of specific lysosomal enzymes.

## 4. The Endoplasmic Reticulum and Golgi Membranes

Although cytoplasm is fluid and in some organisms can undergo rapid streaming, the electron microscope

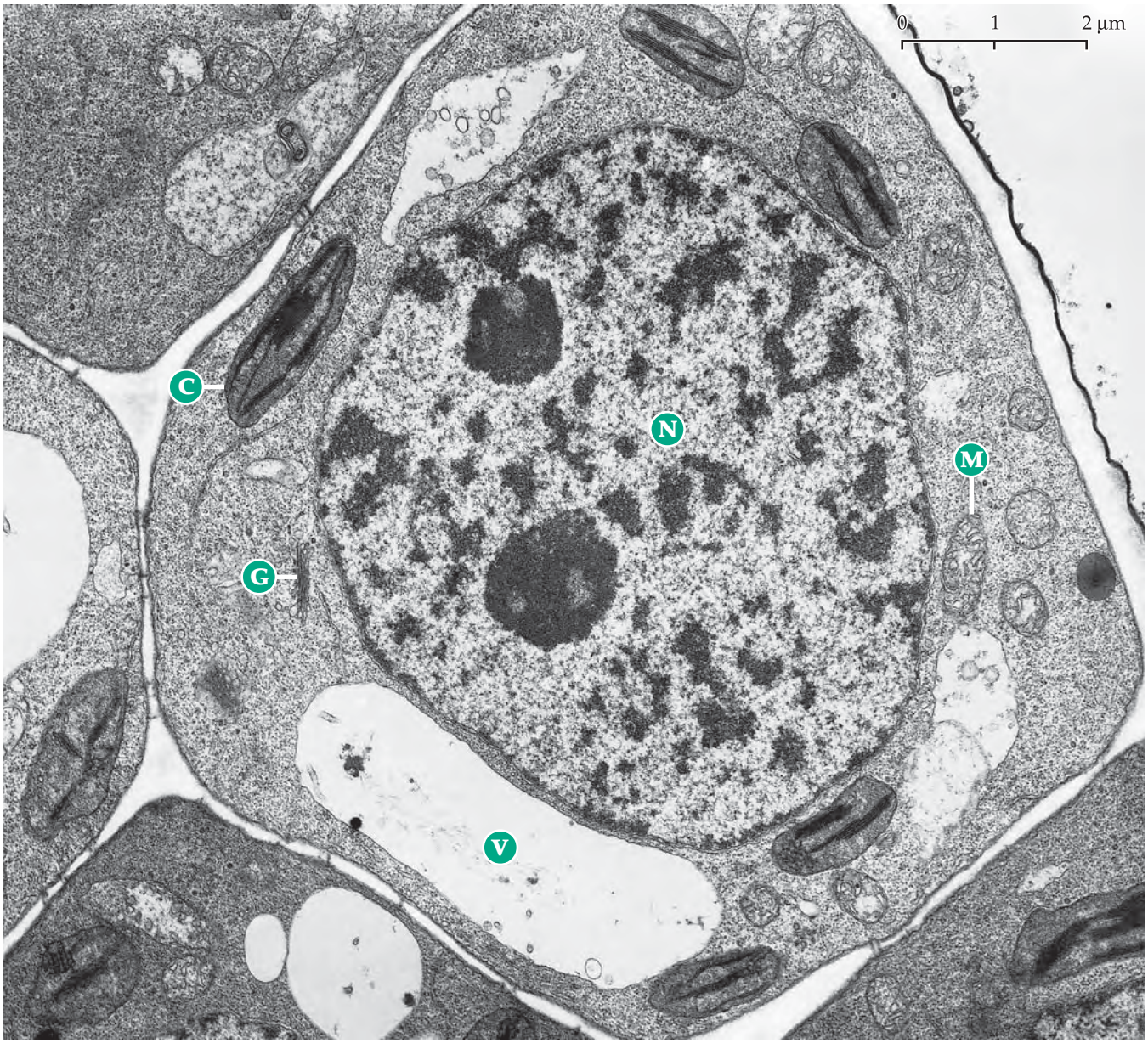
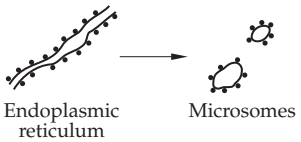



**Figure 1-7** Electron micrograph of a thin section of a young epidermal cell of a sunflower. The tissue was fixed and stained with uranyl acetate and lead citrate. Clearly visible are the nucleus (N), mitochondria (M), chloroplasts (C), a Golgi body dictyosome (G), endoplasmic reticulum, vacuole (V), cell wall, plasmodesmata, and cuticle (upper right, thin dark layer). Micrograph courtesy of H. T. Horner.

has revealed that within the liquid portion, the **cytosol**, there is a complex network of membranes known as the **endoplasmic reticulum** (ER). The membranes of the ER form tubes, vesicles, and flattened sacs called **cisternae**. The intracisternal spaces appear to connect with the perinuclear space and to a series of 3–12 flattened, slightly curved disk-shaped membranes known as the **Golgi apparatus** (Figs. 1-7, 20-8).[42,43] This organelle was first reported by Camillo Golgi in 1898.[44] Its existence was long doubted, but it is known now to play a vital role in metabolism.

The ER, the Golgi membranes, and secretion granules apparently represent an organized system for synthesis of secreted protein and formation of new membranes. Parts of the ER, the **rough endoplasmic reticulum** are lined with many ribosomes of 21–25 nm diameter. While resembling those of bacteria, these eukaryotic ribosomes are about 50% heavier ($4 \times 10^6$ Da). The **smooth endoplasmic reticulum** lacks ribosomes but proteins made in the rough ER may be modified in the smooth ER, e.g., by addition of carbohydrate chains. Small membrane vesicles break off from the smooth ER and pass to the Golgi membranes which lie close to the smooth ER on the side toward the center of the cell. Here additional modification reactions occur (Chapter 20). At the outer edges the membranes of the Golgi apparatus pinch off to form vacuoles which are often densely packed with enzymes or other proteins. These **secretion granules** move to the surface and are released from the cell. In this process of **exocytosis** the membranes surrounding the granules fuse with the outer cell membrane. The rough ER appears to contribute membrane material to the smooth ER and Golgi apparatus, while material from Golgi membranes can become incorporated into the outer cell membrane and into lysosomes. Outer mitochondrial membranes and membranes around vacuoles in plant cells may also be derived directly from the ER. Outer membrane materials are probably "recycled" by endocytosis.

The term **microsome**, frequently met in the biochemical literature, refers to small particles of 50–150 nm diameter which are mostly fragments of the ER together with some material from the plasma membrane. Microsomes are formed when cells are ground or homogenized. Upon centrifugation of the disrupted cells, nuclei and other large fragments sediment first, then the mitochondria. At very high speeds (e.g., at 100,000 times the force of gravity) the microsomes, whose masses are $10^8$–$10^9$ Da, settle. With the electron microscope we see that in the microsomes the membrane fragments have closed to give small sacs to the outside of which the ribosomes still cling:




## 5. Mitochondria, Plastids, and Peroxisomes

**Mitochondria**, complex bodies about the size of bacteria and bounded on the outside by a double membrane (Figs. 1-6 and 1-7), are present in all eukaryotic cells that use oxygen for respiration. The numbers per cell appear to vary from the *one* for certain tiny trypanosomes to as many as $3 \times 10^5$ in some oocytes. Liver cells often contain more than 1000 mitochondria apiece.[45] Study of ultrathin serial sections of a single yeast cell by electron microscopy has shown that under some growth conditions all of the yeast mitochondria are interconnected.[46] More recent evidence from new imaging procedures, e.g. using the green fluorescent protein (Box 23-A)[46a, 46b] also supports the idea that mitochondria are interconnected in a reticulum that can become fragmented under some conditions. The inner membrane of a mitochondrion is often highly folded to form the **cristae** (crests). The outer membrane is porous to small molecules but the passage of substances into and out of the inner space of the mitochondrion, known as the **matrix,** is tightly controlled by the inner membrane. Although some of the oxidative chemical activites of the cells are located in the ER and in peroxisomes, the major energy-yielding reactions for aerobic organisms are found in the mitochondria, which are also the principal site of utilization of oxygen. Within each mitochondrion is a small circular molecule of DNA whose genes encode only a few of the many proteins needed in this organelle. Also present within mitochondria are ribosomes of a size similar to those of bacteria and smaller than those lining the rough ER.

**Plastids** are organelles of plant cells that serve a variety of purposes.[45] Most important are the **chloroplasts**, the chlorophyll-containing sites of photosynthesis. Like mitochondria they contain folded internal membranes (see Fig. 23-19) and several small molecules of DNA.

Fragile organelles, the **peroxisomes** or **microbodies,** occur in many cells.[47–50] In green leaves they may occur in numbers up to one-third those of mitochondria. Peroxisomes are often about the size of mitochondria but have only a single membrane and do not contain DNA. They often contain an apparently crystalline "core." The single membrane of peroxisomes is porous to small molecules such as sucrose. This permits these organelles to be separated from mitochondria by centrifugation in a sucrose gradient where the microbodies assume a density of about 1.25 g/cm$^3$ compared to 1.19 for the impervious mitochondria.

Peroxisomes are rich in enzymes that produce and decompose hydrogen peroxide. They often make a major contribution to the oxidative metabolism of cells. In germinating oilseeds **glyoxysomes**, a type of peroxisome, contain enzymes that catalyze reactions of the biosynthetic "glyoxylate pathway" of metabolism.[51] Organelles that resemble peroxisomes in appearance

but which are functionally more closely related to mitochondria are the **hydrogenosomes** of anaerobic protozoa.[52] As the name suggests, these organelles are the site of formation of molecular hydrogen, a common product of anaerobic metabolism.

## 6. Centrioles, Cilia, Flagella, and Microtubules

Many cells contain **centrioles**,[53] little cylinders about 0.15 µm in diameter and 0.5 µm long, which are *not* enclosed by membranes. Each centriole contains a series of fine **microtubules** of 25 nm diameter. A pair of centrioles are present near the nucleus in most animal cells and play an important role in cell division. Together with surrounding materials they form the **centrosome**. However, centrioles have never been observed in plant cells.

Related in structure to centrioles are the long **flagella** and shorter **cilia** (the two words are virtually synonymous) which are commonly present as organelles of locomotion in eukaryotic cells. Stationary cells of our own bodies also often have cilia. For example, there are $10^9$ cilia/cm$^2$ in bronchial epithelium.[54] Modified flagella form the receptors of light in our eyes and of taste in our tongues. Flagella and cilia have a diameter of about 0.2 µm and a characteristic internal structure. Eleven hollow microtubules of ~24 nm diameter are usually arranged in a "9 + 2" pattern with nine pairs of fused tubules surrounding a pair of single tubules (Figs. 1-8 and 19-23). Each microtubule resembles a bacterial flagellum in appearance, but there are distinct and significant chemical differences. The **basal body** of the flagellum, the **kinetosome** (Fig. 1-8), resembles a centriole in structure, dimensions, and mode of replication. Recently a small 6–9 megabase pair DNA has been found in basal bodies of the protozoan *Chlamydomonas*.[55,56]

Microtubules similar to those found in flagella are also present in the cytoplasm. Together with thinner **microfilaments** of several kinds they form an internal **cytoskeleton** that provides rigidity to cells.[58,59] Microtubules also form the "spindle" of dividing cells. In nerve axons (Chapter 30) the microtubules run parallel to the length of the axons and are part of a mechanical transport system for cell constituents.

## 7. Cell Coats, Walls, and Shells

Like bacteria, most cells of higher plants and animals are surrounded by extracellular materials. Plants have rigid walls rich in cellulose and other carbohydrate polymers. Outside surfaces of plant cells are covered with a **cuticle** containing layers of a polyester called **cutin** and of wax (Fig. 1-6). Surfaces of animal cells are usually lined with carbohydrate molecules which are attached to specific surface proteins to form **glycoproteins**. Spaces between cells are filled with such "cementing substances" as **pectins** in plants and **hyaluronic acid** in animals. Insoluble proteins such as **collagen** and **elastin** surround connective tissue cells. Cells that lie on a surface (epithelial and endothelial cells) are often lined on one side with a thin, collagen-containing **basement membrane** (Figs. 1-6 and 8-31). Inorganic deposits such as calcium phosphate (in bone), calcium carbonate (eggshells and spicules of sponges), and silicon dioxide (shells of diatoms) are laid down, often by cooperative action of several or many cells.

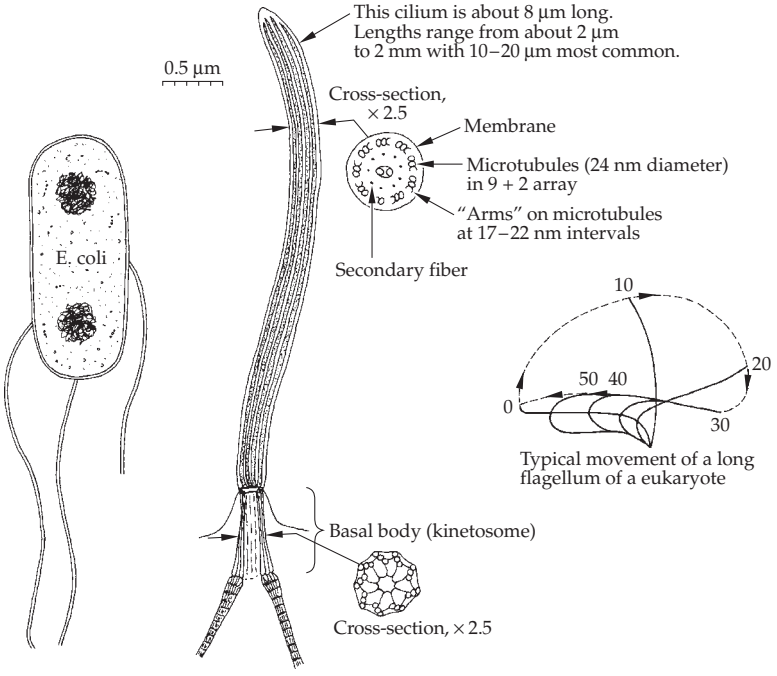



**Figure 1-8** Structure of cilia and flagella of eukaryotes. After P. Satir.[57]

## C. Inheritance, Metabolic Variation, and Evolution of Eukaryotes

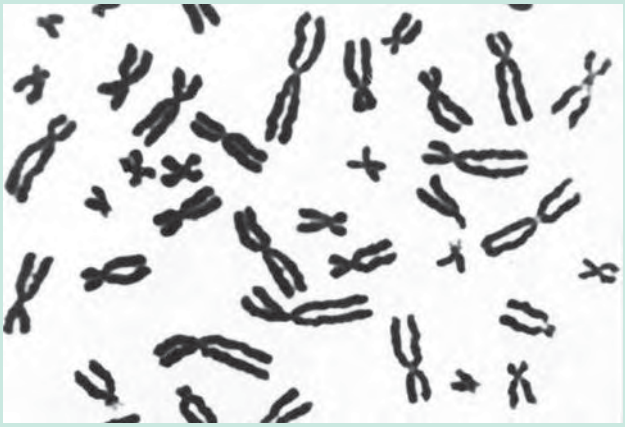
The striking differences between eukaryotic and prokaryotic cells have led to many speculations about the evolutionary relationship of these two great classes of living organisms. A popular theory is that mitochondria, which are characteristic of most eukaryotes, arose from aerobic bacteria. After cyanobacteria had developed and oxygen

## BOX 1-D INHERITED METABOLIC DISEASES

In 1908 Archibald Garrod[a,b] proposed that **cystinuria** (Chapter 8) and several other defects in amino acid and sugar metabolism were "inborn errors of metabolism", i.e. inherited diseases. Since that time the number of recognized genetic defects of human metabolism has increased at an accelerating rate to ~4000.[c–e] Hundreds of other genetic problems have also been identified. For over 800 of these the defective gene has been mapped to a specific chromosome.[f] An example is **sickle cell anemia** (Box 7-B) in which a defective hemoglobin differs from the normal protein at one position in one of its constituent polypeptide chains. Many other defects involve loss of activity of some important enzyme.

Most genetic diseases are rare, affecting about one person in 10,000. However, **cystic fibrosis** affects one in 2500. There are so many metabolic diseases that over 0.5% of all persons born may develop one. Many die at an early age. A much greater number (>5%) develop such conditions as diabetes and mental illness which are, in part, of genetic origin. Since new mutations are always arising, genetic diseases present a problem of continuing significance.

At what rate do new mutations appear? From the haploid DNA content (Table 1-2) we can estimate that the total coding capacity of the DNA in a human cell exceeds two million genes (actually two million *pairs* of genes in diploid cells). However, only a fraction of the DNA codes for proteins. There are perhaps 50,000 pairs of structural genes in human DNA. The *easily detectable* rate of mutation in bacteria is about $10^{-6}$ per gene, or $10^{-9}$ per base per replication.[g] As a result of sophisticated "proofreading" and repair systems, it may be as low as $10^{-10}$ per base in humans.[h] Thus, in the replication of the $3 \times 10^9$ base pairs in diploid human chromosomes we might anticipate about one mistake per cell division. Only about 1/50 of these would be in structural genes and potentially harmful. Thus, if there are $10^{16}$ division cycles in a normal life span[h] each parent may pass on to future generations about 2 mutations in protein sequences. The ~$10^{14}$ body cells (somatic cells) also undergo mutations which may lead to cancer and to other problems of aging. Most mutations may be harmless or nearly so and a few may be beneficial.



*Photomicrograph of human male metaphase chromosomes. © Photo Researchers*

However, many are damaging and some are *lethal*. If a mutation is lethal, a homozygote will not survive and will be lost in an early (and usually undetected) spontaneous abortion. Healthy individuals carry as many as ten lethal recessive mutations as well as at least 3–5 autosomal recessive mutations of a seriously harmful type. Harmful dominant mutations are also frequent in the population. These include an elevated lipoprotein content of the blood and an elevated cholesterol level which are linked to early heart disease.

Biochemical disorders are also important because of the light they shed on metabolic processes. No other species is observed as carefully as *Homo sapiens*. As a consequence frequent reference will be made to genetic diseases throughout the book. A goal is to find ways to prevent or ameliorate the effects of these disorders. For example, in the treatment of **phenylketonuria** (Chapter 25) or of **galactosemia** (Chapter 20), a change in the diet can prevent irreversible damage to the brain, the organ most frequently affected by many of these diseases. Injection of a missing enzyme is giving life to victims of Gaucher's disease (Chapter 20). In many other cases no satisfactory therapy is presently available, but the possibilities of finding some way to supply missing enzymes or to carry out "genetic surgery" are among the most exciting developments of contemporary medical biochemistry (Chapter 26).

---

[a] Garrod, A. E. (1909) *Inborn Errors of Metabolism*, Oxford, London
[b] Bearn, A. G. (1993) *Archibald Garrod and the Individuality of Man*, Oxford, New York
[c] Scriver, C. R., Beaudet, A. L., Sly, W. S., and Valle, D., eds. (1995) *The Metabolic and Molecular Bases of Inherited Disease*, 7th ed., McGraw-Hill, New York
[d] Davies, K. E. (1992) *Molecular Basis of Inherited Disease*, Oxford, New York
[e] McKusick, V. A. (1994) *Mendelian Inheritance in Man*, 11th ed., Johns Hopkins Univ. Press, Baltimore, Maryland
[f] McKusick, V. A., Amberger, J., and Steinberg, J. (1994) *J. NIH Res.* **6**, 115–134
[g] Watson, J. D. (1976) *Molecular Biology of the Gene*, 3rd ed., Benjamin, Menlo Park, California (p. 254)
[h] Koshland, D. E., Jr. (1994) *Science* **266**, 1925

had become abundant, a **symbiotic** relationship could have arisen in which small aerobic bacteria lived within cells of larger bacteria that had previously been obligate anaerobes. Sequence similarities of proteins suggest that these symbionts may have been related to present-day methanogens[60] and thermophilic sulfur bacteria.[61] The aerobes presumably used up any oxygen present, protecting the surrounding anaerobic organisms from its toxicity. The relationship became permanent and led eventually to the mitochondria-containing eukaryotic cell.[62–65] Further symbiosis with cyanobacteria or prochlorophytes could have led to the chloroplasts of the eukaryotic plants.

A fact that supports such ideas is the existence among present-day organisms of many endosymbiotic relationships. For example, the green paramecium (*Paramecium bursaria*) contains, within its cytoplasm, an alga (*Chlorella*), a common green plant that is quite capable of living on its own. Perhaps by accident it took up residence within the paramecium.[62] Some dinoflagellates (Fig. 1-9) contain endosymbiotic cyanobacteria[66] and recently a ciliate that contains endosymbiotic purple photosynthetic bacteria has been discovered.[67] These bacteria do not produce $O_2$ but utilize products of the host ciliates' metabolism such as acetate, lactate, and $H_2$ as electron donors for photosyntheses. They also utilize $O_2$ for respiration and may protect their hosts from the toxicity of $O_2$, just as may have happened in the distant past. According to this theory the symbionts would eventually have lost their photosynthetic ability and have become mitochondria. The relationship of mitochondria to bacteria is also supported by many biochemical similarities.

Fossils of bacteria and blue-green algae have been obtained from rocks whose age, as determined by geochemical dating, is more than $3 \times 10^9$ years.[68,69] However the first eukaryotic cells may have appeared about $1 \times 10^9$ years ago[70] and started to evolve into the more than one million species that now exist.[1,71,72]

## 1. A Changing Genome

How is it possible for the genome of an organism to increase in size as it evolved from a lower form to a higher one? Simple mutations that cause alterations in protein sequences could lead to changes in form and behavior of the organisms but could not, by themselves, account for the increase in genetic material that accompanied evolution. As a result of new techniques of genetic mapping and determining the sequence of nucleotides in DNA we are rapidly acquiring a detailed knowledge of the organization of the genome. It has been found that genes are often present as duplicate but not entirely identical copies. This suggests that there are mechanisms by which cells can acquire extra copies of one or more genes. Indeed it seems probable that at some time in the past the entire genome of bacteria was doubled and that it was later doubled again.[73] Evidence for this is that the masses of bacterial chromosomes group around values of 0.5, 1.4, and $2.7 \times 10^9$ Da. Genes can also be duplicated during the process of genetic recombination, which is discussed in Chapter 27. In addition, the size of the genome may have increased by incorporation of genetic material from extrachromosomal plasmids.

A possible advantage to a cell possessing an extra copy of a gene is that the cell would survive even when mutations rendered unusable the protein encoded by one of the copies. As long as one of the genes remained "good," the organism could grow and reproduce. The extra, mutated gene could be carried for many generations. As long as it produced only harmless, nonfunctioning proteins there might be little selection pressure to eliminate it and it might undergo repeated mutations. After many mutations and many generations later, the protein for which it coded could prove useful to the cell in some new way.

An example of evolution via gene duplication is provided by the oxygen-carrying proteins of blood. It appears that about a billion years ago, the gene for an ancestral **globin**, the protein of hemoglobin, was doubled. One gene evolved into that of present-day **globins** and the other into the gene of the muscle protein **myoglobin**. Still later, the globin gene again doubled leading to the present-day α and β chains of hemoglobin (Chapter 7). These are two distinctly different but related protein subunits whose genes are not even on the same chromosome. To complicate the picture further, most human beings have two or more copies of their α chain gene[74] as well as genes for fetal and embryonic forms of hemoglobin. However, some populations have lost one or more α chain genes. Thus, the genome changes in many details, even today.

## 2. Genetic Recombination, Sex, and Chromosomes

Bacteria usually reproduce by simple fission. The single DNA molecule of the chromosome is duplicated and the bacterium divides, each daughter cell receiving an identical chromosome. However, genetic recombination, which is accomplished in several ways by bacteria (Chapter 27), provides a deliberate process for mixing of genes. This process has been most fully developed in eukaryotic organisms that undergo sexual reproduction. The growth of a multicelled individual begins with the fusion of two haploid **gametes**, an egg and a spermatozoon. Each gamete carries a complete set of genetic instructions, and after the nuclei fuse the fertilized egg or **zygote** is **diploid**. Each diploid cell contains *two* complete sets of genetic blueprints of quite different origin. Even if a gene from one parent

is defective, the chances are that the gene from the other parent will be good. Sexual reproduction and the associated genetic recombination also provide a means for mixing of genes.

When eukaryotic cells prepare to divide in the process called **mitosis** (Fig. 26-11), the DNA molecules of the nucleus, which become spread out through a large volume, coil and fold. Together with proteins and other molecules they form the compact bodies known as chromosomes. Some organisms, such as *Ascaris* (a roundworm), have only two chromosomes, a **homologous pair**, one inheritied from the father and one from the mother. Both chromosomes divide in every mitotic cell division so that every cell of the organism has the homologous pair. Higher organisms usually have a larger number of chromosomes. Thus, humans have 23 homologous pairs. The mouse has 20, the toad 11, onions 8, mosquitos 3, and *Drosophila* 4. Human chromosomes vary in size but are usually 4–6 μm long and ~1 μm in diameter.

By the successive divisions of mitosis, a single fertilized eukaryotic egg cell can grow to an adult. Less than 50 successive mitotic divisions will produce the ~$10^{14}$ cells of a human. However, formation of gametes, which are haploid, requires the special process of **meiosis** (Fig. 26-12), by which the number of chromosomes is divided in half. During meiosis one chromosome of each of the homologous pairs of the diploid cell is passed to each of the gametes that are formed. In an organism such as *Ascaris*, which contains only a single pair of chromosomes, a gamete receives either the chromosome of maternal origin or that of paternal origin but not both. In organisms that have several pairs of chromosomes, one chromosome of each pair is passed to the gamete in a random fashion during meiosis. Most gametes receive some chromosomes of maternal and some of paternal origin. An important feature of meiosis is the genetic recombination that occurs during **crossing-over**. In this process, the strands of DNA are cut and genetic material is exchanged between the chromosomes of maternal and paternal origin. Thus, crossing-over breaks the **linkage** between genes and provides for greater variability in the offspring than would otherwise be possible. Each of us receives half of our genes from our mother and half from our father, but some of these genes have been inherited from each grandparent on both sides of the family, some from each great-grandparent, etc.

Many genes are passed down through many generations without substantial change, but others are evidently designed to be scrambled readily within somatic cells. Cell surface proteins[75] and antibody molecules are among the proteins whose genes undergo alteration during growth and differentiation of the tissues of the body (Chapter 32).

### 3. Haploid and Diploid Phases

In human beings and other higher animals, meiosis leads directly to formation of the gametes, the egg and sperm cells. These fuse to form a diploid nucleus and the adult develops by repeated mitosis of the diploid cells. While meiosis also occurs in the life cycle of all eukaryotic creatures, it is not always at a point corresponding to that in the human life cycle. Thus, the cells of many protozoa and of fungi are ordinarily haploid. When two haploid nuclei fuse to form a diploid cell, meiosis quickly occurs to produce haploid individuals again. Among lower plants and animals there is often an alternation of haploid and diploid phases of the life cycle. For example, gametes of ferns fall to the ground and germinate to form a low-growing green mosslike haploid or **gametophyte** form. The latter produces motile haploid gametes which fuse to a diploid zygote that grows into the larger and more obvious **sporophyte** form of the fern.

It is presumably the ability to survive as a heterozygote, even with one or more highly deleterious mutations, that has led to the dominance of the diploid phase in higher plants and animals.[76] However, to the biochemical geneticist organisms with a haploid phase offer experimental advantages because recessive mutants can be detected readily.

## D. Survey of the Protists

Unicellular eukaryotes have traditionally been grouped together with multicellular organisms in which all cells have similar functions, with little or no differentiation into tissues, as the kingdom **Protista**.[77,78] The fungi may also be included or may be regarded as a separate kingdom.[79] With present-day emphasis on DNA sequence comparisons the traditional classification is changing, however.[26]

### 1. Protozoa

Among the best known of the animal-like protista is the **ameba** (subphylum Sarcodina or Rhizopoda). The most striking feature of the ameba (Fig. 1-9) is its method of locomotion, which involves the transformation of cytoplasm from a liquid state to a semisolid gel. As the ameba moves, the cytoplasm at the rear liquifies and flows to the front and into the extending pseudopodia where it solidifies along the edges. The ameba poses several important biochemical questions: What chemistry underlies the reversible change from liquid to solid cytoplasm? How can the cell membranes break and reform so quickly when an ameba engulfs food particles?[80]

Relatives of the ameba include the **Radiolaria**,

marine organisms of remarkable symmetry with complex internal skeletons containing the carbohydrate polymer chitin together with silica ($SiO_2$) or strontium sulfate. The **Foraminifera** deposit external shells of calcium carbonate or silicon dioxide. Over 20,000 species are known and now as in the distant past their minute shells fall to the bottom of the ocean and form limestone deposits.

Tiny ameboid parasites of the subphylum **Sporozoa** attack members of all other animal phyla. Several genera of **Coccidia** parasitize rabbits and poultry causing enormous damage. Humans are often the victims of species of the genus *Plasmodium* (Fig. 1-9) which invade red blood cells and other tissues to cause **malaria**, one of our most serious ailments on a worldwide basis.[81–84] Throughout history malaria has probably killed more persons than any other disease. *Toxoplasma gondii* is another parasite which, in its haploid phase, is found throughout the world in wild animals and in humans. Although its presence usually elicits no symptoms, it sometimes causes blindness and mental retardation in children and can be fatal to persons with AIDS. Its sexual cycle occurs exclusively in cats.[85,86]

Another subphylum of protozoa, the Mastigophora, are propelled by a small number of flagella and are intermediate between animals and the algae. One of these is *Euglena viridis*, a small freshwater organism with a long flagellum in front, a flexible tapered body, green chloroplasts, and a light-sensitive "eyespot" which it apparently uses to keep itself in the sunshine (Fig. 1-9). Euglena is also able to live as a typical animal if there is no light. Treatment with streptomycin (Box 20-B) causes *Euglena* to lose its chloroplasts and to become an animal permanently. The **dinoflagellates** (Fig. 1-9), some colorless and some green, occur in great numbers among the plankton of the sea. *Giardia lamblia* is a troublesome intestinal parasite.

The **hemoflagellates** are responsible for some of our most terrible diseases. Trypanosomes (genus *Trypanosoma*) invade the cells of the nervous system causing African sleeping sickness. Mutating their surface proteins frequently by gene-scrambling mechanisms, these and other parasites are able to evade the immune response of the host.[87,88] For the same reason it is difficult to prepare vaccines against them. Other flagellates live in a symbiotic relationship within the alimentary canals of termites (Fig. 1-9) and roaches. Termites depend upon bacteria that live within the cells of these symbiotic protozoans to provide the essential enzymes needed to digest the cellulose in wood.

Members of the subphylum Ciliophora, structurally the most complex of the protozoa, are covered with a large number of cilia which beat together in an organized pattern.[89] The following question immediately comes to mind: How

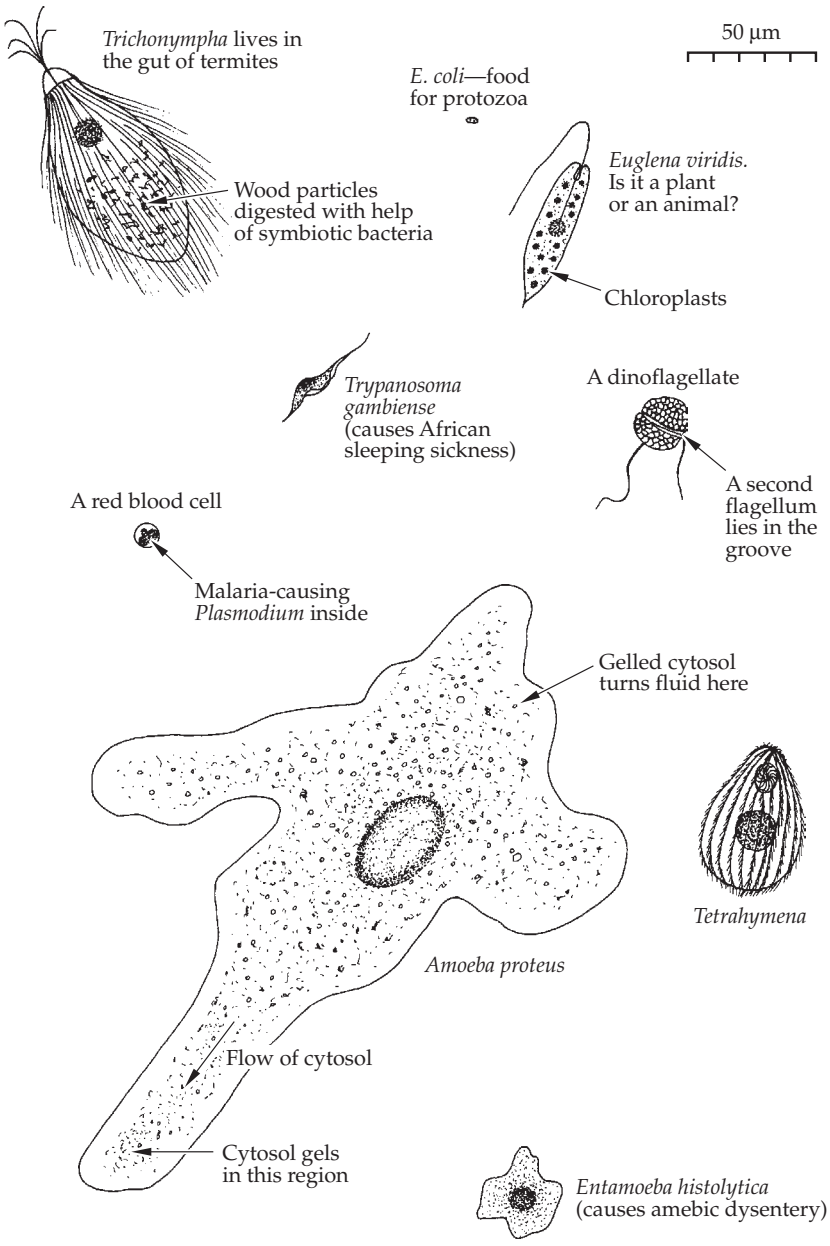



**Figure 1-9** A few well-known protists.

are the cilia able to communicate with each other to provide this organized pattern? Two ciliates that are often studied by biochemists are *Tetrahymena* (Fig. 1-9), one of the simplest, and *Paramecium*, one of the more complex.

The **Myxomycetes** or "slime molds" are more closely related to protozoa than to fungi.[90] Members of the family Acrasieae, the best studied member of which is *Dictyostelium discoideum*, start life as small amebas. After a time, when the food supply runs low, some of the amebas begin to secrete pulses of a chemical attractant **cyclic AMP**. Neighboring amebas respond to the pulses of cyclic AMP by emitting their own pulses about 15 s later, then moving toward the original source.[91,92] The ultimate effect is to cause the amebas to stream to centers where they aggregate and form fungus-like fruiting bodies. Asexual spores are formed and the life cycle begins again. Other Myxomycetes grow as a multinucleate (diploid) **plasmodium** containing millions of nuclei but no individual cell membranes. *Physarum polycephalum*, a species whose plasmodium may spread to a diameter of 30 cm, has become popular with biochemists. The 800,000 nuclei per square millimeter all divide synchronously.

## 2. Fungi

Lacking photosynthetic ability, living most often in soil but sometimes in water, the fungi are represented by almost half as many species (~$10^5$) as are the vascular plants.[93] The distinguishing characteristics of fungi are the lack of chlorophyll and growth as a series of many branched tubules (usually 6–8 μm diameter), the **hyphae**, which constitute the **mycelium**. The hyphae are not made up of separate cells but contain a mass of protoplasm with many nuclei. Only occasional septa divide the tubules. Most fungi are saprophytic, living on decaying plants or animal tissues. However, others are parasites that produce serious and difficult-to-treat infections in humans. An important medical problem is the lack of adequate antibiotics for treating fungal infections (mycoses).[94–96] On the other hand, fungi produce important antibiotics such as **penicillin**. Still others form some of the most powerful toxins known!

The lower fungi or **Phycomycetes** include simple aquatic molds and mildew organisms. Higher fungi are classified as **Ascomycetes** or **Basidiomycetes** according to the manner in which the sexual spores are born. In the Ascomycetes these spores are produced in a small sac called an **ascus** (Fig. 1-10). Each ascus contains four or eight spores in a row, a set of four representing the results of a single pair of meiotic divisions. A subsequent mitotic division will give eight spores. This is one of the features that has made *Neurospora crassa* (Fig. 1-10) a favorite subject for genetic studies.[97] The ascospores can be dissected out in order from the ascus and cultivated separately to observe the results of crossing-over during meiosis.

*Neurospora* also reproduces via haploid spores called **conidia**. The haploid mycelia exist as two mating types and conidia or mycelia from one type can fertilize cells in a special body (the protoperithecium) of the other type to form zygotes. The latter immediately undergo meiosis and mitosis to form the eight ascospores. Among other Ascomycetes are the highly prized edible truffles and morels. However, most mushrooms and puffballs are fruiting bodies of Basidiomycetes. Other Basidiomycetes include the **rusts**, which cause enormous damage to wheat and other grain crops.

Yeasts are fungi adapted to life in an environment of high sugar content and which usually remain unicellular and reproduce by budding (Fig. 1-10). Occasionally the haploid cells fuse in pairs to form diploid cells and sexual spores. Some yeasts are related to the Ascomycetes, others to Basidiomycetes. *Saccharomyces cerevisiae*, the organism of both baker's and brewer's yeast, is an Ascomycete. It can grow indefinitely in either the haploid or diploid phase. The genetics and biochemistry of this yeast have been studied extensively.[98–102] The genome is relatively small with $13.5 \times 10^6$ base pairs in 17 chromosomes. The sequence of the 315,000 base pairs of chromosome III was determined in 1992[101,102] and the sequence of the entire genome is now known.[103]

Fungi often grow in symbiotic association with other organisms. Of special importance are the **mycorrhizae** (fungus roots) formed by colonization of fine roots by beneficial soil fungi. Almost all plants of economic importance form mycorrihizae.[104]

## 3. Algae

Algae are chlorophyll-containing eukaryotic organisms which may be either unicellular or colonial.[105] The colonial forms are usually organized as long filaments, either straight or branched, but in some cases as blades resembling leaves. However, there is little differentiation among cells. The gold-brown, brown, and red algae contain special pigments in addition to the chlorophylls.

The euglenids (**Euglenophyta**) and dinoflagellates (**Pyrrophyta**), discussed in the protozoa section, can equally well be regarded as algae. The bright green **Chlorophyta**, unicellular or filamentous algae, are definitely plants, however. Of biochemical interest is *Chlamydomonas*, a rather animal-like creature with two flagella and a carotenoid-containing eyespot or **stigma** (Fig. 1-11). *Chlamydomonas* contains a single chloroplast. The "pyrenoid", a center for the synthesis of starch, lies, along with the eyespot, within the chloroplast. The organism is haploid with "plus" and "minus" strains

and motile gametes. Zygotes immediately undergo meiosis to form haploid spores. With a well-established genetic map, *Chlamydomonas* is another important organism for studies of biochemical genetics.[106]

The filamentous *Ulothrix* shows its relationship to the animals through formation of asexual spores with four flagella and biflagellate gametes. Only the zygote is diploid. On the other hand, the incomparably beautiful *Spirogyra* (Fig. 1-11) has no motile cells. The ameboid male gamete flows through a tube formed between the two mating cells, a behavior suggesting a relationship to higher green plants.

Some unicellular algae grow to a remarkable size. One of these is *Acetabularia* (Fig. 1-11), which lives in the warm waters of the Mediterranean and other tropical seas. The cell contains a single nucleus which lies in the base or rhizoid portion. In the mature alga, whose life cycle in the laboratory is 6 months, a cap of characteristic form develops. When cap development is complete, the nucleus divides into about $10^4$ secondary nuclei which migrate up the stalk and out into the rays of the cap where they form cysts. After the cap decays and the cysts are released, meiosis occurs and the flagellated gametes fuse in pairs to form zygotes which again grow into diploid algae. Because of its large size and the location of the nucleus in the base, the cells can be cut and grafted. Nuclei can be removed or transplanted and growth and development can be studied in the presence or absence of a nucleus.[107–110] The green algae **Volvox** live in wheel-like colonies of up to several thousand cells and are useful for biochemical studies of differentiation.[111]

Look through the microscope at almost any sample of algae from a pond or aquarium and you will see little boatlike **diatoms** slowly gliding through the water. The most prominent members of the division Chrysophyta, diatoms are characterized by their external "shells" of silicon dioxide. Large and ancient deposits of diatomaceous earth contain these durable silica skeletons which are finely marked, often with beautiful patterns (Fig. 1-11). The slow motion of diatoms is accomplished by streaming of protoplasm through a groove on the surface of the cell. Diatoms are an important part of marine plankton, and it is estimated that three-fourths of the organic material of the world is produced by diatoms and dinoflagellates. Like the brown algae, Chrysophyta contain the pigment **fucoxanthin**.

Other groups of algae are the brown and red marine algae or seaweed. The former (**Phaeophyta**) include the giant kelps from which the polysaccharide **algin** is obtained. The **Rhodophyta** are delicately branched plants containing the red pigment **phycoerythrin**. The polysaccharides, **agar** and **carrageenin**,

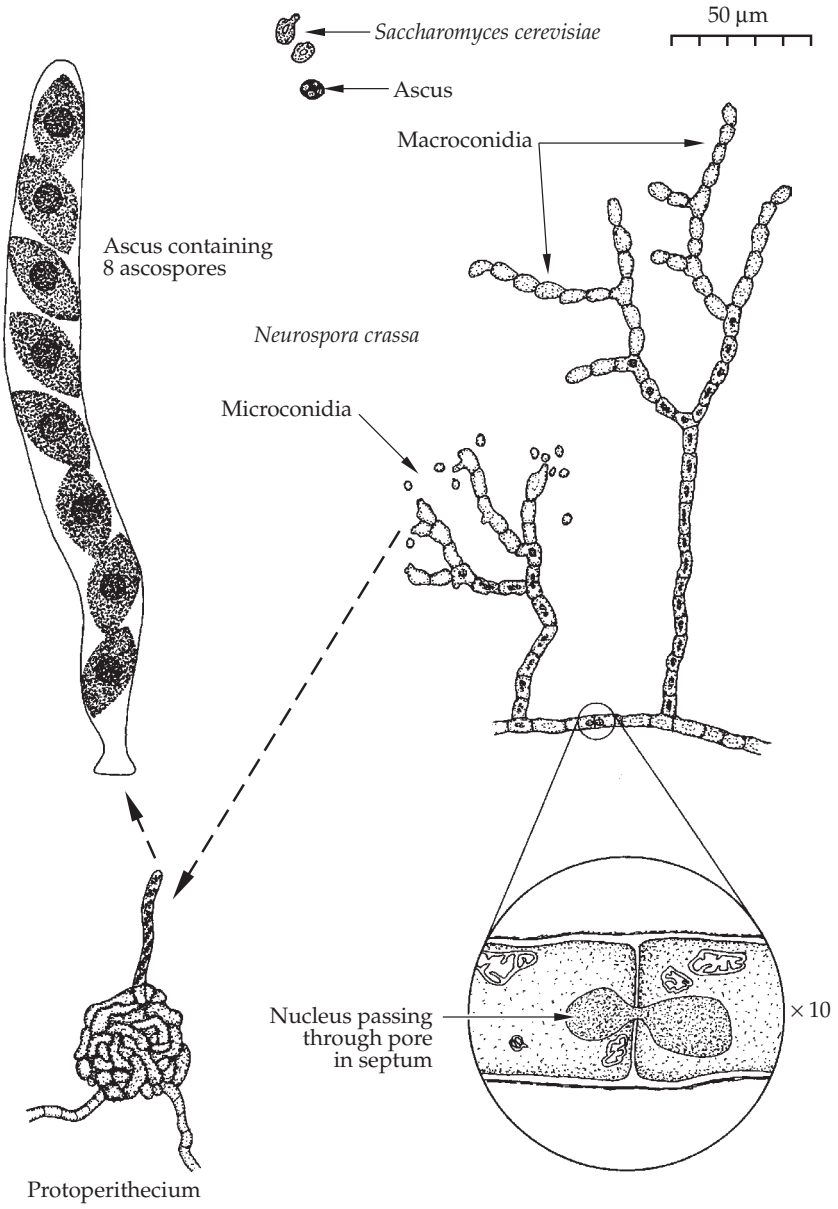



**Figure 1-10** Two frequently studied fungi. Top (including ascus): the yeast *Saccharomyces cerevisiae*. Below: *Neurospora crassa* showing various stages. After J. Webster.[93]

a popular additive to chocolate drinks and other foods, come from red algae.

Symbiotic associations of fungi with either true algae or with cyanobacteria are known as **lichens**. Over 15,000 varieties of lichens grow on rocks and in other dry and often cold places. While the algae appear to benefit little from the association, the fungi penetrate the algae cells and derive nutrients from them.[112] Although either of the two partners in a lichen can be cultured separately, the combination of the two is capable of producing special pigments and phenolic substances known as **depsides** which are not formed by either partner alone.

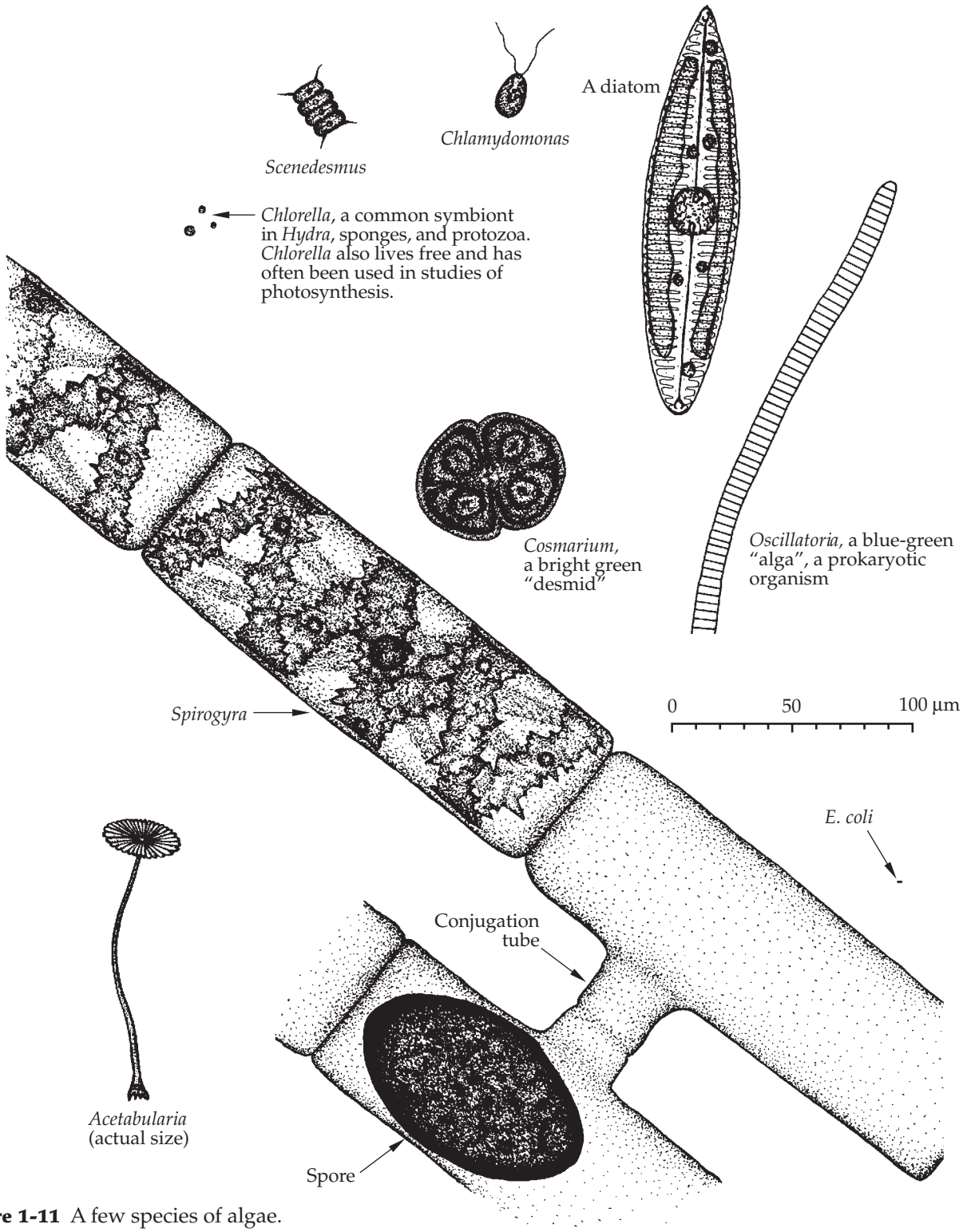



**Figure 1-11** A few species of algae.

## E. The Variety of Animal Forms

In this section, we will consider only a few biochemical and other aspects of multicellular animals or **Metazoa**. The sudden appearance of a large number of Metazoans about 0.5 x $10^9$ years ago[113,114] may have been an outcome of the appearance of split genes (see Section B, 1). As a result of gene duplication the coding pieces of split genes, the **exons**, could be moved to new locations in a chromosome where they could have become fused with other pieces of DNA to form entirely new genes.[115]

### 1. Major Groups of Multicellular Animals

The simplest metazoa are tiny symbiotic worms of the phylum (or subkingdom) **Mesozoa**, which live in the kidneys of deep sea-dwelling cephalopods (octopi and squid). Each worm is made up of only 25 cells in a single layer enclosing one or a small number of elongated axial cells (Fig. 1-12). Mesozoa have been regarded as parasitic, but they appear to facilitate excretion of $NH_3$ by the host through acidification of the urine.[118,119]

**Porifera** or sponges are the most primitive of multicelled animals.[120] They lack distinct tissues but contain several specialized types of cells. The body is formed by stationary cells that pump water through the pores to bring food to the sponge. Within the body **amebocytes** work in groups to form the **spicules** of calcium carbonate, silicon dioxide, or the protein **spongin** (Fig. 1-12). Sponges appear to lack a nervous system.

Individuals of the next most complex major phylum, **Cnidaria** (formerly Coelenterata), are radially symmetric with two distinct cell layers, the **endoderm** and **ectoderm**. Many species exist both as a polyp or **hydra** form (Fig. 1-13) and as a **medusa** or jellyfish. The jellyfish apparently has no brain but the ways in which its neurons interconnect in a primitive radial net are of interest. The Cnidaria have a very simple body form with remarkable regenerative powers. The freshwater hydra, a creature about 1 cm long (Fig. 1-13), contains a total of ~$10^5$ cells. A complete hydra can be regenerated from a small piece of tissue if the latter contains some of both the inner and the outer cell layers.[121,122]

The body of flatworms (phylum **Platyhelminthes**) consists of two external cell layers (endoderm and ectoderm) with a third layer between. A distinct excretory system is present. In addition to a nerve net resembling that of the Cnidaria, there are a cerebral ganglion and distinct eyes. One large group of flatworms, the **planarians** (typically about 15 mm in length, Fig. 1-14), inhabit freshwater streams. They are said to be the simplest creatures in which *behavior* can be studied.

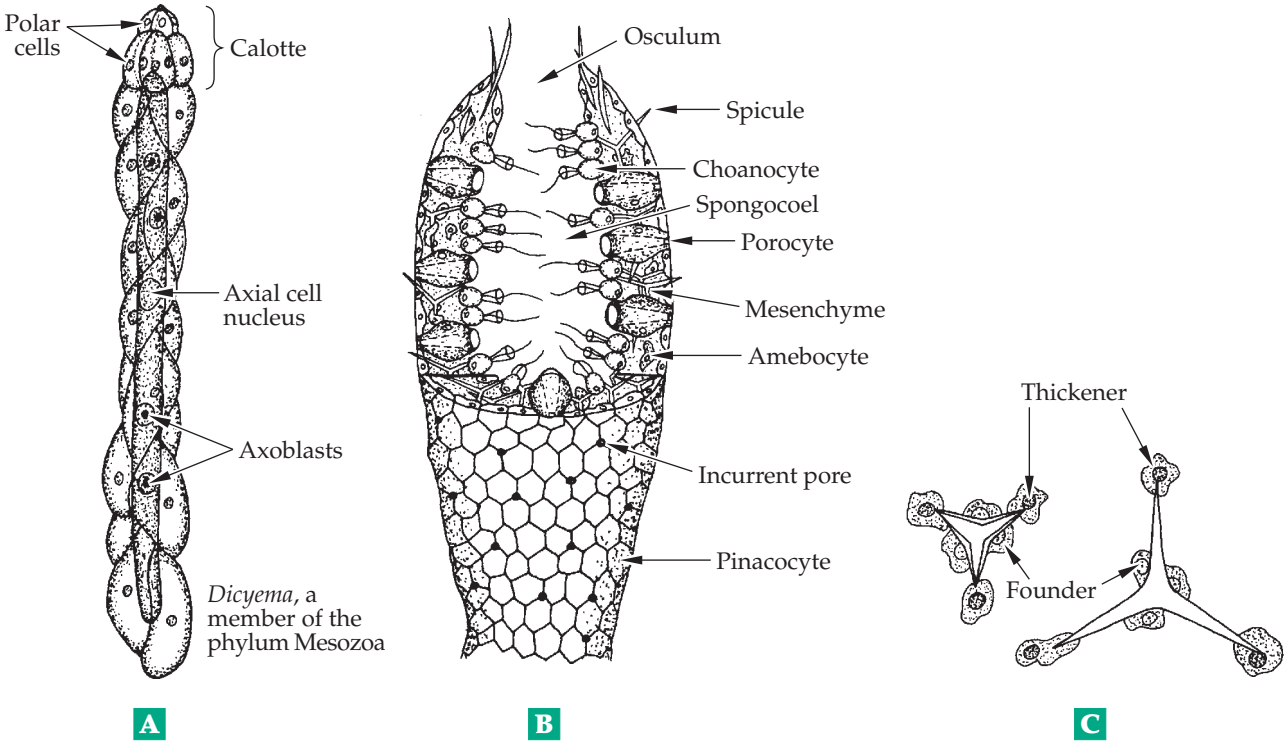



**Figure 1-12** Some lower forms of Metazoa. (A) Mesozoa (25 cells). After C. P. Hickman.[116] (B) A small asconoid sponge. After C. A. Villee, W. F. Walker, Jr., and R. D. Barnes.[117] (C) Ameboid cells of a sponge forming spicules. After Hickman.

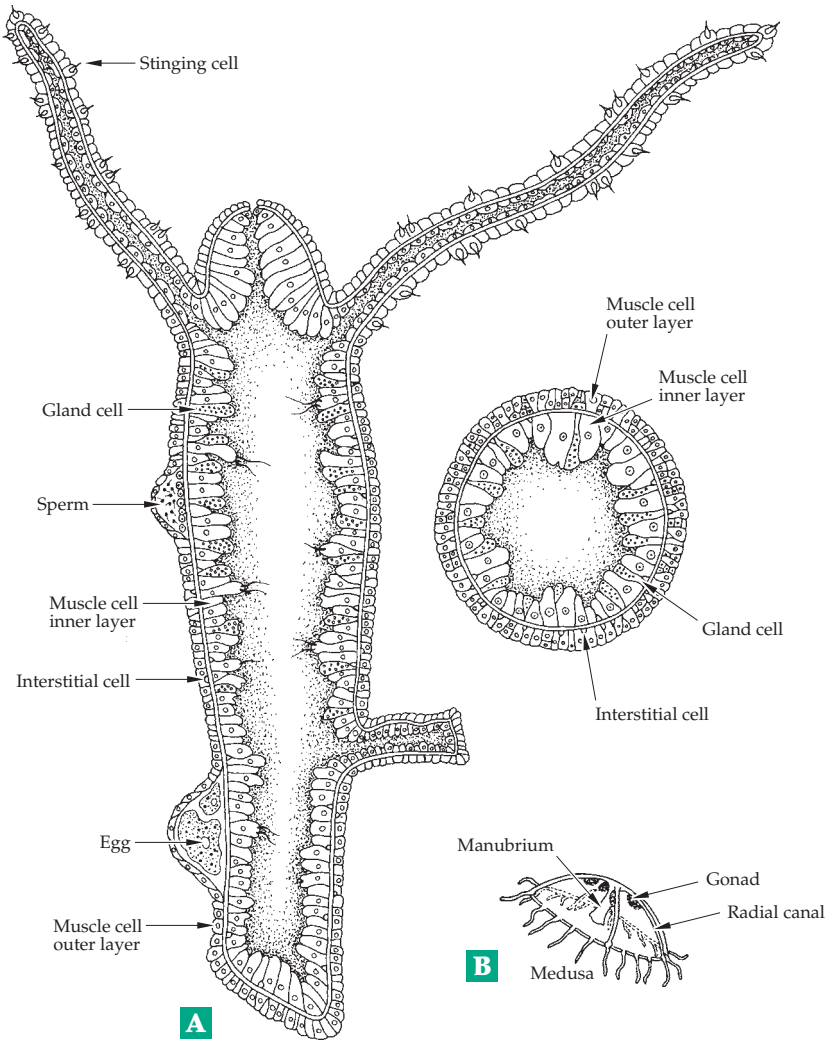



**Figure 1-13** (A) Hydra. After Loomis.[123] (B) The medusa stage of *Obelia*, a hydroid coelenterate.[123a]

Many parasitic flatworms (tapeworms and flukes) attack higher organisms.[124] Among them are the **Schistosoma**, tiny worms that are transmitted to humans through snails and which attack the blood vessels. The resulting **schistosomiasis** is one of the most widespread debilitating diseases on earth today, affecting 200 million people or more.[125,126]

The roundworms (**Nematoda**)[127–129] have, in addition to the **enteron** (alimentary tract), a separate body cavity. Free-living nematodes abound in water and soil but many species are parasitic. They do enormous damage to plants and to some animal species. Trichina, hookworms[129a], and filaria worms attack humans. However, in the laboratory the 1-mm-long, 810-cell nematode, *Caenorhabditis elegans* (Fig. 1-14) has become an important animal. In 1963 Sydney Brenner launched what has become a worldwide effort to make this tiny worm the equivalent in the animal kingdom of *E. coli* in the bacterial world.[129] The $10^8$ nucleotides in the worm's six chromosomes contain ~13,600 genes. *C. elegans* has become an important animal in which to study differentiation. Already the exact lineage of every cell has been traced, as has every connection among the 302 neurons in the animal's nervous system. The related **rotifers**,[130] with whirling "wheels" of cilia on their heads (Fig. 1-14) and transparent bodies, are a delight to the microscopist. Like nematodes, they are "cell constant" organisms. The total number of cells in the body is constant as is that in almost every part of every organ. Part of the developmental plan of such organisms is a "programmed cell death" (Chapter 32).

The **Annelida** (segmented worms)[131] are believed to be evolutionary antecedents of the arthropods. Present-day members include earthworms, leeches, and ~$10^5$ species of marine **polychaetes**. Annelids have a true body cavity separate from the alimentary canal and lined by a peritoneum. They have a well-developed circulatory system and their blood usually contains a type of hemoglobin.

About $10^6$ species of **arthropods** (80% of all known animals) have been described. Most are very small.[72] These creatures, which have a segmented exoskeleton of **chitin** and other materials, include the horseshoe crabs, the Arachnida (scorpions, spiders, and mites), the Crustacea, Myriopoda (centipedes and millipedes), and the Insecta. Important biochemical problems are associated with the development and use of insecticides and with our understanding of the metamorphosis that occurs during the growth of arthropods.[132] The fruit fly *Drosophila melanogaster* has provided much of our basic knowledge of genetics and continues to be the major species in which development is studied.[133–134a]

Among the molluscs (phylum **Mollusca**) the squids and octopuses have generated the most interest among biochemists. The neurons of squid contain giant axons, the study of which has led to much of our knowledge of nerve conduction. Octopuses show signs of intelligence not observed in other invertebrates whose nervous reactions seem to be entirely "preprogrammed." The brains of some snails contain only $10^4$ neurons, some of which are unusually large. The

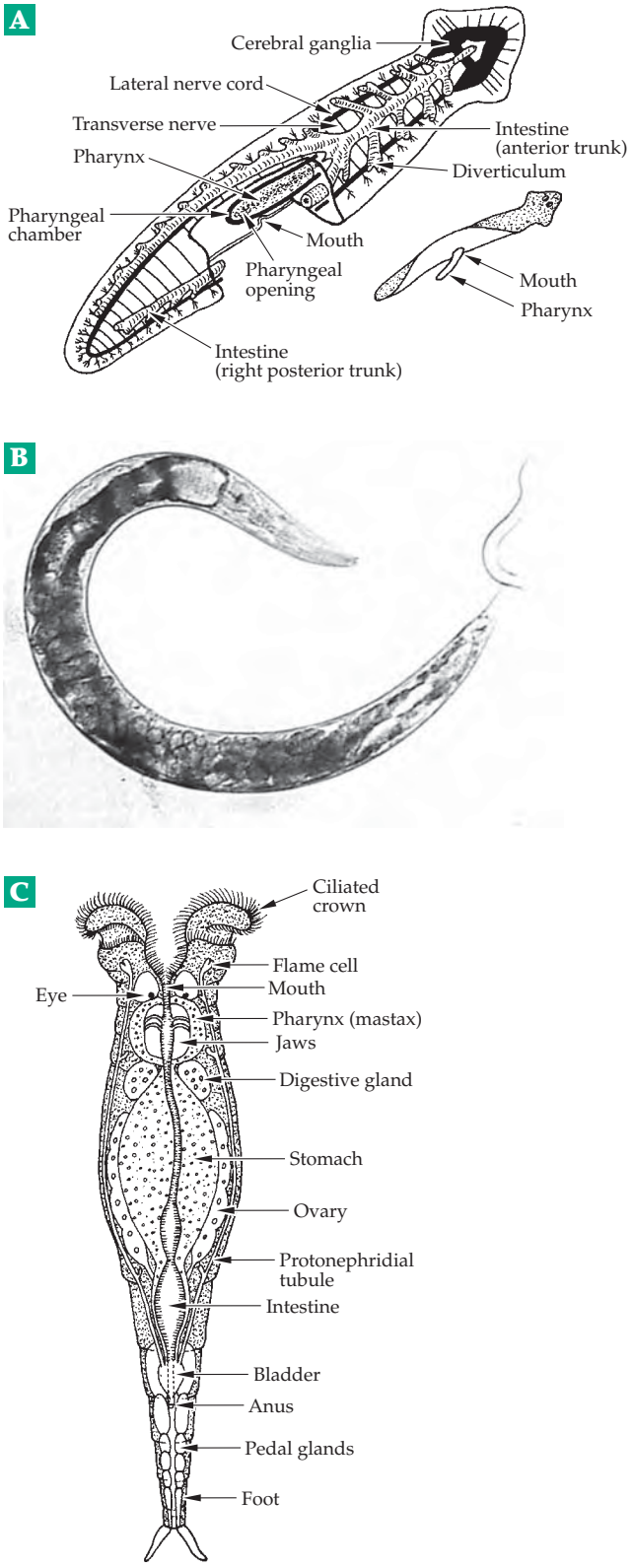



**Figure 1-14** (A) A planarian, length 15 mm. After Hickman.[116] Diagram of digestive and nervous sytems; cutaway section shows ventral mouth. Small drawing shows pharynx extended through ventral mouth. (B) The nematode *Caenorhabditis elegans. Ascaris* is very similar in appearance. From Buchsbaum.[77] (C) A rotifer, *Philodina* (~$10^3$ cells). After C. A. Villee *et al.*[117]

**Echinodermata** or spiny-skinned animals (starfish, sea urchins, and sea cucumbers) are regarded as a highly advanced phylum. Their embryological development has been studied intensively.

The phylum **Chordata**, to which we ourselves belong, includes not only the vertebrates but also more primitive marine animals that have a spinal cord. Among these primitive species, which may be related to early ancestral forms, are the **tunicates** or sea squirts. They have a very high concentration of vanadium in their blood.

## 2. Cell Types and Tissues

Isolated animal cells in tissue culture, no matter how highly differentiated, tend to revert quickly to one of three basic types known as **epitheliocytes**, **mechanocytes**, and **amebocytes**. Epitheliocytes are closely adherent cells derived from epithelial tissues and thought to be related in their origins to the two surface layers of the embryonic blastula. Mechanocytes, often called **fibroblasts** or **fibrocytes**, are derived from muscle, supporting, or connective tissue. Like the amebocytes, they arise from embryonic mesenchymal tissue cells that have migrated inward from the lower side of the blastula (Chapter 32). **Neurons**, **neuroglia**, and **lymphocytes** are additional distinct cell types.

### BOX 1-E ERRORS, MISCONCEPTIONS, AND SPECULATION

Warning: Not everything in this book is true. Despite all efforts to get it right, there are unintentional errors and misinterpretations of experimental results. Indeed, the history of biochemistry is replete with accounts of experimental findings that were interpreted incorrectly. Yet, the ideas expressed often stimulated others to develop a more correct picture later. The same is true today. Students should be critical, should look at experimental details, and *consult original literature as much as possible.*

Progress in science depends both upon careful observations and measurements and upon imaginative interpretations of unexpected findings. Speculative ideas, a number of which are mentioned in this book, provide an important stimulus in science. They should neither be ignored nor accepted as facts. I have tried to write in such a way that established facts will not often be confused with speculation.

***Tissues.*** Cells aggregate to form four major kinds of tissue. Epithelial tissues line the primary surfaces of the body: the skin, the digestive tract, urogenital tract, and glands. External skin is composed of flat platelike squamous epithelial cells whereas internal surfaces are often formed by columnar epithelial cells. Glands (sweat, oil, mammary, and internal secretory) as well as the sensory organs of the tongue, nose, and ear are all composed of epithelial cells. Epithelial cells are among the most highly polarized of cells. One side of each cell faces the outside, either air or water, while the other side is often directly against a basement membrane.

Supporting and connective tissues include the fatty **adipose tissue** as well as **cartilage** and **bone**. Both of the latter contain large amounts of intercellular material or **ground substance** consisting largely of complex polymers. Embryonic fibroblasts differentiate into white fibers, which produce collagen, and yellow fibers, which form elastin. The fibrils of both of these proteins are assembled in the intercellular space where they are embedded in the ground substance. **Osteoblasts** form bone by deposition of calcium phosphate in 3–7 μm thick layers within a ground substance that contains special proteins.

A third tissue is **muscle**, which is classified into three types: **striated** (voluntary skeletal muscle), **cardiac** (involuntary striated muscle), and **smooth** (involuntary) muscle. There are two major groups of cells in **nervous tissue**, the fourth tissue type. **Neurons** are the actual conducting cells whose cell membranes carry nerve impulses. Several kinds of **glial cells** lie between and around the neurons.

***Blood cells.*** Blood and the linings of blood vessels may be regarded as a fifth tissue type.[135,135a] The human body contains $5 \times 10^9$ **erythrocytes** or red blood cells per ml, a total of $2.5 \times 10^{13}$ cells in the five liters of blood present in the body. Erythrocytes are rapidly synthesized in the bone marrow. The nucleus is destroyed, leaving a cell almost completely filled with hemoglobin. With an average lifetime of 125 days, human red blood cells are destroyed by leukocytes in the spleen and liver.

The white blood cells or **leukocytes** are nearly a thousandfold less numerous than red cells. About $7 \times 10^6$ cells are present per ml of blood. There are three types of leukocytes: **lymphocytes** (~26% of the total), **monocytes** (~7% of the total), and **polymorphonuclear leukocytes** or **granulocytes** (~70% of the total). Lymphocytes are about the same size as erythrocytes and are made in lymphatic tissue. Individual lymphocytes may survive for as long as ten years. They function in antibody formation and are responsible for maintenance of long-term immunity.

Monocytes, two times larger, are active in ingesting bacteria. These cells stay in the blood only a short time before they migrate into the tissues where they become **macrophages**,[136] relatively fixed phagocytic cells. Macrophages not only phagocytize and kill invading bacteria, protozoa, and fungi but also destroy cancer cells. They also destroy damaged cells and cellular debris as part of the normal turnover of tissues. They play an essential role in the immune system by "processing" antigens and in releasing stimulatory proteins.

Granulocytes of diameter 9–12 μm are formed in the red bone marrow. Three types are distinguished by staining: **neutrophils**, **eosinophils**, and **basophils**. Neutrophils are the most numerous phagocytic cells of our blood and provide the first line of defense against bacterial infections. The functions of eosinophils and basophils are less well understood. The number of eosinophils rises during attacks of hay fever and asthma and under the influence of some parasites, while the basophil count is increased greatly in leukemia and also by inflammatory diseases. Granules containing histamine, heparin, and leukotrienes are present in the basophils. Blood **platelets** or **thrombocytes** are tiny (2–3 μm diameter) cell-like bodies essential for rapid coagulation of blood. They are formed by fragmentation of the cytoplasm of bone marrow **megakaryocytes**. One mature megakaryocyte may contribute 3000 platelets to the $1–3 \times 10^8$ per ml present in whole blood.

***Cell culture.*** Laboratory growth of isolated animal cells has become very important in biochemistry.[137] Sometimes it is necessary to have many cells with as nearly as possible identical genetic makeup. Such bacterial cells are obtained by plating out the bacteria and selecting a small colony that has grown from a single cell to propagate a "pure strain." Similarly, single eukaryotic cells may be selected for tissue culture and give rise to a **clone** of cells which remains genetically identical until altered by mutations.

The culture of embryonic fibroblasts is used to obtain enough cells to perform prenatal diagnosis of inherited metabolic diseases (Box 1-D). Tissue culture is easiest with embryonic or cancer cells, but many other tissues can be propagated. However, the cells that grow best and which can be propogated indefinitely are not entirely normal; the well-known **HeLa** strain of human cancer cells which was widely grown for many years throughout the world contains 70–80 chromosomes per cell compared with the normal 46.

## 3. Communication

Plants are able to maintain their form because the cells are surrounded by thick walls that cement the cells together. However, animal cells lack rigid walls and must be held together by specialized contacts.[138,139] Contacts between cells of both plants and animals are

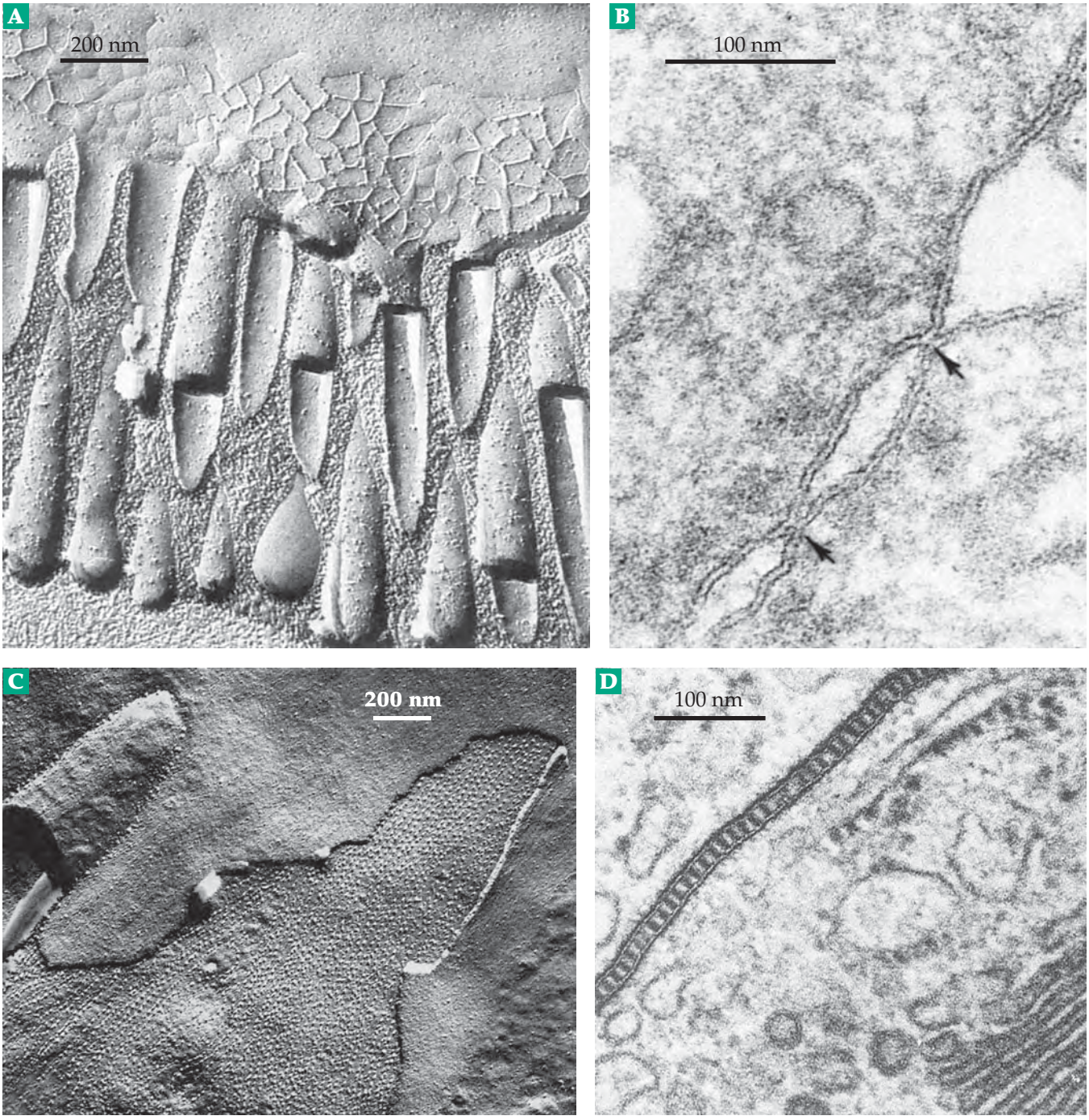
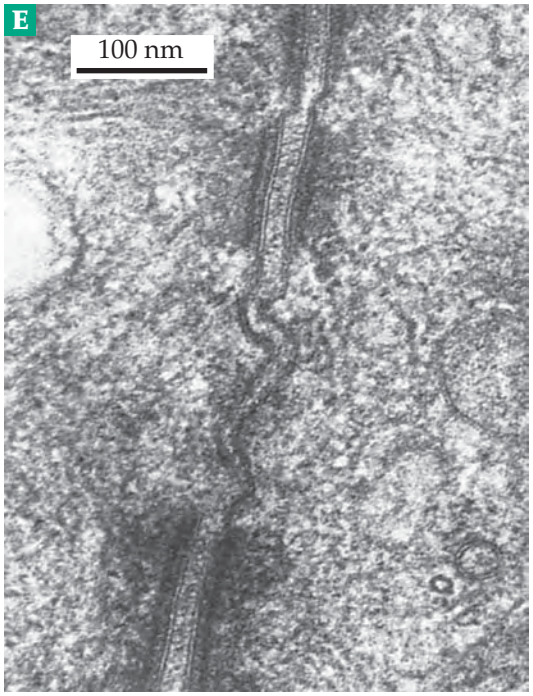
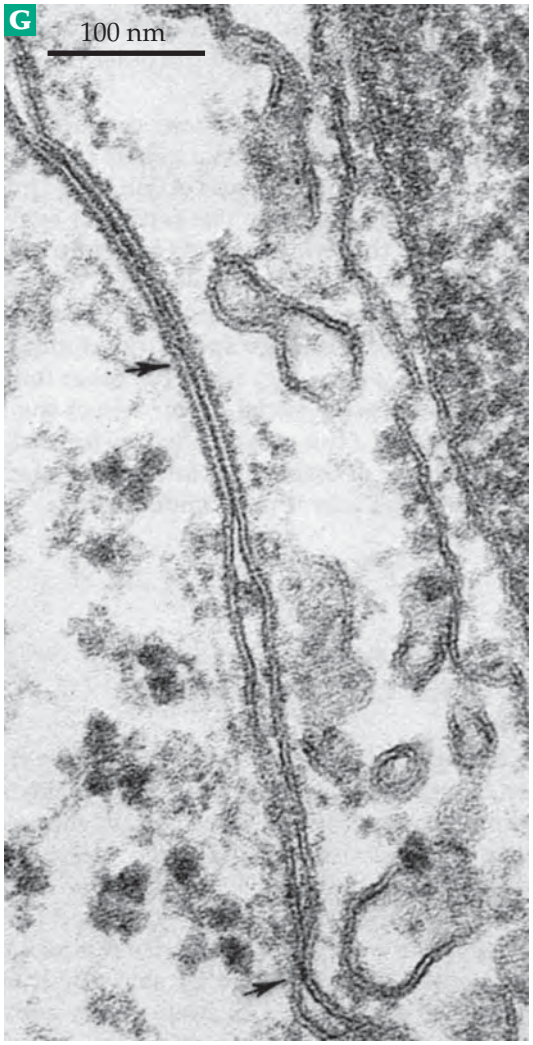



**Figure 1-15** Electron micrographs of cell junctions of three types. (A) Freeze-fractured zona occludens (occlusion zone) between epithelial cells of the rat small intestine. The tight junctions are repesented as a meshwork of ridges (in the P or protoplasmic fracture face) or grooves (in the E fracture face which looks toward the extracellular space). These represent the actual sites of membrane fusion. Microvilli are seen in the lower part of the photograph. From D. S. Friend and N. B. Gilula.[141] (B) Thin cross section of tight junction between mouse hepatocytes. The arrows indicate points of membrane fusion. From Gilula.[142] Copyright 1975 by The Williams & Wilkins Co., Baltimore. (C) A freeze-fractured septate junction from ciliated epithelium of a mollusc. This type of junction forms a belt around the cells. Fracture face P (central depressed area) contains parallel rows of membrane particles that correspond to the arrangmenet of the intracellular septa seen in thin sections. The surrounding fracture face E contains a complementary set of grooves. Particles in nonjunctional membrane regions (upper right corner) are randomly arranged. (D) Thin section of a septate junction of the type shown in (C). The plasma membranes of the two cells are joined by a periodic arrangement of electron-dense bars or septa, which are present within the intercellular space. Note the Golgi membranes in the lower right part of the photograph. (C) and (D) are from N. B. Gilula[143]



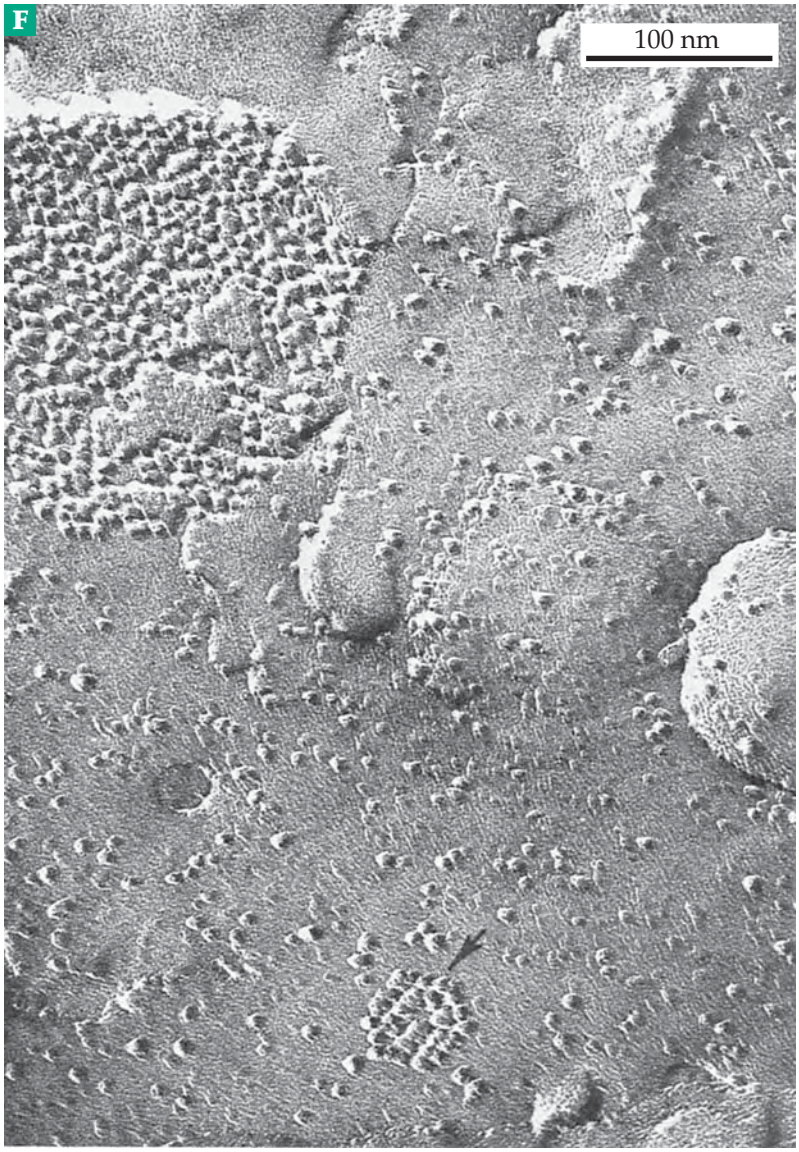




**Figure 1-15** (continued) (E) Desmosomes (macula adherens) in rat intestinal epithelium. Features include a wide (25–35 nm) intercellular space containing dense material, two parallel cell membranes, a dense plaque associated with the cytoplasmic surface and cytoplasmic tonofilaments that converge on the dense plaque. From Gilula.[143] (F) Freeze-fractured surface through gap junctions between communicating cells in culture. Both a large junction and a smaller one below (arrow) can be seen. (G) Gap junctions in thin section. (F) and (G) are from N. B. Gilula.[144]

important for a second reason: Cells must communicate, one with another.

***Cell contacts and junctions.*** Many epithelial cells, e.g., those lining the border of kidney tubules and secretory glands, form **tight junctions** with adjacent cells. Electron microscopy shows that in these junctions the outer portions of the membranes actually fuse in some places (Fig. 1-15). One way in which this has been

demonstrated is to freeze a tissue rapidly and to fracture it in the frozen state within a vacuum chamber. The fractured tissue is kept at about –100°C in a vacuum for a short time while water molecules evaporate from the fractured surfaces. A thin plastic replica is then made of the etched surfaces, which sometimes pass through tight junctions revealing details of their structure (Fig. 1-15, A, E). Study of electron micrographs of such surfaces shows that some cells are completely surrounded by belts of tight junctions, sometimes referred to as **occlusion zones** or **terminal bars**. Tight junctions between endothelial cells of blood capillaries in the brain prevent free diffusion of compounds from the blood stream into brain cells and form the **blood–brain barrier**.[140] Tight junctions between neurons and adjacent cells surround the **nodes of Ranvier** (Chapter 30).

Contacts of another type, known as **septate desmosomes** or **adhesion discs**, form a belt around the cells of invertebrate epithelia. In these contacts a space of ~18 nm between adjacent cell membranes is bridged in a number of places by thin walls. Behind the desmosomes the membrane is often backed up at these points by an electron-dense region to which are attached many fine microfilaments of ~6–10 nm diameter (Fig. 1-15, D).

One method of communication between cells is by passage of chemical substances through special junctions which, because of their appearance in electron micrographs of thin sections (Fig. 1-15, G) are known as **gap junctions**.[139,145,146] Gap junctions may cover substantial areas of the cell interface. In cross section, a thin 3–4 nm gap between the adjacent cell membranes is bridged by a lattice-like structure, which may appear in freeze-fractured surfaces as a hexagonal array of particles (Fig. 1-15, F, lower junction). These particles or **connexons** are each thought to be composed of six protein subunits. A central channel in the connexon is able to pass molecules of molecular mass up to about 500 Da.[147,148] Small molecules may be able to pass freely from one cell to another through the gap junction. Because of their low electrical resistance, gap junctions allow "electrical coupling" of cells. Such junctions form the **electrotonic synapses** that link some neurons to other excitable cells. Heart cells are all electrically coupled through gap junctions.[149]

Another type of communicating junction is also found in **synapses** of the nervous system. At these specialized contacts a nerve impulse transmitted along the membrane of one neuron triggers the release of a **neurotransmitter**, a chemical substance that passes across the gap between cells of the synapse and initiates a nerve impulse in the second neuron (Chapter 30).

***Cell recognition.*** Cells of higher organisms are able to recognize other cells as identical, as belonging to another tissue, or as being "foreign." This ability is developed most highly by cells of the immune system but is possessed to some extent by others. For example, cells of sponges can be separated by partial digestion of the protein "cement" that holds them together. When dissociated cells from orange sponges were mixed with those from yellow sponges, the cells clumped together to reform small sponges.[150,151] Furthermore, orange cells stuck to orange cells and yellow to yellow cells. Similar results have been obtained using a mixture of cultivated liver, kidney, and embryonic brain cells. When a wound heals, epithelial cells grow and move across the wound surface but they stop when they meet. Cells in tissue culture and growing on a glass surface experience this same **contact inhibition**[152] and spread to form a unicellular layer. Cancer cells in culture do not stop but climb one on top of the other, apparently lacking proper recognition and communication. Many chemical signals appear to pass between cells. An important goal of contemporary biochemistry is to understand how cells recognize each other and respond to signals that they receive.

## F. Higher Plants and Plant Tissues

Botanists recognize two divisions of higher plants. The **Bryophyta** or moss plants consist of the Musci (mosses) and Hepaticae (liverworts). These plants grow predominantly on land and are characterized by swimming sperm cells and a dominant gametophyte (haploid) phase. **Tracheophyta**, or vascular plants, contain conducting tissues. About $2 \times 10^5$ species are known. The ferns (class Filicineae, formerly Pteridophyta) are characterized by a dominant diploid plant and alternation with a haploid phase. Seed plants are represented by two classes: **Gymnosperms** (cone-bearing trees) and **Angiosperms**, the true flowering plants.

Genetically the simplest of the angiosperms is the little weed *Arabidopsis thaliana*, whose generation time is as short as five weeks. Its five chromosomes contain only $10^8$ base pairs in all, the smallest known genome among angiosperms[153] and one whose complete nucleotide sequence is being determined. Its biochemistry, physiology, and developmental biology are under intensive study. It may become the "fruit fly" of the plant kingdom.

There are several kinds of plant tissues. Undifferentiated, embryonic cells found in rapidly growing regions of shoots and roots form the **meristematic tissue**. By differentiation, the latter yields the simple tissues, the parenchyma, collenchyma, and sclerenchyma. **Parenchyma** cells are among the most abundant and least specialized in plants. They give rise through further differentiation to the **cambium layer**, the growing layer of roots and stems. They also

make up the pith or pulp in the center of stems and roots, where they serve as food storage cells.

The **collenchyma**, present in herbs, is composed of elongated supporting cells and the **sclerenchyma** of woody plants is made up of supporting cells with hard lignified cell walls and a low water content. This tissue includes **fiber cells**, which may be extremely long; e.g., pine stems contain fiber cells of 40 µm diameter and 4 mm long.

Two complex tissues, the **xylem** and **phloem**, provide the conducting network or "circulatory system" of plants. In the xylem or woody tissue, most of the cells are dead and the thick-walled tubes (**tracheids**) serve to transport water and dissolved minerals from the roots to the stems and leaves. The phloem cells provide the principal means of downward conduction of foods from the leaves. Phloem cells are joined end to end by **sieve plates**, so-called because they are perforated by numerous minute pores through which cytoplasm of adjoining sieve cells appears to be connected by strands 5–9 µm in diameter.[154] Mature sieve cells have no nuclei, but each sieve cell is paired with a nucleated "companion" cell.

**Epidermal tissue** of plants consists of flat cells, usually containing no chloroplasts, with a thick outer wall covered by a heavy waxy **cuticle** about 2 µm thick. Only a few specialized cells are found in the epidermis. Among them are the paired **guard cells** that surround the small openings known as **stomata** on the undersurfaces of leaves and control transpiration of water. Specialized cells in the root epidermis form **root hairs**, long extensions (~1 mm) of diameter 5–17 µm. Each hair is a single cell with the nucleus located near the tip.

Figure 1-16 shows a section from a stem of a typical angiosperm. Note the thin cambium layer between the phloem and the xylem. Its cells continuously undergo differentiation to form new layers of xylem increasing the woody part of the stem. New phloem cells are also formed, and as the stem expands all of the tissues external to the cambium are renewed and the older cells are converted into bark.

Plant **seeds** consist of three distinct portions. The **embryo** develops from a zygote formed by fusion of a sperm nucleus originating from the pollen and an egg cell. The fertilized egg is surrounded in the gymnosperms by a nutritive layer or **endosperm** which is **haploid** and is derived from the same gametophyte tissue that produced the egg. In angiosperms *two* sperm nuclei form; one of these fertilizes the egg, while the other fuses with *two* haploid **polar nuclei** derived from the female gametophyte. (The polar nuclei are formed by the same mitotic divisions that formed the egg.) From this develops a *3n* **triploid** endosperm.

## G. The Chemical Composition of Cells

**Water** is the major component of living cells, but the amount varies greatly. Thus, the pig embryo is 97% water; at birth a new-born pig is only 89% water. A lean 45-kg pig may contain 67% water but a very fat 135-kg animal only 40% water. Similar variations are encountered with other constitutents.

The water content of a tissue is often determined by thoroughly drying a weighed sample of tissue at low temperature in vacuum and then weighing it a second time. The solid material can then be extracted with a solvent that will dissolve out the fatty compounds. These are referred to collectively as **lipids**. After evaporation of the solvent the lipid residue may be weighed. By this procedure a young leafy vegetable might be found to contain 2–5% lipid on a dry weight

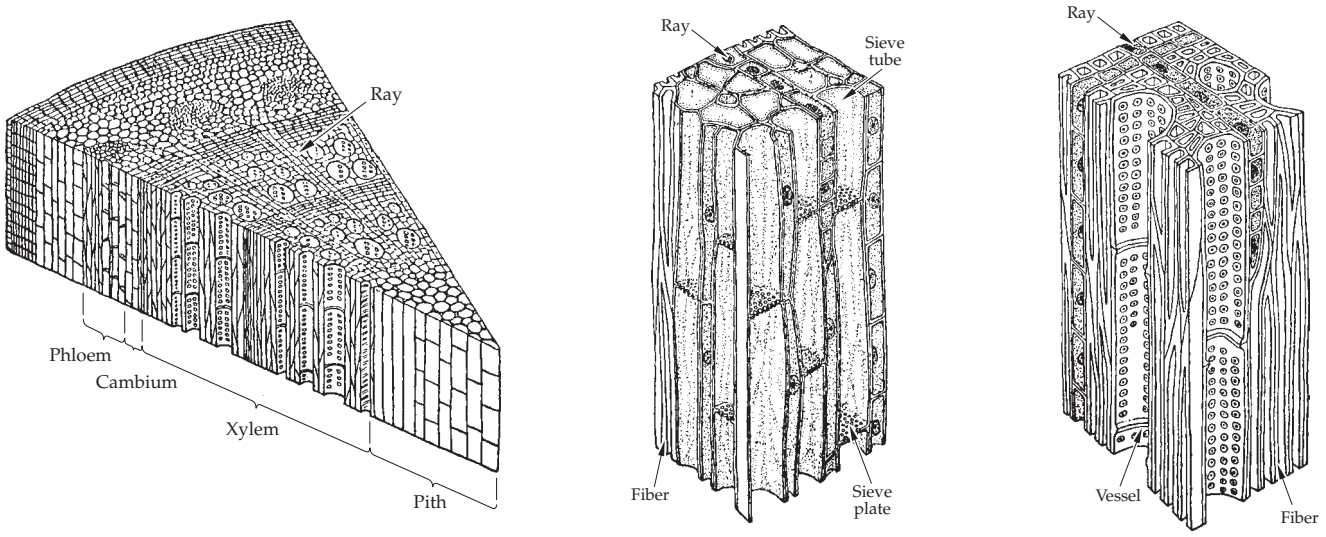



**Figure 1-16** Section of the stem of an angiosperm. Enlarged sections showing tubes of the phloem (left) and xylem (right). From S. Biddulph and O. Biddulph.[155] Drawn by Bunji Tagawa.

basis. Even very lean meats contain 10–30% lipid.

The residue remaining after removal of the lipid consists predominately of three groups of compounds: **proteins**, **nucleic acids**, and **carbohydrates**. Most of the nitrogen present in tissues is found in the proteins and the protein content is sometimes estimated by determining the percentage of nitrogen and multiplying by 6.25. In a young green plant, 20–30% of the dry matter may be protein, while in very lean meat it may reach 50–70%.

**TABLE 1-4**
**Approximate Composition of Metabolically Active Cells and Tissues[a]**

| Component | *E. coli*[b] (%) | Green plant (spinach, *Spinacia oleracea*)[c] | Rat liver[d] (%) |
|---|---|---|---|
| $H_2O$ | 70 | 93 | 69 |
| Protein | 15 | 2.3 | 21 |
| Amino acids | 0.4 | | |
| DNA | 1 | | 0.2 |
| RNA | 6 | | 1.0 |
| Nucleotides | 0.4 | | |
| Carbohydrates | 3 | 3.2 | |
| Cellulose | | 0.6 | |
| Glycogen | | | 3.8 |
| Lipids | 2 | 0.3 | 6 |
| Phospholipids | | | 3.1 |
| Neutral lipids | | | 1.6 |
| Sterols | | | 0.3 |
| Other small molecules | 0.2 | | |
| Inorganic ions | 1 | 1.5 | |
| $K^+$ | | | 0.4 |

| | Equivalents per liter in rat liver |
|---|---|
| Amino acid residues | 2.1 |
| Nucleotide units | 0.03 |
| Glycogen (glucose units) | 0.22 |
| $K^+$ | 0.1 |

[a] Data were not readily available for spaces left blank
[b] From J. D. Watson (1976) *Molecular Biology of the Gene*, 3rd ed., p. 69, Benjamin, New York The amounts of amino acids, nucleotides, carbohydrates, and lipids include precursors present in the cell.
[c] From B. T. Burton (1976) *Human Nutrition*, 3rd ed., McGraw-Hill, New York (p. 505)
[d] From C. Long, ed., (1961) *Biochemists' Handbook*, pp. 677–679, Van Nostrand-Reinhold, Princeton, New Jersey

A dried tissue sample may be burned at a high temperature to an **ash**, which commonly amounts to 3–10% and is higher in specialized tissues such as bone. It is a measure of the inorganic constituents of tissues.

The carbohydrate content can be estimated by the difference of the sum of lipid, protein, and ash from 100%. It amounts to 50–60% in young green plants and only 2–10% in typical animal tissues. In exceptional cases the carbohydrate content of animal tissues may be higher; the glycogen content of oysters is 28%.

The amount of nucleic acid in tissues varies from 0.1% in yeast and 0.5–1% in muscle and in bacteria to 15–40% in thymus gland and sperm cells. In these latter materials of high nucleic acid content it is clear that multiplication of % N by 6.25 is not a valid measure of protein content. For diploid cells of the body the DNA content per cell is nearly constant.

Table 1-4 compares the composition of a bacterium, of a green plant, and of an active animal tissue (rat liver). Although the solid matter of cells consists principally of C, H, O, N, S, and P, many other chemical elements are also present. Among the cations, $Na^+$, $K^+$, $Ca^{2+}$, and $Mg^{2+}$ are found in relatively large amounts. Thus, the body of a 70 kg person contains 1050 g Ca (mostly in the bones), 245 g K, 105 g Na, and 35 g Mg. Iron (3 g), zinc (2.3 g), and rubidium (1.2 g) are the next most abundant. Of these iron and zinc are essential to life but rubidium is probably not. It is evidently taken up by the body together with potassium.

The other metallic elements in the human body amount to less than 1 g each, but at least seven of them play essential roles. They include copper (100 mg), manganese (20 mg), and cobalt (~5 mg). Others, such as chromium (<6 mg), tin, and vanadium, have only recently been shown essential for higher animals.[156,157] Nickel, lead, and others may perhaps be needed.

Nonmetallic elements predominating in the ash are phosphorus (700 g in the human body), sulfur (175 g), and chlorine (105 g). Not only are these three elements essential to all living cells but also selenium, fluorine, silicon (Box 4-B), iodine, and boron are needed by higher animals and boron by plants (Fig. 1-17). Iodine deficiency may affect one billion human beings and may cause 20 million cases per year of **cretinism**, or less severe brain damage.[158]

What is the likelihood that other elements will be found essential? Consider a human red blood cell, an object of volume ~80 $\mu m^3$ and containing about $3 \times 10^8$ protein molecules (mostly hemoglobin). About $7 \times 10^5$ atoms of the "trace metal" copper and $10^5$ atoms of the nutritionally essential tin are present in a single red cell. Also present are $2 \times 10^4$ atoms of silver, a toxic metal. Its concentration, over $10^{-7}$ M, is sufficient that it could have an essential catalytic function. However, we know of none and it may simply have gotten into our bodies from handling money, jewelry, and other

## BOX 1-F ABOUT THE REFERENCES

The lists of references at the ends of chapters are provided to encourage readers to look at original research articles. The lists are neither complete nor critically selected, but they do increase the information given in this book many-fold. I apologize for the important papers omitted. However, the references that are here will help a student to get started in reading the literature. Each reference contains other references and names of persons active in the field. By searching recent journal indices or a computer database it is easy to find additional articles by the same authors or on the same subject.

Look at the various types of scientific articles including reviews, preliminary reports and full research papers. Be sure to examine those in the *primary source journals* which publish detailed research results. These articles have always been sent to referees, active scientists, who check to see that the experiments are described accurately, that the authors have cited relevant literature, and that the conclusions are logical. Some journals, e.g., *Biochem. Biophys. Res. Comm.* and *FEBS Letts.* are dedicated to rapid publication of short reports but are also refereed. Other journals provide mostly reviews or a mixture of reviews. Periodical review series, such as *Advances in Nucleic Acid Chemistry and Related Topics*, often appear annually. Every student who intends to become a professional biochemist should consider purchasing the *Annual Review of Biochemistry* each year. This indispensible source of current information on most aspects of biochemistry is available to students at a very low price.

Many journal papers are difficult to read. To start, pick papers that have an understandable introduction. Choose reviews that are short, such as those in *Trends in Biochemical Sciences*. Then go on to the more comprehensive ones. Never sit back and hope that your computer will automatically fetch just what you need! Many journals carry papers of biochemical importance. Those specializing in biochemistry include the following:

| Full Title | Abbreviation |
|---|---|
| Advances in Carbohydrate Chemistry and Biochemistry[a] | Adv. Carbohydr. Chem. Biochem. |
| Advances in Protein Chemistry[a] | Adv. Protein Chem. |
| Analytical Biochemistry | Anal. Biochem. |
| Annual Review of Biophysics and Biomolecular Structure[a] | Ann. Rev. Biophys. Biomolec. Struct. |
| Annual Review of Biochemistry[a] | Ann. Rev. Biochem. |
| Archives of Biochemistry and Biophysics | Arch. Biochem. Biophys. |
| Biochemical and Biophysical Research Communications | Biochem. Biophys. Res. Commun. |
| Biochemical Journal | Biochem. J. |
| Biochemistry | |
| Biochimica et Biophysica Acta | Biochim. Biophys. Acta |
| Bioorganic Chemistry | Bioorg. Chem. |
| Carbohydrate Research | Carbohydr. Res. |
| EMBO Journal[b] | EMBO J. |
| European Journal of Biochemistry | Eur. J. Biochem. |
| FASEB Journal[c] | FASEB J. |
| Journal of Bacteriology | J. Bacteriol. |
| Journal of Biochemistry | J. Biochem. |
| Journal of Biological Chemistry | J. Biol. Chem. |
| Journal of Lipid Research | J. Lipid Res. |
| Journal of Molecular Biology | J. Mol. Biol. |
| Journal of the American Chemical Society | J. Am. Chem. Soc. |
| Journal of Theoretical Biology | J. Theor. Biol. |
| Methods in Enzymology[a] | Methods Enzymol. |
| Nature | |
| Nucleic Acids Research | Nucleic Acids Res. |
| Proceedings of the National Academy of Sciences, USA | Proc. Natl. Acad. Sci. U.S.A. |
| Science | |
| Structure | |
| Trends in Biochemical Sciences | Trends Biochem. Sci. or TIBS |

[a] These are not journals but series of review and reference books. There are many other series of "Advances in..." and "Annual Reviews of ..." that are not listed here.
[b] EMBO–European Molecular Biology Organization
[c] FASEB–Federation of American Societies for Experimental Biology

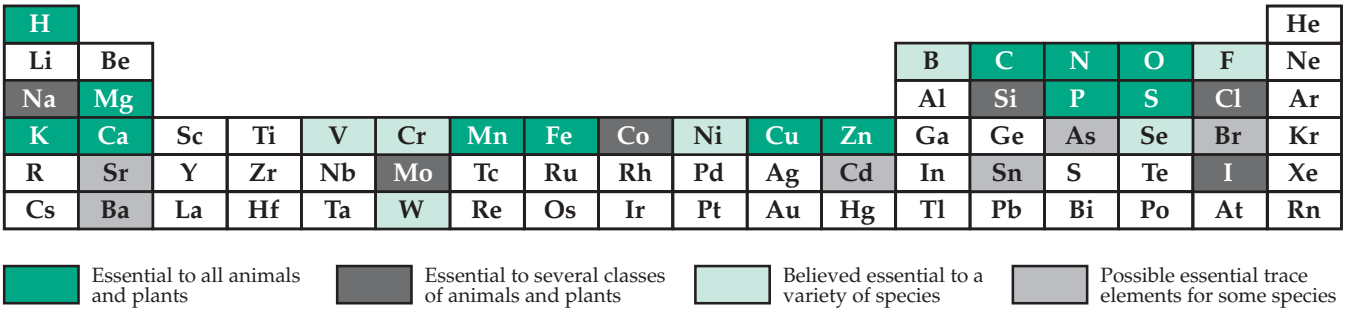



**Figure 1-17** Elements known to be essential to living things (after da Silva and Williams[157]). Essential elements are enclosed within shaded boxes. The 11 elements–C, H, O, N, S, P, Na, K, Mg, Ca, and Cl–make up 99.9% of the mass of a human being. An additional 13 are known to be essential for higher animals in trace amounts. Boron is essential to higher plants but apparently not to animals, microorganisms, or algae.

silver objects. The red blood cell also contains boron and aluminum ($3 \times 10^5$ atoms each), arsenic ($7 \times 10^5$ atoms), lead ($7 \times 10^4$ atoms), and nickel ($2 \times 10^4$ atoms). Of the elements (uranium and below) in the periodic table, only four (Ac, Po, Pa, and Ra) are present, on the average, in quantities less than one atom per cell.[156]

Of the apparently nonessential elements, several, e.g., Cs, Rb, Sr, and Ni (possibly essential) are not toxic at low concentrations. Others, such as Sb, As, Ba, Be, Cd, Pb, Hg, Ag, Tl, and Th, are highly toxic.

The ionic compositions of tissues and of body fluids vary substantially. Blood of marine organisms is similar to that of seawater in its content of $Na^+$, $Cl^-$, $Ca^{2+}$, and $Mg^{2+}$. Blood of freshwater and terrestrial organisms contains about ten times less $Na^+$ and $Cl^-$ and several times less $Ca^{2+}$ and $Mg^{2+}$ than is present in seawater, but it is nevertheless relatively rich in these ions.

In general, cells are rich in $K^+$ and $Mg^{2+}$, the $K^+$ predominating by far, and are poor in $Na^+$ and $Ca^{2+}$. Chloride is the principal inorganic anion, but organic carboxylate and phosphate groups contribute most of the negative charges (Table 1-4), many of which are fixed to proteins or other macromolecules. Ling estimated that cells typically contain about 1.66 M of amino acid residues in their proteins. Of these residues, 10% have negatively charged side chains and 8% positively charged. The difference is a net negative charge amounting to 33 mM within cells.[159.]

## References


1. May, R. M. (1992) *Sci. Am.* **267**(Oct), 42–48
2. Stanier, R. Y., Ingraham, J. L., Wheelis, M. L., and Painter, P. R. (1986) *The Microbial World*, 5th ed., Prentice-Hall, Englewood Cliffs, New Jersey
3. Brock, T. D., Smith, D. W., and Madigan, M. T. (1988) *Biology of Microorganisms*, 5th ed., Prentice-Hall, Englewood Cliffs, New Jersey
4. Neidhardt, F. C., ed. (1987) *Escherichia coli and Salmonella typhimurium*, Am. Soc. for Microbiology, Washington, D.C.
5. Neidhardt, F. C., Ingraham, J. L., and Schaechter, M. (1990) *Physiology of the Bacterial Cell*, Sinauer, Sunderland, Mass.
6. Maniloff, J., and Morowitz, H. J. (1972) *Bacteriol. Rev.* **36**, 263–290
7. Razin, S. (1978) *Microbiol. Rev.* **42**, 414–470
8. Barile, M. F., Razin, S., Tully, J. G., and Whitcomb, R. F., eds. (1979) *The Mycoplasmas*, Vol. I, Academic Press, New York
9. Watson, J. D. (1976) *Molecular Biology of the Gene*, 3rd ed., Benjamin, Menlo Park, California (p. 61)
10. Horowitz, D. M., and Sanders, J. K. M. (1994) *J. Am. Chem. Soc.* **116**, 2695–2702

10a. Gansheroff, L. J., and O'Brien, A. D. (2000) *Proc. Natl. Acad. Sci. U.S.A.* **97**, 2959–2961

11. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., III, and Venter, J. C. (1995) *Science* **270**, 397–403
12. Sonea, S. (1988) *Nature (London)* **331**, 216
13. Costerton, J. W., Geesey, G. G., and Cheng, K.-J. (1978) *Sci. Am.* **238**(Jan), 86–95
14. Macnab, R. M. (1987) in *Escherichia coli and Salmonella typhimurium* (Niedhardt, F. C., ed), pp. 70–83, Am. Soc. for Microbiology, Washington, D.C.
15. Margulis, L., Ashen, J. B., Sonea, S., Solé, M., and Guerrero, R. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 6966–6970
16. Goldstein, S. F., Charon, N. W., and Kreiling, J. A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 3433–3437
17. Levine, M. M. (1985) *N. Engl. J. Med.* **313**, 445–447
18. Kuehn, M. J., Heuser, J., Normark, S., and Hultgren, S. J. (1992) *Nature (London)* **356**, 252–255
19. Alper, J. (1993) *Science* **262**, 1817
20. Holt, J. G., Krieg, N. R., Sneath, P. H. A., Staley, J. T., and Williams, S. T., eds. (1994) *Bergey's Manual of Determinative Bacteriology*, 9th ed., Williams & Wilkins, Baltimore, Maryland
21. Doolittle, R. F. (1992) *Protein Sci.* **1**, 191 – 200
22. Schwartz, R. M., and Dayhoff, M. O. (1978) *Science* **199**, 395–403
23. Cavalier-Smith, T. (1986) *Nature (London)* **324**, 416–417
24. Jones, W. J., Nagle, D. P., Jr., and Whitman, W. B. (1987) *Microbiol. Rev.* **51**, 135–177
25. Woese, C. R. (1987) *Microbiol. Rev.* **51**, 221–271
26. Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A., and Peattie, D. A. (1989) *Science* **243**, 75–77
27. Garrett, R. A., Dalgaard, J., Larson, N., Kjems, J., and Mankin, A. S. (1991) *Trends Biochem. Sci.* **16**, 22–26
28. Searcy, D. G. (1982) *Trends Biochem. Sci.* **7**, 183–185
29. Wheelis, M. L., Kandler, O., and Woese, C. R. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2930–2934
30. Woese, C. R., Kandler, O., and Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576–4579

30a. Doolittle, W. F. (1999) *Science* **284**, 2124–2128

30b. Pennisi, E. (1999) *Science* **284**, 1305–1307

31. Cavalier-Smith, T. (1992) *Nature (London)* **356**, 570
32. Angert, E. R., Clements, K. D., and Pace, N. R. (1993) *Nature (London)* **362**, 239–241
33. Thimann, K. V. (1963) *The Life of Bacteria*, 2nd ed., Macmillan, New York
34. Stanier, R. Y., and Cohen-Bazire, G. (1977) *Ann. Rev. Microbiol.* **31**, 225–274
35. Morden, C. W., and Golden, S. S. (1989) *Nature (London)* **337**, 382–385
36. Marme, D., Marre, E., and Hertel, R., eds. (1982) *Plasmalemma and Tonoplast: Their Functions in the Plant Cell*, Elsevier, Amsterdam
37. Collins, F. S. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10821–10823
38. Hinshaw, J. E., Carraghen, B. O., and Milligan, R. A. (1992) *Cell* **69**, 1133–1141
39. Dingwall, C., and Lasky, R. (1992) *Science* **258**, 942–947
40. Nolta, K. V., Padh, H., and Steck, T. L. (1991) *J. Biol. Chem.* **266**, 18318–18323
41. de Duve, C. (1975) *Science* **189**, 186–194
42. Griffiths, G., and Simons, K. (1986) *Science* **234**, 438–443
43. Hartley, S. M. (1992) *Trends Biochem. Sci.* **17**, 325–327
44. Golgi, C. (1898) *Arch. Ital. Biol.* **30**, 60 and 278
45. Kirk, J. T. O., and Tilney-Bassett, R. A. E. (1978) *The Plastids: Their Chemistry, Structure, Growth, and Inheritance*, 2nd ed., Elsevier-North Holland Biomedical Press, Amsterdam
46. Hoffmann, H. P., and Avers, C. J. (1973) *Science* **181**, 749–751

46a. Capaldi, R. A. (2000) *Trends Biochem. Sci.* **25**, 212–214

46b. Rutter, G. A., and Rizzuto, R. (2000) *Trends Biochem. Sci.* **25**, 215–221

47. Fukui, S., and Tanaka, A. (1979) *Trends Biochem. Sci.* **4**, 246–249
48. Kindl, H., and Lazarow, P. B., eds. (1982) *Peroxisomes and Glyoxysomes*, Vol. 386, Annals of the New York Academy of Sciences, New York
49. de Duve, C. (1983) *Sci. Am.* **248**(May), 74–84
50. Masters, C., and Crane, D. (1995) *The Peroxisome: A Vital Organelle*, Cambridge Univ. Press, London
51. Lord, J. M., and Roberts, L. M. (1980) *Trends Biochem. Sci.* **5**, 271–274
52. Cerkasov, J., Cerkasovova, A., Kulda, J., and Vilhelmova, D. (1978) *J. Biol. Chem.* **253**, 1207–1214
53. Glover, D. M., Gonzalez, C., and Raff, J. W. (1993) *Sci. Am.* **268**(Jun), 62–68
54. Olmsted, J. B., and Borisy, G. G. (1973) *Ann. Rev. Biochem.* **42**, 507–540
55. Hall, J. L., Ramanis, Z., and Luck, D. J. L. (1989) *Cell* **59**, 121–132
56. Hyams, J. S. (1989) *Nature (London)* **341**, 485–486
57. Satir, P. (1961) *Sci. Am.* **204**(Feb), 108–116
58. Lazarides, E. (1980) *Nature (London)* **283**, 244–256
59. Amos, L. A., and Amos, W. B. (1991) *Molecules of the Cytoskeleton*, Guilford, New York
60. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9355–9359
61. Lake, J. A. (1991) *Trends Biochem. Sci.* **16**, 46–50
62. Margulis, L. (1981) *Symbiosis in Cell Evolution*, Freeman, San Francisco, California
63. Hurst, L. D. (1994) *Nature (London)* **369**, 451
64. Gupta, R. S., and Golding, G. B. (1996) *Trends Biochem. Sci.* **21**, 166–171
65. de Duve, C. (1996) *Sci. Am.* **274** (Apr), 50–57
66. Markert, C. L., Shaklee, J. B., and Whitt, G. S. (1975) *Science* **189**, 102–114
67. Fenchel, T., and Bernard, C. (1993) *Nature (London)* **362**, 300
68. Schopf, J. W. (1993) *Science* **260**, 640–646
69. Horgan, J. (1993) *Sci. Am.* **269**(Aug), 24
70. Doolittle, W. F. (1980) *Trends Biochem. Sci.* **5**, 146–149
71. May, R. M. (1986) *Nature (London)* **324**, 514–515
72. Seger, J. (1989) *Nature (London)* **337**, 305–306
73. Riley, M., and Anilionis, A. (1978) *Ann. Rev. Microbiol.* **32**, 519–560
74. Goossens, M., Dozy, A. M., Embury, S. H., Zachariades, Z., Hadjiminas, M. G., Stamatoyannopoulos, G., and Kan, Y. W. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 518–521
75. Williams, R. O., Young, J. R., and Majiwa, P. A. O. (1979) *Nature (London)* **282**, 847–849
76. Kondrashov, A. S. (1988) *Nature (London)* **336**, 435–440
77. Buchsbaum, R., Buchsbaum, M., Pearse, J., and Pearse, V. (1987) *Animals Without Backbones*, 3rd ed., Univ. Chicago Press, Chicago, Illinois
78. Margulis, L., and Schwartz, K. V. (1999) *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth*, 3rd ed., Freeman, New York
79. Laver, W. G., Air, G. M., Dopheide, T. A., and Ward, C. W. (1980) *Nature (London)* **283**, 454–457
80. McKanna, J. A. (1973) *Science* **179**, 88–90
81. Miller, L. H. (1992) *Science* **257**, 36–37
82. Wyler, D. J. (1993) *N. Engl. J. Med.* **329**, 31–37
83. Miller, L. H., Good, M. F., and Milon, G. (1994) *Science* **264**, 1878–1883
84. Nussenzweig, R. S., and Long, C. A. (1994) *Science* **265**, 1381–1382
85. Sibley, L. D., and Boothroyd, J. C. (1992) *Nature (London)* **359**, 82–85
86. Soldati, D., and Boothroyd, J. C. (1993) *Science* **260**, 349–352
87. Donelson, J. E., and Turner, M. J. (1985) *Sci. Am.* **252**(Feb), 44–51
88. Blum, M. L., Down, J. A., Gurnett, A. M., Carrington, M., Turner, M. J., and Wiley, D. C. (1993) *Nature (London)* **362**, 603–609
89. Gall, J. G., ed. (1986) *The Molecular Biology of Ciliated Protozoa*, Academic Press, Orlando, Florida
90. Olive, L. S. (1975) *The Mycetozoans*, Academic Press, New York
91. Kimmel, A. R., ed. (1989) *Molecular Biology of Dictyostelium Development*, Liss, New York
92. Traynor, D., Kessin, R. H., and Williams, J. G. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8303–8307
93. Webster, J. (1980) *Introduction to Fungi*, 2nd ed., Cambridge Univ. Press, New York
94. Davies, D. A. L., and Pope, A. M. S. (1978) *Nature (London)* **273**, 235–236
95. Georgopapadakou, N. H., and Walsh, T. J. (1994) *Science* **264**, 371–373
96. Sternberg, S. (1994) *Science* **266**, 1632–1634
97. Davis, R. H., and DeSerres, F. J. (1970) *Meth. Enzymol.* **17A**, 79–148
98. Botstein, D., and Fink, G. R. (1988) *Science* **240**, 1439–1443
99. Pringle, J. R., Broach, J. R., and Jones, E. W., eds. (1991–1994) *The Molecular and Cellular Biology of the Yeast Saccharomyces*, Cold Springs Harbor Laboratory Press, Cold Spring Harbor, New York (3 Vols.)

## References


100. Wheals, A. E., Rose, A. H., and Harrison, S. J., eds. (1995) *The Yeasts*, 2nd ed., Vol. 6, Academic Press, San Diego, California
101. Gimeno, C. J., Ljungdahl, P. O., Styles, C. A., and Fink, G. R. (1992) *Cell* **68**, 1077–1090
102. Oliver, S. G., and 146 other authors (1992) *Nature (London)* **357**, 38–46
103. Williams, N. (1995) *Science* **268**, 1560–1561
104. Ruehle, J. L., and Marx, D. H. (1979) *Science* **206**, 419–422
105. Canter-Lund, H., and Lund, J. W. G. (1995) *Freshwater Algae: Their Microscopic World Explored*, Biopress, Bristol, England
106. Harris, E. H. (1989) *The Chlamydomonas Sourcebook*, Academic Press, San Diego, California
107. Brachet, J. L. A. (1965) *Endeavour* **24**, 155–161
108. Bonotto, S., Kefeli, V., and Puiseux-Dao, S. (1979) *Developmental Biology of Acetabularia*, Elsevier-North Holland, Amsterdam
109. Berger, S., and Kaever, M. J. (1992) *Dasycladales: An Illustrated Monograph of a Fascinating Algal Order*, Oxford Univ. Press, New York
110. Menzel, D., and Elsner-Menzel, C. (1990) *Protoplasm* **157**, 52–63
111. Jaenicke, L. (1982) *Trends Biochem. Sci.* **7**, 61–64
112. Ahmadjian, V. (1962) in *Physiology and Biochemistry of Algae*, Vol. 2 (Lewin, R. A., ed), pp. 817–822, Academic Press, New York
113. Morris, S. C. (1993) *Nature (London)* **361**, 219–221
114. Levinton, J. S. (1992) *Sci. Am.* **267**(Nov), 84–91
115. Keese, P. K., and Gibbs, A. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9489–9493
116. Hickman, C. P. (1973) *Biology of the Invertebrates*, Mosby, St. Louis, Missouri
117. Villee, C. A., Walker, W. F., Jr., and Barnes, R. D. (1973) *General Zoology*, Saunders, Philadelphia, Pennsylvania
118. Lapan, E. A., and Morowitz, H. (1972) *Sci. Am.* **227**(Dec), 94–101
119. Lapan, E. A. (1975) *Comp. Biochem. Physiol.* **52A**, 651–657
120. Bergquist, P. R. (1978) *Sponges*, Hutchinson, London
121. Grimmelikhuijzen, C. J. P., and Schaller, H. C. (1979) *Trends Biochem. Sci.* **4**, 265–267
122. Lenhoff, H. M., and Lenhoff, S. G. (1988) *Sci. Am.* **258**(Apr), 108–113
123. Loomis, W. F. (1959) *Sci. Am.* **200**(Apr), 150
123a. Hickman, C. P. *Integrated Principles of Zoology* (1966) Mosby, St. Louis
124. Trager, W. (1986) *Living Together: The Biology of Animal Parasitism*, Plenum, New York
125. Cherfas, J. (1991) *Science* **251**, 630–631
126. Kolberg, R. (1994) *Science* **264**, 1859–1861
127. Wood, W. B., ed. (1988) *The Nematode Caenorhabditis elegans*, Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York
128. Kenyon, C. (1988) *Science* **240**, 1448–1453
129. Roberts, L. (1990) *Science* **248**, 1310–1313
129a. Hotez, P. J., and Pritchard, D. I. (1995) *Sci. Am.* **272**(Jun), 68–74
130. Dumont, H. J., and Green, J., eds. (1980) *Rotatoria*, p. 42, Junk, The Hague
131. Mill, P. J., ed. (1978) *Physiology of Annelids*, Academic Press, New York
132. Florkin, M., and Scheer, B. T., eds. (1970, 1971) *Chemical Zoology*, Vol. 5 and 6, Academic Press, New York
133. Goldstein, L. S. B., and Fyrberg, E. A., eds. (1994) *Drosophila melanogaster: Practical Uses in Cell and Molecular Biology*, Academic Press, San Diego, California
134. Lawrence, P. (1992) *The Making of a Fly: The Genetics of Animal Design*, Blackwell, Oxford
134a. Rubin, G. M., and Lewis, E. B. (2000) *Science* **287**, 2216–2218
135. Maddy, A. H. (1992) *Trends Biochem. Sci.* **17**, 125–126
135a. Harris, J. R., ed. (1991) *Blood Cell Biochemistry*, Plenum, New York (Three volumes)
136. Johnston, R. B., Jr. (1988) *N. Engl. J. Med.* **318**, 747–754
137. Freshney, R. I. (1994) *Culture of Animal Cells*, 3rd ed., Wiley-Liss, New York
138. Feldman, J., Gilula, N. B., and Pitts, J. D., eds. (1978) *Intercellular Junctions and Synapses*, Halsted Press, New York
139. Staehelin, L. A., and Hull, B. E. (1978) *Sci. Am.* **238**(May), 141–153
140. Bradbury, M. (1979) *The Concept of a Blood-Brain Barrier*, Wiley, Chichester, New York
141. Friend, D. S., and Gilula, N. B. (1972) *J. Cell Biol.* **53**, 771
142. Gilula, N. B. (1975) *Cellular Membranes and Tumor Cells*, Williams & Wilkins Co., Baltimore, Maryland (p. 221)
143. Gilula, N. B. (1974) in *Cell Communication* (Cox, R. P., ed), pp. 1–29, Wiley, New York
144. Gilula, N. B. (1972) *Nature (London)* **235**, 262–265
145. Zimmer, D. B., Green, C. R., Evans, W. H., and Gilula, N. B. (1987) *J. Biol. Chem.* **262**, 7751–7763
146. Hertzberg, E. L., and Johnson, R. G., eds. (1988) *Gap Junctions*, Liss, New York
147. Robinson, S. R., Hampson, E. C. G. M., Munro, M. N., and Vaney, D. I. (1993) *Science* **262**, 1072–1074
148. Stauffer, P. L., Zhao, H., Luby-Phelps, K., Moss, R. L., Star, R. A., and Muallem, S. (1993) *J. Biol. Chem.* **268**, 19769–19775
149. Veenstra, R. D., and DeHaan, R. L. (1986) *Science* **233**, 272–274
150. Moscona, A. A. (1961) *Sci. Am.* **205**(Sep), 143–162
151. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1983) *Molecular Biology of the Cell*, pp. 679–680, Garland Publ., New York
152. Heaysman, J. E. M. (1978) *Intl. Rev. Cytol.* **55**, 49–66
153. Palca, J. (1989) *Science* **245**, 131
154. Lloyd, A. M., Barnason, A. R., Rogers, S. G., Byrne, M. C., Fraley, R. T., and Horsch, R. B. (1986) *Science* **234**, 464–466
155. Biddulph, S., and Biddulph, O. (1959) *Sci. Am.* **200**(Feb), 44–49
156. Bowen, H. J. M. (1966) *Trace Elements in Biochemistry*, Academic Press, New York
157. Fraústo da Silva, J. J. R., and Williams, R. J. P. (1991) *The Biological Chemistry of the Elements: The Inorganic Chemistry of Life*, Clarendon Press, Oxford
158. Hetzel, B. S. (1994) *N. Engl. J. Med.* **331**, 1770–1771
159. Ling, G. N. (1962) *A Physical Theory of the Living State*, Ginn (Blaisdell), Boston, Massachusetts

## Study Questions

1. Describe the principal structural or organizational differences between prokaryotic and eukaryotic cells.

2. Describe two or more principal functions of proteins within cells, one function of DNA, two or more functions of RNA, and one function of lipids.

3. Compare the chemical makeup of ribosomes, of cell membranes, and of bacterial flagella.

4. Assume the following dimensions: Mycoplasma, sphere, 0.33 µm diameter; *E. coli*, cylinder, 0.8 µm diameter x 2 µm; liver cell, sphere 20 µm; root hair, cylinder, 10 µm diameter x 1 mm.

   a. Calculate for each cell the total volume, the mass in grams and in daltons (assume a specific gravity of 1.0).

   b. Assume that bacterial ribosomes are approximately spherical with a diameter of 23 nm. What is their volume? If the mass of a bacterial ribosome is $2.7 \times 10^6$ daltons, what is its apparent density (divide mass by volume)? Experimentally the buoyant density of bacterial ribosomes in a cesium chloride gradient (Chapter 5) is about 1.6 g/cm$^3$. How can this difference be explained? If eukaryotic ribosomes are 1.17 times larger than bacterial ribosomes in linear dimensions, what is the volume of a eukaryotic ribosome?

   c. What fraction of volume of *E. coli* consists of cell wall, of plasma membrane, of ribosomes (assume 15,000 are present)? If a cell of *E. coli* is 80% water, what fraction by weight of the total solids consists of ribosomes? Of DNA (assuming 2 chromosomes per cell)?

   d. What fraction by volume of a liver cell is composed of ribosomes, of plasma membrane, of mitochondria (assume 1000 mitochondria)? What fraction is accounted for by the nucleus?

5. a. What is the molar concentration of an enzyme of which only one molecule is present in an *E. coli* cell?

   b. Assume that the concentration of $K^+$ within an *E. coli* cell is 150 mM. Calculate the number of $K^+$ ions in a single cell.

   c. If the pH inside the cell is 7.0, how many $H^+$ ions are present?

6. If chromosomes (and chromatin) are 15% DNA, what will be the mass of 23 pairs of chromosomes in a human diploid cell? If the nucleus has a diameter of 5 mm and a density of 1.1 g/cm$^3$, what fraction by weight of the nucleus is chromatin?

7. Compare the surface to volume ratios for an *E. coli* cell, a liver cell, the nucleus of a eukaryotic cell, a root hair. If a cell of 20 µm diameter is 20% covered with microvilli of 0.1 µm diameter and 1 µm length centered on a 0.2 µm spacing, how much will the surface/volume ratio be increased?

8. It has been shown that the code for specifying a particular amino acid in a protein is determined by a sequence of three nucleotides (a codon) in a DNA chain. There are four different kinds of nucleotide units in DNA. How many different codons exist? Note that this is larger than the number of different amino acids (20) that are incorporated into proteins plus the three stop (termination) codons (see Tables 5-5 and 5-6 for a list of codons).

## Study Questions

9. State two similarities and two differences between cyanobacteria (blue-green algae) and green algae.

10. Compare the sizes and structures of bacterial and eukaryotic flagella.

11. How much larger in volume is a typical eukaryotic cell compared to a bacterium?

12. Compare the structure and properties of mitochondria, chloroplasts, and peroxisomes.

13. What are the possible origins of mitochondria and chloroplasts? What evidence can you cite to support your answer?

14. How many different kinds of polymers, e.g. proteins, RNA, that are present in or around living cells, can you name? Can you name some subgroups in any of your categories?

15. Compare the composition of these three, especially with respect to C, H, O, N, S, P, Fe, Cu, Al, and Si:

    a. The earth's crust

    b. Ocean water

    c. Cytoplasm

A spider's orb-web is formed by extrusion of a concentrated protein solution and stretching of the resulting fiber. The cross-strands, which are stronger than steel, resemble silkworm silk. The molecules contain microcrystalline β sheet domains that are rich in Gly-Ala repeats as well as polyalanine segments. The capture spiral is formed from much more elastic molecules that contain many β-turn-forming sequences. These assume a spring-like β spiral. See Box 2-B.



# Contents

# Amino Acids, Peptides, and Proteins

# 2

Thousands of different proteins make up a very large fraction of the "machinery" of a cell. Protein molecules catalyze chemical reactions, carry smaller molecules through membranes, sense the presence of hormones, and cause muscle fibers to move. Proteins serve as structural materials within cells and between cells. Proteins of blood transport oxygen to the tissues, carry hormones between cells, attack invading bacteria, and serve in many other ways. No matter what biological process we consider, we find that a group of special proteins is required.

The amino acid units that make up a protein molecule are joined together in a precise sequence when the protein is made on a ribosome. The chain is then folded, often into a very compact form. Sometimes the chain is then cut in specific places. Pieces may be discarded and parts may be added. A metal ion, a coenzyme derived from a vitamin, or even a single methyl group may be attached to form the biologically active protein. The final product is a complex and sophisticated machine, often with moving parts, that is exquisitely designed for its particular role.

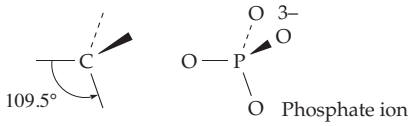
The biological functioning of a protein is determined both by the properties of the chemical groups in the amino acids that are joined to form the protein chain and by the way the chain is folded. The ways in which the different parts of the protein interact with each other and with other molecules are equally important. These interactions play a major role in determining the folding pattern and also provide much of the basis for the biological functioning of proteins. Similar considerations apply also to carbohydrates, nucleic acids, and other biopolymers. For these reasons it is appropriate to review some fundamentals of molecular structure and geometry.

## A. Structural Principles for Small Molecules

Stable organic molecules are held together by covalent bonds which are usually very strong. The standard Gibbs energies of formation ($\Delta G_f^\circ$)[†] of many covalent single bonds are of the order of −400 kJ/mol (96 kcal/mol). The bonds have definite directions, which are measured by **bond angles** and definite **bond lengths**.

### 1. Bond Angles

Because of the tetrahedral arrangement of the four bonds around single-bonded carbon atoms and most phosphorus atoms, all six of the bond angles about the central atom have nearly the same tetrahedral angle of 109.5°.




Bond angles within chains of carbon atoms in organic compounds vary only slightly from this, and even atoms that are attached to fewer than four groups usually have similar angles; for example, the H–O–H angle in a water molecule is 105°, and the H–N–H angles of ammonia are 107°. In ethers the C–O–C angle is 111°. However, bond angles of only 101° are present in $H_2O_2$ and of 92° in $H_2S$ and $PH_3$.

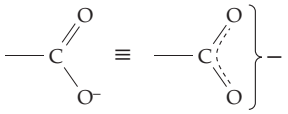
[†]**See Chapter 6 for a review of thermodynamics**

The presence of **double bonds** leads to **planarity** and to compounds with bond angles of 120°, the internal angle in a hexagon. The planar geometry imposed upon an atom by a double bond is often transmitted to an adjacent nitrogen or oxygen atom as a result of **resonance** (Section 6). For example, the amide groups that form the peptide linkage in proteins (see Fig. 2-5) are nearly planar and the angles all fall within four degrees of 120°.

## 2. Bond Lengths

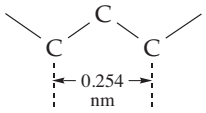
Chemists describe bond lengths as the distances between the nuclei of bonded atoms. The C–C single bond has a length of 0.154 nm (1.54 Å). The C–O bond is ~0.01 nm shorter (0.143 nm), and the typical C–H bond has a length of ~0.109 nm. The C–N bond distance is halfway between that for C–C and C–O (0.149 nm). Other lengths, such as that of O–H, can be estimated from the covalent radii given in Table 2-1.

The length of a double bond between any two atoms (e.g., C=C) is almost exactly 0.020 nm less than that for a single bond between the same atoms. If there is resonance, hence only partial double bond character, the shortening is less. For example, the length of the C–C bond in benzene is 0.140 nm; the C–O distances in the carboxylate anion are 0.126 nm.

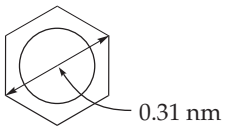


Carboxylate ion

Using simple geometry, it is easy to calculate overall lengths of molecules; here are two distances worth remembering:

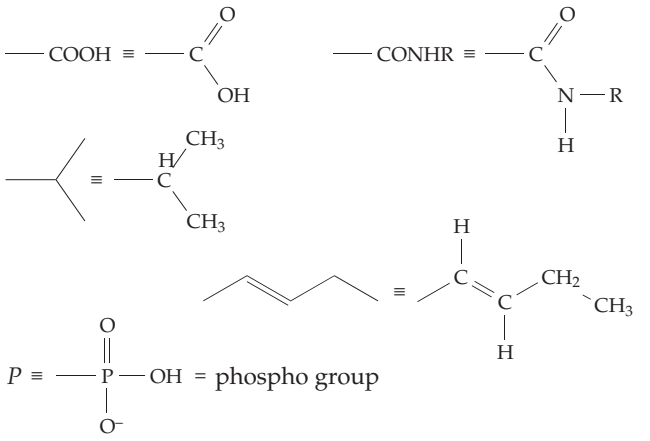


Distance between alternate C atoms in fully extended hydrocarbon chain



Distance across a benzene ring

In the preceding simplified structural formula for benzene the six hydrogen atoms have been omitted. Resonance between the two possible arrangements of the three double bonds[1] is indicated by the circle. Chemical shorthand of the following type is used throughout the book. Carbon atoms may be represented by an angle or the end of a line, but other atoms will always be shown.




## 3. Contact Distances

Covalent bond distances and angles tell us how the atomic nuclei are arranged in space but they do not tell us anything about the outside surfaces of molecules. The distance from the center of an atom to the point at which it contacts an adjacent atom in a packed structure such as a crystal (Fig. 2-1) is known as the **van der Waals radius**. The ways in which biological molecules fit together are determined largely by the van der Waals contact radii. These, too, are listed in Table 2-1. In every case they are approximately equal to the *covalent radius plus 0.08 nm*. Van der Waals radii

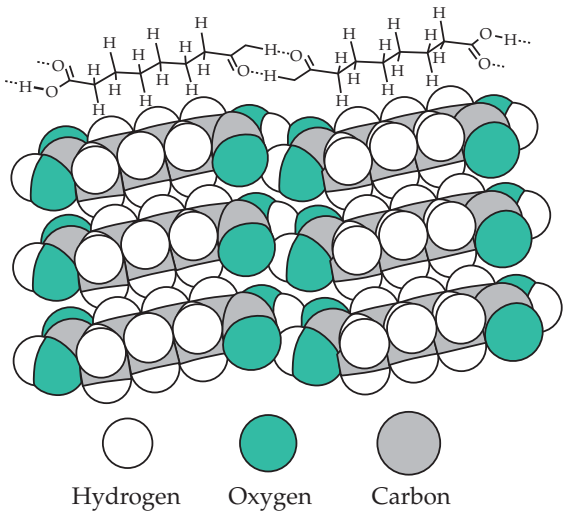


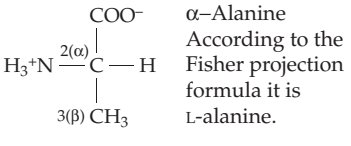

**Figure 2-1** Packing of molecules of suberic acid HOOC–$(CH_2)_6$–COOH in a crystal lattice as determined by neutron diffraction.[2] Notice the pairs of hydrogen bonds that join the carboxyl groups at the ends of the molecules and also the close contact of hydrogen atoms between the chains. Only the positions of the hydrogen nuclei were determined; the van der Waals radii have been drawn around them. However, the radii were originally determined from X-ray and neutron diffraction data obtained from many different crystalline compounds.

are not as constant as covalent radii because atoms can be "squeezed" a little, but only enough to decrease the contact radii by 0.005–0.01 nm. The radii of space-filling molecular models are usually made a little smaller than the actual scaled van der Waals radii to permit easier assembly.

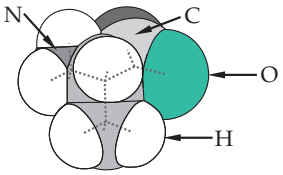
## 4. Asymmetry: Right-Handed and Left-Handed Molecules

The left hand looks much like the right hand, but they are different. One is the mirror image of the other. A practical difference is that your right hand will not fit into a left-handed glove. Despite our daily acquaintance with "handedness" it may seem difficult to explain in words how a right and a left hand differ. However, since most biochemical compounds are asymmetric,[3] it is important to be able to visualize these molecules in three dimensions and to draw their structures on paper. One of the best ways of learning to do this is to study molecular models. You may learn the most by making your own models (see Appendix).

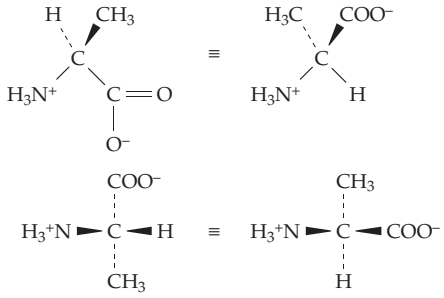
Whenever four different groups are bonded to a central carbon atom, the molecule is asymmetric and the four groups can be arranged in two different **configurations**. Consider alanine, one of the alpha (α)-amino acids from which proteins are built.




α–Alanine
According to the Fisher projection formula it is L-alanine.




It is called an α-amino acid because the amino group is attached to the α (or number 2) carbon atom. To indicate its three-dimensional structure on a flat piece of paper, the bonds that project out of the plane of the paper and up toward the reader are often drawn as elongated triangles, while bonds that lie behind the plane of the paper are shown as dashed lines. The isomer of alanine having the configuration about the α-carbon atom shown in the following structural formulas is called *S*-alanine or L-alanine. The isomer which is a mirror image of *S*-alanine is *R*-alanine or D-alanine. Pairs of *R* and *S* compounds (see Section B for definitions) are known as **enantiomorphic** forms or **enantiomers**.




*S*-Alanine (L-Alanine)

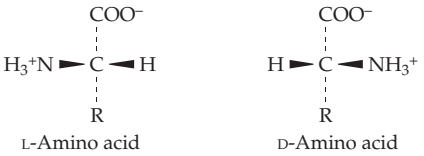
Notice that in the foregoing drawings, the carboxyl group ( —C(=O)OH ), abbreviated –COOH, is shown as having lost a proton to form the carboxylate ion $-COO^-$. Likewise, the amino group ($-NH_2$) has gained a proton to form the $-NH_3^+$ ion. The resulting **dipolar ionic** or **zwitterionic** structure is the one that actually exists for amino acids both in solution and in crystals.

**TABLE 2-1**
**The Sizes of Some Atoms[a–c]**

| Element | Covalent radius (nm)[a,b,d] | van der Waals radii (nm)[d] ***a*** Sideways contact[c] | ***b*** Polar contact[c] |
|---|---|---|---|
| H | 0.030 | 0.12 | |
| F | 0.064 | 0.138 | 0.130 |
| C | 0.077 | 0.16 | |
| N | 0.070 | 0.160 | 0.160 |
| O | 0.066 | 0.154 | 0.154 |
| Cl | 0.099 | 0.178 | 0.158 |
| Si | 0.117 | | |
| P | 0.110 | 0.19 | |
| S | 0.104 | 0.203 | 0.160 |
| Br | 0.114 | 0.184 | 0.154 |
| I | 0.133 | 0.213 | 0.176 |
| Se | | 0.215 | 0.170 |
| "Radius" of methyl group | | 0.20 | |
| Half-thickness of aromatic molecules | | 0.170 | |

[a] From Pauling, L. (1960) *The Nature of the Chemical Bond*, 3rd ed., Cornell Univ. Press, Ithaca, New York (pp. 224–227 and 260).
[b] Covalent radii for two atoms can be summed to give the interatomic distance. The van der Waals radii determine how closely molecules can pack. The closest observed contacts between atoms in macromolecules are approximately 0.02 nm less than the sum of the van der Waals radii. From Sasisekharian, V., Lakshminarayanan, A. V., and Ramachandran, G. N. (1967) in *Conformation of Biopolymers*, Vol. 2 (Ramachandran, G. N., ed), Academic Press, New York, (p. 641)
[c] Nyburg, S. C and Faerman, C. H. (1985) *Acta Crystal*. **B41**, 274–279 Shapes of many atomic surfaces are elliptical. The major radius ***a*** applies to sideways contacts and the minor radius ***b*** to "polar" contacts along a covalent bond axis. Distances are for atoms singly bonded to C and may differ slightly if bonds are to other atoms.
[d] For distances in Å multiply by 10.

***The D- and L- families of amino acids.*** The amino acids of which proteins are composed are related to L-alanine but have various side chains (R groups) in place of the methyl group of alanine. In the preceding section the structure of L-alanine was given in four different ways. To recognize them all as the same structure, we can turn them in space to an orientation in which the carboxyl group is at the top, the side chain ($-CH_3$) is down, and both project behind the paper. The amino group and hydrogen atom will then project upward from the paper at the sides as shown below. According to a convention introduced at the beginning of this century by Emil Fischer, *an amino acid is L if, when oriented in this manner, the amino group lies to the left and D if it lies to the right.*

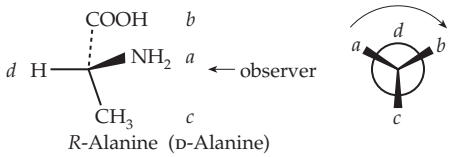



Fischer further proposed that the amino acid in this orientation could be projected onto the paper and drawn with ordinary lines for all the bonds. This gives the previously shown **Fischer projection formula** of L-alanine.

Although the D and L system of designating configuration is old it is still widely used. Remember that D and L refer to the absolute configurations about a selected reference atom in the molecule; for an amino acid this is the number 2 or α-carbon. A quantity that is related to the asymmetry of molecules is the experimentally measurable **optical rotation** (Chapter 23). The sign of the optical rotation (+ or –) is sometimes given together with the name of a compound, e.g., D(+)-glucose. The older designations *d* (dextro) and *l* (levo) indicated + and –, respectively. However, compounds with the D configuration may have either + or – optical rotation.

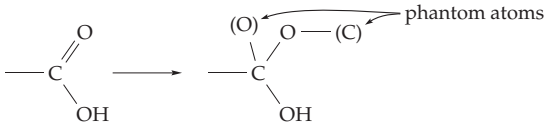
In older literature optical isomerism of the type represented by D and L pairs was usually discussed in terms of "asymmetric carbon atoms" or "asymmetric centers." Now the terms **chiral** (pronounced *ki-ral*) **molecules**, chiral centers, and **chirality** (Greek: "handedness") are preferred.

***The RS notation for configuration.*** This notation, devised by Cahn, Ingold, and Prelog, provides an unambiguous way of specifying configuration at any chiral center.[4,5] It is especially useful for classes of compounds for which no well-established DL system is available. The groups or atoms surrounding the central carbon atom, or other central atom, are ranked according to a **priority sequence**. The priority of a group is determined by a number of sequence rules, the first of which is (1) *Higher atomic number precedes lower*. In the following illustration, the priorities of the groups in D-alanine are indicated by the letters $a > b > c > d$. The highest priority (*a*) is assigned to the $NH_2$ groups which contain nitrogen bonded to the central atom. To establish the configuration, the observer views the molecule down the axis connecting the central atom to the group having the lowest priority, i.e., to group *d*. Viewed in this way, the sequence of groups *a*, *b*, and *c* can either be that of a right-handed turn (clockwise) as shown in the drawing or that of a left-handed turn (counterclockwise).




The view down the axis and toward the group of lowest priority (d), which lies behind the page. The right-handed turn indicates the configuration R (rectus = right); the opposite configuration is S (sinister = left).

To establish the priority sequence of groups first look at the atoms that are bonded directly to the central atom, arranging them in order of decreasing atomic number. Then if necessary, move outward to the next set of atoms, again comparing atomic numbers. In the case of alanine, groups *b* and *c* must be ordered in this way because they both contain carbon directly bonded to the central atom. When double bonds are present at one of the atoms being examined, e.g., the carboxyl group in alanine, imagine that **phantom atoms** that replicate the real ones are present at the ends of the bonds:




These phantom atoms fill out the valences of the atoms involved in the multiple bonds and are considered to have zero atomic number and zero mass. They are not considered in establishing priorities.

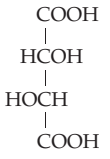
If the first rule and the expansion of multiple bonds are not sufficient to establish the priority, use these additional rules: (2) *Higher atomic mass precedes lower.* (3) When a double bond is present *Z precedes E* (see Geometrical isomers). For ring systems a *cis* arrangement of the highest priority substituents precedes *trans*. (4) When a pair of chiral centers is present *R,R* or *S,S* precedes *R,S* or *S,R*. (5) *An R chiral center precedes S.* For further details see Eliel *et al.*[5] and Bentley.[6] The following groups are ordered in terms of *decreasing priority*[6]: SH > OR > OH > NH–$COCH_3$ > $NH_2$ > COOR > COOH > CHO > $CH_2OH$ > $C_6H_5$ > $CH_3$ > $^3H$ > $^2H$ > H.

Although the *RS* system is unambiguous, closely

related compounds that belong to the same configurational family in the DL system may have opposite configurations in the *RS* system. Thus, L-cysteine (side chain $-CH_2SH$) has the *R* configuration. This is one of the reasons that the DL system is still used for amino acids and sugars.

***Diastereoisomers.*** Whereas compounds with one chiral center exist as an enantiomorphic pair, molecules with two or more chiral centers also exist as diastereoisomers (diastereomers). These are pairs of isomers with an opposite configuration at one or more of the chiral centers, but which are not complete mirror images of each other. An example is L-threonine which has the 2*S*, 3*R* configuration. The diastereoisomer with the 2*S*, 3*S* configuration is known as L-*allo*-threonine. L-isoleucine, whose side chain is $-CH(CH_3)\,CH_2CH_3$, has the 2*S*, 3*R* configuration. It can be called 2(*S*)-amino-3(*R*)-methyl-valeric acid but the simpler name L-isoleucine implies the correct configuration at both chiral centers.

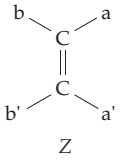
Sometimes the subscript s or g is added to a D or L prefix to indicate whether the chirality of a compound is being related to that of serine, the traditional configurational standard for amino acids, or to that of glyceraldehyde. In the latter case the sugar convention (Chapter 4) is followed. In this convention the configurations of the chiral centers furthest from C1 are compared. Ordinary threonine is $L_s$- or $D_g$-threonine. The configuration of dextrorotatory (+)-tartaric acid can be described as 2*R*, 3*R*, or as $D_s$, or as $L_g$.



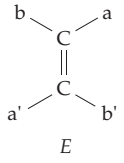
2*R*, 3*R*-tartaric acid

Biochemical reactions are usually stereospecific and a given enzyme will catalyze reactions of molecules of only a single configuration. A related fact is that proteins ordinarily consist entirely of amino acids of the L series.

***Geometrical isomers.*** The *RS* system also gives an unambiguous designation of geometrical isomers containing a double bond.[5,7] At each end of the bond, select the group of highest priority. If these two groups lie on the same side of the double bond the configuration is ***Z*** (from the German **zusammen**, "together"); if on opposite sides ***E*** (**entgegen**, "opposite").



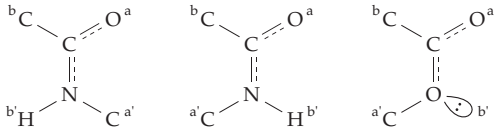
*Z*



*E*

a and a' higher priority than b or b'

Configurations of amide or ester linkages may also be specified in this manner. This is possible because the C–N bond of an amide has partial double-bond character, as to a lesser extent does the C–O bond to the bridge oxygen in an ester. In this case, assign the lowest priority to the unshared electron pair on the ester bridge oxygen.



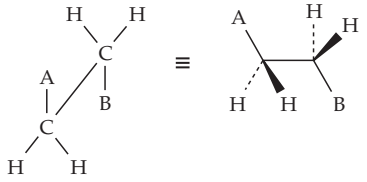
*Z* or *trans* *E* or *cis*

An amide of the *Z* configuration is ordinarily referred to as *trans* in protein chemistry because the main chain atoms are trans.

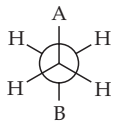
## 5. Conformations: The Shapes That Molecules Can Assume

As important to biochemists as configurations, the stable arrangements of bonded atoms, are **conformations**, the various orientations of groups that are caused by rotation about single bonds.[5,8] In many molecules such rotation occurs rapidly and freely at ordinary temperatures. We can think of a $-CH_3$ group as a kind of erratic windmill, turning in one direction, then another. However, even the simplest molecules have *preferred conformations*, and in more complex structures rotation is usually very restricted.

Consider a molecule in which groups A and B are joined by two $CH_2$ (methylene) groups. If A and B are pulled as far apart as possible, the molecule is in its fully extended **anti** or **staggered** conformation:




The antiperiplanar conformation

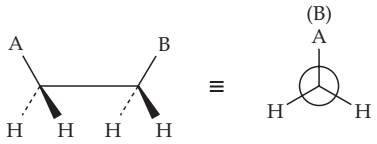


≡

View down the axis joining the C atoms (Newton projection)

Groups A and B are said to be **antiperiplanar** (*ap*) in this conformation. Not only are A and B as far apart as possible but also all of the hydrogen atoms are at their maximum distances one from the other. This can be seen by viewing the molecule down the axis joining

the carbon atoms (Newman projection). Rotation of the second carbon atom 180° around the single bond yields the **eclipsed** conformation in which groups A and B are synperiplanar.

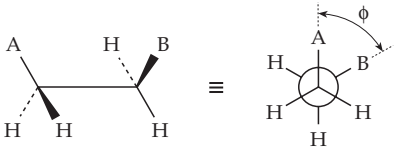



Eclipsed or synperiplanar conformation

If A and B are large bulky groups they will bump together, attainment of the eclipsed conformation will be almost impossible, and rotation will be severely restricted. Even if A and B are hydrogen atoms (ethane), there will be a rotational barrier in the eclipsed conformation which amounts to ~12 kJ (3 kcal) per mole because of the crowding of the hydrogen atoms as they pass each other.[5,9] This can be appreciated readily by examination of space-filling molecular models.

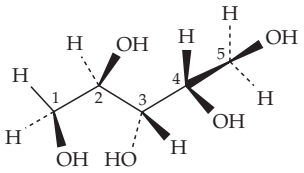
If groups A and B are methyl groups (butane), the steric hindrance between A and B leads to a rotational barrier of ~25 kJ ( 6 kcal) per mole. The consequence of this simple fact is that in fatty acids and related substances and in polyethylene the chains of $CH_2$ groups tend to assume fully extended zigzag conformations.

In addition to this extended conformation there are two **gauche** (**skewed** or synclinal) conformations which are only slightly less stable than the staggered conformation and in which A and B interfere only if they are very bulky. In one of the two gauche conformations B lies to the right of A and in the other to the left of A when viewed down the axis.




One of the two gauche or synclinal conformations

These two conformations are related to right-handed and left-handed screws, respectively. The threads on an ordinary right-handed household screw, when viewed down the axis from either end, move backward from left to right in the same fashion as do the groups A and B in the illustration. The angle ϕ is the **torsion angle** and is positive for right-handed conformations. Gauche conformations are important in many biological molecules; for example, the sugar alcohol **ribitol** stacks in crystals in a "sickle" conformation,[6] in which the chain starts out (at the left) in the zigzag arrangement but shifts to a gauche conformation around the fourth carbon atom, thereby minimizing steric interference between the OH groups on the second and fourth carbons.




Ribitol in sickle conformation

The complete series of possible conformations is shown in Fig. 2-2.

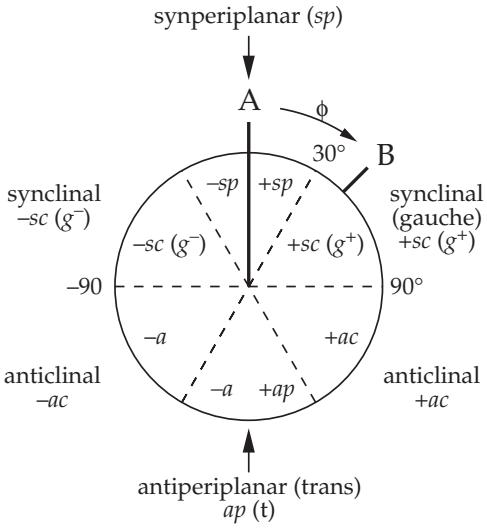
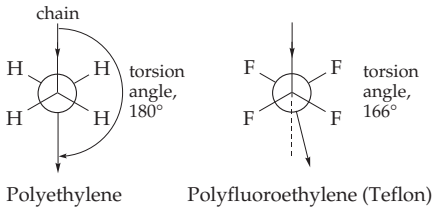



**Figure 2-2** Description of conformations about a single bond in the terminology of Klyne and Prelog[10,11] using the Newman projection. Group A is on the front atom at the top: the conformation is given for each possible position of group B on the other atom.

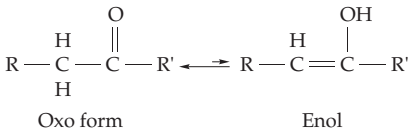
In the chain of methylene units, the hydrogen atoms on alternate carbon atoms of the fully extended chain barely touch (Fig. 2-1) but larger atoms cannot be accommodated. Thus, when fluorine atoms of van der Waals radius 0.135 nm replace the hydrogen atoms of radius 0.12 nm, a fully extended chain is no longer possible. For this reason the torsion angle in polyfluoroethylene is changed from the 180° of polyethylene to 166°, enough to relieve the congestion but not enough to cause severe eclipsing of the fluorines on adjacent carbons. The resulting **helical structure** is reminiscent of those occurring in proteins and other biopolymers. We see that helix formation can be a natural result of steric hindrance between groups of atoms.




Conformations of ring-containing molecules are dealt with in Chapter 4.

## 6. Tautomerism and Resonance

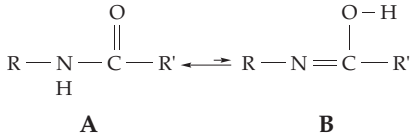
Many simple organic compounds exist as mixtures of two or more rapidly interconvertible isomers or **tautomeric forms**. Tautomers can sometimes be separated one from the other at low temperatures where the rate of interconversion is low. The classic example is the **oxo-enol** (or keto-enol) equilibrium (Eq. 2-1).



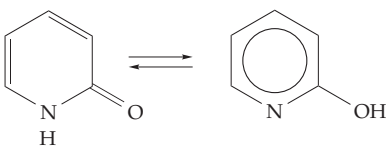

(2-1)

Although usually less stable than the oxo (keto) form, the enol is present in a small amount. It is formed readily from the oxo tautomer by virtue of the fact that hydrogen atoms attached to carbon atoms that are immediately adjacent to carbonyl (C=O) groups are remarkably acidic. Easy dissociation of a proton is a prerequisite for tautomerism. Since most hydrogen atoms bound to carbon atoms do not dissociate readily, tautomerism is unusual unless a carbonyl or other "activating group" is present.

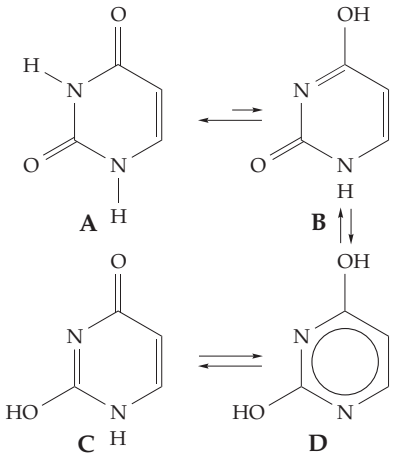
Since protons bound to oxygen and nitrogen atoms usually *do* dissociate readily, tautomerism also exists in amides and in ring systems containing O and N (Eqs. 2-2 to 2-5).



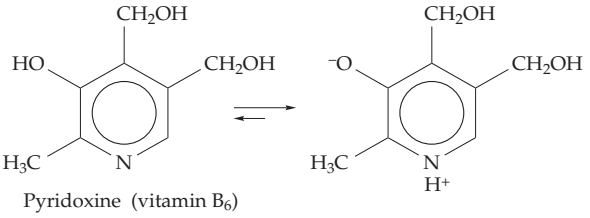

(2-2)




(2-3)




(2-4)



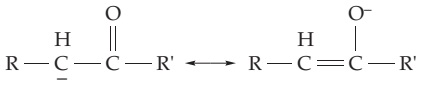

(2-5)

The tautomerism in Eq. 2-2 is the counterpart of that in the oxo-enol transformation. However, the equilibrium constant for aqueous conditions favors form A very strongly. 2-Pyridone is tautomerized to 2-hydroxypyridine (Eq. 2-3) to a greater extent. Pyrimidines (Eq. 2-4) and purines can form a variety of tautomers. The existence of form D of Eq. 2-4 is the basis for referring to uracil as dihydroxypyrimidine. However, the di-oxo tautomer A predominates. Pyridoxine (vitamin $B_6$) exists in water largely as the dipolar ionic tautomer B (Eq. 2-5) but in methanol as the uncharged tautomer A. In a pair of tautomers, a hydrogen atom always moves from one position to another and the lengths and bond character of these bonds also change.

The equilibrium constant for a tautomeric interconversion is simply the ratio of the mole fractions of the two forms; for example, the ratio of enol to oxo forms of acetone[12] in water at 25°C is $6.0 \times 10^{-9}$, while that for isobutyraldehyde is $1.3 \times 10^{-4}$. The ratio of 2-hydroxypyridine to 2-pyridone is about $10^{-3}$ in water but increases to 0.6 in a hydrocarbon solvent and to 2.5 in the vapor phase.[13,14] The ratio of dipolar ion to uncharged pyridoxine (Eq. 2-5) is ~4 at 25°C in water.[15] The ratios of tautomers B, C, and D to the tautomer A of uracil (Eq. 2-4) are small, but it is difficult to measure them quantitatively.[16] These tautomeric ratios are defined for given overall states of protonation (see Eq. 6-82). The constants are independent of pH but will change if the overall state of protonation of the molecule is changed. They may also be altered by

changes in temperature or solvent or by binding to a protein or other molecule.

It is important to distinguish tautomerism from **resonance**, a term used to indicate that the properties of a given molecule cannot be represented by a *single* valence structure but can be represented as a hybrid of two or more structures in which all the nuclei remain in the same places. Only bonding electrons move to convert one resonance form into another. Examples are the **enolate anion**, which can be thought of as a hybrid of structures A and B, and the amide linkage, which can be represented by a similar pair of resonance forms.

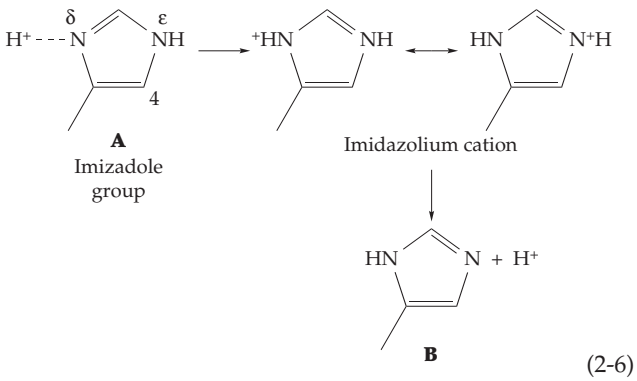


**A** **B**

Enolate anion

A double-headed arrow is often used to indicate that two structures drawn are resonance structures rather than tautomers or other separable isomers.

Although they are distinctly different, tautomerism and resonance are related. Thus, the acidity of carbon-bound hydrogens in ketones, which allows formation of enol tautomers, results from the fact that the enolate anion produced by dissociation of one of these hydrogens is stabilized by resonance. Similarly, tautomerism in the imidazole group of the amino acid histidine is related to resonance in the imidazolium cation. Because of this resonance, if a proton approaches structure A of Eq. 2-6 and becomes attached to the left-hand nitrogen atom ($N^{\delta}$), the positive charge in the resulting intermediate is distributed over both nitrogen atoms. This makes the proton on $N^{\varepsilon}$ acidic, permitting it to dissociate to tautomer B.



**A** **B** (2-6)

Imizadole group

Imidazolium cation

Note: The nitrogen atom designated $N^{\delta}$ (or ND1)[11] in Eq. 2-6 may also be called N1 or $N^{\pi}$ (*pros*-N). Likewise, $N^{\varepsilon}$ (NE2) may be designated N3 or $N^{\tau}$ (*tele*- N). Since $N^{\sigma}$ has sometimes also been called N3, it is best not to use the numerical designations for the nitrogen atoms.

The tautomeric ratio of B to A for histidine in water (Eq. 2-6) has been estimated, using $^{15}N$- and $^{13}C$-NMR, as 5.0 when the α-amino group is protonated and as 2.5 when at high pH it is unprotonated.[17] This tautomerism of the imidazole group is probably important to the function of many enzymes and other proteins; for example, if $N^{\varepsilon}$ of structure A (Eq. 2-6) is embedded in a protein, a proton approaching from the outside can induce the tautomerism shown with the release of a proton in the interior of the protein, perhaps at the active site of an enzyme. The form protonated on $N^{\delta}$ (B of Eq. 2-6), which is the minor form in solution, predominates in some positions within proteins.[18]

## B. Forces between Molecules and between Chemical Groups

The structure of living cells depends very much on the covalent bonds within individual molecules and on covalent crosslinks that sometimes form between molecules. However, weaker forces acting between molecules and between different parts of the same molecule are responsible for many of the most important properties of biochemical substances. These are described as **van der Waals forces**, **electrostatic forces**, **hydrogen bonds**, and **hydrophobic interactions**. In the discussion that follows the thermodynamic quantities $\Delta H$, $\Delta S$, and $\Delta G$ will be used. If necessary, please see Chapter 6 for definitions and a brief review.

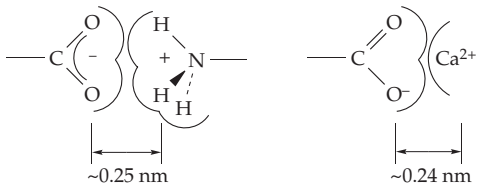
### 1. Van der Waals Forces

All atoms have a weak tendency to stick together, and because of this even helium liquifies at a low enough temperature. This is a result of the van der Waals or "London dispersion forces" which act strongly only at a very short distance. These forces arise from electrostatic attraction between the positively charged nucleus of one atom and the negatively charged electrons of the other.[19–21] Because nuclei are screened by the electron clouds surrounding them, the force is weak. The energy (enthalpy) of binding of one methylene ($-CH_2-$) unit into a monomolecular layer of a fatty acid is about $-\Delta H^{\circ} = 1.7$ kJ/mol.[22] Although this is a small quantity, when summed over the 16 or more carbon atoms of a typical fatty acid the binding energy is substantial. When a methylene group is completely surrounded in a crystalline hydrocarbon, its van der Waals energy, as estimated from the heat of sublimation, is 8.4 kJ/mol; that of $H_2O$ in liquid water at the melting point of ice is 15 kJ/mol.[22]

While van der Waals forces between individual atoms act over very short distances, they can be felt at surprisingly great distances when exerted by large molecules or molecular aggregates.[23] Forces between very smooth surfaces have been measured experimentally at distances as great as 10 nm and even to 300 nm.[23a] However, these "long-range van der Waals forces" probably depend upon layers of oriented water molecules on the plates[23] (see also Section 5).

## 2. Attraction between Charged Groups (Salt Linkages)

Fixed positive and negative charges attract each other strongly. Consider a carboxylate ion in contact with $-NH_3^+$ or with an ion of calcium:




From the van der Waals radii of Table 2-1 and the ionic crystal radius of $Ca^{2+}$ of 0.10 nm, we can estimate an approximate distance between the centers of positive and negative charge of 0.25 nm in both cases. It is of interest to apply Coulomb's law to compute the force $F$ between two charged particles which are almost in contact. Let us choose a distance of 0.40 nm (4.0 Å) and apply Eq. 2-7.

$$F = 8.9875 \times 10^9 \times \frac{qq'}{r^2\varepsilon} \text{ newtons} \qquad (2\text{-}7)$$

In this equation $r$ is the distance in meters, $q$ and $q'$ are the charges in coulombs (one electronic charge = $1.6021 \times 10^{-19}$ coulombs), ε is the dielectric constant, and $F$ is the force in newtons (N). The force per mole is $NF$ where $N$ is Avogadro's number.

An uncertainty in this kind of calculation is in the dielectric constant ε, which is 1.0 for a vacuum, about 2 for hydrocarbons, and 78.5 for water at 25°C. If ε is taken as 2, the force for $r$ = 0.40 nm is $4.3 \times 10^{14}$ N/mol. The force would be twice as great for the $Ca^{2+} - COO^-$ case. To move two single charges further apart by just 0.01 nm would require 4.3 kJ/mol, a substantial amount of energy. However, if the dielectric constant were that of water, this would be reduced almost 40-fold and the electrostatic force would not be highly significant in binding. It is extremely difficult to assign a dielectric constant for use in the interior of proteins.[23b] For charges spaced far apart within proteins the effective dielectric constant is usually as high as 30–60.[24] For closely spaced charges in hydrophobic niches it may be as low as 2–4.[25–27]

A calculation that is often made is the work required to remove completely two charges from a given distance apart (e.g., 0.40 nm) to an infinite distance (Eq. 2-8).

$$W \text{ (kJ mol}^{-1}) = 8.9875 \times 10^6 \times \frac{qq'}{r\varepsilon} \text{ N} = \frac{138.9}{\varepsilon r \text{ (in nm)}} \qquad (2\text{-}8)$$

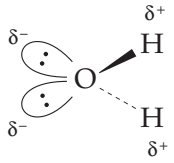
If ε = 2, this amounts to 174 kJ/mol for single charges at a distance of 0.40 nm; 69 kJ/mol at 1 nm; and only 6.9 kJ/mol at 10 nm, the distance across a cell membrane. We see that very large forces exist between closely spaced charges.

Electrostatic forces are of great significance in interactions between molecules and in the induction of changes in conformations of molecules. For example, attraction between $-COO^-$ and $-NH_3^+$ groups occurs in interactions between proteins. Calcium ions often interact with carboxylate groups, the doubly charged $Ca^{2+}$ bridging between two carboxylate or other polar groups. This occurs in carbohydrates such as agarose, converting solutions of these molecules into rigid gels (Chapter 4).

Individual macromolecules as well as cell surfaces usually carry a net negative charge at neutral pH. This causes the surfaces to repel each other. However, at a certain distance of separation the van der Waals attractive forces will balance the electrostatic repulsion.[21] Protruding hydrophobic groups may then interact and may "tether" bacteria or other particles at a fixed distance, often ~5 nm, from a cell surface.[28]

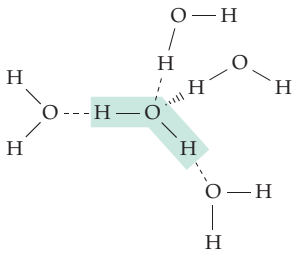
## 3. Hydrogen Bonds

One of the most important weak interactions between biologically important molecules is the hydrogen bond (H-bond).[29,30,30a] These "bonds" are the result of electrostatic attraction caused by the uneven distribution of electrons within covalent bonds. For example, the bonding electron pairs of the H–O bonds of water molecules are attracted more tightly to the oxygen atoms than to the hydrogen atoms. A small net positive charge is left on the hydrogen and a small net negative charge on the oxygen. Such **polarization** of the water molecules can be indicated in the following way:



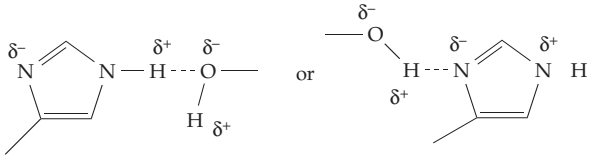
Here the $\delta^+$ and $\delta^-$ indicate a fraction of a full charge present on the hydrogen atoms and on the nonbonded electron pairs of the oxygen atom, respectively. Molecules such as $H_2O$, with strongly polarized bonds, are referred to as **polar molecules** and functional groups with such bonds as **polar groups**. They are to be contrasted with such nonpolar groups as the $-CH_3$ group in which the electrons in the bonds are nearly equally shared by carbon and hydrogen.

A hydrogen bond is formed when the positively charged end of one of the **dipoles** (polarized bonds) is attracted to the negative end of another dipole. Water molecules tend to hydrogen bond strongly one to another; each oxygen atom can be hydrogen-bonded to two other molecules and each hydrogen to another water molecule. Thus, every water molecule can have up to four hydrogen-bonded neighbors.



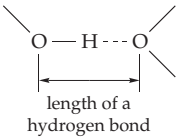
A water molecule hydrogen bonded to four other water molecules; note the tetrahedral arrangement of bonds around the central oxygen.

Many groups in proteins, carbohydrates, and nucleic acids form hydrogen bonds to one another and to surrounding water molecules. For example, an imidazole group of a protein can bond to an OH group of an amino acid side chain or of water in the following ways:



Remember that hydrogen bonds are always formed between pairs of groups, with one of them, often C=O or C=N-, containing the negative end of a dipole and the other providing the proton. The proton acceptor group, often OH or NH and occasionally SH, and even CH in certain structures,[31,31a,31b] donates an unshared pair of electrons. Dashed arrows are sometimes drawn from the hydrogen atom to the electron donor atom to indicate the direction of a hydrogen bond. Do not confuse these arrows with the curved arrows that indicate flow of electrons in organic reactions.

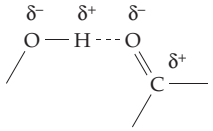
The strength of hydrogen bonds, as measured by the bond energy, varies over the range 10–40 kJ/mol. The stronger the hydrogen bond the shorter its length. Because hydrogen atoms can usually not be seen in X-ray structures of macromolecules, the lengths of hydrogen bonds are often measured between the surrounding heavy atoms:



length of a hydrogen bond

A typical —OH···O hydrogen bond will have a length of about 0.31 nm; a very strong hydrogen bond may be less than 0.28 nm in length, while weak hydrogen bonds will approach 0.36 nm, which is the sum of the van der Waals contact distances plus the O–H bond length. Beyond this distance a hydrogen bond cannot be distinguished easily from a van der Waals contact.

Hydrogen bonds are strongest when the hydrogen atom and the two heavy atoms to which it is bonded are in a straight line. For this reason hydrogen bonds tend to be linear. However, *the dipoles forming the hydrogen bond do not have to be colinear for strong hydrogen bonding:*[32,32a] There is some preference for hydrogen bonding to occur in the direction of an unshared electron pair on the oxygen or nitrogen atom.[33–35]



A linear O–H- - -O hydrogen bond with dipoles at an angle one to another.

Both ammonia, $NH_3$, and the $-NH_2$ groups of proteins are good electron donors for hydrogen bond formation. However, the hydrogen atoms of uncharged $-NH_2$ groups tend to be poor proton donors for H-bonds.[36] Do hydrogen bonds have some covalent character? The answer is controversial.[36a,36b,36c]

Hydrogen bonding is important both to the internal structure of biological macromolecules and in interactions between molecules. Hydrogen bonding often provides the specificity necessary to bring surfaces together in a complementary way. Thus, the location of hydrogen-bond forming groups in surfaces between molecules is important in ensuring an exact alignment of the surfaces.[37] The hydrogen bonds do not always have to be strong. For example, Fersht and coworkers, who compared a variety of mutants of an enzyme of known three-dimensional structure, found that deletion of a side chain that formed a good hydrogen bond to the substrate weakened the binding energy by only 2–6 kJ/mol. However, loss of a hydrogen bond to a charged group in the substrate caused a loss of 15–20 kJ/mol of binding energy.[37] Study of mutant

proteins created by genetic engineering is now an important tool for experimentally investigating the biological roles of hydrogen bonding.[37–39]

## 4. The Structure and Properties of Water

Water is the major constituent of cells and a remarkable solvent whose chemical and physical properties affect almost every aspect of life. Many of these properties are a direct reflection of the fact that most water molecules are in contact with their neighbors entirely through hydrogen bonds.[40–48] Water is the only known substance for which this is true.

In ordinary ice all of the water molecules are connected by hydrogen bonds, six molecules forming a hexagonal ring resembling that of cyclohexane. The structure is extended in all directions by the formation of additional hydrogen bonds to adjacent molecules (Fig. 2-3). As can be seen in this drawing, the molecules in ice assume various orientations in the hexagonal array, and frequently rotate to form their hydrogen bonds in different ways. This randomness remains as the temperature is lowered, and ice is one of few substances with a residual entropy at absolute zero.[49,50] Ice is unusual also in that the molecules do not assume closest packing in the crystal but form an open structure. The hole through the middle of the hexagon and on through the hexagons lying below it is ~0.06 nm in diameter.

The short hydrogen-bond length (averaging 0.276 nm) in ice indicates of strong bonding. The heat of sublimation ($\Delta H°$) of ice is −48.6 kJ/mol. If the van der Waals dispersion energy of −15 kJ/mol is subtracted from this, the difference of −34 kJ/mol can be attributed entirely to the hydrogen bonds—two for each molecule. Their average energy is 17 kJ/mol apiece. However, some of the hydrogen bonds are stronger and others weaker than the average.[51]

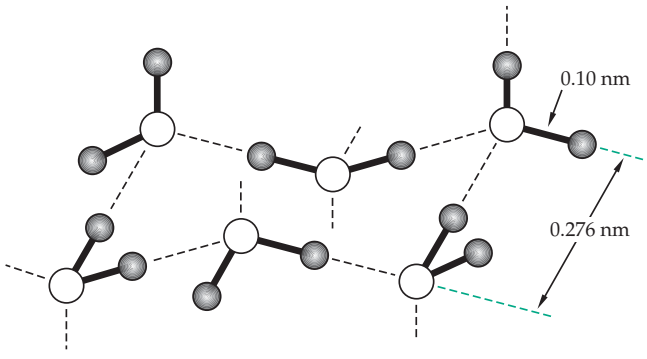
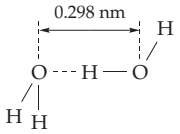
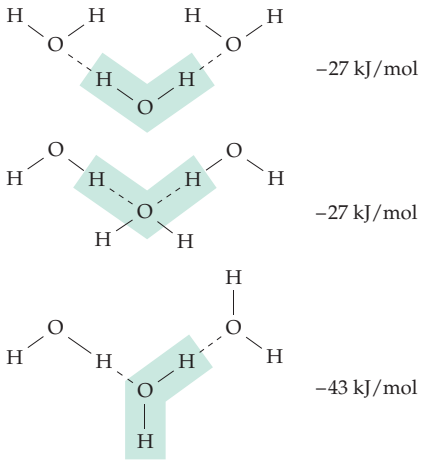



**Figure 2-3** Six water molecules in the lattice of an ice crystal. The hydrogen bonds, which connect protons with electron pairs of adjacent molecules, are shown as dashed lines.

In a gaseous water dimer the hydrogen bond is linear, a fact that suggests some covalent character.[52]




Its length is distinctly greater than that in ice. This is one of a number of pieces of evidence suggesting cooperativity in formation of chains of hydrogen bonds.[40,53,54] Consider the following three trimers for which theoretical calculations have predicted the indicated hydrogen bond energies.[40] In the first case the central water molecule *donates* two protons for hydrogen-bond formation; in the second it *accepts* the protons. In the third case it is both an electron acceptor and a donor. The OH dipoles are oriented "head to tail" and the hydrogen bonds are stronger than in the other




cases. Long chains of similarly oriented hydrogen bonds exist in ice and this may account for the short hydrogen bond lengths. Closed rings of hydrogen bonds oriented to give a maximum cooperative effect also exist in liquid water clusters[55] and within proteins, carbohydrates, and nucleic acids.[53,54]

The nature of liquid water is still incompletely understood,[46-48,56] but we know that water contains ice-like clusters of molecules that are continually breaking up and reforming in what has been called a "flickering cluster" structure. Judging by the infrared spectrum of water, about 10% of the hydrogen bonds are broken when ice melts.[41] A similar conclusion can be drawn from the fact that the heat of melting of ice is −5.9 kJ/mol. It has been estimated that at 0°C the average cluster contains about 500 water molecules.[41] At 50°C there are over 100 and at the boiling point about 40. Although most molecules in liquid water are present in these clusters, the hydrogen bonds are rapidly broken and reformed in new ways, with the average lifetime of a given hydrogen bond being ~$10^{-12}$ s.

## 5. Hydration of Polar Molecules and Ions

Water molecules are able to hydrogen bond not only to each other but also to polar groups of dissolved compounds. Thus, every group that is capable of forming a hydrogen bond to another organic group is also able to form hydrogen bonds of a somewhat similar strength with water. For this reason, hydrogen bonding is usually not a significant force in holding small molecules together in aqueous solutions. Polar molecules that stick together through hydrogen bonding when dissolved in a nonpolar solvent often do not associate in water. How then can biochemists assert that hydrogen bonding is so important in biochemistry? Part of the answer is that proteins and nucleic acids can be either properly folded with hydrogen bonds formed internally or denatured with hydrogen bonds from those same groups to water. The Gibbs energy change between these two states is small.

Every ion in an aqueous solution is surrounded by a shell of oriented water molecules held by the attraction of the water dipoles to the charged ion. The hydration of ions has a strong influence on all aspects of electrostatic interactions and plays a dominant role in determining such matters as the strength of acids and bases, the Gibbs energy of hydrolysis of ATP, and the strength of bonding of metal ions to negatively charged groups. For example, the previously considered interaction between carboxylate and calcium ions would be much weaker if both ions retained their hydration shells.

Consider the following example. $\Delta G°$ for dissociation of acetic acid in water is +27.2 kJ/mol (Table 6-5). The enthalpy change $\Delta H°$ for this process is almost zero (–0.1 kJ/mol) and $\Delta S°$ is consequently –91.6 J $K^{-1}$. This large entropy decrease reflects the increased amount of water that is immobilized in the hydration spheres of the $H^+$ and $acetate^-$ ions formed in the dissociation reaction. In contrast, dissociation of $NH_4^+$ to $NH_3$ and $H^+$ converts one positive ion to another. $\Delta H°$ is large (+52.5 kJ/mol) but the entropy change $\Delta S°$ is small (–2.0 J $K^{-1}$, Table 6-5).

Although effects of hydration are important in almost all biochemical equilibria, they are difficult to assess quantitatively. It is hard to know how many molecules of water are freed or immobilized in a given reaction. Charged groups in proteins are often hydrated. However, if they are buried in the interior of the protein, they may be solvated by polarizable protein side chain groups such as –OH or by backbone or side chain amide groups.[57,58]

## 6. Hydrophobic Interactions

Fats, hydrocarbons, and other materials whose molecules consist largely of nonpolar groups have a low solubility in water and a high solubility in nonpolar solvents. Similarly, the long alkyl groups of fatty acid esters aggregate within membranes and nonpolar side chains of proteins are often packed together in the centers of protein molecules. Because it is as if the nonpolar groups "fear" water, this is known as the **hydrophobic** effect.[59–70] The terms hydrophobic forces, hydrophobic interactions, and hydrophobic bonding have also been used. However, the latter term can be misleading because the hydrophobic effect arises not out of any special attraction between nonpolar groups but primarily from the strong internal cohesion of the hydrogen-bonded water structure.

How strongly do nonpolar groups "attract" each other in water? A partial answer can be obtained by measuring the standard Gibbs energy $\Delta G$ of transfer of a hydrocarbon molecule from a dilute aqueous solution into a dilute solution in another hydrocarbon. By studying a series of alkanes, Abraham[62] calculated the Gibbs energy change per $CH_2$ unit (Eq. 2-9) as: $\Delta G° = -3.8$ kJ/mol.

$$-CH_2-(\text{water}) \longrightarrow -CH_2-(\text{hexane}) \qquad \Delta G° = -3.8 \text{ kJ/mol} \qquad (2\text{-}9)$$

This equation is a quantitative statement of the fact that the $CH_2$ group prefers to be in a nonpolar environment than to be surrounded by water. A similar Gibbs energy change would be expected to accompany the bringing together of a methylene unit from a small molecule and a hydrophobic surface on a protein molecule. However, in the latter case the accompanying decrease in entropy would make $\Delta G°$ less negative.

What causes the decrease in Gibbs energy when nonpolar groups associate in water? Jencks[60] suggested that we think of the transfer of a nonpolar molecule from a nonpolar solvent into water in two steps: (1) Create a cavity in the water of about the right size to accommodate the molecule. Since many hydrogen bonds will be broken, the Gibbs energy of cavity formation will be high. It will be principally an enthalpy ($\Delta H$) effect. (2) Allow the water molecules in the solvent to make changes in their orientations to accommodate the nonpolar molecule that has been placed in the cavity. The water molecules can move to give good van der Waals contacts and also reorient themselves to give the maximum number of hydrogen bonds. Since hydrogen bonds can be formed in many different ways in water, there may be as many or even more hydrogen bonds after the reorientation than before. This will be true especially at low temperature where most water exists as large icelike clusters. For dissolved hydrocarbons, the enthalpy of formation of the new hydrogen bonds often almost exactly balances the enthalpy of creation of the cavity initially so that $\Delta H$ for the overall process (transfer from inert solvent into water) is small. For the opposite transfer (from water to hydrocarbon hexane; Eq. 2-9) $\Delta H°$ is usually a small

positive number for aliphatic hydrocarbons and is nearly zero for aromatic hydrocarbons.

Since $\Delta G° = \Delta H° - T\Delta S°$ (see Chapter 6), it follows that the negative value of $\Delta G°$ for hydrophobic interactions must result from a positive entropy change, which may arise from the restricted mobility of water molecules that surround dissolved hydrophobic groups. When two hydrophobic groups come together to form a "hydrophobic bond," water molecules are freed from the structured region around the hydrophobic surfaces and the entropy increases. The $\Delta S°$ for Eq. 2-9 is about 12 J $deg^{-1}$ $mol^{-1}$. Attempts have been made to relate this value directly to the increased number of orientations possible for a water molecule when it is freed from the structured region.[64] However, interpretation of the hydrophobic effect is complex and controversial.[65–71a]

The formation constant $K_f$ for hydrophobic associations often increases with increasing temperature. This is in contrast to the behavior of $K_f$ for many association reactions that involve polar molecules and for which $\Delta H°$ is often strongly negative (heat is released). An example of the latter is the protonation of ammonia in an aqueous solution (Eq. 2-10).

$$NH_3 + H^+ \rightarrow NH_4^+ \qquad \Delta H° = -52.5 \text{ kJ/mol} \qquad (2\text{-}10)$$

Since $R1nK_f = -\Delta G°/T = -\Delta H°/T + \Delta S°$, $K_f$ decreases with increasing temperature if $\Delta H°$ is negative. Because for a hydrophobic interaction with a positive value of $\Delta H°$ $K_f$ *increases* with increasing temperature, an increase in stability at higher temperatures is sometimes used as a criterion for hydrophobic bonding. However, this criterion does not always hold. For example, base stacking interactions in polynucleotides (Chapter 5), whose strength does not increase with increasing temperature, are still thought to be hydrophobic.

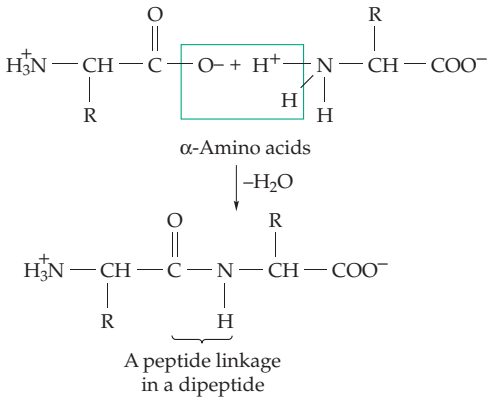
The water molecules that are in immediate contact with dissolved nonpolar groups are partially oriented. They form a cagelike structure around each hydrophobic group. When particles surrounded by such hydration layers are 1–2 nm apart, they sometimes experience either a fairly strong repulsion or an enhanced attraction caused by these hydration layers.[21,64–66,72] Direct experimental measurements have shown that these effects extend to distances of 10 nm[21,63] and can account for the previously mentioned long-range van der Waals forces.

Various efforts have been made to develop scales of **hydrophobicity** that can be used to predict the probability of finding a given amino acid side chain buried within a protein or in a surface facing water.[59,73] A new approach has been provided by the study of mutant proteins. For example, deletion of a single $-CH_2-$ group from an interior hydrophobic region of a protein was observed to decrease the stability of the protein by 4.6 kJ/mol.[74]

## C. Amino Acids and Peptides

Twenty α-amino acids are the monomers from which proteins are made. These amino acids share with other biochemical monomers a property essential to their role in polymer formation: *They contain at least two different chemical groups able to react with each other to form a covalent linkage*. In the amino acids these are the protonated amino ($NH_3^+$) and carboxylate ($COO^-$) groups. The characteristic linkage in the protein polymer is the **peptide** (amide) linkage whose formation can be imagined to occur by the splitting out of water between the carboxyl of one amino acid and the amino group of another (Eq. 2-11).

This equation is not intended to imply a mechanism for peptide synthesis. The equilibrium position for this reaction in an aqueous solution favors the free amino acids rather than the peptide. Therefore, both biological and laboratory syntheses of peptides usually do not involve a simple splitting out of water. Since the dipeptide of Eq. 2-11 still contains reactive carboxyl and amino groups, other amino acid units can be joined by additional peptide linkages to form **polypeptides**. These range from short-chain **oligomers** to polymers of from ~50 to several thousand amino acid units, the proteins.[75–77]




(2-11)

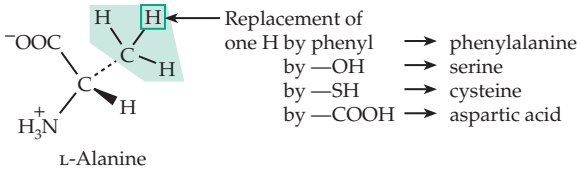
### 1. Properties of α-Amino Acids

The amino acids have in common a dipolar ionic structure and a chiral center. They are differentiated, one from another, by the structures of their **side chain groups**, designated R in the foregoing formulas. These groups are of varying size and chemical structure. The side chain groups fill much of the space in the interior of a protein molecule and also protrude from the external surfaces of the protein where they determine many of the chemical and physical properties of the molecule.

Table 2-2 shows the structures of the side chains

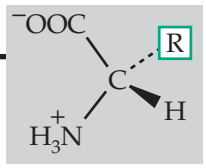
of the amino acids commonly found in proteins. The complete structure is given for proline. Both the three-letter abbreviations and one-letter abbreviations used in describing sequences of amino acids in proteins are also given in this table. Amino acids of groups **a**–**c** of Table 2-2 plus phenylalanine and methionine are sometimes grouped together as *nonpolar*. They tend to be found in a hydrophobic environment on the inside of a protein molecule. Groups **f** and **i** contain *polar*, *charged* side chains which usually protrude into the water surrounding the protein. The rest are classified as *polar* but *noncharged*.

To get acquainted with amino acid structures, learn first those of **glycine**, **alanine**, **serine**, **aspartic acid**, and **glutamic acid**. The structures of many other amino acids can be related to that of alanine (R=$CH_3$) by replacement of a β hydrogen by another group. Metabolic interrelationships will make it easier to learn structures of the rest of the amino acids later.




Since the –COOH groups of glutamic and aspartic acids are completely dissociated to –$COO^-$ at neutral pH, it is customary in the biochemical literature to refer to these amino acids as **glutamate** and **aspartate** without reference to the nature of the cation or cations present as counter ions. Such "-ate" endings are also used for most other acids (e.g., malate, oxaloacetate,

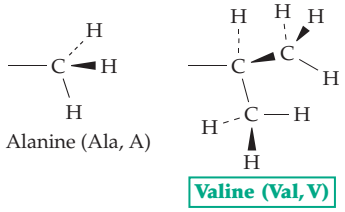
## TABLE 2-2
## Structure and Chemical Properies of Side Chain Groups (R) of Amino Acids



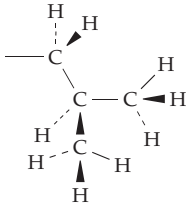
**a.** Glycine "side chain" = —H (Gly, G)[a]

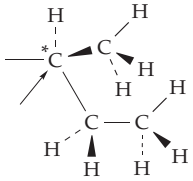
Strictly a link in the peptide chain, glycine provides a minimum of steric hindrance to rotation and to placement of adjacent groups.

**b.** Amino acids with alkyl groups as side chains


Alanine (Ala, A)

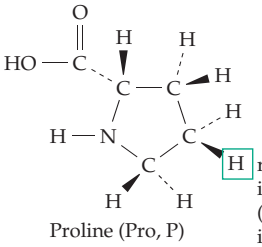

Valine (Val, V)


Leucine (Leu, L)


Isoleucine (Ile, I)

These bulky groups of distinctive shapes participate in hydrophobic interactions in protein interiors and in forming binding sites of specific shapes.

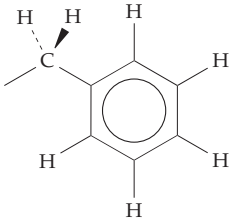
**c.** The imino acid proline. Because the side chain is fused to the α-amino group, the entire structure, not just the side chain, is shown.

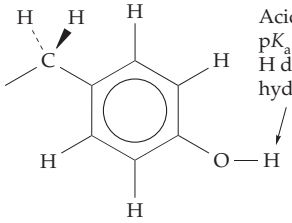

Proline (Pro, P)

H replaced by —OH in hydroxyproline (Hyp) which occurs in collagen

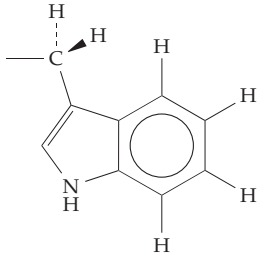
Note the secondary amino group and the relatively rigid conformation. The presence of proline strongly influences the folding of protein chains.

**d.** Aromatic amino acids


Phenylalanine (Phe, F)



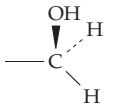
Tyrosine (Tyr, Y)

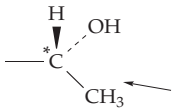

Tryptophan (Trp, W)

Tyrosine, phenylalanine, and tryptophan can form hydrophobic bonds and may be especially effective in bonding to other flat molecules.

**TABLE 2-2**
*(continued)*

**e.** Amino acid alcohols

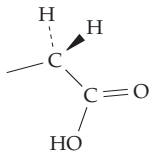

Serine (Ser, S)

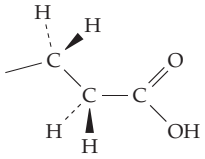

**Threonine (Thr, T)**

Note the second chiral center. The L-amino acid with the opposite configuration in the side chain is L-*allo*-threonine.

The —OH group is very weakly acidic (p$K_a$ ~13.6). It can form esters with phosphoric acid or organic acids and is a site of attachment of sugar rings in glycoproteins. Hydroxyl groups of serine are found at the active centers of some enzymes.

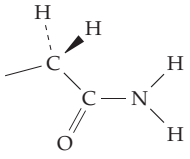
**f.** Acidic amino acids

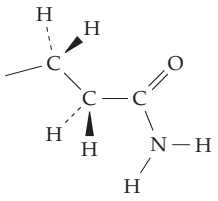

Aspartic acid (Asp, D)


Glutamic acid (Glu, E)

Carboxyl groups of these side chains are dissociated at neutral pH (p$K_a$ values are 4.3–4.7) and provide anionic (–) groups on the surfaces of proteins.

**g.** Amides of aspartic acid and glutamic acid

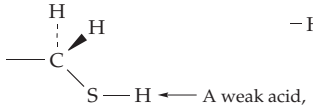

Asparagine (Asn, N)


Glutamine (Gln, Q)

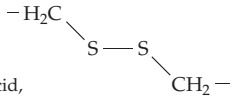
The amide group is not acidic but is polar and participates in hydrogen bonding.

If it is uncertain whether a position in a protein is occupied by aspartic acid or asparagine, it may be designated Asx or B. If glutamic or glutamine, it may be designated as Glx or Z.

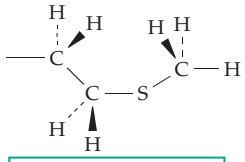
**h.** Sulfur-containing amino acids


Cysteine (Cys, C)

A weak acid, p$K_a$ ~8.5



Two cysteine SH groups can be oxidatively joined to form a disulfide bridge in the "double-headed" amino acid cystine.


**Methionine (Met, M)**

**Key**

**Essential in the human diet**

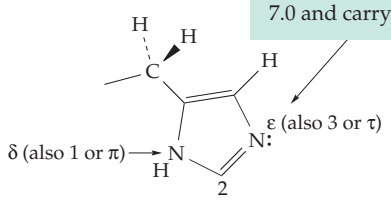
**Essential if phenylalanine or cysteine is inadequate**

**Essential for rapid growth**

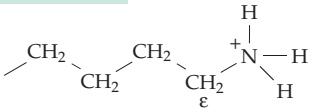
[a] The three-letter and one-letter abbreviations used for the amino acid residues in peptides and proteins are given in parentheses. B, J, U, X, and Z can be used to indicate modified or unusual amino acids.

**i.** Basic amino acids

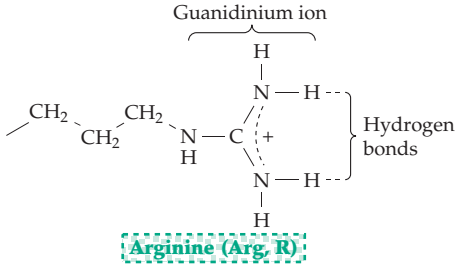
This basic site accepts a proton to form a conjugate acid of p$K_a$ ~6.4–7.0 and carrying a positive charge.


**Histidine (His, H)**

The **imidazole** groups in histidine side chains are parts of the active sites of many enzymes. Like other basic groups in proteins they also may bind metal ions.

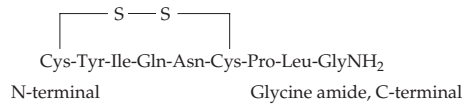

**Lysine (Lys, K)**

A flexible side arm with a potentially reactive amino group at the end. The high p$K_a$ of ~10.5 means that lysine side chains are ordinarily protonated in neutral solutions.

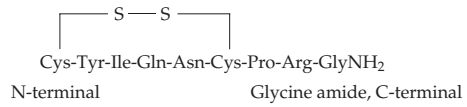


**Arginine (Arg, R)**

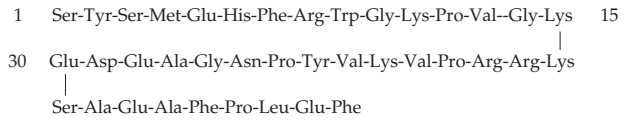
The guanidinium group has a high p$K_a$ of over 12 and remains protonated under most circumstances. It is stabilized by resonance as indicated by the dashed line. Guanidinium groups are often sites for binding of phosphate or carboxylate groups by pairs of hydrogen bonds.



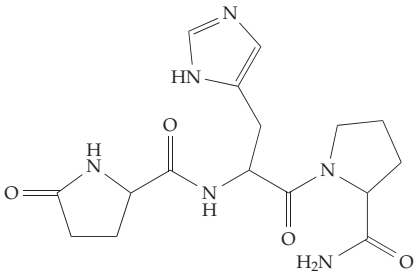

Human oxytocin



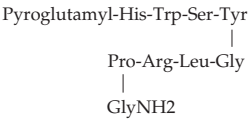

Human vasopressin



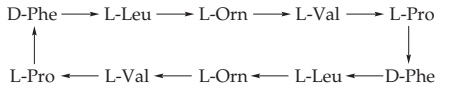

Human adrenocorticotropin (ACTH)



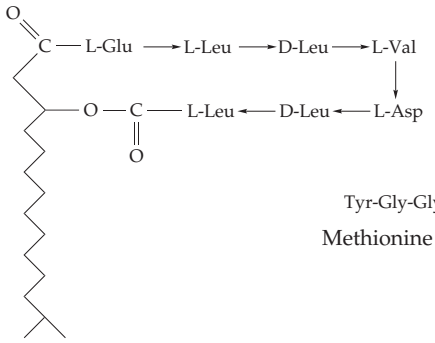
Hypothalamic thyrotropic hormone releasing factor (TRF); pyroglutamyl-histidyl-prolinamide



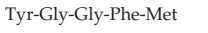

Luteinizing hormone releasing factor (LRF)




Gramicidin S




Surfactin




Methionine enkephalin

**Figure 2-4** Structures of some naturally occurring peptides. Oxytocin and vasopressin are hormones of the neurohypophysis (posterior lobe of the pituitary gland). Adrenocorticotropin is a hormone of the adenohypophysis (anterior pituitary). Hormones of the adenohypophysis are released under the influence of releasing factors (regulatory factors) produced in the neighboring hypothalamus (a portion of the brain) in response to neural stimulation. Structures of two releasing factors are shown. Note that the γ-carboxyl groups of the N-terminal glutamine residues have reacted, with loss of $NH_3$, with the neighboring terminal $-NH_2$ groups to form cyclic amide (pyroglutamyl) groups.[78] Gramicidin S is an antibiotic made by *Bacillus brevis*, and surfactin is a depsipeptide (containing an ester linkage), a surface active antibiotic of *Bacillus subtilis*. Methionine enkephalin is a brain peptide with opiate-like activity.[79]

phosphate, and adenylate) and in names of enzymes (e.g., lactate dehydrogenase).

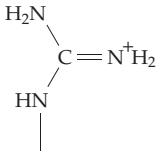
During the formation of polypeptides, the α-amino and carboxyl groups of the amino acids are converted into the relatively unreactive and uncharged amide (peptide) groups except at the two chain termini. In many cases the terminal amino and carboxyl groups are also converted within cells into uncharged groups (Chapter 10). Immediately after the protein is synthesized its terminal carboxyl group is often converted into an amide. The N terminus may be acetylated or cyclized to a pyroglutamyl group. Sometimes a cyclic peptide is formed (Fig. 2-4).

The properties of polypeptides and proteins are determined to a large extent by the chemistry of the side chain groups, which may be summarized briefly as follows. Glycine in a peptide permits a maximum of conformational mobility. The nine relatively nonpolar amino acids–alanine, valine, leucine, isoleucine, proline, methionine, phenylalanine, tyrosine, and tryptophan–serve as building blocks of characteristic shape. Tyrosine and tryptophan also participate in hydrogen bonding and in aromatic: aromatic interactions within proteins.

Much of the chemistry of proteins involves the side chain functional groups –OH, –SH, $-COO^-$, $-NH_3^+$, and imidazole (Eq. 2-6) and the guanidinium group of arginine. The side chains of asparagine and glutamine both contain the amide group $CONH_2$, which is relatively inert chemically but which can undergo hydrogen-bonding interactions. The amide linkages of the polypeptide backbone must also be regarded as important functional groups. Most polar groups are found on the outside surfaces of proteins where they can react chemically in various ways. When inside proteins they form H-bonds to the peptide backbone and to other polar groups.

## 2. Acidic and Basic Side Chains

The side chains of aspartic and glutamic acids carry negatively charged carboxylate groups at pH 7 while those of lysine and arginine carry the positively charged $-NH_3^+$ and guanidinium ions, respectively.



Guanidinium group

At pH 7 the weakly basic imidazole group of histidine may be partially protonated. Both the –SH group of cysteine and the phenolic –OH of tyrosine are weakly acidic and will dissociate and thereby acquire negative charges at a sufficiently high pH.

The number of positive and negative charges on a protein at any pH can be estimated approximately from the acid dissociation constants (usually given as $pK_a$ values) for the amino acid side chains. These are given in Table 2-2. However, $pK_a$ values of buried groups are often greatly shifted from these, especially if they associate as **ion pairs**. In addition, many proteins have free amino and carboxyl-terminal groups at the opposite ends of the peptide chain. These also participate in acid–base reactions with approximately the following $pK_a$ values.

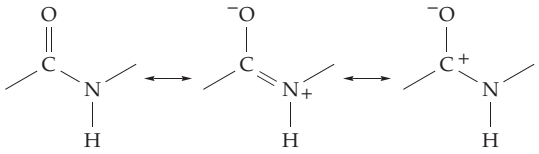
terminal —COOH, $pK_a = 3.6–3.7$
terminal $—NH_3^+$, $pK_a = 7.5–7.9$

The acid–base properties of an amino acid or of a protein are described by titration curves of the type shown in Figs. 3-1 and 3-2. In these curves the number of equivalents of acid or base that have reacted with an amino acid or protein that was initially at neutral pH are plotted against pH. The net negative or positive electrical charge on the molecule can be read directly from the curves. Both the net electrical charges and the distribution of positively and negatively charged groups are often of crucial importance to the functioning of a protein.

Additional aspects of the acid–base chemistry of amino acids and proteins are considered in Chapter 3, Section A and Chapter 6, Section E. The student may find it appropriate to study these sections at this time and to work the associated study problems.

## 3. The Peptide Unit

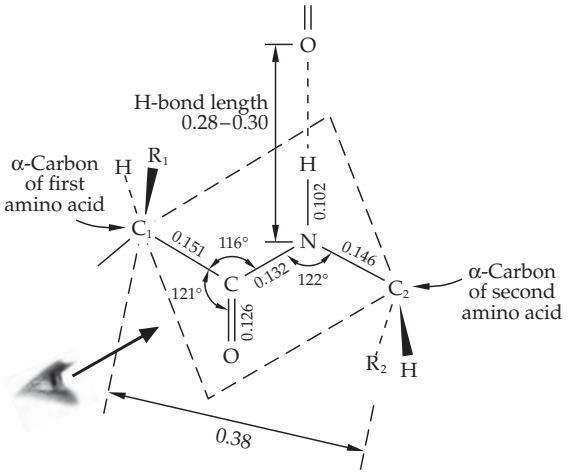
The very ability of a protein to exist as a complex three-dimensional structure depends upon the properties of the amide linkages between the amino acid units. Many of these properties follow from the fact that an amide can be viewed as a resonance hybrid of the following structures. Because of the partial double-bond character, the C–N bond is shorter than that of a normal single bond and the C=O bond is lengthened.



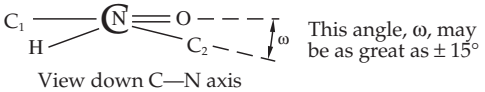
The observed lengths in nanometers determined by X-ray diffraction measurements are given in Fig. 2-5 (top). The partial double-bond character of the C–N bond has important consequences. The peptide unit is nearly planar as is indicated by the dashed parallelogram in Fig. 2-5.

However, the bonds around the nitrogen retain some pyramidal character (Fig. 2-5, bottom). Even more important is the fact that there is flexibility. As a result, the torsion angle ω may vary over a range of ± 15° or even more from that in the planar state.[80,81,81a]

The resonance stabilization of the amide linkage is thought to be about 85 kJ/mol. Rotation around the C–N bond through 90° would be expected to require



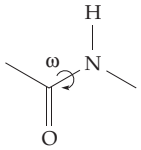

Distance (nm) between successive α-carbon atoms in a protein.



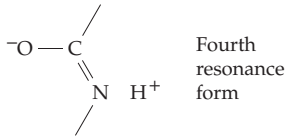

View down C—N axis
(as indicated by heavy arrow above)

**Figure 2-5** Dimensions of the peptide linkage. Interatomic distances in nm, including the hydrogen bond length to an adjacent peptide linkage, are indicated. The atoms enclosed by the dotted lines all lie *approximately* in a plane. However, as indicated in the lower drawing, the nitrogen atom tends to retain some pyramidal character.

about this much energy. This fact immediately suggests a way in which proteins may sometimes be able to store energy—by having one or more peptide units twisted out of complete planarity.[69]

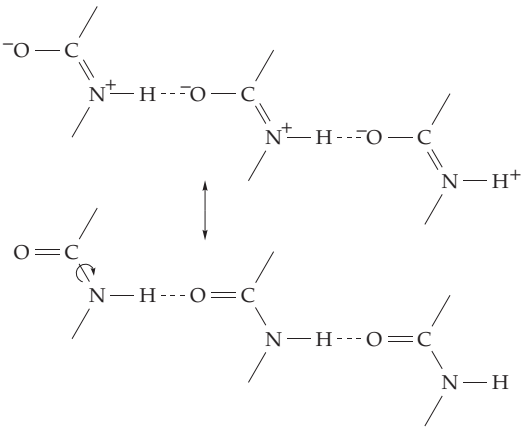


An important effect of the resonance of the amide linkage is that the oxygen atom acquires some negative charge and the NH group some positive charge. Some of the positive charge is usually depicted as residing on the nitrogen, but some is found on the hydrogen atom. The latter can be pictured as arising from a contribution of a fourth resonance form that contains no bond to hydrogen.



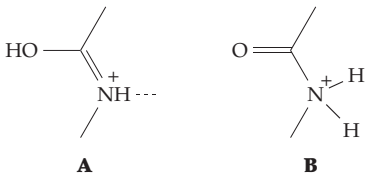

Nevertheless, this picture is inadequate. Various evidence indicates that the nitrogen actually carries a net negative charge.[81a]

The positive and negative ends of the dipoles in the amide group tend to associate to form strong hydrogen bonds. These hydrogen bonds together with the connecting amide linkages can form chains that may run for considerable distances through proteins. The tendency for cooperativity in hydrogen bond formation may impart unusual stability to these chains. As with individual amide linkages, these chains of hydrogen-bonded amides can also be thought of as resonance hybrids:

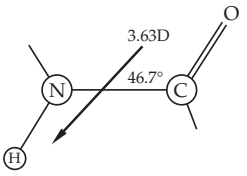


The two structures pictured are extreme forms, the true structure being something in between. In the lower form, rotation about the C–N bond would be permitted but then the charge separation present in the upper structure would no longer exist. Thus, the hydrogen bonds would be weakened. We can conclude that if an amide linkage in such a chain becomes twisted, the hydrogen bonds that it forms will be weakened. If there is cooperativity, the hydrogen bonds will all be strongest when there is good planarity in all of the amides in the chain.

Amides have very weak basic properties and protonation is possible either on the oxygen (A) or on the nitrogen (B).[82,83]

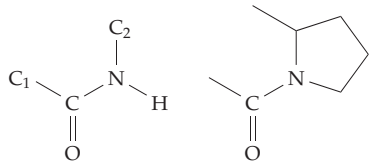



The p$K_a$ values for such protonation are usually less than zero, but it is possible that a correctly placed acidic group in a protein could protonate either oxygen or nitrogen transiently during the action of a protein. Protonation on oxygen would strengthen hydrogen bonds from the nitrogen whereas protonation on nitrogen would weaken hydrogen bonds to oxygen and might permit rotation. The amide group has a permanent dipole moment of 3.63 debyes oriented as follows:




Here the arrow points toward the *positive* end of the dipole.

While the *trans* peptide linkage shown in Fig. 2-4 is usual, the following *cis* peptide linkage, which is ~8 kJ/mol less stable than the *trans* linkage, also occurs in proteins quite often. The nitrogen atom is usually but not always from proline.[81,84]

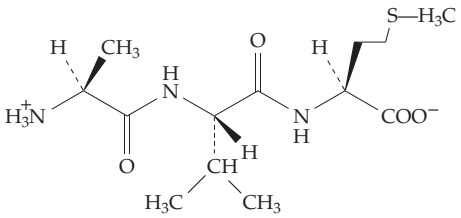


*cis* Peptide linkages

## 4. Polypeptides

The chain formed by polymerization of amino acid molecules provides the **primary structure** of a protein. Together with any covalent crosslinkages and other modifications, this may also be called the

**covalent structure** of the protein. Each monomer unit in the chain is known as an amino acid **residue**. This term acknowledges the fact that each amino acid has lost one molecule of $H_2O$ during polymerization. To be more precise, the number of water molecules lost is *one less* than the number of residues. Peptides are named according to the amino acid residues present and beginning with the one bearing the terminal amino group. Thus, L-alanyl-L-valyl-L-methionine has the following structure:




Like amino acids, this tripeptide is a dipolar ion. The same structure can be abbreviated Ala-Val-Met or, using one-letter abbreviations, AVM. It is customary in describing amino acid sequences to place the amino-terminal (N-terminal) residue at the left end and the carboxyl-terminal (C-terminal) residue at the right end. Residues are numbered sequentially with the N-terminal residue as 1. An example is shown in Fig. 2-6.

The sequence of amino acid units in a protein is always specified by a gene. The sequence determines how the polypeptide chain folds and how the folded protein functions. For this reason much effort has gone into "sequencing," the determination of the precise order of amino acid residues in a protein. Sequences of several hundreds of thousands of proteins and smaller peptides have been established and the number doubles each year.[75,87,88,88a] Most of these

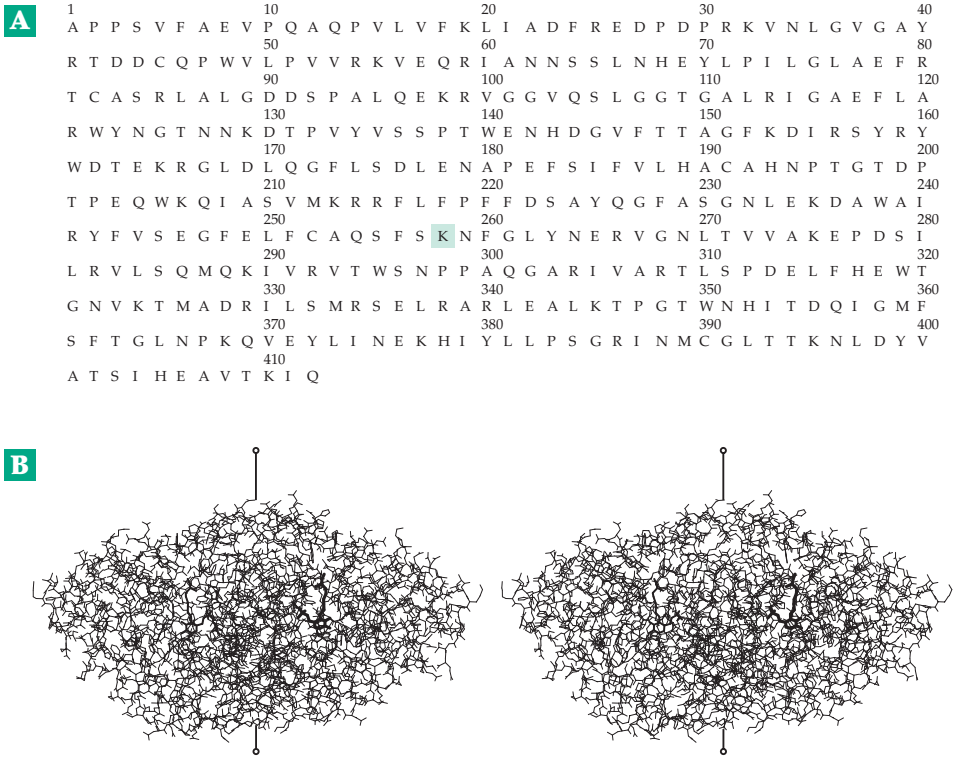


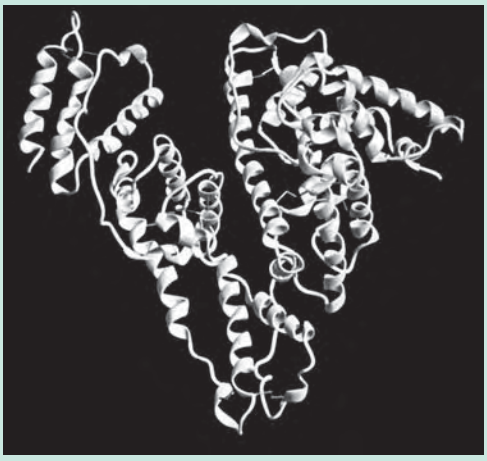

**Figure 2-6** (A) The complete amino acid sequence of the cytoplasmic enzyme aspartate aminotransferase from pig heart. The peptide has the composition $Lys_{19}$, $His_8$, $Arg_{26}$, $(Asp + Asn)_{42}$, $Ser_{26}$, $Thr_{26}$, $(Glu + Gln)_{41}$, $Pro_{24}$, $Gly_{28}$, $Ala_{32}$, $Cys_5$, $Val_{29}$, $Met_6$, $Ile_{19}$, $Leu_{38}$, $Tyr_{12}$, $Phe_{23}$, $Trp_9$. The molecular mass is 46.344 kDa and the complete enzyme is a 93.147-kDa dimer containing two molecules of the bound coenzyme pyridoxal phosphate attached to lysine-258 (enclosed in box).[85,86] (B) A stereoscopic view of a complete enzyme molecule which contains two identical subunits with the foregoing sequence. Coordinates from Arthur Arnone. In this "wire model" all the positions of all of the nearly 7000 atoms that are heavier than hydrogen are shown. The > 8000 hydrogen atoms have been omitted. The view is into the active site of the subunit on the right. The pyridoxal phosphate and the lysine residue to which it is attached are shown with heavy lines. The active site of the subunit to the left opens to the back side as viewed here. The drawing may be observed best with a magnifying viewer available from Abrams Instrument Corp., Lansing, Michigan or Luminos Photo Corp., Yonkers, New York. However, with a little practice, it is possible to obtain a stereoscopic view unaided. Hold the book with good illumination about 20–30 cm from your eyes. Allow your eyes to relax as if viewing a distant object. Of the four images that are visible, the two in the center can be fused to form the stereoscopic picture. Drawings by program MolScript (Kraulis, 1991).

## BOX 2-A PROTEINS OF BLOOD PLASMA

Among the most studied of all proteins are those present in blood plasma.[a–f] Their ready availability and the clinical significance of their study led to the early development of electrophoretic separations. Electrophoresis at a pH of 8.6 (in barbital buffer) indicated six main components. The major and one of the fastest moving proteins is **serum albumin**. Trailing behind it are the **$\alpha_1$-**, **$\alpha_2$-**, and **β-globulins**, **fibrinogen**, and **γ-globulins**. Each of these bands consists of several proteins and two-dimensional separation by electrophoresis and isoelectric focusing (Chapter 3) reveals over 30 different proteins.[e] Many of these contain varying numbers of attached carbohydrate units and appear as families of spots.

Fractionation of large quantities of plasma together with immunochemical assays has led to identification of over 200 different proteins. Sixty or more are enzymes, some in very small quantitites which may have leaked from body cells. Normally plasma contains 5.7 – 8.0 g of total protein per 100 ml (~1 mM). Albumin accounts for 3.5 – 4.5 g/100 ml. An individual's liver synthesizes about 12 g each day. Next most abundant are the **immunoglobulins**. One of these (IgG or γ-globulin) is present to the extent of 1.2–1.8 g/100 ml. Also present in amounts greater than 200 mg per 100 ml are **α-** and **β-lipoproteins**, the **$\alpha_1$ antitrypsin**, **$\alpha_2$-macroglobulin**, **haptoglobin**, **transferrin**, and fibrinogen.

Plasma proteins have many functions. One of them, fullfilled principally by serum albumin, is to impart enough osmotic pressure to plasma to match that of the cytoplasm of cells. The heart-shaped human serum albumin consists of a single 65 kDa chain of 585 amino acid residues coiled into 28 helices.[g] Three homologous repeat units or domains each contain six disulfide bridges, suggesting that gene duplication occurred twice during the evolution of serum albumins. The relatively low molecular mass and high density of negative charges on the surface make serum albumin well adapted for the role of maintaining osmotic pressure. However, serum albumin is not essential to life.



Human serum albumin. From He and Carter. [g]

Over 50 mutant forms have been found and at least 30 persons have been found with no serum albumin in their blood.[h,i] These analbuminemic individuals are healthy and have increased concentrations of other plasma proteins.

A second major function of plasma proteins is transport. Serum albumin binds to and carries many sparingly soluble metabolic products, including fatty acids, tryptophan, cysteine, steroids, thyroid hormones, $Ca^{2+}$, $Cu^{2+}$, $Zn^{2+}$, other metal ions, bilirubin, and various drugs. There are also many more specialized transporter proteins. **Transferrin** carries iron and **ceruloplasmin** (an $\alpha_2$ globulin) transports copper. **Transcortin** carries corticosteroids and progesterone, while another protein carries sex hormones. **Retinol-binding protein** carries vitamin A and **cobalamin-binding proteins** vitamin $B_{12}$. **Hemopexin** carries heme to the liver, where the iron can be recovered.[j] Haptoglobin binds hemoglobin released from broken red cells and also assists in the recycling of the iron in the heme.[k] **Lipoproteins** (see Table 21-1) carry phospholipids, neutral lipids, and cholesterol esters. Most of the mass of these substances is lipid.

Immunoglobulins, $\alpha_1$-trypsin inhibitor and **$\alpha_2$-macroglobulin**,[k] ten or more blood clotting factors; and proteins of the **complement system** all have protective functions that are discussed elsewhere in this book. Hormones, many of them proteins, are present in the blood as they are carried to their target tissues. Many serum proteins have unknown or poorly understood functions. Among these are the **acute phase proteins**, whose concentrations rise in response to inflammation or other injury.

---


[a] Allison, A. C., ed. (1974) *Structure and Function of Plasma Proteins*, Vol. 1, Plenum, New York

[b] Allison, A. C., ed. (1976) *Structure and Function of Plasma Proteins*, Vol. 2, Plenum, New York

[c] Putnam, F. W., ed. (1975) *The Plasma Proteins*, 2nd ed., Vol. 1 and 2, Academic Press, New York

[d] Blomback, B., and Hanson, L. A., eds. (1979) *Plasma Proteins*, John Wiley, Chichester, Oklahoma

[e] Geisow, M. J., and Gordon, A. H. (1978) *Trends Biochem. Sci.* **3**, 169–171

[f] Smith, E. L., Hill, R. L., Lehman, I. R., Lefkowitz, R. J., Handler, P., and White, A. (1983) in *Principles of Biochemistry, Mammalian Biochemistry*, 7th ed., McGraw-Hill, New York, pp. 3–37

[g] He, X. M., and Carter, D. C. (1992) *Nature* **358**, 209–215

[h] Peters, T., Jr. (1996) *All About Albumin: Biochemistry, Genetics and Medical Applications*, Academic Press, San Diego, California

[i] Madison, J., Galliano, M., Watkins, S., Minchiotti, L., Porta, F., Rossi, A., and Putnam, F. W. (1994) *Proc. Natl. Acad. Sci: USA* **91**, 6476–6480

[j] Satoh, T., Satoh, H., Iwahara, S.-I., Hrkal, Z., Peyton, D. H., and Muller-Eberhard, U. (1994) *Proc. Natl. Acad. Sci: U.S.A.* **91**, 8423–8427

[k] Feldman, S. R., Gorias, S. L., and Pizzo, S. V. (1985) *Proc. Natl. Acad. Sci: U.S.A.* **82**, 5700–5704

have been deduced from the sequences of nucleotides in DNA. Sequences of some small peptide hormones and antibiotics are shown in Fig. 2-4 and that of a 412-residue protein in Fig. 2-6. The molecular mass of a protein can be estimated from the chain length by assuming that each residue adds 100–115 Da.

The amino acid composition varies greatly among proteins. A typical globular protein contains all or most of the 20 amino acids. The majority are often present in roughly similar amounts but His, Cys, Met, Tyr, and Trp tend to be less abundant than the others. Specialized proteins sometimes have unusual amino acid compositions. For example, collagen of connective tissue contains 33 mole% glycine and 21% of proline + hydroxyproline residues; the major proteins of saliva contain 22% of glutamate + glutamine and 20–45% proline.[89] Cell walls of plants contain both high proline and high glycine polypeptides. One from petunias is 67% glycine.[90] Silk fibroin contains 45% glycine and 29% alanine. A DNA repair protein of yeast has 13 consecutive aspartate residues.[91] The tough eggshell (chorion) of the domesticated silkmoth *Bombyx mori* contains proteins with ~30% cysteine.[92] Many proteins consist, in part, of repeated short sequences. For example, the malaria-causing *Plasmodium falciparum* in its sporozoite stage is coated with a protein that contains 37 repeats of the sequence NANP interspersed with 4 repeats of NDVP.[93] These two sequences have been indicated with single-letter abbreviations for the amino acids.

With a large number of protein and DNA sequences available, it has become worthwhile to compare sequences of the same protein in different species or of different proteins within the same or different species. Computer programs make it possible to recognize **similarities** and **homologies** between sequences even when deletions and insertions have occurred.[88,88a,94–97] The term homology has the precise biological definition "*having a common evolutionary origin*," but it is often used to describe any close similarity in sequence.[98] Among a pair of homologous proteins, a change at a given point in a sequence may be either **conservative**, meaning that a residue of similar character (large, small, positively charged, nonpolar, etc.) has been substituted, or it may be **nonconservative**.

## D. The Architecture of Folded Proteins

All proteins are made in the same way but as the growing peptide chains peel off from the ribosome, each of the thousands of different proteins in a living cell folds into its own special **tertiary structure**.[88a] The number of possible conformations of a protein chain is enormous. Consider a 300-residue polypeptide which could stretch in fully extended form for ~100 nm. If the chain were folded back on itself about 13 times it could form a 7-nm square sheet about 0.5 nm thick. The same polypeptide could form a thin helical rod 45 nm long and ~1.1 nm thick. If it had the right amino acid sequence it could be joined by two other similar chains to form a collagen-type triple helix of 87 nm length and about 1.5 nm diameter (Fig. 2-7).

The highly folded **globular proteins** vary considerably in the tightness of packing and the amount of internal water of hydration.[99,100] However, a density of ~1.4 g $cm^{-3}$ is typical. With an average mass per residue of 115 Da our 300-residue polypeptide would have a mass of 34.5 kDa or 5.74 x $10^{-20}$ g and a volume of 41 $nm^3$. This might be approximated by a cube 3.45 nm in width, a "brick" of dimensions 1.8 x 3.6 x 6.3 nm, or a sphere of diameter 4.3 nm. Although protein molecules are usually very irregular in shape,[101] for purposes of calculation idealized ellipsoid and rod shapes are often assumed (Fig. 2-7).

It is informative to compare these dimensions with those of the smallest structures visible in cells; for example, a bacterial flagellum is ~13 nm in diameter and a cell membrane ~8–10 nm in thickness. Bricks of the size of the 300-residue polypeptide could be used to assemble a bacterial flagellum or a eukaryotic microtubule. Helical polypeptides may extend through cell membranes and project on both sides, while a globular protein of the same chain length may be almost completely embedded in the membrane.

### 1. Conformations of Polypeptide Chains

To understand how a polypeptide chain folds we need to look carefully at the possible conformations of the peptide units. Since each peptide unit is nearly planar, we can think of a polypeptide as a chain of flat units fastened together as in Fig. 2-8. Every peptide unit is connected to the next by the α-carbon of an amino acid. This carbon provides two single bonds to the chain and rotation can occur about both of them (except in the cyclic amino acid proline). To specify the conformation of an amino acid unit in a polypeptide chain, we must describe the torsion angles about both of these single bonds.[11,76,102] These angles are indicated by the symbols **ϕ (phi)** and **ψ (psi)** and are assigned the value 180° for the fully extended chain as shown in Fig. 2-8. Each angle is taken as zero for the impossible conformation in which the two chain ends are in the eclipsed conformation. By the same token, the torsion angle **ω (omega)** around the C–N bond of the amide is 0° for a planar *cis* peptide linkage and 180° for the usual *trans* linkage.

Since both ϕ and ψ can vary for each residue in a protein, there are a large number of possible conformations. However, many are excluded because they bring certain atoms into collision. This fact can be established readily by study of molecular models.

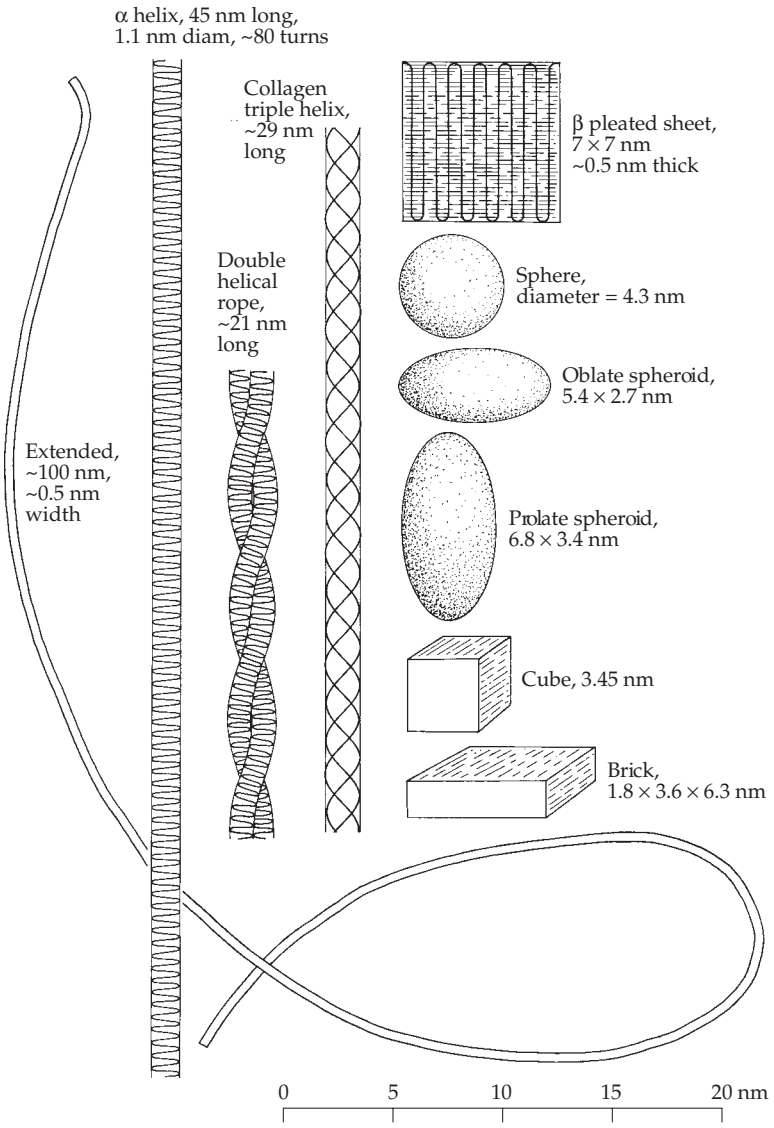



**Figure 2-7** Some idealized shapes that a 34.5 kDa protein molecule of 300 amino acids might assume.

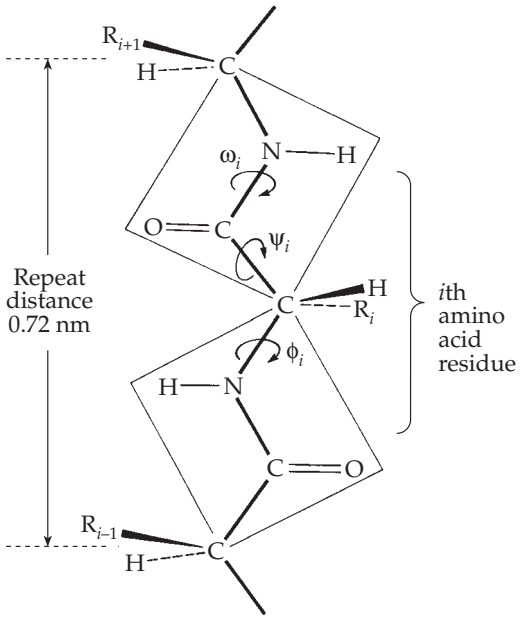



**Figure 2-8** Two peptide units in the completely extended β conformation. The torsion angles $\phi_i$, $\psi_i$, and $\omega_i$ are defined as 0° when the main chain atoms assume the *cis* or eclipsed conformation. The angles in the completely extended chain are all 180°. The distance from one α carbon atom ($C_\alpha$) to the next in a peptide chain is always 0.38 nm, no matter how the chain is folded.

Using a computer, it is possible to study the whole range of possible combinations of ϕ and ψ. This has been done for the peptide linkage by Ramachandran. The results are often presented as plots of ϕ vs ψ (Ramachandran plots or **conformational maps**)[102a,103,103a] in which possible combinations of the two angles are indicated by blocked out areas. The original Ramachandran plots were made by representing the atoms as hard spheres of appropriate van der Waals radii, but the version shown in Fig. 2-9 was calculated using a complex potential energy function to represent the van der Waals attraction and the repulsion from close contact.[103] This map was calculated for poly-L-alanine but it would be very similar for most amino acids.

Notice the two areas connected by a higher energy bridge on the left side of Fig. 2-9. The upper area contains the pairs of torsion angles for the extended **β structures** as well as for **collagen**. The lower area contains allowed conformations for the *right-handed helices.* As can be seen from Fig. 2-10, most of the observed conformations of peptide units in a real protein fall into these regions. Glycyl residues are an exception. Since glycine has no β-carbon atom, the conformations are less restricted. Out of nearly 1900 non-glycine residues in well-determined protein structures, 66 were found in disallowed areas of the Ramachandran diagram.[104a] These were often accommodated by local distortions in bond angles. The positions at which such steric strain occurs are often in regions concerned with function.[104b] One residue, which lies in a disallowed region in Fig. 2-10 is asparagine 297 of aspartate aminotransferase (Fig. 2-6). It is located

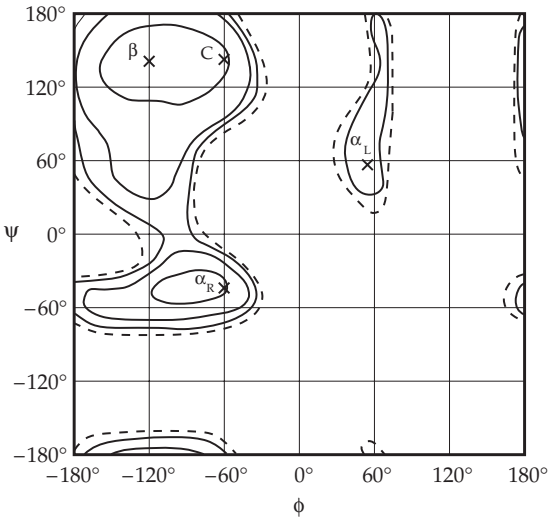



**Figure 2-9** Potential energy distribution in the ϕ–ψ plane for a pair of peptide units with alanyl residues calculated using potential parameters of Scheraga and Flory. Contours are drawn at intervals of 1 kcal (4.184 kJ) per mol going down from 0 kcal per mol. The zero contour is dashed. From Ramachandran *et al*.[104] The points marked x are for the four ideal structures: twisted β structure (β), collagen (C), right-handed α helix ($\alpha_R$), and the less favored left-handed α helix ($\alpha_L$).

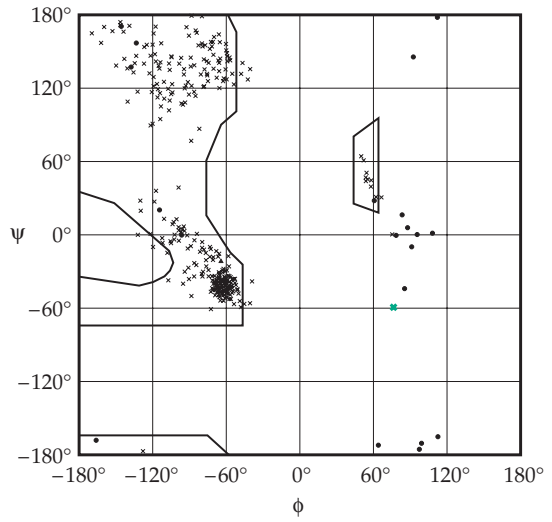



**Figure 2-10** Ramachandran plot for cytosolic aspartate aminotransferase. The angles ϕ and ψ were determined experimentally from X-ray diffraction data at 0.16 nm resolution and model building. The majority of conformations are those of α helices or of β structure. Glycine residues are indicated by filled circles, while all other residues are denoted by an "×". One of these (green) lies quite far from an allowed area and must give rise to localized strain.[104b] Extreme lower limit "allowed" regions by the hard-sphere criteria are shown in outline. From coordinates of Arthur Arnone *et al*. (unpublished).[106]

adjacent to the coenzyme site. The possible conformations of proline residues are limited. The angle ϕ is always −60 ± 20°, while ψ for the residue adjacent to the proline N can be either ~150° or ~ −30°.[105] Typical ϕ, ψ angles for some regular peptide structures are given in Table 2-3.

Conformations of side chain groups are designated by a series of torsion angles designated $\chi_i$.[11] Within proteins there are strong preferences for certain $\chi_i$,$\chi_2$ pairs.[107] Torsion angles within proline rings are considered in Chapter 5,A,6.

## 2. The Extended Chain β Structures

As was first pointed out by Pauling and Corey,[108,109] an important structural principle is that within proteins the maximum possible number of hydrogen bonds involving the C=O and N–H groups of the peptide chain should be formed. One simple way to do this is to line up fully extended chains (ϕ = ψ = 180°) and to form hydrogen bonds between them. Such a structure

**TABLE 2-3**
**Approximate Torsion Angles for Some Regular Peptide Structures[a]**

| Structure | | | ϕ (deg.) | ψ (deg.) |
|---|---|---|---|---|
| Hypothetical fully extended polyglycine chain[b] | | | −180 | +180 |
| β-Poly(L-alanine) in antiparallel-chain pleated sheet | | | −139 | +135 |
| Parallel-chain pleated sheet | | | −119 | +113 |
| Twisted β strand | | | −120 | +140 |
| Polyglycine II | | | −80 | +150 |
| Poly(L-proline) II | | | −78 | +149 |
| Collagen[c] | | | −60 ± 15 | −140 ± 15 |
| Right-handed α helix[d] | | | −57 | −47 |
| Left-handed α helix | | | +57 | +47 |
| β Bends: | Type I, | residue 2 | −60 | −30 |
| | | residue 3 | −90 | 0 |
| | Type II, | residue 2 | −60 | 120 |
| | | residue 3 | 80 | 0 |
| | Type III, | residue 2 & 3 | −60 | −30 |
| β Bulges[e]: | "Classical" β bulge | residue 1 | −100 | −45 |
| | G1 bulge | residue 1 | 85 | 0 |

[a] From Liébecq, C., ed. (1992) *Biochemical Nomenclature*, Portland Press, London and Chapel Hill (for the International Union of Biochemistry and Molecular Biology).

[b] Torsion angles for the fully extended chain can be designated either +180° or –180°, the two being equivalent. They are given as ϕ = –180° and ψ = +180° to facilitate comparison with the other structures.

[c] Ramachandran, G. N. (1967) *Treatise on Collagen*, Vol. 1, p. 124, Academic Press, New York

[d] Both ϕ and ψ are quite variable but ϕ + ψ = –104°.

[e] For residues other than these indicated the torsion angles are about those of typical β structures.

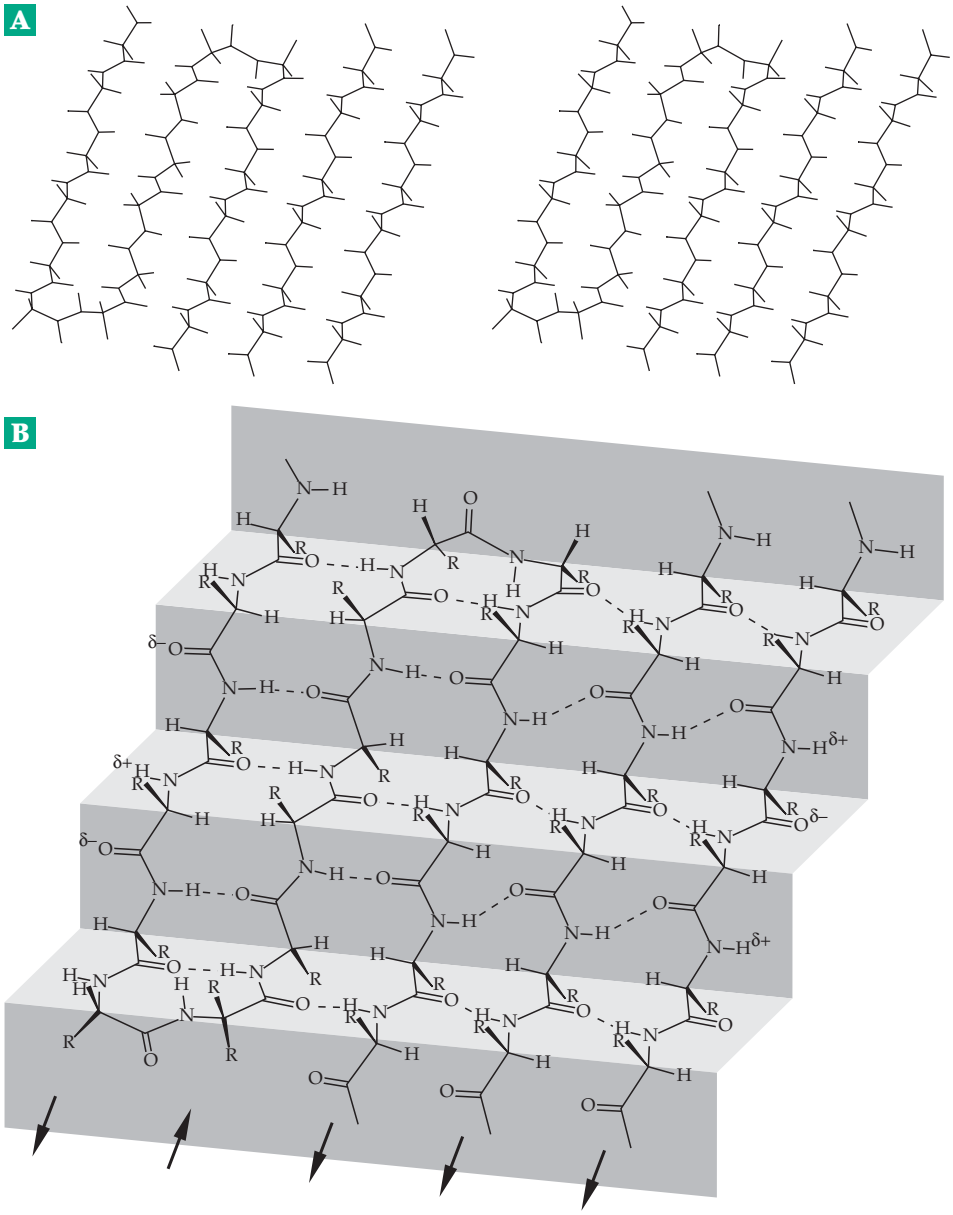



**Figure 2-11** The extended chain β pleated sheet structures. (A) Stereoscopic drawing without atomic symbols. (B) Drawing with atomic symbols. At the left is the *antiparallel* structure. The 0.70 nm spacing is slightly decreased from the fully extended length. The amino acid side chains (R) extend alternately above and below the plane of the "accordion pleated" sheet. The pairs of linear hydrogen bonds between the chains impart great strength to the structure. The chain can fold back on itself using a "β turn" perpendicular to the plane of the pleated sheet. The *parallel* chain structure (right side) is similar but with a less favorable hydrogen bonding arrangement. Arrows indicate chain directions.

exists for polyglycine and resembles that in Fig. 2-11. Notice that on the left side of this figure, the adjacent chains run in opposite directions; hence, the term **antiparallel β structure**. The antiparallel arrangement not only gives the best hydrogen bond formation between chains but also permits a single chain to fold back on itself giving a compact hairpin loop.

***Pleated sheets.*** While a fully extended polyglycine chain is possible, the side chains of other amino acids cannot be accommodated without some distortion of the structure. Thus, the peptide chains in silk fibroin have a repeat distance of 0.70 nm compared with the 0.72 nm for the fully extended chain (Fig. 2-8). Pauling and Corey[108] showed that this shortening of the chain could result from rotation of angle ϕ by ~40° (to –140°) and rotation of ψ in the opposite direction by ~45° (to +135°) to give a slightly puckered chain. The resulting multichain structure (shown in Fig. 2-11) is known as a **pleated sheet**. As in this figure, both

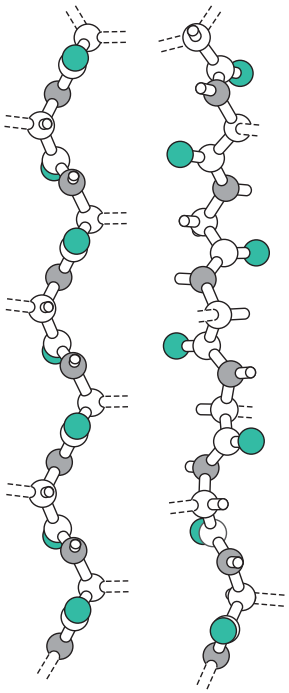


**Figure 2-12** Straight (left) and twisted (right) peptide chains in extended β conformations. From Chothia.[110]

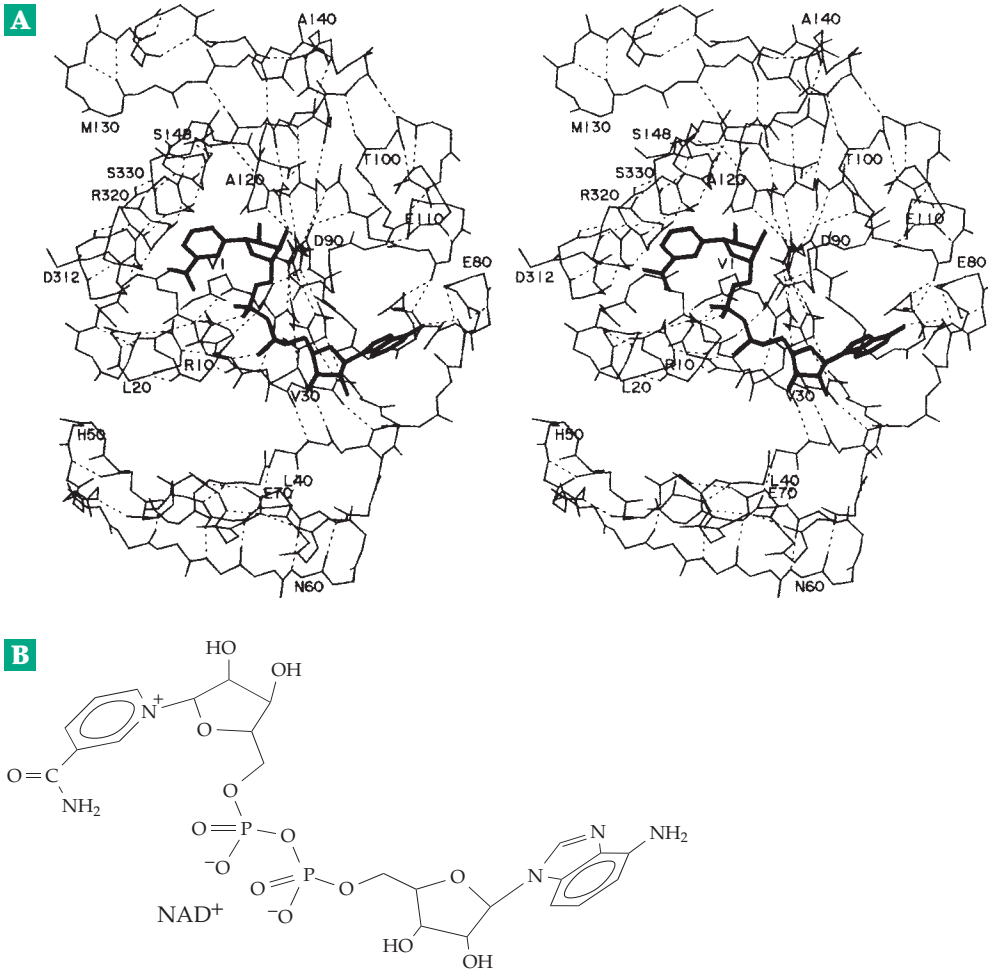



**Figure 2-13** (A) Stereoscopic view of the nucleotide binding domain of glyceraldehyde phosphate dehydrogenase. The enzyme is from *Bacillus stearothermophilus* but is homologous to the enzyme from animal sources. Residues are numbered 0–148. In this wire model all of the main chain C, O, and N atoms are shown but side chains have been omitted. The large central twisted β sheet, with strands roughly perpendicular to the page, is seen clearly; hydrogen bonds are indicated by dashed lines. Helices are visible on both sides of the sheet. The coenzyme $NAD^+$ is bound at the end of the β sheet toward the viewer. Note that the two phosphate groups in the center of the $NAD^+$ are H-bonded to the N terminus of the helix beginning with R10. From Skarzynski *et al.*[111a] (B) Structural formula for $NAD^+$.

parallel and antiparallel strands are often present in a single β sheet within a protein.

The β structure is one of the most important **secondary structures** in proteins. It occurs in about 80% of the soluble globular proteins whose structures have been determined. In many cases almost the entire protein is made up of β structure. Single strands of extended polypeptide chain are sometimes present within globular proteins but more often a chain folds back on itself to form a hairpin loop. A second fold may be added to form an antiparallel "β meander"[102] and additional folds to form β sheets. Beta structures are found in silk fibers (Box 2-B) as well as in soluble proteins.

***Twisted sheets.*** X-ray diffraction studies have shown that β pleated sheets are usually not flat but are twisted. In a twisted sheet the individual polypeptide chains make a shallow left-handed helix. However, when successive carbonyl groups are viewed along the direction of the chain, a **right-handed twist** is seen (Fig. 2-12).[110] Such twisted β sheets are often found in the globular proteins. An example (Fig. 2-13) is the "nucleotide-binding" domain of a dehydrogenase enzyme. The twist of the sheet is seen clearly in this stereoscopic view. When such chains are associated into β sheets, whether parallel or antiparallel, and are viewed in a direction perpendicular to the chains and looking down the edge of the sheet, a left-handed "propeller" is seen. Such a propeller is visible in the

drawing of carboxypeptidase A (Fig. 2-14).

The cause of the twist in β sheets appears to lie in noncovalent interactions between hydrogen atoms on the β-carbon atoms of side chains and the peptide backbone atoms. For side chains of most L- amino acids these interactions provide a small tendency towards the observed right-handed twist.[111] Nonplanarity in the amide groups (Fig. 2-5) may also contribute. Interstrand interactions seem to be important.[111b]

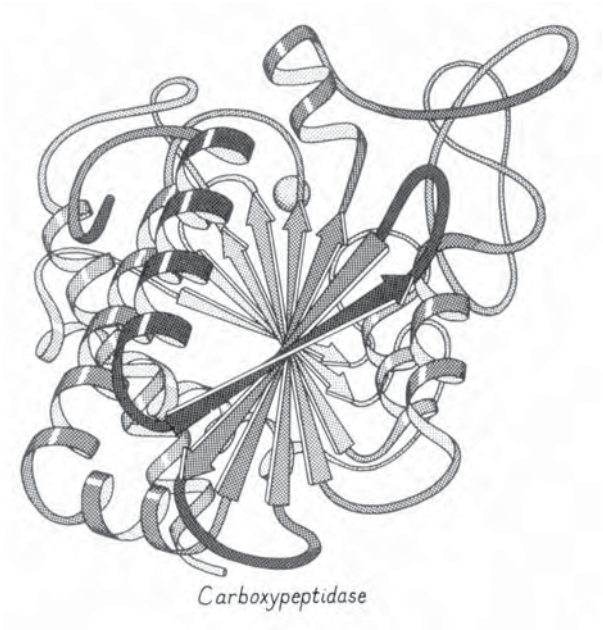



**Figure 2-14** A "ribbon" drawing of the 307- residue protein-hydrolyzing enzyme carboxypeptidase A. In this type of drawing wide ribbons are used to show β strands and helical turns while narrower ribbons are used for bends and loops of the peptide chains. The direction from the N terminus to C terminus is indicated by the arrowheads on the β strands. No individual atoms are shown and side chains are omitted. Courtesy of Jane Richardson.[117]

***Properties of β sheets.*** In antiparallel β sheets, nonpolar residues are often present on one side of the sheet and polar residues on the other. The nonpolar side may be buried in the protein, perhaps backed up against another β sheet of similar structure as in the carbohydrate-binding **lectin** shown in Fig 2-15 to give a **β sandwich**. To accommodate this packing arrangement, nonpolar and polar residues tend to alternate in the amino acid sequence. The facing nonpolar surfaces of the two sheets are essentially smooth. The β strands in one sheet lie at an angle of ~30° to those in the other sheet. The twist of the strands allows pairs of nonpolar side chains from the two sheets to maintain good van der Waals contact with each other for a considerable distance along the strands. Beta sheets of silks (Box 2-B) and of immunoglobulin domains (Fig 2-16B) are also thought to be associated in a back-to-back fashion. *Parallel* β structures have been found only when there are five or more strands, some of which may be antiparallel to the others. The parallel β structure is apparently less stable than the antiparallel structure. Parallel strands are usually buried in the protein, being surrounded by either other extended strands or helices.

Chains of hydrogen bonds and amide linkages pass across β sheets perpendicular to the chain direction. There are partial positive and negative charges along the outside edges of the sheet where these hydrogen bond chains terminate (Fig. 2-11). The polarity of these chains alternates and the positive end of one peptide bond is relatively near to the negative end of the next along the edges of the sheets. These "unsatisfied ends" of hydrogen bond chains are often sites of

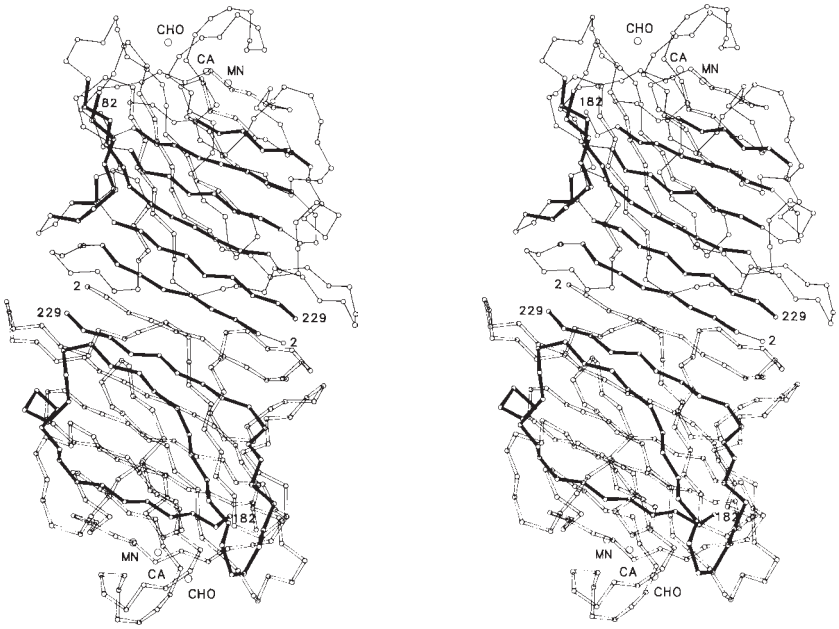



**Figure 2-15** A stereoscopic **alpha-carbon plot** showing the three-dimensional structure of favin, a sugar-binding lectin from the broad bean (*Vicia faba*). In this plot only the α-carbon atoms are shown at the vertices. The planar peptide units are represented as straight line segments. Side chains are not shown. The protein consists of two identical subunits, each composed of a 20-kDa α chain and a 20-kDa β chain. The view is down the twofold rotational axis of the molecule. In the upper subunit the residues involved in the front β sheet are connected by double lines, while those in the back sheet are connected by heavy solid lines. In the lower subunit the α chain is emphasized. Notice how the back β *sheet* (not the chain) is continuous between the two subunits. Sites for bound $Mn^{2+}$ (MN), $Ca^{2+}$ (CA), and sugar (CHO) are marked by larger circles. From Reeke and Becker.[112]

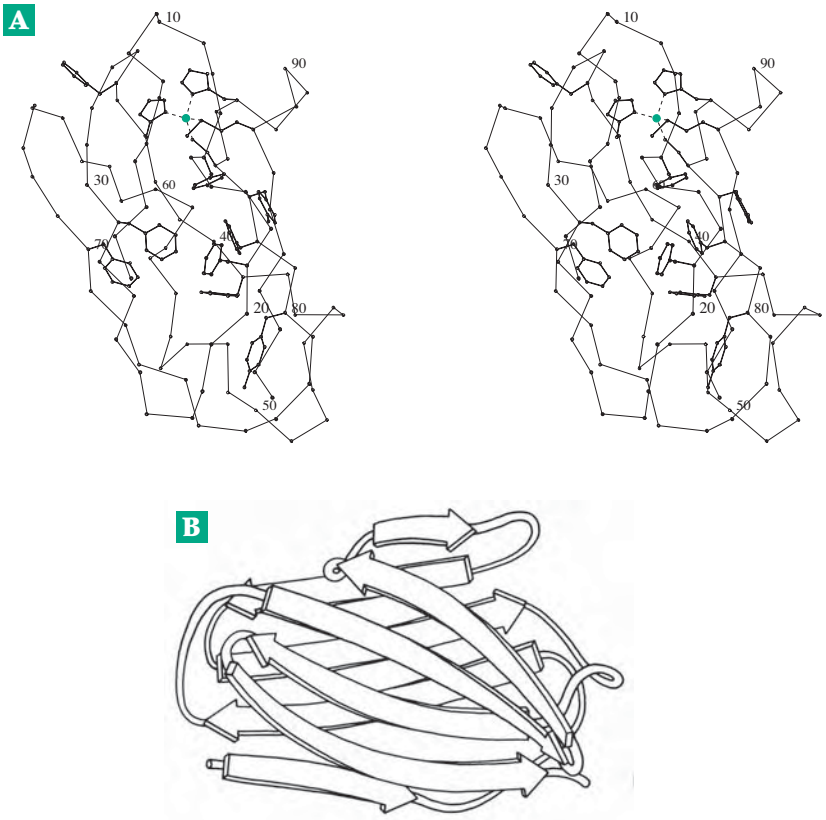



**Figure 2-16** Beta cylinders. (A) Stereoscopic α-carbon plot of plastocyanin, a copper-containing electron-transferring protein of chloroplasts. The copper atom at the top is also shown coordinated by the nitrogen atoms of two histidine side chains. The side chains of the aromatic residues phenylalanines 19, 29, 35, 41, 70, and 82 and tyrosines 80 and 83 are also shown. Most of these form an internal cluster. From Guss and Freeman.[116] (B) Ribbon drawing of immunoglobulin fold. This is a common structure in domains of the immunoglobulins and in many other extracellular proteins. Two layers of antiparallel β sheet are stacked face to face to form a flattened barrel. One disulfide bridge is always present and is represented as a thick rod. From J. Richardson.[117] (C) Five tandem fibronectin type III domains. These domains, which are found in the muscle protein titin as well as in fibronectin, resemble immunoglobulin domains but lack disulfide bridges. From Erickson.[117a] Figure courtesy of Harold P. Erickson.

interaction of polar groups from side chains. Thus, OH groups of serine or threonine residues, amide groups of asparagine and glutamine residues, etc., often fold back and hydrogen-bond to these ends.

Edges of β sheets can also serve as binding sites for other polar molecules. For example, substrates bind to an edge of a β sheet in the active sites of trypsin and other proteases (Chapter 12). Some proteins, e.g. the lectin shown in Fig. 2-15, form dimers by joining identical edges of a β sheet in antiparallel orientation.[112]

***Cylinders and barrels.*** The twisted β sheets of proteins are often curved to form structures known as **β cylinders** or **β barrels** (Fig. 2-16).[113,114] Simple cylinders formed by parallel β strands form the backbones of the electron transport protein plastocyanin, the enzyme superoxide dismutase, the oxygen carrier hemerythin (Fig. 16-20), transporter proteins that carry hydrophobic ligands,[115] and the immunoglobulins in which each domain contains a β barrel.

The eight-stranded β cylinder of plastocyanin (Fig. 2-16A) is somewhat flattened and can also be regarded as a β sandwich.[116,118] However, the β barrel of triose phosphate isomerase (see Fig. 2-28) is surrounded by eight α helices which provide additional stability and a high symmetry. Bacterial outer membranes contain pores created by very large β cylinders within proteins called **porin**s.[119,120] The one shown in Fig. 8-20 has 16 strands.

A single β strand can also be wound into a cylinder with the hydrogen bonds running parallel to the helix axis. A right-handed **parallel β helix** of this type has been found in the bacterial enzyme **pectate lyase**.[121,122] The polypeptide chains of the 353-residue protein contain seven complete turns of about 22

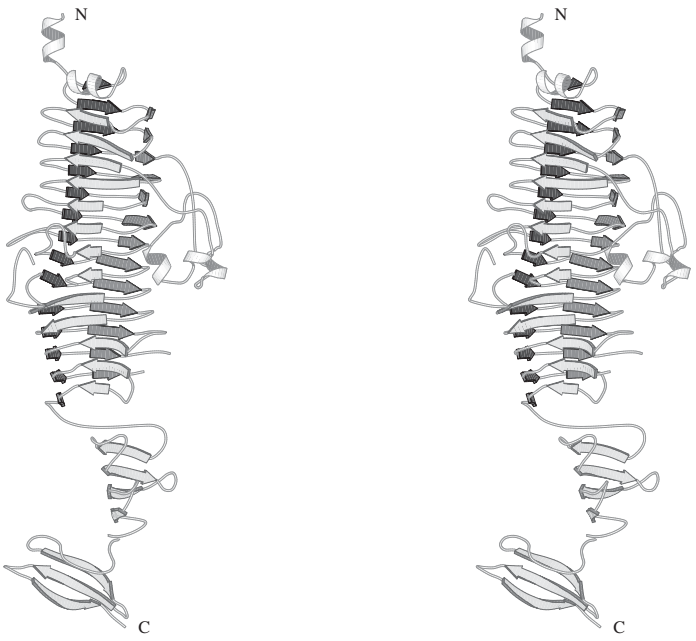



**Figure 2-17** Wire model of the tailspike protein of bacteriophage P22 of *Salmonella*. Three of these fish-shaped molecules associate as a trimer to form the spike. From Steinbacher *et al*.[123]

residues each which form a cylindrical parallel β sheet. The cylinder is folded inward into roughly an L-shape with packed side chains filling the remaining inside space. The tailspikes of bacteriophage P22, a virus of *Salmonella*, are formed from 666-residue protein subunits which contain 13 turns of parallel β helix (Fig. 2-17).[123]

The tendency for a β sheet to fold into a cylinder is encouraged in antiparallel β structures by the existence of a common irregularity called the β **bulge**.[124,125] As illustrated in Fig. 2-18, a β bulge contains an extra residue inserted into one of the chains. In the second type of bulge shown in Fig. 2-18, the extra residue is glycine with torsion angles of about $\phi = 85°$, $\psi = 0°$, which are possible only for glycine. In the two β cylinders of trypsin, chymotrypsin, and elastase (Fig. 12-9), there are seven β bulges.

Beta structures are found in many small peptides. The hormone oxytocin (Fig. 2-4), the antibiotics gramicidin S (Fig. 2-4) and valinomycin (Fig. 8-22), and the mushroom peptide antamanide (Box 28-B) are among these. The cyclic structures of these compounds favor formation of antiparallel β strands with sharp turns at the ends. Polypeptide antibiotics that have alternating

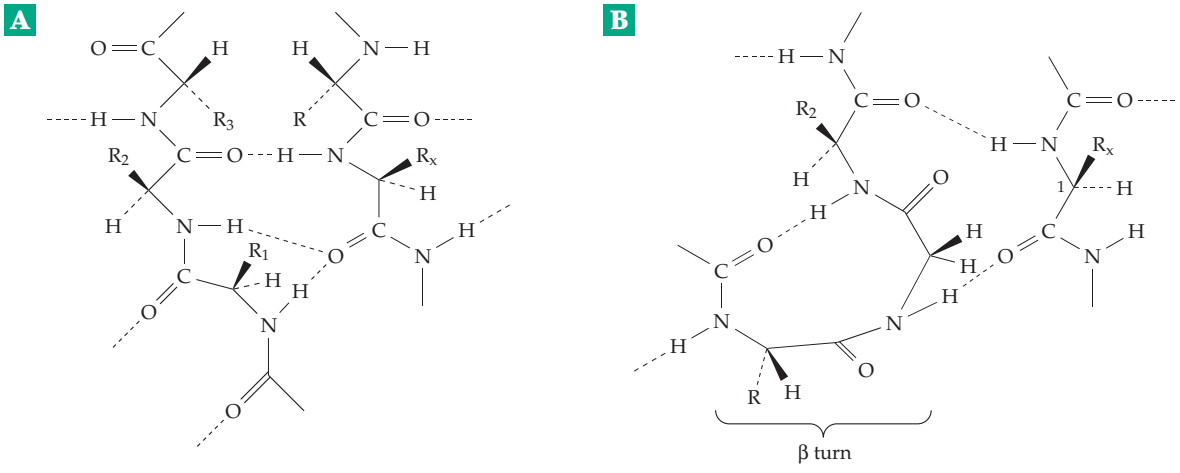



**Figure 2-18** Typical β bulges in antiparallel pleated sheets. The residues $R_1$, $R_2$, and $R_x$ identify the bulges. (A) A "classic" β bulge, in which $\phi_1$ and $\psi_1$ are nearly those of an α helix while other torsion angles are approximately those of regular β structures. (B) The G1 bulge in which the first residue is glycine with $\phi_1 = 85°$, $\psi_1 = 0°$. It is attached to a type II β turn of which the glycine (labeled 1) is the third residue.

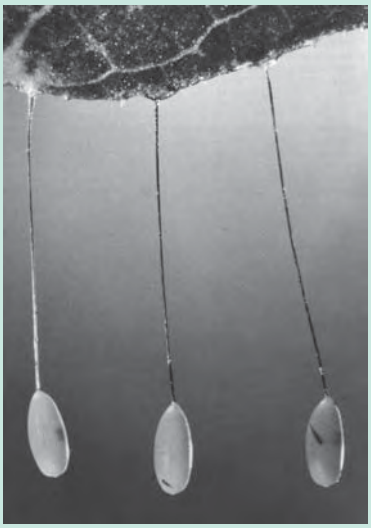
sequences of D and L residues can be coiled into a β helix[126,127] or a pair of polypeptide chains arranged in an antiparallel fashion can form a double stranded β helix with an 0.3 to 0.4-nm hydrophilic pore through the center. These peptide antibiotics appear to exert their antibacterial action by creating pores through cell membranes and allowing ions to pass through without control.

***Beta propellers.*** Another major folding pattern is a circular array of four to eight "blades" that form a propeller-like structure. Each blade is a small, roughly triangular four-stranded antiparallel β sheet (See Figs. 11-7 and 15-23). Sequences that fold into these blades can often, but not always, be recognized as **WD repeats**. These are typically 44- to 60-residue sequences that have the sequence GH (Gly-His) about 11–24 residues from the N terminus and WD (Trp-Asp) at the C terminus.[127a,127b] This repeat sequence encodes the

## BOX 2-B SILKS

The green lacewing fly *Chrysopa flava* lays its eggs on 1 cm silk stalks glued to the undersides of leaves. It has been proposed, as for other silks, that the peptide chains in the stalks are aligned perpendicular to the long dimension of the fiber and are folded back on themselves many times to form a β sheet with only ~8 residues between folds.[a] The chains of **silk fibroin**, the major protein of silkworm silk, contain 50 repeats of the sequence:[b]

$$GAGAGSGAAG(SGAGAG)_8Y$$



Courtesy of Ralph Buchsbaum

All of the alanine and serine side chains of this sequence presumably protrude on one side of a β sheet while the other side has only the hydrogen atoms of the glycine. This permits an efficient stacking of the sheets with interdigitation of the side chains of the alanine and serine residues of adjacent sheets.

However, DNA sequences of silk genes have revealed a greater complexity. The silk fibroin sequence suggests that at bends between β strands there are often S – S bridges and the crystalline β sheet domains are interspersed with 100–200 residue segments of amorphous protein[b–d] whose coiled chains can be stretched greatly. The protein from *Bombyx mori* consists of 350-kDa heavy chains linked by a disulfide bond to 25-kDa light chains.[e] A silkworm **housing silk** contains polyalanine sequences, e.g. $(Ala)_{10-14}$. These evidently form α-helical regions.[f] Silk molecules are synthesized and stored in glands as a concentrated solution of apparently globular molecules. The silk is extruded through a spinneret, whose diameter is ~10 μm, after which the silkworm stretches the silk, causing it to form a stiff fiber. Apparently the stretching of the folded polypeptide chains permits the β-sheet-forming sequences to find each other and form oriented crystalline regions.[d]

Spiders may produce silk from as many as seven different types of glands. Dragline silk, by which a spider itself may descend, is stronger than steel but because of the coiled amorphous regions may be stretched by 35%.[g–l] The most elastic silks are those of the catching spirals of orb-webs.[i,j,l] They can be stretched 200% and contain a variety of repeated sequences, including $GPCC(X)_n$. The latter is similar to sequences in elastin (Section 4), and is able to form type II β bends, which may have proline in position 2 and must have glycine in position 3 (Fig. 2-24). A chain of repeated β bend motifs can form a flexible spring, a β spiral as proposed for elastin. Genes for spider silk have been cloned and are being used to engineer new proteins with commercial uses, e.g., to help anchor cells in regenerating body tissues.[d,h,j]


[a] Geddes, A. J., Parker, K. D., Atkins, E. D. T., and Beighton, E. (1968) *J. Mol. Biol.* **32**, 343–368
[b] Garel, J.-P. (1982) *Trends Biochem. Sci.* **7**, 105–108
[c] Vollrath, R. (1992) *Sci. Am.* **266**(Mar), 70–76
[d] Calvert, P. (1998) *Nature (London)* **393**, 309–311
[e] Mori, K., Tanaka, K., Kikuchi, Y., Waga, M., Waga, S., and Mizuno, S. (1995) *J. Mol. Biol.* **251**, 217–228
[f] van Beek, J. D., Beaulieu, L., Schäfer, H., Demura, M., Asakura, T., and Meier, B. H. (2000) *Nature (London)* **405**, 1077–1079
[g] Simmons, A. H., Michal, C. A., and Jelinski, L. W. (1996) *Science* **271**, 84–87
[h] Hinman, M. B., and Lewis, R. V. (1992) *J. Biol. Chem.* **267**, 19320–19324
[i] Hayashi, C. Y., and Lewis, R. V. (1998) *J. Mol. Biol.* **275**, 773–784
[j] Guerette, P. A., Ginzinger, D. G., Weber, B. H. F., and Gosline, J. M. (1996) *Science* **272**, 112–115
[k] Spek, E. J., Wu, H.-C., and Kallenbach, N. R. (1997) *J. Am. Chem. Soc.* **119**, 5053–5054
[l] Hayashi, C. Y., and Lewis, R. V. (2000) *Science* **287**, 1477–1479

fourth β strand of one blade followed by the first three strands of the next blade. This overlap snaps the blades together to form the propeller. At least one tight turn is present in each blade. A seven-bladed propeller is present in the β subunits of the regulatory GTP-hydrolyzing **G proteins**, which couple extracellular signals to intracellular enzymes and ion channels (Fig. 11-7). A similar β propeller is present in **clathrin**, which forms cage-like enclosures around endocytic vesicles (Chapter 8).[127c] Six-bladed propellers are predicted to occur in many extracellular proteins.[127d] A β propeller binds the coenzyme PQQ in bacterial dehydrogenases (Fig. 15-23).

## 3. Helices

The **alpha helix** represents the second major structural element of soluble proteins[108,128] and is also found in many fibrous proteins, including those of muscle and hair. In the α helix both ϕ and ψ are about −50 to −60° and except at the N-terminal helix end, *each NH is hydrogen bonded to the fourth C=O further down the chain*. All of the N–H and C=O groups of the peptide linkages lie roughly parallel to the helix axis; the N–H groups point toward the N terminus of the chain and the carbonyl groups toward the C terminus (Fig. 2-19A).

The number of amino acid units per turn of the helix is ~3.6, with five turns of the helix containing 18 residues. The **pitch** (repeat distance) of the helix, which can be determined experimentally from X-ray diffraction data, is 0.54 nm. Polar coordinates for the α helix have been tabulated.[130] With L-amino acids, the right-handed helix, is more stable than the left-handed helix which has so far not been found in proteins. Frequently, however, a few residues have the ϕ, ψ angles of this helix. The ϕ, ψ angles of the α helix are given in Table 2-3 as −57°, −47°, but are much more variable in real helices. In erythrocruorin, for which an accurate structure determination has been made,[131]

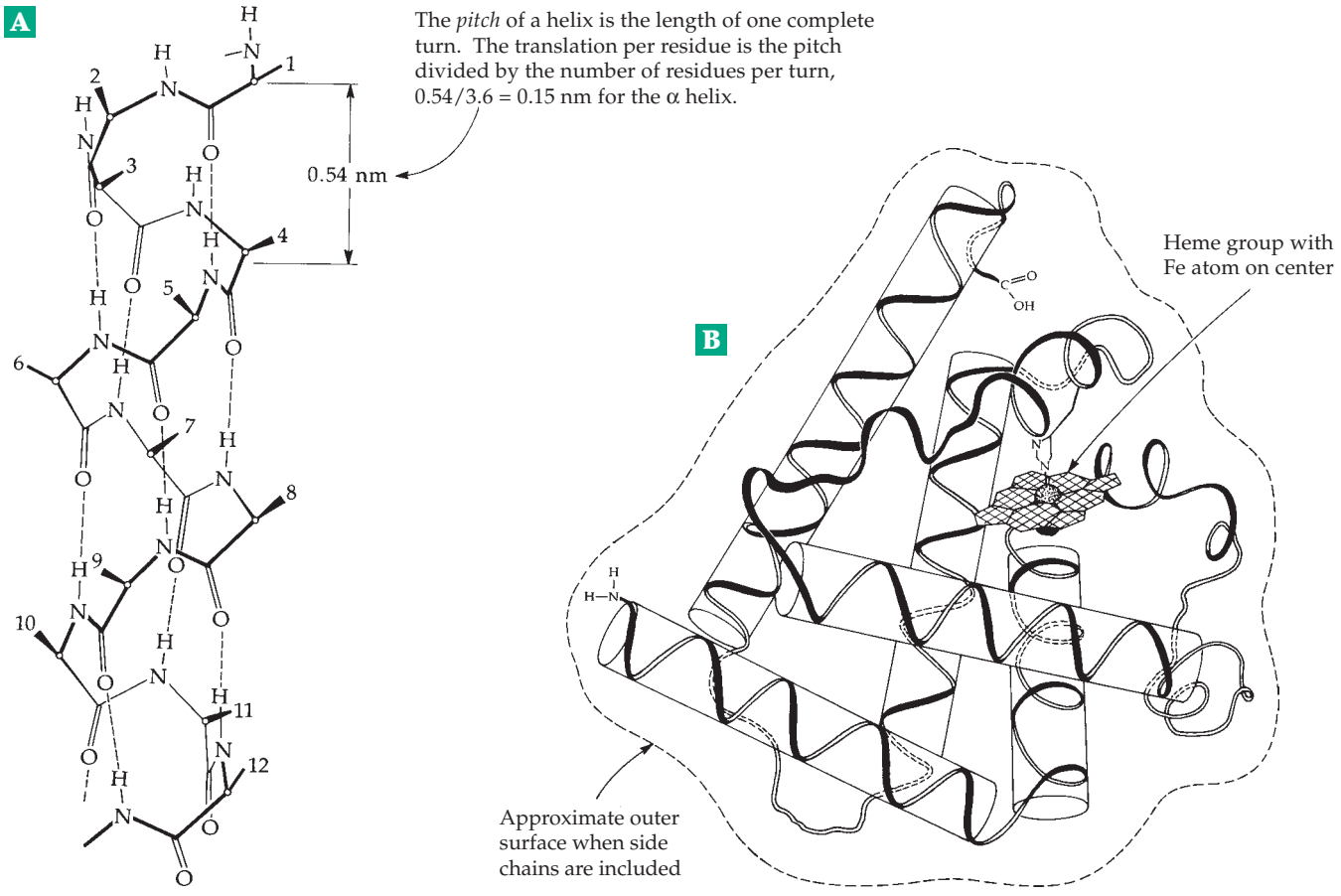



**Figure 2-19** The α helix. (A) The right-handed α helix with vertical hydrogen bonds indicated by dotted lines. The positions of the amino acid side chains are indicated by the numbers. (B) The conformation of the peptide backbone of myoglobin.[129] Five long α helices are indicated as rods. Several shorter helices can also be seen. The overall size of the molecule is approximately 4.4 × 4.4 × 2.5 nm.

the average values are $\phi = -64°$, $\psi = -40°$. More important is the observation that $\phi + \psi = 104°$ within about $\pm 10°$ for most residues in helices of this and other proteins.[132,133] The deviation from the ideal dimensions results in part from the hydrogen bonding of water molecules or polar protein side chains to the carbonyl oxygen atoms in the helix.[131]

Helix formation is spontaneous for peptides as short as 13 residues in water.[134,135] Although the difference in thermodynamic stability between an α helix and an unfolded **"random coil"** conformation is small, poly-L-alanine peptides form helices in water. Glycine destabilizes helices, presumably because of the increased entropy of the unfolded chain which results from the wider range of the conformational angles $\phi$ and $\psi$ for glycyl residues. Proline destabilizes helices even more because its restricted $\phi$ and $\psi$ angles cause the helix to be kinked.[136,137] However, prolyl residues are often present at ends of helices. Other amino acids all fit into helices but may stabilize or destabilize the helix depending upon immediately neighboring groups.[138–140] Bulky side chain groups with a low dielectric constant stabilize helices by strengthening the hydrogen bonds within the helices.[141]

Helices of smaller and larger diameter than that of the α helix are possible. The most important is the **$3_{10}$ helix** (or $3.0_{10}$ helix), which has exactly three residues per turn.[140,142–144] Each NH forms a hydrogen bond to the third C=O on down the chain; thus, the $3_{10}$ helix is tighter than the α helix. Although long $3_{10}$ helices are seldom found in proteins, a single turn of this tighter helix frequently occurs as a "defect" at the end of an α helix. A polymer of α-aminoisobutyric acid forms long $3_{10}$ helices because the α-dimethyl side chains constrain $\phi$ and $\psi$ to appropriate values.[145] Short helical peptide chains in water may exist as a mixture of α and $3_{10}$ forms in equilibrium.[146] The **π helix**, with 4.4 residues per turn is of a larger diameter than the α helix and has only been found in proteins as a single turn, usually at a C terminus.[132,132a]

***Properties of helices.*** The dipoles of the backbone amide linkages of an α helix are all oriented in the same direction. The positive end of each dipole is associated through hydrogen bonding with the negative end of another. This leaves three partial positive charges at the N terminus of the helix and three partial

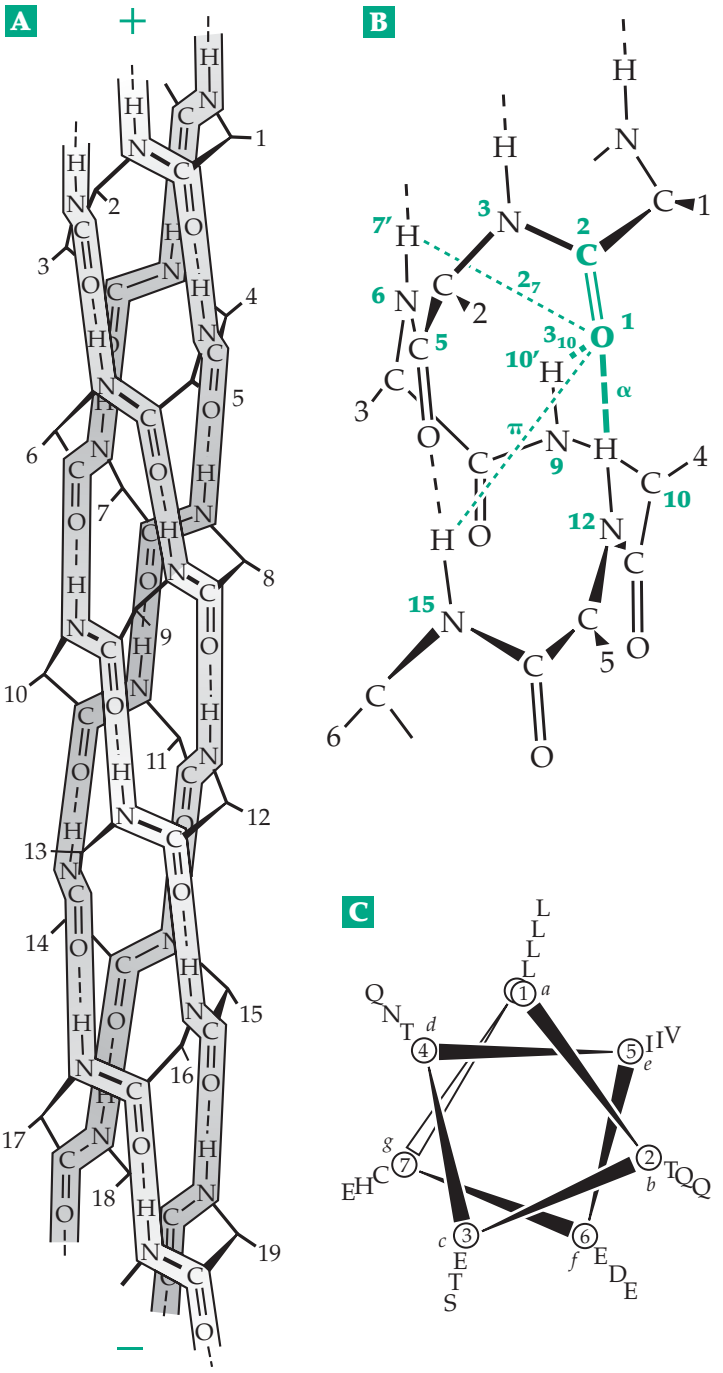



**Figure 2-20** (A) The α helix showing the three chains of interconnected amide groups and hydrogen bonds with partial net positive and negative charges at the ends. These chains run across the turns of the helical polypeptide backbone. (B) Scheme illustrating hydrogen-bonding pattern for $2_7$ ribbon, $3_{10}$ helix, α helix, and π helix. (C) The α helix represented as a helical wheel. Imagine viewing the helix from the N-terminal end of the segment with the lines corresponding to the backbone of the peptide. Residues, which are designated by single letters, are spaced 100° apart since there are 3.6 residues/turn. The peptide shown is a 22-residue sequence from a "leucine-zipper" domain of a protein that participates in gene regulation. From Fathallah-Shaykh *et al.*[145a]

negative charges at the C terminus and creates a "macrodipole." It has been estimated that each end of the helix carries one-half an elementary unit of charge.[147–148b] However, this may be an overestimate. The partial charges are connected by three chains of hydrogen bonds that run across the turns of the helix as is indicated in Fig. 2-20A. These chains are polarized and also possess a large **hyperpolarizability**,[149] i.e., the polarization of the helix increases more rapidly than in direct proportion to an applied electrical field.

In a $3_{10}$ helix, there are just two chains of hydrogen bonds across the polypeptide chain and running the length of the helix. The characteristic hydrogen bonding of the α helix and that of the $3_{10}$ helix are often both present within a helix. Irregularities with $3_{10}$ type hydrogen bonds, arising from interactions with amino acid side chains or with solvent molecules, may cause a helix to be kinked or curved.[132] Side chains of polar residues, including those of Asn, Asp, Thr, and Ser (and less often Glu, Gln, or His), frequently fold back and hydrogen bond to the NH and CO groups that carry the partial charges at the helix ends (Fig. 2-20). The side chain of the third residue in the helix may also hydrogen bond to the NH of the first residue.[133,150–157] The hydrogen bonding of a negatively charged side chain group to the N-terminal end of the helix or of a positively charged group to the C-terminal end provides an **N-cap** or a **C-cap** which helps to stabilize the helix by strengthening its hydrogen bonds.[158] However, the most frequent residue at the C-terminal end is glycine.[151,155,159] Helices often point toward active sites of enzymes and interactions of the helix dipoles with substrates undergoing reaction may be important to the mechanism of action of these catalysts.[148,149,160]

***Stacking of helices in proteins.*** Many proteins are made up almost entirely of α helices. One of these, **myoglobin**, was the first protein for which the complete three-dimensional structure was worked out by X-ray crystallography.[129] Myoglobin is a small oxygen-carrying protein of muscle. Its structure is closely related to that of hemoglobin of blood. Its 153 amino acid residues are arranged in eight different α-helical segments containing from 7 to 26 residues each. These rodlike helices are stacked together in an irregular fashion as shown in Fig. 2-19. Serum albumin (Box 2-A) has 28 helices organized into three homologous domains. In contrast, the filamentous bacterial viruses have protein coats made up of small subunits, each coiled as a single α helix. These are packed in a regular manner to form the rodlike virus coats (Fig. 7-7).

***Coiled coils.*** In a large family of proteins, two right-handed α helices are coiled around each other in a left-handed **superhelix** (Fig. 2-21).[161–167] This **coiled coil** structure was first suggested by Francis Crick[166] to account for the fact that the X-ray diffraction pattern of the **keratins** of skin and hair indicated a pitch of 0.51 nm rather than the 0.54 nm expected for an α helix.

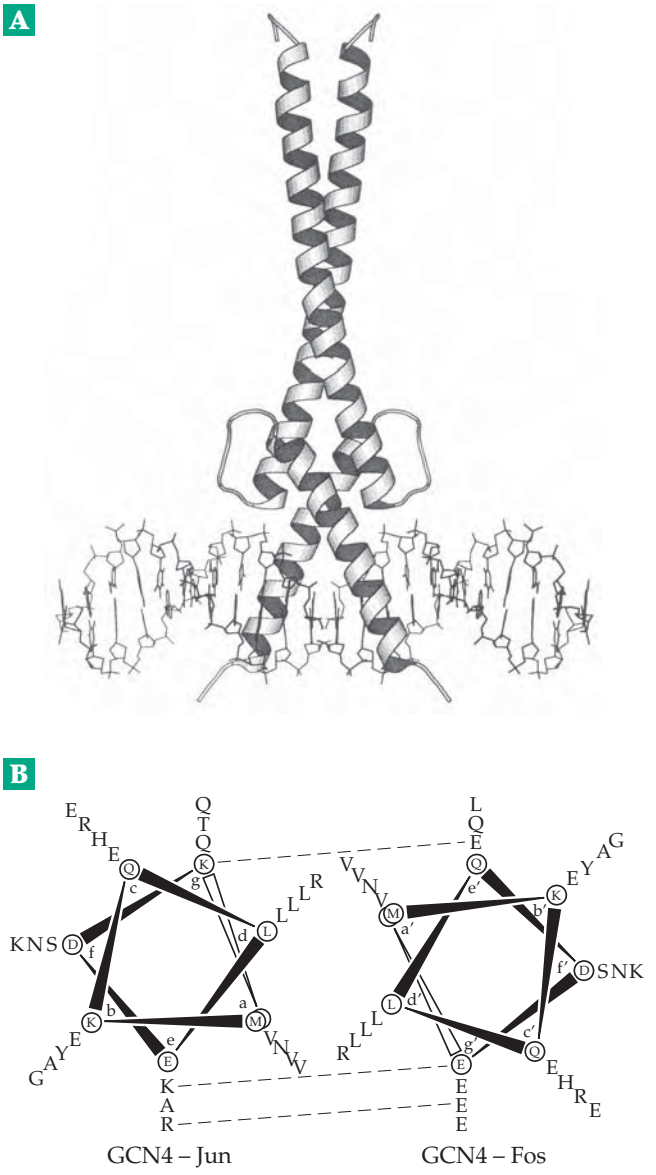



**Figure 2-21** (A) Ribbon drawing of the transcription factor called Max in a complex with DNA. The C termini of the peptide chains are at the top.[169] Courtesy of S. K. Burley. (B) Helical wheel representation of residues 2–31 of the coiled coil portion of the leucine zipper (residues 249–281) of the related transcription factor GCN4 from yeast. The view is from the N terminus and the residues in the first two turns are circled. Heptad positions are labeled *a*–*g*. Leucine side chains at positions *d* interact with residues *d′* and *e′* of the second subunit which is parallel to the first. However, several residues were altered to give a coiled coil that mimics the structure of the well-known heterodimeric oncoproteins Fos and Jun (see Chapter 11). This dimer is stabilized by ion pairs which are connected by dashed lines. See John et al.[172]

Crick suggested that two supercoiled α helices inclined at an angle of 20° to each other would produce the apparent shortening in the helix pitch and would also permit nonpolar side chains from one strand to fit into gaps in the surface of the adjacent strand, a **knobs-into-holes** bonding arrangement. Helical strands tend to coil into ropes because a favorable interstrand contact can be repeated along the length of the strands only if the strands coil about each other. A coiled coil can often be recognized by **heptad repeats**:

$$(\mathbf{a}\text{-b-c-}\mathbf{d}\text{-e-f-}\mathbf{g})_n$$
$$1\ 2\ 3\ 4\ 5\ 6\ 7$$

Here, residues *a*, *d*, and *g* (1, 4, and 7) often carry nonpolar side chains. These come together in the coiled coil as is illustrated in the **helical wheel** representations in Fig. 2-21B and provide a longitudinal hydrophobic strip along the helix.[167] Charged groups are often present in other locations and in such a way as to provide electrostatic stabilization through interactions between residues within a single α helix[170-171a] or between the pair of helices. The latter type of interaction also determines whether the coiled-coil consists of a parallel or antiparallel pair and whether the two helices are of identical or of differing amino acid sequence.[171b]

Recent attention has been focused on a DNA-binding structure called the **leucine zipper**. A pair of parallel α helices are joined as a coiled coil at one end but flare out at the other end to bind to DNA. In the yeast transcription factor GCN4, whose three-dimensional structure has been determined to high resolution (Fig. 2-21B),[172] the *d* position of the coiled coil is occupied by leucine and the *e* and *g* positions are often occupied by charged groups that form stabilizing ion pairs. Residues at positions *b*, *c*, and *f* are generally on the outside and exposed to solvent.[168,171,173] The coiled coil flares out at the C-terminal ends and carries DNA-binding groups. The structure of a related transcription factor is shown in Fig. 2-21A.[169]

The muscle proteins **myosin**[174] and **tropomyosin** also both consist of pairs of identical chains oriented in the same direction. The two 284-residue tropomyosin chains each contain 40 heptads and are linked by a single disulfide bridge. X-ray crystallographic studies[175,176] and electron microscopy[177] show that the molecule is a rod of 2.0 nm diameter and 41 nm length, the dimensions expected for the coiled coil. However, as with other "regular" protein structures, there are some irregularities. Myosin chains (Chapter 19) contain 156 heptads.

Another group of proteins have parallel coiled coils flanked by nonhelical domains in subunits that associate as filaments. These include the **keratins** of skin as well as the intermediate fiber proteins of the cytoskeleton (Chapter 7).[164,178] Natural coiled coils often have a parallel orientation, but synthetic peptides have been designed to form antiparallel coiled coils.[179,179a]

***Helix bundles.*** A third peptide chain can be added to a coiled coil to form a triple-stranded bundle.[180–183] An example is the glycoprotein **laminin** found in basement membranes. It consists of three peptide chains which, for ~600 residues at their C-terminal ends, form a three-stranded coil with heptad repeats.[182,184] Numerous proteins are folded into four helical segments that associate as **four-helix bundles** (Fig. 2-22).[185–188] These include electron carriers, hormones, and structural proteins. The four-helix bundle not only is a simple packing arrangement, but also allows interactions between the + and – ends of the macro-dipoles of the helices.

Membranes contain many largely α-helical proteins. Cell surface receptors often appear to have one, two, or several membrane-spanning helices (see Chapter 8). The single peptide chain of the bacterial light-operated ion pump **bacteriorhodopsin** (Fig. 23-45) folds back upon itself to form seven helical rods just long enough to span the bacterial membrane in which it functions.[189] Photosynthetic reaction centers contain an α helix bundle which is formed from two different protein subunits (Fig. 23-31).[190] A recently discovered **α,α barrel** contains 12 helices. Six parallel helices form an inner barrel and 6 helices antiparallel to the first 6 form an outer layer (see Fig. 2-29).[191–193]

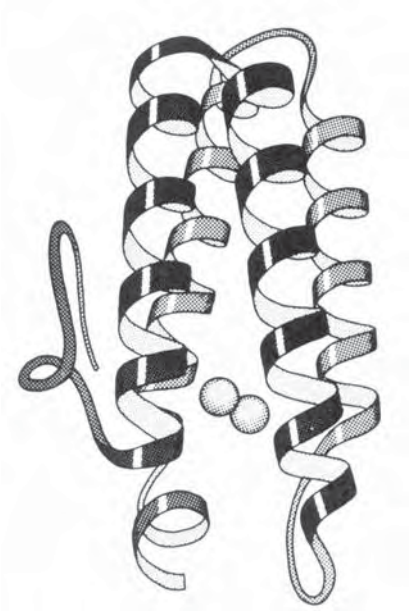


**Figure 2-22** Ribbon drawing of an up-and-down four-helix bundle in **myohemerythrin**. The two spheres represent the two iron atoms which carry an $O_2$ molecule. They are coordinated by histidine and aspartate side chains. Courtesy of J. Richardson.[117]

## 4. Polyglycine II and Collagen

In a second form of polyglycine each amino acid residue is rotated 120° from the preceding one about a 3-fold screw axis as is shown in the end view of Fig. 2-23A. The angle ψ is about 150° while ϕ is about –80° for each residue. The distance along the axis is 0.31 nm/residue and the repeat distance is 0.93 nm. The molecules can coil into either a right-handed or a left-handed helix. In this structure, the N–H and C=O groups protrude perpendicular to the axis of the helix and, as in the β structure, can form H-bonds with adjacent chains.

Poly (L-proline) assumes a similar helical structure. However, the presence of the bulky side chain groups induces a *left-handed* helical twist. **Collagen**, the principal protein of connective tissue, basement membranes, and other structures, is the most abundant protein in the animal body. Its fundamental unit of structure is a triple helix of overall dimensions 1.4 x 300 nm which resembles the structure of polyglycine II but contains only three chains.[195] The left-handed helices of the individual chains are further wound into a right-handed superhelix (Fig. 2-23B,C). Collagen contains 33% glycine and 21% (proline + hydroxyprolines). The reason for the high glycine content is that bulky side chain groups cannot fit inside the triple helix. A hydrogen-bonded cylinder of hydration surrounds each triple helix and is anchored by hydrogen bonds to the –OH groups of the hydroxyproline residues.[196] Sequences of several collagen chains have been established. One of these contains ~1050 residues, ~1020 of which consist of ~340 triplets of sequence GlyXY where Y is usually proline and X is often hydroxyproline (Hyp).[197,198] The commonest triplet is Gly-Hyp-Pro.

There are several types of collagen. In one type, two identical chains of one kind are coiled together with a third dissimilar chain to form the triple helix. Several of these triple helices associate to form 8-nm microfibrils (Fig 2-23D).[199] Once synthesized, collagen is extensively modified and crosslinked. See Chapter 8.

Collagen-like triple helices also occur within other proteins. One of these is protein C1q, a component of the **complement system** of blood (Chapter 31). This protein interacts with antibodies to trigger a major aspect of the immune response. C1q has six subunits, each made up of three different polypeptide chains of about 200 residues apiece. Beginning a few residues from the N termini, there are over 80 residues in each chain with collagen-like sequences. The three chains apparently form a triple helix within each subunit. However, the C-terminal portions are globular in nature.[200] Collagen-like tails also are present on some forms of the enzyme acetylcholinesterase (see Chapter 12C,10). The **extensins** of plant cell walls contain 4-hydroxyproline and evidently have a structure related to that of collagen.[201,202] Shorter 4- to 8-residue segments of left-handed polyproline helix are found in many proteins.[203]

## 5. Turns and Bends

To form a globular protein, a polypeptide chain must repeatedly fold back on itself. The turns or bends by which this is accomplished can be regarded as a third major secondary structural element in proteins. Turns often have precise structures, a few of which are illustrated in Fig. 2-24. As components of the loops of polypeptide chains in active sites, turns have a special importance for the functioning of enzymes and other proteins. In addition, tight turns are often sites for modification of proteins after their initial synthesis (Section F).

The **β turn** (Fig. 2-24) is often found in hairpin or reverse turns at the edges of β sheets (Fig. 2-11) and at other locations.[204–212] If all four residues that contribute to β bends are counted, they constitute about one-third of the amino acid residues in most proteins.[124] In many β turns, the C=O of the first residue hydrogen bonds to the NH of the fourth residue. This hydrogen bond may be part of the hydrogen bond network of a β pleated sheet. The peptide unit between α-carbon atoms 2 and 3 of the turn is perpendicular to the sheet. There are two possibilities for the orientation of this peptide unit. In a *type I turn*, the C=O is down when the turn is viewed as in Fig. 2-24, while the side chains of residues 2 and 3 point upwards or outward on the opposite side of the bend. In a *type II turn*, the C=O is up and the NH down. Residue 3 is always glycine in a type II turn because the side chain would collide with the C=O group if any other amino acid were present. As is seen in Fig. 2-24, a *trans*-proline can fit at position 2 in a type II turn[206] as well as in type I turns. A *cis*-proline residue can fit at position 2 or position 3 in a type I β turn.[137,213]

The *type III β turn* is similar to a type I turn but has the ϕ, ψ angles of a $3_{10}$ helix and the two chains emerging from the turn are not as nearly parallel as they are in type I turns. Beta turns of the less common types I′, II′, and III′ have a left-handed twist. As can be seen in Fig. 2-24, this permits a better match to the twist of strands in a β sheet. Unless glycine is present, these bends are less stable because of steric hindrance.[214,215] Polar side chain groups such as those of aspartate or asparagine often form hydrogen bonds to the central peptide units of β turns.[214]

The tight γ turn[215] and the proline-containing β turn shown in Fig. 2-24 are thought to be major components of the secondary structure of **elastin**.[216–218] This stretchable polymer, which consists largely of nonpolar amino acids, is the most abundant protein of the elastic fibers of skin, lungs, and arteries. The

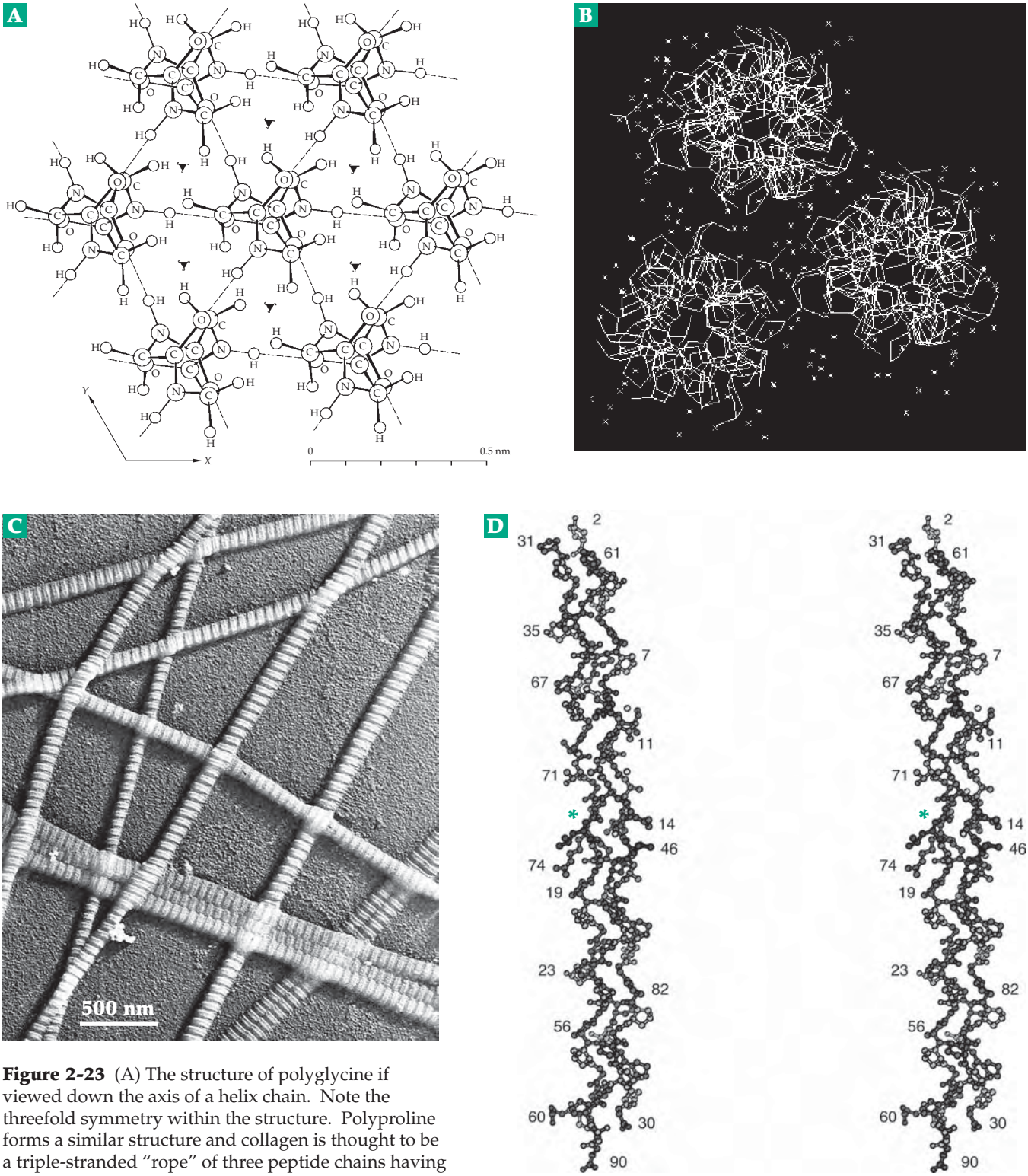



**Figure 2-23** (A) The structure of polyglycine if viewed down the axis of a helix chain. Note the threefold symmetry within the structure. Polyproline forms a similar structure and collagen is thought to be a triple-stranded "rope" of three peptide chains having essentially the same structure but in addition containing a right-handed supercoil. This is illustrated by the crystal structure of a collagen-like peptide shown in B and C. (B) View similar to that in (A), but showing how three chains form the triple-stranded ropes separated by a cylinder of hydration. The structure is shown as a wire model. The x's are water molecules. They form an extensive network of H-bonds to one another and to peptide groups. From Bella *et al.*[194] C) Electron micrograph of collagen fibrils, each of which consists of many triple-helical units. Deposited from suspension and shadow cast with chromium. Courtesy of Jerome Gross, M.D. (D) Stereoscopic view of a collagen-like model peptide. Each of the three parallel 30-residue chains contains a $(Pro\text{-}Hyp\text{-}Gly)_3$ "cap" at each end and the 12-residue sequence Ile-Thr-Gly-Ala-Arg-Gly-Leu-Ala-Gly-Pro-Hyp-Gly in the center. This sequence is found in human type III collagen and includes a site (green asterisk) of known mutations (see Chapter 8). From Kramer *et al.*[194a] Courtesy of Helen Berman.

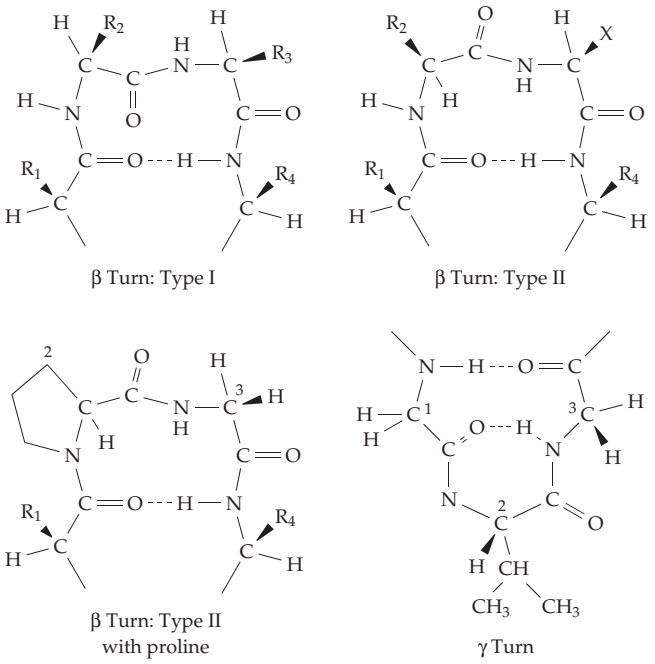



**Figure 2-24** Tight turns found in polypeptide chains. Two types of β turn are shown. A third variant, the type III or $3_{10}$ turn resembles the type I turn but has the ϕ, ψ angles of a $3_{10}$ helix. Type II β turns containing proline and tighter γ turns are thought to be major structural components of elastin. Another β turn, lacking the hydrogen bond has a *cis*-proline residue at position 3.

70-kDa chains are thought to have 70- to 75-residue regions in which the polypeptide folds back on itself repeatedly with a large number of bends in a broad left-handed **β spiral**.[216,218a] The consensus sequence VPGVP, which tends to form a type II β turn with proline in position 2 (Fig. 2-24) is present in long tandem repeats e.g., $(VPGVG)_{11}$. These extensible regions alternate with short α helices which are crosslinked to other chains. Similar structures are present in silks (Box 2-B) and in proteins of wheat gluten.[217a,218b]

Besides hairpin turns and broader U-turns, a protein chain may turn out and fold back to reenter a β sheet from the opposite side. Such **crossover connections**, which are necessarily quite long, often contain helices. Like turns, crossover connections have a handedness and are nearly always right-handed (Fig. 2-25).[117,219] Most proteins also contain poorly organized loops on their surfaces. Despite their random appearance, these loops may be critical for the functioning of a protein.[220] In spite of the complexity of the folding patterns, peptide chains are rarely found to be knotted.[221]

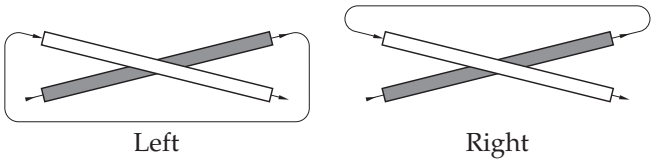



**Figure 2-25** Right- and left-handed crossover connections in proteins. These connections are nearly always right-handed. The broad arrows represent β strands. The crossover often contains a helix. Units of two adjacent β strands (βαβ units) with an α helix between are found frequently in globular proteins.

## 6. Domains, Subunits, and Interfaces

Many proteins are organized into tightly folded globular "domains" consisting of 50–150 amino acid residues.[117,222–227] Smaller proteins may have 2 or 3 domains but large proteins may have more. For example, the immunoglobulin "heavy chains" (Chapter 31) have four or five domains similar to that shown in Fig. 2-16B. The enormous 3000-kDa muscle protein **titin** contains 260 domains, many of which are of the immunoglobulin typeFig. 2-16,C.[228,117a] In most proteins the polypeptide chain folds to form one domain, then passes through a "hinge" to the next. In others, the C terminus or the N terminus of the polypeptide folds back across two or more domains as a kind of "strap" that helps to hold the protein together. Even when a protein contains only one domain, it often consists of two distinct lobes with a cleft between them. Many proteins, e.g. hemoglobin (Fig. 7-25), consist of subunits about the size of the globular domains in larger proteins.

Structural domains of proteins are sometimes encoded by a single coding segment of DNA i.e., by a single exon in a split gene. Domains of this type may have served as **evolutionarily mobile modules** that have spread to new proteins and multiplied during evolution. For example, the immunoglobulin structural domain is found not only in antibodies but also in a variety of cell surface proteins.[229–232]

Whether we deal with domains connected by a flexible hinge or with subunits, there are **interfaces** between the different parts of the protein. These interfaces are often formed largely of nonpolar groups. However, they frequently contain a small number of hydrogen bonds that bridge between one domain and another or between one subunit and another. In the case of hemoglobin, important changes occur in this hydrogen bonding and movement occurs along one of the interfaces between two subunits. Likewise the active sites of enzymes are often located at interfaces between domains. During catalysis, movement and reorganization of the hydrogen bonds and side chain packing in the interfaces may take place.[232a]

## 7. Packing of Side Chains

In Fig. 2-19B myoglobin is pictured as a cluster of α-helical rods surrounding the heme core. This picture is incomplete because the space between the rods and inside the molecule is tightly packed with amino acid side chains almost all of which are hydrophobic. The same is true for the β barrels of Fig. 2-16, which are filled largely with nonpolar side chains. As the structures of more and more proteins have been determined, a consistent pattern has emerged. Within the interior of proteins the side chain groups are packed together remarkably well.[99,100,222,233–236] Although occasional holes are present, they are often filled by water molecules.[99,237a] The **packing density**, the volume enclosed by the van der Waals envelope divided by the total volume, is ~0.75 for the interior of the lysozyme and ribonuclease molecules compared with the theoretical value of 0.74 for close-packed spheres. However, regions with many hydrogen bonds may be less tightly packed.[238]

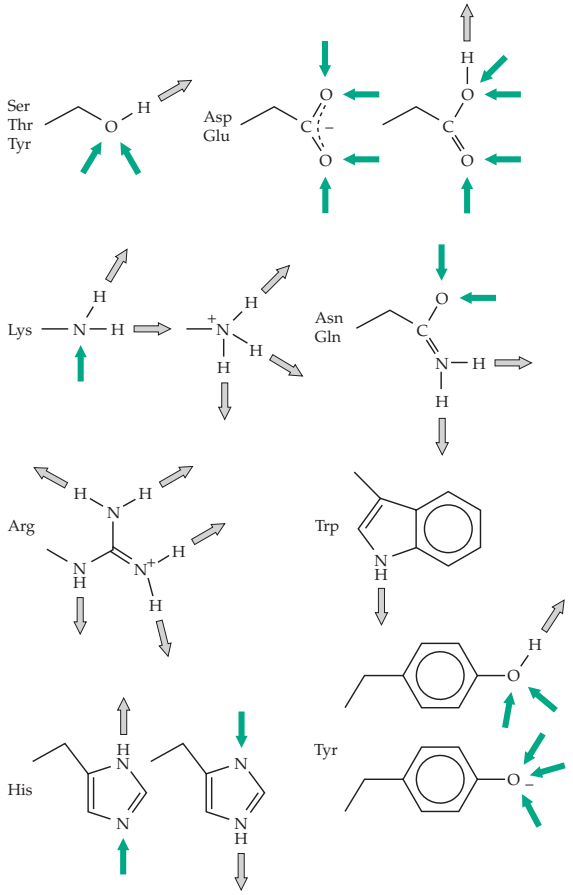



**Figure 2-26** Some of the possibilities for hydrogen bonding of side chain groups in proteins. Oxygen atoms can and frequently do form up to three hydrogen bonds at once. Open arrows point *from* H-atoms and *toward* electron donor pairs.

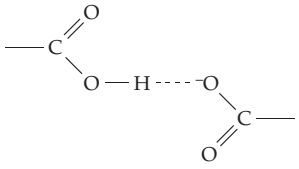
The interiors of proteins often contain large numbers of aromatic side chains which are frequently associated as pairs or as **aromatic clusters**.[239,240] For example, see the structure of plastocyanin in Fig. 2-16A. Rings may lie perpendicular one to another or be "stacked" face to face but offset. Oxygen atoms often lie in contact with the edges of aromatic rings.[241] The planar guanidinium groups of arginine side chains often stack against aromatic rings and amide groups may sometimes do the same.[242] It has been suggested that both aromatic: aromatic and aromatic:oxygen interactions may be associated with additional stabilization of the protein by ~4- to 8- kJ/mol. Tyrosine side chain –OH groups often stabilize ends of β strands by forming H-bonds to backbone atoms.[243]

Most polar groups are on the surfaces of proteins, and those that are not are almost always hydrogen bonded to other groups in the interior.[244] While most nonpolar groups are inside proteins, they are also present in the outer surfaces where they are often clustered into **hydrophobic regions** or "**patches**." The latter may be sites of interaction with other proteins or with lipid portions of membranes.

## 8. The Network of Internal Hydrogen Bonds

The fact that nonpolar groups tend to be buried in the interior of proteins suggests that the inside of a protein might be a flexible blob of oily material. In fact, the nonpolar groups tend to be densely packed and aromatic rings often impart considerable rigidity to the hydrophobic cores. Buried polar groups, which are invariably hydrogen bonded to other side chain groups, to amide groups of the polypeptide backbone,[244–247] or to buried water molecules, form a well-defined internal network. When a series of closely related proteins, for example, those having the same function in several different species, are compared, the hydrogen bonded network is often nearly invariant. This suggests a functional significance. The hydrogen bonding possibilities for some of the side chain groups in proteins are indicated in Fig. 2-26.

Charged side chains are sometimes buried in the interior of proteins, usually together with an ion of the opposite charge to form a **hydrogen-bonded ion pair**.[248,248a,248b] However, there is sometimes a single charge that is "neutralized" by the interaction with dipoles of polar groups.[217] Sometimes an undissociated carboxyl group is hydrogen bonded to a carboxylate ion.[249]



A buried carboxyl group of this type will display a $pK_a$ value far higher than the normal value.

Arginine side chains are large and able to form multiple hydrogen bonds[247,250] as well as salt linkages[251] to different parts of a folded peptide. Cations of arginine and lysine are never buried unless in ion pairs. Protons from strong acids have long been known to bind to centers of benzene rings and water can form weak hydrogen bonds to the centers of aromatic rings.[252,253,253a] Such bonds also occur within proteins and often involve the binding of guanidinium groups or of inorganic cations to indole rings of tryptophans.[253b,253c] Protonated imidazole groups may also bind to aromatic rings.[253d]


Another important aspect of the structure of proteins is the presence of **hydrogen-bonded water molecules** in pockets and cracks. These molecules, as well as a much larger number of water molecules bound at the outer surface, are clearly visible from X-ray studies. They often occur singly, bonded to the ends of amide groups, especially to the carbonyl ends. Internal bends of the peptide chain are almost always hydrated.[254] These water molecules often make two or more hydrogen bonds to different parts of the protein or to other water molecules. Clusters of water molecules,[49,255,255a] sometimes in the form of pentagonal rings,[256] are often present. NMR spectroscopy has shown that water molecules bound to protein *surfaces* exchange rapidly with the bulk water in which the protein is dissolved; the "residence time" on the protein is typically less than a nanosecond. *Interior* water molecules have much longer residence times of $10^{-2}$ to $10^{-8}$ s for a small protein.[257] They may be regarded as part of the protein structure.

## E. Folding Patterns and Protein Families

Proteins are folded in many ways. We have already considered several simple patterns: the antiparallel β cylinder (Fig. 2-16), the 2-helix coiled coil (Fig. 2-21) and the 3- and 4-helix bundles (Fig. 2-22). Another simple motif that has been found repeatedly is the **helix–turn–helix** or **helix–loop–helix** in which two helices at variable angles, one to another and with a turn or short loop between them, form a structural unit. DNA-binding **repressors** and **transcription factors** (see Fig. 2-21 and also Chapter 5) often contain this motif as do many $Ca^{2+}$-binding proteins. Proteins containing 3–6 helical segments, often fold into a roughly polyhedral shape.[258,259] An example is myoglobin (Fig. 2-19B).

### 1. Complex Folding Patterns

Proteins often contain elements of both α and β structure. One of the first of the complex folding motifs to be recognized was a nucleotide-binding domain identified by Rossmann and associates.[260–262] This Rossman fold contains six parallel β strands which alternate with six helices. The result of the parallel β structure is that the helices are also parallel and that their amino-terminal ends, which carry partial positive charges, are aligned in approximately the same directions. The positive end of one of the helices lies behind the negatively charged phosphate groups[148] which characteristically bind at the edge of the sheet containing the C termini of the β strands. This can be seen in Fig. 2-13, which shows binding of the coenzyme NAD to the nucleotide binding domain of **glyceraldehyde phosphate dehydrogenase**. Similar nucleotide-binding domains are found in many other dehydrogenases whose members constitute a **protein family**.

Figure 2-27 depicts **topology diagrams** for the Rossman fold and for two related families of proteins. These families bind the nucleotides called GTP and ATP, respectively. Both are structural relatives of NAD. A major part of the structure of all of the proteins in these families consist of β–α units, each one containing a β strand followed by a helix. They are

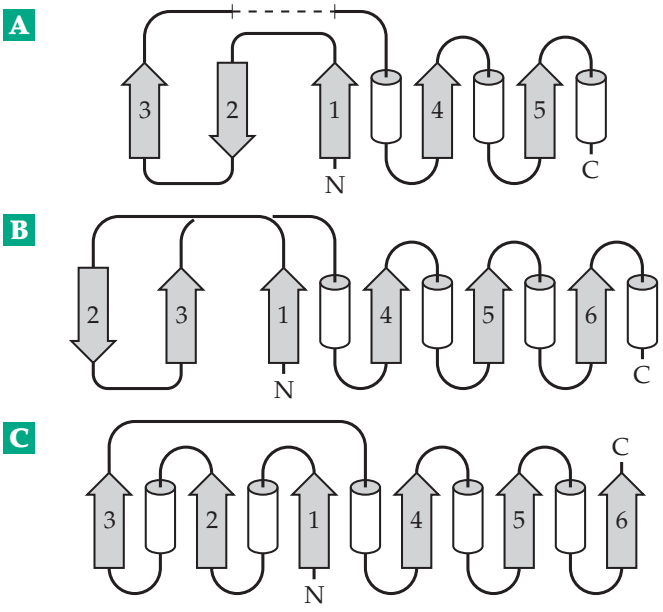



**Figure 2-27** Topologies of the folds of three families of nucleotide binding α/β proteins. Cylinders represent α helices and arrows β strands. (A) The ATPase fold for the clathrin-uncoating ATPase; (B) The G-protein fold that binds GTP and is found in *ras* proteins; (C) The Rossmann fold that binds NAD in several dehydrogenases. From Brändén.[262]


**Figure 2-28** The eight-fold α/β barrel structure of triose phosphate isomerase. From Richardson. (A) Stereoscopic view. (B) Ribbon drawing. Courtesy of Jane Richardson.[117]

classified as α/β proteins. The Rossman fold is composed of six β–α units. Recently a ribonuclease inhibitor protein with 15 consecutive β–α units has been characterized.[263] Each β–α unit contains several residues of leucine. This **leucine-rich repeat** occurs in many other proteins as well.[264,264a,264b]

The **α/β barrel** shown in Fig. 2-28 consists of 8 consecutive β–α units in a symmetric array.[265,266] By 1995 over 40 of these barrels had been identified in a diverse group of enzymes. One bifunctional enzyme contains two α/β barrels. Although the nature of the reaction catalyzed varies, the active site is always found in the center of the barrel at the C-terminal ends of the 8 parallel β strands and therefore between the N termini of the surrounding helices. The enzyme sequences show no homology and frequent occurrence of the 8-stranded barrel may reflect the fact that it is a natural packing arrangement of β–α units. However, a 10-stranded barrel of this type has also been found.[267]

This barrel can be compared with that of the 12-helix α,α barrel of a fungal glucoamylase whose structure is shown in Fig. 2-29. Numerous more complex folding patterns have been discovered. They have been classified by Jane Richardson.[117,122] Many of the proteins described by these folding patterns can be grouped into families and "superfamilies."[227] Chothia suggested that there may be about 1,000 families in nature;[268] over 700, with over 360 distinct folds have been identified.[268a]

## 2. Symmetry

A sometimes puzzling feature of protein structure is the widespread occurrence of an approximate two-fold axis of symmetry. This often arises as a natural result of association of a pair of irregular subunits (Chapter 7). The association is such that rotation

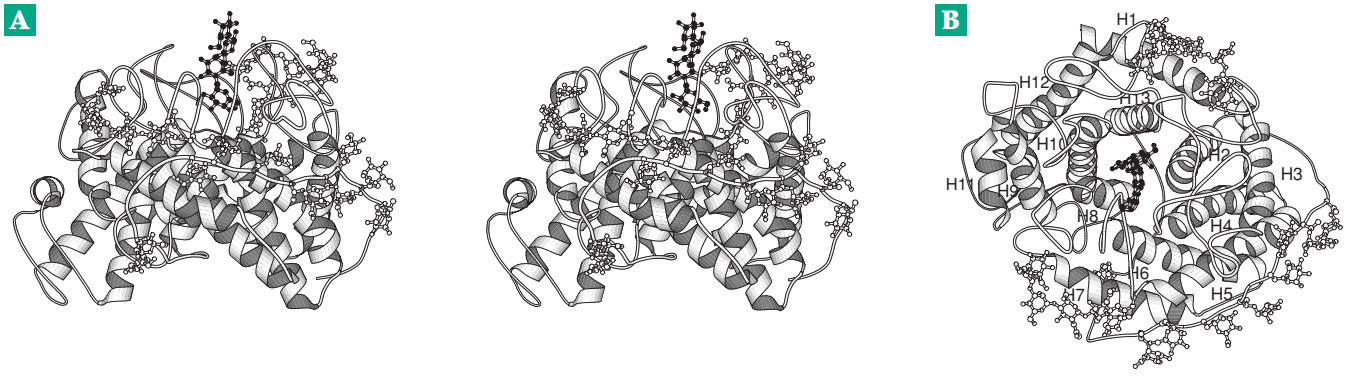



**Figure 2-29** Structure of the α,α barrel of a fungal enzyme glucoamylase. (A) side view (stereoscopic); (B) top view. The active site, which cleaves glucose units from the ends of starch chains, is in the depression in the center of the barrel. Here it is occupied by an inhibitor. See Aleshin *et al*.[192] Courtesy of Alexander Aleshin.

about the twofold axis will cause the two subunits to exchange positions and to remain in an identical chemical environment. Approximate symmetry is often observed also *within* single peptide chains. For example, in the Rossman fold (Figs. 2-13, 2-27), an approximate twofold axis passes between the center strands of the β sheet residues and relates the two flanking helices, which begin with residues R10 and T100, respectively. The bound $NAD^+$ also possesses an approximate twofold axis, but it is not quite symmetrically placed at the end of the β sheet. Both phospho groups are seen to interact with the N terminus of the helix beginning at residue 10. The small bacterial protein ferredoxin (Fig. 16-16B) contains two iron-sulfur clusters related by an approximate 2-fold axis. The two β cylinders of elastase (Fig. 12-9) as well as the two sides of the flattened β barrel of copper–zinc superoxide dismutase[269] are approximately related by twofold axes. The enzyme thiosulfate: cyanide sulfurtransferase (Eq. 24-46) is remarkably symmetric but the active site is located in just one half. The widespread existence of this approximate symmetry suggests a biological significance that remains to be discovered.

## 3. Effects of Sequence on Folding

Studies of synthetic polypeptides as well as examination of known protein structures reveal that some amino acids, e.g., Glu, Ala, Leu, tend to promote α helix formation. Others, such as Tyr, Val, and Ile, are more often present in β structure, while Gly, Pro, and Asn are likely to be found in bends.[270,270a,270b] The frequencies with which particular amino acids appear in helices, β structure, or turns were first compiled by

**TABLE 2-4**
**Classification of Protein Residues According to Their Tendencies to Form α Helix, β Structure, and β Turns[a]**

| Amino acid | $P_\alpha$ | Helix-forming tendency | $P_\beta$ | β structure-forming tendency | $P_t$ |
|---|---|---|---|---|---|
| $Glu^-$ | 1.51 | ++ | 0.37 | br+ | 0.44 |
| Met | 1.45 | ++ | 1.05 | + | 0.67 |
| Ala | 1.42 | ++ | 0.83 | i | 0.57 |
| Leu | 1.21 | ++ | 1.30 | + | 0.53 |
| Lys+ | 1.16 | + | 0.74 | br | 1.01 |
| Phe | 1.13 | + | 1.38 | + | 0.71 |
| Gln | 1.11 | + | 1.10 | + | 0.56 |
| Trp | 1.08 | + | 1.37 | + | 1.11 |
| Ile | 1.08 | + | 1.60 | ++ | 0.58 |
| Val | 1.06 | + | 1.70 | ++ | 0.30 |
| $Asp^-$ | 1.01 | w | 0.54 | br+ | 1.26 |
| $His^+$ | 1.00 | w | 0.87 | i | 0.69 |
| $Arg^+$ | 0.98 | i | 0.93 | i | 1.00 |
| Thr | 0.83 | i | 1.19 | + | 1.00 |
| Ser | 0.77 | i | 0.75 | br | 1.56 |
| Cys | 0.70 | i | 1.19 | + | 1.17 |
| Tyr | 0.69 | br | 1.47 | ++ | 1.25 |
| Asn | 0.67 | br | 0.89 | i | 1.68 |
| Pro | 0.57 | br+ | 0.55 | br+ | 1.54 |
| Gly | 0.57 | br+ | 0.75 | br | 1.68 |

++ = strong former; + = former; w = weak former; i = indifferent; br = breaker; br+ = strong breaker

[a] The conformational parameters $P_\alpha$, $P_\beta$, and $P_t$ (β turn) are the frequencies of finding a particular amino acid in an α helix, β structure, or β turn (in 29 proteins of known structure) divided by the average frequency of residues in those regions. Residues are arranged in order of decreasing tendency toward helix formation. From Chou, P. V. and Fasman, G. D (1974) *Biochemistry* **13**, 222–245.

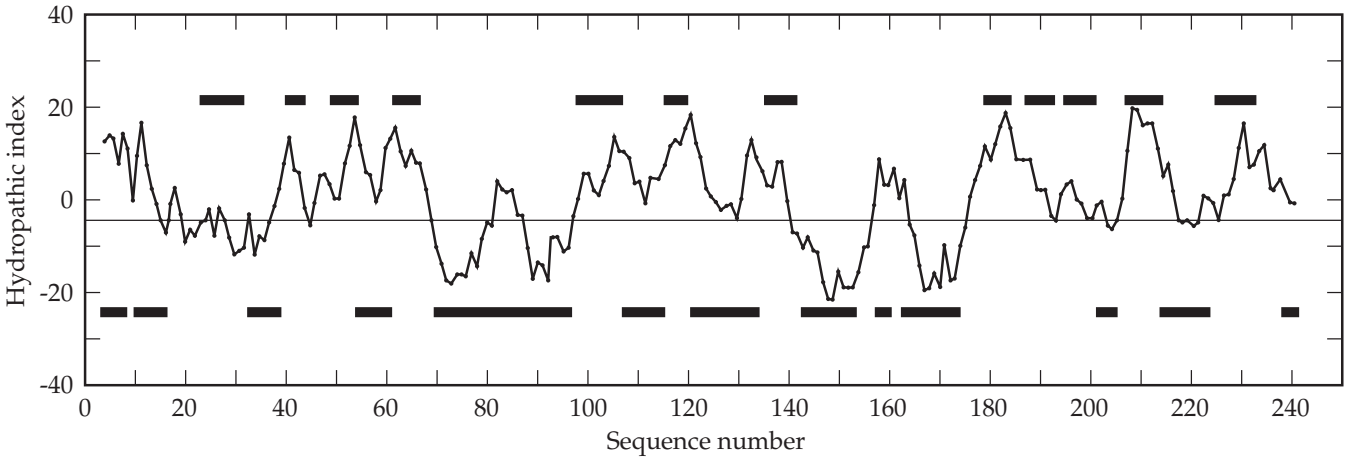



**Figure 2-30** Plot of hydropathy index versus sequence number for bovine chymotrypsinogen. The indices for individual residues have been averaged nine at a time. The solid bars at the top of the plot mark interior regions as determined by crystallography. The solid bars below the plot indicate regions that are on the outside of the molecule. From Kyte and Doolittle.[280]

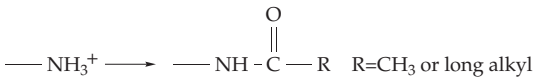
Chou and Fasman[270] (Table 2-4). These frequencies also differ significantly between parallel and antiparallel β sheets. Some of the preferences are readily understandable. For example, glycine is too flexible to stabilize a helix, but it can occur in helices between residues that are better helix formers. Because the peptide linkage to its nitrogen lacks an NH group, proline can fit into only one of three positions at the N terminus of a helix. For the strong helix formers, the β-$CH_2$ groups fit into the helix well and provide stabilization through van der Waals interactions. However, the β-methyl groups of valine and isoleucine cause crowding. For these residues, a β structure is more favorable than a helix.[271] Side chains of Asp, Asn, Ser, Thr, and Cys can hydrogen bond to backbone amide groups and can either stabilize or destabilize a helix or β sheet depending upon their location.

It appears that the folding pattern of a peptide is encoded in the sequence itself.[272–274] Thus, when several residues that favor helix formation are clustered together, a helix may form. Chou and Fasman suggested that when four helix formers out of six residues are clustered, nucleation of a helix takes place. The helix can then be elongated in both directions until terminated by a proline or other "helix breaker"; additional folding can then occur. If three out of five β formers are clustered, a β strand may form. If random folding brings two or more of these strands together they may associate to form the nucleus for a β sheet. Some success has been achieved using this approach in predicting whether a given residue will be found in a helix, a β strand or a loop.[275] However, prediction of complete folding patterns is much more difficult. Many new approaches are being explored.[275–279d] The problem is an important one. Although three-dimensional structures have been determined for thousands of proteins, sequences are known for hundreds of thousands. The number is growing rapidly. Being able to predict correctly a three-dimensional protein structure will be a major scientific accomplishment with many practical consequences.

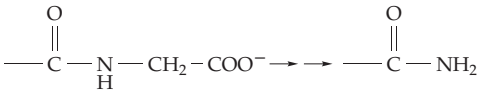
A way of examining the entire sequence for polar or nonpolar character of the side chains was introduced by Kyte and Doolittle.[280,281] A **hydropathy index** based on the polarity of the side chains of a given residue and of its neighbors in the sequence is plotted against residue number (Fig. 2-30). Helices are often found to have an **amphipathic** character, hydrophobic on one side and hydrophilic on the other. Such helices can be characterized by a plot such as that in Fig. 2-30. The hydrophobic side of an amphipathic helix can pack against a hydrophobic core of a protein, can lie against a membrane, or can be aligned with other helices to form coiled coils (Fig. 2-21) or to give a hydrophilic channel with an outer hydrophobic surface that can fit into a cell membrane.

## F. Chemical Modification and Crosslinking

Some newly synthesized proteins, upon folding, are ready to "go to work" immediately but others must be modified. Pieces are frequently cut from the ends or out of the center of a folded peptide chain. Sometimes the amino and carboxyl termini are converted to nonionic groups, e.g.,



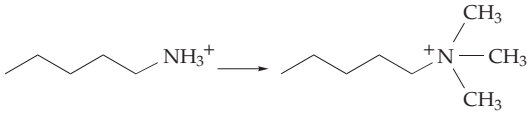
(2-12)



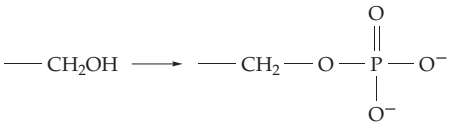
(2-13)

Side chains may be modified in a very large number of different ways.[282–284] These include acetylation and other kinds of acylation (Eq. 2-12),[285–287] **methylation** (Eqs. 2-14, 2-15), **phosphorylation** (Eq. 2-16), **phosphoadenylation**,[288] formation of **sulfate esters** (Eq. 2-17),[289,290] and hydrolysis (Eqs. 2-18, 2-19).[291] In at least a few proteins some L-amino acid residues are converted to D-.[292–294]

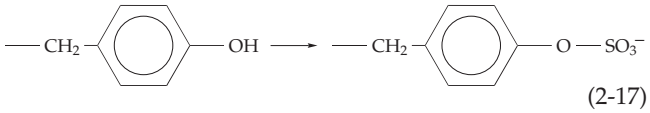
(2-14)



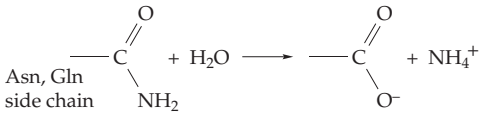
(2-15)



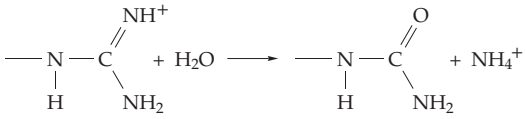
(2-16)



(2-17)

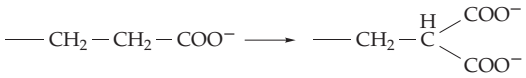


(2-18)



(2-19)

Usually, only one or a small number of side chains of a protein is modified. However, glycoproteins may contain many different attached sugar molecules. An example is the glucoamylase shown in Fig. 2-29. In several proteins involved in blood clotting (Fig. 12-17), as many as 10 glutamic acid side chains are carboxylated (Eq. 2-20).

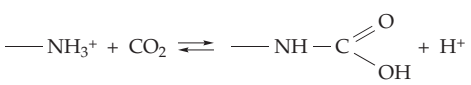



(2-20)

The 216-residue hen egg yolk storage protein **phosvitin** contains 123 serine residues, most of which have been phosphorylated (Eq. 2-16).[295] A basic protein of the myelin sheath of neurons contains as many as 6 specific residues of **citrulline** (Eq. 2-19).[296] An adhesive protein from the foot of a marine mollusk contains ~80 repeated sequences containing hydroxy-proline 2,3-dihydroxyproline and **3,4-dihydroxyphenylalanine** (Dopa).[297,298]

Some modification reactions alter the electrical charge on a side chain and, as a consequence, can affect the ways in which the protein interacts with other molecules. Negative charges added in formation of clusters of γ-carboxyglutamates (Eq. 2-20) create strong **calcium ion-binding centers** in the modified proteins. Acylation of N termini or of Ser, Thr, or Cys side chains by long-chain fatty acids provides hydrophobic "tails" able to anchor proteins to membrane surfaces.[285,286] Addition of polyprenyl groups to cysteine side chains near the C-termini has a similar effect.[299,300] These and more complex membrane anchors are considered in Chapter 8. While modified amino acids can be found at many places in a protein, they are often located at turns. For example, serine or threonine residues in turns are often phosphorylated or glycosylated. Modification of proteins is dealt with further in Chapter 10 and at other points in the book. Glycosylation of proteins is considered in Chapter 20.

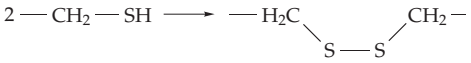
In addition to deliberate enzyme-catalyzed processes, there are nonenzymatic processes that alter proteins. These include the degradative reactions described in Section 5 and also *reversible* reactions that may be physiologically important. For example, the N-terminal amino groups of peptides, and other amino groups of low p$K_a$ can form **carbamates** with bicarbonate (Eq. 2-21).[301–303] This provides an important mechanism of carbon dioxide transport in red blood cells (Chapter 7) and a way by which $CO_2$ pressure can control some metabolic processes.




(2-21)

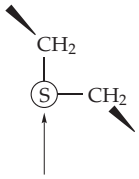
## 1. Disulfide Crosslinkages

The linking together of two different parts of a peptide chain or of two different peptide chains is extremely important to living beings.[304,305] One of the most widespread of crosslinkages is the **disulfide bridge**. It forms spontaneously when two –SH groups of cysteine side chains are close together and are oxidized by $O_2$ or some other reagent (Eq. 2-22).




(2-22)

The disulfide group is inherently chiral, with the preferred torsion angle about the S–S bond being ~90°. Both right- and left-handed disulfide bridges occur in proteins.[305,306]

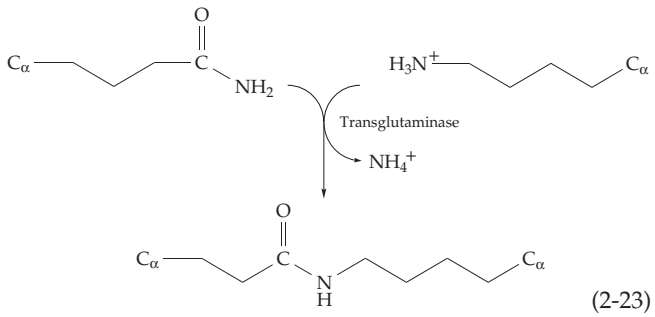



Two S atoms viewed down S—S axis

Disulfide linkages are frequently present in proteins that are secreted from cells but are less common in enzymes that stay within cells. Perhaps because the latter are in a protected environment, the additional stabilization provided by disulfide bridges is not needed. Disulfide bridges are not only found within single-subunit proteins but they also link different peptide chains. For example, the four chains of each immunoglobulin molecule (Chapter 31) are joined by disulfide bridges and each domain within the chains is stabilized by a disulfide bridge. One of the most highly crosslinked proteins known occurs in the keratin matrix of hair. Breakage of the –S–S– linkages of this protein is an essential step in the chemical "permanent wave" process. A thiol compound is used to reductively cleave the crosslinks and after the hair is reset new crosslinks are formed by air oxidation.

## 2. Other Crosslinkages

Another common crosslink is an amide formed between the γ-carboxyl group of a glutamic acid side chain and an amino group from a lysine residue.[307] This **isopeptide** linkage is formed from a residue of gluta-mine through the action of the enzyme **transglutaminase** (Eq. 2-23). Isopeptide crosslinks are found in hair, skin, connective tissue, and blood clots.




(2-23)

Occasionally an isopeptide linkage joins amino groups of lysine side chains to the C-terminal carboxyl groups of other peptide chains to give branched chains, e.g. see **ubiquitin** (Box 10-C). Elaborate crosslinks derived from lysine are found in collagen and elastin (Chapter 8). Dityrosine linkages formed by oxidative joining of the aromatic rings of tyrosine are found in insect cuticle and in the plant cell wall extensins[202] (Chapters 20, 25).

## G. Dynamic Properties of Proteins

Sometimes using energy from the cleavage of chemical bonds, sometimes depending only upon energy provided by the random bombardment by solvent molecules, proteins perform their specific functions with amazing speed. A question that has long intrigued biochemists is "to what extent do proteins stretch or flex, unfold and refold, or undergo other conformational movements during their action?" To answer this and related questions, many techniques are being applied to study the dynamic properties of proteins.[308,309]

### 1. Motion of Backbone and Side Chains

Even in the *crystalline state* there is evidence of movement. In the images constructed from X-ray or neutron diffraction experiments side chains on the surfaces of protein molecules are often not clearly visible because of rapid rotational movement. Some segments of the polypeptide chain may be missing from the image. However, side chain groups within the core of a domain are usually seen clearly. They probably move only in discrete steps. However, they may sometimes shift rapidly between different conformations, all of which maintain a close-packed interior.[310–312]

Studies of nuclear magnetic resonance spectra (Chapter 3) and of polarization of fluorescence (Chapter 23), have shown that there is rapid though restricted rotational movement of side chains of proteins *in solution*. Even buried phenylalanine and tyrosine side chains often rotate rapidly whereas movement of the bulkier tryptophan rings is more limited. Peptide NH protons in unfolded polypeptide chains undergo rapid acid or base catalyzed exchange with $H_2O$.[313] However, in globular proteins, the rate of exchange of buried NH protons is often orders of magnitude slower.[314–316] For most proteins there appears to be little tendency to unfold completely and then refold; the major domains hold together tightly. However, there may be local unfolding, for example, of a helix at the surface of a protein, which will allow more rapid exchange. Using NMR techniques (Chapter 3), rates of exchange of all of the individual peptide NH protons within small proteins can be measured.[315] Cracks may open up in proteins. This is suggested by the fact that $O_2$, $I_2$, and certain other small molecules are apparently able to penetrate the protein freely and to quench the fluorescence of buried tryptophan side chains (Chapter 23).

Since packing density tends to be lower at active sites than in the bulk of the protein, it is probable that more conformational alterations occur near active sites than elsewhere.[317] Lumry and Rosenberg[318] suggested that the "defects" of poor packing and poor hydrogen bonding of some regions of a folded peptide chain provide a store of potential energy that can be important to the functioning of a protein. Even in a very well-packed protein domain there are defects. Some atoms are compressed by the folding of the peptide chain and are closer together by over 0.04 nm than predicted by the van der Waals radii. These packing defects have been estimated to destablize the protein by as much as 250 kJ/mol.[319] Details of structural heterogeneity within several proteins for which very precise structural data are available have been described.[311]

### 2. Conformational Changes

We have seen that some polypeptides assume an extended β conformation while others form helices. In some cases, the same protein can do both. For example, hair can be stretched greatly, the α helices of the keratin molecules uncoiling and assuming a β conformation with hydrogen bonds *between* chains instead of *within* a single chain. Thus, a polymer may have more than one conformational state in which the folding and hydrogen bonding are different.[320,321] With soluble proteins more than one folded conformation is possible with different sets of hydrogen bonds and internal hydrophobic interactions. Some of the conformations of a globular protein are more stable than others, and a protein will ordinarily assume one of the energetically most favorable conformations. However, there may be other conformations of almost equal energy.

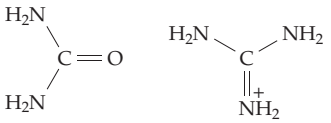
A large body of evidence suggests that many proteins do exist in two or more different but well-defined

conformational states and that the ability of a protein to undergo easy conversion from one to another is of great biological significance. In some cases such as that of hemoglobin (Chapter 7), there are changes in the interactions between subunits. Alterations in the hinge regions between domains of immunoglobulins have been seen. In many enzymes, including the dehydrogenases (Chapter 15), kinases (Chapter 12), and aspartate aminotransferase (Fig. 2-6), a cleft between two domains appears to open and close.[322] In other cases, more subtle alterations in conformational state, involving mainly changes in the internal hydrogen bonded network together with small localized changes in chain folding, have been observed.

Conformational alterations in proteins are probably facilitated by the fact that some hydrogen-bonded groups are found within the hydrophobic interior. *All of the buried hydrogen atoms suitable for hydrogen bond formation are ordinarily hydrogen bonded to an electron donor group. However, because oxygen atoms in proteins each have two unshared electron pairs, there are, in general, more electron donor groups than there are hydrogen atoms to which they can bind. This sets the stage for a competition between electronegative centers for particular proton suitable for hydrogen bonding and provides a molecular basis for the easy triggering of conformational changes.*[323]

## 3. Denaturation and Refolding

An extreme conformational alteration is the **denaturation** of proteins, which may be caused by heating or by treatment with reagents such as strong acids and bases, **urea**, **guanidinium chloride**, and **sodium dodecyl sulfate** (SDS).




Urea — Guanidinium chloride

$CH_3-(CH_2)_{11}-OSO_3^- Na^+$

Sodium dodecyl sulfate (SDS)

Denaturation leads to unfolding of a protein to a more random conformation. In the denatured state the amide groups of the peptide chain form hydrogen bonds with surrounding water molecules and with denaturants such as urea or the guanidinium ion rather than with each other.[324,325] Denaturants also diminish the strength of the hydrophobic interactions that promote folding.[326] Characteristic biochemical activities are lost and physical properties such as sedimentation constant, viscosity, and light absorption are altered. The ease of denaturation of proteins and the fact that denaturation is sometimes reversible show that the energy differences between the folded conformations and the open random coil conformation are usually not great.[274,327,328] However, it has been difficult to establish the amount of stabilization of a folded polypeptide chain provided by buried hydrogen bonds[328a,328b,328c] or the role of cooperatively formed hydrogen-bonded chains[328c,328d] or clusters.[328c,328d,328e]

Complete denaturation of a protein was generally regarded as an irreversible process prior to 1956 when Anfinsen showed that denatured ribonuclease (Chapter 12) could refold spontaneously.[273] This 124-residue protein contains four disulfide (–S–S–) bridges and thus is tied firmly together. When these bridges are broken reductively in the presence of a denaturing agent, the enzyme becomes inactive. Anfinsen found that upon reoxidation under appropriate conditions, complete activity reappeared. The molecules had folded spontaneously into the correct conformation, the one in which the correct *one* of 105 (7 x 5 x 3 x 1) possible pairings of the eight –SH groups present needed to reform the four disulfide bridges had taken place. This observation has had an important influence on thinking about protein synthesis and folding of polypeptide chains into biologically active molecules.

A puzzling problem was posed by Levinthal many years ago.[329] We usually assume that the peptide chain folds into one of the most stable conformations possible. However, proteins fold very rapidly. Even today, no computer would be able, in our lifetime, to find by systematic examination the thermodynamically most stable conformation.[328] It would likewise be impossible for a folding protein to "try out" more than a tiny fraction of all possible conformations. Yet folded and unfolded proteins often appear to be in a thermodynamic equilibrium! Experimental results indicate that denatured proteins are frequently in equilibrium with a **compact denatured state** or "molten globule" in which hydrophobic groups have become clustered and some secondary structures exists.[330–336] From this state the polypeptide may rearrange more slowly through other **folding intermediates** to the final "native conformation."[336a,336b]

It is generally assumed that within cells the folding of the peptide chain commences while the chain is still being synthesized on a ribosome. The growing chain probably folds rapidly in a random way until it finds one or more stable conformations which serve as folding intermediates for slower conversion to the finished protein.[337,338] However, any process within a cell is affected by the complex intracellular environment.[339] Folding can be catalyzed or inhibited by proteins known as **molecular chaperones**. Folding may sometimes require isomerization of one or a few proline residues from *trans* to *cis*.[340,341] For example, during the refolding of ribonuclease the isomerization of Pro 93 appears to be a rate-limiting step.[342,343] Such

isomerizations may be assisted by **peptidyl-prolyl-(*cis–trans*) isomerases** (Box 9-F). Disulfide linkages are sometimes formed incorrectly.[344,345] A **protein-disulfide isomerase** catalyzes cleavage and reformation of these bridges, allowing the protein to find the most stable crosslinking arrangement. The actions of these enzymes and of molecular chaperones are considered further in Chapters 10 and 12.

### 4. Effects of pH and Solvent

Because polypeptide chains contain many acidic and basic amino acid side chains, the properties of proteins are greatly influenced by pH. At low pH carboxylates, $-S^-$, and imidazole groups accept protons causing the overall net charge on the macromolecule to be strongly positive. At high pH protons are lost and the protein becomes negatively charged. Electrostatic repulsion between like charges may cause proteins to denature at low or high pH. More stable proteins may be very soluble at low or high pH. Proteins often have a minimum solubility and a maximum stability near the **isoelectric point,** the pH at which the net charge is zero.[336,346,347] Activities of enzymes, abilities to bind specifically to other proteins, and

## BOX 2-C THE NOBEL PRIZES

Many young scientists dream of one day winning a Nobel Prize. Although denounced by some, the much sought and highly publicized award has, since 1901, been given to an outstanding group of scientists. Many of these have made major contributions to biochemistry or to techniques important to biochemists. Here is a partial list.

| Year | Name | Prize[a] | Discovery or subject of study |
|---|---|---|---|
| 1901 | Wilhelm C. Röntgen | Physics | Discovery of X-rays |
| 1902 | Emil H. Fischer | Chemistry | Synthesis of sugars and purines |
| 1903 | Svante A. Arrhenius | Chemistry | Electrolytic dissociation; a founder of physical chemistry |
| 1903 | Antoine Henri Becquerel; Marie S. Curie, and Pierre Curie | Physics | Discovery and study of radioactivity |
| 1906 | Camillo Golgi and S. Ramon y Cajal | Physiology/Medicine | Discovery of Golgi apparatus |
| 1907 | Edward Buchner | Chemistry | Biochemistry, cell-free fermentation |
| 1910 | Albrecht Kossel | Physiology/Medicine | Isolation of nucleic acid bases |
| 1914 | Max von Laue | Physics | Discovery of X-ray diffraction by crystals |
| 1915 | Richard M. Willstätter | Chemistry | Plant pigments, especially chlorophyll |
| 1915 | William H. and William L. Bragg | Physics | Analysis of crystal structure by X-rays |
| 1919 | Jules Bordet | Physiology/Medicine | Discovery of blood complement; complement fixation test |
| 1920 | Walther H. Nernst | Chemistry | Thermochemistry |
| 1922 | Archibald V. Hill and Otto F. Meyerhof | Physiology/Medicine | Chemistry of muscle contraction |
| 1923 | Frederick G. Banting and John J. R. Macleod | Physiology/Medicine | Discovery of insulin and treatment of diabetes |
| 1926 | Theodor Svedburg | Chemistry | Study of high $M_r$ compounds, development of ultracentrifuge |
| 1927 | Heinrich O. Wieland | Chemistry | Bile acids |
| 1928 | Adolf O. R. Windaus | Chemistry | Sterols and vitamins |
| 1929 | Frederick G. Hopkins and Christiaan Eijkman | Physiology/Medicine | Discovery of vitamins, tryptophan, vitamin $B_1$ |
| 1929 | Arthur Harden and Hans A. S. von Euler-Chelpin | Chemistry | Fermenting enzymes, fermentation of sugars |
| 1930 | Karl Landsteiner | Physiology/Medicine | Blood groups A, B, O, Rh |
| 1930 | Hans Fischer | Chemistry | Structures and chemistry of porphyrins, chlorophyll |
| 1931 | Otto H. Warburg | Physiology/Medicine | Respiratory enzymes |
| 1933 | Thomas H. Morgan | Physiology/Medicine | Chromosome theory and chromosome maps |
| 1934 | George R. Minot, William P. Murphy, and George H. Whipple | Physiology/Medicine | Treatment of pernicious anemia |
| 1936 | Henry H. Dale and Otto Loewi | Physiology/Medicine | Acetylcholine release at nerve endings |
| 1937 | Albert von Szent-Györgyi | Physiology/Medicine | Vitamin C |
| 1937 | Walter N. Haworth and Paul Karrer | Chemistry | Carbohydrate structures, structures of carotenoids, flavins, vitamin A |
| 1938 | Richard Kuhn | Chemistry | Carotenoids and vitamins |
| 1939 | Gerhard Domagk | Physiology/Medicine | Prontosil, first antibacterial sulfa drug |
| 1939 | Adolf F. J. Butenandt and L. Ruzicka | Chemistry | Isolation and study of sex hormones, study of polymethylenes, terpenes |
| 1943 | E. A. Doisy and Carl Henrik Dam | Physiology/Medicine | Isolation and study of vitamin K |
| 1945 | Alexander Fleming, Ernst B. Chain, and Howard W. Florey | Physiology/Medicine | Discovery and structure of penicillin |
| 1945 | A. J. Virtanen | Chemistry | Nutritional chemistry |
| 1946 | James B. Sumner, J. H. Northrop, and W. M. Stanley | Chemistry | Crystallization of enzymes and virus proteins |
| 1947 | Carl F. Cori and Gerty T. Cori | Physiology/Medicine | Glycogen metabolism, the Cori cycle |
| 1947 | B. A. Houssay and Robert Robinson | Chemistry | Investigation of plant alkaloids |
| 1948 | Arne W. K. Tiselius | Chemistry | Electrophoresis, study of serum proteins |
| 1950 | Phillip S. Hench, Edward C. Kendall, and Tadeus Reichstein | Physiology/Medicine | ACTH |
| 1952 | Selman A. Waksman | Physiology/Medicine | Discovery of streptomycin |
| 1952 | A. J. P. Martin and Richard L. M. Synge | Chemistry | Paper chromatography |
| 1953 | Fritz A. Lipmann and Hans A. Krebs | Physiology/Medicine | Coenzyme A, citric acid cycle |
| 1954 | Linus C. Pauling[b] | Chemistry | The nature of the chemical bond |
| 1955 | A. H. T. Theorell | Physiology/Medicine | Oxidative enzymes |
| 1955 | Vincent du Vigneaud | Chemistry | Synthesis of biotin and oxytocin |
| 1957 | Daniel Bovet | Physiology/Medicine | First antihistamine drug |
| 1957 | Alexander R. Todd | Chemistry | Work on nucleotides coenzymes |
| 1958 | Joshua Lederberg, George W. Beadle, and Edward L. Tatum | Physiology/Medicine | One gene-one enzyme hypothesis from genetic studies of neurospora |
| 1958 | Frederick Sanger[b] | Chemistry | Protein sequencing, insulin |

numerous physiological processes are controlled by effects of pH on proteins.

The solvent for most proteins in nature is water. However, many enzymes function well in organic solvents if they retain only a small amount of essential structural and catalytic water.[348]

## 5. Irreversible Damage to Proteins

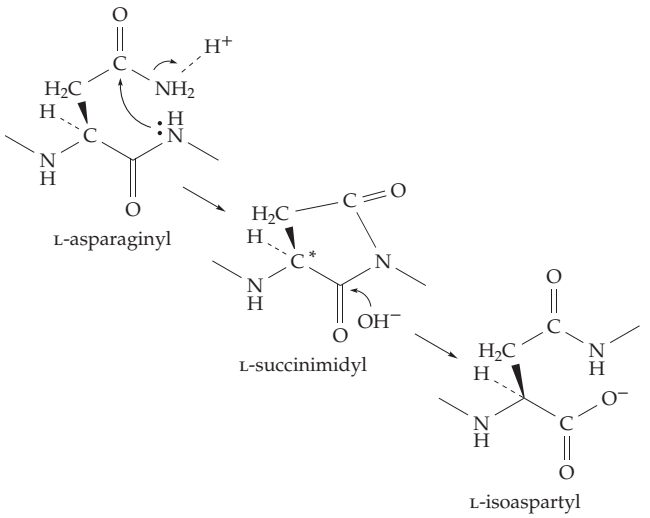
Every protein in an organism has its own characteristic lifetime. No sooner is it synthesized than degradation begins. Enzyme-catalyzed cleavage of the peptide linkages leads to turnover of proteins but before this occurs a number of spontaneous damaging reactions may alter the protein. Prevalent among these is **deamidation** of asparaginyl residues to aspartyl or

## BOX 2-C THE NOBEL PRIZES (continued)

| Year | Name | Prize[a] | Discovery or subject of study |
|---|---|---|---|
| 1959 | Arthur Kornberg and Severo Ochoa | Physiology/Medicine | Enzymatic synthesis of DNA |
| 1960 | F. M. Burnet and P. Medawar | Physiology/Medicine | Immunological tolerance in animals |
| 1961 | Melvin Calvin | Chemistry | Photosynthesis using $^{14}CO_2$ |
| 1962 | Francis H. C. Crick, James D. Watson, and Maurice H. F. Wilkins | Physiology/Medicine | Molecular structure of DNA |
| 1962 | John C. Kendrew and Max F. Perutz | Chemistry | Structures of protein by X-ray diffraction |
| 1962 | Linus C. Pauling[b] | Peace | Ending atmospheric testing of nuclear weapons |
| 1964 | Konrod E. Bloch and Feodor Lynen | Physiology/Medicine | Pathways of cholesterol biosynthesis |
| 1964 | Dorothy M. Crowfoot Hodgkin | Chemistry | X-ray structures, vitamin $B_{12}$ |
| 1965 | Andre Lwolf, Jacques Monod, and Francois Jacob | Physiology/Medicine | Messenger RNA, regulation of transcription |
| 1967 | Ragnar Granit, Haldan Keffer Hartline, and George Wald | Physiology/Medicine | Chemistry of vision |
| 1967 | Manfred Eigen and Norrish Porter | Chemistry | Study of high-speed chemical reactions |
| 1968 | Robert W. Holley | Physiology/Medicine | RNA sequence |
| 1968 | H. Gobind Khorana and Marshall W. Nirenberg | Physiology/Medicine | The genetic code |
| 1969 | Max Delbrück, Alfred D. Hershey, and Salvador E. Luria | Physiology/Medicine | Replication and genetic structures of viruses |
| 1970 | Julius Axelrod, Bernard Katz, and Ulf von Euler | Physiology/Medicine | Transmission of nerve impulses, noradrenaline |
| 1970 | Luis F. Leloir | Chemistry | Role of sugar nucleotides in biosynthesis of carbohydrates |
| 1971 | Earl W. Sutherland, Jr. | Physiology/Medicine | Cyclic AMP |
| 1972 | Gerald M. Edelman and Rodney R. Porter | Physiology/Medicine | Structure of antibodies |
| 1972 | Christian B. Anfinsen, Sanford Moore, and William H. Stein | Chemistry | Ribonuclease, structure and activity |
| 1974 | Albert Claude, Christian R. de Duve, and George E. Palade | Physiology/Medicine | Cell structure |
| 1975 | David Baltimore, Renato Dulbecco, and Howard M. Temin | Physiology/Medicine | Reverse transscriptase |
| 1975 | John W. Cornforth and Vladimir Prelog | Chemistry | Stereochemistry of organic molecules and of enzymatic reactions |
| 1977 | Rosalyn S. Yalow, Roger C. L. Guillemin, and Andrew V. Schally | Physiology/Medicine | Radioimmunoassay thyrotropin-releasing hormone |
| 1978 | Daniel Nathans, Werner Arber, and Hamilton O. Smith | Physiology/Medicine | Restriction enzymes |
| 1978 | Peter Mitchell | Chemistry | Biological energy transfer |
| 1980 | Paul Berg, Walter Gilbert, and Frederick Sanger[b] | Chemistry | Recombinant DNA, methods of sequence determination for DNA |
| 1982 | Sune K. Bergstrom, Bergt I. Samuelsson, and John R. Vane | Physiology/Medicine | Isolation and study of prostaglandins |
| 1982 | Aaron Klug | Chemistry | Development of crystallographic electron microscopy |
| 1983 | Barbara McClintock | Physiology/Medicine | Gene transposition |
| 1984 | Niels K. Jerne, Georges J. F. Köhler, and Cesar Milstein | Physiology/Medicine | Cellular basis of immunology |
| 1984 | R. Bruce Merrifield | Chemistry | Solid-phase synthesis of peptides |
| 1985 | Joseph L. Goldstein and Michael S. Brown | Physiology/Medicine | Control of cholesterol metabolism |
| 1986 | Rita Levi-Montalcini and Stanley Cohen | Physiology/Medicine | Nerve growth factor |
| 1987 | Susumu Tonegawa | Physiology/Medicine | Genetics of antibody formation |
| 1988 | Johan Diesenhofer, Robert Huber, and Hartmut Michel | Chemistry | Three-dimensional structure of a photosynthetic reaction center |
| 1988 | Gertrude Elion, George Hitchings, and James Black | Physiology/Medicine | Principles of drug treatment and design of many important drugs |
| 1989 | J. Michael Bishop, Harold E. Varmus, and Joseph E. Murray | Physiology/Medicine | Origin of retroviral oncogenes |
| 1989 | Sidney Altman and Thomas R. Cech | Chemistry | Catalytic RNA |
| 1991 | Erwin Neher and Bert Sakmann | Physiology/Medicine | Functioning of single ion channels in cells |
| 1991 | Richard R. Ernst | Chemistry | High-resolution NMR |
| 1992 | Edmond H. Fischer and Edwin G. Krebs | Physiology/Medicine | Reversible protein phosphorylation in biological regulation |
| 1992 | Rudolph A. Marcus | Chemistry | Theory of electron transfer reactions |
| 1993 | Richard J. Roberts and Phillip A. Sharp | Physiology/Medicine | Discovery of split genes |
| 1993 | Michael Smith and Kary B. Mullis | Chemistry | Oligonucleotide-directed mutagenesis and polymerase chain reaction |
| 1994 | Alfred Gilman and Martin Rodbell | Physiology/Medicine | "G-proteins" |
| 1995 | Edwin B. Lewis, Christiane Nusslein-Volhard, and Eric Wieschaus | Physiology/Medicine | Homeotic mutations in *Drosophila* |
| 1995 | Paul Crutzen, Sherwood Rowland, and Mario Molina | Chemistry | Damage to the stratospheric ozone layer |
| 1996 | Peter C. Doherty and Rolf M. Zinkernagel | Physiology/Medicine | Specificity of cell-mediated immune response |
| 1997 | Paul D. Boyer and John E. Walker | Chemistry | Mechanism of ATP synthesis |
| 1997 | Jens C. Skou | Chemistry | Discovery of $Na^+$, $K^+$-ATPase |
| 1997 | Stanley B. Prusiner | Physiology/Medicine | Discovery of prions |
| 1998 | Robert F. Furchgott, Louis J. Ignarro, and Ferid Murad | Physiology/Medicine | Nitric oxide as a signaling molecule |
| 1999 | Günter Blobel | Physiology/Medicine | Intrinsic signals that govern transport and localization of proteins |
| 2000 | Arvid Carlsson | Physiology/Medicine | Identification of dopamine as signaling molecule in brain |
| | Paul Greengard | Physiology/Medicine | Discovery of the dopamine signaling cascade |
| | Eric R. Kandel | Physiology/Medicine | Molecular basis of learning |

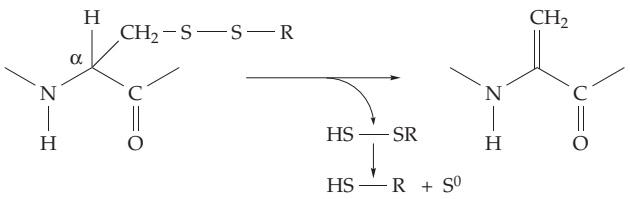
[a] Prizes are given in Physics, Chemistry, Physiology or Medicine, Literature and Peace.
[b] Sanger and Pauling each have been awarded two Nobel Prizes.

**isoaspartyl** groups (Eq. 2-24).[349,350] Aspartyl residues can undergo the same type of cyclization at low pH. The sequence Asn-Gly is especially susceptible to rearrangement according to Eq. 2-24. However, the peptide torsion angles may be more important in determining whether deamidation occurs. The intermediate succinimide may be racemized easily at the chiral center marked by the asterisk in Eq. 2-24. Thus, the Asn-Gly sequence represents a weak linkage, which is nevertheless present in many proteins.




(2-24)

Cystine residues in disulfide bridges are also not completely stable but undergo β elimination[305,351] (Chapter 13) at slightly alkaline pH values according to Eq. 2-25. The free thiol group formed (HSR) is still attached to the protein but may sometimes be in a position to enter into thiol-disulfide exchange reactions with other S–S bridges causing further degradation. Methionine side chains in proteins can be oxidized to sulfoxides: $-CH_2-CH_2-\overset{\overset{O}{\|}}{S}-CH_3$ and hydroxyl groups can be introduced into aromatic rings by oxidation. Hydrogen peroxide and other oxidants may be responsible for such oxidation within cells (see Chapter 18). The very long-lived proteins of the lens of the eye are especially susceptible to deamidation, racemization, oxidation, and accumulation of covalently attached blue fluorescent compounds.[352–354]



(2-25)

## H. Design and Engineering of Proteins

Methods of chemical synthesis of polypeptides and of cloning and mutating genes now allow us to alter peptide sequences at will and to design completely new proteins.[355–356b] The methods are discussed in Chapters 3, 5, and 26. The following are examples.

Peptides that form α helices that associate as coiled coils,[357] or as three- or four-helix tetrameric bundles[179a,358–360] or amphipathic helices that associate with lipid bilayers have been made.[355,361] More difficult has been the design of proteins that form β sheets.[362–364a] Many efforts are being made to understand protein stability[365–367] by systematic substitutions of one residue for another. Addition of new disulfide linkages at positions selected by study of three-dimensional structures sometimes stabilizes enzymes.[368–371] On the other hand elimination of unnecessary cysteine residues can enhance stability by preventing β elimination[351] and replacement of asparagine by threonine can improve the thermostability of enzymes by preventing deamidation.[372,373]

Artificial mutants of subtilisin[373] and of many other enzymes are helping us to understand the mechanisms of catalysis. Artificially prepared temperature-sensitive mutants (Chapter 26) and naturally occurring mutants are providing new insight into pathways of folding of proteins whose three-dimensional structures are known. For these studies it is necessary to follow a rational strategy in deciding which of the astronomical number of possible mutant forms may be of interest. For example, residues within active sites may be changed. Introduction of mutations at random locations can also be useful in finding regions of interest.[374] When mutant proteins can be crystallized, an exact understanding of the effect of the mutation on the structure is possible. Sometimes two or more mutations may have to be introduced to obtain the desired modification.

It is also possible to incorporate a range of "unnatural" amino acids in specific sites in a polypeptide sequence and to observe resulting effects on a protein's properties.[375] A quite different approach is to design polypeptides that mimic a natural peptide but consist of D-amino acids. The peptide chain is *reversed*, i.e., the N terminus becomes the C terminus and every peptide linkage is also reversed. The amino acid side chains preserve their relationships one to another and the backbone atoms tend to preserve their hydrogen bonding pattern. Peptides made in this way tend to be resistant to cleavage by enzymes. Some may be useful as drugs.[376,377]

## References


1. Glendening, E. D., Faust, R., Streitwieser, A., Vollhardt, K. P. C., and Weinhold, F. (1993) *J. Am. Chem. Soc.* **115**, 10952–10957
2. Gao, Q., Weber, H.-P., Craven, B. M., and McMullan, R. K. (1994) *Acta Crystallogr.* **B50**, 695–703
3. Hegstrom, R. A., and Kondepudi, D. K. (1990) *Sci. Am.* **262**(Jan), 108–115
4. Cahn, R. S., Ingold, C., and Prelog, V. (1966) *Angew. Chem. Int. Ed. Engl.* **5**, 385–415
5. Eliel, E. L., Wilen, S. H., and Mander, L. N. (1994) *Stereochemistry of Organic Compounds*, Wiley, New York
6. Bentley, R. (1969) *Molecular Asymmetry in Biology*, Vol. 1, Academic Press, New York (pp. 49–56)
7. Blackwood, J. E., Gladys, C. L., Loening, K. L., Petrarca, A. E., and Rush, J. E. (1968) *J. Am. Chem. Soc.* **90**, 509–510
8. Hanson, K. R. (1966) *J. Am. Chem. Soc.* **88**, 2731–2742
9. Lambert, J. B. (1970) *Sci. Am.* **222**(Jan), 58–70
10. Klyne, W., and Prelog, V. (1960) *Experientia* **16**, 521–523
11. Liébecq, C., ed. (1992) *Biochemical Nomenclature*, Portland Press, London and Chapel Hill, North Carolina (for the International Union of Biochemistry and Molecular Biology)
12. Chiang, Y., Kresge, A. J., and Tang, Y. S. (1984) *J. Am. Chem. Soc.* **106**, 460–462
13. Beak, P., Fry, F. S., Jr., Lee, J., and Steele, F. (1976) *J. Am. Chem. Soc.* **98**, 171–179
14. Beak, P. (1977) *Acc. Chem. Res.* **10**, 186–192
15. Metzler, D. E., Harris, C. M., Johnson, R. J., Siano, D. B., and Thomson, J. A. (1973) *Biochemistry* **12**, 5377–5392
16. Katritzky, A. R., and Karelson, M. (1991) *J. Am. Chem. Soc.* **113**, 1561–1566
17. Blomberg, F., Maurer, W., and Rüterjans, H. (1977) *J. Am. Chem. Soc.* **99**, 8149–8159
18. Walters, D. E., and Allerhand, A. (1980) *J. Biol. Chem.* **255**, 6200–6204
19. Fersht, A. (1977) *Enzyme Structure and Mechanism*, Freeman, San Francisco, California
20. Gabler, R. (1978) *Electrical Interactions in Molecular Biophysics*, Academic Press, New York (p. 188)
21. Israelachvili, J. N., and McGuiggan, P. M. (1988) *Science* **241**, 795–800
22. Salem, L. (1962) *Can. J. Biochem. Physiol.* **40**, 1287–1298
23. Tsao, Y.-H., Fennell Evans, D., and Wennerström, H. (1993) *Science* **262**, 547–550

23a. Kurihara, K., and Kunitake, T. (1992) *J. Am. Chem. Soc.* **114**, 10927–10933

23b. Yang, L., Valdeavella, C. V., Blatt, H. D., and Pettitt, B. M. (1996) *Biophys. J.* **71**, 3022–3029

24. Warshel, A. (1987) *Nature (London)* **330**, 15–16
25. Honig, B., and Nicholls, A. (1995) *Science* **268**, 1144–1149
26. Meot-Ner, M. M. (1988) *J. Am. Chem. Soc.* **110**, 3075–3080
27. Loewenthal, R., Sancho, J., Reinikainen, T., and Fersht, A. R. (1993) *J. Mol. Biol.* **232**, 574–583
28. Gristina, A. G., Oga, M., Webb, L. X., and Hobgood, C. D. (1985) *Science* **228**, 990–993
29. Jeffrey, G. A., and Saenger, W. (1991) *Hydrogen Bonding in Biological Structures*, Springer-Verlag, Berlin
30. Gilli, P., Bertolasi, V., Ferretti, V., and Gastone, G. (1994) *J. Am. Chem. Soc.* **116**, 909–915

30a. Jeffrey, G. A. (1997) *An Introduction to Hydrogen Bonding*, Oxford Univ. Press, New York

31. Derewenda, Z. S., Derewenda, U., and Kobos, P. M. (1994) *J. Mol. Biol.* **241**, 83–93

31a. Wahl, M. C., and Sundaralingam, M. (1997) *Trends Biochem. Sci.* **22**, 97–102

31b. Bella, J., and Berman, H. M. (1996) *J. Mol. Biol.* **264**, 734–742

32. Schulz, G. E., and Schirmer, R. H. (1979) *Principles of Protein Structure*, Springer, New York (p. 36)

32a. Platts, J. A., Howard, S. T., and Bracke, B. R. F. (1996) *J. Am. Chem. Soc.* **118**, 2726–2733

33. Taylor, R., Kennard, O., and Versichel, W. (1983) *J. Am. Chem. Soc.* **105**, 5761–5766
34. Murray-Rust, P., and Glusker, J. P. (1984) *J. Am. Chem. Soc.* **106**, 1018–1025
35. Görbitz, C. H., and Etter, M. C. (1992) *J. Am. Chem. Soc.* **114**, 627–631
36. Nelson, D. D., Jr., Fraser, G. T., and Klemperer, W. (1987) *Science* **238**, 1670–1674

36a. Ghanty, T. K., Staroverov, V. N., Koren, P. R., and Davidson, E. R. (2000) *J. Am. Chem. Soc.* **122**, 1210–1214

36b. Cornilescu, G., Ramirez, B. E., Frank, M. K., Clore, G. M., Gronenborn, A. M., and Bax, A. (1999) *J. Am. Chem. Soc.* **121**, 6275–6279

36c. Benedict, H., Shenderovich, I. G., Malkina, O. L., Malkin, V. G., Denisov, G. S., Golubev, N. S., and Limbach, H.-H. (2000) *J. Am. Chem. Soc.* **122**, 1979–1988

37. Fersht, A. R., Shi, J., Knill-Jones, J., Lowe, D. M., Witkinson, A. J., Blow, D. M., Brick, P., Carter, P., Woye, M. M. Y., and Winter, G. (1985) *Nature (London)* **314**, 235–238
38. Eriksson, A. E., Baase, W. A., Zhang, X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P., and Matthews, B. W. (1992) *Science* **255**, 178–183

38a. Takano, K., Yamagata, Y., Kubota, M., Funahashi, J., Fujii, S., and Yutani, K. (1999) *Biochemistry* **38**, 6623–6629

39. Lim, W. A., Hodel, A., Sauer, R. T., and Richards, F. M. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 423–427
40. Schuster, P., Zundel, G., and Sandorfy, C. (1976) *The Hydrogen Bond*, North Holland Publ. Co., Amsterdam (3 vols.)
41. Luck, W. A. P., and Kleeberg, H. (1978) in *Photosynthetic Oxygen Evolution* (Metzner, H., ed), pp. 1–29, Academic Press, London
42. Edsall, J. T. (1983) *Trends Biochem. Sci.* **8**, 29–31
43. Grunwald, E. (1986) *J. Am. Chem. Soc.* **108**, 5719–5726
44. Westhof, E. (1993) *Water and Biological Macromolecules*, CRC Press, Boca Raton, Florida
45. Kusalik, P. G., and Svishchev, I. M. (1994) *Science* **265**, 1219–1221
46. Libnau, F. O., Toft, J., Christy, A. A., and Kvalheim, O. M. (1994) *J. Am. Chem. Soc.* **116**, 8311–8316
47. Benson, S. W., and Siebert, E. D. (1992) *J. Am. Chem. Soc.* **114**, 4269–4276
48. Liu, K., Cruzan, J. D., and Saykally, R. J. (1996) *Science* **271**, 929–933
49. Pauling, L. (1935) *J. Am. Chem. Soc.* **57**, 2680
50. Wittebort, R. J., Usha, M. G., Ruben, D. J., Wemmer, D. E., and Pines, A. (1988) *J. Am. Chem. Soc.* **110**, 5668–5671
51. Li, J., and Ross, D. K. (1993) *Nature (London)* **365**, 327–329
52. Dyke, T. R., Mack, K. M., and Muenter, J. S. (1977) *J. Chem. Phys.* **66**, 498–510
53. Saenger, W. (1979) *Nature (London)* **279, 280**, 343–344, 848
54. Saenger, W. (1987) *Ann. Rev. Biophys. Biophys. Chem.* **16**, 93–114
55. Pribble, R. N., and Zwier, T. S. (1994) *Science* **265**, 75–79
56. Gregory, J. K., Clary, D. C., Liu, K., Brown, M. G., and Saykally, R. J. (1997) *Science* **275**, 814–817
57. Quiocho, F. A., Sack, J. S., and Vyas, N. K. (1987) *Nature (London)* **329**, 561–564
58. Pflugrath, J. W., and Quiocho, F. A. (1988) *J. Mol. Biol.* **200**, 163–180
59. Kauzmann, W. (1959) *Adv. Prot. Chem.* **14**, 1–63
60. Jencks, W. P. (1987) *Catalysis in Chemistry and Enzymology*, Dover, Mineola, New York (p. 393)
61. Tanford, C. (1980) *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd ed., Wiley, New York
62. Abraham, M. H. (1980) *J. Am. Chem. Soc.* **102**, 5910–5912
63. Ben-Náim, A. (1980) *Hydrophobic Interactions*, Plenum, New York
64. Murphy, K. P., Privalov, P. L., and Gill, S. J. (1990) *Science* **247**, 559–561
65. Muller, N. (1992) *Trends Biochem. Sci.* **17**, 459–463
66. Spolar, R. S., Livingstone, J. R., and Record, J., MT. (1992) *Biochemistry* **31**, 3947–3955
67. Stites, W. E., Meeker, A. K., and Shortle, D. (1994) *J. Mol. Biol.* **235**, 27–32
68. Hecht, D., Tadesse, L., and Walters, L. (1993) *J. Am. Chem. Soc.* **115**, 3336–3337
69. Gill, S. J., and Wadsö, I. (1976) *Proc. Natl. Acad. Sci. U.S.A.* **73**, 2955–2958
70. Sharp, K. A., Nicholls, A., Fine, R. F., and Honig, B. (1991) *Science* **252**, 106–109
71. Van Oss, C. J. (1994) *Interfacial Forces in Aqueous Media*, Dekker, New York

71a. Hummer, G., Garde, S., García, A. E., Pohorille, A., and Pratt, L. R. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8951–8955

72. Israelachvili, J., and Wennerström, H. (1996) *Nature (London)* **379**, 219–225
73. Radzicka, A., and Wolfenden, R. (1988) *Biochemistry* **27**, 1664–1670
74. Kellis, J. T., Jr., Nyberg, K., Sali, D., and Fersht, A. R. (1988) *Nature (London)* **333**, 784–786
75. Sanger, F. (1988) *Ann. Rev. Biochem.* **57**, 1–28
76. Creighton, T. E. (1993) *Proteins, Structures and Molecular Principles*, 2nd ed., Freeman, New York
77. Kyte, J. (1995) *Structure in Protein Chemistry*, Garland Publ., New York
78. Guillemin, R., and Burgus, R. (1972) *Sci. Am.* **227**(Nov), 24–33
79. Hochster, R. M., Kates, M., and Quastel, J. H., eds. (1973) *Metabolic Inhibitors*, Vol. 3, Academic Press, New York (p. 312)
80. Ramachandran, G. N., and Kolaskar, A. S. (1973) *Biochim. Biophys. Acta.* **303**, 385–388
81. MacArthur, M. W., and Thornton, J. M. (1996) *J. Mol. Biol.* **264**, 1180-1195

81a. Milner-White, E. J. (1997) *Protein Sci.* **6**, 2477–2482

82. Moffat, J. B. (1973) *J. Theor. Biol.* **40**, 247–258
83. Fersht, A. R. (1971) *J. Am. Chem. Soc.* **93**, 3504–3515
84. Karle, I. L. (1974) *Biochemistry* **13**, 2155–2162
85. Ovchinnikov, Y. A. (1973) *FEBS Lett.* **29**, 31–33
86. Doonan, S. (1974) *FEBS Lett.* **38**, 229–233
87. Bhown, A. S. (1990) *Handbook of Proteins*, A & M Publ., Birmingham, Alabama
88. Boguski, M., and McEntyre, J. (1994) *Trends Biochem. Sci.* **19**, 71

88a. Branden, C., and Tooze, J. (1991) *Introduction to Protein Structure*, Garland Publ., New York

89. Clements, S., Mehansho, H., and Carlson, D. M. (1975) *J. Biol. Chem.* **260**, 13471–13477
90. Condit, C. M., and Meagher, R. B. (1986) *Nature (London)* **323**, 178–181
91. Reynolds, P., Weber, S., and Prakash, L. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 168–172
92. Rodakis, G. C., and Kafatos, F. C. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3551–3555
93. Young, J. F., Hockmeyer, W. T., Gross, M., Ballou, W. R., Wirtz, R. A., Trosper, J. H., Beaudoin, R. L., Hollingdale, M. R., Miller, L. H., Diggs, C. L., and Rosenberg, M. (1985) *Science* **228**, 958–962

## References


94. Doolittle, R. F. (1987) *Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequences*, Univ. Science Books, Mill Valley, California
95. Doolittle, R. F. (1989) *Trends Biochem. Sci.* **14**, 244–245
96. Pearson, W. R., and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444–2448
97. Holm, L., and Sander, C. (1993) *J. Mol. Biol.* **233**, 123–138
98. Lewin, R. (1987) *Science* **237**, 1570
99. Fleming, P. J., and Richards, F. M. (2000) *J. Mol. Biol.* **299**, 487–498
100. Gekko, K., and Hasegawa, Y. (1986) *Biochemistry* **25**, 6563–6571
101. Goodsell, D. S., and Olson, A. J. (1993) *Trends Biochem. Sci.* **18**, 65–68
102. Schultz, G. E., and Schirmer, R. H. (1979) *Principles of Protein Structure*, Springer-Verlag, New York

102a. Ramachandran, G. N., Ramakrishnam, C., and Sasisekharan, V. (1963) *J. Mol. Biol.*, **7**, 95–99

103. Ramachandran, G. N., Venkatachalam, C. M., and Krimm, S. (1966) *Biophys. J.* **6**, 849–872

103a. Karplus, P. A. (1996) *Protein Sci.* **5**, 1406–1420

104. Sasisekharian, V., Lakshminarayanan, A. V., and Ramachandran, G. N. (1967) in *Conformation of Biopolymers*, Vol. 2 (Ramachandran, G. N., ed), p. 641–654, Academic Press, New York

104a. Gunasekaran, K., Ramakrishnan, C., and Balaram, P. (1996) *J. Mol. Biol.* **264**, 191–198

104b. Herzberg, O., and Moult, J. (1991) *Proteins* **11**, 223–229

105. Milner-White, E. J., Bell, L. H., and Maccallum, P. H. (1992) *J. Mol. Biol.* **228**, 725–734

106. Rhee, S. (1990) *High resolution X-ray diffraction studies of cytosolic aspartate aminotransferase* [Ph.D. Dissertation], University of Iowa, Iowa City, Iowa

107. Dunbrack, R. L., Jr., and Karplus, M. (1993) *J. Mol. Biol.* **230**, 543–574

108. Pauling, L., and Corey, R. B. (1951) *Proc. Natl. Acad. Sci. U.S.A.* **37**, 729–740

109. Rich, A. (1994) *Nature (London)* **371**, 285

110. Chothia, C. (1973) *J. Mol. Biol.* **75**, 295–302

111. Chou, K.-C., Némethy, G., and Scheraga, H. A. (1983) *Biochemistry* **22**, 6213–6221

111a. Skarzynski, T., Moody, P. C. E., and Wonacott, A. J. (1987) *J. Mol. Biol.* **193**, 171–187

111b. Wang, L., O'Connell, T., Tropsha, A., and Hermans, J. (1996) *J. Mol. Biol.* **262**, 283–293

112. Reeke, G. N., Jr., and Becker, J. W. (1986) *Science* **234**, 1108–1111

113. Murzin, A. G., Lesk, A. M., and Chothia, C. (1994) *J. Mol. Biol.* **236**, 1369–1381

114. Murzin, A. G., Lesk, A. M., and Chothia, C. (1994) *J. Mol. Biol.* **236**, 1382–1400

115. LaLonde, J. M., Bernlohr, D. A., and Ranaszak, L. J. (1994) *FASEB J.* **8**, 1240–1241

116. Guss, J. M., and Freeman, H. C. (1983) *J. Mol. Biol.* **169**, 521–563

117. Richardson, J. S. (1981) *Adv. Prot. Chem.* **34**, 167–339

117a. Erickson, H. P. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 10114–10118

118. Redinbo, M. R., Cascio, D., Choukair, M. K., Rice, D., Merchant, S., and Yeates, T. O. (1993) *Biochemistry* **32**, 10560–10567

119. Weiss, M. S., and Schulz, G. E. (1993) *J. Mol. Biol.* **231**, 817–824

120. Kreusch, A., and Schulz, G. E. (1994) *J. Mol. Biol.* **243**, 891–905

121. Yoder, M. D., and Jurnak, F. (1995) *FASEB J.* **9**, 335–342

122. Richardson, J. S. (1994) *FASEB J.* **8**, 1237–1239

123. Steinbacher, S., Seckler, R., Miller, S., Steipe, B., Huber, R., and Reinemer, P. (1994) *Science* **265**, 383–385

124. Richardson, J. S., Getzoff, E. D., and Richardson, D. C. (1978) *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2574–2578

125. Chan, A. W. E., Hutchinson, E. G., Harris, D., and Thornton, J. M. (1993) *Protein Sci.* **2**, 1574–1590

126. Urry, D. W. (1972) *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1610–1614

127. Nicholson, L. K., and Cross, T. A. (1989) *Biochemistry* **28**, 9379–9385

127a. Smith, T. F., Gaitatzes, C., Saxena, K., and Neer, E. J. (1999) *Trends Biochem. Sci.* **24**, 181–185

127b. Garcia-Higuera, I., Fenoglio, J., Li, Y., Lewis, C., Panchenko, M. P., Reiner, O., Smith, T. F., and Neer, E. J. (1996) *Biochemistry* **35**, 13985–13994

127c. ter Naar, E., Harrison, S. C., and Kirchhausen, T. (2000) *Proc. Natl. Acad. Sci. U.S.A.* **97**, 1096–1100

127d. Springer, T. A. (1998) *J. Mol. Biol.* **283**, 837–862

128. Pauling, L. (1993) *Protein Sci.* **2**, 1060–1063

129. Kendrew, J. C. (1969) *Sci. Am.* **205**(Dec), 96–110

130. Arnott, S., and Dover, S. D. (1967) *J. Mol. Biol.* **30**, 209–212

131. Steigemann, W., and Weber, E. (1979) *J. Mol. Biol.* **127**, 309–338

132. Barlow, D. J., and Thornton, J. M. (1988) *J. Mol. Biol.* **201**, 601–619

132a. Rajashankar, K. R., and Ramakumar, S. (1996) *Protein Sci.* **5**, 932–946

133. Harper, E. T., and Rose, G. D. (1993) *Biochemistry* **32**, 7605–7609

134. Chakrabartty, A., Schellman, J. A., and Baldwin, R. L. (1991) *Nature (London)* **351**, 586–588

135. Creamer, T. P., and Rose, G. D. (1995) *Protein Sci. U.S.A.* **4**, 1305–1314

136. Dempsey, C. E. (1992) *Biochemistry* **31**, 4705–4712

137. Ramachandran, G. N., and Mitra, A. K. (1976) *J. Mol. Biol.* **107**, 85–92

138. Blaber, M., Zhang, X.-j, and Matthews, B. W. (1993) *Science* **260**, 1637–1639

139. Chakrabartty, A., Kortemme, T., and Baldwin, R. L. (1994) *Protein Sci.* **3**, 843–852

140. Toniolo, C., and Benedetti, E. (1991) *Trends Biochem. Sci.* **16**, 350–353

141. Avbelj, F., and Moult, J. (1995) *Biochemistry* **34**, 755–764

142. Karpen, M. E., De Haseth, P. L., and Neet, K. E. (1992) *Protein Sci.* **1**, 1333–1342

143. Fiori, W. R., Miick, S. M., and Millhauser, G. L. (1993) *Biochemistry* **32**, 11957–11962

144. Smythe, M. L., Huston, S. E., and Marshall, G. R. (1993) *J. Am. Chem. Soc.* **115**, 11594–11595

145. Bavosa, A., Benedetti, E., Di Blasio, B., Pavone, V., Pedone, C., Toniolo, C., and Bonora, G. M. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1988–1992

145a. Fathallah-Shaykh, H., Wolf, S., Wong, E., Posner, J. B., and Furneaux, H. M. (1991) *Proc. Natl. Acad. Sci. U.S.A* **88**, 3451–3454

146. Miick, S. M., Martinez, G. V., Flori, W. R., Todd, A. P., and Millhauser, G. L. (1992) *Nature (London)* **359**, 653–655

147. Wada, A. (1976) *Adv. Biophys.* **9**, 1–63

148. Hol, W. G. J., van Duijnen, P. T., and Berendsen, H. J. C. (1978) *Nature (London)* **273**, 443–446

148a. Hol, W. G. J., Halie, L. M., and Sander, C. (1981) *Nature (London)* **294**, 532–536

148b. Lockhart, D. J., and Kim, P. S. (1992) *Science* **257**, 947–951

149. Levine, B. F., and Bethea, C. G. (1976) *J. Chem. Phys.* **65**, 1989–1993

150. Gray, T. M., and Matthews, B. W. (1984) *J. Mol. Biol.* **175**, 75–81

151. Richardson, J. S., and Richardson, D. C. (1988) *Science* **240**, 1648–1652

152. Nicholson, H., Anderson, D. E., Dao-pin, S., and Matthews, B. W. (1991) *Biochemistry* **30**, 9816–9828

153. Åqvist, J., Luecke, H., Quiocho, F. A., and Warshel, A. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 2026–2030

154. Lyu, P. C., Wemmer, D. E., Zhou, H. X., Pinker, R. J., and Kallenbach, N. R. (1993) *Biochemistry* **32**, 421–425

155. Viguera, A. R., and Serrano, L. (1995) *J. Mol. Biol.* **251**, 150–160

156. Doig, A. J., Chakrabartty, A., Klingler, T. M., and Baldwin, R. L. (1994) *Biochemistry* **33**, 3396–3403

157. Doig, A. J., Macarthur, M. W., Stapley, B. J., and Thornton, J. M. (1997) *Protein Sci.* **6**, 147–155

158. Sali, D., Bycroft, M., and Fersht, A. R. (1988) *Nature (London)* **335**, 740–743

159. Aurora, R., and Rose, G. D. (1998) *Protein Sci.* **7**, 21–38

160. Metzler, D. E. (1979) *Adv. Enzymol.* **50**, 1–40

161. Lupas, A. (1996) *Trends Biochem. Sci.* **21**, 375–382

161. Lupas, A. (1996) *Trends Biochem. Sci.* **21**, 375–382

162. Kohn, W. D., Mant, C. T., and Hodges, R. S. (1997) *J. Biol. Chem.* **272**, 2583–2586

163. Fraser, R. D. B., MacRae, T. P., Parry, D. A. D., and Suzuki, E. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1179–1183

164. Johnson, L. D., Idler, W. W., Zhou, X.-M., Roop, D. R., and Steinert, P. M. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1896–1900

165. Cooper, D., and Sun, T.-T. (1986) *J. Biol. Chem.* **261**, 4646–4654

166. Crick, F. H. C. (1953) *Acta Crystallogr.* **6**, 689–697

167. Vazquez, S. R., Kuo, D. Z., Bositis, C. M., Hardy, L. W., Lew, R. A., and Humphreys, R. E. (1992) *J. Biol. Chem.* **267**, 7406–7410

168. O'Shea, E. K., Llemm, J. D., Kim, P. S., and Alber, T. (1991) *Science* **254**, 539–544

169. Ferré-D'Amaré, A. R., Prendergast, G. C., Ziff, E. B., and Burley, S. K. (1993) *Nature (London)* **363**, 38–44

170. Lyu, P. C., Marky, L. A., and Kallenbach, N. R. (1989) *J. Am. Chem. Soc.* **111**, 2733–2734

171. Graddis, T. J., Myszka, D. G., and Chaiken, I. M. (1993) *Biochemistry* **32**, 12664–12671

171a. Fairman, R., Chao, H.-G., Lavoie, T. B., Villafranca, J. J., Matsueda, G. R., and Novotny, J. (1996) *Biochemistry* **35**, 2824–2829

171b. Kohn, W. D., Kay, C. M., and Hodges, R. S. (1998) *J. Mol. Biol.* **283**, 993–1012

172. John, M., Briand, J.-P., Granger-Schnarr, M., and Schnarr, M. (1994) *J. Biol. Chem.* **269**, 16247–16253

173. Zhou, N. E., Kay, C. M., and Hodges, R. S. (1992) *Biochemistry* **31**, 5739–5746

174. Cohen, C., and Parry, D. A. D. (1986) *Trends Biochem. Sci.* **11**, 245–248

175. Phillips, G. N., Jr., Fillers, J. P., and Cohen, C. (1986) *J. Mol. Biol.* **192**, 111–131

176. Mo, J., Holtzer, M. E., and Holtzer, A. (1993) *Protein Sci.* **2**, 128–130

177. Xie, X., Rao, S., Walian, P., Hatch, V., Phillips, G. N., Jr., and Cohen, C. (1994) *J. Mol. Biol.* **236**, 1212–1226

178. Eckert, R. L. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1114–1118

179. Gernert, K. M., Surles, M. C., Labean, T. H., Richardson, J. S., and Richardson, D. C. (1995) *Protein Sci.* **4**, 2252–2260

179a. Monera, O. D., Zhou, N. E., Lavigne, P., Kay, C. M., and Hodges, R. S. (1996) *J. Biol. Chem.* **271**, 3995–4001

180. Lovejoy, B., Choe, S., Cascio, D., McRorie, D. K., DeGrado, W. F., and Eisenberg, D. (1993) *Science* **259**, 1288–1293

## References


181. Harbury, P. B., Kim, P. S., and Alber, T. (1994) *Nature (London)* **371**, 80–83
182. Efimov, V. P., Nepluev, I. V., Sobolev, B. N., Zurabishvili, T. G., Schulthess, T., Lustig, A., Engel, J., Haener, M., Aebi, U., Venyaminov, S. Y., Potekhin, S. A., and Mesyanzhinov, V. V. (1994) *J. Mol. Biol.* **242**, 470–486
183. Lieberman, M., Tabet, M., and Sasaki, T. (1994) *J. Am. Chem. Soc.* **116**, 5035–5044
184. Utani, A., Nomizu, M., Timpl, R., Roller, P. P., and Yamada, Y. (1994) *J. Biol. Chem.* **269**, 19167–19175
185. Robinson, C. R., and Sligar, S. G. (1993) *Protein Sci.* **2**, 826–837
186. Steif, C., Weber, P., Hinz, H.-J., Flossdorf, J., Cesareni, G., and Kokkinidis, M. (1993) *Biochemistry* **32**, 3867–3876
187. Redfield, C., Boyd, J., Smith, L. J., Smith, R. A. G., and Dobson, C. M. (1992) *Biochemistry* **31**, 10431–10437
188. Harris, N. L., Presnell, S. R., and Cohen, F. E. (1994) *J. Mol. Biol.* **236**, 1356–1368
189. Chou, K.-C., Carlacci, L., Maggiora, G. M., Parodi, L. A., and Schulz, M. W. (1992) *Protein Sci.* **1**, 810–827
190. Rees, D. C., Komiya, H., Yeates, T. O., Allen, J. P., and Feher, G. (1989) *Ann. Rev. Biochem.* **58**, 607–633
191. Aleshin, A. E., Hoffman, C., Firsov, L. M., and Honzatko, R. B. (1994) *J. Mol. Biol.* **238**, 575–591
192. Aleshin, A. E., Firsov, L. M., and Honzatko, R. B. (1994) *J. Biol. Chem.* **269**, 15631–15639
193. Harris, E. M. S., Aleshin, A. E., Firsov, L. M., and Honzatko, R. B. (1993) *Biochemistry* **32**, 1618–1626
194. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) *Science* **266**, 75–81

194a. Kramer, R. Z., Bella, J., Mayville, P., Brodsky, B., and Berman, H. M. (1999) *Nature Struct. Biol.* **6**, 454–457

195. Brodsky, B., and Shah, N. K. (1995) *FASEB J.* **9**, 1537–1546
196. Redeker, V., Levilliers, N., Schmitter, J.-M., Le Caer, J.-P., Rossier, J., Adoutte, A., and Bré, M.-H. (1994) *Science* **266**, 1688–1691
197. Miller, A. (1982) *Trends Biochem. Sci.* **7**, 13–18
198. Eyre, D. R. (1980) *Science* **207**, 1315–1322
199. Parry, D. A. D., and Craig, A. S. (1979) *Nature (London)* **282**, 213–215
200. Porter, R. R., and Reid, K. B. M. (1979) *Adv. Prot. Chem.* **33**, 1–71
201. Ashford, D., and Neuberger, A. (1980) *Trends Biochem. Sci.* **5**, 245–248
202. Chen, J., and Varner, J. E. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4399–4403
203. Adzhubei, A. A., and Sternberg, M. J. E. (1994) *Protein Sci.* **3**, 2395–2410
204. Richardson, J. S. (1985) *Nature (London)* **316**, 102–103
205. Hutchinson, E. G., and Thornton, J. M. (1994) *Protein Sci.* **3**, 2207–2216
206. Venkatachalam, C. M. (1968) *Biopolymers* **6**, 1425–1436
207. Perczel, A., Foxman, B. M., and Fasman, G. D. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8210–8214
208. Yan, Y., Tropsha, A., Hermans, J., and Erickson, B. W. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7898–7902
209. Perczel, A., McAllister, M. A., Császár, P., and Csizmadia, I. G. (1993) *J. Am. Chem. Soc.* **115**, 4849–4858
210. Mattos, C., Petsko, G. A., and Karplus, M. (1994) *J. Mol. Biol.* **238**, 733–747
211. Hynes, T. R., Hodel, A., and Fox, R. O. (1994) *Biochemistry* **33**, 5021–5030
212. Haque, T. S., Little, J. C., and Gellman, S. H. (1994) *J. Am. Chem. Soc.* **116**, 4105–4106
213. Gierasch, L. M., Deber, C. M., Madison, V., Niu, C.-H., and Blout, E. R. (1981) *Biochemistry* **20**, 4730–4738
214. Rees, D. C., Lewis, M., and Lipscomb, W. N. (1983) *J. Mol. Biol.* **168**, 367–387
215. Sapse, A.-M., Mallah-Levy, L., Daniels, S. B., and Erickson, B. W. (1987) *J. Am. Chem. Soc.* **109**, 3526–3529
216. Urry, D. W. (1993) *Angew. Chem. Int. Ed. Engl.* **32**, 819–841
217. Sandberg, L. B., Soskel, N. T., and Leslie, J. G. (1981) *N. Engl. J. Med.* **304**, 566–579

217a. Tatham, A. S., and Shewry, P. R. (2000) *Trends Biochem. Sci.* **25**, 567–571

218. Gotte, L., Volpin, D., Horne, R. W., and Mammi, M. (1976) *Micron* **7**, 95–102

218a. Urry, D., Luan, C.-H., and Peng, S. (1995) *Ciba. Found. Symp.* **192**, 4–30

218b. van Dijk, A. A., de Boef, E., Bekkers, A., van Wijk, L. L., van Swieten, E., Hamer, R. J., and Robillard, G. T. (1997) *Protein Sci.* **6**, 649–656

219. Sternberg, M. J. E., and Thornton, J. M. (1976) *J. Mol. Biol.* **105**, 367–382
220. Leszczyński, J. F., and Rose, G. D. (1986) *Science* **234**, 849–855
221. Liang, C., and Mislow, K. (1994) *J. Am. Chem. Soc.* **116**, 11189–11190
222. Chothia, C. (1984) *Ann. Rev. Biochem.* **53**, 537–572
223. Resnick, D., Pearson, A., and Krieger, M. (1994) *Trends Biochem. Sci.* **19**, 5–8
224. Shoyab, M., Plowman, G. D., McDonald, V. L., Bradley, J. G., and Todaro, G. J. (1989) *Science* **243**, 1074–1076
225. Doolittle, R. F. (1995) *Ann. Rev. Biochem.* **64**, 287–314
226. Berman, A. L., Kolker, E., and Trifanov, E. N. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4044–4047
227. Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994) *Nature (London)* **372**, 631–634
228. Pan, K.-M., Damodaran, S., and Greaser, M. L. (1994) *Biochemistry* **33**, 8255–8261
229. Doolittle, R. F., and Bork, P. (1993) *Sci. Am.* **269**(Oct), 50–56
230. Baron, M., Norman, D. G., and Campbell, I. D. (1991) *Trends Biochem. Sci.* **16**, 13–17
231. Mulichak, A. M., Tulinsky, A., and Ravichandran, K. G. (1991) *Biochemistry* **30**, 10576–10588
232. Bork, P., Holm, L., and Sander, C. (1994) *J. Mol. Biol.* **242**, 309–320

232a. Bogan, A. A., and Thorn, K. S. (1998) *J. Mol. Biol.* **280**, 1–9

233. Tsai, J., Taylor, R., Chothia, C., and Gerstein, M. (1999) *J. Mol. Biol.* **290**, 253–266

233a. Gerstein, M., and Chothia, C. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10167–10172

234. Varadarajan, R., and Richards, F. M. (1992) *Biochemistry* **31**, 12315–12327
235. Allewell, N. (1987) *Trends Biochem. Sci.* **12**, 417–418
236. Lim, W. A., Farruggio, D. C., and Sauer, R. T. (1992) *Biochemistry* **31**, 4324–4333
237. Hubbard, S. J., and Argos, P. (1994) *Protein Sci.* **3**, 2194–2206

237a. Hubbard, S. J., and Argos, P. (1996) *J. Mol. Biol.* **261**, 289–300

238. Hunt, N. G., Gregoret, L. M., and Cohen, F. E. (1994) *J. Mol. Biol.* **241**, 214–225
239. Hunter, C. A., and Sanders, K. M. (1990) *J. Am. Chem. Soc.* **112**, 5525–5534
240. Burley, S. K., and Petsko, G. A. (1985) *Science* **229**, 23–28
241. Thomas, K. A., Smith, G. M., Thomas, T. B., and Feldmann, R. J. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4843–4847
242. Flocco, M. M., and Mowbray, S. L. (1994) *J. Mol. Biol.* **235**, 709–717
243. Hemmingsen, J. M., Gernert, K. M., Richardson, J. S., and Richardson, D. C. (1994) *Protein Sci.* **3**, 1927–1937
244. Teeter, M. M. (1991) *Ann. Rev. Biophys. Biophys. Chem.* **20**, 577–600
245. McDonald, I. K., and Thornton, J. M. (1994) *J. Mol. Biol.* **238**, 777–793
246. Bordo, D., and Argos, P. (1994) *J. Mol. Biol.* **243**, 504–519
247. Shimoni, L., and Glusker, J. P. (1995) *Protein Sci.* **4**, 65–74
248. Dao-pin, S., Anderson, D. E., Baase, W. A., Dahlquist, F. W., and Matthews, B. W. (1991) *Biochemistry* **30**, 11521–11529

248a. Kumar, S., and Nussinov, R. (1999) *J. Mol. Biol.* **293**, 1241–1255

248b. Strop, P., and Mayo, S. L. (2000) *Biochemistry* **39**, 1251–1255

249. Sawyer, L., and James, M. N. G. (1982) *Nature (London)* **295**, 79–80
250. Borders, C. L., Jr., Broadwater, J. A., Bekeny, P. A., Salmon, J. E., Lee, A. S., Eldridge, A. M., and Pett, V. B. (1994) *Protein Sci.* **3**, 541–548
251. Mrabet, N. T., Van den Broeck, A., Van den brande, I., Stanssens, P., Laroche, Y., Lambeir, A.-M., Matthiissens, G., Jenkins, J., Chiadmi, M., van Tilbeurgh, H., Rey, F., Janin, J., Quax, W. J., Lasters, I., De Maeyer, M., and Wodak, S. J. (1992) *Biochemistry* **31**, 2239–2253
252. Dougherty, D. A. (1996) *Science* **271**, 163–168
253. Mitchell, J. B. O., Nandi, C. L., McDonald, I. K., Thornton, J. M., and Price, S. (1994) *J. Mol. Biol.* **239**, 315–331

253a. De Wall, S. L., Meadows, E. S., Barbour, L. J., and Gokel, G. W. (1999) *J. Am. Chem. Soc.* **121**, 5613–5614

253b. Gallivan, J. P., and Dougherty, D. A. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9459–9464

253c. Wouters, J. (1998) *Protein Sci.* **7**, 2472–2475

253d. Fernández-Recio, J., Romero, A., and Sancho, J. (1999) *J. Mol. Biol.* **290**, 319–330

254. Rose, G. D., Young, W. B., and Gierasch, L. M. (1983) *Nature (London)* **304**, 654–657
255. Watenpaugh, K. D., Margulis, T. N., Sieker, L. C., and Jensen, L. H. (1978) *J. Mol. Biol.* **122**, 175–190

255a. Nakasako, M. (1999) *J. Mol. Biol.* **289**, 547–564

256. Teeter, M. M. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6014–6018
257. Otting, G., Liepinsh, E., and Wüthrich, K. (1991) *Science* **254**, 974–980
258. Murzin, A. G., and Finkelstein, A. V. (1988) *J. Mol. Biol.* **204**, 749–770
259. Chothia, C. (1989) *Nature (London)* **337**, 204–205
260. Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., and Rossmann, M. G. (1974) *J. Mol. Biol.* **90**, 25–49
261. Rossmann, M. G., Liljas, A., Branden, C.–I., and Banaszak, L. J. (1975) *The Enzymes*, 3rd ed., Vol. 11, Academic Press, New York (p. 61)
262. Brändén, C. (1990) *Nature (London)* **346**, 607–608
263. Kobe, B., and Deisenhofer, J. (1995) *Nature (London)* **374**, 183–186
264. Kobe, B., and Deisenhofer, J. (1994) *Trends Biochem. Sci.* **19**, 415–421

264a. Kajava, A. V. (1998) *J. Mol. Biol.* **277**, 519–527

264b. Kobe, B., and Kajava, A. V. (2000) *Trends Biochem. Sci.* **25**, 509–515

265. Farber, G. K., and Petsko, G. A. (1990) *Trends Biochem. Sci.* **15**, 228–234
266. Reardon, D., and Farber, G. K. (1995) *FASEB J.* **9**, 497–503
267. Uhlin, U., and Eklund, H. (1994) *Nature (London)* **370**, 533–539
268. Chothia, C. (1992) *Nature (London)* **357**, 543–544

268a. Zhang, C., and DeLisi, C. (1998) *J. Mol. Biol.* **284**, 1301–1305

269. McLachlan, A. D. (1980) *Nature (London)* **285**, 267–268

## References


270. Chou, P. Y., and Fasman, G. D. (1974) *Biochemistry* **13**, 222–245
270a. Myers, J. K., Pace, C. N., and Scholtz, J. M. (1997) *Biochemistry* **36**, 10923–10929
270b. Spek, E. J., Olson, C. A., Shi, Z., and Kallenbach, N. R. (1999) *J. Am. Chem. Soc.* **121**, 5571–5572
271. Minor, D. L., Jr., and Kim, P. S. (1994) *Nature (London)* **367**, 660–663
272. Richards, F. M. (1991) *Sci. Am.* **264**(Jan), 54–63
273. Anfinsen, C. B. (1973) *Science* **181**, 223–230
274. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. (1995) *Protein Sci.* **4**, 561–602
275. Rost, B., Schneider, R., and Sander, C. (1993) *Trends Biochem. Sci.* **18**, 120–123
276. Wako, H., and Blundell, T. L. (1994) *J. Mol. Biol.* **238**, 693–708
277. Yi, T.-M., and Lander, E. S. (1993) *J. Mol. Biol.* **232**, 1117–1129
278. Monge, A., Friesner, R. A., and Honig, B. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5027–5029
279. Smith, C. K., Withka, J. M., and Regan, L. (1994) *Biochemistry* **33**, 5510–5517
279a. Lacroix, E., Viguera, A. R., and Serrano, L. (1998) *J. Mol. Biol.* **284**, 173–191
279b. Gotoh, O. (1996) *J. Mol. Biol.* **264**, 823–838
279c. Dandekar, T., and Argos, P. (1996) *J. Mol. Biol.* **256**, 645–660
279d. Jones, D. T. (1999) *J. Mol. Biol.* **292**, 195–202
280. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132
281. Reyes, V. E., Phillips, L., Humphreys, R. E., and Lew, R. A. (1989) *J. Biol. Chem.* **264**, 12854–12858
282. Wold, F. (1986) *Trends Biochem. Sci.* **11**, 58–59
283. Graves, D. J., Martin, B. L., and Wang, J. H. (1994) *Co-and post-translational modification of proteins*, Oxford Univ. Press, New York
283a. Angeletti, R. H., ed. (1998) *Protein, Analysis and Design*, Academic Press, San Diego, California
284. Yan, S. C. B., Grinnell, B. W., and Wold, F. (1989) *Trends Biochem. Sci.* **14**, 264–268
285. Berthiaume, L., Deichaite, I., Peseckis, S., and Resh, M. D. (1994) *J. Biol. Chem.* **269**, 6498–6505
286. McIlhinney, R. A. J. (1990) *Trends Biochem. Sci.* **15**, 387–391
287. Violand, B. N., Schlittler, M. R., Lawson, C. Q., Kane, J. F., Siegel, N. R., Smith, C. E., Kolodziej, E. W., and Duffin, K. L. (1994) *Protein Sci.* **3**, 1089–1097
288. Hilz, H., Fanick, N., and Klapproth, K. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6267–6271
289. Hortin, G., Fok, K. F., Toren, P. C., and Strauss, A. W. (1987) *J. Biol. Chem.* **262**, 3082–3085
290. Bateman, A., Solomon, S., and Bennett, H. P. J. (1990) *J. Biol. Chem.* **265**, 22130–22136
291. Takahara, H., Tsuchida, M., Kusubata, M., Akutsu, K., Tagami, S., and Sugawara, K. (1989) *J. Biol. Chem.* **264**, 13361–13368
292. Kreil, G. (1994) *J. Biol. Chem.* **269**, 10967–10970
293. Mor, A., Amiche, M., and Nicolas, P. (1992) *Trends Biochem. Sci.* **17**, 481–485
294. Kreil, G. (1994) *Science* **266**, 996–997
295. Prescott, B., Renugopalakrishnan, V., Glimcher, M. J., Bhushan, A., and Thomas, G. J., Jr. (1986) *Biochemistry* **25**, 2792–2798
296. Wood, D. D., and Moscarello, M. A. (1989) *J. Biol. Chem.* **264**, 5121–5127
297. Waite, J. H., Housley, T. J., and Tanzer, M. L. (1985) *Biochemistry* **24**, 5010–5014
298. Taylor, S. W., Waite, J. H., Ross, M. M., Shabanowitz, J., and Hunt, D. F. (1994) *J. Am. Chem. Soc.* **116**, 10803–10804
299. Clarke, S. (1992) *Ann. Rev. Biochem.* **61**, 355–386
300. Marshall, C. J. (1993) *Science* **259**, 1865–1866
301. Mroz, E. A. (1989) *Science* **243**, 1615
302. Rothgeb, T. M., England, R. D., Jones, B. N., and Gurd, R. S. (1978) *Biochemistry* **17**, 4564–4571
303. Lorimer, G. H. (1983) *Trends Biochem. Sci.* **8**, 65–68
304. Friedman, M. (1977) *Protein Crosslinking*, Vol. 2, Plenum, New York
305. Torchinsky, Y. M. (1981) *Sulfur in Proteins*, Pergamon, Oxford
306. Harrison, P. M., and Sternberg, M. J. E. (1994) *J. Mol. Biol.* **244**, 448–463
307. Folk, J. E., and Finlayson, J. S. (1977) *Adv. Prot. Chem.* **31**, 1–133
308. Gurd, F. R. N., and Rothgeb, T. M. (1979) *Adv. Prot. Chem.* **33**, 73–165
309. Karplus, M., and McCammon, J. A. (1983) *Ann. Rev. Biochem.* **53**, 263–300
310. Caspar, D. L. D., Clarage, J., Salunke, D. M., and Clarage, M. (1988) *Nature (London)* **332**, 659–662
311. Smith, J. L., Hendrickson, W. A., Honzatko, R. B., and Sheriff, S. (1986) *Biochemistry* **25**, 5018–5027
312. Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991) *Science* **254**, 1598–1603
313. Buck, M., Radford, S. E., and Dobson, C. M. (1994) *J. Mol. Biol.* **237**, 247–254
314. Kim, K.-S., Fuchs, J. A., and Woodward, C. K. (1993) *Biochemistry* **32**, 9600–9608
315. Rohl, C. A., and Baldwin, R. L. (1994) *Biochemistry* **33**, 7760–7767
316. Spyracopoulos, L., and O'Neil, J. D. J. (1994) *J. Am. Chem. Soc.* **116**, 1395–1402
317. Huber, R. (1979) *Trends Biochem. Sci.* **4**, 271–276
318. Lumry, R., and Rosenberg, A. (1974) *Colloques Internationaux du C.N.R.S.* **246**, 53
319. Mao, B., Pear, M. R., McCammon, J. A., and Northrup, S. H. (1981) *J. Mol. Biol.* **151**, 199–202
320. Brack, A., and Spach, G. (1981) *J. Am. Chem. Soc.* **103**, 6319–6323
321. Frauenfelder, H., Parak, F., and Young, R. D. (1988) *Ann. Rev. Biophys. Biophys. Chem.* **17**, 451–479
322. Gerstein, M., Lesk, A. M., and Chothia, C. (1994) *Biochemistry* **33**, 6739–6749
323. Kretsinger, R. H., and Nockolds, C. E. (1973) *J. Biol. Chem.* **248**, 3313–3326
324. Liepinsh, E., and Otting, G. (1994) *J. Am. Chem. Soc.* **116**, 9670–9674
325. Scholtz, J. M., Barrick, D., York, E. J., Stewart, J. M., and Baldwin, R. L. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 185–189
326. Kamoun, P. P. (1988) *Trends Biochem. Sci.* **13**, 424–425
327. Dill, K. A., and Shortle, D. (1991) *Ann. Rev. Biochem.* **60**, 795–825
328. Baker, D., and Agard, D. A. (1994) *Biochemistry* **33**, 7505–7509
328a. Sharp, K. A., and Englander, S. W. (1994) *Trends Biochem. Sci.* **19**, 526–529
328b. Pace, C. N., Shirley, B. A., McNutt, M., and Gajiwala, K. (1996) *FASEB J.* **10**, 75–83
328c. Sippl, M. J. (1996) *J. Mol. Biol.* **260**, 644–648
328d. Meyer, E. (1992) *Protein Sci.* **1**, 1543–1562
328e. Hill, R. B., Hong, J.-K., and DeGrado, W. F. (2000) *J. Am. Chem. Soc.* **122**, 746–747
329. Levinthal, C. (1968) *J. Chim. Phys.-Chim. Biol.* **65**, 44–45
330. Lattman, E. E., Fiebig, K. M., and Dill, K. A. (1994) *Biochemistry* **33**, 6158–6166
331. Mark, A. E., and van Gunsteren, W. F. (1992) *Biochemistry* **31**, 7745–7748
332. Peng, Z., and Kim, P. S. (1994) *Biochemistry* **33**, 2136–2141
333. Hagihara, Y., Tan, Y., and Goto, Y. (1994) *J. Mol. Biol.* **237**, 336–348
334. Yu, Y., Makhatadze, G. I., Pace, C. N., and Privalov, P. L. (1994) *Biochemistry* **33**, 3312–3319
335. Dobson, C. M., Evans, P. A., and Radford, S. E. (1994) *Trends Biochem. Sci.* **19**, 31–37
336. Yang, A.-S., and Honig, B. (1994) *J. Mol. Biol.* **237**, 602–614
336a. Baker, D. (2000) *Nature (London)* **405**, 39–42
336b. Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M., and Karplus, M. (2000) *Trends Biochem. Sci.* **25**, 331–339
337. Udgaonkar, J. B., and Baldwin, R. L. (1988) *Nature (London)* **335**, 694–699
338. Roder, H., Elöve, G. A., and Englander, S. W. (1988) *Nature (London)* **335**, 700–704
339. Gething, M.-J., and Sambrook, J. (1992) *Science* **355**, 33–45
340. Odefey, C., Mayr, L. M., and Schmid, F. X. (1995) *J. Mol. Biol.* **245**, 69–78
341. Chazin, W. J., Kördel, J., Drakenberg, T., Thulin, E., Brodin, P., Grundström, T., and Forsén, S. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2195–2198
342. Lin, L.-N., and Brandts, J. F. (1987) *Biochemistry* **26**, 3537–3543
343. Herning, T., Yutani, K., Taniyama, Y., and Kikuchi, M. (1991) *Biochemistry* **30**, 9882–9891
344. Chatrenet, B., and Chang, J.-Y. (1993) *J. Biol. Chem.* **268**, 20988–20996
345. Talluri, S., Rothwarf, D. M., and Scheraga, H. A. (1994) *Biochemistry* **33**, 10437–10449
346. Pace, C. N. (1990) *Trends Biochem. Sci.* **15**, 14–17
347. Yang, A.-S., and Honig, B. (1993) *J. Mol. Biol.* **231**, 459–474
348. Hartsough, D. S., and Merz, K. M., Jr. (1993) *J. Am. Chem. Soc.* **115**, 6529–6537
349. Aswad, D. W., and Johnson, B. A. (1987) *Trends Biochem. Sci.* **12**, 155–158
350. Brennan, T. V., and Clarke, S. (1993) *Protein Sci.* **2**, 331–338
351. Volkin, D. B., and Klibanov, A. M. (1987) *J. Biol. Chem.* **262**, 2945–2950
352. Wells-Knecht, M. C., Huggins, T. G., Dyer, D. G., Thorpe, S. R., and Baynes, J. W. (1993) *J. Biol. Chem.* **268**, 12348–12352
353. Luthra, M., Ranganathan, D., Ranganathan, S., and Balasubramanian, D. (1994) *J. Biol. Chem.* **269**, 22678–22682
354. Lubec, G., Weninger, M., and Anderson, S. R. (1994) *FASEB J.* **8**, 1166–1169
355. Kaiser, E. T. (1987) *Trends Biochem. Sci.* **12**, 305–309
355a. Beasley, J. R., and Hecht, M. H. (1997) *J. Biol. Chem.* **272**, 2031–2034
356. Oxender, D. E., and Fox, C. F., eds. (1987) *Protein Engineering*, Liss, New York
356a. Dahiyat, B. I., and Mayo, S. L. (1997) *Science* **278**, 82–87
356b. DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F., and Lombardi, A. (1999) *Ann. Rev. Biochem.* **68**, 779–819
357. Bryson, J. W., Betz, S. F., Lu, H. S., Suich, D. J., Zhou, H. X., O'Neil, K. T., and DeGrado, W. F. (1995) *Science* **270**, 935–941
358. Nautiyal, S., Woolfson, D. N., King, D. S., and Alber, T. (1995) *Biochemistry* **34**, 11645–11651
359. Betz, S. F., Liebman, P. A., and DeGrado, W. F. (1997) *Biochemistry* **36**, 2450–2458
360. Handel, T. M., Williams, S. A., and DeGrado, W. F. (1993) *Science* **261**, 879–885
361. Struthers, M. D., Cheng, R. P., and Imperiali, B. (1996) *Science* **271**, 342–345
362. Yan, Y., and Erickson, B. W. (1994) *Protein Sci.* **3**, 1069–1073
363. Quinn, T. P., Tweedy, N. B., Williams, R. W., Richardson, J. S., and Richardson, D. C. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 8747–8751
364. Hecht, M. H. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 8729–8730
364a. Kortemme, T., Ramírez-Alvarado, M., and Serrano, L. (1998) *Science* **281**, 253–256

## References


365. Kellis, J. T., Jr., Nyberg, K., and Fersht, A. R. (1989) *Biochemistry* **28**, 4914–4922
366. Goldenberg, D. P. (1988) *Ann. Rev. Biophys. Biophys. Chem.* **17**, 481–507
367. Alber, T. (1989) *Ann. Rev. Biochem.* **58**, 765–798
368. Katz, B. A., and Kossiakoff, A. (1986) *J. Biol. Chem.* **261**, 15480–15485
369. Pantoliano, M. W., Ladner, R. C., Bryan, P. N., Rollence, M. L., Wood, J. F., and Poulos, T. L. (1987) *Biochemistry* **26**, 2077–2082
370. Wetzel, R. (1987) *Trends Biochem. Sci.* **12**, 478–482
371. Mitchinson, C., and Wells, J. A. (1989) *Biochemistry* **28**, 4807–4815
372. Ahern, T. J., Casal, J. I., Petsko, G. A., and Klibanov, A. M. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84**, 675–679
373. Estell, D. A., Graycar, T. P., Miller, J. V., Powers, D. B., Burnier, J. P., Ng, P. G., and Wells, J. A. (1986) *Science* **233**, 659–663
374. Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E., and Hutchison, C. A., III. (1989) *Nature (London)* **340**, 397–440
375. Mendel, D., Ellman, J. A., Chang, Z., Veenstra, D. L., Kollman, P. A., and Schultz, P. G. (1992) *Science* **266**, 1798–1802
376. Brady, L., and Dodson, G. (1994) *Nature (London)* **368**, 692–694
377. Liu, N., Deillon, C., Klauser, S., Gutte, B., and Thomas, R. M. (1998) *Protein Sci.* **7**, 1214–1220

## Study Questions

1. Name all of the isometric tripeptides which could be formed from one molecule each of tyrosine, alanine, and valine.

2. What functional groups are found in protein side chains? Of what importance to protein structure and function are (a) hydrophobic groups, (b) acidic and basic groups, (c) sulfyhydryl groups?

3. If placed in water and adjusted to a pH of 7, will the following migrate toward the anode or the cathode if placed in an electrical field?
   (a) Aspartic acid, (b) alanine, (c) tyrosine, (d) lysine, (e) arginine, (f) glutamine.

4. Draw the following hydrogen-bonded structures:
   (a) A dimer of acetic acid.
   (b) A tyrosine–carboxylate bond in the interior of a protein.
   (c) A phosphate–guanidinium ion pair in an enzyme–substrate complex.

5. Contrast the properties of the amino acids with those of the saturated fatty acids with respect to solubility in water and in ether and to physical state.

6. Describe in as much detail as you can the characteristic properties of (a) β sheets, (b) α helices, (c) turns in peptide chains, and (d) collagen.

7. Predict whether the following peptide segments will be likely to exist as an α helix or as part of a β structure within a protein:
   (a) Poly-L-leucine
   (b) Poly-L-valine
   (c) Pro-Glu-Met-Val-Phe-Asp-Ile
   (d) Pro-Glu-Ala-Leu-Phe-Ala-Ala

8. Describe three ways in which a side chain of a serine residue can fold back and hydrogen bond to a C=O or N–H group of the backbone and two ways by which an asparagine side chain can do the same. There are yet other possibilities.

9. Compare structural features and properties of the following proteins: silk fibroin, α-keratin, collagen, and bovine serum albumin.

10. In what way do the solubilities of proteins usually vary with pH? Why?

11. Compare the following: the diameters of (a) a carbon atom in an organic molecule (b) a bacterial cell, e.g. of *E. coli* (c) a human red blood cell (d) a ribosome (e) the length of a peptide unit in an extended polypeptide chain (f) the length of the carbon atom chain in an 18-carbon fatty acid.

12. Compare: (a) the length of a peptide unit (residue) in a polypeptide in an extented (β) conformation. (b) the length by which an α helix is extended by the addition of one amino acid unit (c) the length of one turn of an α helix. (d) the diameter of an α helix (both using atom centers in the backbones and using van der Waals radii) for a poly-L-alanine helix.

13. What are: albumins, globulins, protamines, scleroproteins, glycoproteins, lipoproteins?

14. Where are the following found and what are their functions? Gamma globulin, hemocyanin, pepsin, glucagon, ferritin, phosphorylase.

15. List the nutritionally essential amino acids for human beings. Compare these needs with those of other species, including lactic acid bacteria, malaria parasites, green plants, etc.

16. Define: chiral, enantomer, diastereomer, epimer, anomer (see Chapter 4), prochiral (see Chapter 9). What is meant by the statement that biochemical reactions are stereochemically specific? Why is such stereospecificity to be expected in organisms (which are constructed from asymmetric units)? See Chapter 9 for further discussion.

17. What are disulfide bridges and of what significance are they in protein structure?

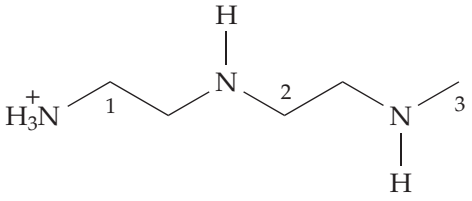
18. What is meant by "denaturation" of a protein? Mention several ways in which denaturation can be brought about. How is denaturation explained in terms of structure?

19. A chain of L-amino acids can form either a right-handed or a left-handed helix. From the Ramachandran diagram in Fig. 2-9, can you say anything about the relative stabilities of right and left-handed helices? What do you predict for polyglycine?

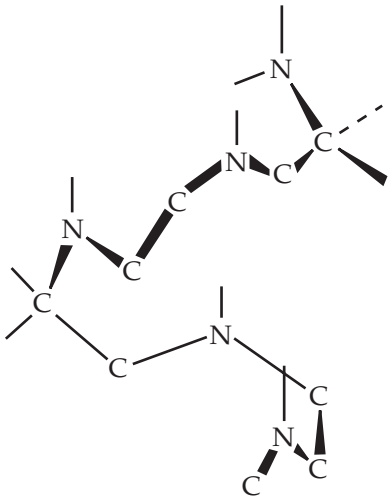
20. What similarities and differences would you predict for two proteins of identical amino acid sequence but one made from all L-amino acids and the other from all D-amino acids?

## Study Questions

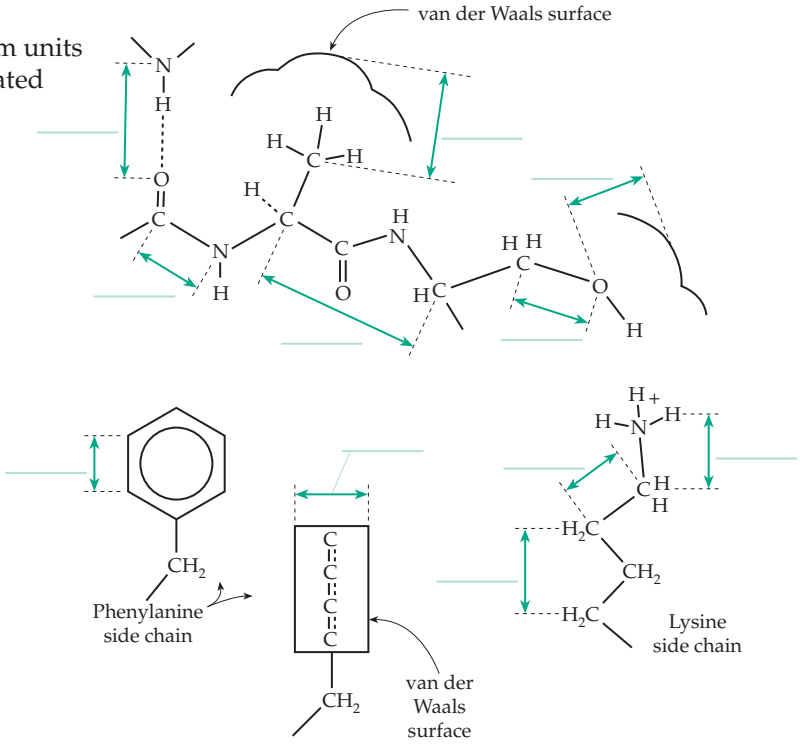
21. Complete the following peptide structure for L-seryl-L-valyl-L-asparaginyl—etc. Extend the chain in the C-terminal direction to form a **beta turn** with the chain coming back to form a **beta sheet**. Add a third segment of peptide *parallel* to the folded back chain to form a 3-stranded beta sheet. Indicate all hydrogen bonds correctly. Draw the side chains of the seryl and asparaginyl residues in positions 1 and 3 so that they form proper hydrogen bonds to groups in the peptide backbone.




22. Complete the following structure to form a short segment of alpha helix. Extend the chain by at least three residues. Form all hydrogen bonds correctly. Add at least three side chains with the correct chirality at the alpha carbon positions. Add one electrically charged side chain and show how it interacts with the peptide backbone at either the C- or N-terminus to help *stabilize* the helix.

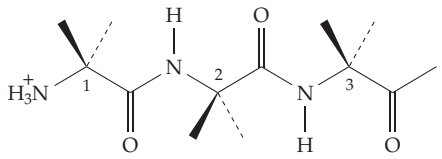


23. What is the relationship between Ångstrom units (Å) and nanometers (nm)? Give the indicated distance in Å or nm.

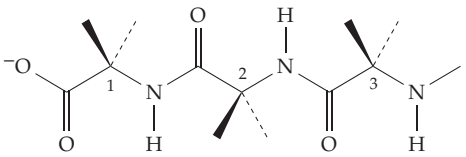


## Study Questions

24. Draw three residues of a polypeptide chain constructed of L-amino acids using the top template of the two below. Now, using the second template, whose polypeptide chain begins at the right, use the same three residues, *numbering from the C-terminus* using D-amino acids. This is know as a *retro*-inverso polypeptide.




Polypeptide

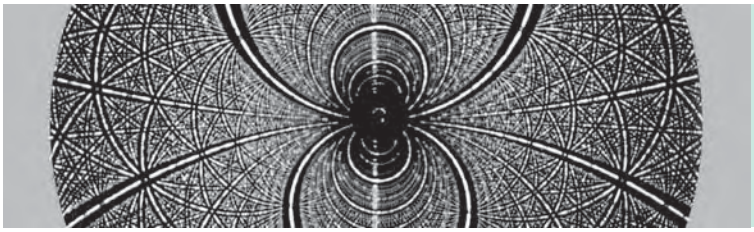



*Retro*-inverso polypeptide

Compare the outer surface of of the standard polypeptide and its *retro*-inverso analog. How would your answer be affected by the presence of threonine or isoleucine in the peptides?

If the two polypeptides were folded into a hairpin loop of β structure how would the exterior surfaces compare? How would the hydrogen bonding compare?

What possible value can you imagine for *retro*-inverso polypeptides in design of drugs? See Brady and Dodson, *Nature* **368**, 692–694 (1994) or Guichard *et al., Proc. Natl. Acad. Sci. USA* **91**, 9765–9769 (1994)

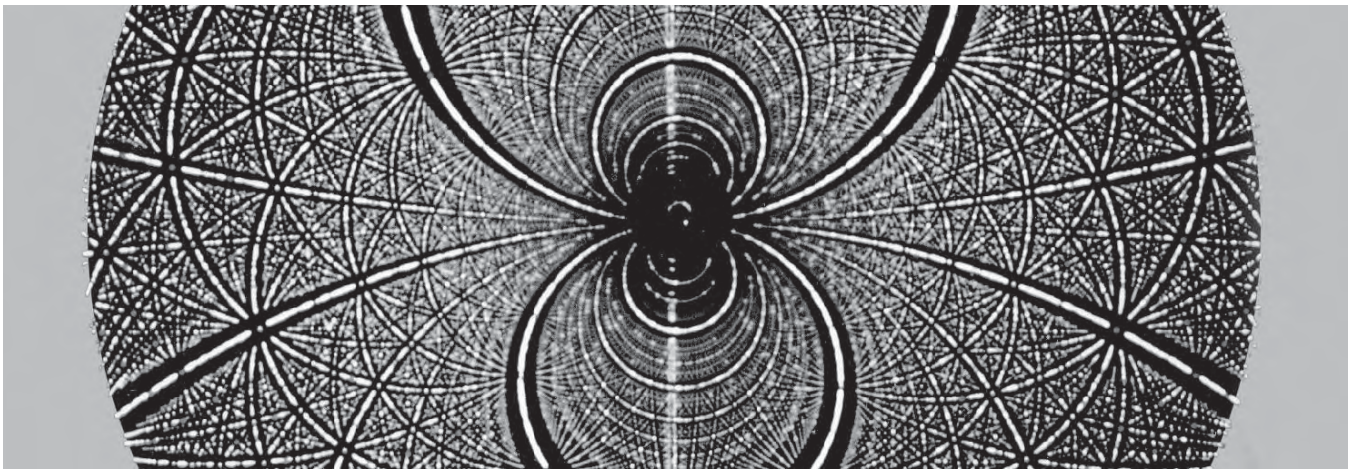


A Laue X-ray diffraction pattern from a protein crystal. A stationary crystal is irradiated with very intense white, multiwavelength X rays from a synchrotron source. The diffraction pattern is rich in information. A single 0.1 ms X-ray pulse may provide a pattern with enough information to determine a three-dimensional structure. The pattern consists of thousands of diffraction spots arranged on intersecting rings. The coordinates of the diffraction spots together with their measured intensities provide the necessary information for structure determination. Courtesy of Louise Johnson.

# Contents

# Determining Structures and Analyzing Cells

# 3



How have chemists deduced the thousands of structural formulas that we write for the substances found in nature? The answer is far too complex to give here in detail. However, the separation of compounds, the analysis of mixtures, and the unraveling of structures remain essential parts of biochemistry. A "minireview" of methods, with emphasis on proteins, follows. Additional procedures having to do primarily with carbohydrates, nucleic acids, or lipids are given in Chapters 4, 5, and 8, respectively.

## A. Understanding pH and Electrical Charges on Macromolecules

Proteins, nucleic acids, and carbohydrates all contain acidic or basic functional groups. The strengths of the acidic groups vary over a broad range from that of the strongly acidic phosphate and sulfate esters to that of the very weakly acidic alcoholic –OH group. The net electrical charge, as well as the spatial distribution of the charged groups, affects the properties of a macromolecule greatly. Therefore, it will be worthwhile for us to consider some aspects of acid–base chemistry before discussing other topics.

### 1. Strengths of Acids and Bases: the p$K_a$'s

The strength of an acid is usually described by the acid dissociation constant $K_a$

$$HA = A^- + H^+ \qquad (3\text{-}1)$$

$$K_a = [A^-][H^+] / [HA] \qquad (3\text{-}2)$$

For strong acids $K_a$ is high and for weak acids it is low. Since the values of $K_a$ vary by many orders of magnitude it is customary to use as a measure of the acid strength p$K_a$. This is the negative logarithm of $K_a$ (p$K_a$ = –log $K_a$) . For very strong acids p$K_a$ is less than zero, while very weak acids have p$K_a$ values as high as 15 or more.

In the biochemical literature *the strength of a base is nearly always given by the p$K_a$ of the conjugate acid.* Thus, $A^-$ in Eq. 3-1 is a base and HA its conjugate acid. The base could equally well be uncharged A and its conjugate acid $HA^+$. In both cases Eq. 3-2 would hold. This defines the strength of both the acid HA and the base $A^-$. It follows that strong bases have weak conjugate acids with high p$K_a$ values, while weak bases have strong conjugate acids with low p$K_a$ values.

For a compound containing several acidic groups we define a series of consecutive dissociation constants $K_{1a}$, $K_{2a}$, $K_{3a}$, etc. For the sake of simplicity we will omit the a's and call these $K_1$, $K_2$, $K_3$, ⋯.

$$H_3A \xrightarrow{K_1} H_2A \xrightarrow{K_2} HA \xrightarrow{K_3} A \qquad (3\text{-}3)$$

If there are $n$ consecutive dissociation constants there will be $n + 1$ ionic species $H_3A$, $H_2A$, etc. Notice that $H_3A$ could be a neutral molecule with $H_2A$, HA, and A bearing changes of –1, –2, and –3, respectively. Alternatively, $H_3A$ might carry 1, 2, or 3 positive changes. In every case the mathematical expressions will be the same. For this reason the charges have been deliberately omitted from Eq. 3-3 and others that follow. Each constant in Eq. 3-3 is defined as in Eq. 3-2, i.e., $K_2 = [HA][H^+] / [H_2A]$, etc. Keep in mind that there

are significant differences between the *apparent equilibrium constants* (concentration equilibrium constants) that we ordinarily use and thermodynamic equilibrium constants that are obtained by extrapolation to zero ionic strength.[1] A related complication is the uncertainity associated with the measurement of pH. This is often considered a measurement of hydrogen ion activity but this is not a correct statement. (The matter is considered briefly in Chapter 6). However, for all practical purposes, in the range of about pH 4–10 the pH can be equated with $-\log [H^+]$.

Often only one of the ionic forms of Eq. 3-3 will be important in a biochemical reaction. A particular ionic species may be the substrate for an enzyme. Likewise, an enzyme–substrate complex in only a certain state of protonation may react to given products. In these cases, and whenever pH affects an equilibrium, it is useful to relate the concentration $[A_i]$ of a given ionic form of a compound to the total of all ionic forms $[A]_t$ using Eqs. 3-4 and 3-5.

$$[A_i] = [A]_t / F_i \tag{3-4}$$

$$[A]_t = [A] + [HA] + [H_2A] + \cdots [H_nA] \tag{3-5}$$

For Eq. 3-4, $A_1 = H_nA$, $A_2 = H_{n-1}A$, etc. and $F_1$, $F_2$, etc. are the *Michaelis pH functions*,[2,3] which were proposed by L. Michaelis in 1914. For the case represented by Eq. 3-3 there are four ionic species and therefore four Michaelis pH functions which have the following form (Eq. 3-6). Here, $K_1$, $K_2$, etc. are the usual consecutive acid dissociation constants.

$$F_1 = 1 + K_1/[H^+] + K_1K_2/[H^+]^2 + K_1K_2K_3/[H^+]^3$$

$$F_2 = [H^+]/K_1 + 1 + K_2/[H^+] + K_2K_3/[H^+]^2$$

$$F_3 = [H^+]^2/K_1K_2 + [H^+]/K_2 + 1 + K_3/[H^+]$$

$$F_4 = [H^+]^3/K_1K_2K_3 + [H^+]^2/K_2K_3 + [H^+]/K_3 + 1 \tag{3-6}$$

If there are only three ionic forms the first three of these equations will apply if the final term is dropped from each. The student should be able to verify these equations and to write the appropriate pH functions for other cases. Since these relationships are met so often in biochemistry it is worthwhile to program a computer to evaluate the Michaelis pH functions and to apply them as needed. From Eq. 3-4 it can be seen that *the reciprocal of the Michaelis pH function for a given ionic form represents the fraction of the total compound in that form* and that the sum of these reciprocals for all the ionic forms is equal to one. Examples of the use of the Michaelis pH functions in this book are given in Eq. 6-50, which relates the Gibbs energy of hydrolysis of ATP to the pH, and in Eqs. 9-55 to 9-57, which describe the pH dependence of enzymatic action.

Notice that in Eq. 3-3 single arrows have been used rather than the pairs ($\rightleftharpoons$) that are often employed to indicate reversible equilibria. This is done so that *the direction of the arrow indicates whether we are using a dissociation constant or an association constant*. The use of single arrows in this manner to indicate how the equilibrium constants are to be written is a good practice when dealing with complex equilibria.

## 2. Titration Curves

When a neutral amino acid is titrated with acid, the carboxylate groups become protonated and acid is taken up. Likewise titration with base removes protons from the protonated amino groups and base is taken up. If we plot the number of equivalents of acid or base that have reacted with the neutral amino acid versus pH, a titration curve such as that shown in Fig. 3-1 is generated. The curve for histidine contains three steps; the first corresponds to the titration of the carboxylate group with acid, the second to the titration of the protonated imidazole of the side chain, and the third to the titration of the protonated amino group with base. Each step is characterized by a midpoint that is equal to the $pK_a$ value for the group being titrated. The ends of the curve at low and high pH, which are drawn with a dashed line, are obtained only after corrections have been applied to the data. If only the equivalents of acid or base *added* rather than the number *reacted* are plotted, we obtain the solid line shown in Fig. 3-1. We see that at the low pH end there is no distinct end point. As we add more acid to try to complete the titration, the correction that must be applied to the data becomes increasingly greater. The difference between the dashed and solid lines reflects the fact that at the low pH end much of the acid added is used to simply lower the pH. Therefore, we have a large free $[H^+]$. Similarly, at the high pH end we have a high free $[OH^-]$.

The exact shapes of the ends of the titration curve depend heavily on the total concentration. Likewise, the magnitude of the correction required to obtain a plot of equivalents of acid or base reacted varies with the concentration and is smaller the higher the concentration of the substance being titrated (see problems 2 and 3 at the end of this chapter). An important rule in doing acid-base titrations, especially when very small samples are available, is to use the highest possible concentration of sample in the smallest possible volume and to titrate with relatively concentrated acid or base. Because of the difficulty of adequately correcting titration curves at low pH it is hard to estimate the $pK_a$ values of the carboxyl groups of amino acids accurately from titration. An additional experimental difficulty exists at the high pH end because of the tendency for basic

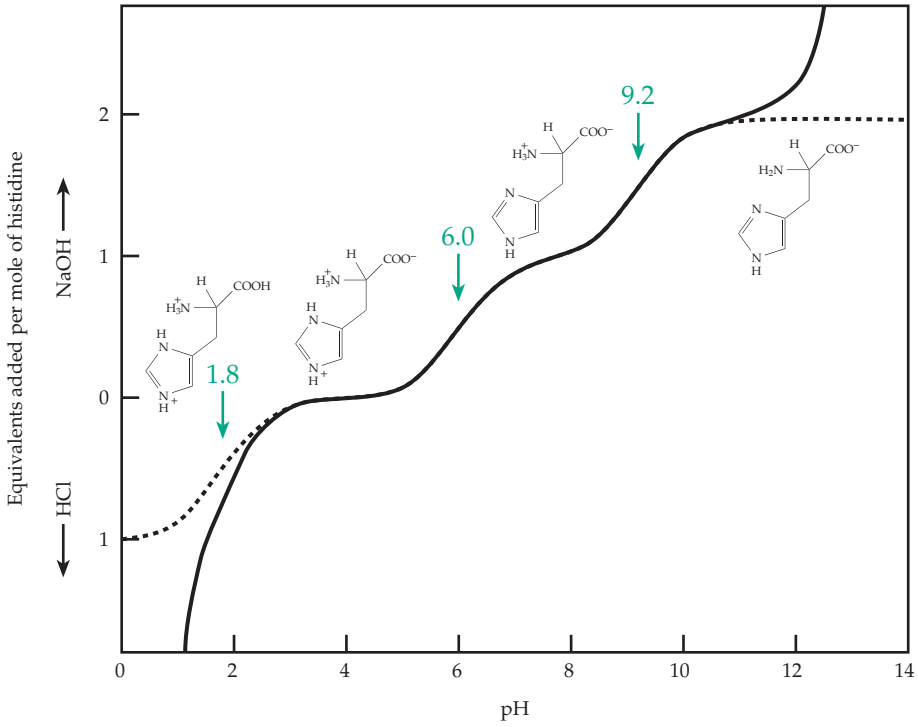



**Figure 3-1** Titration curve for histidine. The solid line represents the uncorrected titration curve for 3 mM histidine monohydrochloride titrated with 0.2 M HCL to lower pH or with 0.2 M NaOH to higher pH assuming p$K_a$ values of 1.82, 6.00, and 9.17. The dashed line represents the corrected curve showing the number of protons bound or lost per mole of histidine monohydrochloride.

solutions to absorb $CO_2$ from the air. Sometimes formaldehyde is added to shift the apparent p$K_a$ of the amino groups to lower values and make the titration more satisfactory.

Despite the difficulties, real proteins can be titrated successfully to estimate the numbers of carboxyl, histidine, tyrosine, and other groups.[4] An example is shown in Fig. 3-2. The experimental data have been fitted with a theoretical curve based on the p$K_a$'s of the carboxyl, histidine, tyrosine, and amino groups as determined by NMR measurements.[5] Account has been taken of the effect of the electrical field created by the many charged groups in distorting the curve from that obtained by summing the theoretical titration curves of the component groups. However, when there are multiple acid–base groups that are close together in a protein, a more complex situation involving tautomerism arises. This is discussed in Chapter 6.

Titration curves based on plots of light absorption versus pH or of NMR chemical shifts versus pH (see Fig. 3-29) are often useful. They have the important advantage that no special correction for free acid or base is needed at low or high pH.

## 3. Buffers

A mixture of a weak acid HA and of its **conjugate base** A constitutes a buffer which resists changes in pH. This can be seen most readily by taking logarithms of both sides of Eq. 3-2. By replacing log $K$ with $-$p$K_a$ and log [H$^+$] with $-$pH and rearranging we obtain Eq. 3-7 (the Henderson–Hasselbalch equation). It is sometimes useful to rewrite this as Eq. 3-8, where $\alpha$ is the fraction of the acid that is dissociated at a given pH.

$$\text{pH} - \text{p}K_a = \log([\text{A}]/[\text{HA}]) \qquad (3\text{-}7)$$

$$\text{pH} - \text{p}K_a = \log[\alpha/(1-\alpha)];$$

$$\alpha = [\text{A}] \,/\, [\text{HA}] + [\text{A}] = \frac{10^{(\text{pH}-\text{p}K_a)}}{1 + 10^{(\text{pH}-\text{p}K_a)}} \qquad (3\text{-}8)$$

Logarithms to the base 10 are used in both equations. These equations are useful in preparing buffers and in thinking about what fraction of a substance exists in a given ionic form at a particular value of pH. From Eq. 3-7 it is easy to show that when the pH is near the p$K_a$ relatively large amounts of acid or base must be added to change the pH if the concentrations of the buffer pair A and HA are high.

Buffers are often added to maintain a constant pH in biochemical research[6] and naturally occurring buffer systems within body fluids and cells are very important (Box 6-A). Among the most important natural buffers are the proteins themselves, with the imidazole groups of histidine side chains providing much of the buffering capacity of cells around pH 7 (Figs. 3-1 and 3-2). Table 3-1 lists some useful biochemical buffers and their p$K_a$ values. Here are a few

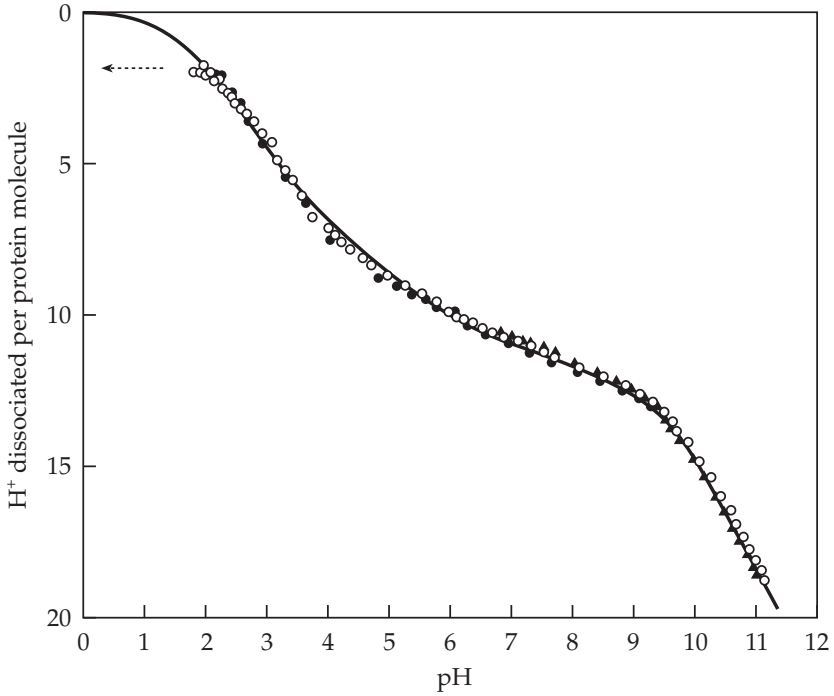



**Figure 3-2** Acid–base titration curve for hen lysozyme at 0.1 ionic strength and 25°C. ○, initial titration from the pH attained after dialysis; ●, back titration after exposure to pH 1.8; ▲, back titration after exposure to pH 11.1. The solid curve was constructed on the basis of "intrinsic" p$K_a$ values based on NMR data. From Kuramitsu and Hamaguchi[5]

practical hints about buffer preparation. Buffers containing monovalent ions tend to change pH with dilution less than do those with multivalent ions such as $HPO_4^{2-}$ or $HP_2O_7^{3-}$. The p$K_a$ values of carboxylic acids and of phosphoric acid, or of its organic derivatives, change very little with change in temperature. The pH of a buffer prepared with such components is nearly independent of temperature (Table 3-1). However, the p$K_a$ of the $-NH_3^+$ group changes greatly with temperature. Buffer composition can be calculated readily from Eq. 3-8 and p$K_a$ values from Table 3-1. It is convenient to keep in the laboratory standardized (to ~1% error) 1 M HCl and 1 M NaOH for use in buffer preparation. Compositions calculated from Eq. 3-8 will usually yield buffers of pH very close to those expected. Final adjustment with HCl may be needed if the pH is more than one unit away from a p$K_a$ value. When two buffering materials are present, the composition should be calculated independently for each. The measurement of pH should always be done with great care because it is easy to make errors. Everything depends upon the reliability of the standard buffers used to calibrate the pH meter.[7] Often, especially during isolation of small compounds, it is desirable to work in the neutral pH region with volatile buffers, e.g., trimethylamine and $CO_2$ or ammonium bicarbonate, which can be removed by vacuum evaporation or lyophilization.[8]

## B. Isolating Compounds

Before structural work can begin, pure substances must be separated from the complex mixtures in which they occur in cells and tissues.[4,13–24] Often, a substance must be isolated from a tissue in which it is present in a very low concentration. After it is isolated in pure form, if it is a large molecule, it must often be cut up into smaller pieces which are separated, purified, and identified. Accurate quantitative analysis is required to determine the ratios of these fragments. Considerable ingenuity may then have to be exercised in putting the pieces of the jigsaw puzzle back together to determine the structure of the "native" molecule. Many books, a few of which are cited here,[4,13–23,25–45] provide instructions. There are also journals and other periodicals dedicated to biochemical methods.[46–53]

### 1. Fractionation of Cells and Tissues

A fresh tissue or a paste of packed cells of a microorganism, usually collected by centrifugation, may be the starting material.[23,54–57] Tissue is often ground in a kitchen-type blender or, for gentler treatment, in a special **homogenizer**. The popular Potter–Elvehjem homogenizer is a small apparatus in which a glass or plastic pestle rotates inside a tight-fitting mortar tube (see standard laboratory equipment catalogs for pictures). Microbial cells are frequently broken with supersonic oscillation (**sonication**) or in special pressure cells. It is important to pay attention to the pH, buffer composition, and, if subcellular organelles are to be separated, the osmotic pressure. To preserve the integrity of organelles, 0.25 M sucrose is frequently used as the suspending medium, and $MgCl_2$ as well as a metal complexing agent such as ethylenediaminetetraacetate (EDTA) (Table 6-10) may be added. Soluble enzymes are often extracted without addition of sucrose, but reducing compounds such as glutathione (Box 11-B), mercaptoethanol, or dithiothreitol (Eq. 3-23) may be added. The crude **homogenate** may be strained and is usually centrifuged briefly to remove cell fragments and other "debris." Large-scale purification of

**TABLE 3-1**
**Practical p$K_a$ Values for Some Useful Buffer Compounds at 25°C and Ionic Strength 0.1**[a]

| Compound | p$K_a$ | Grams per mole | d(p$K_a$) / dT | Charge on conjugate base |
|---|---|---|---|---|
| Citric acid (p$K_1$) | | 192 | | −1 |
| Formic acid | 3.7 | 0 | −1 | |
| Citric acid (p$K_2$) | 4.45 | 192 | −0.0016 | −2 |
| Acetic acid | 4.64 | 60 | 0.0002 | −1 |
| Succinic acid (p$K_2$) | 5.28 | 118 | 0 | −2 |
| Citric acid (p$K_3$) | 5.80 | 192 | 0 | −3 |
| 3,3-Dimethylglutaric acid | 5.98 | 160 | 0.006 | −2 |
| Piperazine (p$K_1$) (5.68) | 6.02 | 86 | | 0 |
| Cacodylic acid (dimethylarsinic acid) | 6.1 | 138 | | −1 |
| MES[b] | 6.1 | 195 | −0.011 | |
| BIS-TRIS[b] | 6.41 | 209 | −.017 | |
| Carbonic acid (p$K_1$) | 6.4[c] | | | |
| Pyrophosphoric acid (p$K_3$) | 6.76 | 178 | | -2 |
| Phosphoric acid (p$K_2$) | 6.84 | 98 | −.0028 | −2 |
| PIPES[b] | 6.90 | 353 | −.0085 | −2 |
| Imidazole | 7.07 | 68 | −.020 | 0 |
| BES[b] | 7.06 | 213 | −.016 | −1 |
| Diethylmalonic acid | 7.2 | 136 | | |
| TES[b] | 7.37 | 279 | −.020 | −1 |
| HEPES[b] | 7.46 | 238 | −.014 | −1 |
| N-Ethylmorpholine | 7.79 | 115 | −.022 | 0 |
| Triethanolamine | 7.88 | 149 | −.020 | 0 |
| TRICINE[b] | 8.02 | 178 | −.021 | −1 |
| TRIS[b] | 8.16 | 121 | −.031 | 0 |
| Glycylglycine | 8.23 | 132 | −.028 | −1 |
| BICINE[b] | 8.26 | 163 | −.018 | −1 |
| 4-Phenolsulfonic acid | 8.70 | 174 | −.013 | −2 |
| Diethanolamine | 9.00 | 105 | −.024 | 0 |
| Ammonia | 9.2 | | −0.031 | 0 |
| Boric acid | 9.2 | | −0.008 | mixed |
| Pyrophosphoric acid (p$K_4$) | 9.41 | 178 | | -3 |
| Ethanolamine | 9.62 | 61 | −.029 | 0 |
| Glycine (p$K_2$) | 9.8 | 75 | −0.025 | −1 |
| Piperazine (p$K_2$) | 9.82 | 86 | | 0 |
| Carbonic acid (p$K_2$) | 10.0 | | −0.009 | −2 |
| Piperidine | 11.1 | 85 | | 0 |

[a] Based on compilation by Ellis and Morrison[9] with additional data from Good *et al.*[10,11] and Dawson *et al.* [12] The Good buffers have dipolar ionic constituents. Since no form is without electrically charged groups they are unlikely to enter and disrupt cells.

[b] Abbreviations used:

| | |
|---|---|
| BES | *N*,*N*-Bis(2-hydroxyethyl)-2-aminoethanesulfonic acid |
| BICINE | *N*,*N*-Bis(2-hydroxyethyl)glycine |
| BIS-TRIS | Bis(2-hydroxyethyl)iminotris(hydroxymethyl)methane |
| HEPES | *N*-2-Hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid |
| MES | 2-(*N*-Morpholino)ethanesulfonic acid |
| PIPES | Piperazine-*N*,*N*′-bis(2-ethanesulfonic acid) |
| TES | *N*-Tris(hydroxymethyl)methyl-2-aminoethanesulfonic acid |
| TRICINE | *N*-Tris(hydroxymethyl)methylglycine |
| TRIS | Tris (hydroxymethyl) aminomethane |

[c] For $CO_2$ (solid) + $H_2O \rightarrow H^+ + HCO_3^-$, apparent p$K_a$.

proteins is often initiated with such a crude homogenate.

Cell organelles are also often separated by centrifugation. In one procedure a homogenate in 0.25 M sucrose (**isotonic** with most cells) is centrifuged for 10 min at a field of 600–1000 times the force of gravity (600–1000 *g*) to sediment nuclei and whole cells. The supernatant fluid is then centrifuged another 10 min at ~10,000 *g* to sediment mitochondria and lysosomes. Finally, centrifugation at ~100,000 *g* for about an hour yields a pellet of microsomes (p. 14),[58] which contains both membrane fragments and ribosomes. Each of the separated components can be resuspended and recentrifuged to obtain cleaner preparations of the organelles. The sedimented particles can often be solubilized by chemical treatment, for example, by the addition of either ionic or nonionic detergents. Membrane proteins can be isolated following solubilization in this way (Chapter 8). The soluble supernatant fluid remaining after the highest speed centrifugation provides the starting material for isolation of soluble enzymes and many small molecules.

## 2. Separations Based on Molecular Size, Shape, and Density

The simplest way to separate very large dissolved molecules is to let the small ones pass through a suitable sieve which may be a membrane with holes or a bed of gel particles. If the size of the particles approaches that of the holes in the sieve, the rate of passage will depend upon shape as well as size.

***Dialysis, ultrafiltration, and perfusion chromatography.*** In dialysis[59] and ultrafiltration,[60] a thin membrane, e.g., made of cellulose acetate (cellophane) and containing holes 1–10 nm in diameter (typically 5 nm), is used as a semipermeable barrier. Small molecules pass through but large ones are retained. Dialysis depends upon diffusion and can be hastened by adequate stirring. Ultrafiltration requires a pressure difference across the membrane. The more sophisticated procedures of gel filtration and perfusion chromatography were introduced in 1959.[61–65] A column is packed with material such as the crosslinked dextran Sephadex, polyacrylamide gels (such as the Bio-Gel P Series), or agarose gels (e.g., Bio-Gels A and Sepharoses). These come in the form of soft beads, the interior network of which is a three-dimensional network of polymer strands (Fig. 4-10). Recently porous beads of hard crosslinked polystyrene, glass, or various other silicate materials have been employed.[66–68] The interstices between strands, whose size depends upon the degree of crosslinking introduced chemically into the gel, are small enough to exclude large molecules but to admit smaller ones. If a mixture of materials of different molecular size is passed through such a column the smaller molecules are retarded because of diffusion into the gel, while the larger molecules pass through unretarded (Fig. 3-3). Sephadex G-25 excludes all but salts and compounds no larger than a simple sugar ring. Sephadex G-200, which is much less crosslinked, permits separation of macromolecules in the range of 5–200 kDa. As is explained in Section B, gel filtration also provides an important way of estimating $M_r$ for proteins and other macromolecules.

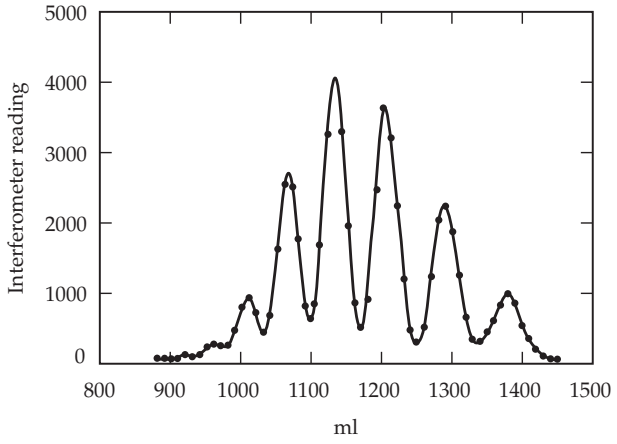



**Figure 3-3** Separation of oligosaccharides by gel filtration. The sugars dissolved in distilled water were passed through a column of Sephadex G-25. The peaks contain (right to left) glucose, cellobiose, cellotriose, etc. From Flodin and Aspberg.[64]

***Centrifugation.*** Centrifuges of many sizes and speeds are used in the laboratory to remove debris as well as to collect precipitated proteins and other materials at various steps in a purification scheme. The most remarkable are the **ultracentrifuges** which produce forces greater than $4 \times 10^5$ times that of gravity. They can be used both for separation of molecules and for determination of $M_r$ (see Section C).

When macromolecules in a solution are subjected to an ultracentrifugal field they are accelerated rapidly to a constant velocity of sedimentation. This is expressed as a **sedimentation constant *s***, which is the rate (cm/s) per unit of centrifugal force. The unit of *s* is the second but it is customary to give it in Svedberg units, S (1S= $10^{-13}$*s*). Sizes of particles are often cited by their S values. The sedimentation constant is affected by the sizes, shapes, and densities of the particles as is discussed further in Section C. If carried out at constant velocity an equilibrium will eventually be attained in which sedimentation is just balanced by diffusion and a smooth concentration gradient forms from the top to the bottom of the centrifuge cell or tube. Concentration gradients can also be formed by centrifuging a

concentrated solution of small molecules.[13,69,70] Such a concentration gradient is also a **density gradient** which can be made very steep. For example, a gradient with over a 10% increase in density from top to bottom can be created using 6 M **cesium chloride** (CsCl) and is widely used in DNA separations. If DNA is added prior to centrifugation it will come to rest in a narrow band or bands determined by the **buoyant densities** of the species of DNA present (see Chapter 5). After centrifugation, which is usually done in a plastic tube, a hypodermic needle is inserted through the bottom of the tube and the contents are pumped or allowed to flow by gravity into a fraction collector.

Another type of gradient centrifugation (**zone centrifugation**) utilizes a preformed gradient to stabilize bands of cell fragments, organelles, or macromolecules as they sediment.[71–74] For example, RNA can be separated into several fractions of differing sedimentation constants in a centrifuge tube that contains **sucrose** ranging in concentration from 25% at the bottom to 5% at the top. This is prepared by a special mixing device or "gradient maker" prior to centrifugation. The solution of RNA is carefully layered on the top, the tube is centrifuged at a high speed for several hours, and the different RNA fractions separate into slowly sedimenting sharp bands. A 20–60% gradient of sucrose or glycerol may be used in a similar way to separate organelles.[58]

## 3. Separations Based on Solubility

Some fibrous proteins are almost insoluble in water and everything else can be dissolved away. More often soluble proteins are precipitated from aqueous solutions by adjustment of the pH or by addition of large amounts of salts or of organic solvents. The solubility of any molecule is determined both by the forces that hold the molecules together in the solid state and by interactions with solvent molecules and with salts or other solutes that may be present.[58a]

Proteins usually have many positively and negatively charged groups on their surfaces. If either a positive or a negative charge predominates at a given pH the protein particles will tend to repell each other and to remain in solution. However, near the isoelectric point (pI), the pH at which the net charge is zero (see Section 7), the solubility will usually be at a minimum. The pH of a tissue extract may be adjusted carefully to the pI of a desired protein. Any protein that precipitates can be collected by centrifugation and redissolved to give a solution enriched in the protein sought. Some proteins, such as those classified (by an old system) as **globulins**, are insoluble in water but are readily **"salted in"** by addition of low concentrations (e.g., up to 0.1 M) of salts. Low concentrations of salts increase the solubility of most proteins because the salt ions interact with the charged groups on the protein surfaces and interfere with strong electrostatic forces that are often involved in binding protein molecules together in the solid state. Some salts, including $CaCl_2$ and NaSCN, which bind to proteins, are especially effective in salting in.[75,76] Addition of *high* concentrations of salt causes precipitation of most proteins from aqueous solutions. The most effective and most widely used materials for this **"salting out"** of proteins are $(NH_4)_2SO_4$ and $Na_2SO_4$. Because the salt ions interact so strongly with water, the protein molecules interact less with water and more with each other. A similar intramolecular effect may cause the stabilization of proteins of halophilic bacteria by 1–4 M KCl.[77]

Different proteins precipitate at different concentrations of an added salt. Hence, a fraction of proteins precipitating between two different concentrations of salt can be selected for further purification (Fig. 3-4). Protein concentrations can be estimated as described in Box 3-A.

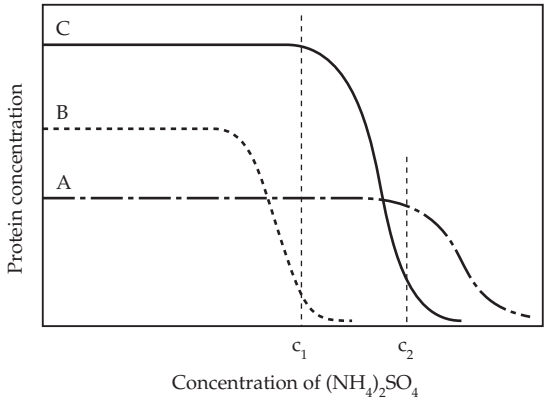



**Figure 3-4** Hypothetical behavior of a solution containing three proteins, A, B, and C, upon ammonium sulfate fractionation. The concentration of protein remaining in the solution is plotted against ammonium sulfate concentration (usually expressed as % saturation). Addition of ammonium sulfate to concentration $c_1$ will precipitate largely protein B, which can be removed by centrifugation. Addition of additional salt to $c_2$ will precipitate largely protein C, while A remains in solution.

Precipitation methods are popular first steps in purification of proteins because they can be carried out on a large "batch" scale. The amounts of ammonium sulfate used are often expressed as percentage saturation, i.e., as a percentage of the amount required to saturate the solution (4.1 M at 25°). Convenient tables are available[12,78] that permit one to weigh out the correct amount of solid ammonium sulfate to give a desired percentage saturation or to go from one

percentage saturation to a higher one.

Proteins are often stabilized by low concentrations of simple alcohols or ketones[76] and by higher concentrations of polyhydroxy alcohols, such as glycerol[77] and sucrose,[78] and also by certain inert, synthetic polymers such as **polyethyleneglycol** (PEG).[79] The latter is a widely used precipitant. The polyhydroxyalcohols and PEG are all hydrated but tend not to interact strongly with the protein molecules. On the other hand, simple alcohols may denature proteins by their interaction with nonpolar regions.[77]

## 4. Separation by Partition

Many of the most important separation methods are based on repeated equilibration of a material between two separate phases, at least one of which is usually liquid. Small molecules may be separated by **countercurrent distribution** in which a material is repeatedly partitioned between two immiscible liquid phases, one more polar than the other. New portions of both liquids are moved by a machine in a "countercurrent manner" between the equilibration steps or are moved continuously through coiled tubes.[80–82,82a]

A similar result is accomplished by using as one phase a solid powder or fine "beads" packed in a vertical column or spread in a thin layer on a plate of glass. The methods are usually referred to as **chromatography**, a term proposed by Tswett to describe separation of materials by color. In 1903 Tswett passed solutions of plant leaf pigments (chlorophylls and carotenes) in nonpolar solvents such as hexane through columns of alumina and of various other adsorbents and observed separation of colored bands which moved down the column as more solvent was passed through. Individual

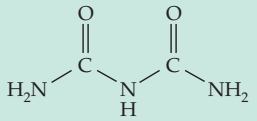
## BOX 3-A QUANTITATIVE ESTIMATION OF PROTEIN CONCENTRATIONS

Biochemists often need to estimate the content of protein in a sample. For example, in devising a purification procedure for an enzyme it is customary to estimate the number of units of enzyme activity (as defined in Chapter 9) per milligram of protein (U/ mg). As progress is made in the purification this ratio increases. It becomes constant with respect to additional purification attempts, when a homogeneous enzyme is obtained.

One of the most widely used and most sensitive protein assays (for 0.1–1 mg/ml of protein) is the colorimetric procedure of **Lowry**.[a–c] It makes use of a phosphomolybdic–phosphotungstic acid reagent (the Folin–Ciocalteu reagent) which is reduced by proteins in the presence of alkaline $Cu^{2+}$ to characteristic "blue oxides" whose color can be monitored at 750 nm. Much of the color comes from the reducing action of tyrosine and tryptophan. The color yield varies greatly from protein to protein and users may have trouble with reproducibility. A related method utilizes **bicinchoninic acid** which forms a purple color (measured at 362 nm) with the $Cu^{+1}$ that is formed by reduction of alkaline $Cu^{2+}$ by the protein.[d–f] This reagent is easier to use than that of the Lowry procedure and gives stable and reproducible readings.

A third widely used procedure, introduced by **Bradford**[g] and modified by others, measures the binding of the dye Coomassie brilliant blue whose peak absorption shifts from 465 nm to 595 nm upon binding. The change occurs within two minutes and is stable. However, the color yield varies from one protein to another.

Less sensitive but very simple and precise is measurement of the light absorption around 280 nm. This is discussed in the main text in Section D,6. For a typical protein an absorbance of 1.0 at 280 nm corresponds to a protein concentration of 1 mg / ml.[h] The very old **biuret method** is also useful for samples containing 1–10 mg / ml of protein. The violet color that arises upon addition of copper sulfate to an alkaline solution of a peptide or protein is recorded at 540–560 nm.[h] The color is especially intense for longer polypeptides. The name of the method arises from the fact that biuret gives a similar color[i] (see also Eq. 6-85).



Biuret

[a] Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951) *J. Biol. Chem.* **193**, 265–275
[b] Peterson, G. L. (1979) *Anal. Biochem.* **100**, 201–220
[c] Larson, E., Howlett, B., and Jagendorf, A. (1986) *Anal. Biochem.* **155**, 243–248
[d] Smith, P. K., Krohn, R. I., Hermanson, G. T., Mallia, A. K., Gartner, F. H., Provenzano, M. D., Fujimoto, E. K., Goeke, N. M., Olson, B. J., and Kenk, D. C. (1985) *Anal. Biochem.* **150**, 76–85
[e] Davis, L. C., and Rodke, G. A. (1987) *Anal. Biochem.* **161**, 152–156
[f] Hill, H. D., and Straka, J. G. (1988) *Anal. Biochem.* **170**, 203–208
[g] Bradford, M. M. (1976) *Anal. Biochem.* **72**, 248–254
[h] Fruton, J. S., and Simmonds, S. (1958) *General Biochemistry*, 2nd ed., Wiley, New York (p. 130)
[i] Layne, E. (1957) *Methods Enzymol.* **III**, 450–451

pure pigments could be **eluted** from the column by continued passage of solvent. This important method is called **adsorption chromatography**. It is assumed that the pigments are absorbed on the surface rather than being disssolved in the solid material. A related method is **hydrophobic interaction chromatography** of proteins.[83–86] The packing material is similar to that described in Section 6 (Affinity Chromatography) but bears long-chain alkyl groups that can interact with the hydrophobic patches on surfaces of proteins. A very different adsorbent that is very useful in separation of proteins is carefully prepared microcrystalline **hydroxylapatite.**[87,88] It presumably functions in part by ion exchange.

Column packing materials such as **silica gel** contain a large amount of water, and separation involves partition between an immobilized aqueous phase in the gel and a mobile, often organic, solvent flowing through the column. Usually materials elute sooner when they are more soluble in the mobile phase than in the aqueous phase. These methods are closely related to perfusion chromatography, which is described in Section 2.

Aromatic amino acids, lipids, and many other materials can be separated on **reversed-phase** columns in which nonpolar groups, usually long-chain alkyl groups, are covalently attached to silica gel, alumina, or other inert materials.[66,80] The mobile phase is a more polar solvent, often aqueous, and gradually made less polar by addition of an organic solvent. In reversed-phase chromatography more polar compounds migrate faster through the system than do nonpolar materials, which experience hydrophobic interaction with the solid matrix.

Many traditional chromatographic methods including reversed-phase chromatography have been adapted for use in automatic systems which employ columns of very finely divided solid materials such as silica, alumina, or ion exchange materials coated onto fine glass beads.[89–91] These **high-performance liquid chromatographic** (**HPLC**) systems often utilize pressures as high as 300 atmospheres. Separations are often sharper and faster than with other chromatographic methods.[92–97] Reversed-phase columns in which the solid matrix may carry long (e.g. $C_{18}$) hydrocarbon chains have been especially popular for separation both of peptides and of proteins. Proteins may also be separated by gel filtration, ion exchange, or other procedures with HPLC equipment.

A sheet of high-quality filter paper containing adsorbed water serves as the stationary phase in **paper chromatography**. However, **thin-layer chromatography**, which employs a layer of silica gel or other material spread on a glass or plastic plate, has often supplanted paper chromatography because of its rapidity and sharp separations (Fig. 3-5).[16,96a,98–100] An approach that requires no stationary phase at all is **field flow fractionation**.[101] Here a suitable external field (e.g., electrical, magnetic, and centrifugal) or a thermal gradient is imposed on the particles flowing through a narrow channel.

For volatile materials **vapor phase chromatography** (gas chromatography) permits equilibration between the gas phase and immobilized liquids at relatively high temperatures. The formation of volatile derivatives, e.g., methyl esters or trimethylsilyl derivatives of sugars, extends the usefulness of the method.[103,104] A method which makes use of neither a gas nor a liquid as the mobile phase is **supercritical fluid chromatography**.[105] A gas above but close to its critical pressure and temperature serves as the solvent. The technique has advantages of high resolution, low temperatures, and ease of recovery of products. Carbon dioxide, $N_2O$, and xenon are suitable solvents.

## 5. Ion Exchange Chromatography

Separation of molecules that contain electrically charged groups is often accomplished best by ion exchange chromatography.[105a] The technique depends upon interactions between the charged groups of the molecules being separated and fixed ionic groups on

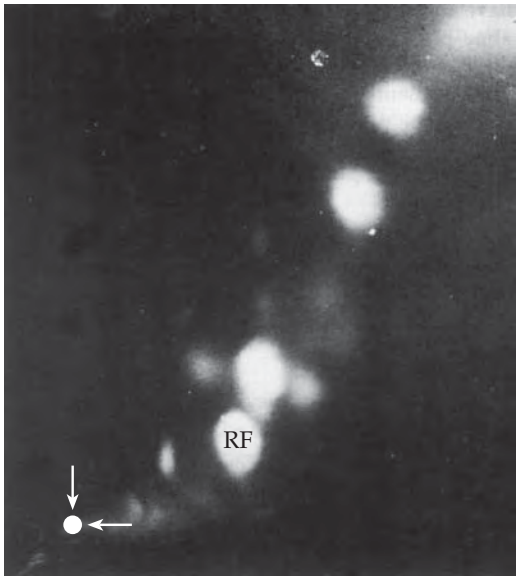



**Figure 3-5** Photograph of a two-dimensional thin layer (silica gel) chromatogram of a mixture of flavins formed by irradiation of ~10 μg of the vitamin riboflavin. The photograph was made by the fluorescence of the compounds under ultraviolet light. Some riboflavin (RF) remains. The arrows indicate the location of the sample spot before chromatography. Chromatography solvents: a mixture of acetic acid, 2-butanone, methanol, and benzene in one direction and *n*-butanol, acetic acid, and water in the other. See Treadwell *et al*.[102]

an immobile matrix. Separation depends upon small differences in p$K_a$ values and net charges and upon varying interactions of nonpolar parts of the molecules being separated with the matrix. Since changes in pH can affect both the charges on the molecules being separated and those of the ion exchange material, the affinities of the molecules being separated are strongly dependent on pH. For example, proteins and most amino acids are held tightly by cation exchangers at low pH but not at all at high pH.

Aqueous solutions are usually employed and the columns are packed with beads of **ion exchange resins**, porous materials containing bound ionic groups such as $-SO_3^-$, $-COO^-$, $-NH_3^+$, or quaternary nitrogen atoms. Synthetic resins based on a crosslinked polystyrene are usually employed for separation of small molecules. For larger molecules chemical derivatives of cellulose or of crosslinked dextrans (Sephadex), agarose, or polyacrylamide are more appropriate. Positively charged ions, such as amino acids in a low pH solution, are placed on a cation exchange resin such as Dowex 50, which contains dissociated sulfonic acid groups as well as counter ions such as $Na^+$, $K^+$, or $H^+$. The adsorbed amino acids are usually eluted with buffers of increasing pH containing sodium or lithium ions. The procedure, which was developed by Moore and Stein,[106–109] is widely used for automatic quantitative analysis of amino acid mixtures obtained by hydrolysis of a protein or peptide (Fig. 3-6).[110–113]

Ion exchange chromatography of proteins and peptides is often done with such ion exchange materials as carboxymethyl-Sephadex and phosphocellulose, which carry negatively charged side chains or diethylaminoethylcellulose (DEAE-cellulose), which carries positively charged amino groups.[114] These materials do not denature proteins or entrap them and have a large enough surface area to provide a reasonable absorptive capacity. The mobile phase is usually buffered. For anion exchangers the pH should be above ~4.4 to keep most carboxylate side chains on the proteins ionized. The pH may be increased to pH 7–10, where most histidine imidazolium ions have dissociated, increasing the mobility of many proteins. For cation exchange the pH is usually buffered below pH 6 or 7 (Fig. 3-6).[115,116]

## 6. Affinity Chromatography

In this technique the chromatographic absorbent is designed to make use of specific biochemical inter-

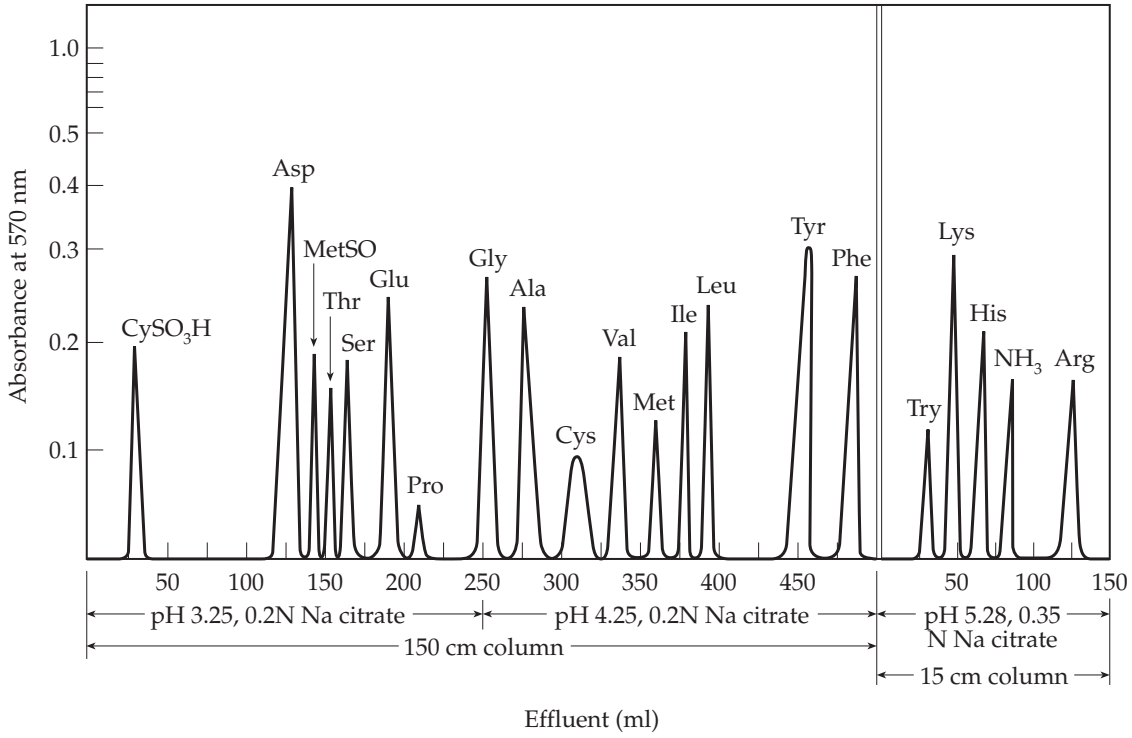
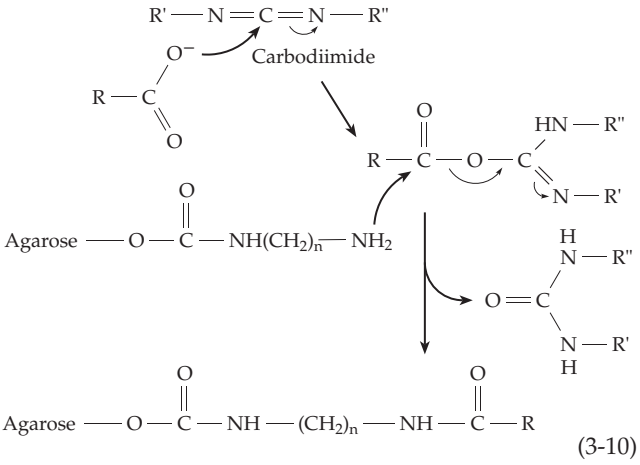



**Figure 3-6** Separation of amino acids by cation-exchange chromatography on a sulfonated polystyrene resin in the $Na^+$ form by the method of Moore and Stein.[110] The amino acids were detected by reaction with ninhydrin (Box 3-C); areas under the peaks are proportional to the amounts. Two buffers of successively higher pH are used to elute the amino acids from one column, while a still higher pH buffer is used to separate basic amino acids on a shorter column. From Robyt and White.[13]

actions to "hook" selectively a particular macromolecule or group of macromolecules.[117–122] Affinity chromatography is used in many ways, including the purification of enzymes, antibodies, and other proteins that bind tightly to specific small molecules.

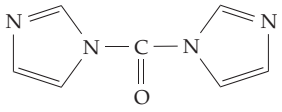
Because of their open gel structure (Fig. 4-10) agarose derivatives in bead form provide a good solid support matrix. The hydroxyl groups of the agarose are often linked to amino compounds. In one widely used procedure[119] the agarose is treated with cyanogen bromide (Br – C ≡N) in base to "activate" the carbohydrate (Eq. 3-9, step *a*). Then the amino compound is added (step *b*). The isourea product shown is the major one with agarose gels but dextran-based matrices tend to react by steps *c*–*e*.

Absorbents containing a large variety of R groups of shapes specifically designed to bind to the desired proteins can be made in this way. If the coupling is done with a diamine [R = $(CH_2)_n$—$NH_2$], the resulting ω-aminoalkyl agarose can be coupled with other compounds by reaction with **carbodiimide** (Eq. 3-10). For reaction in nonaqueous medium dicyclohexylcarbodiimide (R" = cyclohexyl) is often used, but for linking groups to agarose a water-soluble reagent such as 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide is recommended.[120] Carbodiimides are widely used for forming amide or phosphodiester linkages in the laboratory. The formation of an amide with a side chain of ω-aminoalkyl agarose can be pictured as in Eq. 3-10.

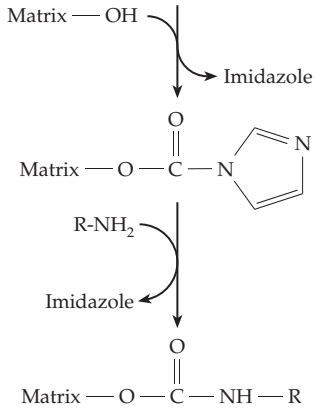



(3-10)

Many other means of preparation of adsorbents for affinity chromatography are also available.[121,122] For example, 1,1'-carbonyl-diimidazole can be used to couple a diamine to the matrix (Fig. 3-11). This reagent has the advantage that it does not depend upon the relatively unstable isourea linkages formed by Eq. 3-9 to hold the specific affinity ligands.[121]

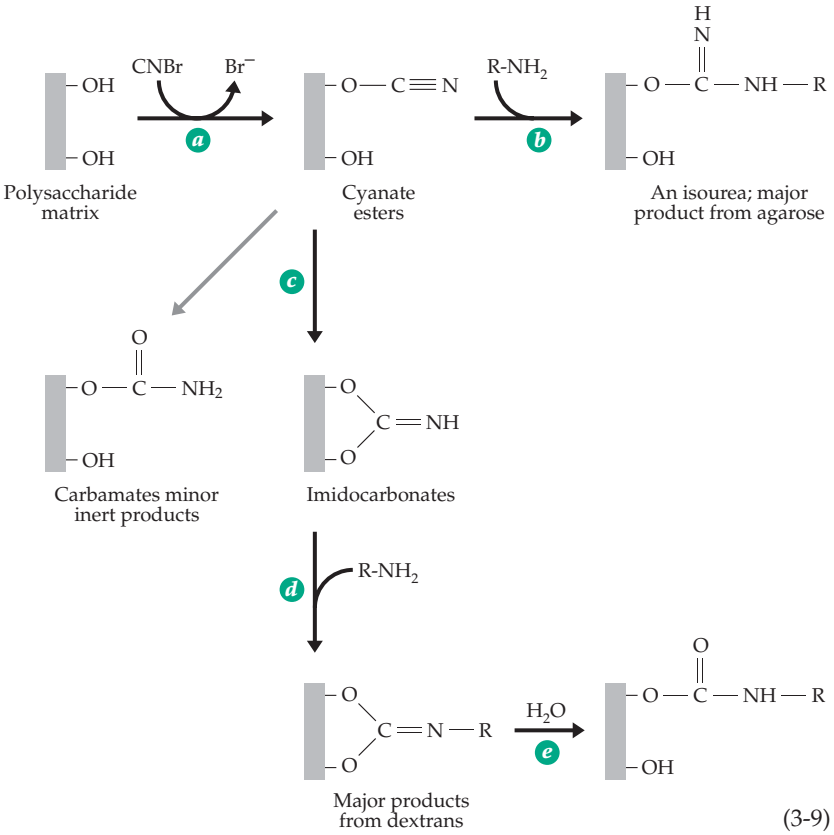


1,1'-carbonyl-diimidazole

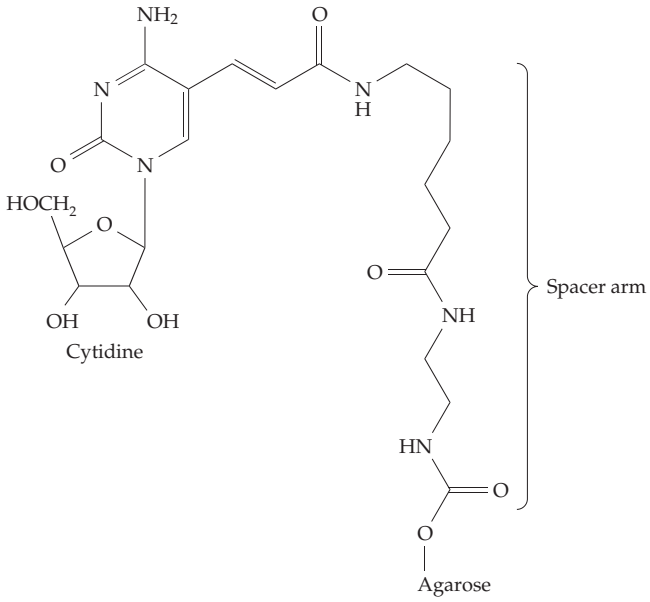



(3-11)

An example of a successful application of affinity chromatography is the isolation of the enzyme cytidine deaminase from cells of *E. coli*. Cytidine was linked covalently via long **spacer** arms to the agarose beads as in the following diagram:



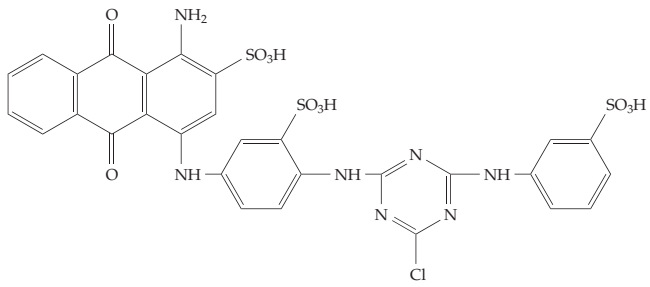

(3-9)




A cell extract was subjected to ammonium sulfate fractionation and the dialyzed protein was then poured through the affinity column which held the cytidine deaminase molecules because of their affinity for the cytidine structures that were bound to the agarose. The protein was eluted with a borate buffer; the borate formed complexes with the adjacent hydroxyl groups of the cytidine and thereby released the protein. After passage of the protein through an additional column of DEAE-Sephadex the deaminase had been purified 1700-fold compared to the crude extract.[123]

Another technique is to engineer genes to place a polyhistidine "tag" at the C terminus of a protein chain. Commercial cloning vehicles and kits are available for this purpose.[124] The protein produced when the engineered gene is expressed can be captured by the affinity of the polyhistidine tag for $Cu^{2+}$, $Ni^{2+}$, $Co^{2+}$, or $Zn^{2+}$ held in chelated form on an affinity column.[124–127]

A surprising discovery was that certain dyes, for example Cibachron blue, when covalently coupled to a suitable matrix, often bind quite specifically to proteins that have a nonpolar binding pocket near a positive charge.[128] This includes many enzymes that act on nucleotides.




Cibacron blue F3GA

## 7. Electrophoresis and Isoelectric Focusing

The methods considered in this section make use of movement of molecules in an electrical field. Separation depends directly upon differences in the net charge carried by molecules at a fixed pH. The net charge for compounds containing various combinations of acidic and basic groups can be estimated by considering the $pK_a$ of each group and the extent to which that group is dissociated at the selected pH using Eqs. 3-3 to 3-5. At some pH, the **isoelectric point** (pI), a molecule will carry no net charge and will be immobile in an electric field. At any other pH it will move toward the anode (+) or cathode (–). The pH at which the protein carries no net charge in the complete absence of added electrolytes is called the **isoionic point**.[129]

**Electrophoresis**, the process of separating molecules, and even intact cells[130] (Box 3-B), by migration in an electrical field, is conducted in many ways.[28,131–140] In **zone electrophoresis**, a tiny sample of protein solution, e.g., of blood serum, is placed in a thin line on a piece of paper or cellulose acetate. The sheet is moistened with a buffer and electrical current is passed through it. An applied voltage of a few hundred volts across a 20-cm strip suffices to separate serum proteins in an hour. To hasten the process and to prevent diffusion of low-molecular-weight materials, a higher voltage may be used. Two to three thousand volts may be applied to a sample cooled by water-chilled plates. Large-scale electrophoretic separations may be conducted in beds of starch or of other gels.

One of the most popular and sensitive methods for separation of proteins is electrophoresis in a column filled with **polyacrylamide** or **agarose gel** or on a thin layer of gel on a plate. The method depends upon both electrical charge and molecular size and has been referred to as **electrophoretic molecular sieving**.[28,133,135–137,141–143] Polyacrylamide gel electrophoresis is often carried out in the presence of ~1% of the denaturing detergent sodium docecyl sulfate which coats the polypeptide chain rather evenly. This method, which is often referred to as **SDS–PAGE,** has the advantage of breaking up complex proteins composed of more than one subunit and sorting the resultant monomeric polypeptide chains according to molecular mass (see Box 3-C). A disulfide-reducing reagent (see Eq. 3-23) such as ~1% 2-mercaptoethanol is usually present but may be omitted to permit detection of crosslinked peptides.

**Capillary electrophoresis** is increasingly popular[144–149a] and can be used to separate attomole amounts.[150] It can be used not only for separation of proteins but also for rapid estimation of the net charge on a protein.[151] The separation is conducted in tubes with internal diameters as small as 10–15 μm and as short as 1 cm. Multiple channels cut into a glass chip

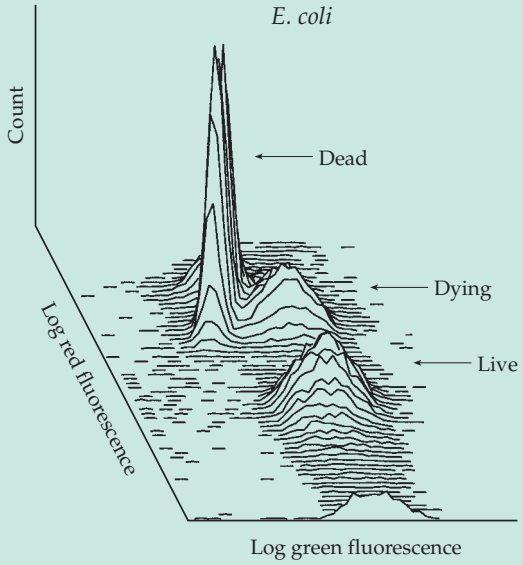
## BOX 3-B SORTING AND ANALYZING SINGLE CELLS

It is often important to examine and analyze individual cells.[a] For example, large numbers of single blood cells can be tested for the presence of specific antigenic determinants that arise by mutation. This permits assessment of the frequency of these mutations.[b] The complex chemical processing of neuropeptides can be studied on the contents of a single neuron (Chapter 30) using mass spectrometry.[c]

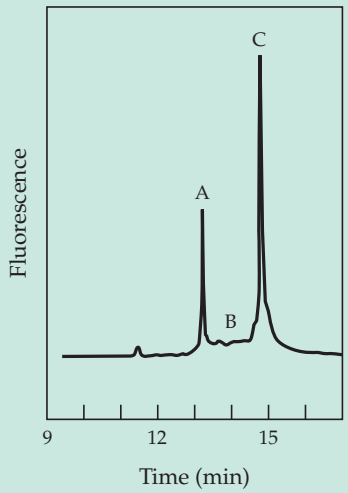
Several methods for separating cells have been devised. These include electrophoresis[d] or use of magnetic microspheres.[b] Micromanipulation can sometimes be used to select single cells for analysis. The most impressive technique is **flow cytometry**,[e,f] which is used daily on human blood samples in clinical laboratories. A suspension of cells is passed at a high rate of flow through a narrow capillary of ~0.2 mM diameter. The sample stream, which is surrounded by a larger "sheath" stream, has a smaller diameter of ~20 μm. One or more laser beams are used to record information about each cell over a period of a few microseconds as the cells pass by at a rate of as much as $10^5$ cells / s.

Flow cytometers developed from simpler cell counters, but now they are used to record cell size (from light scattering), optical absorbance, fluorescence, and phosphorescence. The optical properties are often enhanced by staining. The use of two dyes that fluoresce at different wavelengths permits the construction of two-dimensional plots as in the accompanying figure.

Capillary electrophoresis is one of the techniques able to separate constituents of single cells and is illustrated in the second figure.




Flow cytometric histogram of fluorescently labeled live and dead *E. coli* bacteria. The dye kit (BacLightTM) that was used stains membrane-compromised dead bacteria with a red fluorescing dye and live bacteria with a green fluorescing dye. Cells were analyzed on an EPICS XL cytometer (Coulter Corporation). By integrating the area under each population it was possible to discern the percentage of dead (54), dying (16), and live (30) bacteria within a mixed population. The use of fluorescent dyes such as these has proven useful for studying various mechanisms employed by the food industry for killing microorganisms in food products and for studying a variety of bacteria derived from seawater, soil, plant materials, and laboratory-grown cultures. Courtesy of Kristi Harkins.




Electropherogram of major proteins from a single erythrocyte. Peaks A, B, and C are carbonic anhydrase (~7 amol), methemoglobin (~ 5 amol), and hemoglobin $A_0$ (~ 450 amol), as identified from migration times relative to standards.[a] Courtesy of Edward S. Yeung.


[a] Yeung, E. S. (1994) *Acc. Chem. Res.* **27**, 409–414

[b] Jovin, T. M., and Arndt-Jovin, D. J. (1980) *Trends Biochem. Sci.* **5**, 214–219

[c] Li, K. W., Hoek, R. M., Smith, F., Jiménez, C. R., van der Schors, R. C., van Veelen, P. A., Chen, S., van der Greef, J., Parish, D. C., Benjamin, P. R., and Geraerts, W. P. M. (1994) *J. Biol. Chem.* **269**, 30288–30292

[d] Bauer, J. (1994) *Cell Electrophoresis*, CRC Press, Boca Raton, Florida

[e] Shapiro, H. M. (1995) *Practical Flow Cytometry*, 3rd ed., Wiley-Liss, New York

[f] Darzynkiewicz, Z., Robinson, J. P., and Crissman, H. A., eds. (1994) *Flow Cytometry*, 2nd ed., Academic Press, San Diego

can be used.[152]

Whereas in conventional zone electrophoresis most of the electrical current is carried by the buffer, in **isotachophoresis**[153,154] the ions being separated carry most of the current. In **isoelectric focusing**,[28,155–157] a pH gradient is developed electrochemically in a vertical column or on a thin horizontal plate between an anode and a cathode. The pH gradient in a column is stabilized by the presence of a density gradient, often formed with sucrose, and the apparatus is maintained at a very constant temperature. Proteins within the column migrate in one direction or the other until they reach the pH of the isoelectric point where they carry no net charge and are "focused" into a narrow band. As little as 0.01 pH unit may separate two adjacent protein bands which are located at positions corresponding to their isoelectric points. A newer development is the use of very narrow pH gradients that are immobilized on a polyacrylamide matrix.[158–160] With this technique some hemoglobin mutants differing only in substitution of one neutral amino acid for another have been separated.[161] Special techniques are needed for highly basic proteins.[162]

A two-dimensional method in which proteins are separated by isoelectric focusing (preferably with an immobilized pH gradient) in the first dimension and by SDS-gel electrophoresis in the second has become a popular and spectacularly successful method for studying complex mixtures of proteins (Box 3-C).[163–166] Over 2000 proteins can be separated on a single plate. A similar procedure but without SDS can be used to examine undenatured proteins.[167–169] Computer-assisted methods are being developed to catalog the thousands of proteins being identified in this way[170–172] and also to allow rapid identification of spots by mass spectrometry. The technique can be applied to intact proteins in subpicomole quantities, even in whole cell lysates,[150,173,173a] or an enzyme such as trypsin can be used to cut the proteins into pieces on the gel plate and the mixtures of peptides can be analyzed by mass spectrometry.[174–176] Capillary electrophoresis or capillary isoelectric focusing can be applied before samples are sent to the mass spectrometer.

## C. Determining the Relative Molecular Mass, $M_r$

The evaluation of $M_r$ is often of critical importance. Minimum values of $M_r$ can often be computed from the content of a minor constituent, e.g., the tryptophan of a protein or the iron of hemoglobin. However, physicochemical techniques provide the basis for most measurements.[177] Observations of osmotic pressure or light scattering can also be used and provide determinations of $M_r$ that are simple in principle, but which have pitfalls.[178]

### 1. Ultracentrifugation

Some of the most reliable methods for determining $M_r$ depend upon **analytical ultracentrifuges**. These instruments, capable of generating a centrifugal field as much as $4 \times 10^5$ times that of gravity, were developed in the 1920s and 1930s by T. Svedburg and associates in Uppsala, Sweden.[179–181] Driven by oil turbines, the instruments were expensive and difficult to use, but by 1948 a reliable electrically driven machine, the Beckman Model E ultracentrifuge, came into widespread use. It has had a major impact on our understanding of proteins, on methods of purification of proteins, and on our understanding of interactions of protein molecules with each other and with small molecules.[182,183] Nevertheless, it was not until 1990 that a truly "user-friendly" analytical untracentrifuge became available.[180,184–186] The Beckman Model XL-A centrifuge has a very small rotor driven by an air-cooled induction motor and is computer controlled. Data are recorded automatically in digital form and computer programs are available to carry out the necessary computations. The instrument can record ultraviolet-visible spectra at multiple radial positions in the sample cell (Fig. 3-7).

A straightforward determination of $M_r$ is obtained by centrifuging until an equilibrium distribution of the molecules of a protein or other macromolecular material is obtained and by recording the variation in concentration from the center to the periphery of the centrifuge cell[177,183,185,187–190] (see also Section A,2). Using short cells, this **sedimentation equilibrium** can be attained in 1–5 hours instead of the 1–2 days needed with older instruments. For a single component system the concentration distribution at equilibrium is given by Eq. 3-12.

$$c(r) = c(a) \exp\left[M\,(1-\bar{v}\rho)\,\omega^2\,(r^2-a^2)\,/\,2RT\right] \quad (3\text{-}12)$$

Here $c(r)$ is the concentration c at the radial position $r$ (measured from the centrifuge axis), $a$ is the radial distance of the meniscus, $M$ is the molecular mass in daltons, and $\bar{v}$ is the partial specific volume in ml/gram. For most proteins $\bar{v}$ varies from 0.69–0.75. It is the reciprocal of the density of the particle. Rho (ρ) is the density (g/ml) of the solvent. A plot of log $c(r)$ against $r^2$ is a straight line of slope $M\,(1-\bar{v}\rho)\,/\,2RT$. The computer can also accommodate mixtures of proteins of differing molecular masses, interacting mixtures, etc.[185,191]

***Sedimentation velocity.*** The relative molecular mass $M_r$ can also be measured from observation of the velocity of movement of the boundary (or boundaries for multicomponent systems) between solution and solvent from which the macromolecules have sedi-

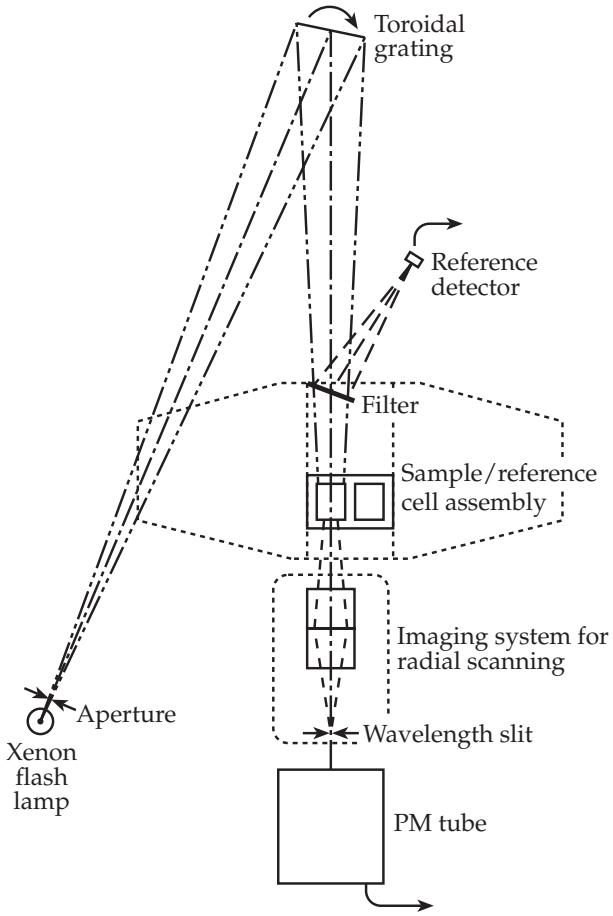



**Figure 3-7** The scanning absorption optical system of the Beckman Optima™ XL-A ultracentrifuge. Courtesy of Beckman Coulter.

mented. This boundary, which can be visualized by optical methods, is quite sharp initially, but it broadens with time, because of diffusion, as the macromolecules sediment.

A molecule in a centrifuge is acted upon not only by the applied centrifugal force but also by an opposing **buoyant force** that depends upon the difference in density of the sedimenting particles and the solvent and by a frictional drag, which is proportional to a **frictional coefficient** $f$. Setting the sum of these forces to zero for the hydrodynamic steady state yields Eq. 3-13, which defines the **sedimentation constant** $s$.

$$s = \frac{v}{\omega^2 r} = \frac{M(1-\bar{v}\rho)}{N f} \tag{3-13}$$

Here $f$ is the frictional coefficient which is difficult to predict or to measure but is often assumed to be the same as the frictional coefficient that affects **diffusion**. It can be obtained from the diffusion coefficient $D$ using Eq. 3-14.

$$D(\mathrm{cm^2\ s^{-1}}) = k_B T/f = RT/N f \tag{3-14}$$

By combining Eqs. 3-13 and 3-14 we obtain the Svedberg equation:

$$M = \frac{RTs}{D(1-\bar{v}\rho)} \tag{3-15}$$

Here $R$ is in the cgs (cm-gram-second) unit of $8.31 \times 10^7$ erg $\mathrm{mol^{-1}\ deg^{-1}}$. Using this equation the relative molecular mass $M_r$, which is numerically the same as $M$, can be evaluated from the sedimentation constant $s$. Since $s$, $D$, and $\bar{v}$ must all be measured with care, the method is demanding. It is often necessary to measure $s$ and $D$ at several concentrations and to extrapolate to infinite dilution. It is also customary to correct the data to give the values $s^\circ_{20,w}$ and $D^\circ_{20,w}$ expected at 20°C in pure water at infinite dilution.

To the extent that we can regard protein molecules as spherical we can substitute for $f$ in Eq. 3-13 the frictional coefficient of a sphere:

$$f_{\mathrm{sphere}} = 6\pi\eta r_h \tag{3-16}$$

Here $r_h$ is the hydrated radius or **Stokes radius** of the protein. On this assumption $s$ will be expected to increase with the relative molecular mass approximately as $M_r^{2/3}$. A plot of log $s$ against log $M_r$ should be a straight line. Figure 3-8 shows such a plot for a number of proteins. The plots for nucleic acids, which can often be approximated as rods rather than spheres, fall on a different line from those of proteins. Furthermore, the sedimentation constant falls off more rapidly with increasing molecular mass than it should for spheres.

From analysis of a variety of well-characterized proteins, Squire and Himmel[192] observed that if proteins are assumed to contain 0.53 g $H_2O$ per gram of protein and to have a mean value for $\bar{v}$ of 0.730 g / $\mathrm{cm^3}$ the value of $M_r$ can be predicted by Eq. 3-17 with the standard deviation indicated. Here, S is the sedimentation constant in Svedberg units ($10^{-13}$s).

$$M_r = 6850\ S^{3/2}\ )\ 0.090\ M_r \tag{3-17}$$

For proteins with various values of $\bar{v}$, Eq. 3-18 applies.

$$M_r = 922\ [S / (1-\bar{v}\rho)]^{3/2}\ ) \pm 0.066\ M_r \tag{3-18}$$

## BOX 3-C ISOTOPES IN BIOCHEMICAL INVESTIGATIONS

Both stable[a] and radioactive[b–e] isotopes are widely used in chemical and biological investigations. The study of metabolism was revolutionized by the introduction of isotopic tracers. In one of the first biological experiments with the stable isotope $^{15}N$ (detected by mass spectrometry), Schoenheimer and associates in 1937 established the previously unsuspected turnover of protein in living tissues (Chapter 24, Section B). In 1937 Ruben *et al.* reported the uptake of radioactive $^{11}CO_2$ by plants.[f] A few years later Calvin and associates first traced the pathway of carbon in photosynthesis using the much longer lived $^{14}CO_2$ (Box 17-F). Wood and Werkman, in 1941, employed the stable isotope $^{13}C$ in studies of bacterial and mammalian metabolism (Box 17-C). The radioactive $^{32}P$ and $^{35}S$ have served to elucidate the metabolism of phosphorus and sulfur. Titrium ($^{3}H$) has been used to label many organic substances including thymine, which has been used extensively in the study of nucleic acid metabolism. Radioactive isotopes provide the basis for sensitive analytical procedures such as **radioimmunoassays** of minute quantities of hormones (Box 31-D). Through **radioautography** these isotopes facilitate numerous analytical procedures (see accompanying photo) and have provided the basis for important end-group methods used in sequence determination of polynucleotides (Eq. 5-24).

Several isotopes used in biochemistry are listed in the following table. For each radioactive isotope, the half-life is given, as is the type of particle emitted, and the energy of the particle. Gamma rays, such as those given off in decay of $^{125}I$ or $^{131}I$, are very penetrating and easy to count precisely, as is the energetic β radiation from $^{32}P$. On the other hand, $^{3}H$ (tritium) is relatively difficult to detect[g] but its weak β particle, which can travel only a short distance through a sample, makes it uniquely suitable for radioautography on a microscopic scale. Positrons ($\beta^+$) travel some distance, e.g., up to a few millimeters in the case of $^{13}N$. They are then destroyed by reacting with an electron to produce a pair of γ rays of energy 0.511 MeV, equal to the sum of the rest masses of an electron plus a positron. The half-life (Eq. 9-4) determines the isotopic abundance needed to achieve a given radiation rate, a practical matter in providing a sufficient rate of decay to permit counting with an acceptably low statistical error. Even very short-lived isotopes such as $^{13}N$, have proved useful as tracers.[h] The amount of an isotope giving $3.7 \times 10^{10}$ disintegrations per second (this is 1 g of pure radium, 0.3 mg of $^{3}H$, or 0.22 g of $^{14}C$) is known as the **curie** (Ci). One millicurie (mCi) provides $2.22 \times 10^{9}$ disintegrations / min (dpm). Radiolabeled substances ordinarily contain only a small fraction of the unstable isotope together with a larger number of unlabeled molecules. Compounds are usually sold in millicurie or microcurie quantities and with a stated specific activity as mCi $mmol^{-1}$. For example, a compound labeled at a single position with $^{3}H$ and having a specific activity of 50 mCi $mmol^{-1}$ would contain about 0.17% $^{3}H$ at that position.

| Isotope | Half-life | Maximum energy of radiation (MeV) β | γ |
|---|---|---|---|
| $^{2}H$ (deuterium) | Stable | | |
| $^{3}H$ (tritium) | 12.26 years | 0.018 | |
| $^{11}C$ | 20.4 min | | ($\beta^+$) |
| | 0.511 | | |
| $^{13}C$ | Stable | | |
| $^{14}C$ | 5730 years | 0.156 | |
| $^{13}N$ | 9.96 min | 1.2 | ($\beta^+$) |
| | 0.511 | | |
| $^{15}N$ | Stable | | |
| $^{15}O$ | 20.4 min | | ($\beta^+$) |
| | 0.511 | | |
| $^{18}O$ | Stable | | |
| $^{18}F$ | 110 min | | ($\beta^+$) |
| | 0.511 | | |
| $^{22}Na$ | 2.6 years | 0.55 | |
| | 1.28 | | |
| $^{32}P$ | 14.3 days | 1.71 | |
| $^{35}S$ | 87.2 days | 0.167 | |
| $^{36}Cl$ | $3 \times 10^{5}$ years | 0.716 | |
| $^{40}K$ | $1.3 \times 10^{9}$ years | 1.4 | |
| | 1.5 | | |
| $^{45}Ca$ | 165 days | 0.26 | |
| $^{59}Fe$ | 45 days | 0.46 | |
| | 1.1 | | |
| $^{65}Zn$ | 250 days | 0.32 | |
| | 1.14 | | |
| $^{90}Sr$ | 29 years | 0.54 | |
| $^{125}I$ | 60 days | 0.036 | |
| $^{131}I$ | 8.06 days | 0.61 | |
| | 0.36 | | |

Because of the development of new NMR techniques and improvements in mass spectrometry stable isotopes, such as $^{2}H$, $^{13}C$, $^{19}F$, and $^{31}P$, are being used more frequently to study metabolism.[j] Carbon-13 containing compounds can fulfill many

## BOX 3-C ISOTOPES IN BIOCHEMICAL INVESTIGATIONS (continued)


Radioautogram showing the separation of proteins of *E. coli* labeled with $^{14}C$ amino acids. From O'Farrell.[i] Twenty-five µl of sample containing 180,000 cpm and ~10 µg of protein were subjected to isoelectric focusing in a 2.5 x 130-mm tube containing polyacrylamide gel to separate proteins according to isoelectric point. The gel was then extruded from the column and was placed on one edge of a slab of polyacrylamide gel. Then SDS electrophoresis in the second dimension separated the proteins according to size. Over 1000 spots could be seen in the original radioautogram, which was obtained by placing a piece of photographic film over the gel slab and exposing it to the radiation for 875 hours. For details see O'Farrell.[i]

of the same tracer functions as $^{14}C$. Even the radioactive $^{3}H$ nucleus can be utilized for *in vivo* NMR.[k,l] Although radioisotope labeling is very sensitive, it gives little information unless compounds are isolated and laboriously degraded to determine the positions of the labels. NMR spectroscopy is less sensitive but can give direct chemical information about the positions of $^{13}C$ in compounds within living cells. A compound containing only $^{13}C$ in one or in many positions can be safely administered to human individuals as well as to other organisms and spectra of products that arise can be observed. High-resolution deuterium NMR spectroscopy has been used to follow $^{2}H$ incorporated at C-1, C-2, or C-6 positions in glucose.[m]

As a result of metabolic reactions an isotope may appear at more than one position in a product, yielding two or more isotope isomers or **isotopomers**. These are seen individually by NMR spectroscopy and the concentration and isotope labeling patterns of the labeled compounds can be followed over a period of time. The use of this **isotopomer analysis** in studies of the citric acid cycle is illustrated in Box 17-C and its use in studies of glucose metabolism is considered in Chapter 17, Section L.

A change in isotopic mass, especially from $^{1}H$ to $^{2}H$ or $^{3}H$, often produces a strong effect on reaction rates and the study of **kinetic isotope effects** has provided many insights into the mechanisms of enzymatically catalyzed reactions. Isotopes have permitted a detailed understanding of the stereochemistry of enzymatic reactions, an impressive example being the synthesis and use of chiral acetate (Chapter 13)[n] and chiral phosphate groups (Chapter 12). Specific isotopic properties provide the basis for NMR (Section G).


[a] Matwiyoff, N. A., and Ott, D. G. (1973) *Science* **181**, 1125–1132
[b] Wang, C. H., Willis, D. L., and Loveland, W. D. (1975) *Radiotracer Methodology in the Biological, Environmental, and Physical Sciences*, Prentice-Hall, Englewood Cliffs, New Jersey
[c] Wang, Y., ed. (1969) *Handbook of Radioactive Nuclides*, CRC Press, Cleveland, Ohio
[d] Thornburn, C. C. (1972) *Isotopes and Radiation in Biology*, Butterworth, London
[e] Slater, R. J., ed. (1990) *Radioisotopes in Biology: A Practical Approach*, IRL Press, Oxford
[f] Ruben, S., Hassid, W. Z., and Kamen, M. D. (1939) *J. Am. Chem. Soc.* **61**, 661–663
[g] Bransome, J., ed. (1970) *Liquid Scintillation Counting*, Grune & Stratton, New York
[h] Cooper, A. J. L. (1985) *Adv. Enzymol.* **57**, 251–356
[i] O'Farrell, P. H. (1975) *J. Biol. Chem.* **250**, 4007–4021
[j] Wolfe, R. R. (1992) *Radioactive and Stable Isotope Tracers in Biomedicine*, Wiley, New York
[k] Newmark, R. D., Un, S., Williams, P. G., Carson, P. J., Morimoto, H., and Klein, M. P. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 583–587
[l] Bergerat, A., Guschlbauer, W., and Fazakerley, G. V. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 6394–6397
[m] Aguayo, J. B., Gamcsik, M. P., and Dick, J. D. (1988) *J. Biol. Chem.* **263**, 19552–19557
[n] Cornforth, J. W. (1976) *Science* **193**, 121–125

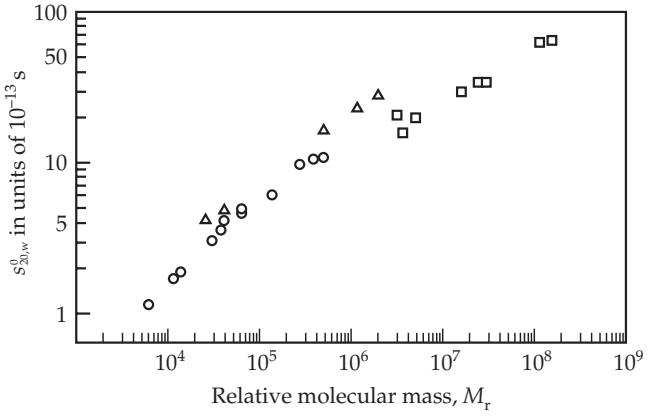



**Figure 3-8** Plots of the logarithm of the sedimentation constant $s$ against the logarithm of the molecular weight for a series of proteins and nucleic acids: (○) globular proteins, (△) RNA, and (□) DNA. Proteins include a lipase (milk), cytochrome *c*, ribonuclease (pancreatic), lysozyme (egg white), follicle-stimulating hormone, bacterial proteases, human hemoglobin, prothrombin (bovine), malate dehydrogenase, γ-globulin (horse), tryptophanase (*E. coli*), glutamate dehydrogenase (chicken), and cytochrome *a*. Double-stranded DNA molecules are those of bacteriophage ϕX174 (replicative form), $T_7$, $\lambda_{b2}$, $T_2$, and $T_4$, and that of a papilloma virus. The RNA molecules are tRNA, rRNA, and mRNA of *E. coli*, and that of turnip yellow mosaic virus.[190a,191a]

## 2. Gel Filtration and Gel Electrophoresis

Several newer methods of molecular mass determination were developed in the 1960s–1980s. One is gel filtration. A column of gel beads such as Sephadex is prepared carefully and is calibrated by passing a series of protein solutions through it. The volume $V_e$ at which a protein peak emerges from the column can be expressed as the sum of two terms (Eq. 3-19) in

$$V_e = V_o + \sigma V_i \qquad (3\text{-}19)$$

which $V_o$ is the **void volume**, i.e., the elution volume that is observed for very large particles that are completely excluded from the gel, and $V_i$ is the internal volume within the beads of gel. The value of σ is inversely related to the diffusion constant $D$, which for a spherical particle $D$ is related by Eq. 3-20 to the Stokes radius $r_h$. This equation comes directly from Eqs. 3-16 and 3-14.

$$r_h = \frac{k_B T}{6\pi\eta D} \qquad (3\text{-}20)$$

This suggests a proportionality between σ and the molecular radius. In fact, Eq. 3-21, in which a and b are constants provides a fairly good approximation for σ and $V_e$ is correlated approximately with log $M_r$ as shown in Fig. 3-9.[193,194]

$$\sigma = a \log r_h + b \qquad (3\text{-}21)$$

A series of reference proteins of known molecular masses are used to calibrate the column and $M_r$ for an unknown protein is estimated from its position on the graph.[195,196] Another modification of the method depends upon chromatography in a high concentration of the denaturing salt guanidinium chloride. The assumption is made that proteins are denatured into random coil conformations in this solvent.[196]

Probably the most widely used method for determining the molecular mass of protein subunits is gel electrophoresis in the presence of the denaturing detergent sodium dodecyl sulfate (SDS). The protein molecules are not only denatured but also all appear to become more or less evenly coated with detergent.[197] The resulting rodlike molecules usually show a uniform dependence of electrophoretic mobility on molecular mass (plotted as log $M_r$). An example is shown in Fig. 3-10. Again, the molecular mass of the protein under investigation is estimated by comparison of its rate of migration with that of a series of marker proteins.[195,198]

## 3. Mass Spectrometry

Mass spectrometry has played a role in biochemistry since the early 1940s when it was introduced for use in following isotopic labels during metabolism.[199–200c] However, it was not until the 1990s that suitable commercial instruments were developed to permit mass spectrometry using two new methods of ionization. The techniques are called **matrix-assisted laser desorption / ionization time-of-flight (MALDI-TOF)** and **electrospray ionization (ESI)** mass spectrometry.

In the MALDI technique a pulsed laser beam strikes a solid sample and heats, vaporizes, and ionizes compounds with little decomposition.[201–209] Proteins or other biopolymers are mixed with a "matrix" that absorbs the heat of the laser beam. The protein sample together with the matrix is dried. Most proteins form crystals and the laser beam is directed toward individual protein crystals or aggregates. Various materials are used for the matrix. Compounds as simple as glycerol, succinic acid, or urea can be used with an infrared laser. For proteins an ultraviolet nitrogen laser tuned to 337 nm is usually employed with an ultraviolet light-absorbing matrix such as hydroxybenzoic acid, 2,5-dihydroxybenzoic acid, α-hydroxy-

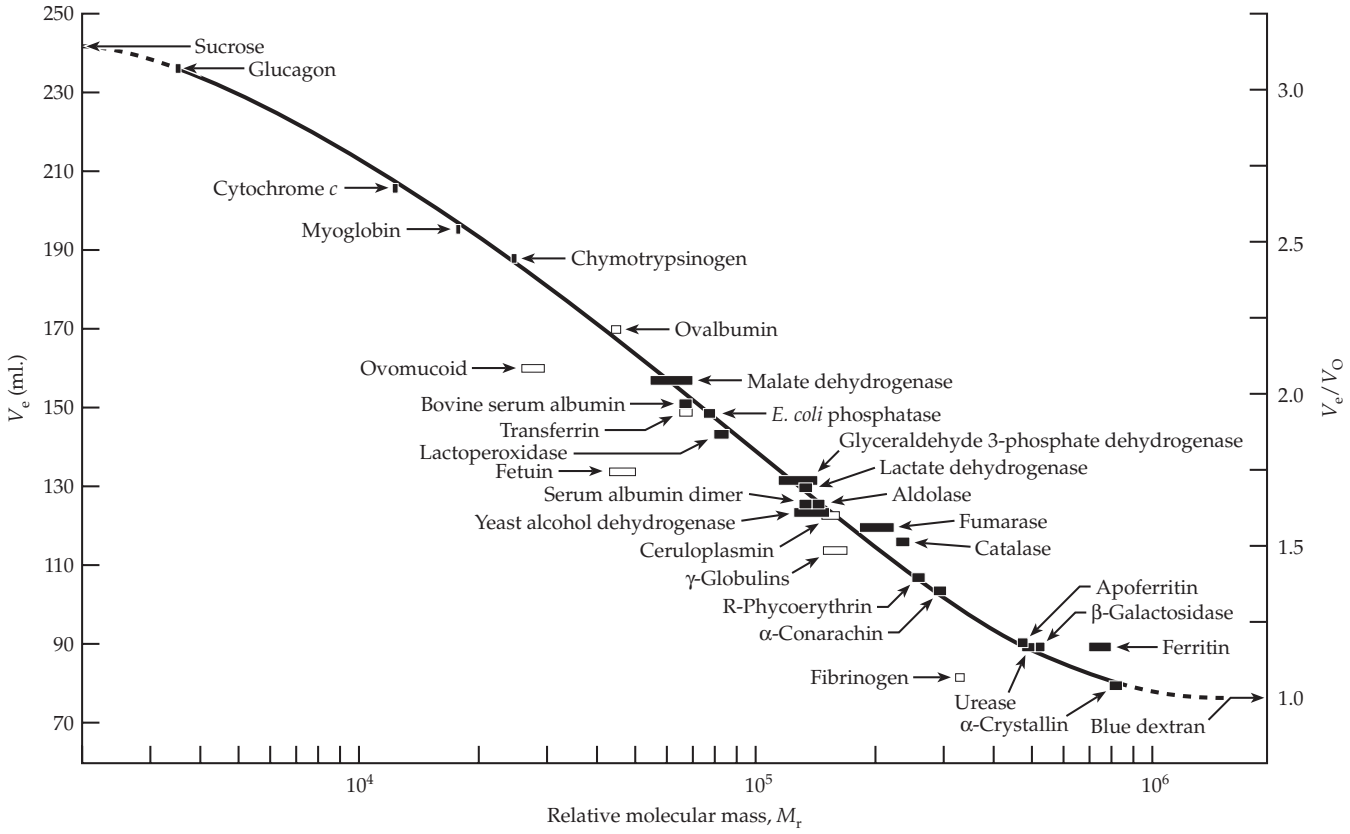



**Figure 3-9** Elution volume of various proteins on a column of Sephadex G-200 as a function of molecular mass. The right-hand vertical axis shows the ratio of the elution volumes to that of blue dextran, a high-molecular-mass polysaccharide that is excluded from the internal volume. After Andrews.[193]

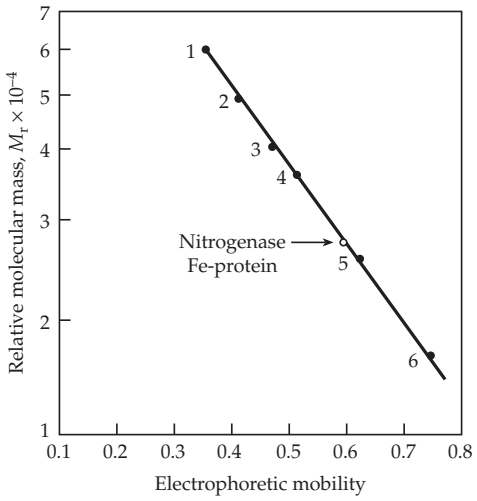



**Figure 3-10** Estimation of the molecular mass of the polypeptide chain of the nitrogenase Fe-protein using SDS-polyacrylamide electrophoresis; from a set of four standard curves. The marker proteins are (1) catalase, (2) fumarase, (3) aldolase, (4) glyceraldehyde-phosphate dehydrogenase, (5) α-chymotrypsinogen A, and (6) myoglobin. (o) indicates position of azoferredoxin. From Nakos and Mortenson.[195]

cinnamic acid, or sinapinic acid (Chapter 25). The matrix ionizes, desorbs from the surface, and transfers energy to the crystalline protein, causing it to ionize and desorb from the surface. Oligosaccharides and oligonucleotides can be ionized in a similar way.

MALDI spectra are relatively simple (Fig. 3-11), often containing a single major peak corresponding to the singly charged molecular ion $[M + H]^+$ of mass $m + 1$ and perhaps a doubly charged molecular ion $[M + 2H]^{2+}$ of mass $(m + 2)/2$. For oligomeric proteins the major peak is often that of the monomer with weaker peaks for oligomers. The instrument can also be adjusted to generate negative ions whose detection is useful for study of phosphorylated peptides, many oligosaccharides, and oligonucleotides. With a TOF spectrometer there is no upper limit to the mass range and masses of over 100 kDa can be measured to about ±0.1%. Femtomole quantities can be detected.

The MALDI method is especially useful for complex mixtures of peptides and can be utilized in peptide sequencing. The technique is also appropriate for studying mixtures of glycoproteins. Negative-ion MALDI can be applied to oligonucleotide mixtures. Further improvements in resolution in both MALDI

and ESI methods are anticipated as a result of developement of Fourier transform mass spectrometers.[202]

In ESI mass spectrometry[201,203 – 205,210 – 213] the sample, dissolved in an appropriate solvent (usually a 50:50 mixture of methanol and water for proteins), is infused directly into the ionization chamber of the spectrometer through a fused silica capillary. At the end of the capillary the solution is subjected to electrical stress created by a voltage difference of about 5 kV between the electrospray needle and the sampling orifice (the counter-electrode). The process results in the formation of singly and / or multiply charged molecular ions which are guided into the analyzer for mass analysis. For proteins every arginine, lysine, and histidine may bind a hydrogen ion to form a variety of positive ions. A 100-kDa protein may easily bind 100 protons bringing

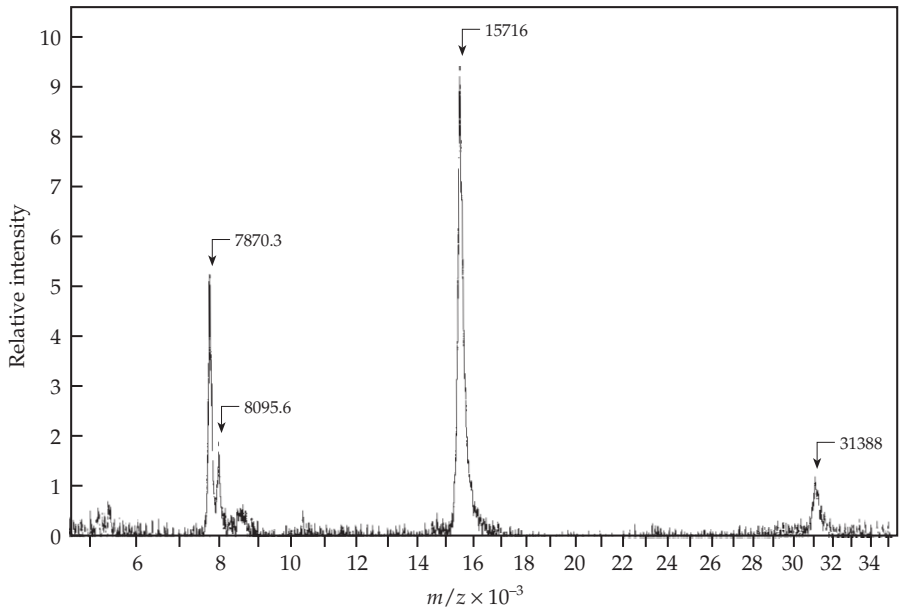



**Figure 3-11** Matrix-assisted laser desorption / ionization time-of-flight (MALDI-TOF) mass spectrum of bovine erythrocyte Cu-Zn superoxide dismutase averaged over ten shots with background smoothing. One-half µl of solution containing 10 pmol of the enzyme in 5 mM ammonium bicarbonate was mixed with 0.5 µl of 50 mM α-cyanohydroxycinnamic acid dissolved in 30% (v / v) of acetonitrile-0.1% (v / v) of trifluoroacetic acid. The mixture was dried at 37° C before analysis. The spectrum shows a dimer of molecular mass of 31,388 Da, singly charged and doubly charged molecular ions at 15,716, and 7870 Da, respectively. The unidentified ion at mass 8095.6 may represent an adduct of the matrix with the doubly charged molecular ion. Courtesy of Louisa Tabatabai.

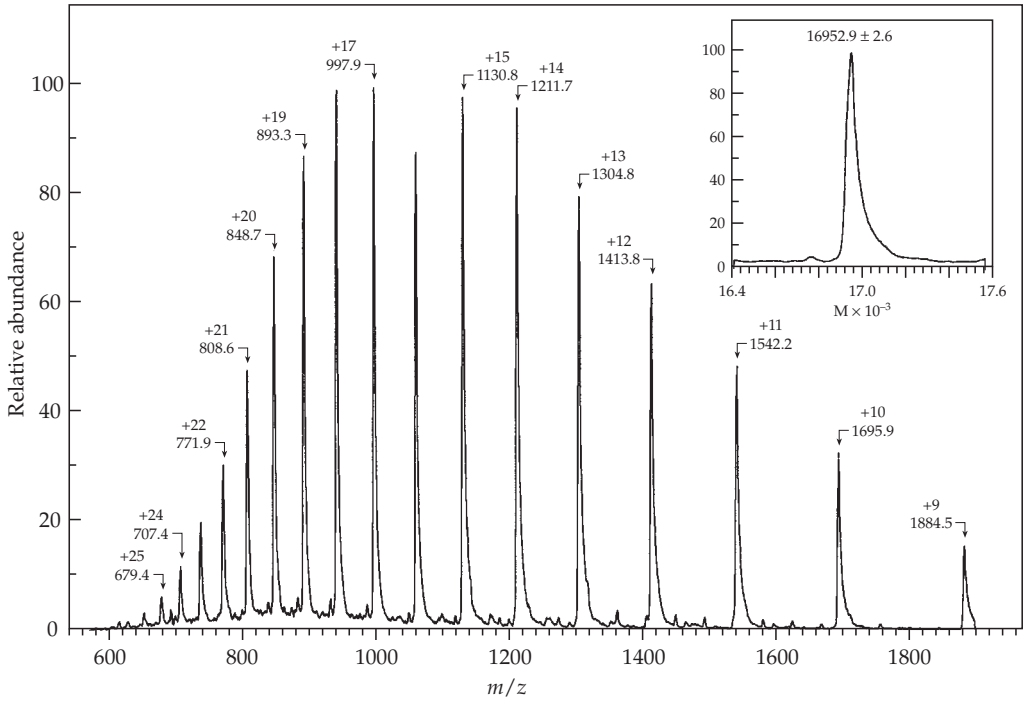



**Figure 3-12** Positive ion electrospray mass spectrum of horse apomyoglobin ($M_r$ 16,950.4). The net charge on each ion as well as the mass to charge ratio $m/z$ is indicated at the top of each peak. The inset shows a computer "deconvolution" of the spectrum with the calculated value of molecular mass. Courtesy of Kamel Harrata.

the $m/z$ ratio for the fully charged protein to 1000, well below the maximum $m/z$ ratio of ~2400 for a typical quadrupole mass spectrometer. Since not all basic groups are protonated the spectra consist of families of peaks of differing $m/z$ (Fig. 3-12). For a single pure protein the molecular mass can be calculated from the ratio of $m/z$ values from any pair of adjacent peaks. It is better, especially if there is a mixture of proteins, to use a computer to extract $M_r$.[214] The accuracy can be quite high, typically ±0.01%: one mass unit in 10,000.

Some complexity arises from the fact that each carbon atom in a protein contains about 1% of $^{13}C$. This means that for a protein of mass >10 kDa there will be a confusing array of peaks in the mass spectrum and it may be difficult to pick out the relatively minor "monoisotopic" peak that arises from molecules containing only $^{12}C$, $^{1}H$, $^{14}N$, $^{16}O$, and $^{32}S$. In fact, the peak representing the most abundant mass will be a few mass units higher than the monoisotopic peak.[215,216] (see Study Question 15). New computer programs have been devised to assist in the analysis. Use of **$^{13}C$ and $^{15}N$-depleted nutrients** also extends the applicability of mass spectrometry.[217]

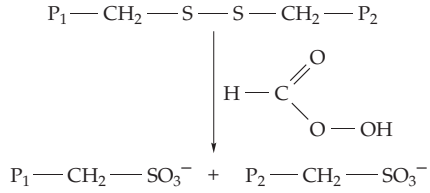
Electrospray mass spectrometry utilizes a soft ionization technique at nearly atmospheric pressure. As a result, intact molecular ions are formed in high yield. The instrument can be interfaced readily to HPLC or capillary electrophoresis columns and subfemtomole amounts of proteins can be detected. A disadvantage is that salt concentrations must be kept low (< mM) and that the protein tends to bind $Na^+$, $K^+$, and anions that may confuse interpretation of spectra.

## D. Determining Amino Acid Composition and Sequence

When they are first isolated, proteins are usually characterized by $M_r$, isoelectric point, and other easily measured properties. Among these is the amino acid composition[112] which can be determined by completely hydrolyzing the protein to the free amino acids. Later, it is important to establish the primary structure or amino acid sequence.[51] This has been accomplished traditionally by cutting the peptide chain into smaller pieces that can be characterized easily. However, most protein sequences are now deduced initially from the corresponding DNA sequences, but further chemical characterization is often needed.

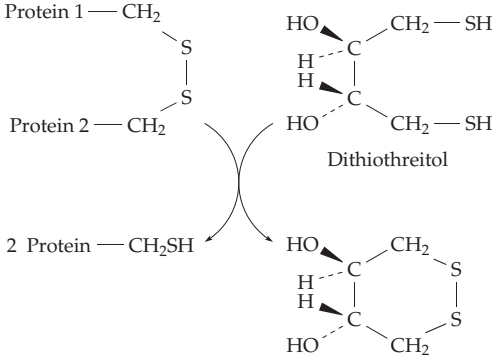
### 1. Cleavage of Disulfide Bridges

Before a polypeptide chain can be degraded it is usually necessary to break any disulfide bridges.[218–220] For some proteins such as the keratins of hair, these linkages must be broken even to get the protein into solution. Oxidation with performic acid (Eq. 3-22) has been used on ribonuclease but is not often employed because the performic acid also oxidizes tryptophan residues.



(3-22)

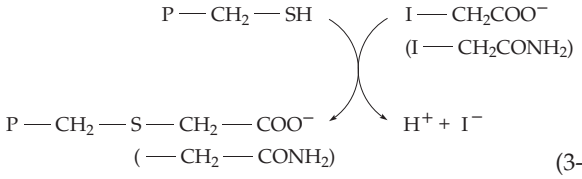
Reduction by **dithiothreitol** or **dithioerythritol** (Eq. 3-23) is usually successful.



(3-23)

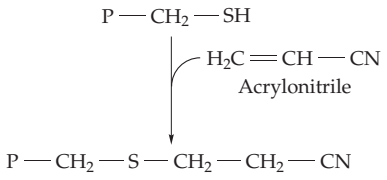
Upon oxidation these dithiols cyclize to form stable disulfides, driving the reaction to completion. These same compounds are also widely used to protect SH groups in enzymes against accidental oxidation by oxygen and to dissolve highly crosslinked insoluble proteins. Mercaptoethanol, $HS–CH_2–CH_2–OH$, may be used for the same purposes but requires higher concentrations and has a disagreeable odor.

Once cleaved, disulfide bridges may be prevented from reforming by conversion of the resulting thiol groups to stable derivatives, e.g., with iodoacetate and iodacetamide (Eq. 3-24).



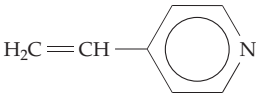
(3-24)

A better reagent is acrylonitrile (Eq. 3-25).




(3-25)

In modern sequencing methods vinylpyridine, which reacts in a similar way, is often used. It can be detected during amino acid analysis or sequencing after derivatization with phenylisothiocyanate (Eq. 3-30).



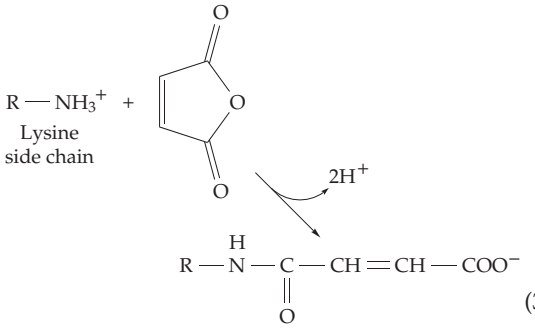

Vinylpyridine

## 2. Hydrolysis and Other Chain Cleavage Reactions

Most biopolymers are inherently unstable with respect to cleavage to monomer units by reaction with water. Hydrolysis can be catalyzed by protons, by hydroxyl ions, or by the protein-hydrolyzing enzymes that are discussed in Chapter 12. Complete hydrolysis of proteins is usually accomplished by heating under nitrogen with 6 M HCl at 150°C for 65 min. Some amino acids, especially tryptophan, are destroyed and the amides in the side chains of asparagine and glutamine are converted to the free acids. Some peptide linkages such as Val-Val are very resistant and tend to be incompletely hydrolyzed. No procedure has been found which gives the ideal complete hydrolysis. Use of 4 M methanesulfonic acid containing 3-(2-aminoethyl)indole instead of 6 M HCl gives less decomposition of tryptophan.[221] Base-catalyzed hydrolysis of proteins also gives good yields of tryptophan but causes extensive racemization of amino acids.[222]

Complete enzymatic digestion of proteins can be accomplished with a mixture of enzymes including proteases produced by fungi (Pronase). However, the enzymes attack each other, making quantitative analysis difficult. The problem can be circumvented by immobilizing the hydrolytic enzymes in a column of agarose gel. The protein to be hydrolyzed is passed through the gel and the constituent amino acids emerge from the bottom of the column.[223,224]

***Selective enzymatic hydrolysis.*** The traditional strategy in sequence determination is to cut protein chains into smaller pieces which can be separated by chromatography or electrophoresis and sequenced individually. Enzymatic cleavage is especially useful because of its specificity. **Trypsin**, a so-called **endopeptidase**, cleaves peptide chains at a rapid rate only if the carbonyl group of the amide linkage cleaved is contributed by one of the basic amino acids lysine, arginine, or aminoethylcysteine (see Fig. 12-10). If the protein is treated with maleic anhydride (Eq. 3-26),




(3-26)

### BOX 3-D THE PROTEOME

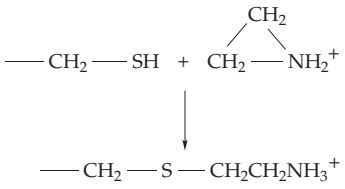
The ability to separate rapidly and detect minute amounts of proteins has spawned a new concept: The **proteome** is envisioned as a record of all proteins being actively synthesized by a cell – or of all genes being actively "expressed."[a–c] The concept evolved from efforts to automate the cataloging of spots on two-dimensional gels such as that shown in Box 3-B where each spot represents a single protein. The ability to unambiguously identify the spots by mass spectroscopy has brought new optimism to the attempt to use gels automatically to analyze all of the proteins formed by a cell.[a,c] The methods are potentially very important to developing new diagnostic procedures for human medicine. Watching changes in the proteome, including posttranslational modifications in proteins, as cells develop and grow will provide new insights into biochemical regulation.

A similar concept is that of a **complete transcriptional map**, a record of all of the different RNA molecules being synthesized by a cell.[d,e] These include many different mRNAs, each of which may give rise to more than one protein, as well as many RNAs with other functions.

---

[a] Kahn, P. (1995) *Science* **270**, 369–370
[b] Swinbanks, D. (1995) *Nature (London)* **378**, 653
[c] Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko.A, Boucherie, H., and Mann, M. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14440–14445
[d] Richard, G.-F., Fairhead, C., and Dujon, B. (1997) *J. Mol. Biol.* **268**, 303–321
[e] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10614–10619

the lysine residues are protected and trypsin will cleave only at the Arg-X positions. If the resultant peptides are separated and held at pH 3.5 overnight, the blocking groups are hydrolyzed off and a second trypsin treatment can be used to cleave at the Lys-X positions.[225] The number of cleavage sites for trypsin can be increased by converting an –SH group to a positively charged one by aminoethylation (Eq. 3-27).[226] The reaction can be accomplished either with ethyleneimine (caution: carcinogen) as shown in this equation or by bromoethylamine, which eliminates $Br^-$ to form ethyleneimine.



(3-27)

Because of its specificity for basic residues, trypsin converts a protein into a relatively small number of **tryptic peptides** which may be separated and characterized. Trypsin acts primarily on denaturated proteins, and to obtain good results the disulfide bridges must be broken first. **Chymotrypsin** is less specific than trypsin and **pepsin** is even less specific (Table 3-2). Nevertheless, they can be used to cut a peptide chain into smaller fragments whose sequences can be determined. To establish the complete amino acid sequence for a protein, "overlapping" peptide fragments must be found that contain sequences from ends of two different tryptic fragments. In this way the tryptic peptides can be placed in the order in which they occurred in the native protein. This tedious procedure is rarely used today. Peptide sequencing is still important but is usually coordinated with gene sequencing, X-ray structure determination, or mass spectroscopy which minimize the need for overlapping fragments.

While trypsin cuts the peptide linkages Lys-X and Arg-X, a fungal protease cleaves only X-Lys.[227] A protease from the submaxillary glands of mice cleaves only Arg-X,[228] one from *Staphylococcus* specifically at Glu-X,[229,230] and one from kidneys at Pro-X.[231]

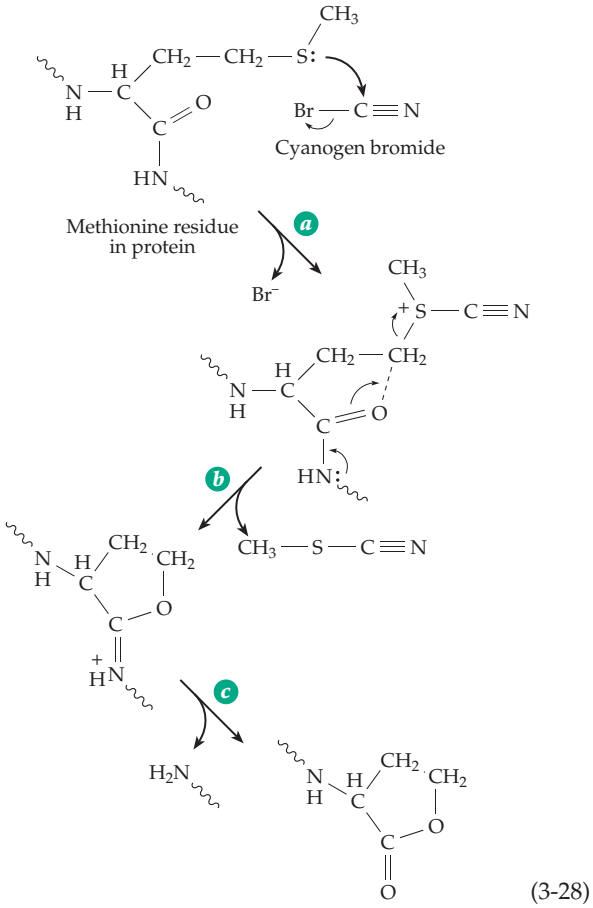
Several enzymes catalyze stepwise removal of amino acids from one or the other end of a peptide chain. **Carboxypeptidases**[232] remove amino acids from the carboxyl-terminal end, while **aminopeptidases** attack the opposite end. Using chromatographic methods, the amino acids released by these enzymes may be examined at various times and some idea of the sequence of amino acids at the chain ends may be obtained. A **dipeptidyl aminopeptidase** from bovine spleen cuts dipeptides one at a time from the amino terminus of a chain. These can be converted to volatile trimethylsilyl derivatives and identified by mass spectrometry.[233] If the chain is shortened by one residue using the Edman degradation (Section 3) and the dipeptidyl aminopeptidase is again used, a different set of dipeptides that overlaps the first will be obtained and a sequence can be deduced. Carboxypeptidase Y can be used with MALDI mass spectrometry to deduce the C-terminal amino acid sequence for a peptide. However, Ile and Leu cannot be distinquished.

**TABLE 3-2**
**Specificities of Commonly Used Protein-Hydrolyzing Enzymes**

| | | |
|---|---|---|
| Trypsin | Lys-X, Arg-X | X not Pro |
| Chymotrypsin | | |
| rapidly: | Phe-X, Tyr-X, Trp-X | X not Pro |
| slowly: | Y-X<br>Y=Leu, Asn, Gln, His, Met, Ser, Thr | X not Pro |
| *Staphylococcus aureus* protease V-8 | Glu-X | X not Pro |
| Clostripain | Arg-X | |
| Pepsin | | |
| preferentially: | X-Phe-X, X-Tyr-X, X-Leu-X | |
| less so: | X-Ala-X | |
| Thermolysin | | |
| rapidly: | X-Y<br>Y=Ile, Leu, Val, Ala, Phe, Met | |
| slowly: | X-Y<br>Y=Tyr, Gly, Thr, Ser | |

***Nonenzymatic cleavages.*** Of the various nonenzymatic methods that have been proposed, one has been outstandingly useful. Cyanogen bromide, N ≡ C – Br, cleaves peptide chains adjacent to methionine residues. The sulfur of methionine displaces the bromide ion (Eq. 3-28) and because of a favorable spatial relationship, the resulting sulfonium compound undergoes C–S bond cleavage through participation of the adjacent peptide group (Eq. 3-28, step *b*). The C=N of the product is then hydrolyzed with cleavage of the peptide chain in step *c*.

The linkage Asp-Gly can often be cleaved specifically by treatment with hydroxylamine at high pH.[195] Procedures for specific cleavage of tryptophanyl bonds have been de-




(3-28)

vised.[234] The Asp-Pro linkage is susceptible to cleavage by trifluoroacetic acid, which is used in the automated Edman degradation employed in peptide sequencing (Eq. 3-30). Cleavage by trifluoroacetic acid can be used to generate peptides for subsequent sequence determination.

***Separating the peptides.*** A procedure that has been very important in the development of protein chemistry is **peptide mapping** or "fingerprinting." The procedure begins with cleavage of the disulfide linkages, denaturation, and digestion with trypsin or some other protease. The sizes and amino acid compositions of the resulting series of peptides are characteristic of the protein under study. The mixture of peptides is placed on a thin layer plate and subjected to chromatography in one direction, then to electrophoresis in the other direction, with the peptides separating into a characteristic pattern or fingerprint. Fingerprinting has been especially useful in searching for small differences in protein structure, for example, between genetic variants of the same protein (Fig. 7-27). Currently, the peptides are usually separated on ion exchange or gel filtration columns, by reversed-phase HPLC, or by capillary electrophoresis and are then often passed, in subpicomole amounts, into a mass spectrometer.[176,235,236]

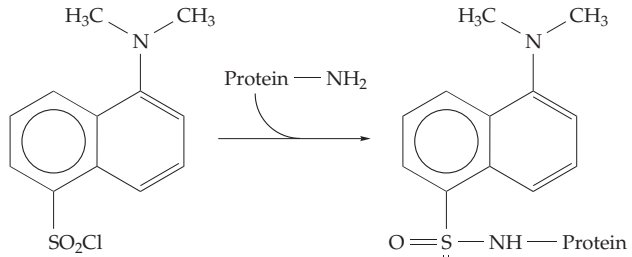
## 3. Determining Amino Acid Sequence

The covalent structure of insulin was established by Frederick Sanger in 1953 after a 10-year effort. This was the first protein sequence determination.[237,238] Sanger used partial hydrolysis of peptide chains whose amino groups had been labeled by reaction with 2,4-dinitrofluorobenzene[239] to form shorter end-labeled fragments. These were analyzed for their amino acid composition and labeled and hydrolyzed again as necessary. Many peptides had to be analyzed to deduce the sequence of the 21-residue and 30-residue chains that are joined by disulfide linkages in insulin.[237,238]

The Sanger method is mainly of historic interest, although end-labeling may still be used for various purposes. A more sensitive labeling reagent than was used by Sanger is **dansyl chloride**. It reacts to form a sulfonamide linkage that is stable to acid hydrolysis and is brilliantly fluorescent (Eq. 3-29). The related reagent dimethylaminoazobenzene-4'-sulfonyl chloride gives highly colored derivatives easily seen on thin-layer chromatography plates.[240]

***The Edman degradation.*** One of the most important reagents for sequence analysis is **phenylisothiocyanate,** whose use was developed by P. Edman.[241–243] This reagent also reacts with the N-terminal amino group of peptides (Eq. 3-30, step *a*). The resulting adduct undergoes cyclization with cleavage of the peptide linkage (Eq. 3-30, step *b*) under acidic conditions. After rearrangement (step *c*) the resulting **phenylthiohydantoin** of the N-terminal amino acid can be identified. The procedure can then be repeated on the shortened peptide chain to identify the amino acid residue in the second position. With careful work the Edman degradation can be carried down the chain for several tens of residues.

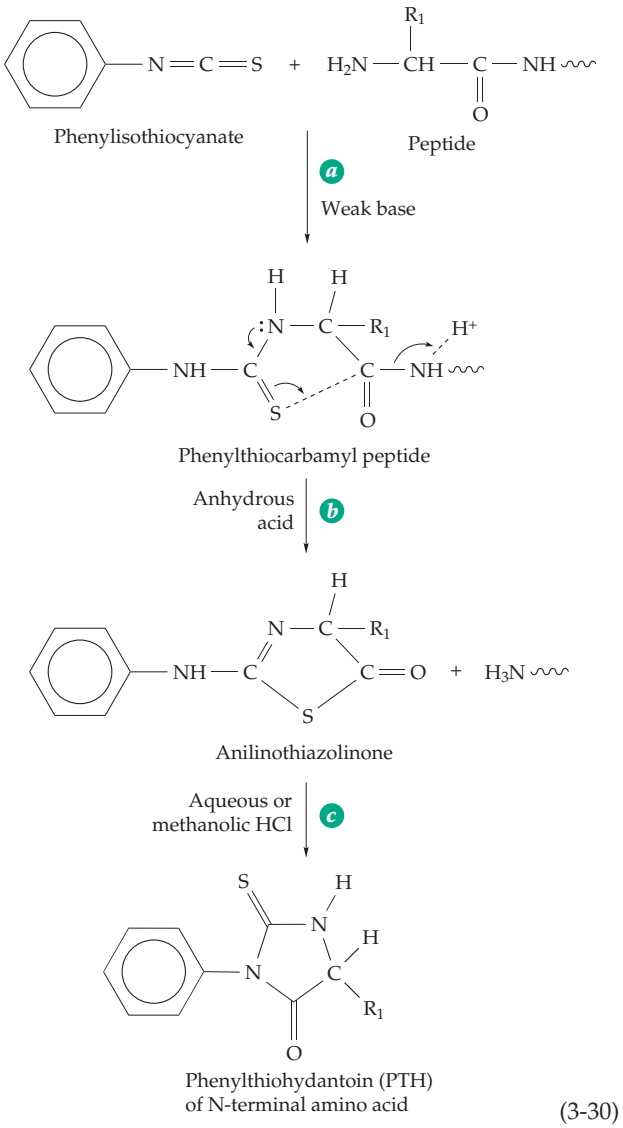
Ingenious **protein sequenators** have been devised to carry out the Edman degradation automatically.[242,244–246] Each released phenylthiohydantoin is then identified by HPLC or other techniques. Commercial sequenators have often required 5–20 nmol of peptide but new microsequenators can be used with amounts as low as 5–10 picomoles or less.[247,248]




(3-29)

Microsequencers permit sequence analysis on minute amounts of protein. Microsequencing can be used in conjunction with two-dimensional electrophoretic separations of proteins such as that shown in Box 3-C. The proteins in the polyacrylamide gel are electrophoretically transferred onto a porous sheet (membrane) of an inert material such as polyvinyl difluoride.[249–251] After staining, a selected spot is cut out and placed into the sequencer. To avoid the problems associated with blocked N termini, the protein may be treated with proteases on the membrane and the resulting peptide fragments may then be separated on a narrow-bore HPLC column and sequenced.[240]

Because many proteins are modified at the N terminus, blocking application of the Edman degradation, it would be useful to have a similar method for sequencing from the C terminus. It has been difficult to devise a suitable strategy, but there has been some success.[252–254]




(3-30)

***Protein sequences from the genes.*** Complete sequences of large numbers of genes have been determined and the corresponding sequences of proteins can be read directly from those of the corresponding genes. One method for sequencing a gene is to isolate a specific messenger RNA, which does not contain intervening sequences. A DNA copy (cDNA) is made from the mRNA and is used to ascertain the sequence of the encoded protein. The genomic DNA is also often sequenced. Introns are recognized by the nucleotide sequences at their ends and the correct amino acid sequence for the encoded protein is deduced.

In many instances, however, a gene can be identified only after part of the protein, often an N-terminal portion has been sequenced. This knowledge permits synthesis of an **oligonucleotide probe** that can be used to locate the gene (Chapter 5). Nucleotide sequences can be verified by comparison with sequences of tryptic or other fragments of a protein. Similarly, protein sequences are often checked by sequencing the corresponding genes as well as by study of X-ray structures. Substantial numbers of errors are made in sequencing of both DNA and protein so that checking is important.

***Mass spectrometry in sequencing.*** Proteins can also be sequenced by mass spectrometry or by a combination of Edman degradation and mass spectrometry.[213,255,255a] Until recently the peptides had to be converted to volatile derivatives by extensive methylation and acetylation or by other procedures. However, newer ionization methods including MALDI (Fig. 3-11) and ESI (Fig. 3-12) have made it possible to obtain mass spectra on unmodified peptides. In one procedure a nonspecific protease cleaves a peptide chain into a mixture of small oligopeptides which are separated by HPLC into 20–40 fractions, each of which may contain 10–15 peptides but which can be sent directly into the ionization chamber of the mass spectrometer.[256] Peptides can be generated from a protein using immobilized enzymes, separated on a chromatographic column, and introduced sequentially into the mass spectrometer. Examination of peptide mixtures by mass spectrometry provides a way of verifying sequences deduced from DNA sequencing.[257] Mass spectrometry is also used widely to study covalently modified proteins.[173,216,257a,257b] As a rule, these cannot be recognized from gene sequences.

## 4. Locating Disulfide Bridges

A final step in sequencing is often the location of S–S bridges. The reduced and alkylated protein can be cleaved enzymatically (e.g., with elastase, pepsin, or thermolysin) to relatively small fragments, each of which contains no more than one modified cysteine.

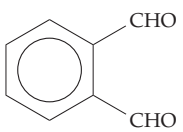
The same enzymatic cleavage can then be applied to the unreduced enzyme. Pairs of peptide fragments remain linked by the S–S bridges. These crosslinked pairs can be separated, the disulfide bridges cleaved, and the resulting peptides identified, each as one of the already sequenced fragments. Mass spectrometry provides a rapid method for their identification.[258]

Another elegant way of locating S–S bridges employs **diagonal electrophoresis**. Electrophoresis of the digest containing the crosslinked pairs is conducted in one direction on a sheet of filter paper. Then the paper is exposed to performic acid vapor to cleave the bridges according to Eq. 3-22 and electrophoresis is conducted in the second direction and the paper is sprayed with ninhydrin. The spots falling off the diagonal are those that participated in S–S bridge formation. They can be associated in pairs from their positions on the paper and can be identified with peptides characterized during standard sequencing procedures.[259]

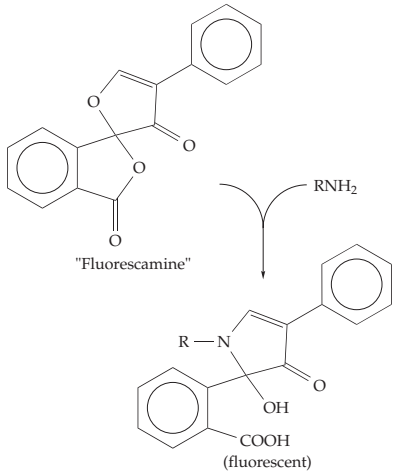
Diagonal electrophoresis and its relative diagonal chromatography are useful for other purposes as well. After electrophoresis or chromatography is conducted in one direction, the paper or thin-layer plate may be sprayed with a reagent that will react with some components or may be irradiated with light before the separation is repeated in the second direction (Fig. 3-5).[102,260]

## 5. Detecting Products

Important to almost all biochemical activity is the ability to detect, and to measure quantitatively, tiny amounts of specific compounds. "Color reagents," which develop characteristic colors with specific compounds, are especially popular. For example, **ninhydrin** (Box 3-E) can be used as a "spray reagent" to detect a small fraction of a micromole of an amino acid or peptide in a spot on a chromatogram. It can also be used for a quantitative determination, the color being developed in a solution. More sensitive than absorption of light (color) is fluorescence. **Fluorescamine** (Eq. 3-31) reacts with any primary amine to form a highly fluorescent product. As little as 50 pmol of amino acid can be determined quantitatively.[261] A yet more sensitive fluorogenic reagent for detection of amino acids, peptides, and amines of all types is *o*-phthaldialdehyde.[262,263]

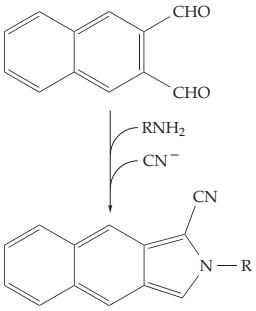



*o*-Phthaldialdehyde



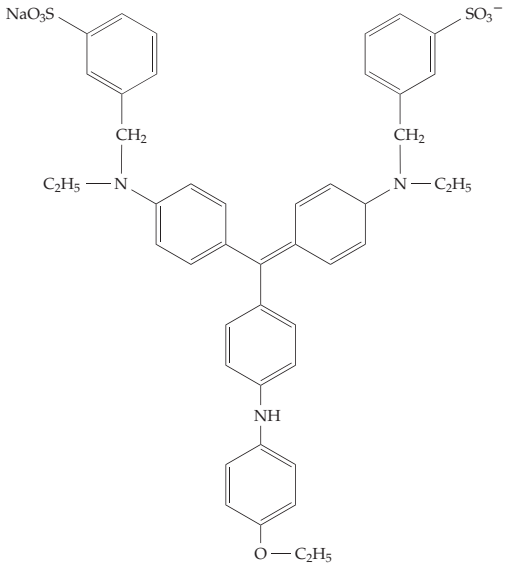

(3-31)

Reaction with naphthalene 2,3-dicarboxaldehyde (Eq. 3-32) increases the limit of detection 100-fold or more.[264]




(3-32)

Detection of proteins on thin-layer plates, gel slabs, or membranes is often accomplished by staining with a dye,[265–267] the most widely used being Coomassie brilliant blue.[268] Various silver-containing stains may also be used. After separation of a protein mixture by electrophoresis and transfer to an inert membrane,

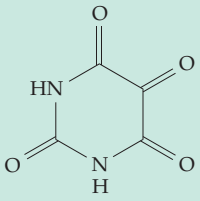



Coomassie brilliant blue

## BOX 3-E NINHYDRIN

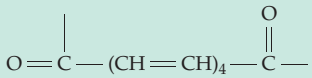
Ninhydrin (1,2,3-indantrione monohydrate) forms Schiff bases (ketimines) with amino acids. These react in ways similar to those of Schiff bases of pyridoxal phosphate (Chapter 14). Decarboxylation of the ketimines followed by hydrolysis of the resulting aldimines yields an intermediate amine that can couple with a second molecule of ninhydrin to form a characteristic purple color.[a,b] First reported by Ruhemann in 1910, the intermediate amine can also be hydrolyzed to free ammonia. Therefore, to ensure maximum color yield ninhydrin solutions for quantitative analysis usually contain reduced ninhydrin, which can react with free $NH_3$ and ninhydrin to form Ruhemann's purple (see scheme). The reaction has been widely used in chromatography and in quantitative amino acid analysis and also as a convenient spray reagent for paper and thin-layer chromatography. While α-amino acids react most readily, primary amines and peptides also form Ruhemann's purple. In these cases a proton rather than $CO_2$ is lost from the ketimine. When pyridoxamine (Chapter 14) on chromatograms reacts, a bright orange product, presumably the aldimine, appears. Secondary amines, such as proline, give a yellow color.

Both the ninhydrin reaction and pyridoxal phosphate-catalyzed decarboxylation of amino acids (Chapter 14) are examples of the **Strecker degradation**. Strecker reported in 1862 that alloxan causes the decarboxylation of alanine to acetaldehyde, $CO_2$, and ammonia.[c]

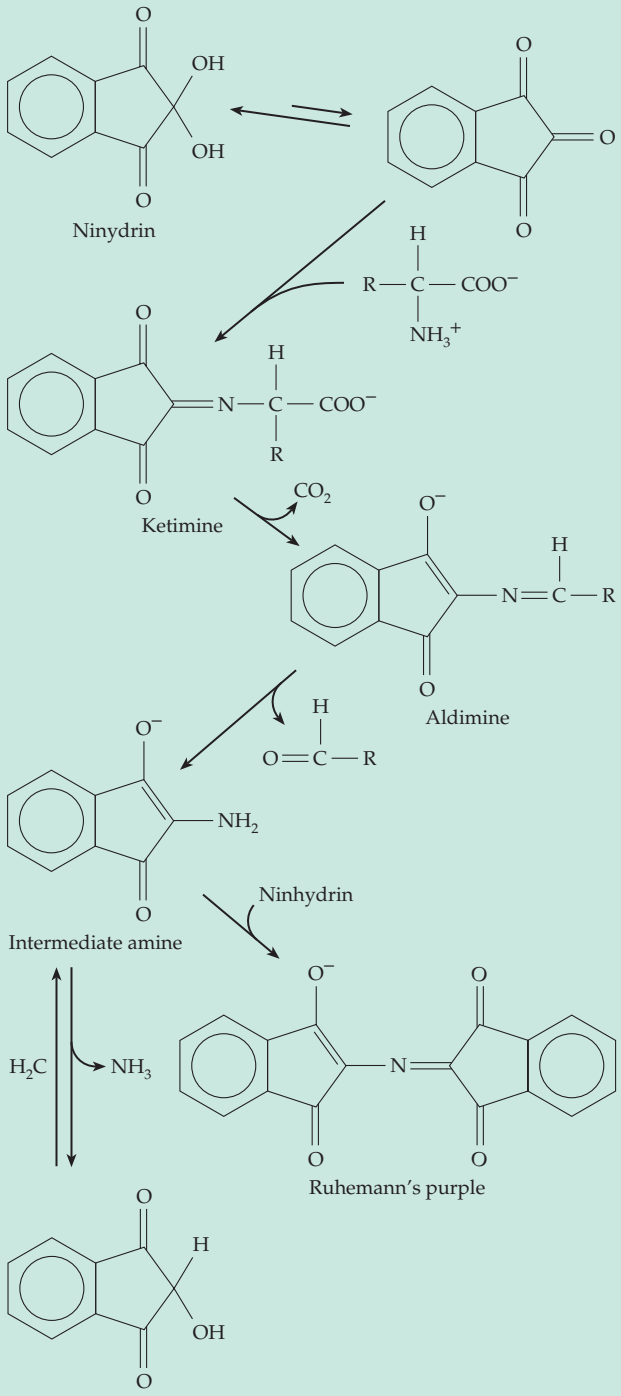


Alloxan

Many other carboxyl compounds, e.g., those of the general structure



and *p*-nitrosalicylaldehyde also cause the Strecker degradation.[d]




[a] Wigfield, D. C., and Croteau, S. M. (1980) *Biochem. Edu.* **8**, 26–27
[b] Friedman, M., and Williams, L. D. (1974) *Bioorg. Chem.* **3**, 267–280
[c] Strecker, A. (1862) *Annalen* **123**, 363–365
[d] Schonberg, A., and Moubacher, R. (1952) *Chem. Rev.* **50**, 261–277

the resulting protein "blots" can be stained with specific antibodies.[269,270] Flame ionization detectors can measure as little as a few picomoles of almost any substance leaving a vapor-phase chromatographic column. The importance of developing new, more sensitive analytical methods by which the quantity of material investigated can be scaled down can hardly be overemphasized. Increasingly sensitive methods of detection, including mass spectrometry, now permit measurement of fmol ($10^{-15}$ mol) quantities in some cases. With this ability the output of neurotransmitters from a single neuron in the brain can be measured and the contents of single cells can be analyzed.

## 6. Absorption of Light

Side chains of the three aromatic amino acids phenylalanine, tyrosine, and tryptophan absorb ultraviolet light in the 240- to 300-nm region, while histidine and cystine absorb to a lesser extent. Figure 3-13 shows the absorption spectrum of a "reference compound" for tyrosine. There are three major absorption bands, the first one at 275 nm being a contributor to the well-known 280-nm absorption band of proteins. There is a much stronger absorption band at about 240 nm. Sensitive methods for estimating protein concentration depend upon the measurement of this absorption together with that from other side chains at around 280 or 230 nm.[13,271–274] There is an even stronger absorption band at 192 nm. However, at these wavelengths even air absorbs light and experimental difficulties are extreme. At 280 nm, and even more at 230 nm, it is easy to contaminate samples with traces of light-absorbing material invisible to the eye. Therefore, most estimations of protein concentration from light absorption depend upon the 280-nm band.

Figure 3-14 shows the spectra of *N*-acetyl ethyl esters of all three of the aromatic amino acids and of cystine. To a first approximation, the absorption spectra of proteins can be regarded as a summation of the spectra of the component amino acids. However, the absorption bands of some residues, particularly of tyrosine and tryptophan, are shifted to longer wavelengths than those of the reference compounds in water. This is presumably a result of being located within nonpolar regions of the protein. Notice that the spectra for tyrosine, phenylalanine, and cystine in Fig.

### BOX 3-F BIOSENSORS AND ELECTRONIC NOSES

A new approach to detection of molecules of biological interest is the development of biosensors. These are small devices that detect the binding of specific molecules to a **receptor** which is in intimate contact with a specially prepared surface that serves as a **transducer**. The receptor might be a layer of enzyme, antibody, hormone receptor, lectin, or oligonucleotide. Binding of substrate, antigen, hormone, sugar, or complementary polynucleotide strand, respectively, induces a response consisting of some kind of electrical or optical signal.[a–d] If the sensor is constructed on a semiconductor chip changes in an imposed potential difference may be detected.[a,e] However, changes in optical properties are more often observed. Fluorescence of dyes incorporated into the tranducing layer may be induced by binding of a molecule to a protein that undergoes an allosteric modification (see Chapter 9).[f] Many biosensors measure **surface plasmon resonance**, a change in the evanescent wave that develops in a surface when a light beam at the angle of total reflectance strikes the surface. This induces a change in the dielectric constant which can be measured.[f–m] Biosensors are used to estimate binding constants and also rate constants. However, read the article by Schuck and Milton[m] for tests of the validity of kinetic data. Biosensors can serve as "electronic noses." One possible application is in the analysis of compounds in human breath as an aid to medical diagnosis. Over 400 volatile organic compounds have been identified in breath using gas chromatography and mass spectrometry.[n]


[a] Briggs, J. (1987) *Nature (London)* **329**, 565–566
[b] Zurer, P. (1997) *Chem. Eng. News* **September 15**, 7
[c] Kress-Rogers, E., ed. (1997) *Handbook of Biosensors and Electronic Noses: Medicine, Food, and the Environment*, CRC Press, Boca Raton, Florida
[d] Cunningham, A. J. (1998) *Introduction to Bioanalytical Sensors*, Wiley, New York
[e] McConnell, H. M., Owicki, J. C., Parce, J. W., Miller, D. L., Baxter, G. T., Wada, H. G., and Pitchford, S. (1992) *Science* **257**, 1906–1912
[f] Marvin, J. S., Corcoran, E. E., Hattangadi, N. A., Zhang, J. V., Gere, S. A., and Hellinga, H. W. (1997) *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4366–4371
[g] Raether, H. (1988) *Surface Plasmons, Springer Tracts in Modern Physics*, Vol. 111, Springer-Verlag, Berlin
[h] Peterlinz, K. A., Georgiadis, R. M., Herne, T. M., and Tarlov, M. J. (1997) *J. Am. Chem. Soc.* **119**, 3401–3402
[i] Hendrix, M., Priestley, E. S., Joyce, G. F., and Wong, C.-H. (1997) *J. Am. Chem. Soc.* **119**, 3641–3648
[j] Salamon, Z., Brown, M. F., and Tollin, G. (1999) *Trends Biochem. Sci.* **24**, 213–219
[k] Chao, H., Houston, M. E., Jr., Grothe, S., Kay, C. M., O'Connor-McCourt, M., Irvin, R. T., and Hodges, R. S. (1996) *Biochemistry* **35**, 12175–12185
[l] McNally, A. J., Mattsson, L., and Jordan, F. (1995) *J. Biol. Chem.* **270**, 19744–19751
[m] Schuck, P., and Minton, A. P. (1996) *Trends Biochem. Sci.* **21**, 458–460
[o] Phillips, M. (1992) *Sci. Am.* **267**(July), 74–79

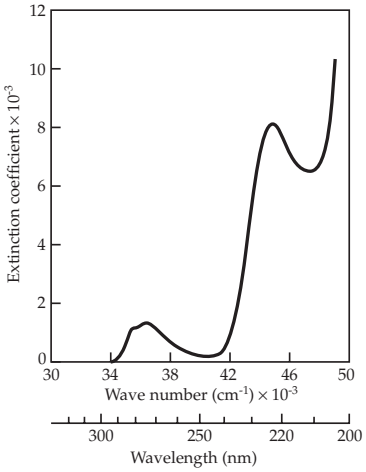



**Figure 3-13** The absorption spectrum of *N*-acetyltyrosine ethyl ester in an aqueous phosphate buffer of pH 6.8. Absorbance (as molar extinction coefficient, Eq. 23-5) is plotted against increasing energy of light quanta in units of wave number. The more commonly used wavelength scale is also given. Spectra are most often presented with the low wavelength side to the left. In the convention adopted here the energy of a quantum increases to the right. There are three $\pi-\pi^*$ electronic transitions that give rise to absorption bands of increasing intensity. The third $\pi-\pi^*$ transition of the aromatic ring is at ~ 52,000 $cm^{-1}$ (192 nm) and reaches a molar extinction coefficient of ~ 40,000. The $n-\pi^*$ and $\pi-\pi^*$ transitions of the amide group in this compound also contribute to the high energy end of the spectrum (see Chapter 23 for further discussion).

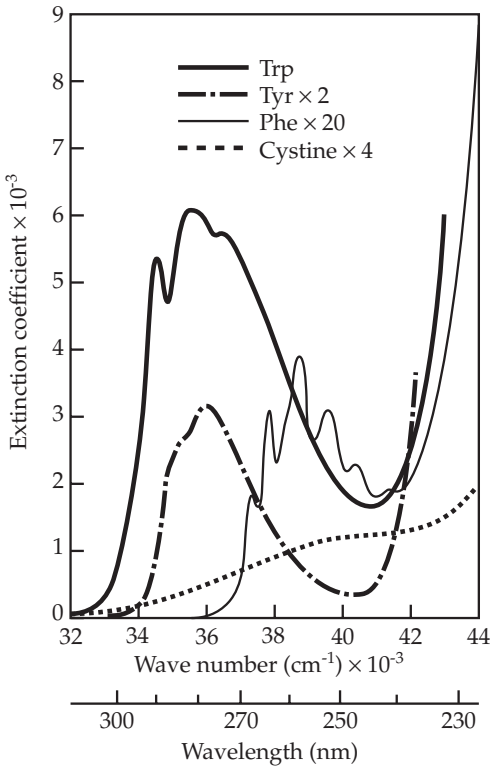



**Figure 3-14** The spectra of the first electronic transitions of the *N*-acetyl derivatives of the ethyl esters of phenylalanine, tyrosine, and tryptophan together with that of the dimethyl ester of cystine in methanol at 25°C. The spectra for the Tyr, Phe, and cystine derivatives have been multiplied by the factors given on the graph.[272]

3-14 have been multiplied by factors of 2 to 20. It is evident that if all of the light-absorbing side chains were present in equal numbers tryptophan would dominate the absorption band and that phenylalanine would contribute little except some small wiggles. The molar extinction coefficient ε can be estimated from the numbers of residues of each type per molecule as follows:[274]

$$\varepsilon_{280}\ (M^{-1}cm^{-1}) = 5500\ (\text{no. Trp}) + 1490\ (\text{no. Tyr}) + 125\ (\text{no. cystine}) \qquad (3\text{-}33)$$

For proteins of unknown composition, a useful approximation is that a solution containing 1 mg / ml of protein has an absorbance at 280 nm of about 1.0.

## E. Quantitative Determinations and Modification Reactions of Side Chain Groups

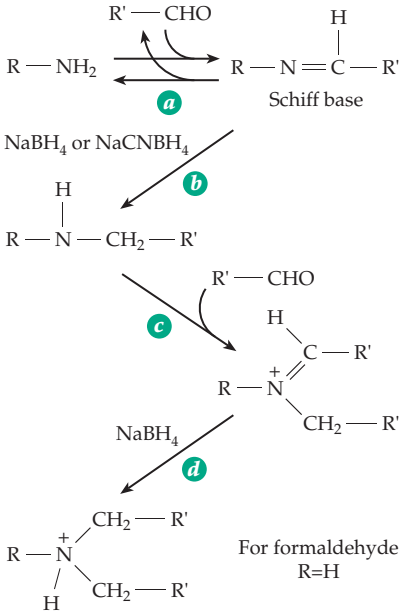
The functional groups present in the side chains of proteins include $-NH_2$, –SH, S—S, –OH, $-COO^-$, the imidazole group of histidine, the guanidine group of arginine, the phenolic group of tyrosine, the indole ring of tryptophan, and the $-S-CH_3$ group of methionine. These are able to enter into a great variety of chemical reactions, most of which make use of the nucleophilic properties of these groups. The reactions are most often those of nucleophilic **addition** or nucleophilic **displacement**. The basic chemistry of these reactions often parallels biochemical reactions that are discussed in Chapters 12 and 13. In many instances, the reactions are nonspecific; amino, thiol, and hydroxyl groups may all react with the same reagent. The usefulness of the reactions depends to a large extent on the discovery of conditions under which there is some selectivity. It is also important that the reactions be complete. Only a few reactions will be considered here; these and others have been reviewed by Glazer *et al.*[225,275]

### 1. Reactions of Amino Groups

The numerous amino groups of lysine residues and of the N termini of peptide chains usually pro-

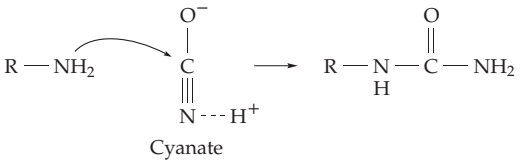
trude into the aqueous surroundings of a protein. Chemical modification can be done in such a way as to preserve the net positive charge which amino groups carry at most pH values, to eliminate the positive charge leaving a neutral side chain, or to alter the charge to a negative value. Alterations of these charges can greatly affect interactions of the protein molecules with each other and with other substances.

Amino groups react reversibly with carbonyl compounds to form Schiff bases.[275–277] Reduction of the latter by sodium borohydride or sodium cyanoborohydride causes an irreversible change (Eq. 3-34, steps *a* and *b*). Cyanoborohydride is specific for Schiff bases and does not reduce the carbonyl compound. However, side products may cause problems.[277,278] Depending upon which carbonyl compound is used, the net positive charge on the amino group may be retained or it may be replaced with a different charge by this "reductive alkylation" sequence. Formaldehyde will react according to Eq. 3-34 in two steps to give a dimethyl amino group with no change of net charge.[279] Pyridoxal phosphate (Chapter 14) is converted by Eq. 3-34 into a fluorescent label. With a limited amount of pyridoxal phosphate only one or a few lysine residues may be labeled, often at active centers of enzymes. Schiff bases formed from glyceraldehyde in Eq. 3-34 can undergo the Amadori rearrangement (Eq. 4-8) to form stable products which, however, can be reconverted to the original amino groups upon acid hydrolysis. The borohydride reduction product of Eq. 3-34 with glyceraldehyde can be reconverted to the original amine by periodate oxidation (Eq. 4-11).[277]



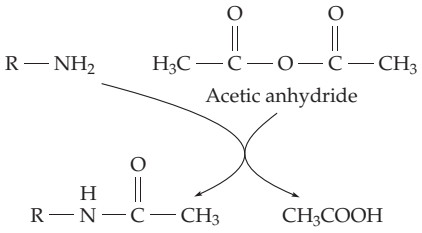

(3-34)

Another addition reaction of amino groups is **carbamoylation** with sodium cyanate (Eq. 3-35). A displacement reaction by an amino group on an acid




(3-35)

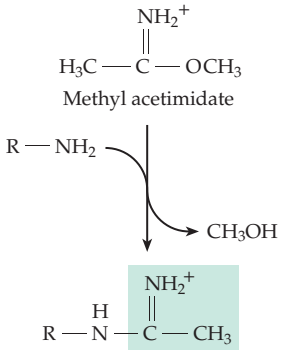
anhydride such as acetic anhydride (Eq. 3-36) leads to **acylation**, a nonspecific reaction which is also undergone by thiol, hydroxyl, and other groups. When acetic anhydride is used, the net positive charge of an



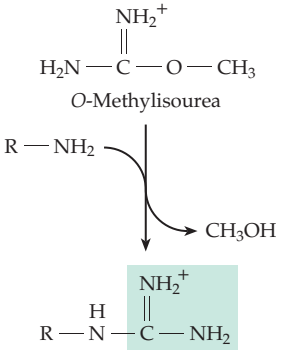

(3-36)

amino group is lost. However, the product obtained with succinic anhydride or maleic anhydride (Eq. 3-26) carries a negative charge. In the latter case, the modification can readily be reversed by altering the pH.

Both **amidination** (Eq. 3-37) and **guanidination** (Eq. 3-38) lead to retention of the positive charge.

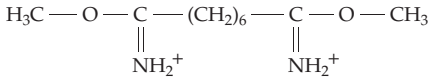



(3-37)




(3-38)

Bifunctional imidoesters such as **dimethylsuberimidate** may be used to establish whether or not two different proteins or subunits are close together in a complex or in a supramolecular structure such as a membrane or ribosome.

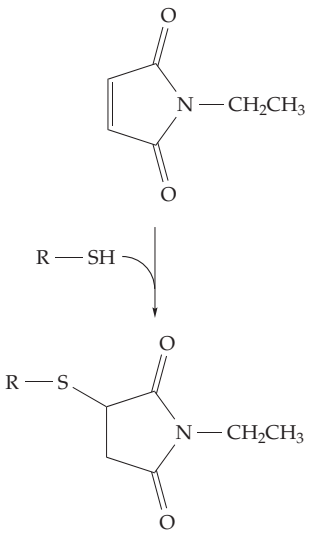


Dimethylsuberimidate

Another useful reaction of amino side chains is that with dansyl chloride (Eq. 3-29). Many lysine derivatives can be determined quantitatively by amino acid analysis.[280]

## 2. Reactions of SH Groups

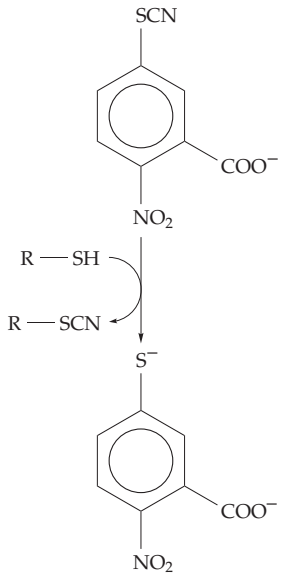
In addition to the alkylation with iodoacetate (Eq. 3-24), sulfhydryl groups can react with *N*-ethylmaleimide (Eq. 3-39).[281] This reaction blocks the SH groups irreversibly and has often been used in attempts to establish whether or not a thiol group plays a role in the functioning of a protein. Loss of function in the



(3-39)

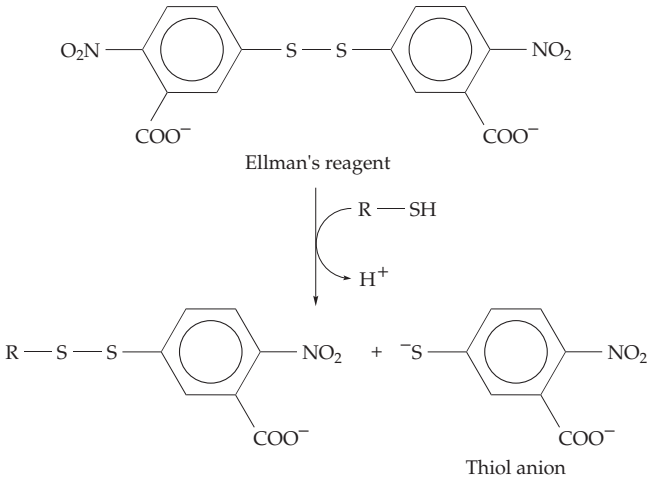
presence of this "sulfhydryl reagent" may mean that an SH group has an essential role or it could be a result of the bulk of the group added. The *N*-ethylmaleimide group is large and could prevent proper contact between an enzyme and substrate or between two proteins. To avoid the possible effect of excessive bulk, it is useful to convert the SH to the small thiocyanate group – SCN (Eq. 3-40).[282]

Ellman's reagent 5,5'-dithiobis(2-nitrobenzoic acid; DTNB) reacts quantitatively with – SH groups (Eq. 3-41) to form mixed disulfides with release of a thiolate anion that absorbs light at 412 nm with a molar extinction



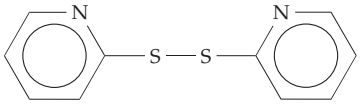
(3-40)

coefficient of 14,150 $M^{-1}$ $cm^{-1}$.[283] While DTNB has been widely used to determine the content of –SH groups in proteins, there are some disadvantages.



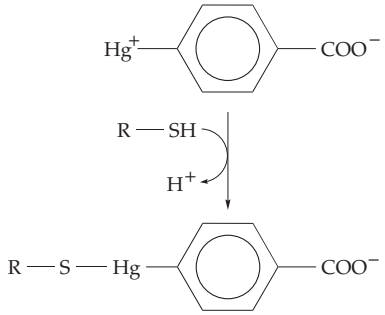
(3-41)

Pyridyldisulfides such as 2-pyridyldisulfide or the isomeric 4-pyridyldisulfide react more completely and with greater selectivity.[281]




Thiol groups have a high affinity for mercury ions including organic mercury derivatives, which are widely used in the determination of protein structures by X-ray crystallography (Section F). Titration of SH groups in proteins is often accomplished with

*p*-mercuribenzoate (Eq. 3-42). The reaction may be followed spectrophotometrically at 250 – 255 nm, a region in which the mercaptide product absorbs strongly.

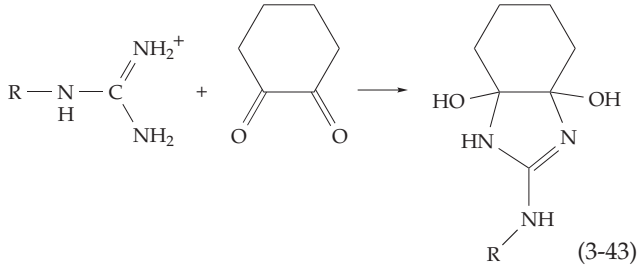



(3-42)

## 3. Reactions of Other Side Chains

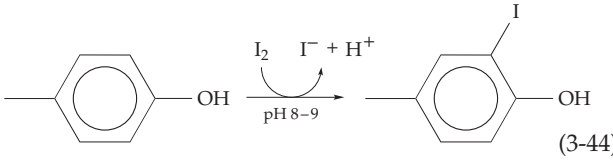
There are no highly selective reactions for –OH, –$COO^-$, or imidazole groups. However, some hydroxyl groups in active sites of enzymes are unusually reactive in nucleophilic addition or displacement and can be modified by acylation, phosphorylation, or in other ways. Carboxyl groups, which are exceedingly numerous on protein surfaces, can be modified by treating with a water-soluble carbodiimide (Eq. 3-10) in the presence of a high concentration of an amine such as the ethyl ester of glycine. The imidazole groups of residues of histidine can often be selectively destroyed by dye-sensitized photooxidation (Ch. 12, Section D,5) or can be acylated with ethoxyformic anhydride.

Compounds with two adjacent carbonyl groups such as 1,2-cyclohexanedione (Eq. 3-43) react selectively with guanidinium groups from arginine residues in proteins. Under certain conditions the product indicated in Eq. 3-43 predominates. Related reagents are derived from camphorquinone.[284]



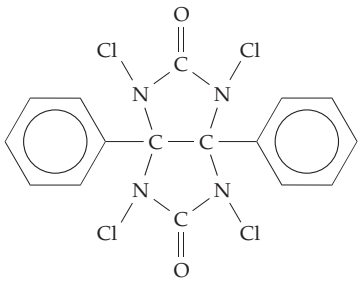

(3-43)

The phenolic group of tyrosine undergoes iodination (Eq. 3-44), acylation, coupling with diazonium compounds, and other reactions.



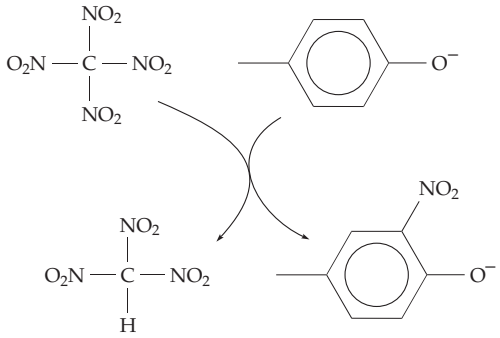

(3-44)

The following sparingly soluble chloroamide together with $I^-$ will also iodinate tyrosine and can be used to incorporate radiolabeled iodine into proteins.[285,286]




1,3,4,6-Tetrachloro-3α,6α-diphenylglycouril

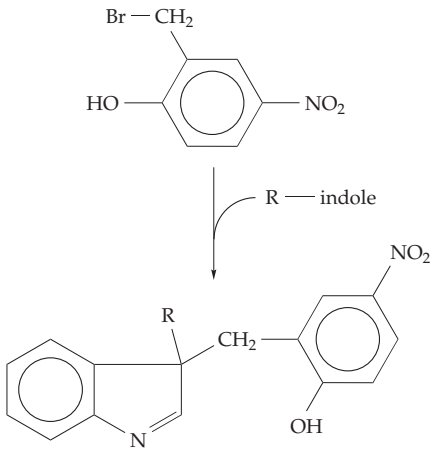
Tetranitromethane reacts slowly with tyrosyl groups to form 3-nitrotyrosyl groups (Eq. 3-45). The by-product **nitroform** is intensely yellow with $\varepsilon_{350}$ = 14,400. The reagent also oxidizes SH groups and reacts with other anionic groups.




Nitroform

(3-45)

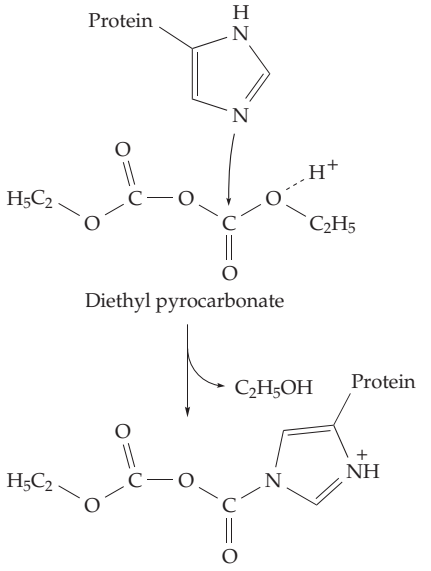
Koshland devised the following reagent for the indole rings of tryptophan residues (Eq. 3-46).




(3-46)

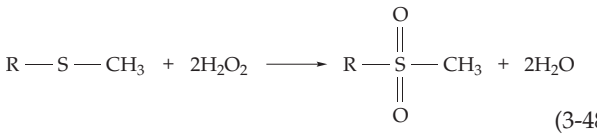
Imidazole, lysine amino groups, and tyrosine hydroxyl groups react with **diethylpyrocarbonate**

(Eq. 3-47) at low enough pH (below 6) that the reaction becomes quite selective for histidine.[70,287] Reactivity with this reagent is often used as an indication of histidine in a protein.[288–290]. The reaction may be monitored by observation of NMR resonances of imidazole rings.[288,290]




(3-47)

The thioether side chains of methionine units in proteins can be oxidized with hydrogen peroxide to the corresponding sulfones (Eq. 3-48). They can also be alkylated, e.g., by $CH_3I$ to form $R—S^+(CH_3)_2$.




(3-48)

## 4. Affinity Labeling

To identify groups that are part of or very near to the active site of a protein, reagents can be designed that carry a reactive chemical group into the active site.[291] The related **photoaffinity labeling**[292,293] is also widely used (see also Chapter 23).

# F. Synthesis of Peptides

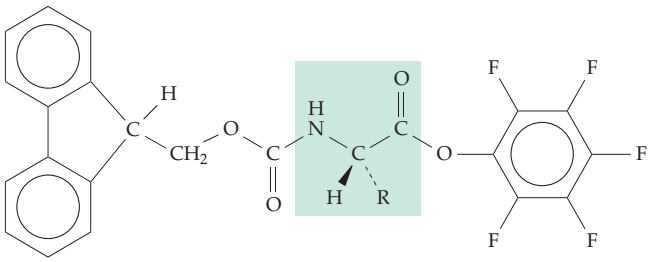
The synthesis of peptides of known sequence in the laboratory is extremely important to biochemical research. For example, we might want to know how the effects of a peptide hormone are altered by replacement of one amino acid in a particular position by another. The synthetic methods must be precise[294–298] and because there are so many steps the yield should be 98% or better for every step. Even so, it is still impractical to synthesize very large peptides. Those that have been made, such as the hormone insulin and the enzyme ribonuclease, have been obtained in low yields and have been difficult to purify.[298] It is usually more practical to obtain large peptides from natural sources. It is often practical to clone a suitable piece of DNA in a bacterial plasmid and to set up biological production of the desired peptide (Chapter 26). On the other hand, for smaller peptides, laboratory synthesis is feasible. Even for large peptides it is useful because it permits incorporation of unnatural amino acids as well as isotopic labels.

The general procedure for making a peptide in the laboratory is to "block" the amino group of what will become the N-terminal amino acid with a group that can be removed later. The subsequent amino acid units "activated" at their carboxyl end are then attached one by one. The chemical activation is often accomplished by conversion of the carboxyl group of the amino acid to an anhydride. At the end of the synthesis, the blocking group must be removed from the N terminus and also from various side chain groups such as those of cysteine and lysine residues. In many respects, this procedure is analogous to the biological synthesis of proteins whose basic chemistry is discussed in Chapter 29.

## 1. Solid-Phase Peptide Synthesis

Modern methods of peptide synthesis began with the solid-phase method introduced by Merrifield[299] in 1962 (Fig. 3-15). To begin the synthesis a suitably protected amino acid is covalently linked to a polystyrene bead. The blocking *t*-butoxycarbonyl (Boc) group is removed as isobutene by an elimination reaction to give a bound amino acid with a free amino group. This can then be coupled to a second amino acid with a blocked amino group using dicyclohexylcarbodiimide (Eq. 3-10). The removal of the blocking group and addition of a new amino acid residue can then be repeated as often as desired. The completed peptide is removed from the polystyrene by action of a strong acid such as HF.

Advantages of the Merrifield procedure are that the peptide is held tightly and can be washed thoroughly at each step. Problems arise from repeated use of trifluoroacetic acid and the need to use HF or other strong acid to cleave the peptide from the matrix and also to remove blocking benzyl groups that must be present on many side chain groups. Newer variations of the procedure include a more labile linkage to a polyamide type of polymer and use of blocked amino acids.[297,300–301a] These "active esters" will spontaneously condense with the free amino group of the growing peptide and with suitable catalysis will eliminate



Blocked amino acid

pentafluorophenol. The fluorenylmethoxycarbonyl (Fmoc) blocking group is removed under mildly basic conditions. The whole procedure has been automated in commercially available equipment.

Smaller peptides may be joined to form longer ones.[298] Also useful is enzymatic synthesis. Protein-hydrolyzing enzymes under appropriate conditions will form peptide linkages, for example, joining together oligopeptides.[302,303] Other new methods have been devised to join unprotected peptides.[304,305] Semisynthetic approaches can also be used to place unnatural amino acids into biologically synthesized proteins through the use of suppressor transfer RNAs (Chapter 29).[306]

## 2. Combinatorial Libraries

Many chemists devote all of their efforts to the synthesis of new compounds, including polypeptides, that might be useful as drugs. Traditionally, this has involved the tedious preparation of a large number of compounds of related structure which can be checked individually using various biochemical or biological tests. In recent years a new approach using "combinatorial chemistry" has become very popular and is continually being adapted for new purposes.[307–310] There are several approaches to creating a combinatorial library.

In "split synthesis" procedures a solid-phase

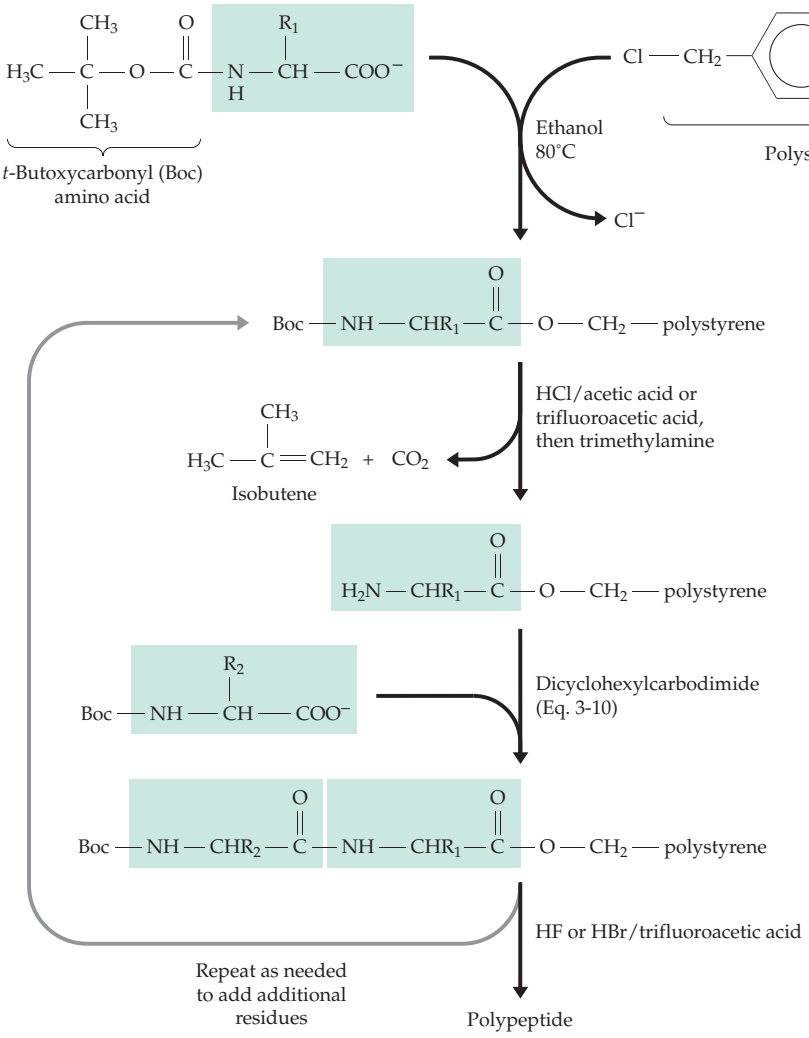



**Figure 3-15** Procedure for solid-phase peptide synthesis devised by Merrifield.[299]

synthesis is conducted on beads. For example, a family of peptides, each with the same C terminus, can be started on a large number of beads. After the first amino acid residue is attached the beads are divided into up to 20 equal portions and different amino acids are added to each portion. The beads can then be mixed and again subdivided. The third residue will again contain many different amino acids attached to each of the different amino acids in the second position. By repeating the procedure again, perhaps for many steps, a "library" of random peptides with each bead carrying a single compound will be formed.

To test whether a polypeptide or other compound carried on a given bead has a derived biological activity, such as the ability to inhibit a certain enzyme, various assays that require only one bead can be devised. However, if a particular bead carries a compound of interest, how can it be identified? The bead carries only a small amount of compound but it may be possible using microsequencing procedures to identify it. An alternative procedure is to use an encoding method to identify the beads.

An alternative to the "one bead–one peptide" approach is to incorporate random sequences of a DNA segment into a gene that can be used to "display" the corresponding peptide sequence. This is illustrated in Fig. 3-16. A protein segment (which may be a random sequence) can be displayed either on the major coat proteins along the shaft or on the minor coat proteins at the end of the bacterial virus fd (see Fig. 7-7). In the case of random insertions, each virus particle may display a different sequence (as many as $10^8$). Peptides may be selected by binding to a desired receptor or monoclonal antibody and the DNA encapsulated in the virus particle can be used to produce more peptides for identification purposes.[311–316] Many other ingenious systems for constructing and testing libraries of peptides[317–320] and other molecules[319,321,322] are being devised. One of these involves a photolithography procedure for immobilizing macromolecules in a regular addressable array, e.g., in a 0.5-mm checkerboard pattern, on a flat surface.[323,324]

## G. Microscopy

The light microscope[325] was developed around 1600 but serious studies of cell structure (histology) did not begin until the 1820s. By 1890, microscope lenses had reached a high state of perfection[326] but the attainable resolution was limited by the wavelength of light. For 450 nm blue light the limit is about 300 nm and for ultraviolet light, viewed indirectly, about 200 nm.[325,327] By the 1940s the electron microscope with its far superior resolving power had overshadowed the light microscope.

For both light and electron microscopy, the preparation of thin sections of cells is a very important technique. Only with very thin sections is the image sufficiently focused. However, **confocal scanning optical microscopy**, invented in the 1950s but not used commercially until much later,[328,329] provided an alternative solution to the focusing problem. A conical beam of light focused to a point is scanned across the sample and the transmitted light (or light emitted by fluorescence) passes through a small "pinhole" aperture located in the primary image plane to a photomultiplier tube where its intensity is recorded. The illuminating beam is moved to scan the entire field sequentially. A series of pinholes in a spinning disk may accomplish the same result. The focal plane can

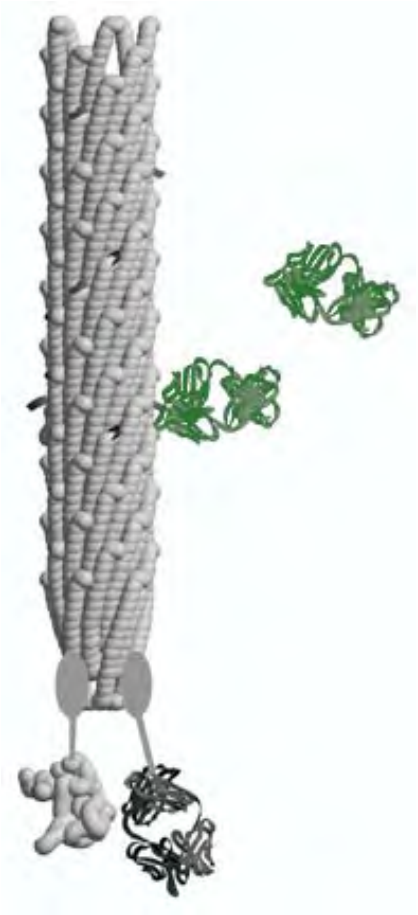


**Figure 3-16** Model of bacteriophage fd engineered to display peptides as inserts in the coat proteins of the virus. The native virus structure is shown in gray; proteins not present in the native virus are shown black or green. Inserted near the N-termini of some major coat proteins is a 6-residue peptide. To one of these peptides a specific Fab antibody fragment (green) has bound from solution, and a second Fab is shown nearby. The N-terminal region of a minor coat protein at the end of the virion has been engineered to display a (different) Fab fragment. Steric constraints are less stringent for inserts in the minor proteins, but fewer copies per virion are possible. Reprinted with permission from Barbas, *et al.*[313a]

be varied so that an image of a thick object such as a cell can be optically sectioned into layers of less than 1 µm thickness. Stereoscopic pairs can also be generated (Fig. 3-17).[330]

A newer development in confocal microscopy is the use of two-photon and three-photon excitation of the fluorescent molecules that occur naturally within cells using short pulses of short-wavelength high-energy laser light.[331] Distribution of such compounds as NADH,[332] DNA, and the neurotransmitter serotonin[333] can be observed without damaging cells. Individual storage granules, each containing ~5 x $10^8$ molecules of serotonin in a concentration of ~50 mM, can be seen.[334] Another new instrument, the **near-field scanning optical microscope** (NSOM),[335,336] is a lensless instrument in which the illuminating beam passes through a very small (e.g., 100 nm diameter) hole in a probe that is scanned in front of the sample. It may extend the limit of optical microscopy to ~1/50 the wavelength of the light.

Since it first became commercially available in 1939, the electron microscope has become one of the most important tools of cell biology.[337,338] The practical resolution is about 0.4 nm, but recent developments in scanning electron microscopy have resulted in resolution of 0.14 nm.[339] Of major importance was the development around 1950 of microtomes and knives capable of cutting thin (20–200 nm) sections of tissues embedded in plastic.[340] A bacterium such as *E. coli* can be sliced into as many as 10 thin longitudinal slices (see Fig. 1-4) and a eukaryotic cell of 10 µm diameter into 100 slices. Serial sections can be examined to determine three-dimensional structures. For some results see Bubel.[341]

If a slice of fresh (frozen) tissue is examined directly, little is seen because most of the atoms found in cells are of low atomic mass and scatter electrons weakly and uniformly. Therefore, thin sections must be "stained' with atoms of high atomic mass, e.g., by treatment with potassium permanganate or osmium tetroxide. Tissues must also be "fixed" to prevent disruption of cell structures during the process of removal of water and embedding in plastic. Fixatives such as formaldehyde react with amino groups and other groups of proteins and nucleic acids. Some proteins are precipitated in place and digestive enzymes that otherwise would destroy much of the fine structure of the cell are inactivated. Glutaraldehyde (a five-carbon dialdehyde) is widely used to fix and crosslink protein molecules in the tissue. The methods continue to be improved.[342]

Small particles, including macromolecules, may be "shadowed." Chromium or platinum can be evaporated in a vacuum from an angle onto the surface of the specimen, Individual DNA molecules can be "seen" in this way.[343] In fact, only the "shadows" are seen and they are 2–3 times wider than the DNA molecules. In the **negative contrast** method a thin layer of a solution containing the molecules to be examined, together with an electron-dense material such as 1% sodium phosphotungstate, is spread on a thin carbon support film. Upon drying, a uniform electron-dense layer is formed. Where the protein molecules lie, the phosphotungstate is excluded, giving an image of the protein molecule.

Surfaces of cells, slices, or intact bacteria can be coated with a deposit of platinum or carbon. The coating, when removed, provides a "negative" **replica** which can be examined in the microscope. Alternatively, a thin plastic replica can be made and can be shadowed to reveal topography. In "freeze fracturing" and "freeze etching," fresh tissue, which may contain glycerol to prevent formation of large ice crystals, is frozen rapidly. Such frozen cells can often be revived; hence, they may be regarded as still alive until the moment that they are sliced! The frozen tissue is placed in a vacuum chamber within which it is sliced or fractured with a cold knife. If desired, the sample can be kept in the vacuum chamber at about –100°C for a short time, during which some water molecules evaporate from the surface. The resultant etching reveals a fine structure of cell organelles and membranes in sharp relief. After etching, a suitable replica is made and examined (Fig. 1-15A and E). Fracturing tends to take place through lipid portions of cell membranes.

Small viruses, bacterial flagella, ribosomes, and even molecules can be seen by electron microscopy. However, to obtain a clear image in three dimensions requires a computer-based technique of **image reconstruction** or **electron microscope tomography**, which was developed initially by Aaron Klug and associates.[344–349] A sample is mounted on a goniometer, a device that allows an object to be tilted at exact angles. Electron

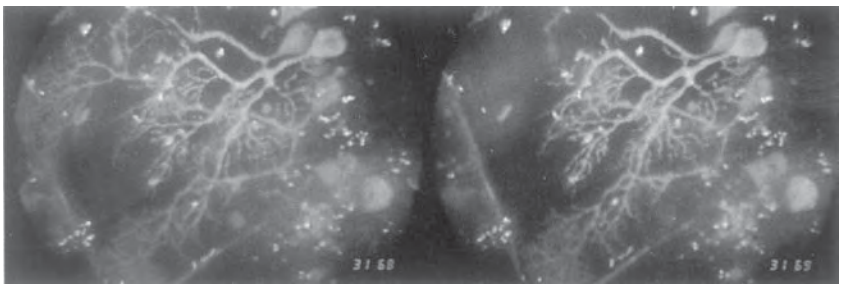


**Figure 3-17** Confocal micrograph showing a forty-micrometer stereo slice in a 90-µm thick section of mouse cerebellum.[330] Courtesy of A. Boyde.

micrographs are prepared with the sample untilted and tilted in several directions at various angles, e.g., up to 90° in 10° increments. The micrographs are digitized and a computer is used to reconstruct a three-dimensional image.

In **electron crystallography**[350] micrographs of two-dimensional crystalline arrays of molecules or larger particles are prepared. A Fourier transform of the micrograph gives a diffraction pattern which can be treated in a manner similar to that usual for X-ray diffraction to give a three-dimensional image. An important milestone in use of this technique was the determination of the structure of bacteriorhodopsin (Fig. 23-45) at 0.3-nm resolution. Bear in mind that X-ray crystallography can also be viewed as a form of microscopy.

Invention of the **scanning tunneling microscope** (STM) by Binnig and Roher[351 – 353] initiated a new revolution in microscopy. The STM and similar scanned probe microscopes examine surfaces by moving a fine probe mounted in a piezoelectric x,y,z-scanner[354] across the surface to be examined. The tiny tungsten probe of the STM is so fine that its tip may consist of a single atom. When a small voltage is applied a minuscule quantum mechanical tunnelling current flows across the small gap between the probe and the surface and a high-resolution image, sometimes at atomic resolution, is created from the recorded variation in current.[354-359] The STM theoretically responds only to surfaces that conduct electrons, but nonconducting samples have been imaged at high humidity; presumably by conductance of electrons or ions through the surface water layer. The success of the STM spurred the development of many other types of scanned probe microscopes. Among these, the **atomic force microscope** (AFM; Fig. 3-18) has been especially useful for biological materials, including proteins and nucleic acids. The AFM moves a fine-tipped stylus directly across the sample surface or, alternatively, vibrates the probe above the surface. The small up-and-down movements of the stylus are recorded[359–366] and thereby create a topographic or force-field map of the sample. AFM images contain three-dimensional information and can be used to view individual molecules (Fig. 3-19).[367,367a] **Chemical force microscopy** is sensitive to adhesion and friction as a function of the interaction between defined chemical groups on the tip and sample.[368]

An emerging field is force spectroscopy, in which the AFM measures interaction forces between and within individual molecules.[369 – 371a] Under development are NMR microscopes (Section I). There is continual effort to see small objects more directly and more clearly!

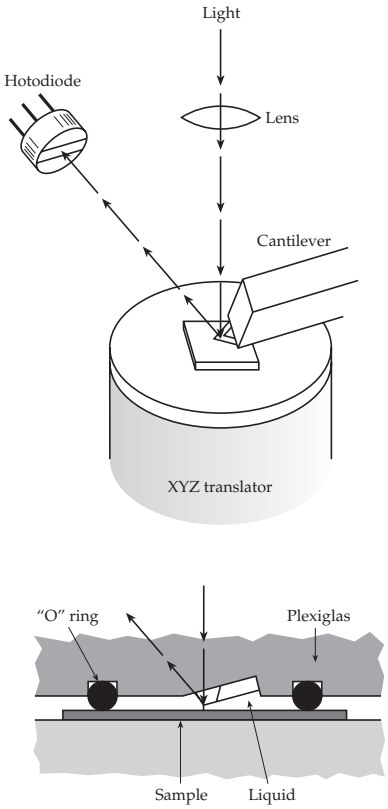



**Figure 3-18** Schematic diagram of the atomic force microscope.[360] Courtesy of Paul Hansma.

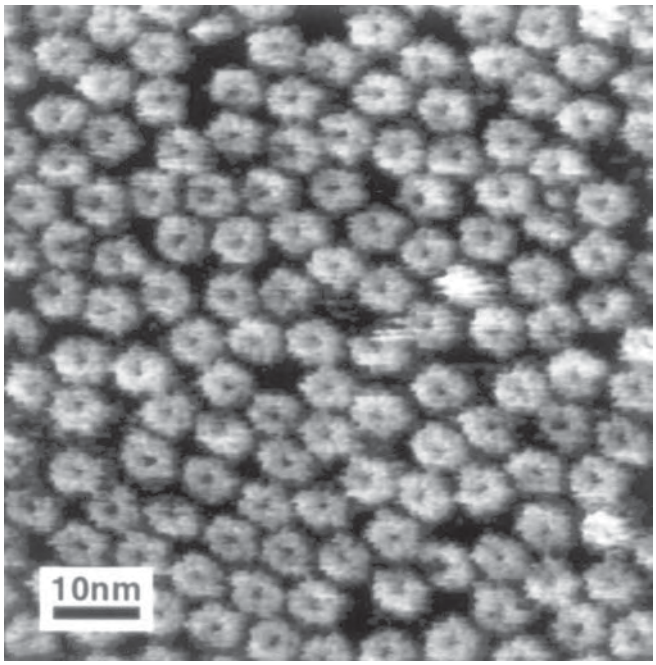



**Figure 3-19** AMF images of cholera toxin[364] (see also Box 11-A). Courtesy of Z. Shao.

## H. X-ray and Neutron Diffraction

One of the most important techniques by which we have learned bond lengths and angles and precise structures of small molecules is X-ray diffraction. Today, this technique, which involves measurements of the scattering of X-rays by crystalline arrays of molecules, is being used with spectacular success to study macromolecules of biochemical and medical importance.[372–378]

X-rays were described by Röntgen in 1896 but there was uncertainty as to their wave nature.[274] It was not until 1912 that the wavelengths of X-rays had been measured and it was recognized that they were appropriate for the use of X-rays in structure determination. Consider the fact that with a conventional light microscope we cannot distinguish two small objects that are much closer together than the distance represented by the wavelength of the light with which we observe them. This is about 460 nm for blue light. By comparison the wavelength of the copper $K_\alpha$ radiation, which is used in protein crystallography, is 0.1542 nm entirely appropriate for seeing the individual atoms of which matter is composed. Recognizing this, W. L. Bragg in Cambridge, England, in 1913 used X-ray diffraction to establish the structures of NaCl, KCl, and KBr in the crystalline state. The science of X-ray crystallography had been founded.[379]

In 1926, James Summer crystallized the enzyme urease (Chapter 16) and crystallization of other enzymes soon followed.[380] In 1934, J. B. Bernal brought back to Cambridge from Uppsala, Sweden, some crystals of pepsin almost 2 mm long that had been grown in T. Svedberg's laboratory. Bernal and Crowfoot showed that these delicate crystals, which contained almost 50% water, gave a sharp diffraction pattern when they were protected by enclosure in a narrow capillary tube containing some of the mother liquor from which the crystals had been grown.[381,382] After this, diffraction patterns were obtained for many protein crystals, and in 1937 Max Perutz chose for his thesis work at Cambridge the X-ray crystallography of hemoglobin. The project seemed hopeless at times, but in 1968, 31 years later, Perutz had determined the structure of hemoglobin.[383–386]

If X-rays could be focused easily, could one build an X-ray microscope that would permit the immediate viewing of molecular structures? X-ray holograms at the molecular level have been obtained.[387,388] However, currently the only practical X-ray microscope for protein structures involves the measurement of diffraction patterns created by the scattering of X-rays (or of neutrons) from the crystalline lattice. The details of this procedure can be found in other sources.[372–375,378,389–390] It is sufficient to point out here that a pattern of many

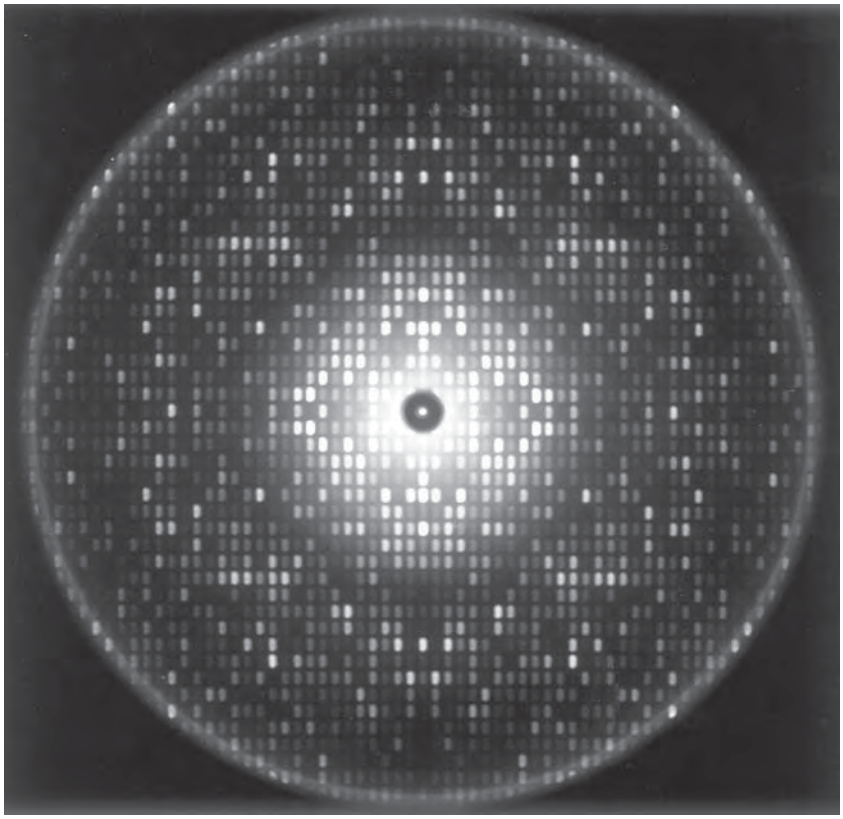


**Figure 3-20** An X-ray diffraction photograph such as was used to determine the structure of hemoglobin. This precession photograph was obtained from a crystal of human deoxyhemoglobin by rotating the crystal along two different axes in a defined manner before a narrow X-ray beam. The film also was moved synchronously. The periodicity observed in the photograph is a result of a diffraction phenomenon arising from the periodic arrangement of atoms in the crystal. The distances of the spots from the origin (center) are inversely related to the distances between planes of atoms in the crystal. In this photograph (which shows only two dimensions of the three-dimensional diffraction pattern) the spots at the periphery represent a spacing of 0.28 nm. If the intensities of the spots are measured, and if the phases of the harmonic functions required for the Fourier synthesis can be assigned correctly, the structure can be deduced to a resolution of 0.28 nm from a set of patterns of this type. In the case of human deoxyhemoglobin, a complete set of data would consist of about 16,000 spots. With modern equipment the resolution limit can be extended to better than 0.15 nm (or more than 100,000 unique spots). Photograph courtesy of Arthur Arnone.

spots, such as that in Fig. 3-20, is obtained. In a typical determination of a protein structure from 10,000 to several times that number of spots must be measured. The needed information is contained in the coordinates of the spots (i.e., in the angles through which the scattering occurs) and in the intensities of the spots. The structure is obtained by a Fourier synthesis in which a large number of mathematical functions in the form of three-dimensional sinusoidal waves are summed. The wavelengths of these functions are determined by the positions of the spots in the diffraction pattern. The amplitudes of the waves are related to the intensities of the spots.[374] However, the waves are not all in phase and the phases must be learned in some other way. This presents a difficult problem. Mathematical methods have been devised that automatically determine the phases for small molecules and usually allow the structures to be established quickly. For proteins it is more difficult.

In 1954, Perutz introduced the **isomorphous replacement method** for determining phases. In this procedure a heavy metal, such as mercury or platinum, is introduced at one or more locations in the protein molecule. A favorite procedure is to use mercury derivatives that combine with SH groups. The resulting heavy metal-containing crystals must be isomorphous with the native, i.e., the molecules must be packed the same and the dimensions of the crystal lattice must be the same. However, the presence of the heavy metal alters the intensities of the spots in the diffraction pattern and from these changes in intensity the phases can be determined. Besides the solution to the phase problem, another development that was absolutely essential was the construction of large and fast computers. It would have been impossible for Perutz to determine the structure of hemoglobin in 1937, even if he had already known how to use heavy metals to determine phases.

The first protein structure to be learned was that of myoglobin, which was established by Kendrew *et al*. in 1960.[391–393] That of the enzyme lysozyme was deduced by Blake *et al*. in 1965.[394] Since then, new structures have appeared at an accelerating rate so that today we know the detailed architecture of over 6000 different proteins[395] with about 300 distinctly different folding patterns.[396] New structures are being determined at the rate of about one per day. X-ray diffraction has also been very important to the study of naturally or artifically oriented fibrous proteins[397] and provided the first experimental indications of the β structure of proteins.

Suppose that you have isolated a new protein. How can you learn its three-dimensional structure? The first step is crystallization of the protein, something that a biochemist may be able to do. Crystallization is done in many ways, often by the slow diffusion of one solution into another or by the slow removal of solvent through controlled evaporation.[374,398] Ammonium sulfate and polyethylene glycol are two commonly used precipitants. The presence of the neutral detergent β-octyl glucoside improves some crystals.[399] Droplets of protein solution mixed with the precipitant are often suspended on microscope cover glasses in small transparent wells or are placed in depression plates within closed plastic boxes.[374,400,401] In either case, a reservoir of a solution with a higher concentration of precipitant is present in the same compartment. Water evaporates from the samples into the larger reservoir, concentrating the protein and causing its crystallization. Crystals grow slightly better in a spacecraft than on Earth.[402,403] Some proteins, notably myosin from muscle, crystallize well only after reductive methylation of all lysine side chains to dimethyllysine with formaldehyde and sodium borohydride (Eq. 3-34).[278]

The next step is for a protein crystallographer to mount a small perfect crystal in a closed silica capillary tube and to use an X-ray camera to record diffraction patterns such as that in Fig. 3-20. These patterns indicate how perfectly the crystal is formed and how well it diffracts X-rays. The patterns are also used to calculate the dimensions of the unit cell and to assign the crystal to one of the seven **crystal systems** and one of the 65 enantiomorphic **space groups**. This provides important information about the relationship of one molecule to another within the unit cell of the crystal. The unit cell (Fig. 3-21) is a parallelopiped

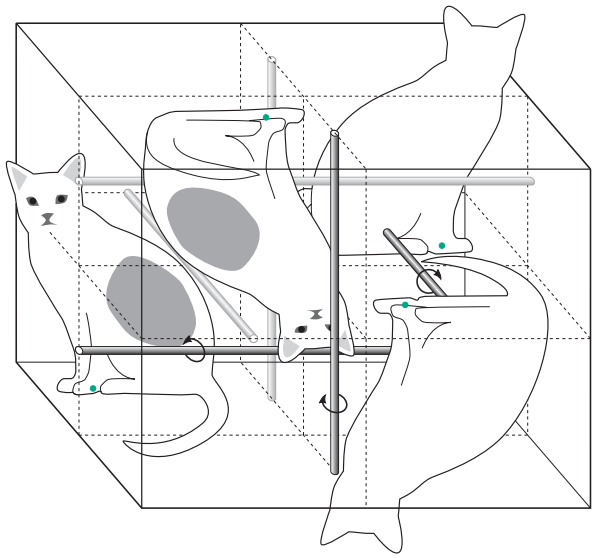


**Figure 3-21** Diagram showing an asymmetric unit as it might appear in a unit cell of space group $P2_12_12_1$. This unit cell has three pairs of nonintersecting twofold **screw axes** which are marked by the shaded rods. These are designated by arrows and the symbol. Two asymmetric units are related one to another by rotation around a twofold axis together with translation by one-half the dimension of the unit cell.

whose sides are parallel to crystallographic axes and which, by translation in three dimensions, gives rise to the entire crystal. The unit cell must contain at least one **asymmetric unit**, the smallest unit of structure that lacks any element of crystal symmetry. The asymmetric unit may consist of one or a small number of protein subunits.

In crystals of the triclinic system all of the asymmetric units are aligned in the same manner and there are no axes of symmetry. Monoclinic crystals have a single **twofold** or **dyad axis** (see Fig. 7-11). The unit cell might contain two molecules related one to another by the dyad or one dimeric molecule with the dyad located as in Fig. 7-11C. Orthorhombic crystals have three mutually perpendicular twofold axes and the unit cell is a rectangular solid. Trigonal, tetragonal, and hexagonal crystal systems have three, four and sixfold axes of symmetry, respectively, while the cubic crystal contains four threefold axes along with diagonals of the cube as well as two-fold axes passing through the faces (see Fig. 15-14). Within each crystal system there are several space groups. An example is the orthorhombic space group $P2_12_12_1$, which is often met with small organic molecules and proteins. In this space group the unit cell contains three mutually perpendicular but nonintersecting twofold screw axes (Fig. 3-21). The position of one molecule in the unit cell is related to the next by both a 180° degree rotation about the twofold axes and a translation of one-half the length of the unit cell. There is such a screw axis for each of the three directions.[372]

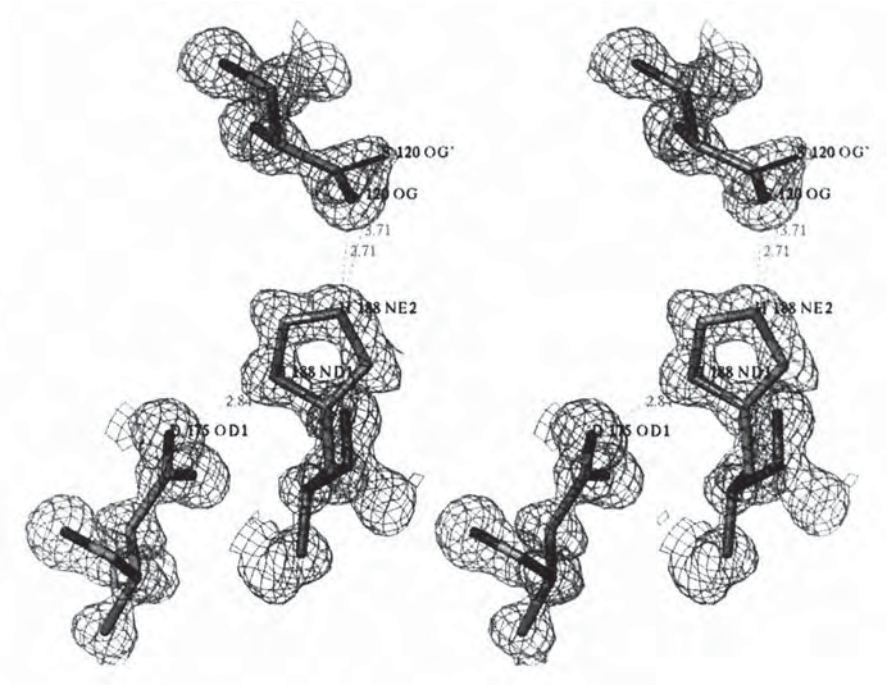



**Figure 3-22** Stereoscopic view of a section of the structure of cutinase from the fungus *Fusarium solani* determined to a resolution of 0.10 nm. The three amino acid residues shown are serine 120 (top), histidine 188, and aspartate 175 (lower left). The structure is presented as a contour map with a "wire mesh" drawn at a "cutoff" level of density equal to 1σ above the average, where σ is the root mean square density of the entire map. The side chains of these three residues constitute the "catalytic triad" in the active site of this enzyme (see Chapter 12). At this resolution more than one conformation of a group may often be seen. For example, the gamma oxygen (OG) of S120 is seen in two positions, the major one being toward His 188. When the map is drawn with a lower contour level the N-H proton on His 188 that is hydrogen bonded to Asp 175 can also be seen.[410] Courtesy of Christian Cambillau.

The third step in the structure determination is collection of the X-ray diffraction data. This may be done with a **diffractometer** in which a narrow collimated pencil source of X-rays is aimed at the crystal and the intensities and positions of the diffracted beams are measured automatically. The computer-controlled diffractometer is able to measure the angles to within less than one-hundredth of a degree. If sufficient time is allowed, very weak spots can be counted. Today, diffractometers are more likely to be used for preliminary measurements, while the major data collection is done with an **area detector**, an instrument that collects many reflections simultaneously. There are some difficulties. For example, during the long periods of irradiation needed to measure the weak spots with a diffractometer the protein crystals decompose and must be replaced frequently. Data collection may last for months. The newest methods utilize more powerful X-ray sources and often **synchrotron radiation**, which delivers very short and extremely intense pulses of X-rays and allows data to be collected on very small well-formed crystals.[404]

The fourth step is the preparation of isomorphous crystals of heavy metal-containing derivatives. The heavy metal may be allowed to react with the protein before crystallization or may be diffused into preformed crystals. A variety of both cationic and anionic metal complexes, even large $Ta_6Br_{12}^{2+}$ tantalum clusters, have been used.[405] Two or more different heavy metal derivatives are often required for calculation of the phases. The heavy metal atoms must be present at only a very small number of locations in the unit cell.

An entire data set must be collected for each of these derivatives. The evaluation of the phases from these data is a complex mathematical process which usually involves the calculation first of a "difference Patterson projection."[406] This is derived by Fourier transformation of the differences between the scattering intensities from the native and heavy atom-containing crystals. The Patterson map is used to locate the coordinates of the heavy metal atoms which are then refined and used to compute the phases for the native protein.

Alternative methods of solving the phase problem are also used now. When a transition metal such as Fe, Co, or Ni is present in the protein, anomolous scattering of X-rays at several wavelengths (from synchrotron radiation) can be used to obtain phases. Many protein structures have been obtained using this multiple wavelength anomalous diffraction (MAD phasing) method.[404,407,408] Selenocysteine is often incorporated into a protein that may be produced in bacteria using recombinant DNA procedures. Crystals are prepared both with protein enriched in Se and without enrichment. In some ways it is better to incorporate tellurium ($^{127}$Te) in telluromethionine.[409]

In the fifth step of an X-ray structure determination the **electron density map** is calculated using the intensities and phase information. This map can be thought of as a true three-dimensional image of the molecule revealed by the X-ray microscope. It is usually displayed as a stereoscopic view on a computer graphics system (Fig. 3-22). It is also often prepared in the form of a series of transparencies mounted on plastic sheets. Each sheet represents a layer, perhaps 0.1 nm thick, with contour lines representing different levels of electron density.

Using the electron density map a three-dimensional model of the protein can be built. For years, the customary procedure was to construct a model at a scale of 2 cm = 0.1 nm using an **optical comparator**, but

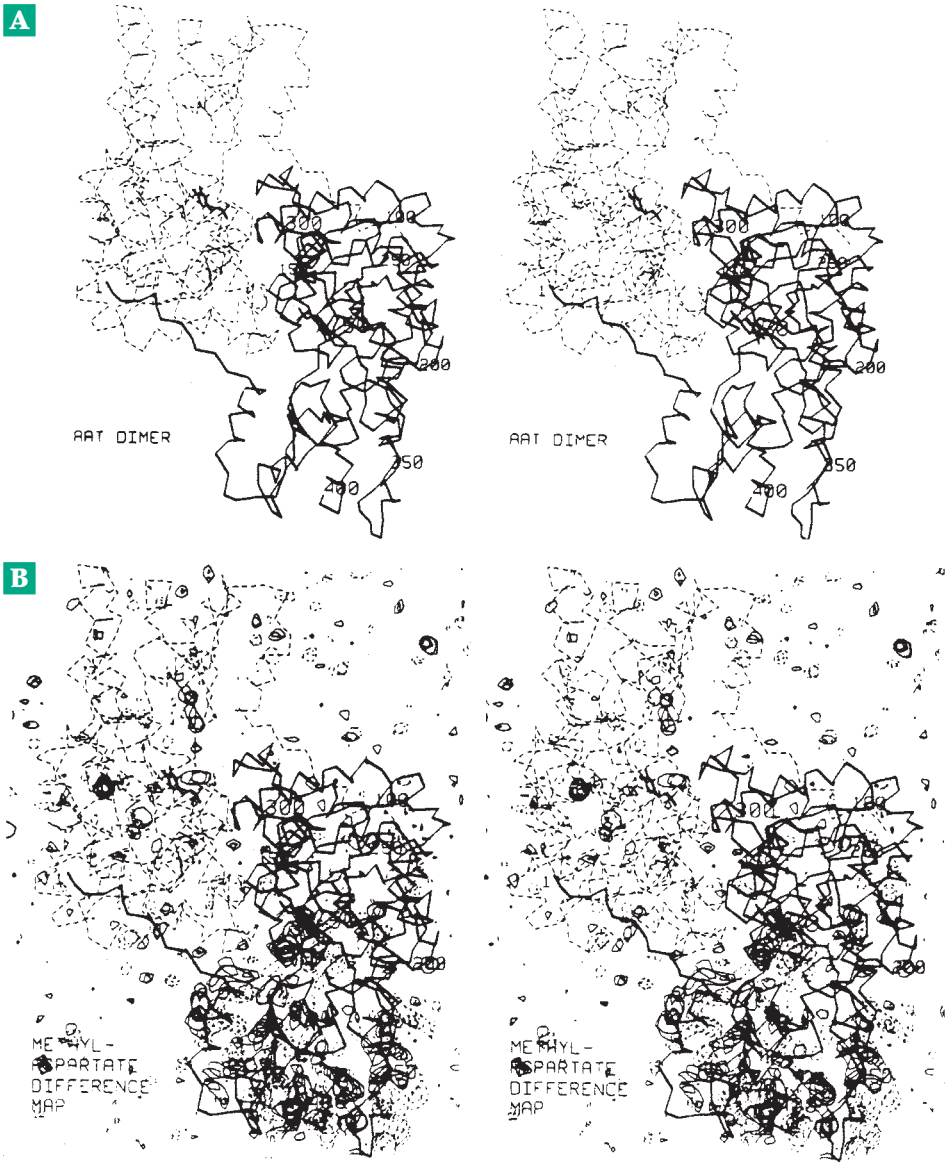



**Figure 3-23** (A) Stereoscopic α-carbon plot of the cystolic aspartate aminotransferase dimer viewed down its dyad symmetry axis. Bold lines are used for one subunit (subunit 1) and dashed lines for subunit 2. The coenzyme pyridoxal 5'-phosphate (Fig. 3-24) is seen most clearly in subunit 2 (center left). (B) Thirteen sections, spaced 0.1 nm apart, of the 2-methylaspartate difference electron density map superimposed on the α-carbon plot shown in (A). The map is contoured in increments of ± 2σ (the zero level omitted), where σ = root mean square density of the entire difference map. Positive difference density is shown as solid contours and negative difference density as dashed contours. The alternating series of negative and positive difference density features in the small domain of subunit 1 (lower right) show that the binding of L-2-methylaspartate between the two domains of this subunit induces a right-to-left movement of the small domain. (Continues)

crystallographers now use computer graphics systems. The three-dimensional image of the electron density map and a computer-generated atomic model are superimposed on the computer screen. An example is shown in Fig. 3-22. When the superposition has been completed, the coordinates of all atoms are present in the memory of the computer.

The final step in the structure determination is **refinement** using various mathematical methods. From the coordinates of the atoms in the model the expected diffraction pattern is computed and is compared with that actually observed. The differences between predicted and observed density are squared and summed. The sum of the squares constitutes an error function which is then minimized by moving the various atoms in the model short distances while keeping bond lengths, angles, and van der Waals distances within acceptable limits and recalculating the error function. This complex "refinement" procedure must be repeated literally hundreds of thousands of times with every part of the structure being varied. Structures at very high resolution (0.07–0.1 nm) may reveal multiple positions for hydrogen-bonded side chains[410a,b] as well as hydrogen atoms (Fig. 3-22) and even bonding electrons.[410b]

Once the three-dimensional structure of the protein is known, further experiments are usually done using the X-ray diffraction technique. Since protein crystals contain channels of solvent between the packed molecules, it is usually possible to diffuse small molecules into the crystal. These may be substrates, inhibitors, or allosteric effectors. Diffraction data are collected after diffusion of the small molecules into the crystal and a **difference electron density map** may be calculated. This may show exactly where those molecules were bound. An example is shown in Fig. 3-23. Sometimes difference maps not only show the binding to an enzyme of a substrate or other small molecule but also reveal conformational changes in proteins. Such is the case for Fig. 3-23, which shows the binding of α-methylaspartate, an inhibitor that behaves initially like a substrate and goes part way through the reaction sequence for aspartate aminotransferase until further reaction is blocked by the methyl group. This difference map also shows that the part of the protein to the left of the binding site in the figure has moved.[411] Using the X-ray data it was possible subsequently to deduce the nature of the conformational change.

Examination of the effect of temperature on the diffraction pattern of a protein can give direct information about the mobility of different parts of the mole-

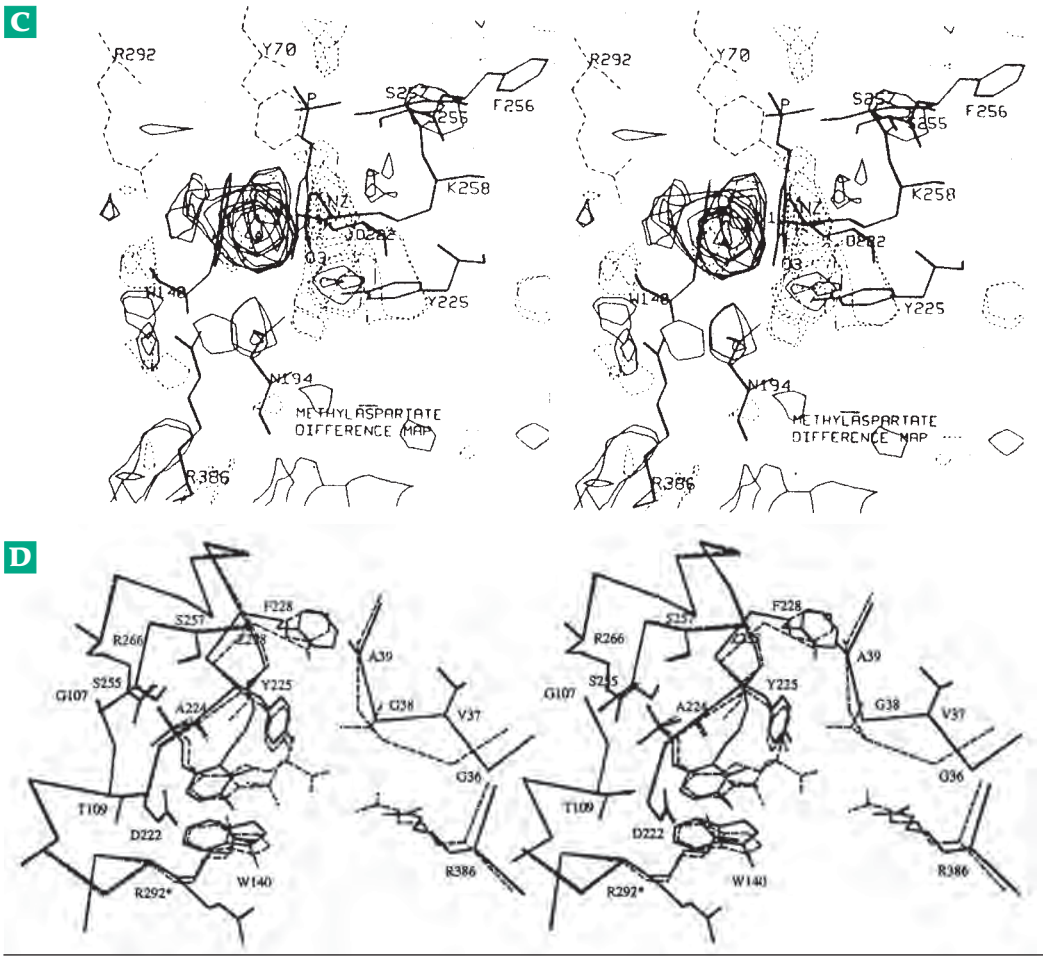
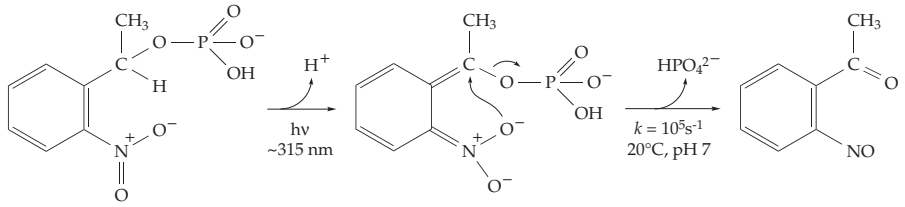



**Figure 3-23** (Cont.) (C) Nine sections, spaced 0.1 nm apart, of a part of the 2-methylaspartate difference map superimposed on the atomic model shown in (A) and (D). The coenzyme is shown as the internal aldimine with Lys 258 (see Fig. 14-6, 14-10). The positive and negative contours on the two sides of the coenzyme ring indicate that the coenzyme tilts over to form the external aldimine when substrates react.[413] (D) Superimposed structure of the active site of the enzyme in its free form as in (A) (bold lines) and the refined structure of the α-methylaspartate complex, (dashed lines).[411] This illustrates the tilting of the coenzyme ring, which is also shown in Eq. 14-39 and Fig. 14-10. Courtesy of Arthur Arnone and Sangkee Rhee.

cules.[412] This complements information obtainable in solution from NMR spectra.

Because hydrogen atoms contain only one electron, and therefore scatter X-rays very weakly, they are usually not seen at all in X-ray structures of proteins. However, neutrons are scattered strongly by hydrogen atoms and **neutron diffraction** is a useful tool in protein structure determination.[414,415] It has been used to locate tightly bonded protons that do not exchange with $^2H_2O$ as well as bound water ($^2H_2O$).

The development of synchrotron radiation as an X-ray source[404,416–418] has permitted accumulation of data for electron density difference maps in less than 1 s and it is expected that such data can eventually be acquired in ~1 ps.[419–421] If a suitable photochemical reaction can be initiated by a picosecond laser flash, a substrate within a crystalline enzyme can be watched as it goes through its catalytic cycle. An example is the release of inorganic phosphate ions from a "caged phosphate" (Eq. 3-49) and study of the reaction of the released phosphate with glycogen phosphorylase (Chapter 12).[422,423]




"Caged phosphate" (3-49)

However, caged substrates usually must diffuse some distance before reacting, so very rapid events cannot be studied. An alternative approach is to diffuse substrates into crystals at a low temperature at which reaction is extremely slow but a substrate may become seated in an active site ready to react. In favorable cases such "frozen" Michaelis complexes may be heated by a short laser pulse to a temperature at which the reaction is faster and the steps in the reaction may be observed by X-ray diffraction.[424,425]

## I. Nuclear Magnetic Resonance (NMR)

Organic chemists and biochemists alike have long relied on NMR spectroscopy to assist in identification and determination of structures of small compounds. Most students have some familiarity with this technique and practical information is available in many places.[426–428] Measurements can be done on solids, liquids, or gases but are most often done on solutions held in special narrow NMR tubes. Volumes of samples are typically 0.5 ml (for a 1–5-mM solution containing 1–5 µmol of protein) but less for small molecules and higher concentrations. Newer techniques allow use of a volume as small as 5 nl and containing <0.1 nmol of sample.[429] For many years progress in biochemical application of NMR was slow, but a dramatic increase in the power of the spectrometers, driven in great measure by the revolution in computer technology, has made NMR spectroscopy a major force in the determination of structures and functions of proteins and nucleic acids.[430,431]

### 1. Basic Principles of NMR Spectroscopy

The basis of NMR spectroscopy lies in the absorption of electromagnetic radiation at radiofrequencies by atomic nuclei.[426,427,432–437] All nuclei with odd mass numbers (e.g., $^1H$, $^{13}C$, $^{15}N$, $^{17}O$, $^{19}F$, and $^{31}P$), as well as those with an even mass number but an odd atomic number, have magnetic properties. Absorption of a quantum of energy $E = h\nu$ occurs only when the nuclei are in the strong magnetic field of the NMR spectrometer and when the frequency ν of the applied electromagnetic radiation is appropriate for "resonance" with the nucleus being observed. In the widely used "500-megahertz" NMR spectrometers the liquid helium-cooled superconducting electromagnet has a field strength of 11.75 tesla (T). In this field a proton resonates at ~500 megahertz (MHz) and nuclei of $^{31}P$, $^{13}C$, and $^{15}N$ at ~202, 125, and 50 MHz, respectively. At 500 MHz the energy of a quantum is only $E = 3.3 \times 10^{-33} \times 10^8$ J = 0.2 J mol$^{-1}$, more than four orders of magnitude less than the average energy of thermal motion of molecules (3.7 kJ mol$^{-1}$). Thus, the spin transitions induced in the NMR spectrometer have no significant effect on the chemical properties of molecules.

The resonance frequency ν at which absorption occurs in the spectrometer is given by Eq. 3-50, where $H_o$ is the strength of the external magnetic field, $\mu$ is the magnetic moment of the nucleus being investigated, and $h$ is Planck's constant. The basis for NMR spectroscopy lies in the fact that nuclei in different positions in a molecule resonates at slightly different frequencies. In a protein each one of the hundreds or thousands of protons resonate at its own frequency. With older NMR instruments a spectrum at the constant magnetic field $H_o$ can be obtained by varying the frequency and observing the values at which absorption occurs, much as is done for ultraviolet, visible, and infrared spectra (Chapter 23). With newer pulsed NMR spectrometers the measurement is done differently but the spectra

$$\nu = \mu H_o / h \qquad (3\text{-}50)$$

look the same. The higher the magnetic field, the greater the variation in resonance frequency and the higher the sensitivity. The most powerful commercial NMR spectrometers currently available operate at about 750 MHz for $^1H$ and a few higher frequency instruments have been built.

The **proton NMR** spectrum of the coenzyme pyridoxal phosphate in $^2H_2O$ is shown in Fig. 3-24 as obtained with a 60-MHz spectrometer. Four things can be measured from such a spectrum: (1) the **intensity** (area under the band). In a proton NMR spectrum, areas are usually proportional to the numbers of equivalent protons giving rise to absorption bands; (2) the **chemical shift**, the difference in frequency between the peak observed for a given proton and a peak of some standard reference compound. In Fig. 3-24 the reference peak is at the right edge; (3) the **width** at half-height (in hertz), a quantity that can provide information about molecular motion and about chemical exchange; and (4) **coupling constants** which measure interactions between nearby magnetic nuclei. These are extremely important to the determination of structures of both small and large molecules.

With a magnetic field of $H_o$ = 11.75 T and a 500-MHz oscillator, the positions of proton resonances in organic compounds are spread over a range of ~10,000 Hz. This is 20 parts per million (ppm) relative to 500 MHz. Positions of individual resonances are usually given in ppm and are always measured in terms of a shift from the resonance position of some standard substance. For protons this is most often **tetramethylsilane (TMS)**, an inert substance that can be added directly to the sample in its glass tube. Biochemists often use $^2H_2O$ as solvent and the water-soluble sodium 3-trimethylsilyl 1-propane sulfonate (**DSS** or Tier's salt) as a standard. Its position is insignificantly different from that of TMS. For NMR spectra measured in $^2H_2O$, the "pD" of the medium is sometimes indicated. It has often been taken as the pH meter reading plus 0.4. However, because of uncertainty about the meaning of pD, most workers cite the apparent pH measured with a glass electrode and standardized against aqueous buffers.[438] It is important to describe how the measurement was made when publishing results.

***Band widths.*** The narrowness of a band in an NMR spectrum is limited by the **Heisenberg uncertainty principle**, which states that $\Delta E \times \Delta t = h/2\pi$, where $h$ is Planck's constant, $\Delta E$ is the uncertainty in the energy, and $\Delta t$ is the lifetime of the magnetically excited state. Since $E = h\nu$ for electromagnetic radiation, $E$ is directly proportional to the width of the absorption band (customarily measured at one-half its full height). The magnetic nucleus is well shielded from external influences and the lifetime of its excited state tends to be long. Hence, $\Delta\nu$ is small, often amounting to less than 0.2 Hz. This fact is very favorable for the success of high-resolution proton magnetic resonance. However, bands are often much broader for large macromolecules.

***The chemical shift.*** In a molecule such as TMS, the electrons surrounding the nuclei "shield" the nucleus so that it does not experience the full external magnetic field. For this reason, absorption occurs at a high frequency (high energy). Protons that are bound

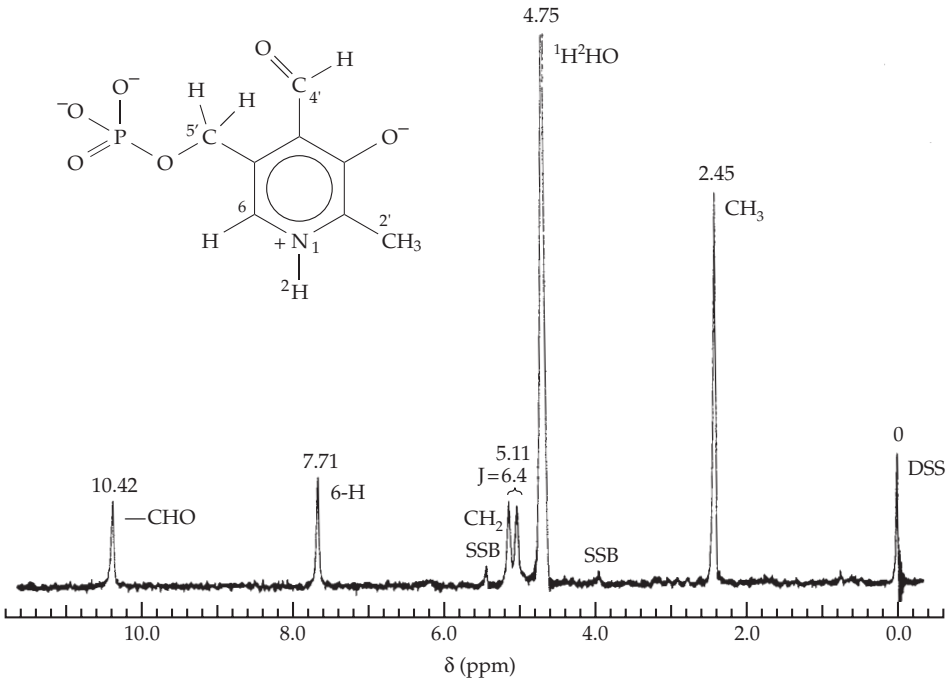



**Figure 3-24** The 60-MHz proton magnetic resonance spectrum of pyridoxal 5'-phosphate at neutral pH (apparent pH = 6.65). The internal standard is DSS. Chemical shifts in parts per million are indicated beside the peaks. Spectrum courtesy of John Likos.

to an atom deficient in electrons (because of attachment to electron withdrawing atoms or groups) are **deshielded**. The greater the deshielding, the further **downfield** from the TMS position is the NMR peak.

The magnitude of this chemical shift may be stated in hertz, but it is most often expressed in ppm as δ (Eq. 3-51). The value of δ is the shift in frequency relative to frequency of the oscillator in parts per million and is independent of the field strength. It still depends upon use of a particular reference standard which must be stated when a δ value is given. In the spectrum shown in Fig. 3-24, the methyl protons appear 2.45 ppm below the DSS peak but still at a relatively high field. Characteristic chemical shift ranges for other protons (Table 3-3) extend to ~20 ppm.

$$\delta\,(\text{ppm}) = \frac{\Delta\nu\,(\text{Hz}) \times 10^6}{\nu\,(\text{Hz})\text{ of oscillator}} \qquad (3\text{-}51)$$

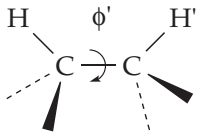
Aromatic rings lead to strong deshielding of attached protons because of a **ring current** induced in the circulating π electrons. Thus, in Fig. 3-24 the peaks of the methylene protons which are adjacent to the aromatic ring occur at 5.10 and 5.12 ppm. The 6-H, which is bound directly to the ring, is more strongly deshielded and appears at 7.71 ppm. The hydrogen of the aldehyde groups is deshielded as a result of a similar "diamagnetic electronic circulation" in the carbonyl group. Its peak is even further downfield at 10.4 ppm. Ring current and other effects on chemical shifts are important in NMR spectroscopy of proteins. Aromatic proton resonances sometimes stand out because they have been shifted far downfield. Ring current effects on chemical shifts can be predicted quite accurately if three-dimensional structures are known.[439] Additionally, computer programs are available for predicting them.[440–443]

Hydrogen bonding has a very large effect on the chemical shift of protons. The resonance of a strongly hydrogen-bonded proton is usually shifted downfield from its position in non-hydrogen-bonding media. This is especially true for hydrogen bonds to charged groups, e.g., the NH of a histidine or tryptophan side chain hydrogen bonded to a carboxylate group, a situation often met in the active sites of proteins. The chemical shift of $^{13}C$ in the carboxyl group is also affected.[444,445] While **ring current shifts** can be predicted quite well, it is much more difficult to predict the total chemical shift.[446,447]

***Scalar coupling (J coupling).*** The energy of the spin transition of a hydrogen nucleus is strongly influenced by the local presence of other magnetic nuclei, e.g., other protons that are covalently attached to the same or an adjacent atom. These neighboring protons can be in either of the two spin states, a fact that results in easily measured differences in the energy of the NMR transition under consideration. This spin–spin interaction (coupling) leads to a splitting of NMR bands of protons into two or more closely spaced bands. The ethyl group often appears in NMR spectra as a "quartet" of four evenly spaced peaks that arise from the $CH_2$ group and a triplet of peaks arising from the $CH_3$ protons. Protons attached to the same carbon (**geminal** protons), and in similar environments, do not ordinarily split each other's peaks, while the protons on the neighboring carbon do.

The **coupling constant** *J* is the difference in hertz between the successive peaks in a multiplet. It is a field-independent quantity and the same no matter what the frequency of the spectrometer. In Fig. 3-24 the peak of the methylene protons is split by $^1H–^{31}P$ coupling, with a value of $J \sim 6.4$ Hz. While spin–spin coupling is most pronounced when magnetic nuclei are close together in a structure, the effect can sometimes be transmitted through up to five covalent bonds. The technique of **double irradiation** or **spin decoupling** can be used to detect spin coupling. The sample is irradiated at the resonance frequency of one of the nuclei involved in the coupling, while the spectrum is observed in the frequency region of the other nucleus of the coupled pair. Under these conditions the multiplet collapses into a singlet and the mutual coupling of the two nuclei is established. The coupling can be seen directly in appropriate two-dimensional NMR spectra.

The coupling constant between two **vicinal protons** which are attached to adjacent carbon atoms (or other atoms) depends upon the torsion angle.



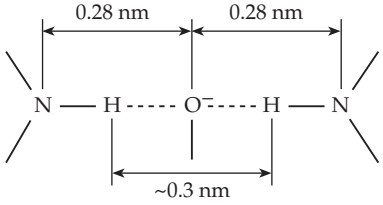

The **Karplus equation** (Eq. 3-52) relates *J* to the torsion angle ϕ′ (so labeled to distinguish it from peptide torsion angle ϕ; Fig. 2-8).

$$J_{H,H'} \quad A\ \cos^2\phi' + B\cos\phi' + C \qquad (3\text{-}52)$$

This equation was predicted on theoretical grounds, but the constants, *A*, *B*, and *C* are empirical.[450] Other forms of the equation, some of them simplified, have also been proposed.[451] The Karplus relationship is often used to estimate time-averaged torsion angles in peptides. For a $C_\alpha H–NH$ torsion angle the parameters *A*, *B* and *C* of Eq. 3-52 are ~ 6.4, 1.4, and 1.9, respectively. For a $C_\alpha H - C_\beta H$ they are ~ 9.5, 1.6, and

1.8, respectively.[452] Coupling constants between $^{1}H$ and $^{13}C$ or $^{15}N$ are of importance in determination of three-dimensional structure of proteins.[453]

**The nuclear Overhauser effect (NOE)** is the result of transfer of magnetization from one nucleus to a nearby nucleus directly through space rather than via *J*-coupling.[427,454] This was observed first as a result of irradiation of a resonance in a one-dimensional spectrum resulting in an increased intensity of the resonance of the nearby nucleus. Magnetization transfer can occur from a given nucleus to one or more nearby nuclei. Each such transfer that is detected is usually referred to simply as an NOE. For an NOE to be observable the two nuclei must be very close together, <0.5 nm (5 Å). The strength of the magnetization transfer falls off approximately as the sixth power of the interatomic distance. Consider two hydrogen atoms, both tightly hydrogen-bonded to an intervening oxygen atom, e.g., of a carboxylate or phosphate group. The expected H–H distance would be ~0.3 nm and a strong NOE between them would be anticipated. Two nonbonded hydrogen atoms (e.g., on methyl groups of amino acid side chains) can be as close together as 0.24 nm at van der Waals contact and could show a very strong NOE. However, contact is rarely this close in proteins unless in a hydrogen bond.




## 2. Nuclei Other than Hydrogen

***Deuterium ($^{2}H$).*** The natural abundance is very low so that use of $^{2}H$-labeled compounds is practical for study of metabolism, e.g., for following an $^{2}H$ label in glucose into products of fermentation[455] or in mammalian blood flow.[456] Deuterium NMR has been used extensively to study lipid bilayers (Chapter 8).

***Carbon - 13.*** Use of $^{13}C$ in NMR developed slowly because of the low natural abundance of this isotope. Another complication was the occurrence of $^{13}C$–$^{1}H$ coupling involving the many protons normally present in organic compounds. The latter problem was solved by the development of **wide-band proton decoupling** (noise decoupling). With a natural abundance of only 1.1%, $^{13}C$ is rarely present in a molecule at adjacent positions. Thus, $^{13}C$–$^{13}C$ coupling does not introduce complexities and in a noise-decoupled natural abundance spectrum each carbon atom gives rise to a single peak. Even so, $^{13}C$ NMR spectroscopy was not practical until pulsed Fourier transform (FT) spectrometers were developed (Section 2).

Chemical shifts in $^{13}C$ spectra are often 200 ppm or more downfield relative to TMS. The effects of substituents attached to a carbon atom are often additive when two or more substituents are attached to the same atom.[457,458] It is often necessary (but costly) to prepare compounds enriched in $^{13}C$ beyond the natural abundance. For proteins this may be done by growing an organism on a medium containing [$^{13}C$]glucose or a single amino acid enriched in $^{13}C$.[459,460] Using metabolites enriched in $^{13}C$, it is also possible to observe metabolism of living tissues directly. For example, glycogen synthesis from $^{13}C$-containing glucose has been observed in a human leg muscle using a wide-bore magnet and surface coils for transmitting and receiving.[461] This topic is discussed further in Box 17-C.

***Nitrogen - 15.*** Despite difficulties associated with low natural abundance (0.37%) and low sensitivity, $^{15}N$ NMR is practical and with isotopically enriched samples has become very important. Proteins with a high content of $^{15}N$ can be produced easily and inexpensively from cloned genes in bacterial plasmids. For example, cells of *E. coli* can be grown on a minimal medium containing [$^{15}N$] $NH_4Cl$. Since $^{13}C$ can also be introduced in a similar way it is possible to incorporate both isomers simultaneously. Production of uniformly labeled protein containing $^{15}N$ and / or $^{13}C$ provides the basis for multidimensional isotope-edited spectra necessary for protein structure determination (next section) and for study of tautomerization of histidine rings (Eq. 2-6).[460,462–464] $^{15}N$ chemical shifts of groups in proteins are spread over a broad range (Table 3-3).[465]

***Phosphorus - 31.*** NMR spectroscopy using $^{31}P$, the ordinary isotope of phosphorus, also has many uses.[466] Application of $^{31}P$ NMR to living tissues has been extraordinarily informative[467] and is dealt with in Chapter 6. The many phosphorus nuclei in nucleotides, coenzymes, and phosphorylated metabolites and proteins are all suitable objects of investigation by NMR techniques.

***Fluorine - 19.*** Although not abundant in nature, $^{19}F$ gives an easily detected NMR signal and can be incorporated in place of hydrogen atoms into many biochemical compounds including proteins.[468,469] In one study genetic methods were used to place 3-fluorotyrosine separately into eight positions in the lac repressor protein (Chapter 29).[470] Measurements of $^{19}F$ NMR spectra were used to study domain movement. Active site groups of enzymes can be modified to incorporate $^{19}F$.[471] Binding of fluorinated substrates can be studied.[472] Nontoxic $^{19}F$-containing compounds are useful as intracellular pH indicators, the NMR spectrometer

serving as the pH meter (Box 6-A).[473,474] An atom of fluorine attached to an aromatic ring is highly and predictibly sensitive to inductive effects of substituents in the para position to the fluorine.[475] For example, fluorine at the 6-position in pyridoxal phosphate (Fig. 14-4) can be observed in enzymes and reports changes in coenzyme structure.[476,477]

***Some other nuclei.*** Here are a few reported uses of NMR on other nuclei. $^{3}He$, binding into little cavities in fullerenes;[478] $^{11}B$, binding of boronic acids to active sites;[479] $^{23}Na$, measurement of intracellular $[Na^+]$;[480–482] $^{35}Cl$ and $^{37}Cl$, binding to serum albumin;[483] $^{113}Cd$, reporter that can replace $Zn^{2+}$ (no magnetic moment) in active sites of many enzymes and in nonenzymatic systems as well;[484] Tl, replacing $K^+$ in enzyme binding sites;[485] $^{17}O$, study of dynamics of protein hydration;[486] and $^{77}Se$, observation of acetylchymotrypsin intermediate.[487]

**TABLE 3-3**
**Approximate Chemical Shift Ranges in $^{1}H$- and in $^{15}N$- NMR Spectra**

| Group | $^{1}H$ chemical shift (ppm from TMS) | $^{15}N$ chemical shift (ppm from liquid $NH_3$) |
|---|---|---|
| $-CH_3$ | 0–4.0 | |
| $-CH_2-$ | 1.1–4.4 | |
| –CH | 2.4–5 | |
| Peptide αH random coil[a] | 3.9–5.0 | |
| –OH[b] | ~5–6 | |
| $>C=C<^{H}$ | 5–8 | |
| $-NH_2$ | | 31–37 |
| NH | | |
| Peptide | 7–12 | 103–142 |
| Aromatic H | 7–9 | |
| Imidazole | | |
| $C^{\varepsilon 1}-H$ | ~7.7 | |
| $C^{\delta 2}-H$ | ~7.0 | |
| N–H | ~10 | 165–180 |
| Imidazolium | | |
| $C^{\varepsilon 1}-H$ | ~8.7 | |
| $C^{\delta 2}-H$ | ~7.4 | |
| N–H | 10–18 | |
| –CHO Aldehyde | 9.4–10.4 | |
| –COOH | 11.3–12.2 | |
| Indole NH | ~10 | 130–145 |
| Guanidinium | | |
| $-N^{\eta}H_2$ | | 69–77 |
| $-N^{\varepsilon}H$ | | 31–37 |

[a] See Wishart *et al.*[448]
[b] See Linderstrøm-Lang[449]

## 3. Fourier Transform Spectrometers and Two-Dimensional NMR

Although NMR spectroscopy was widely used by the 1950s it was revolutionized by two developments, pioneered by Richard Ernst in the mid 1960s and 1970s.[488] The first of these was pulsed Fourier transform (FT) spectroscopy, which permits rapid accumulation of high-resolution spectra. In an FT NMR spectrometer a strong pulse of radiofrequency (RF) radiation is delivered to the sample over a period of a few microseconds and its effects are observed at all frequencies simultaneously. Although 1–2 or more seconds must be allowed before the next pulse is delivered, one complete NMR spectrum is obtained with each pulse. Often the results of hundreds, thousands, or even hundreds of thousands of pulses are added to provide greater sensitivity. With good temperature control this may be accomplished over periods of minutes to days.[489,490]

***Free induction decay.*** The strong exciting RF pulse is delivered with an orientation at right angles to that of the static field $H_o$ of the magnet and whose direction defines the *z* axis. As a result of this pulse, the magnetization of a nucleus is tilted away from the *z* axis and *precesses around the z axis* at its resonance (Larmor) frequency, which is ~500 MHz for $^{1}H$ in a 500-MHz spectrometer. The frequency that is measured is actually a difference from the "carrier frequency" of the RF pulse. The precessing magnetization of the nuclei has a component in the *xy* plane which induces an electrical signal in the coil of the NMR probe which defines the *y* axis. This signal, which contains the Larmor frequencies of all of the nuclei of a given element, is recorded as a function of time over a period of a few seconds. The signal decays away exponentially. However, this curve of **free induction decay** (FID) is not smooth but contains within it all of the Larmor frequencies. If enough points (perhaps 500 in a 2-s acquisition) are recorded and stored in the computer's memory, Fourier transformation of the data will produce the frequency-dependent NMR spectrum.[489,490]

***Relaxation times $T_1$ and $T_2$.*** When a very strong pulse of electromagnetic radiation is applied in the NMR spectrometer, virtually all of the nuclei are placed in the magnetically excited state. If another pulse were applied immediately, little energy would be absorbed because the system is **saturated**. In the

older "continuous-wave" NMR spectrometers, the energy is always kept small so that little saturation occurs. However, in FT NMR instruments, the strong pulses lead to a high degree of saturation. Application of repeated pulses would produce no useful information were it not for the fact that the excited nuclei soon relax back to their equilibrium energy distribution. Relaxation occurs through interactions of the nuclei with fluctuating magnetic fields in the environment. For organic molecules in solution the fluctuations that are most often effective in bringing about relaxation are the result of moving electrical dipoles in the immediate vicinity. Even so, relaxation of protons in water requires seconds.

Relaxation of nuclear magnetic states is characterized by two relaxation times. The longitudinal or **spin-lattice relaxation time** $T_1$ measures the rate of relaxation of the net magnetic vector of the nuclei in the direction of $H_o$. The transverse or **spin–spin relaxation time** $T_2$ measures the relaxation in the $xy$ plane perpendicular to the direction of $H_o$. The two relaxation times can be measured independently. In general, $T_2 < T_1$. For solids, $T_2$ is quite short ($\sim 10^{-5}$ s) whereas relaxation times of seconds are observed in solutions. This lengthening of the lifetime of the excited state in going from solid to liquid leads to a narrowing of absorption lines and explains why NMR bands in liquids are often narrow. However, an increase in viscosity or a loss of fluidity in a membrane leads to broadening.

How can $T_2$ and $T_1$ be measured? $T_2$ for fluids can often be estimated from the width of the band $\Delta\nu$ at half-height (Eq. 3-53). However, pulsed NMR methods are usually employed, the measurement of $T_1$ being especially easy.

$$T_2 \approx 1/\pi\,\Delta\nu \tag{3-53}$$

An attempt is often made to relate $T_1$ and $T_2$ to the molecular dynamics of a system. For this purpose a relationship is sought between $T_1$ or $T_2$ and the **correlation time $\tau_c$** of the nuclei under investigation. The correlation time is the time constant for exponential decay of the fluctuations in the medium that are responsible for relaxation of the magnetism of the nuclei. In general, $1/\tau_c$ can be thought of as a rate constant made up of the sum of all the rate constants for various independent processes that lead to relaxation. One of the most important of these ($1/\tau_1$) is for molecular tumbling.

$$1/\tau_1 \approx (3k_BT)/4\pi\eta r^3 \tag{3-54}$$

This equation is closely related to that of rotational diffusion (Eq. 9-35). Another term is the reciprocal of the **residence time** $\tau_m$, the mean time that a pair of dipoles are close enough together to lead to relaxation.

In the usual solvents at room temperature, $\tau_c$ is of the order of $10^{-12}$ to $10^{-10}$ s. Thus, relaxation rates in solutions are considerably faster than the frequencies of radiation absorbed in the NMR spectrometer ($\sim 10^8$ $s^{-1}$). Relaxation is relatively ineffective and $T_1$ and $T_2$ are usually large and equal. Bands remain sharp. As the correlation time increases (as happens, for example, if the viscosity is increased), $T_1$ and $T_2$ decrease with $T_1$ reaching a minimum when $\tau_c \sim \upsilon$, the frequency of the absorbed radiation. Lines are broadened and hyperfine lines (from coupling between nuclei) cannot be resolved. As $\tau_c$ is increased further, $T_2$ reaches a constant low value, while $T_1$ rises again. NMR measurements can be made in the region where $\tau_c$ exceeds $\nu$, a circumstance that is favored by the use of high-frequency spectrometers. On the other hand, in fluids it is more customary to work in the range of "extreme motional narrowing" at low values of $\tau_c$. Both $T_1$ and $T_2$ rise as the mobility of the molecules increases.

A limitation of use of NMR measurements of proteins comes from the increase in tumbling time with increasing size of the molecules. Since $1/\tau_r$ is often the most important term in the relaxation rate constant, only small proteins of mass <20 kDa give very sharp bands. Nevertheless, usable spectra are often obtainable on proteins ten times this size.

A practical problem in $^{13}C$ NMR arises from slow relaxation (long $T_1$). Partial saturation is attained and signal intensities are reduced for those carbon atoms for which relaxation is especially ineffective. Relaxation times can be measured separately for each carbon atom in a molecule and can yield a wealth of information about the **segmental motion** of groups within a molecule. Although the relationships between relaxation times and molecular motion are complex, they are often relatively simple for $^{13}C$. Carbon atoms are usually surrounded by attached hydrogen atoms, and dipole–dipole interactions with these hydrogen atoms cause most of the nuclear relaxation. For a carbon atom attached to N equivalent protons in a molecule undergoing rapid tumbling, Eq. 3-55 holds. Where $h = h/2\pi$ and $\gamma_C$ and $\gamma_H$ are the magnetogyric ratios of carbon and hydrogen nuclei.

$$1/T_1 \approx \frac{Nh^2\gamma_C{}^2\gamma_H{}^2\tau_{eff}}{r^6} \tag{3-55}$$

This equation permits a calculation of an effective correlation time $\tau_{eff}$ for each carbon atom.[491] $T_1$ and $T_2$ can also be evaluated for individual $^{15}N$ or $^{13}C$ nuclei in labeled proteins.[491]

***Two-dimensional and multidimensional NMR spectra.*** Proteins have such complex NMR spectra that, except for small regions at the upfield and downfield

ends (see Fig. 3-26A), it is impossible to interpret one-dimensional spectra. A solution to this problem came from the development by Jeener, Ernst, and Freeman of methods of displaying NMR spectra in two dimensions.[428,490,492,493] All two-dimensional and multidimensional NMR methods make use of one basic procedure: After the initial RF pulse, a second or a series of subsequent RF pulses are introduced. This is done before the nuclei have had time to relax completely. The time $t_1$ from the initial pulse to the second pulse is called the **evolution period** and allows accumulation of information about NOEs or *J*- coupling, whether homonuclear (e.g. $^1H–^1H$ or $^{13}C–^{13}C$) or heteronuclear (e.g., $^1H–^{13}C$ or $^1H–^{15}N$).

Two-dimensional spectra usually require hours or days of acquisition because separate FIDs are collected for a series of many different values of $t_1$. This provides a second timescale for the experiment. The second pulse is often used to rotate the directions of magnetization of the nuclei that are being observed from the *xy* plane into the *yz* plane but with the *z* component at 180° to the H° vector. One very important type of two-dimensional plot is the **NOESY** (NOE spectroscopy) spectrum, which detects NOEs between all excited nuclei that are close enough together (Fig. 3-25B). The transfer of magnetization occurs during a **mixing time** Δ, which follows the second pulse. For a protein Δ may be 25–300 ms. A third 90° pulse returns the *z* components of the magnetization to be parallel with the *y* axis and the FID is collected over the period $t_2$. The amplitude of the resonances detected is modulated by the frequencies that existed during the evolution

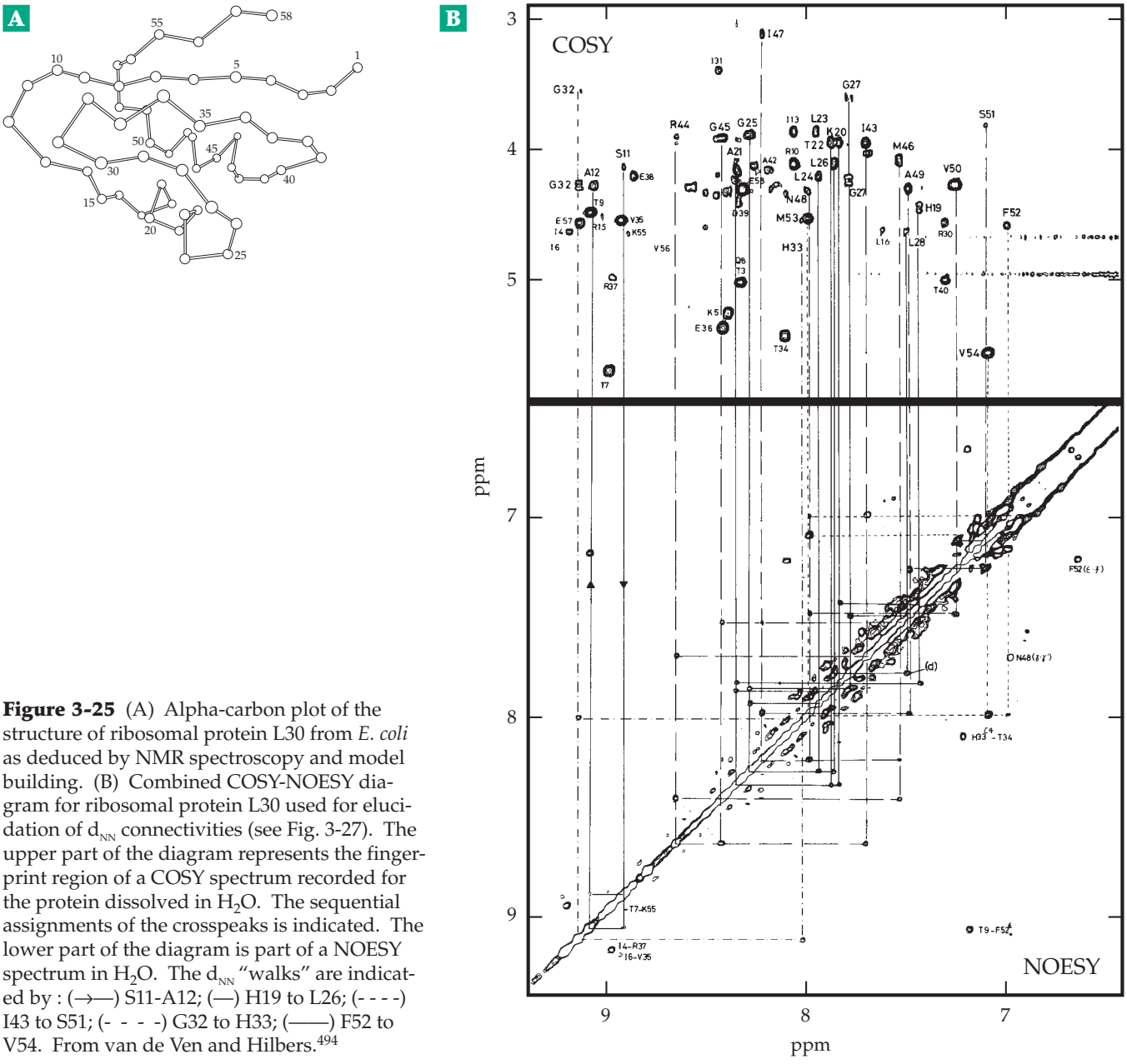



**Figure 3-25** (A) Alpha-carbon plot of the structure of ribosomal protein L30 from *E. coli* as deduced by NMR spectroscopy and model building. (B) Combined COSY-NOESY diagram for ribosomal protein L30 used for elucidation of $d_{NN}$ connectivities (see Fig. 3-27). The upper part of the diagram represents the fingerprint region of a COSY spectrum recorded for the protein dissolved in $H_2O$. The sequential assignments of the crosspeaks is indicated. The lower part of the diagram is part of a NOESY spectrum in $H_2O$. The $d_{NN}$ "walks" are indicated by : (→—) S11-A12; (—) H19 to L26; (- - - -) I43 to S51; (- - - -) G32 to H33; (——) F52 to V54. From van de Ven and Hilbers.[494]

period. After Fourier analysis[436,490] a two-dimensional plot with two frequency axes is generated and is usually displayed as a contour plot. Along the diagonal of Fig. 3-25B are peaks representing the one-dimensional spectrum. All of the peaks off of the diagonal are NOEs which can be related back to the peaks on the diagonal as shown by the horizontal and vertical lines.

A second important two-dimensional method is **correlation spectroscopy (COSY)**, in which the pairs of off-diagonal peaks result from spin–spin coupling. A related method called **TOCSY** provides correlations that extend through more than three bonds. The COSY plot in Fig. 3-26B is for the synthetic cyclic decapeptide *cyclo*-($\Delta^3$-Pro–D-*p*-Cl-Phe–D-Trp–Ser–Tyr–D-Trp–N-Me-Leu–Arg–Pro–β-Ala). It was obtained in six hours on a 500-MHZ instrument. The region marked I reveals couplings of protons on α-carbons to those on adjacent β carbons within the same residue ($J_{\alpha\beta}$; Fig. 3-27) and other couplings within the side chain. Each amino acid has a characteristic pattern. From careful study of this region it is possible to correlate each α-H resonance with a particular amino acid side chain. However, some residues are difficult to distinguish, e.g., His, Trp, Phe, and Tyr have similar $J_{\alpha\beta}$ values. Region II of Fig. 3-26B reveals connectivities of α and β hydrogens to N-H protons in the 7– 9 ppm region. Each α-H is coupled to the N-H of the same residue ($J_{\alpha H}$, Fig. 3-27). A section of a COSY plot is also shown in Fig. 3-25B and indicates how resonances can be related to those in the NOESY plot made on the same sample.

Of great importance in the determination of protein structures is the use of $^{15}$N- or $^{13}$C-enriched samples to obtain **isotope-edited** spectra. For example in **HSQC** or in **$^{15}$N-multiple quantum cohenence (HMQC)** spectra we see only NH protons in a plot of $^1$H chemical shift in one dimension versus the $^{15}$N chemical shift of the attached

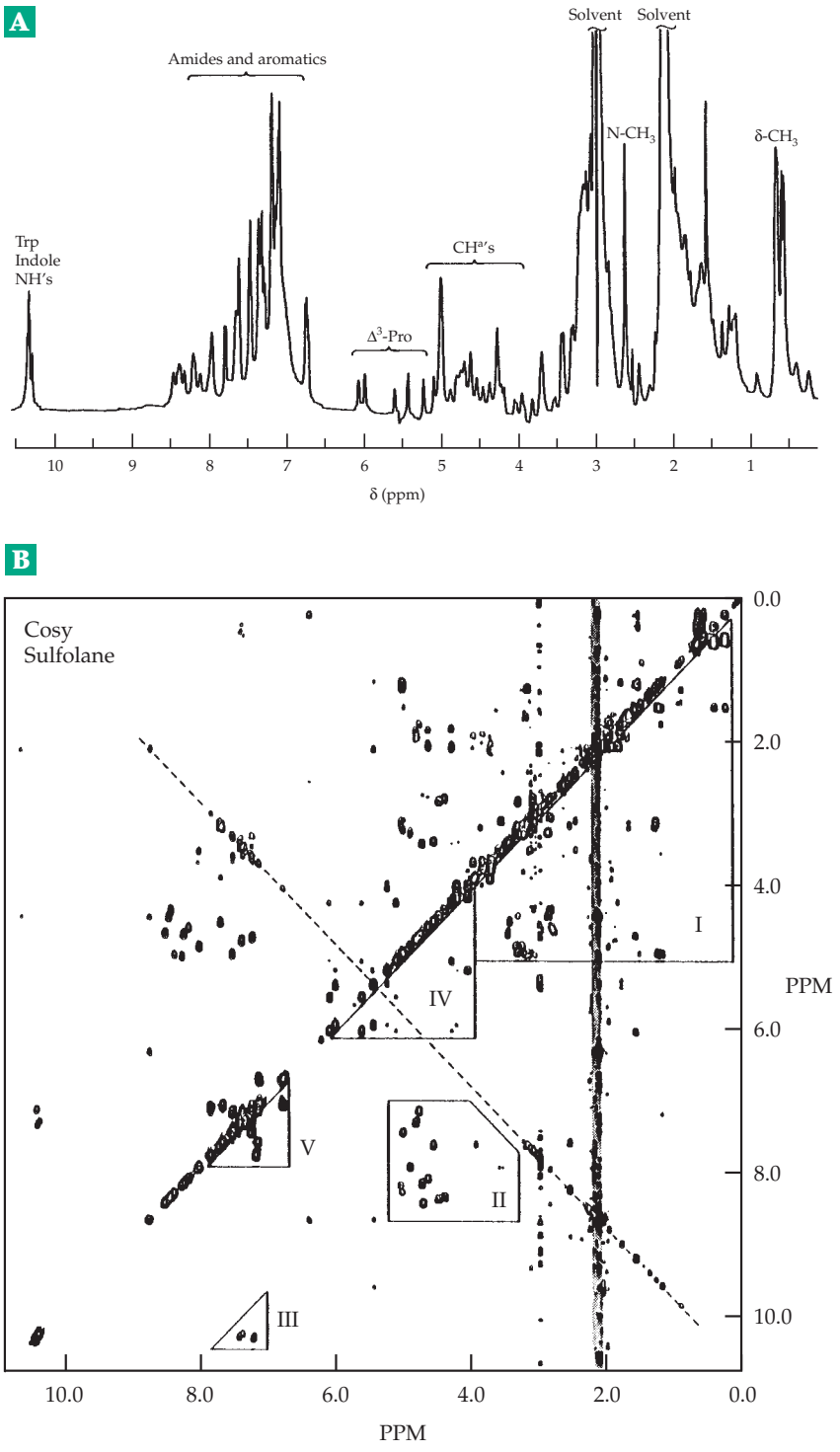



**Figure 3-26** Proton NMR spectra of a cyclic decapeptide analog of the gonadotropin-releasing hormone in the solvent sulfolane at 500 MHz. (A) One-dimensional spectrum. This figure also illustrates upfield methyl group regions, α-hydrogen, amide, aromatic, and downfield (10–20 ppm) regions. The indole NH resonances are shifted downfield by the ring current of the indole. (B) COSY spectrum plotted as a contour map. The outlined areas represent five unique *J*-coupled regions; area I, the $C^\alpha$ to $C^\beta$ to $C^\gamma$ etc., side chain connectivities; area II, the NH to $C^\alpha$ connectivities; area III, the indole NH of tryptophan; area IV, the connectivities of the $\Delta^3$-Prol residue; and area V, the aromatic resonances. Peaks that appear on the solid vertical line and the dashed diagonal lines are artifacts. From Baniak *et al.*[495]

nitrogen atom in the other (Fig. 3-28). Furthermore, as shown in this figure, particular types of NH bonds (peptide, $-NH_2$, imidazole, indole, amide, and guanidinium) appear in different regions. There is only one peptide NH per residue and, for a small protein, each may be separately visible.

## 4. Three-Dimensional Structures and Dynamics of Proteins

The first three-dimensional structure of a small protein was determined solely from NMR measurements in 1984. To date hundreds of "NMR structures" have been deduced. For small proteins of $M_r < 10{,}000$ two-dimensional COSY and NOESY spectra can suffice. For larger proteins use of $^{15}N$- and/or $^{13}C$-enriched proteins is essential.[428,430,496] This permits generation of a third and even a fourth frequency axis and three- and four dimensional NMR. For example, using various complex pulse sequences an $^{15}N$-correlated $^1H - {}^2H$ NOESY spectrum can be generated. This shows directly which NOEs arise from NH protons with resonances in the amide region of an HSQC or HMQC spectrum (Fig. 3-28). With both $^{15}N$ and $^{13}C$ present, many additional coupling patterns and $J$ values can be observed.[497–502] It is also possible to measure NOEs from atoms in a protein to those of a relatively weakly bound ligand such as a coenzyme or substrate analog and to determine the conformation of the bound ligand from this **transferred NOE**.[503–505]

***Assignment of resonances.*** After acquisition of the necessary data, which may require 10–15 mg of protein, the observed resonances must be assigned to specific amino acid residues in the peptide chain. The connectivities of the individual CH and NH groups that have been identified in the COSY spectrum and information about the relationship of one atom to another atom nearby in space are required. The closest neighbors to either αH or peptide NH protons are often protons in a neighboring residue ($d_{\alpha N}$, $d_{NN}$, Fig. 3-27). As was pointed out in the preceding section, NOE correlations obtained from plots such as that in Fig. 3-25B provide much of the information needed to establish which resonances belong to each residue in a known sequence.[494] The three-dimensional structure of the ribosomal protein L30 of *E. coli*. (Fig. 3-25A) was deduced entirely by NMR spectroscopy. A downfield part of the NOESY and COSY plots used is shown in Fig. 3-25B. This figure also shows how cross-peaks in the NOESY spectrum were correlated with identified COSY peaks. It is helpful initially to hunt for unique dipeptides that can be identified in the NOESY spectrum. Ambiguities that arise can be resolved by use of various additional techniques.

***NOEs and distance constraints.*** NOESY plots also contain the essential information needed to determine which side chains *distant* in the sequence are close together in space. A NOE observed for a pair of nuclei falls off as the inverse sixth power of the distance between them. For this reason, NOEs are observed only for pairs of atoms closer than about 0.4 nm. It is possible, in principle, to calculate distances between nuclei from the NOE intensity, but this is not accurate. Often, the NOE cross-peaks are grouped into three categories that correspond to maximum possible distances of 0.25, 0.30, and 0.40 nm. These can be related for the most part to intraresidue (e.g., $d_{N\alpha'}$ of Fig. 3-27) sequential (e.g., $d_{\alpha N'}$ and $d_{NN}$, Fig. 3-27) and long range backbone–backbone distances. These values constitute a series of **distance constraints** which are applied while making an automated computer search for a folding pattern that will meet these constraints and at the same time have acceptable torsion angles and good side chain packing throughout. This process makes use of distance geometry algorithms and other methods.[430,506,507] An early success was the solution of a 75-residue amylase inhibitor independently by crystallographers[508] and NMR spectroscopists.[509] The NMR structure was based on 401 NOE distance constraints, 168 distance constraints imposed by hydrogen bonds and 50 torsion angles deduced from $J$ values. Recently, refinement of NMR structures has been done as in X-ray crystallography.[507] Some structures have been refined using both NMR and X-ray data.[442]

The spectra in Figs. 3-25, 3-26 and 3-28 are for relatively small proteins. Spectra of larger proteins are more complex and lines are broader. Many techniques are used to simplify spectra. The NMR spectrum of a protein is simplified considerably if the protein is denatured by heating, and $^1H$ NMR spectra of "random coil" proteins can be predicted well from tables of standard chemical shifts for the individual amino acids.[510] Many amide NH protons exchange with solvent rapidly, making it easier to assign the remaining peaks. However, nearly all NH peaks will be seen in an HMQC or HSQC (Fig. 3-28) spectrum. Partial, or even complete, substitution of deuterium for hydrogen will also simplify spectra.[511–513] Microorganisms that will grow in a medium rich in $D_2O$ can be used as sources of partially deuterated proteins. Because the remaining protons usually have $^2H$ rather than $^1H$ as a neighbor, dipolar line broadening is reduced and sharper resonances are observed.[514] Substitution of $^{15}N$ for $^{14}N$ in the backbone amide groups can also yield spectra with narrower lines.[515] **Isotope-edited** NMR spectra allow simplification of complex two-dimensional spectra by observation of only those protons attached to an isotopically labeled nucleus, e.g., $^{13}C$ or $^{15}N$.[516–518] Measurement of $^{15}NH\text{-}C_{\alpha H}$ $J$ couplings facilitates structure determinations.[519]

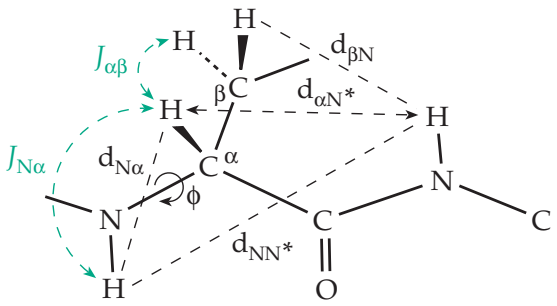



**Figure 3-27** Illustration of some distances ($d_{NN}$, $d_{\alpha N}$ and $d_{\beta N}$ obtained from NOESY spectra and some coupling constants ($J_{\alpha\beta}$ and $J_{N\alpha}$ obtained from $^1$H COSY spectra or *J*-resolved spectra. The coupling constants and proton chemical shifts provide a "fingerprint" for each residue and $J_{N\alpha}$ may also provide an estimated value for torsion angle f. The distances establish residue-to-residue connectivities as well as distance constraints that may permit a calculation of three-dimensional structure. Additional coupling constants can be measured for $^{15}$N- or $^{13}$C- enriched proteins. Coupling from β-hydrogens to other side chain hydrogens provides "fingerprint" information about individual residues.

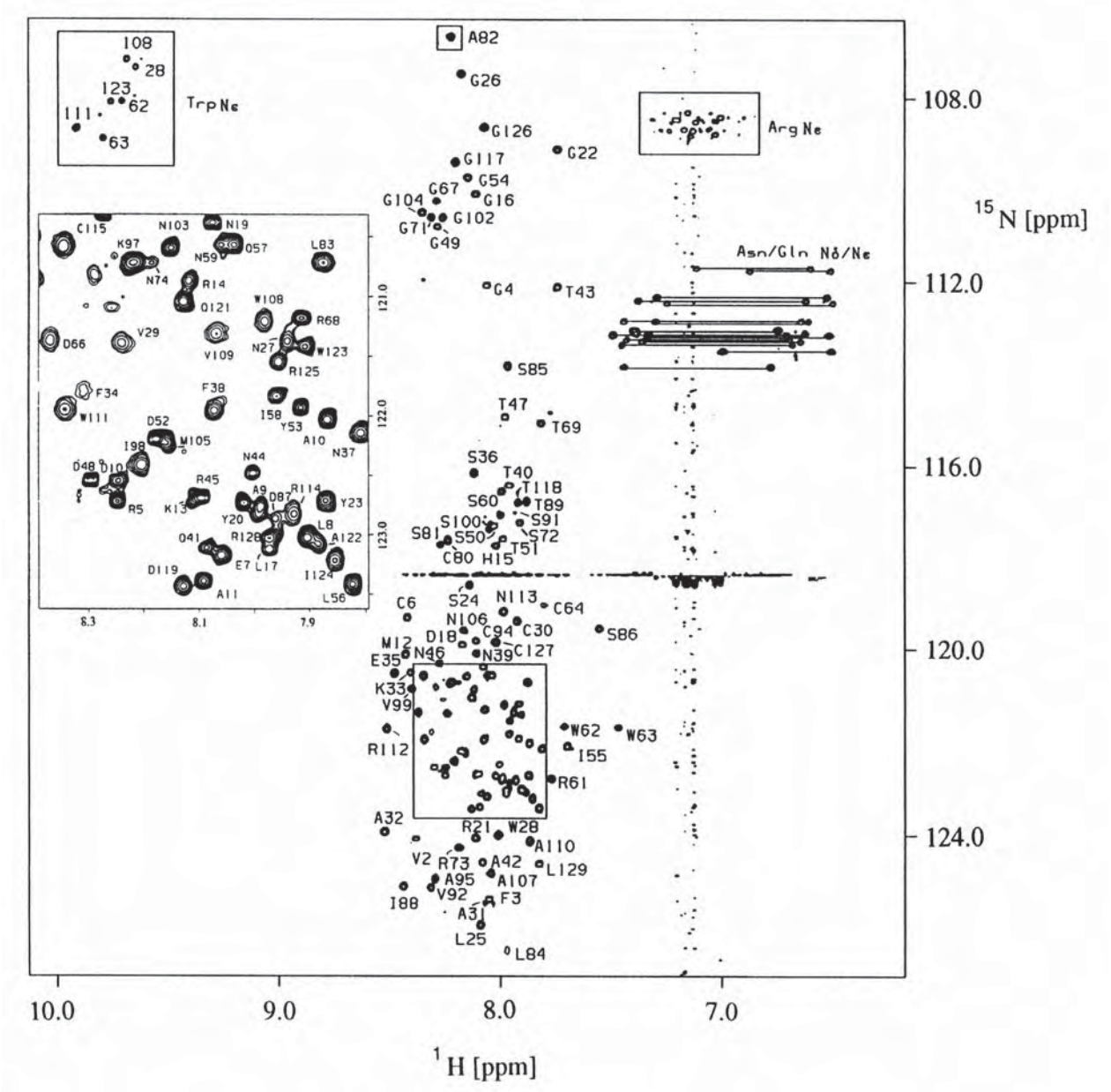



**Figure 3-28** A $^{15}$N – $^1$H HSQC spectrum of partially denatured 129-residue hen lysozyme. Boxes enclose the tryptophan indole region (upper left), the arginine side chain $N^{\varepsilon}$ region (upper left), and a portion of the amide NH region (lower center and enlarged in the insert). Resonances of pairs of hydrogen atoms in side chain (Asn and Gln) amide groups are indicated by horizontal lines. From Buck *et al.*[524]

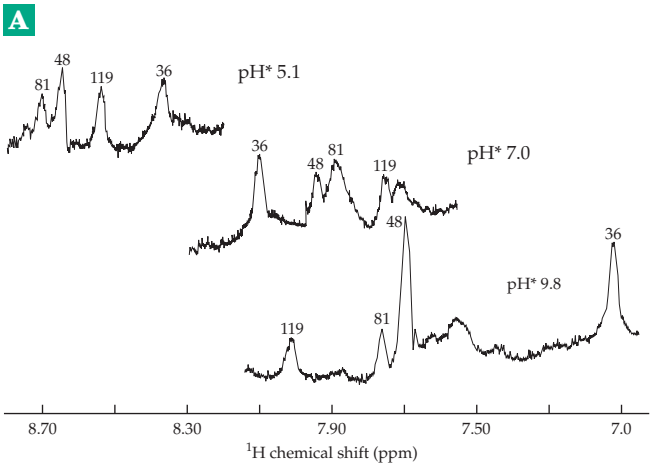
Many aspects of the dynamics of the action of enzymes of this size can be studied by NMR spectroscopy. New approaches allow study of proteins up to ~30 kDa in size.[519a] Computer programs can analyze data and make automatic assignments of resonances.[520,521]

Also useful are techniques for measuring NMR spectra on solids, including microcrystalline proteins[522] (see also Box 9-C). In some cases, e.g., for heme proteins,[522,523] the crystals can be oriented in a magnetic field permitting measurements of NMR spectra with more than one orientation of the crystals. This potentially affords more information than the usual techniques. For example, growth of bacteria in $^2H_2O$ media containing $^1$H-containing pyruvate yields proteins with almost complete deuteration in the C-α an C-β positions but with highly protonated methyl groups. This gives rise to good – $CH_3$ to – $CH_3$ NOEs and provides other advantages in both NMR spectroscopy and mass spectrometry.[512] Isotopically enriched amino acids, e.g., $^{13}$C-enriched leucine,[525] or isoleucine containing $^{15}$N, $^{13}$C, and $^2$H as well[526] can be fed to growing bacteria. The use of paramagnetic shifts by ions such as $Gd^{3+}$ may be helpful.[527] Very large shifts are sometimes induced in heme proteins by the $Fe^{3+}$ of the heme that is embedded within the protein.[528]

## 5. Other Information from NMR Spectra

***NMR titrations.*** How do pH changes affect NMR resonances? Resonances of $^1$H nuclei close enough to a proton with a p$K_a$ in the pH region under study will experience a shift, which may be either upfield or downfield when the proton dissociates. The resonance of $^{13}$C in a carboxyl group will shift downfield when the proton on the carboxyl group dissociates. If the proton that dissociates is tightly hydrogen bonded in a protein, its rate of dissociation may be *slow* compared to the NMR frequency used. If so, the original peak will decrease in value as the pH is raised and a new peak will appear at a position characteristic of the dissociated form. However, protons attached to N or O are usually in *rapid* exchange with the solvent. In this case, the NMR resonance of the nucleus being observed will move continuously from one chemical shift value at low pH to a different one at high pH. In an intermediate case the resonance will shift and broaden.

Both the $C^{\delta 2}$-H and $C^{\varepsilon 1}$-H protons of histidine can often be seen in proteins (Fig. 3-29A).[529–531] As is shown in Fig. 3-29 A and B, their chemical shifts are strongly dependent upon the state of protonation of the ring nitrogen atoms. At low pH, the positive charge that is shared by the two NH groups attracts electrons away from both CH positions, causing deshielding of the CH protons. The effect is greater for the $C^{\varepsilon 1}$-H than for $C^{\delta 2}$-H protons (Table 3-3). If the groups being titrated do not interact strongly with other nearby basic or



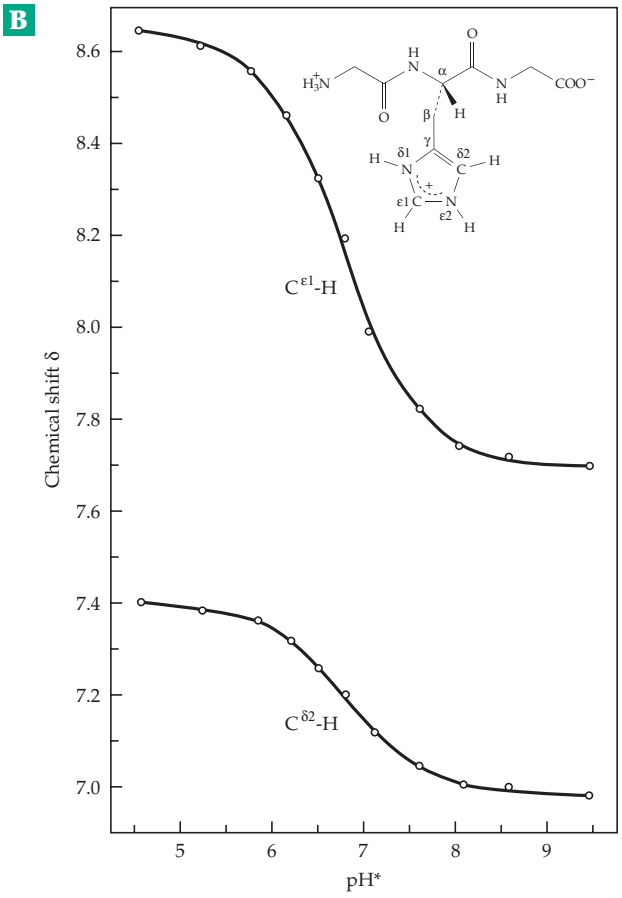




**Figure 3-29** (A) $^1$H NMR spectra of human myoglobin in $D_2O$ showing the C-H$^{\varepsilon 1}$ resonances of the imidazole rings of histidines 36, 48, 81, and 119 at three values of the apparent pH (pH*). From Bothelho and Gurd.[529] (B) $^1$H NMR titration curves for the histidine $C^{\varepsilon 1}$ and C-$^{\delta 2}$ ring protons of glycyl-L-histidylglycine. Data obtained at 100 MHz using 0.1 M tripeptide in $D_2O$ containing 0.3 M NaCl. The pH values (labeled pH*) are uncorrected glass electrode pH meter readings of $D_2O$ solutions using an electrode standardized with normal $H_2O$ buffers. Chemical shifts are downfield from TMS. Courtesy of J. L. Markley.[438]

acidic groups, p$K_a$ values can be estimated for individual histidines.[530,532] With $^{15}$N-containing proteins the tautomeric states ($NH^{\varepsilon 2}$ vs $NH^{\delta 1}$) of the imidazole rings can also be deduced.[533] Observation of imidazole rings shows that in proteins p$K_a$ values of imidazolium groups are sometimes less than five and sometimes greater than ten.[463,534]

***Observing exchangable protons.*** $^1$H spectra of proteins are often recorded in $D_2O$ because of interference from the very strong absorption of $H_2O$ at approximately 4.8 ppm. However, resonances of some rapidly exchanging protons are lost. Special pulse sequences, as well as improvements in spectrometer design, allow many of these resonances to be seen in $H_2O$.[535–537] At the far downfield end (δ>10) of $^1$H NMR spectra there are often weak peaks arising from NH protons of imidazole or indole side chains that can be observed in $H_2O$.[463,538–540] These resonances are often shifted 2–5 ppm downfield from the positions given in Table 3-3. The $^1$H resonance for a carboxyl (–COOH) proton is shifted to 20 ppm or more in spectra of very strongly hydrogen-bonded anionic complexes such as the malonate dianion.[541–544] (see Chapter 9, Section D,4). The strongest hydrogen bonds cause the greatest downfield shifts because the negative charge pulls the $^1$H proton away from the electrons of the atom to which it is attached, deshielding the proton.[545]

A good example is provided by the imidazole NH proton of the active site of trypsin and related serine proteases (Chapter 12), which is seen at ~16 ppm. Another example is provided by aspartate aminotransferase, whose $^1$H NMR spectrum in a dilute aqueous phosphate buffer is shown in Fig. 3-30. The peak labeled A is the resonance of the NH proton on the ring of the pyridoxal phosphate coenzyme (marked in Fig. 14-6) and peak B belongs to an adjacent imidazole group of histidine 143. Both of these protons move with pH changes around a p$K_a$ of ~ 6.2 which is associated with the Schiff base proton 6–8 nm away.[534,546–551] Peak A moves upfield 2.0 ppm and peak B downfield 1.0 ppm when the pH is raised around this p$K_a$. These hydrogen-bonded protons act as sensitive "reporters" of the electronic environment of the active site. Many proteins contain carboxylate or phosphate groups in their active sites and observation of NMR resonances of protons hydrogen bonded to them or to groups in substrates or inhibitors may be a useful technique for study of many enzymes and other proteins.

***Exchange rates of amide protons.*** The NH protons of the peptide backbone can be observed in $H_2O$ in the 6–11 ppm region. If the spectrum is recorded for a sample in $D_2O$, many of these resonances disappear gradually as the protons of the peptide units exchange with the deuterium ions of the medium. A study of the observed exchange rates can shed light on the dynamics of protein molecules in solution. Exchange of amide protons in proteins with $^2$H or $^3$H has been studied on a relatively slow timescale by such techniques as observation of infrared vibrations of the amide group (see Chapter 23).[552,553] However, the development of two-dimensional NMR dramatically improved the ability to study proton exchange.[552,554–556] The measurements may involve use of quenching by rapid solvent change,[556] special pulse sequences,[555] and study of $T_1$ relaxation rates (seconds timescale). Recently, electrospray mass spectrometry has also been exploited.[557,558] Exchange patterns for large proteins may be followed,[559] and proteolytic fragmentation into short peptides may be used after various lengths of exchange time to investigate the exchange in specific regions of a protein.[216,560] Unfolding of a protein under denaturing conditions can be studied, as can refolding of a completely denatured protein.[561]

Amide hydrogen exchange is usually discussed in terms of a model proposed by Linderstrøm-Lang in 1955.[554,556] He suggested that portions of protein molecules unfold sporadically to allow rapid exchange which can be catalyzed by $H^+$, $HO^-$, or other acids or

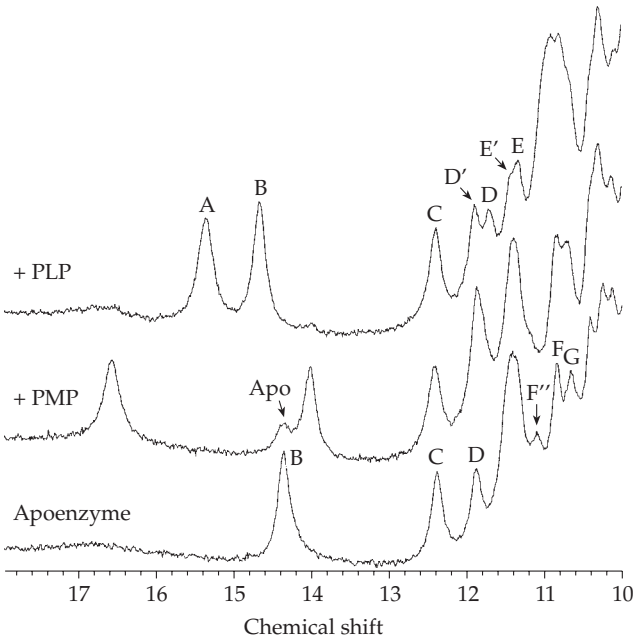



**Figure 3-30** Spectra of the pyridoxal phosphate (PLP), pyridoxamine phosphate (PMP) and apoenzyme forms of pig cytosolic aspartate aminotransferase at pH 8.3, 21°C. Some excess apoenzyme is present in the sample of the PMP form. Spectra were recorded at 500 $MH_2$. Chemical shift values are in parts per million relative to that of $H_2O$ taken as 4.80 ppm at 22°C. Peak A is from a proton on the ring nitrogen of PLP or PMP, peaks B and D are from imidazole NH groups of histidines 143 and 189 (see Fig. 14-6), and peaks C and D′ are from amide NH groups hydrogen bonded to carboxyl groups.

bases. The pH dependence of exchange for a given amide NH can be described as follows:

$$k_{ex} = k_H [H^+] + k_{OH} [OH^-] + k_W \qquad (3\text{-}56)$$

where $k_H$, $k_{OH}$, and $k_W$ are rate constants for acid-catalyzed, base-catalyzed, and the very slow water-catalyzed exchange.[554,556,562] While many amide protons exchange rapidly, hydrogen-bonded NH protons in well-packed hydrophobic core regions exchange slowly,[554,563] sometimes remaining in the protein for years in a $D_2O$ solution. Some unusually stable small proteins such as the seed protein **crambin** show little exchange. The $C^{\varepsilon 1}$ protons of histidine imidazoles also exchange slowly with $D_2O$ from the medium. The exchange rates[529,564] are rapid at higher temperatures, with an average half-life of about 11 min at 65°C. Different residues may exchange at different rates. Binding of substrates or inhibitors can stabilize the protein, slowing all exchange rates of both amide and imidazole groups.

***Solid-state NMR and other topics.*** Little can be said about these topics, but NMR measurement on solid crystalline materials is now practiced and is providing a wealth of information. It is being applied more often to biochemically related problems.[465,542,565 – 567]

NMR spectroscopy of nucleic acids is discussed briefly in Chapter 5. An important medical application of NMR is in **imaging**, a topic dealt with in Box 30-A.

## J. The Protein Data Bank, Three-Dimensional Structures, and Computation

The x,y,z coordinates of all atoms in published, refined three-dimensional structures have been deposited in the Protein Data Bank (Table 3-4).[568–571] Many other related databases are available,[572] e.g., covering molecular modeling,[573] gene sequences, proteome data,[574] and much, much more. A good way to keep up to date is to read the "computer corner" in *Trends in Biochemical Sciences (TIBS).* Most databases can be reached on the World Wide Web.[572] A selected list is

**TABLE 3-4**
**Selected World Wide Web Servers Related to Protein Structures and Sequences**[a]

| Question / area | Tool | Access / URL |
|---|---|---|
| Database search by comparison of 3D structures | Dali server | *http://www.embl-heidelberg.de/dali/dali.html* |
| Structural classification of proteins | SCOP | *http://www.bio.cam.ac.uk/scop/* |
| Summary and analysis of PDP structures | PDBSum | *http://www.biochem.ucl.ac.uk/bsm/pdbsum* |
| Retrieve 3D coordinates | Protein Data Bank | *http://www.rcsb.org/pdb/* |
| SWISS-PROT sequence database, Swissmodel homology modeling, etc. | ExPASy | *http://expasy.hcuge,ch/* |
| Molecular graphics viewer for PCs and workstations | RasMol | *http://www.bernstein-plus-sons.com/software/rasmol* |
| World Wide Web -Entrez and Molecular Modeling Database access | | *http://www.ncbi.nlm.nih.gov* |
| *Protein Science* Kinemages | MAGE and PREKIN | *http://www.prosci.uci.edu/kinemages/Kinemagelndex.html* |
| Pedro's Biomolecular Research Tools | | *http://www.fmi.ch/biology/research_tools.html* |
| Predict secondary structure from sequence | Predict Protein server | *http://www.sander.embl-heidelberg.de* |
| Browse databanks in molecular biology | SRS server | *http://www.embl-heidelberg/srs/srsc* |
| The Human Genome Database | | *http://gdbwww.gdb.org/* |
| Image Library of Biological Macromolecules | | *http://www.imb-jena.de/IMAGE.html* |
| Bacterial Nomenclature | | *http://www.gbf-braunschweig.de/DSMZ/bactnom/bactnam.htm* |

[a] From Holm and Sander,[570] Hogue *et al.*,[573] Laskowski *et al.*,[580] and Walsh *et al.*[572]

given in Table 3-4. The widely used viewer called RasMol can be used with your PC or MacIntosh computer[575–576a] or UNIX workstation[569,575]. Another way to view macromolecules is to use the *Protein Science* Kinemages ("kinetic images") using the program MAGE.[577,578]

Although it is mentioned in a few places, this book does not begin to describe the rapid growth of computation in biochemistry, biophysics, and biology in general. Very fast methods of protein structure determination are being developed.[579,582] One of the major goals in current computation is to predict folding patterns of proteins from their sequences.[582,583] By comparing sequences we can often guess an approximate structure but accurate predictions are still not possible. Having a structure, we would like to predict properties and reactivities and to be able to guess how two more macromolecules interact to form macromolecular complexes. We would like to understand the complex chain of nonpolar and electrostatic interactions that underlie the fundamental properties of catalysis, movement, and responsiveness of organisms. Many computers and ingeneous minds are working to help us match theory with reality in these areas. Read the current journals!

## References


1. Strang, R. (1981) *Trends Biochem. Sci.* **6**, VII–VIII
2. Michaelis, L. (1922) *Die Wasserstoffionenkonzentration*, p. 48, Springer-Verlag, Berlin
3. Dixon, M., and Webb, E. C., eds. (1979) *Enzymes*, 3rd ed., pp.138–163, Academic Press, New York
4. Dryer, R. L., and Lata, G. F. (1989) *Experimental Biochemistry*, Oxford Univ. Press, New York
5. Kuramitsu, S., and Hamaguchi, K. (1980) *J. Biochem.* **87**, 1215 – 1219
6. Beynon, R. J., and Easterby, J. S. (1996) *Buffer Solutions: The Basics*, IRL Press, Oxford
7. Bates, R. G. (1964) *Determination of pH*, Wiley, New York
8. Stoll, V. S., and Blanchard, J. S. (1990) in *Guide to Protein Purification* (Deutscher, M. P., ed), pp. 24–38, Academic Press, San Diego, California
9. Ellis, K. J., and Morrison, J. F. (1982) *Methods Enzymol.* **87**, 405–426
10. Good, N. E., Winget, G. D., Winter, W., Connolly, T. N., Izawa, S., and Singh, R. M. M. (1966) *Biochemistry* **5**, 467–477
11. Ferguson, W. J., Braunschweiger, K. I., Braunschweiger, W. R., Smith, J. R., McCormick, J. J., Wasmann, C. C., Jarvis, N. P., Bell, D. H., and Good, N. E. (1980) *Anal. Biochem.* **104**, 300–310
12. Dawson, R. M. C., Elliot, D. C., Elliot, W. H., and Jones, K. M. (1986) *Data for Biochemical Research*, 3rd ed., Oxford Science Pub., London
13. Robyt, J. F., and White, B. J. (1990) *Biochemical Techniques, Theory and Practice*, 2nd ed., Waveland Press, Inc., Prospect Height, Ill
14. Cooper, T. G. (1977) *The Tools of Biochemistry*, Wiley, New York
15. Alexander, R. R., and Griffiths, J. M. (1992) *Basic Biochemical Methods*, 2nd ed., Wiley, New York
16. Wilson, K., and Walker, J., eds. (1994) *Principles and Techniques of Practical Biochemistry*, 4th ed., Cambridge Univ. Press, New York
17. Boyer, R. F. (1992) *Modern Experimental Biochemistry*, 2nd ed., Benjamin-Cummings Publ., Redwood City, California
18. le Maire, M., Chabaud, R., and Herve, G. (1991) *Laboratory Guide to Biochemistry, Enzymology, and Protein Physical Chemistry*, Plenum, New York
19. Fini, C., Floridi, A., and Finelli, V. N., eds. (1989) *Laboratory Methodology in Biochemistry*, CRC Press, Boca Raton, Florida
20. Scopes, R. K. (1993) *Protein Purification: Principles and Practice*, 3rd ed., Springer, New York
21. Harris, E. L. V., and Angal, S. (1990) *Protein Purification: A Practical Approach*, IRL Press, Oxford, 2 volumes
22. Deutscher, M. P., ed. (1990) *Guide to Protein Purification: Methods in Enzymology*, Vol. 182, Academic Press, San Diego, California
23. Celis, J. E., ed. (1994) *Cell Biology: A Laboratory Handbook*, Academic Press, New York (3 volumes)
24. Fasman, G. D., ed. (1989) *CRC Practical Handbook of Biochemistry and Molecular Biology*, CRC Press, Boca Raton, Florida
25. Rosenberg, I. M. (1996) *Protein Analysis and Purification*, Birkhäuser, Cambridge, Massachusetts
26. Walker, J. M., ed. (1996) *The Protein Protocols Handbook*, Humana Press, Totowa, New Jersey
27. Marshak, D. R., Kadonaga, J. T., Burgess, R. R., Knuth, M. W., Brennan, W. A., Jr., and Lin, S.-H. (1996) *Strategies for Protein Purification and Characterization*, Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York
28. Creighton, T. E., ed. (1989) *Protein Structure: A Practical Approach*, IRL Press, Oxford
29. Marshak, D. R., Kadonaga, J. T., Burgess, R. R., Knuth, M. W., Brennan, W. A., and Lin, S.-H. (1995) *Strategies for Protein Purification and Characterization*, Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York
30. Crabb, J. W., ed. (1995) *Techniques in Protein Chemistry*, Vol. VI, Academic Press, San Diego, California
31. Walker, J. M., ed. (1994) *Basic Protein and Peptide Protocols*, Humana Press, Totowa, New Jersey
32. Bollag, D. M., and Edelstein, S. J. (1990) *Protein Methods*, Wiley-Liss, New York
33. Creighton, T. E., ed. (1989) *Protein Structure and Protein Function: A Practical Approach*, IRL Press, Oxford
34. Copeland, R. A. (1994) *Methods for Protein Analysis. A Practical Guide to Laboratory Protocols*, Chapman and Hall, New York
35. Matsuidaira, P. T., ed. (1995) *A Practical Guide to Protein and Peptide Purification for Microsequencing*, Academic Press, San Diego, CA
36. Hugli, T. E., ed. (1989) *Techniques in Protein Chemistry*, Academic Press, New York
37. Means, G. E., and Feeney, R. E. (1973) *Chemical Modification of Proteins*, Holden-Day, San Francisco
38. Keleti, G., and Lederer, W. H. (1974) *Handbook of Micromethods for the Biological Sciences*, Van Nostrand-Reinhold, New York
39. Switzer, R. L., and Garrity, L. F. (1999) *Experimental Biochemistry:Theory and Exercises in Fundamental Methods*, 3rd ed., Freeman, New York
40. Schulz, G. E., and Schirmer, R. H. (1979) *Principles of Protein Structure*, Springer-Verlag, New York
41. Scopes, R. K. (1994) *Protein Purification. Principles of Protein Structure*, Springer-Verlag, New York
42. Villafranca, J., ed. (1990) *Current Research in Protein Chemistry: Techniques, Structure, and Function*, Academic Press, New York
43. Coligan, J., Dunn, B., Ploegh, H., Speicher, D., and Wingfield, P., eds. (1995) *Current Protocols in Protein Science*, Wiley, New York
44. Matsudaira, P. T., ed. (1989) *A Practical Guide to Protein and Peptide Purification for Microsequencing*, Academic Press, San Diego, California
45. Doonan, S., ed. (1996) *Protein Purification Protocols*, Humana Press, Totowa, New Jersey
46. Little, P., and Hughes, S., eds. (1992) *Technique*, Vol. 4, Academic Press, London
47. *Methods: A Companion to Methods in Enzymology*, 1990–present, Academic Press, San Diego, California
48. Wankat, P. C., Van Oss, C. J., and Henry, J. D., Jr., eds. (1992) *Separation and Purification Methods*, Dekker, New York
49. *protein expression and purification*, (1990 – present), Academic Press, San Diego, California
50. Coligan, J., Dunn, B., Ploegh, H., Speicher, D., and Wingfield, P., eds. (1996–present) *Current Protocols in Protein Science*, Wiley, New York
51. Hirs, C. H. W., and Timasheff, S. N., eds. (1983) *Methods in Enzymology, Enzyme Structure*, Vol. 91, Academic Press, New York
52. Doolittle, R. F., ed. (1990) *Methods in Enzymology, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid*, Vol. 183, Academic Press, New York
53. Bollag, D. E., Rozycki, M. D., and Edelstein, S. J., eds. (1996) *Protein Methods*, 2nd ed., Wiley, New York
54. Birnie, G. D., and Fox, S. M., eds. (1969) *Subcellular Components*, Butterworth, London
55. Estabrook, R. W., and Pullman, M. E., eds. (1967) *Methods in Enzymology*, Vol. 10, Academic Press, New York
56. Kaback, H. R. (1971) in *Methods in Enzymology*, Vol. 22 (Jakoby, W. B., ed), pp. 99–120, Academic Press, New York
57. Dingle, J. T. (1972) *Lysosomes. A Laboratory Handbook*, North-Holland Publ., Amsterdam
58. Rasmussen, N. (1996) *Trends Biochem. Sci.* **21**, 319–321

58a. Jenkins, W. T. (1998) *Protein Sci.* **7**, 376–382

59. McPhie, P. (1971) in *Methods in Enzymology*, Vol. 22 (Jakoby, W. B., ed), pp. 23–32, Academic Press, New York
60. Blatt, W. F. (1971) in *Methods in Enzymology*, Vol. 22 (Jakoby, W. B., ed), pp. 39–49, Academic Press, New York
61. Porath, J., and Flodin, P. (1959) *Nature (London)* **183**, 1657–1659

## References


62. Porath, J. (1978) *Trends Biochem. Sci.* **3**, N100
63. Lathe, G. H. (1978) *Trends Biochem. Sci.* **3**, N99–N100
64. Flodin, P., and Aspberg, K. (1961) *Biol. Struct. Funct. Proc. IUB/IUBS Int. Sym*, 1st ed., Vol. 1, pp. 345-349 (p. 346)
65. Weir, M., and Sparks, J. (1987) *Biochem. J.* **245**, 85
66. Regnier, F. E. (1991) *Nature (London)* **350**, 634–635
67. Afeyan, N. B., Gordon, N. F., and Mazsaroff, I. (1990) *J. Chromatogr.* **519**, 1–29

67a. Wu, C.-s. (1999) *Column Handbook for Size Exclusion Chromatography*, Academic Press, San Diego, California

68. Liapis, A. I., and McCoy, M. A. (1992) *J. Chromatogr.* **599**, 87–104
69. Cantor, C. R., and Schimmel, P. R. (1980) *Biophysical Chemistry, Part II*, Freeman, San Francisco, California (pp. 549–685)
70. Kyte, J. (1995) *Structure in Protein Chemistry*, Garland Publ., New York
71. Brakke, M. K. (1953) *Arch. Biochem. Biophys.* **45**, 275–290
72. Britten, R. J., and Roberts, R. B. (1960) *Science* **131**, 32–33
73. Martin, R. G., and Ames, B. N. (1961) *J. Biol. Chem.* **236**, 1372–1379
74. Meselson, M., Stahl, F. W., and Vinograd, J. (1957) *Proc. Natl. Acad. Sci. U.S.A.* **43**, 581–588
75. da Silva, J. R. R. F. (1978) in *New Trends in Bio-organic Chemistry* (Williams, R. J. P., and da Silva, J. R. R. F., eds), Academic Press, London
76. Arakawa, T., and Timasheff, S. N. (1982) *Biochemistry* **21**, 6545–6552
77. Rao, J. K. M., and Argos, P. (1981) *Biochemistry* **20**, 6536–6543
78. Green, A. A., and Hughes, W. L. (1955) in *Methods in Enzymology*, Vol. 1 (Colowick, S. P., and Kaplan, N. O., eds), pp. 67–90, Academic Press, New York
79. Dawson, R. M. C., Elliott, D. C., Elliott, W. H., and Jones, K. M. (1969) *Data for Biochemical Research*, 2nd ed., Clarendon Press, Oxford
80. Ito, Y. (1982) *Trends Biochem. Sci.* **7**, 47–50
81. Ito, Y. (1987) *Nature (London)* **326**, 419–420
82. Mandava, N. B., and Ito, Y., eds. (1988) *Countercurrent Chromatography*, Dekker, New York

82a. Menet, J.-M., and Thiébaut, D., eds. (1999) *Countercurrent Chromatography*, Dekker, New York

83. Er-el, Z., Zaidenzaig, Y., and Shaltiel, S. (1972) *Biochem. Biophys. Res. Commun.* **49**, 383–390
84. Arakawa, T., and Timasheff, S. N. (1982) *Biochemistry* **21**, 6536–6544
85. Atha, D. H., and Ingham, K. C. (1981) *J. Biol. Chem.* **256**, 12108–12117
86. Kenney, A., and Fowell, S., eds. (1992) *Practical Protein Chromatography*, Humana Press, Totowa, New Jersey
87. Tiselius, A., Hjerten, S., and Levin, O. (1956) *Arch. Biochem. Biophys.* **65**, 132
88. Wolf, W. J., and Sly, D. A. (1964) *J. Chromatogr.* **15**, 247–250
89. Regnier, F. E. (1987) *Science* **238**, 319–323
90. Hutchens, T. W., and Porath, J. (1987) *Biochemistry* **26**, 7199–7204
91. Molnár, I., Horváth, C., and Jatlow, P. (1978) *Chromatographia* **11**, 260–265
92. Hamilton, R. J., and Sewell, P. A. (1982) *Introduction to High Performance Liquid Chromatography*, 2nd ed., Chapman and Hall, London
93. Berridge, J. C. (1985) *Techniques for the Automated Optimization of HPLC Separations*, Wiley, New York
94. Simpson, C. F. (1982) *Techniques in Liquid Chromatography*, Wiley, New York
95. Mant, C. T., and Hodges, R. S. (1991) *High-Performance Liquid Chromatography of Peptides and Proteins: Separation, Analysis, and Conformation*, CRC Press, Boca Raton, Florida
96. Gooding, K. M., and Regnier, F. E., eds. (1990) *HPLC of Biological Macromolecules*, Dekker, New York

96a. Beesley, T. E., Buglio, B., and Scott, R. P. W. (2000) *Quantitative Chromatographic Analysis*, Dekker, New York

97. Oliver, R. W. A., ed. (1989) *HPLC of Macromolecules: A Practical Approach*, IRL Press, Oxford
98. Fried, B., and Sherma, J., (1999) *Thin-Layer Chromatography*, 4th ed., Dekker, New York
99. Zlatkis, A., and Kaiser, R. E. (1977) *High Performance Thin-Layer Chromatography*, Elsevier Scientific-North Holland, Amsterdam
100. Sherma, J., and Fried, B., eds. (1996) *Practical Thin-Layer Chromatography: A Multidisciplinary Approach*, CRC Press, Boca Raton, Florida
101. Liu, M.-K., Li, P., and Giddings, J. C. (1993) *Protein Sci.* **2**, 1520–1531
102. Treadwell, G. E., Cairns, W. L., and Metzler, D. E. (1968) *J. Chromatogr.* **35**, 376–388
103. Novotny, M. V. (1989) *Science* **264**, 51–57
104. Schomvurg, G. (1990) *Gas Chromatography: A Practical Course*, VCH Publ., Cambridge, UK
105. Lee, M. L., and Markides, K. E. (1987) *Science* **235**, 1342–1347

105a. Muraviev, D., Gorshkov, V., and Warshawsky, A. (1999) *Ion Exchange*, Dekker, New York

106. Moore, S., and Stein, W. H. (1949) *Cold Spring Harbor Symp. Quant. Biol.* **14**, 179–190
107. Moore, S., and Stein, W. H. (1949) *J. Biol. Chem.* **178**, 53–77
108. Manning, J. M. (1993) *Protein Sci.* **2**, 1188–1191
109. Svasti, J. (1980) *Trends Biochem. Sci.* **5**, 8–9
110. Moore, S., and Stein, W. H. (1951) *J. Biol. Chem.* **192**, 663–681
111. Moore, S., and Stein, W. H. (1954) *J. Biol. Chem.* **211**, 893–906
112. Moore, S., and Stein, W. H. (1963) in *Methods in Enzymology*, Vol. 6 (Colowick, S. P., and Kaplan, N. O., eds), pp. 819–831, Academic Press, New York
113. Yamamoto, S., Nakanishi, K., and Matsuno, R. (1988) *Ion Exchange Chromatography of Proteins*, Dekker, New York
114. Peterson, E. A. (1970) *Cellulosic Ion Exchangers*, Elsevier-North Holland, Amsterdam
115. Nölting, B., Golbik, R., Neira, J. L., Soler-Gonzalez, A. S., Schreiber, G., and Fersht, A. R. (1997) *Proc. Natl. Acad. Sci. U.S.A.* **94**, 826–830
116. Lundell, N., and Markides, K. (1992) *Chromatographia* **34**, 369–375
117. Dean, P. D. G., Johnson, W. S., and Middle, F. A., eds. (1985) *Affinity Chromatography, A Practical Approach*, IRL Press, Oxford
118. Parikh, I., and Cuatrecasas, P. (1985) *Chem. Eng. News* **63** (Aug 26), 17–32
119. Porath, J. (1968) *Nature (London)* **218**, 834–838
120. Inman, J. K. (1974) in *Methods in Enzymology*, Vol. 34 (Jakoby, W. B., and Wilchek, M., eds), pp. 30–58, Academic Press, New York
121. Bethell, G. S., Ayers, J. S., Hearn, M. T. W., and Hancock, W. S. (1981) *J. Chromatogr.* **219**, 361–372
122. Wilchek, M., and Miron, T. (1987) *Biochemistry* **26**, 2155–2161
123. Ashley, G. W., and Bartlett, P. A. (1984) *J. Biol. Chem.* **259**, 13615–13620
124. Hengen, P. N. (1995) *Trends Biochem. Sci.* **20**, 285–286
125. Loo, T. W., and Clarke, D. M. (1995) *J. Biol. Chem.* **270**, 21449–21452
126. Janknecht, R., de Martynoff, G., Lou, J., Hipskind, R. A., Nordheim, A., and Stunnenberg, H. G. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8972–8976
127. Hochuli, E., Dobeli, H., and Schacher, A. (1987) *J. Chromatogr.* **411**, 177–184
128. Turner, A. J. (1981) *Trends Biochem. Sci.* **6**, 171–173
129. Vesterberg, O. (1971) in *Methods in Enzymology*, Vol. 22 (Jakoby, W. B., ed), pp. 389–412, Academic Press, New York
130. Bauer, J. (1994) *Cell Electrophoresis*, CRC Press, Boca Raton, Florida
131. Allison, S. A., and Tran, V. T. (1995) *Biophys. J.* **68**, 2261–2270
132. Shuster, L. (1971) in *Methods in Enzymology*, Vol. 22 (Jakoby, W. B., ed), pp. 412–433, Academic Press, New York
133. Chrambach, A., and Rodbard, D. (1971) *Science* **172**, 440–451
134. Andrews, A. T. (1986) *Electrophoresis: Theory, Techniques and Biochemical and Clinical Applications*, 2nd ed., Clarendon-Oxford Univ. Press, Oxford
135. Gordon, A. H. (1980) *Electrophoresis of Proteins in Polyacrylamide and Starch Gels*, Revised ed., Elsevier-North Holland, Amsterdam
136. Chrambach, A., Dunn, M. J., and Radola, B. J., eds. (1987) *Advances in Electrophoresis*, Vol. I, VCH Publ., Cambridge, UK
137. Serwer, P., Easom, R. A., Hayes, S. J., and Olson, M. S. (1989) *Trends Biochem. Sci.* **14**, 4–7
138. *electrophoresis*, (1995–present), VCH Publ., New York
139. *advances in electrophoresis*, (1987–present), VCH Publ., Cambridge
140. Hawcroft, D. M. (1997) *Electrophoresis: The Basics*, IRL Press, Oxford
141. Hengen, P. N. (1995) *Trends Biochem. Sci.* **20**, 202–203
142. Garfin, D. E. (1990) *Methods Enzymol.* **182**, 425–441
143. Hames, B. D., and Rickwood, D., eds. (1990) *Gel Electrophoresis of Proteins: A Practical Approach*, 2nd ed., IRL Press, Oxford
144. Righetti, P. G., ed. (1996) *Capillary Electrophoresis in Analytical Biotechnology*, CRC Press, Boca Raton, Florida
145. Camilleri, P., ed. (1993) *Capillary Electrophoresis: Theory and Practice*, CRC Press, Inc., Boca Raton
146. Altria, K. D., ed. (1996) *Capillary Electrophoresis Guidebook*, Humana Press, Totowa, New Jersey
147. Landers, J. P., ed. (1994) *Handbook of Capillary Electrophoresis*, CRC Press, Boca Raton, Florida
148. Landers, J. P. (1993) *Trends Biochem. Sci.* **18**, 409–414
149. Wehr, T., Rodríguez-Díaz, R., and Zhu, M. (1998) *Capillary Electrophoresis of Proteins*, Dekker, New York

149a. Yao, S., Anex, D. S., Caldwell, W. B., Arnold, D. W., Smith, K. B., and Schultz, P. G. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5372–5377

150. Valaskovic, G. A., Kelleher, N. L., and McLafferty, F. W. (1996) *Science* **273**, 1199–1202
151. Gao, J., Gomez, F. A., Härter, R., and Whitesides, G. M. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 12027–12030
152. Harrison, D. J., Fluri, K., Seiler, K., Fan, Z., Effenhauser, C. S., and Manz, A. (1993) *Science* **261**, 895–897
153. Radola, B. J., and Graesslin, D., eds. (1977) *Electrofocusing and Isotachophoresis*, de Gruyter, Berlin
154. Bocek, P., Dem, M., Gebauer, P., and Dolnik, V. (1987) *Analytical Isotachophoresis*, VCH Publ., Cambridge, UK

## References


155. Righetti, P. G. (1983) *Isoelectric Focusing: Theory, Methodology and Applications*, Elsevier - North Holland Biomedical Press, Amsterdam
156. Catsimpoolas, N., and Drysdale, J., eds. (1977) *Biological and Biomedical Applications of Isoelectric Focusing*, Plenum, New York
157. Egen, N. B., Bliss, M., Mayersohn, M., Owens, S. M., Arnold, L., and Bier, M. (1988) *Anal. Biochem.* **172**, 488–494
158. Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Westermeier, R., and Postel, W. (1982) *J. Biochem. Biophys. Methods* **6**, 317–339
159. Righetti, P. G., Gianazza, E., and Gelfi, C. (1988) *Trends Biochem. Sci.* **13**, 335–338
160. Righetti, P. G. (1990) *Immobilized pH Gradients: Theory and Methodology*, Elsevier, Amsterdam
161. Whitney, J. B., III, Cobb, R. R., Popp, R. A., and O'Rourke, T. W. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 7646–7650
162. Madden, M. S. (1995) *Anal. Biochem.* **229**, 203–206
163. Anderson, N. (1979) *Nature (London)* **278**, 122–123
164. Wade, N. (1981) *Science* **211**, 33–35
165. Henslee, J. G., and Srere, P. A. (1979) *J. Biol. Chem.* **254**, 5488–5491
166. Hanash, S. M., and Strahler, J. R. (1989) *Nature (London)* **337**, 485–486
167. Neukirchen, R. O., Schlosshauer, B., Baars, S., Jackle, H., and Schwarz, U. (1982) *J. Biol. Chem.* **257**, 15229–15234
168. Racine, R. R., and Langley, C. H. (1980) *Nature (London)* **283**, 855–857
169. Celis, J. E., and Bravo, R., eds. (1984) *Two-Dimensional Gel Electrophoresis of Proteins: Methods and Applications*, Academic Press, New York
170. Celis, J. E., and Bravo, R. (1981) *Trends Biochem. Sci.* **6**, 197–201
171. Garrels, J. I. (1989) *J. Biol. Chem.* **264**, 5269–5282
172. Krauss, M. R., Collins, P. J., and Blose, S. H. (1989) *Nature (London)* **337**, 669–670
173. Wood, T. D., Chen, L. H., Kelleher, N. L., Little, D. P., Kenyon, G. L., and McLafferty, F. W. (1995) *Biochemistry* **34**, 16251–16254

173a. McLafferty, F. W., Fridriksson, E. K., Horn, D. M., Lewis, M. A., and Zubarev, R. A. (1999) *Science* **284**, 1289 – 1290

174. Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko.A, Boucherie, H., and Mann, M. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14440–14445
175. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) *Nature (London)* **379**, 466–469
176. Hess, D., Covey, T. C., Winz, R., Brownsey, R. W., and Aebersold, R. (1993) *Protein Sci.* **2**, 1342–1351
177. Van Holde, K. E. (1985) *Physical Biochemistry*, 2nd ed., Prentice-Hall, Engelwood Cliffs, New Jersey
178. Cantor, C. R., and Schimmel, P. R. (1980) *Biophysical Chemistry*, Freeman, San Francisco, California
179. Svedburg, T., and Pederson, K. O. (1940) *The Ultracentrifuge*, Oxford Univ. Press, London
180. Schachman, H. K. (1989) *Nature (London)* **341**, 259–260
181. Schachman, H. K. (1992) in *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (Harding, S. E., Rowe, A. J., and Horton, J. C., eds), pp. 3–15, Royal Society for Chemistry, Cambridge, UK
182. Schachman, H. K. (1959) *Ultracentrifugation in Biochemistry*, Academic Press, New York
183. Rivas, G., and Minton, A. P. (1993) *Trends Biochem. Sci.* **18**, 284–287
184. Harding, S. E., Rowe, A. J., and Horton, J. C., eds. (1992) *Analytical Ultracentrifugation in Biochemistry and Polymer Science*, CRC Press, Boca Raton, Florida
185. Hansen, J. C., Lebowitz, J., and Demeler, B. (1994) *Biochemistry* **33**, 13155–13163
186. Ralston, G. (1993) *Introduction to Analytical Ultracentrifugation*, Beckman Instruments, Fullerton, California
187. Stafford, W. F., and Schuster, T. M. (1995) in *Introduction to Biophysical Methods for Protein and Nucleic Acid Research* (Glasel, J. A., and Deutscher, M. P., eds), pp. 111–145, Academic Press, San Diego
188. Bowen, T. J. (1970) *An Introduction to Ultracentrifugation*, Wiley (Interscience), New York
189. Bothwell, M. A., Howlett, G. J., and Schachman, H. K. (1978) *J. Biol. Chem.* **253**, 2073–2076
190. Pollet, R. J., Haase, B. A., and Standaert, M. L. (1979) *J. Biol. Chem.* **254**, 30–33

190a. *CRC Handbook of Biochemistry*, (1968) Chem. Rubber Publ. Co., Cleveland, Ohio

191. Jacobsen, M. P., Wills, P. R., and Winzor, D. J. (1996) *Biochemistry* **35**, 13173–13179

191a. Colowick, S. P., and Kaplan, N. O., eds., (1968) *Methods Enzymol.* **12B**, 388–389

192. Squire, P. G., and Himmel, M. E. (1979) *Arch. Biochem. Biophys.* **196**, 165–177
193. Andrews, P. (1965) *Biochem. J.* **96**, 595–606
194. Cantor, C. R., and Schimmel, P. R. (1980) *Biophysical Chemistry*, Freeman, San Fransico, California (pp. 670–684)
195. Nakos, G., and Mortenson, L. (1971) *Biochemistry* **10**, 455–458
196. Mann, K. G., and Fish, W. W. (1972) *Methods Enzymol.* **26C**, 28–42
197. Igou, D. K., Lo, J. T., Clark, D. S., Mattice, W. L., and Younathan, E. S. (1974) *Biochem. Biophys. Res. Commun.* **60**, 140–145
198. Weber, K., Pringle, J. R., and Osborn, M. (1972) *Methods Enzymol.* **26C**, 3–29
199. Wood, H. G., Werkman, C. H., Hemingway, A., and Nier, A. O. (1941) *J. Biol. Chem.* **139**, 377–381
200. Biemann, K. (1995) *Protein Sci.* **4**, 1920–1927

200a. Niessen, W. M. A. (1998) *Liquid Chromatography - Mass Spectrometry*, 2nd ed., Dekker, New York

200b. Chapman, J. R., ed. (2000) *Mass Spectrometry of Proteins and Peptides*, Hamana Press, Totowa, New Jersey

200c. Burlingame, A. L., Carr, S. A., and Baldwin, M. A., eds. (2000) *Mass Spectrometry in Biology and Medicine*, Hamana Press, Totowa, New Jersey

201. Senko, M. W., and McLafferty, F. W. (1994) *Annu. Rev. Biophys. Biomol. Struct.* **23**, 763–785
202. Li, Y., Hunter, R. L., and McIver, R. T., Jr. (1994) *Nature (London)* **370**, 393–395
203. Siuzdak, G. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11290–11297
204. Caprioli, R. M., and Suter, M. J.-F. (1995) in *Introduction to Biophysical Methods for Protein and Nucleic Acid Research* (Glasel, J. A., and Deutscher, M. P., eds), pp. 147-204, Academic Press, San Diego
205. Burlingame, A. L., and Carr, S. A., eds. (1996) *Mass Spectrometry in the Biological Sciences*, Humana Press, Totowa, New Jersey
206. Karas, M., and Hillenkamp, F. (1988) *Anal. Chem.* **60**, 2299–2303
207. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., and Yoshida, T. (1988) *Rapid Commun. Mass Spectrom.* **2**, 151–156
208. Beavis, R. C., and Chait, B. T. (1989) *Rapid Commun. Mass Spectrom.* **3**, 233–237
209. Hillenkamp, F., Karas, M., Beavis, R. C., and Chait, B. T. (1991) *Anal. Chem.* **63**, 1193A–1198A
210. Smith, R. D., Loo, J. A., Edmonds, C. G., Barinaga, C. J., and Udseth, H. R. (1990) *Anal. Chem.* **62**, 882–899
211. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1990) *Mass Spectrom. Rev.* **9**, 37
212. Dole, M., Mack, L. L., Hines, R. L., Mobley, R. C., Ferguson, L. D., and Alice, M. B. (1968) *J. Chem. Phys.* **49**, 2240
213. Mann, M., and Wilm, M. (1995) *Trends Biochem. Sci.* **20**, 219–224
214. Mann, M., Meng, C. K., and Fenn, J. B. (1989) *Anal. Chem.* **61**, 1702–1708
215. Covey, T. R., Bonner, R. F., Shushan, B. I., and Henion, J. (1988) *Rapid Commun. Mass Spectrom.* **2**, 249–256
216. Miranker, A., Kruppa, G. H., Robinson, C. V., Aplin, R. T., and Dobson, C. M. (1996) *J. Am. Chem. Soc.* **118**, 7402–7403
217. Marshall, A. G., Senko, M. W., Li, W., Li, M., Dillon, S., Guan, S., and Logan, T. M. (1997) *J. Am. Chem. Soc.* **119**, 433–434
218. Torchinsky, Y. M. (1981) *Sulfur in Proteins*, Pergamon, Oxford
219. Friedman, M. (1973) *The Chemistry and Biochemistry of the Sulfhydryl Group in Amino Acids, Peptides and Proteins*, Pergamon, Oxford
220. Jocelyn, P. C. (1972) *Biochemistry of the SH Group*, Academic Press, New York
221. Simpson, R. J., Neuberger, M. R., and Liu, T.-Y. (1976) *J. Biol. Chem.* **251**, 1936–1940
222. Hugli, T. E., and Moore, S. (1972) *J. Biol. Chem.* **247**, 2828–2834
223. Chin, C. C. Q., and Wold, F. (1974) *Anal. Biochem.* **61**, 379–391
224. Royer, G. P., Schwartz, W. E., and Liberatore, F. A. (1977) *Methods Enzymol.* **47E**, 40–45
225. Glazer, A. N. (1976) in *The Proteins*, 3rd ed., Vol. II (Neurath, H., and Hill, R. L., eds), pp. 1–103, Academic Press, New York
226. Plapp, B. V., Raftery, M. A., and Cole, R. D. (1967) *J. Biol. Chem.* **242**, 265–270
227. Doonan, S., and Fahmy, H. M. A. (1975) *Eur. J. Biochem.* **56**, 421–426
228. Mitchell, W. M. (1977) *Methods Enzymol.* **47E**, 165–169
229. Gadasi, H., Maruta, H., Collins, J. H., and Korn, E. D. (1979) *J. Biol. Chem.* **254**, 3631–3636
230. Drapeau, G. R. (1977) *Methods Enzymol.* **47E**, 189–191
231. Koida, M., and Walter, R. (1976) *J. Biol. Chem.* **251**, 7593–7599
232. Hayashi, R. (1977) *Methods Enzymol.* **47E**, 84–93
233. Krutzsch, H. C., and Pisano, J. J. (1977) *Methods Enzymol.* **47E**, 391–404
234. Mahoney, W. C., and Hermodson, M. A. (1979) *Biochemistry* **18**, 3810–3814
235. Hancock, W. S., ed. (1996) *New Methods in Peptide Mapping for the Characterization of Proteins*, CRC Press, Boca Raton, Florida
236. Lundblad, R. L. (1995) *Techniques in Protein Modification*, CRC Press, Boca Raton, Florida
237. Sanger, F. (1949) *Biochem. J.* **45**, 563–574
238. Sanger, F. (1988) *Ann. Rev. Biochem.* **57**, 1–28
239. Ruthen, R. (1993) *Sci. Am.* **269**(Oct), 30–31
240. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. H. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6970–6974
241. Edman, P. (1956) *Acta Chem. Scand.* **10**, 761–768
242. Edman, P., and Begg, C. (1967) *Eur. J. Biochem.* **1**, 80–91
243. Niall, H. D. (1977) *Nature (London)* **268**, 279–280
244. Schlesinger, D. H., ed. (1988) *Macromolecular Sequencing and Synthesis*, Liss, New York
245. Shively, J. E., ed. (1986) *Methods of Protein Microcharacterization*, Humana Press, Clifton, New Jersey

## References


246. Bhown, A. S. (1988) *Protein/Peptide Sequence Analysis: Current Methodoligies*, CRC Press, Boca Raton, Florida
247. Hunkapiller, M. W., and Hood, L. E. (1983) *Science* **219**, 650–659
248. Totty, N. F., Waterfield, M. D., and Hsuan, J. J. (1992) *Protein Sci.* **1**, 1215–1224
249. Gershoni, J. M. (1985) *Trends Biochem. Sci.* **10**, 103–106
250. Moos, M., Jr., Nguyen, N. Y., and Liu, T.-Y. (1988) *J. Biol. Chem.* **263**, 6005–6008
251. Baldo, B. A., and Tovey, E. R., eds. (1989) *Protein Blotting*, S. Karger, Farmington, CT
252. Wellner, D., Panneerselvam, C., and Horecker, B. L. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 1947–1949
253. Bailey, J. M., Nikfarjam, F., Shenoy, N. R., and Shively, J. E. (1992) *Protein Sci.* **1**, 1622–1633
254. Bailey, J. M., Shenoy, N. R., Ronk, M., and Shively, J. E. (1992) *Protein Sci.* **1**, 68–80
255. Johnson, R. S., and Walsh, K. A. (1992) *Protein Sci.* **1**, 1083–1091
255a. Keough, T., Youngquist, R. S., and Lacey, M. P. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96**, 7131–7136
256. Hunt, D. F., Shabanowitz, J., Yates, J. R., III, Zhu, N.-Z., Russell, D. H., and Castro, M. E. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84**, 620–623
257. Takao, T., Hitouji, T., Shimonishi, Y., Tanabe, T., Inouye, S., and Inouye, M. (1984) *J. Biol. Chem.* **259**, 6105–6109
257a. Jaffe, H., Veeranna, and Pant, H. C. (1998) *Biochemistry* **37**, 16211–16224
257b. Whitelegge, J. P., Gundersen, C. B., and Faull, K. F. (1998) *Protein Sci.* **7**, 1423–1430
258. Yazdanparast, R., Andrews, P. C., Smith, D. L., and Dixon, J. E. (1987) *J. Biol. Chem.* **262**, 2507–2513
259. Brown, J. R., and Hartley, B. S. (1966) *Biochem. J.* **101**, 214–228
260. Svobodova, S., Hais, I. M., and Kostir, J. V. (1953) *Chem. Listy* **47**, 205
261. Chen, R. F., Smith, P. D., and Maly, M. (1978) *Arch. Biochem. Biophys.* **189**, 241–250
262. Lee, K. S., and Drescher, D. G. (1979) *J. Biol. Chem.* **254**, 6248–6251
263. Chen, R. F., Scott, C., and Trepman, E. (1979) *Biochim. Biophys. Acta.* **576**, 440–455
264. Lunte, S. M. *et al.* (1989) *Anal. Biochem.* **178**, 202–207
265. Sharma, Y., Rao, C. M., Rao, S. C., Krishna, A. G., Somasundaram, T., and Balasubramanian, D. (1989) *J. Biol. Chem.* **264**, 20923–20927
266. Wilson, C. M. (1983) *Methods Enzymol.* **91**, 236–247
267. Merril, C. R. (1990) *Methods Enzymol.* **182**, 477–488
268. Finn, F. M., and Hofmann, K. (1976) in *The Proteins*, 3rd ed., Vol. II (Neurath, H., and Hill, R. L., eds), pp. 105–253, Academic Press, New York
269. Burnette, W. N. (1981) *Anal. Biochem.* **112**, 195–203
270. Karey, K. P., and Sirbasku, D. A. (1989) *Anal. Biochem.* **178**, 255–259
271. Strickland, E. H., Billups, C., and Kay, E. (1972) *Biochemistry* **11**, 3657–3662
272. Metzler, D. E., Harris, C., Yang, I.-Y., Siano, D., and Thomson, J. A. (1972) *Biochem. Biophys. Res. Commun.* **46**, 1588–1597
273. Morton, R. A. (1975) *Biochemical Spectroscopy*, Wiley, New York
274. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) *Protein Sci.* **4**, 2411–2423
275. Glazer, A. N., Delange, R. J., and Sigman, D. S. (1976) *Chemical Modification of Proteins: Selected Methods and Analytical Procedures*, Elsevier, New York
276. Feeney, R. E., Blankenhorn, G., and Dixon, H. B. F. (1975) *Advances in Protein Chemistry*, Vol. 29, Academic Press, New York (pp. 135–203)
277. Acharya, A. S., and Manjula, B. N. (1987) *Biochemistry* **26**, 3524–3530
278. Rypniewski, W. R., Holden, H. M., and Rayment, I. (1993) *Biochemistry* **32**, 9851–9858
279. Jentoft, N., and Dearborn, D. G. (1979) *J. Biol. Chem.* **254**, 4359–4365
280. Hennecke, M., and Plapp, B. N. (1984) *Anal. Biochem.* **136**, 110–118
281. Brocklehurst, K. (1979) *Int. J. Biochem* **10**, 259–274
282. Degani, Y., and Patchornik, A. (1974) *Biochemistry* **13**, 1–11
283. Riddles, P. W., Blakeley, R. L., and Zerner, B. (1983) *Methods Enzymol.* **91**, 49–60
284. Pande, C. S., Pelzig, M., and Glass, J. D. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 895–899
285. Fraker, P. J., and Speck, J. C., Jr. (1978) *Biochem. Biophys. Res. Commun.* **80**, 849–857
286. Tolan, D. R., Lambert, J. M., Boileau, G., Fanning, T. G., Kenny, J. W., Vassos, A., and Traut, R. R. (1980) *Anal. Biochem.* **103**, 101–109
287. Miles, E. W. (1977) *Methods Enzymol.* **47**, 431–442
288. Kawata, Y., Sakiyama, F., Hayashi, F., and Kyogoku, Y. (1990) *Eur. J. Biochem.* **187**, 255–262
289. Secundo, F., Carrea, G., D'Arrigo, P., and Servi, S. (1996) *Biochemistry* **35**, 9631–9636
290. Miura, S., Tomita, S., and Ichikawa, Y. (1991) *J. Biol. Chem.* **266**, 19212–19216
291. Colman, R. F. (1983) *Ann. Rev. Biochem.* **52**, 67–91
292. Chowdhry, V., and Westheimer, F. H. (1979) *Ann. Rev. Biochem.* **48**, 293–325
293. Jelenc, P. C., Cantor, C. R., and Simon, S. R. (1978) *Proc. Natl. Acad. Sci. U.S.A.* **75**, 3564–3568
294. Simmonds, R. J. (1992) *Chemistry of Biomolecules*, CRC Press, Boca Raton, Florida
295. Bodanszky, M., Bodanszky, A., and Trost, B. M., eds. (1994) *The Practice of Peptide Synthesis*, 2nd ed., Springer-Verlag, New York
296. Lloyd-Williams, P., Albericio, F., and Giralt, E. (1997) *Chemical Approaches to the Synthesis of Peptides and Proteins*, CRC Press, Boca Raton, Florida
297. Kent, S. B. H. (1988) *Ann. Rev. Biochem.* **57**, 957–989
298. Kaiser, E. T., Mihara, H., Laforet, G. A., Kelly, J. W., Walters, L., Findeis, M. A., and Sasaki, T. (1989) *Science* **243**, 187–192
299. Merrifield, B. (1986) *Science* **232**, 341–347
300. Atherton, E., Logan, C. J., and Sheppard, R. C. (1981) *J. Chem. Soc., Perkin, I.*, 538–546
301. Atherton, E., and Sheppard, R. C. (1985) *J. Chem. Soc. Chem. Commun.*, 165–166
301a. Miller, S. C., and Scanlan, T. S. (1998) *J. Am. Chem. Soc.* **120**, 2690–2691
302. You-shang, Z. (1983) *Trends Biochem. Sci.* **8**, 16–17
303. Kullman, W. (1987) *Enzymatic Peptide Synthesis*, CRC Press, Boca Raon, Florida
304. Liu, C.-F., and Tam, J. P. (1994) *J. Am. Chem. Soc.* **116**, 4149–4153
305. Tam, J. P., Lu, Y.-A., Liu, C.-F., and Shao, J. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 12485–12489
306. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C., and Schultz, P. G. (1989) *Science* **244**, 182–188
307. Borman, S. (1996) *Chem. Eng. News* **74** (Feb 12), 29–35
308. Youngquist, R. S., Fuentes, G. R., Lacey, M. P., and Keough, T. (1995) *J. Am. Chem. Soc.* **117**, 3900–3906
309. Cortese, R., ed. (1996) *Combinatorial Libraries: Synthesis, Screening and Application Potential*, Walter de Gruyter & Co., New York
310. Ellman, J., Stoddard, B., and Wells, J. (1997) *Proc. Natl. Acad. Sci. U.S.A.* **94**, 2779–2782
311. Kishchenko, G., Batliwala, H., and Makowski, L. (1994) *J. Mol. Biol.* **241**, 208–213
312. Scott, J. K. (1992) *Trends Biochem. Sci.* **17**, 241–245
312a. Barbas, C. F., Burton, D. B., Scott, J. K., and Silverman, G. J. (2000) *Phage Display: A Laboratory Manual* Cold Spring Harbor Press, Cold Spring Harbor, New York
313. Marvin, D. A., (1998) *Current Opinions in Structural Biology* **8**, 150–158
313a. Barbas, C. F., Burton, D. B., Scott, J. K., and Silverman, G. J. (2000) *Phage Display: A Laboratory Manual* Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York
314. McConnell, S. J., and Hoess, R. H. (1995) *J. Mol. Biol.* **250**, 460–470
315. Katz, B. A. (1995) *Biochemistry* **34**, 15421–15429
316. Allen, J. B., Walberg, M. W., Edwards, M. C., and Elledge, S. J. (1995) *Trends Biochem. Sci.* **20**, 511–516
317. Gates, C. M., Stemmer, W. P. C., Kaptein, R., and Schatz, P. J. (1996) *J. Mol. Biol.* **255**, 373–386
318. Bastos, M., Maeji, N. J., and Abeles, R. H. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6738–6742
319. Han, H., Wolfe, M. M., Brenner, S., and Janda, K. D. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6419–6423
320. Chu, Y.-H., Dunayevskiy, Y. M., Kirby, D. P., Vouros, P., and Karger, B. L. (1996) *J. Am. Chem. Soc.* **118**, 7827–7835
321. Déprez, B., Williard, X., Bourel, L., Coste, H., Hyafil, F., and Tartar, A. (1995) *J. Am. Chem. Soc.* **117**, 5405–5406
322. Kim, R. M., Manna, M., Hutchins, S. M., Griffin, P. R., Yates, N. A., Bernick, A. M., and Chapman, K. T. (1996) *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10012–10017
323. Sundberg, S. A., Barrett, R. W., Pirrung, M., Lu, A. L., Kiangsoontra, B., and Holmes, C. P. (1995) *J. Am. Chem. Soc.* **117**, 12050–12057
324. Baum, R. M. (1991) *Chem. Eng. News* **69**, 21–22
325. Slayter, E. M., and Slayter, H. S. (1992) *Light and Electron Microscopy*, Cambridge Univ. Press, New York
326. Bracegirdle, B. (1989) *Trends Biochem. Sci.* **14**, 464–468
327. Matthews, C. K., and van Holde, K. E. (1996) *Biochemistry*, 2nd ed., Benjamin, Menlo Park, California
328. Lichtman, J. W. (1994) *Sci. Am.* **271**(Aug), 40–45
329. Shotton, D., and White, N. (1989) *Trends Biochem. Sci.* **14**, 435–439
330. Boyde, A. (1985) *Science* **230**, 1270–1276
331. Gura, T. (1997) *Science* **276**, 1988–1990
332. Bennett, B. D., Jetton, T. L., Ying, G., Magnuson, M. A., and Piston, D. W. (1996) *J. Biol. Chem.* **271**, 3647–3651
333. Pennisi, E. (1997) *Science* **275**, 480–481
334. Maiti, S., Shear, J. B., Williams, R. M., Zipfel, W. R., and Webb, W. W. (1997) *Science* **275**, 530–532
335. Pool, R. (1988) *Science* **241**, 25–26
336. Betzig, E., and Chichester, R. J. (1993) *Science* **262**, 1422–1425
337. Dykstra, M. J. (1992) *Biological Electron Microscopy*, Plenum, New York
338. Hand, A. R. (1995) in *Introduction to Biophysical Methods for Protein and Nucleic Acid Research* (Glasel, J. A., and Deutscher, M. P., eds), pp. 205-260, Academic Press, San Diego
339. Nellist, P. D., McCallum, B. C., and Rodenburg, J. M. (1995) *Nature (London)* **374**, 630–632

## References


340. Schliwa, M. (1997) *Nature (London)* **387**, 764
341. Bubel, A., ed. (1989) *Microstructure and Function of Cells*, 1st ed., Ellis Harwood, New York
342. Penman, S. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 5251–5257
343. Morel, G., ed. (1995) *Visualization of Nucleic Acids*, CRC Press, Boca Raton, Florida
344. Caspar, D. L. D., and DeRosier, D. J. (1982) *Science* **218**, 653–655
345. Frank, J., ed. (1992) *Electron Tomography: Three-Dimensional Imaging with the Transmission Electron Microscope*, Plenum, New York
346. Skoglund, U., and Daneholt, B. (1986) *Trends Biochem. Sci.* **11**, 499–503
347. Lambert, O., Boisset, N., Taveau, J.-C., and Lamy, J. N. (1994) *J. Mol. Biol.* **244**, 640–647
348. Taveau, J.-C., Boisset, N., Lamy, J., Lambert, O., and Lamy, J. N. (1997) *J. Mol. Biol.* **266**, 1002–1015
349. De Rosier, D. J. (1997) *Nature (London)* **386**, 26–27
350. Dorset, D. L. (1997) *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1791–1794
351. Binnig, G., and Rohrer, H. (1987) *Rev. Mod. Phys.* **59**, 615–625
352. Binnig, G., and Rohrer, H. (1985) *Sci. Am.* **253**, 50–56
353. Wiesendanger, R. (1997) *Proc. Natl. Acad. Sci. U.S.A.* **94**, 12749–12750
354. Wickramasinghe, H. K. (1989) *Sci. Am.* **261**(Oct), 98–105
355. Hansma, P. K., Elings, V. B., Marti, O., and Bracker, C. E. (1988) *Science* **242**, 209–216
356. Pool, R. (1990) *Science* **247**, 634–636
357. Edstrom, R. D., Yang, X., Lee, G., and Evans, D. F. (1990) *FASEB J.* **4**, 3144–3151
358. Engel, A. (1991) *Ann. Rev. Biophys. Biophys. Chem.* **20**, 79–108
359. Amato, I. (1997) *Science* **276**, 1982–1985
360. Prater, C. B., Butt, H. J., and Hansma, P. K. (1990) *Nature (London)* **345**, 839–840
361. Hansma, H. G., and Hoh, J. H. (1994) *Annu. Rev. Biophys. Biomol. Struct.* **23**, 115–139
362. Radmacher, M., Tillmann, R. W., Fritz, M., and Gaub, H. E. (1992) *Science* **257**, 1900–1905
363. Karrasch, S., Hegerl, R., Hoh, J. H., Baumeister, W., and Engel, A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 836–838
364. Mou, J., Yang, J., and Shao, Z. (1995) *J. Mol. Biol.* **248**, 507–512
365. Yip, C. M., and Ward, M. D. (1996) *Biophys. J.* **71**, 1071–1078
366. Mou, J., Sheng, S., Ho, R., and Shao, Z. (1996) *Biophys. J.* **71**, 2213–2221
367. Kuznetsov, YuG., Malkin, A. J., Land, T. A., DeYoreo, J. J., Barba, A. P., Konnert, J., and McPherson, A. (1997) *Biophys. J.* **72**, 2357–2364

367a. Czajkowsky, D. M., and Shao, Z. (1998) *FEBS Letters* **430**, 51–54

368. Noy, A., Frisbie, C. D., Rozsnyai, L. F., Wrighton, M. S., and Lieber, C. M. (1995) *J. Am. Chem. Soc.* **117**, 7943–7951
369. Boland, T., and Ratner, B. D. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 5297–5301
370. Miles, M. (1997) *Science* **277**, 1845–1847
371. Hansma, H. G., Laney, D. E., Bezanilla, M., Sinsheimer, R. L., and Hansma, P. K. (1995) *Biophys. J.* **68**, 1672–1677

371a. Fisher, T. E., Oberhauser, A. F., Carrion-Vazquez, M., Marszalek, P. E., and Fernandez, J. M. (1999) *Trends Biochem. Sci.* **24**, 379–384

372. Blundell, T. L., and Johnson, L. N. (1976) *Protein Crystallography*, Academic Press, New York
373. Cantor, C. R., and Schimmel, P. R. (1980) *Biophysical Chemistry*, Freeman, San Francisco, California (pp. 687–811)
374. McPherson, A. (1982) *Preparation and Analysis of Protein Crystals*, Wiley, New York
375. Wyckoff, H. W., Hirs, C. H. W., and Timasheff, S. N., eds. (1985) *Methods in Enzymology*, Vol. 114,115, Academic Press, Orlando, Florida
376. Allewell, N. M., and Trikha, J. (1995) in *Introduction to Biophysical Methods for Protein and Nucleic Acid Research* (Glasel, J. A., and Deutscher, M. P., eds), pp. 381–431, Academic Press, San Diego
377. Verlinde, C. L. M. J., Merritt, E. A., Van Den Akker, F., Kim, H., Feil, I., Delboni, L. F., Mande, S. C., Sarfaty, S., Petra, P. H., and Hol, W. G. J. (1994) *Protein Sci.* **3**, 1670–1686
378. Woolfson, N. M. (1997) *An Introduction to X-ray Crystallography*, 2nd ed., Cambridge Univ. Press, New York
379. Thomas, J. M. (1993) *Nature (London)* **364**, 478–482
380. Dounce, A. L., and Allen, P. Z. (1988) *Trends Biochem. Sci.* **13**, 317–320
381. Bernal, J. D., and Crowfoot, D. (1934) *Nature (London)* **133**, 794–795
382. Glusker, J. P. (1994) *Protein Sci.* **3**, 2465-2469
383. Perutz, M. F., Muirhead, H., Cox, J. M., Goaman, L. G. G., Mathews, F. S., McGandy, E. L., and Webb, L. E. (1968) *Nature (London)* **219**, 29–32
384. Perutz, M. (1992) *Protein Structure: New Approaches to Disease and Therapy*, W.H. Freeman, San Fransico and New York
385. Eisenberg, D. (1994) *Protein Sci.* **3**, 1625–1628
386. Perutz, M. F. (1978) *Sci. Am.* **239**(Dec), 92–125
387. Tegze, M., and Faigel, G. (1996) *Nature (London)* **380**, 49–51
388. Fadley, C. S., and Len, P. M. (1996) *Nature (London)* **380**, 27–28
389. Rhodes, G. (1993) *Crystallography Made Crystal Clear*, Academic Press, New York

389a. Glusker, J. P., and Trueblood, K. N. (1985) *Crystal Structure Analysis A Primer*, 2nd ed., Oxford Univ. Press, Oxford

389b. Drenth, J. (1994) *Principles of Protein X-ray Crystallography*, Springler-Verlag, New York

389c. McRee, D. E. (1999) *Practical Protein Crystallography*, Academic Press, San Diego, California

390. Sweet, R. M. (1985) *Methods Enzymol.* **114**, 19–46
391. Kendrew, J. C. (1963) *Science* **139**, 1259–1266
392. Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., and Shore, V. C. (1960) *Nature (London)* **185**, 422–427
393. Rossman, M. G. (1994) *Protein Sci.* **3**, 1731–1733
394. Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R. (1965) *Nature (London)* **206**, 757–761
395. Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., and Thornton, J. M. (1997) *Trends Biochem. Sci.* **22**, 488–490
396. Gerstein, M., and Levitt, M. (1997) *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11911–11916
397. Marvin, D. A., and Nave, C. (1982) in *Structural Molecular Biology* (Davies, D. B., Saenger, W., and Danyluk, S. S., eds), pp. 3–44, Plenum, New York
398. Ducruix, A., and Giegé, R., eds. (1999) *Crystallization of Nucleic Acids and Proteins*, 2nd ed., IRL Press, Oxford
399. McPherson, A., Koszelak, S., Axelrod, H., Day, J., Williams, R., Robinson, L., McGrath, M., and Cascio, D. (1986) *J. Biol. Chem.* **261**, 1969–1975
400. McPherson, A., Jr. (1976) *J. Biol. Chem.* **251**, 6300–6303
401. Ray, W. J., Jr., and Puvathingal, J. M. (1986) *J. Biol. Chem.* **261**, 11544–11549
402. Day, J., and McPherson, A. (1992) *Protein Sci.* **1**, 1254–1268
403. Koszelak, S., Day, J., Leja, C., Cudney, R., and McPherson, A. (1995) *Biophys. J.* **69**, 13–19
404. Service, R. F. (1999) *Science* **285**, 1342–1346
405. Knäblein, J., Neuefeind, T., Schneider, F., Bergner, A., Messerschmidt, A., Löwe, J., Steipe, B., and Huber, R. (1997) *J. Mol. Biol.* **270**, 1–7
406. Glusker, J. P. (1984) *Trends Biochem. Sci.* **9**, 328–330
407. Guss, J. M., Merritt, E. A., Phizackerley, R. P., Hedman, B., Murata, M., Hodgson, K. O., and Freeman, H. C. (1988) *Science* **241**, 806–811
408. Walter, R. L., Ealick, S. E., Friedman, A. M., Blake, R. C., II, Proctor, P., and Shoham, M. (1996) *J. Mol. Biol.* **263**, 730–751
409. Budisa, N., Karnbrock, W., Steinbacher, S., Humm, A., Prade, L., Neuefeind, T., Moroder, L., and Huber, R. (1997) *J. Mol. Biol.* **270**, 616–623
410. Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A., and Cambillau, C. (1997) *J. Mol. Biol.* **268**, 779–799

410a. Yamano, A., Heo, N.-H., and Teeter, M. M. (1997) *J. Biol. Chem.* **272**, 9597–9600

410b. Lamzin, V. S., Morris, R. J., Dauter, Z., Wilson, K. S., and Teeter, M. M. (1999) *J. Biol. Chem.* **274**, 20753–20755

411. Rhee, S., Silva, M. M., Hyde, C. C., Rogers, P. H., Metzler, C. M., Metzler, D. E., and Arnone, A. (1997) *J. Biol. Chem.* **272**, 17293–17302
412. Frauenfelder, H., Petsko, G. A., and Tsernoglou, D. (1979) *Nature (London)* **280**, 558–563
413. Arnone, A., Rogers, P. H., Hyde, C. C., Briley, P. D., Metzler, C. M., and Metzler, D. E. (1985) in *Transaminases* (Christen, P., and Metzler, D. E., eds), pp. 138–155, Wiley, New York
414. Wlodawer, A. (1982) *Prog. Biophys. and Mol. Biol.* **40**, 115–159
415. Kossiakoff, A. A. (1985) *Ann. Rev. Biochem.* **54**, 1195–1227
416. Butler, D. (1994) *Nature (London)* **371**, 469
417. Hellemans, A. (1997) *Science* **277**, 1214–1215
418. Glanz, J. (1995) *Science* **267**, 1904–1906
419. Pool, R. (1988) *Science* **241**, 295
420. Eisenberger, P., and Suckewer, S. (1996) *Science* **274**, 201–202
421. Neutze, R., and Hajdu, J. (1997) *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5651–5655
422. Hajdu, J., and Johnson, L. N. (1990) *Biochemistry* **29**, 1669–1678
423. Duke, E. M. H., Wakatsuki, S., Hadfield, A., and Johnson, L. N. (1994) *Protein Sci.* **3**, 1178–1196
424. Jacoby, M. (1997) *Chem. Eng. News* **Dec 8**, 5
425. Rischel, C., Rousse, A., Uschmann, I., Albouy, P.-A., Geindre, J.-P., Audebert, P., Gauthier, J.-C., Förster, E., Martin, J.-L., and Antonetti, A. (1997) *Nature (London)* **390**, 490–492
426. Rattle, H. (1995) *An NMR Primer for Life Scientists*, Partnership Press, Tuckerton, NJ
427. Roberts, G. C. K., ed. (1993) *NMR of Macromolecules. A Practical Approach*, IRL Press, Oxford
428. Evans, J. N. S. (1995) *Biomolecular NMR Spectroscopy*, Oxford University Press, New York
429. Olson, D. L., Peck, T. L., Webb, A. G., Magin, R. L., and Sweedler, J. V. (1995) *Science* **270**, 1967–1969
430. Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York
431. Nageswara Rao, B. D., and Kemple, M. D., eds. (1996) *NMR as a structural tool for macromolecules: Current status and future directions*, Plenum, New York
432. Bovey, F. A. (1969) *Nuclear Magnetic Resonance Spectroscopy*, Academic Press, New York

## References


433. Sanders, J. K. M., and Hunter, B. K. (1993) *Modern NMR Spectroscopy*, 2nd ed., Oxford Univ. Press, London
434. Dybowski, C., and Lichter, R. L., eds. (1987) *NMR Spectroscopy Techniques*, Vol. 5, Dekker, New York
435. Bangerter, B. W. (1995) in *Introduction to Biophysical Methods for Protein and Nucleic Acid Research* (Glasel, J. A., and Deutscher, M. P., eds), pp. 317-379, Academic Press, San Diego
436. Freeman, R. (1997) *A Handbook of Nuclear Magnetic Resonance*, 2nd ed., Addison Wesley Longman Limited, Essex, England
437. Emsley, J. W., ed. (1996) *Encyclopedia of NMR*, John Wiley & Sons, Ltd., Sussex, England
438. Markley, J. L. (1975) *Acc. Chem. Res.* **8**, 70–80
439. Perkins, S. J. (1982) in *Biological Magnetic Resonance*, Vol. 4 (Berliner, L. J., and Reuben, J., eds), pp. 193-334, Plenum Press, New York and London
440. Osapay, K., and Case, D. A. (1991) *J. Am. Chem. Soc.* **113**, 9436–9444
441. Szilágyi, L. (1995) *Prog. Nucl. Magn. Reson. Spectrosc.* **27**, Part 4, 326–443
442. Williamson, M. P., Kikuchi, J., and Asakura, T. (1995) *J. Mol. Biol.* **247**, 541–546
443. Sitkoff, D., and Case, D. A. (1997) *J. Am. Chem. Soc.* **119**, 12262–12273
444. Gu, Z., Zambrano, R., and McDermott, A. (1994) *J. Am. Chem. Soc.* **116**, 6368–6372
445. Gu, Z., Ridenour, C. F., Bronnimann, C. E., Iwashita, T., and McDermott, A. (1996) *J. Am. Chem. Soc.* **118**, 822–829
446. de Dios, A. C., Pearson, J. G., and Oldfield, E. (1993) *Science* **260**, 1491–1496
447. Sulzbach, H. M., Schleyer, P. V. R., and Schaefer, H. F., III (1994) *J. Am. Chem. Soc.* **116**, 3967–3972
448. Wishart, D. S., Sykes, B. D., and Richards, F. M. (1992) *Biochemistry* **31**, 1647–1651
449. Linderstrøm-Lang, K. (1955) *Chemical Society Special Publication*, No. 2, 1–20
450. Barfield, M., and Karplus, M. (1969) *J. Am. Chem. Soc.* **91**, 1–16
451. Karplus, M. (1963) *J. Am. Chem. Soc.* **85**, 2870–2871
452. Bax, A. (1989) *Ann. Rev. Biochem.* **58**, 223–256
453. Wang, A. C., and Bax, A. (1996) *J. Am. Chem. Soc.* **118**, 2483–2494
454. Stonehouse, J., Adell, P., Keeler, J., and Shaka, A. J. (1994) *J. Am. Chem. Soc.* **116**, 6037–6038
455. Aguayo, J. B., Gamcsik, M. P., and Dick, J. D. (1988) *J. Biol. Chem.* **263**, 19552–19557
456. Ackerman, J. J. H., Ewy, C. S., Kim, S.-G., and Shalwitz, R. A. (1987) *Ann. N.Y. Acad. Sci.* **508**, 89–98
457. Breitmaier, E., and Voelter, W. (1987) *Carbon-13 NMR Spectroscopy*, 3rd ed., VCH Publishers, New York
458. Beckmann, N. (1995) *Carbon-13 NMR Spectroscopy of Biological Systems*, Academic Press, New York
459. Gronenborn, A. M., Clore, G. M., Schmeissner, U., and Wingfield, P. T. (1986) *Eur. J. Biochem.* **161**, 37–43
460. Clore, G. M., Bax, A., Driscoll, P. C., Wingfield, P. T., and Gronenborn, A. M. (1990) *Biochemistry* **29**, 8172–8184
461. Jue, T., Rothman, D. L., Shulman, G. I., Tavitian, B. A., DeFronzo, R. A., and Shulman, R. G. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4489–4491
462. Markley, J. L. (1989) *Methods Enzymol.* **176**, 12–64
463. Pelton, J. G., Torchia, D. A., Meadow, N. D., and Roseman, S. (1993) *Protein Sci.* **2**, 543–558
464. Rajagopal, P., Waygood, E. B., and Klevit, R. E. (1994) *Biochemistry* **33**, 15271–15282
465. Solum, M. S., Altmann, K. L., Strohmeier, M., Berges, D. A., Zhang, Y., Facelli, J. C., Pugmire, R. J., and Grant, D. M. (1997) *J. Am. Chem. Soc.* **119**, 9804–9809
466. Verkade, J. G., and Quin, L. D., eds. (1987) *Phosphorus-31 NMR Spectroscopy in Stereochemical Analysis*, VCH Publishers, Deerfield Beach, Florida
467. Ugurbil, K., Kingsley-Hickman, P. B., Sako, E. Y., Zimmer, S., Mohanakrishnan, P., Robitaille, P. M. L., Thoma, W. J., Johnson, A., Foker, J. E., From, A. H. L. (1987) *Ann. N.Y. Acad. Sci.* **508**, 265–286
468. Feeney, J., McCormick, J. E., Bauer, C. J., Birdsall, B., Moody, C. M., Starkmann, B. A., Young, D. W., Francis, P., Havlin, R. H., Arnold, W. D., and Oldfield, E. (1996) *J. Am. Chem. Soc.* **118**, 8700–8706
469. Lau, E. Y., and Gerig, J. T. (1997) *Biophys. J.* **73**, 1579–1592
470. Jarema, M. A. C., Lu, P., and Miller, J. H. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 2707–2711
471. Slebe, J. C., and Martinez-Carrion, M. (1990) *J. Biol. Chem.* **253**, 2093–2097
472. Briley, P. A., Eisenthal, R., Harrison, R., and Smith, G. D. (1990) *Biochem. J.* **163**, 325–331
473. Roberts, J. K. M., Wade-Jardetzky, N., and Jardetzky, O. (1981) *Biochemistry* **20**, 5389–5394
474. Deutsch, C. J., and Taylor, J. S. (1987) *Ann. N.Y. Acad. Sci.* **508**, 33
475. Gutowsky, H. S., McCall, D. W., McGarvey, B. R., and Meyer, L. H. (1952) *J. Am. Chem. Soc.* **74**, 4809–4817
476. Chang, Y. C., and Graves, D. J. (1985) *J. Biol. Chem.* **260**, 2709–2714
477. Scott, R. D., Chang, Y.-C., Graves, D. J., and Metzler, D. E. (1985) *Biochemistry* **24**, 7668–7681
478. Saunders, M., Jiménez-Vázquez, H. A., Cross, R. J., Mroczkowski, S., Freedberg, D. I., and Anet, F. A. L. (1994) *Nature (London)* **367**, 256–257
479. Adebodun, F., and Jordan, F. (1988) *J. Am. Chem. Soc.* **110**, 309–310
480. Springer, C. S., Jr. (1987) *Ann. Rev. Biophys. Biophys. Chem.* **16**, 375–399
481. Wittenberg, B. A., and Gupta, R. K. (1985) *J. Biol. Chem.* **260**, 2031–2034
482. Jelicks, L. A., and Gupta, R. K. (1989) *J. Biol. Chem.* **264**, 15230–15235
483. Price, W. S., Ge, N.-H., Hong, L.-Z., and Hwang, L.-P. (1993) *J. Am. Chem. Soc.* **115**, 1095–1105
484. Wang, S. M., and Gilpin, R. K. (1983) *Anal. Chem.* **55**, 493–497
485. Markham, G. D. (1986) *J. Biol. Chem.* **261**, 1507–1509
486. Denisov, V. P., and Halle, B. (1995) *J. Mol. Biol.* **245**, 682–687
487. Mullen, G. P., Dunlap, R. B., and Odom, J. D. (1986) *Biochemistry* **25**, 5625–5632
488. Aldhous, P. (1991) *Nature (London)* **353**, 689
489. Becker, E. D., and Farrar, T. C. (1972) *Science* **178**, 361–368
490. Bax, A., and Lerner, L. (1986) *Science* **232**, 960–967
491. Yamazaki, T., Muhandiram, R., and Kay, L. E. (1994) *J. Am. Chem. Soc.* **116**, 8266–8278
492. Farrar, T. C., and Becker, E. D. (1971) *Pulse and Fourier Transform NMR*, Academic Press, New York
493. Wüthrich, K. (1989) *Science* **243**, 45–50
494. van de Ven, F. J. M., and Hilbers, C. W. (1986) *J. Mol. Biol.* **192**, p. 419-441
495. Baniak, E. L., Rivier, J. E., Struthers, R. S., Hagler, A. T., and Gierasch, L. M. (1987) *Biochemistry* **26**, 2642–2656
496. Wüthrich, K. (1990) *J. Biol. Chem.* **265**, 22059–22062
497. Bax, A., and Grzesiek, S. (1993) in *NMR of Proteins* (Clore, G. M., and Gronenborn, A. M., eds), pp. 33–52, CRC Press, Boca Raton, Florida
498. Clore, G. M., and Gronenborn, A. M. (1994) *Protein Sci.* **3**, 372–390
499. Hoitnik, C. W. G., Driscoll, P. C., Hill, H. A. O., and Canters, G. W. (1994) *Biochemistry* **33**, 3560–3571
500. Yamazaki, T., Tochio, H., Furui, J., Aimoto, S., and Kyogoku, Y. (1997) *J. Am. Chem. Soc.* **119**, 872–880
501. Yamazaki, T., Lee, W., Arrowsmith, C. H., Muhandiram, D. R., and Kay, L. E. (1994) *J. Am. Chem. Soc.* **116**, 11655–11666
502. Hu, J.-S., and Bax, A. (1997) *J. Am. Chem. Soc.* **119**, 6360–6368
503. Song, S., Velde, D. V., Gunn, C. W., and Himes, R. H. (1994) *Biochemistry* **33**, 693–698
504. Barsukov, I. L., Lian, L.-Y., Ellis, J., Sze, K.-H., Shaw, W. V., and Roberts, G. C. K. (1996) *J. Mol. Biol.* **262**, 543–558
505. Wu, W.-J., Anderson, V. E., Raleigh, D. P., and Tonge, P. J. (1997) *Biochemistry* **36**, 2211–2220
506. Metzler, W. J., Hare, D. R., and Pardi, A. (1989) *Biochemistry* **28**, 7045–7052
507. McDowell, L. M., Lee, M., McKay, R. A., Anderson, K. S., and Schaefer, J. (1996) *Biochemistry* **35**, 3328–3334
508. Pflugrath, J. W., Wiegand, G., Huber, R., and Vértesy, L. (1986) *J. Mol. Biol.* **189**, 383–386
509. Kline, A. D., Braun, W., and Wüthrich, K. (1986) *J. Mol. Biol.* **189**, 377–382
510. McDonald, C. C., and Phillips, W. D. (1969) *J. Am. Chem. Soc.* **91**, 1513–1521
511. Nietlispach, D., Clowes, R. T., Broadhurst, R. W., Ito, Y., Keeler, J., Kelly, M., Ashurst, J., Oschkinat, H., Domaille, P. J., and Laue, E. D. (1996) *J. Am. Chem. Soc.* **118**, 407–415
512. Rosen, M. K., Gardner, K. H., Willis, R. C., Parris, W. E., Pawson, T., and Kay, L. E. (1996) *J. Mol. Biol.* **263**, 627–636
513. Venters, R. A., Farmer, B. T., II, Fierke, C. A., and Spicer, L. D. (1996) *J. Mol. Biol.* **264**, 1101–1116
514. LeMaster, D. M., and Richards, F. M. (1988) *Biochemistry* **27**, 142–150
515. Bax, A., Kay, L. E., Sparks, S. W., and Torchia, D. A. (1989) *J. Am. Chem. Soc.* **111**, 408–409
516. Torchia, D. A., Sparks, S. W., and Bax, A. (1989) *Biochemistry* **28**, 5509–5524
517. Lowry, D. F., Redfield, A. G., McIntosh, L. P., and Dahlquist, F. W. (1988) *J. Am. Chem. Soc.* **110**, 6885–6886
518. Stockman, B. J., Reily, M. D., Westler, W. M., Ulrich, E. L., and Markley, J. L. (1989) *Biochemistry* **28**, 230–236
519. Qian, H., Mayo, S. L., and Morton, A. (1994) *Biochemistry* **33**, 8167–8171

519a. Riek, R., Pervushin, K., and Wüthrich, K. (2000) *Trends Biochem. Sci.* **25**, 462–468

520. Günter, P., Braun, W., Billeter, M., and Wüthrich, K. (1989) *J. Am. Chem. Soc.* **111**, 3997–4004
521. Oh, B.-H., Westler, W. M., and Markley, J. L. (1989) *J. Am. Chem. Soc.* **111**, 3083–3085
522. Cross, T. A., and Opella, S. J. (1983) *J. Am. Chem. Soc.* **105**, 306–308
523. Rothgeb, T. M., and Oldfield, E. (1981) *J. Biol. Chem.* **256**, 1432–1446
524. Buck, M., Schwalbe, H., and Dobson, C. M. (1995) *Biochemistry* **34**, 13219–13232
525. Nicholson, L. K., Kay, L. E., Baldisseri, D. M., Arango, J., Young, P. E., Bax, A., and Torchia, D. A. (1992) *Biochemistry* **31**, 5253–5263
526. Gardner, K. H., and Kay, L. E. (1997) *J. Am. Chem. Soc.* **119**, 7599–7600
527. Lenkinski, R. E., Dallas, J. L., and Glickson, J. D. (1979) *J. Am. Chem. Soc.* **101**, 3071–3077

## References


528. Carver, J. A., and Bradbury, J. H. (1984) *Biochemistry* **23**, 4890–4905
529. Bothelho, L. H., and Gurd, F. R. N. (1978) *Biochemistry* **17**, 5188–5196
530. Bothelho, L. H., Friend, S. H., Matthew, J. B., Lehman, L. D., Hanania, G. I. H., and Gurd, F. R. N. (1978) *Biochemistry* **17**, 5197–5205
531. Morino, Y., Yamasaki, M., Tanase, S., Nagashima, F., Akasaka, K., Imoto, T., and Miyazawa, T. (1984) *J. Biol. Chem.* **259**, 3877–3882
532. Cocco, M. J., Kao, Y.-H., Phillips, A. T., and Lecomte, J. T. J. (1992) *Biochemistry* **31**, 6481–6491
533. Bachovin, W. W., Wong, W. Y. L., Farr-Jones, S., Shenvi, A. B., and Kettner, C. A. (1988) *Biochemistry* **27**, 7689–7697
534. Mollova, E. T., Metzler, D. E., Kintanar, A., Kagamiyama, H., Hayashi, H., Hirotsu, K., and Miyahara, I. (1997) *Biochemistry* **36**, 615–625
535. Guéron, M., Plateau, P., and Decorps, M. (1991) *Progr. in NMR Spectr.* **23**, 135–209
536. Plateau, P., and Guéron, M. (1982) *J. Am. Chem. Soc.* **104**, 7310–7311
537. Sklenar, V., and Bax, A. (1987) *J. of Magnetic Resonance* **74**, 469–479
538. Bullitt, E., and Makowski, L. (1995) *Nature (London)* **373**, 164–167
539. Griffin, J. H., Cohen, J. S., and Schechter, A. N. (1973) *Biochemistry* **12**, 2096–2099
540. Tong, H., and Davis, L. (1995) *Biochemistry* **34**, 3362–3367
541. Gunnarsson, G., Wennerström, H., Egan, W., and Forsén, S. (1976) *Chem. Phys. Lett.* **38**, 96–99
542. McDermott, A., and Ridenour, C. F. (1996) in *Encyclopedia of NMR* (Emsley, J. W., ed), pp. 3820–3824, J. Wiley & Sons, Ltd., Sussex, UK
543. Zhao, Q., Abeygunawardana, C., Gittis, A. G., and Mildvan, A. S. (1997) *Biochemistry* **36**, 14616–14626
544. Perrin, C. L., and Nielson, J. B. (1997) *J. Am. Chem. Soc.* **119**, 12734–12741
545. Tjandra, N., and Bax, A. (1997) *J. Am. Chem. Soc.* **119**, 8076–8082
546. Kintanar, A., Metzler, C. M., Metzler, D. E., and Scott, R. D. (1991) *J. Biol. Chem.* **266**, 17222–17229
547. Metzler, C. M., Metzler, D. E., Kintanar, A., Scott, R. D., and Marceau, M. (1991) *Biochem. Biophys. Res. Commun.* **178**, 385–392
548. Firsov, L. M., Neustroev, K. N., Aleshin, A. E., Metzler, C. M., Metzler, D. E., Scott, R. D., Stoffer, B., Christensen, T., and Svensson, B. (1994) *Eur. J. Biochem.* **223**, 293–302
549. Metzler, D. E., Metzler, C. M., Mollova, E. T., Scott, R. D., Tanase, S., Kogo, K., Higaki, T., and Morino, Y. (1994) *J. Biol. Chem.* **269**, 28017–28026
550. Metzler, D. E., Metzler, C. M., Scott, R. D., Mollova, E. T., Kagamiyama, H., Yano, T., Kuramitsu, S., Hayashi, H., Hirotsu, K., and Miyahara, I. (1994) *J. Biol. Chem.* **269**, 28027–28033
551. Metzler, D. E. (1997) *Methods Enzymol.* **280**, 30–40
552. Englander, S. W., Mayne, L., Bai, Y., and Sosnick, T. R. (1997) *Protein Sci.* **6**, 1101–1109
553. Kossiakoff, A. A. (1982) *Nature (London)* **296**, 713–721
554. Rohl, C. A., and Baldwin, R. L. (1994) *Biochemistry* **33**, 7760–7767
555. Gemmecker, G., Jahnke, W., and Kessler, H. (1993) *J. Am. Chem. Soc.* **115**, 11620–11621
556. Zhang, Y.-Z., Paterson, Y., and Roder, H. (1995) *Protein Sci.* **4**, 804–814
557. Zhang, Z., and Smith, D. L. (1993) *Protein Sci.* **2**, 522–531
558. Wagner, D. S., Melton, L. G., Yan, Y., Erickson, B. W., and Anderegg, R. J. (1994) *Protein Sci.* **3**, 1305–1314
559. Zhang, Z., Post, C. B., and Smith, D. L. (1996) *Biochemistry* **35**, 779–791
560. Zappacosta, F., Pessi, A., Bianchi, E., Venturini, S., Sollazzo, M., Tramontano, A., Marino, G., and Pucci, P. (1996) *Protein Sci.* **5**, 802–813
561. Kim, K.-S., Fuchs, J. A., and Woodward, C. K. (1993) *Biochemistry* **32**, 9600–9608
562. Eriksson, M. A. L., Härd, T., and Nilsson, L. (1995) *Biophys. J.* **69**, 329–339
563. Spyracopoulos, L., and O'Neil, J. D. J. (1994) *J. Am. Chem. Soc.* **116**, 1395–1402
564. Markley, J. L. (1975) *Biochemistry* **14**, 3546–3554
565. Schmidt-Rohr, K., and Spiess, H. W. (1994) *Multidimensional Solid-State NMR and Polymers*, Academic Press, San Diego, California
566. Schmidt-Rohr, K. (1996) *J. Am. Chem. Soc.* **118**, 7601–7603
567. Wu, C. H., Ramamoorthy, A., Gierasch, L. M., and Opella, S. J. (1995) *J. Am. Chem. Soc.* **117**, 6148–6149
568. Meyer, E. F. (1997) *Protein Sci.* **6**, 1591–1597
569. Stampf, D. R., Felder, C. E., and Sussman, J. L. (1995) *Nature (London)* **374**, 572–573
570. Holm, L., and Sander, C. (1995) *Trends Biochem. Sci.* **20**, 478–480
571. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540
572. Walsh, L., Newell, M., Ruczaj, H., Reid, T., and Desmond, P., eds. (1997) *Trends Guide to the Internet*, Elsevier, Cambridge, UK
573. Hogue, C. W. V., Ohkawa, H., and Bryant, S. H. (1996) *Trends Biochem. Sci.* **21**, 226–229
574. Wilkins, M. R., Hochstrasser, D. F., Sanchez, J.-C., Bairoch, A., and Appel, R. D. (1996) *Trends Biochem. Sci.* **21**, 496–497
575. Peitsch, M. C., Wells, T. N. C., Stampf, D. R., and Sussman, J. L. (1995) *Trends Biochem. Sci.* **20**, 82–84
576. Sayle, R. A., and Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376

576a. Bernstein, H. J. (2000) *Trends Biochem. Sci.* **25**, 453–454

577. Sokolik, C. W. (1995) *Trends Biochem. Sci.* **20**, 122–124
578. Hunt, T. (1996) *Trends Biochem. Sci.* **21**, 74
579. Service, R. F. (2000) *Science* **287**, 1954–1956
580. Laskowski, R. A., Hutchinson E. G., Michie, A. D., Wallace, A. C., Jones, M. L., and Thornton, J. M. (1997) *Trends Biochem. Sci* **22**, 488–490
581. Cohn, E. J., and Edsall, J. T. (1943) *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, Van Nostrand-Reinhold, Princeton, New Jersey (pp. 90–93)
582. Abbott, A. (2000) *Nature (London)* **408**, 130–132
583. Simmerling, C., Lee, M. R., Ortiz, A. R., Kolinski, A., Skolnick, J., and Kollman, P. A. (2000) *J. Am. Chem. Soc.* **122**, 8392–8402

## Study Questions

1. a. From the expression for the dissociation constant of an acid, HA (Eq. 3-1), derive the logarithmic form $pH - pK_a = \log[A^-] / [HA] = \log_{10} \alpha/(1-\alpha)$ (Eq. 3-11), where $\alpha$ is the fraction of the acid in the ionized form.

   b. The apparent dissociation constant $K_a$ for the $H_2PO_4^-$ ion at 25°C and 0.5 M total phosphate concentration is $1.380 \times 10^{-7}$ M. What will be the pH of a solution 0.025 M in $KH_2PO_4$ and 0.25 M in $Na_2HPO_4$? This is a National Bureau of Standards buffer (see Bates[7]).

   c. Suppose that you wanted to prepare a buffer of pH = 7.00 at 25°C from anhydrous $KH_2PO_4$ ($M_r$ = 136.09) and $Na_2HPO_4$ ($M_r$ = 141.98). If you placed 3.40 g of $KH_2PO_4$ in a 1 liter volumetric flask, how much anhydrous $Na_2HPO_4$ would you have to weigh and add before making to volume to obtain the desired pH? If you wanted to have the correct pH to ± 0.01 unit, how accurately would you have to weigh your salts? NOTE: It is quicker to prepare a buffer of precise pH this way than it is to titrate a portion of buffer acid to the desired pH with sodium hydroxide.

2. The apparent $pK_a$ for 0.1 M formic acid is 3.70 at 25°C.

   a. Concentrated HCl was added to a liter of 0.1 M sodium formate until a pH of 1.9 was attained. Calculate the concentration of formate ion and that of unionized formic acid.

   b. Calculate the hydrogen ion concentration.

   c. How many equivalents of HCl had to be added in part a to bring the pH to 1.9?

3. Exactly 0.01-mol portions of glycine were placed in several 100-ml volumetric flasks. The following exact amounts of HCl or NaOH were added to the flasks, the solutions were made to volume with water, mixed, and the pH measured. Calculate the $pK_a$ values for the carboxyl and amino groups from the following, making as many independent calculations of each p*K*a as the data permit. At low pH values *you must correct for the free hydrogen ion concentration* (see question 2c).[581]

| Flask No. | Mol HCl | Mol NaOH | pH |
|---|---|---|---|
| 1 | 0.010 | | 1.71 |
| 2 | 0.009 | | 1.85 |
| 3 | 0.006 | | 2.25 |
| 4 | 0.002 | | 2.94 |
| 5 | | 0.002 | 9.00 |
| 6 | | 0.004 | 9.37 |
| 7 | | 0.005 | 9.60 |

4. Using the $pK_a$ values from problem 3, construct the theoretical titration curve showing the equivalents of $H^+$ or $OH^-$ reacting with 1 mol of glycine as a function of pH. Note that the shape of this curve is independent of the $pK_a$. Sketch similar curves for glutamic acid ($pK_a$'s equal 2.19, 4.25, and 9.67), histidine ($pK_a$'s equal 1.82, 6.00, and 9.17) and lysine ($pK_a$'s equal 2.18, 8.95, and 10.53).

   Compare your plot for glycine with a plot of 1 M acid or base added to 0.01 mol of glycine in 100 ml of water. You may also compare your curves with those for glycine published in other textbooks.

5. Make a table of characteristic $pK_a$ values for acidic and basic groups in proteins. Which of these groups contribute most significantly to the titration curves of proteins?

6. If placed in water and adjusted to a pH of 7, will the following migrate toward the anode or the cathode if placed in an electrical field? (a) Aspartic acid, (b) alanine, (c) tyrosine, (d) lysine, (e) arginine, and (f) glutamine

7. The tripeptide L-Ala – L-His – L-Gln had the following $pK_a$ values: 3.0 ($\alpha$-COOH), 9.1 ($\alpha$-$NH_3^-$), and 6.7 (imidazolium).

   a. What is the isoelectric pH (pI) of the peptide, i.e., the pH at which it will carry no net charge? Hint, the pI for amino acids is usually given approximately as the arithmetic mean of two $pK_a$ values.[581]

   b. Draw the structures of the ionic forms of the peptide that occur at pH 5 and at pH 9. At each pH compute the fraction of the peptide in each ionic form.

## Study Questions

8. a. Write the structure for glycyl-L-tryptophanyl-L-prolyl-L-seryl-L-lysine.

   b. What amino acids could be isolated from it following acid hydrolysis?

   c. Following alkaline hydrolysis?

   d. After nitrous acid treatment followed by acid hydrolysis

   e. In an electrolytic cell at pH 7.0 would the peptide migrate toward the cathode or toward the anode? What is the approximate isoelectric point of the peptide?

   f. If a solution of this peptide were adjusted to pH 7, and then titrated with sodium hydroxide in the presence of 10% formaldehyde, how many equivalents of base would be required per mole of peptide to raise the pH to 10?

9. A peptide was shown to contain only L-lysine and L-methionine. Titration of the peptide showed 3 free amino groups for each free carboxyl group present, and each amino group liberated 1 mole of $N_2$ when the peptide was treated with $HNO_2$ in the Van Slyke apparatus. When the deaminated peptide was hydrolyzed completely in acid and the hydrolyzate again treated with $HNO_2$, the same amount of $N_2$ is liberated as that derived from the intact peptide. A sample of the original peptide was treated with excess dinitrofluorobenzene to give a dinitrophenyl (DNP) peptide, which was shown spectrophotometrically to contain three DNP groups per free carboxyl group. When this DNP-peptide was completely hydrolyzed, the following products were found: a colorless compound containing S ($A_1$) and a yellow compound containing S ($A_3$). Partial hydrolysis of the DNP-peptide yields $A_1$, $A_2$, $A_3$, plus 4 additional yellow compounds, $B_1$, $B_2$, $B_3$, and $B_4$. On complete hydrolysis, $B_1$ yields $A_1$, $A_2$, and $A_3$; $B_2$ yields $A_1$ and $A_2$; $B_3$ yields $A_1$ and $A_3$; and $B_4$ yields $A_3$ only.

   What is the most probable structure of the original peptide?

10. (a) Explain two advantages of the isotope dilution method of analysis. (b) From the following data, calculate the amount of cyclic AMP (cAMP) present per ml of human gluteus maximus muscle cells. Cells were treated with $^{32}$P-enriched cAMP, S.A. = 50 μCi / μmol for 0.2 h (all cAMP was taken up by cells), cells were homogenized, and the soluble cAMP was isolated and purified. The specific activity of the isolated cAMP = 10 μCi / μmol. The total amount of cAMP added was $1.0 \times 10^{-7}$ mol per ml of cells.

11. The figure in Box 3-C shows a high-resolution separation of the soluble proteins of *E. coli*. The investigator labeled the proteins with $^{14}$C-containing amino acids.

   a. How would you carry out the labeling experiment?

   b. What other isotope(s) could be used to label proteins? What chemical form(s) would you use? What limitations might there be?

   c. What soluble components of an *E. coli* cell sonicate might interfere with the two-dimensional separation, and how could they be removed?

   d. What technique(s) other than radioactive labeling could be used for locating proteins?

   e. If *all* the soluble proteins of *E. coli* were detected, about how many separate proteins would you expect to see?

   f. Indicate two or more properties of the resolution technique which are most significant in making it applicable to a system containing a very large number of proteins.

12. $^{35}$S is a beta emitter, with no gamma or other type of radiation. It has the following properties: $t_{1/2}$ = 86.7 days, $\varepsilon_{max}$ = 0.168 MeV.

   a. Write the equation for the radiochemical decomposition of $^{35}$S.

   b. Discuss the advantages and limitation of the use of $^{35}$S as an isotopic tracer.

Each cotton fiber is a single cell seed hair, ~30 mm in length. The dry fiber is ~95% cellulose, which constitutes the secondary cell wall (See Fig. 20-4,D) and is also present in the primary cell wall. The fibers, which are ~30 nm in length, consist of many parallel chains (Fig. 4-5) ~5 µm in length, each containing ~10,000 glucose residues. Van der Waals forces, together with one hydrogen bond per glucose, contribute to the stability of the tightly packed fiber. Cotton is one of the major agricultural crops, ~87 million bales, each ~220 kg, being produced annually world-wide. Photo and information from A. D. French and M. A. Godshall, Southern Regional Research Center, USDA, New Orleans, LA.

# Contents

# Sugars, Polysaccharides, and Glycoproteins

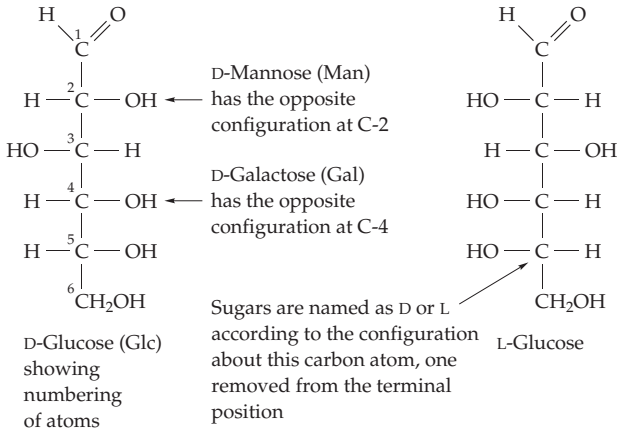
# 4

We are all familiar with sugars, important components of our diet which are present in fruits, honey, table sugar, and syrups. Within our bodies the simple sugar **D-glucose** is an essential source of energy. It is present in blood at a relatively constant concentration of 5.5 mM and is carried to all tissues. A **polysaccharide** called **glycogen**, a polymeric form of glucose, provides a reserve of readily available energy within our cells. **Starches** and other polysaccharides store energy within plants. Polysaccharides also have major *structural* functions in nature. **Cellulose**, another polymer of glucose, forms the fibers of cotton, plant cell walls, and wood. Both the tough exoskeletons of anthropods and the cell walls of fungi depend for their strength on the nitrogen-containing polysaccharide **chitin**. Polysaccharides that carry many negative charges, such as **hyaluronan,** form a protective layer between animal cells, while **pectins** play a similar role in plants.

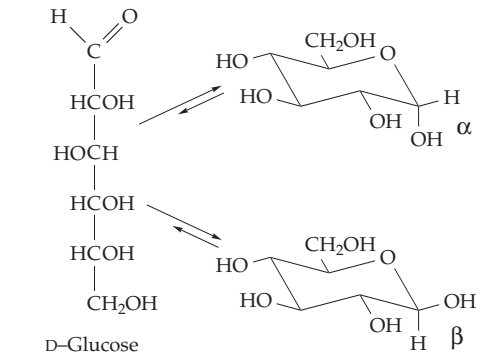
A third function of sugar residues (**glycosyl groups**) is in biological recognition and communication. The outer surfaces of cells are nearly covered by covalently attached **oligosaccharides**, small polymeric arrays of sugar rings. Some are attached to side chain groups of proteins and others to lipids to form **glycoproteins** and **glycolipids,** respectively. Because of the variety of different sugars, the various ways in which they can be linked, and their ability to form oriented hydrogen bonds, these oligosaccharides provide much of the chemical code for identifying cells. This coding enables cells to attach to each other in correct ways during development of a multicelled organism. It helps to activate our immune system to attack parasites and also helps bacteria to attack us! We are only beginning to understand the numerous critical functions of the glycosyl groups of cell surfaces.

## A. Structures and Properties of Simple Sugars

The simple sugars, or **monosaccharides**, are polyhydroxyaldehydes (**aldoses**) or polyhydroxyketones (**ketoses**).[1–5] All have the composition $(CH_2O)_n$, hence the family name **carbohydrate**. A typical sugar, and the one with the widest distribution in nature, is glucose.




The carbonyl group of this and other sugars is highly reactive and a characteristic reaction is **addition** of electron-rich groups such as –OH. If a sugar chain is long enough (4–6 carbon atoms) one of the hydroxyl groups of the same molecule can add to the carbonyl group to form a cyclic **hemiacetal** or ring form, which reaches an equilibrium with the free aldehyde or ketone form (Eq. 4-1). The six-membered rings formed in this way (**pyranose** rings) are especially stable, but five-membered **furanose** rings also exist in many carbohydrates.

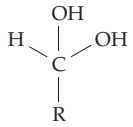



Pyranose ring forms (hemiacetal forms)

D–Glucose (free aldehyde form)

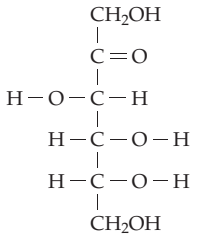
(4-1)

The ring forms of the sugars are the monomers used by cells to form polysaccharides. Indeed, it is the natural tendency of 5- and 6-carbon sugars (**pentoses** and **hexoses**) to cyclize that permits formations of stable sugar polymers from the reactive and unstable monomers. When a sugar cyclizes a new chiral center is formed at the **anomeric carbon atom**, the atom that was present in the original carbonyl group. The two configurations about this carbon atom are designated α and β as indicated in Eq. 4-1. In an equilibrium mixture, *ring forms of most sugars predominate over open chains.*[6–13] Thus, at 25°C in water glucose reaches an




Covalent hydrate of R—CHO

equilibrium with ~0.001% free aldehyde, 0.004% covalent hydrate of the aldehyde (see also Eq. 13-1), 39% α-pyranose form, 61% β-pyranose form, and 0.15% each of the much less stable α and β furanose forms.[9,12] The ketose sugar **fructose** in solution exists as ~73% β-pyranose, 22% β-furanose, 5% α-furanose, and 0.5% open chain form (Fig. 4-1).[6,10,13] Although polysaccharides are composed almost exclusively of sugar residues in ring forms, the open chain forms are sometimes metabolic intermediates.




D-Fructose (Fru), a ketose that is a close structural and metabolic relative of D-glucose. It occurs in honey and fruit juices in free form, in the disaccharide sucrose (table sugar) as a 5-membered furanose ring, and in other oligosaccharides and polysaccharides.

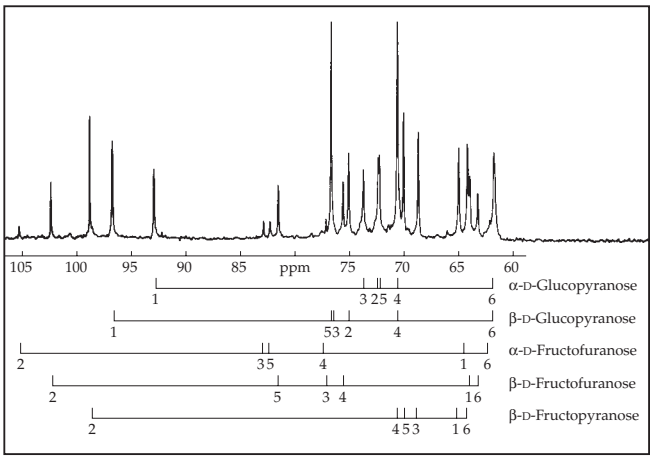



**Figure 4-1** Natural abundance $^{13}$C-NMR spectrum of honey showing content of α and β pyranose ring forms of glucose and both pyranose and furanose ring forms of fructose. Open chain fructose, with a peak at 214 ppm, was present in a trace amount. From Prince *et al.*[6]

Because of their many polar hydroxyl groups, most sugars are very soluble in water. However, hydrogen bonds between molecules stabilize sugar crystals making them insoluble in nonpolar solvents. Intermolecular hydrogen bonds between chains of sugar rings in cellulose account for much of the strength and insolubility of these polysaccharides.

## 1. The Variety of Monosaccharides

Sugars contain several chiral centers and the various diastereomers are given different names. The commonly occurring sugars **D-glucose**, **D-mannose**, and **D-galactose** are just three of the 16 diastereomeric aldohexoses. The Fischer projection formulas for the entire family of eight D-aldoses with 3–6 carbon atoms are given in Fig. 4-2. Several of these occur only rarely in nature.

Monosaccharides are classified as D or L according to the configuration at the chiral center farthest from the carbonyl group. If the –OH group attached to this carbon atom lies to the right when the sugar is oriented according to the Fischer convention, the sugar belongs to the D family. The simplest of all the chiral sugars is glyceraldehyde. The family of aldoses in Fig. 4-2 can be thought of as derived from D-glyceraldehyde by upward extension of the chain. Besides D-glucose, D-mannose, and D-galactose, the most abundant naturally occurring sugar shown in Fig. 4-2 is **D-ribose**, a major component of RNA. Another abundant aldose is **D-xylose**, a constituent of the polysaccharides of wood.

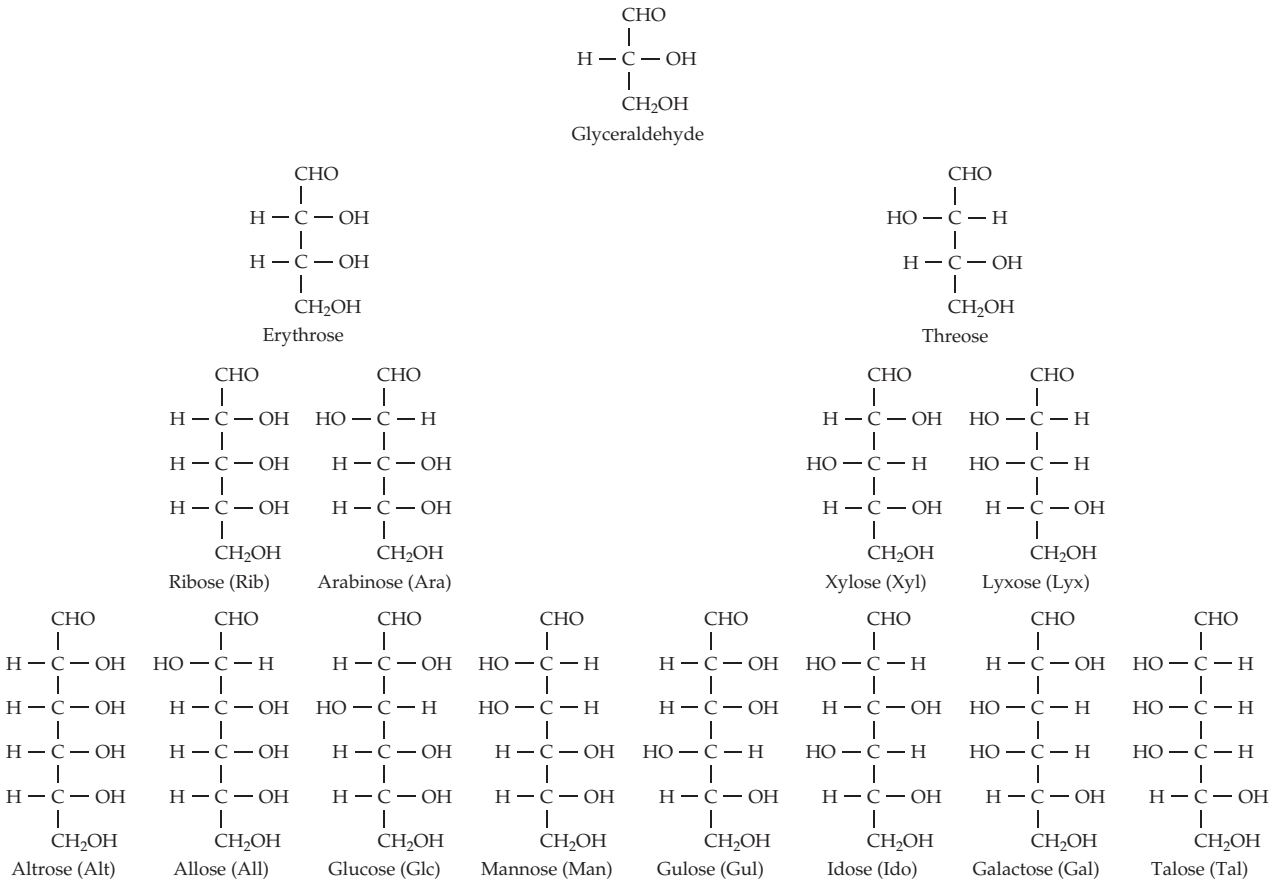
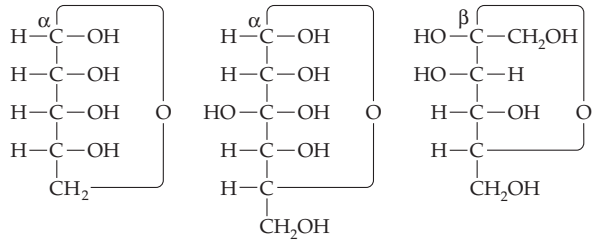



**Figure 4-2** Formulas for the D-aldoses. Prefixes derived from the names of these aldoses are used in describing various other sugars including ketoses. The prefixes include *erythro*, *threo*, *arabino*, *ribo*, *galacto*, *manno*, etc., D-Fructose can be described in this manner as D-*arabino*-hexulose. The prefixes refer to the configurations of a series of consecutive but not necessarily contiguous chiral centers.[2] Thus, 3-*arabino*-hexulose is an isomer of fructose with the carbonyl group at the 3 position. The prefix deoxy, which means "lacking oxygen," is often used to designate a modified sugar in which an –OH has been replaced by –H, e.g., 2-deoxyribose.

For each D sugar there is an L sugar which is the complete mirror image or enantiomer of the D form. Although **L-arabinose** occurs widely in plants, and some derivatives of L sugars are present in glycoproteins, most naturally occurring sugars have the D configuration. A pair of sugars, such as glucose and galactose, differing in configuration at only one of the chiral centers are known as **epimers**. The D-ketose sugars with the carbonyl group in the 2 position (Fig. 4-3) are also abundant in nature, often occurring in the form of phosphate esters as intermediates in metabolism.

***Ways of indicating configuration.*** The Fischer projection formulas used in Figs. 4-2 and 4-3 are convenient in relating the sugar structures by their individual carbon configurations to each other, but they give an unrealistic three-dimensional picture. According to the Fischer convention each carbon atom must be viewed with both vertical bonds projecting behind the atom viewed. In fact, the molecule cannot assume such a conformation because the chain folds back on itself, bringing many atoms into collision.

Ring forms of sugars are also often drawn according to the Fischer convention; making use of elongated bent lines to represent ordinary simple bonds:




In this representation the hydroxyl at the anomeric

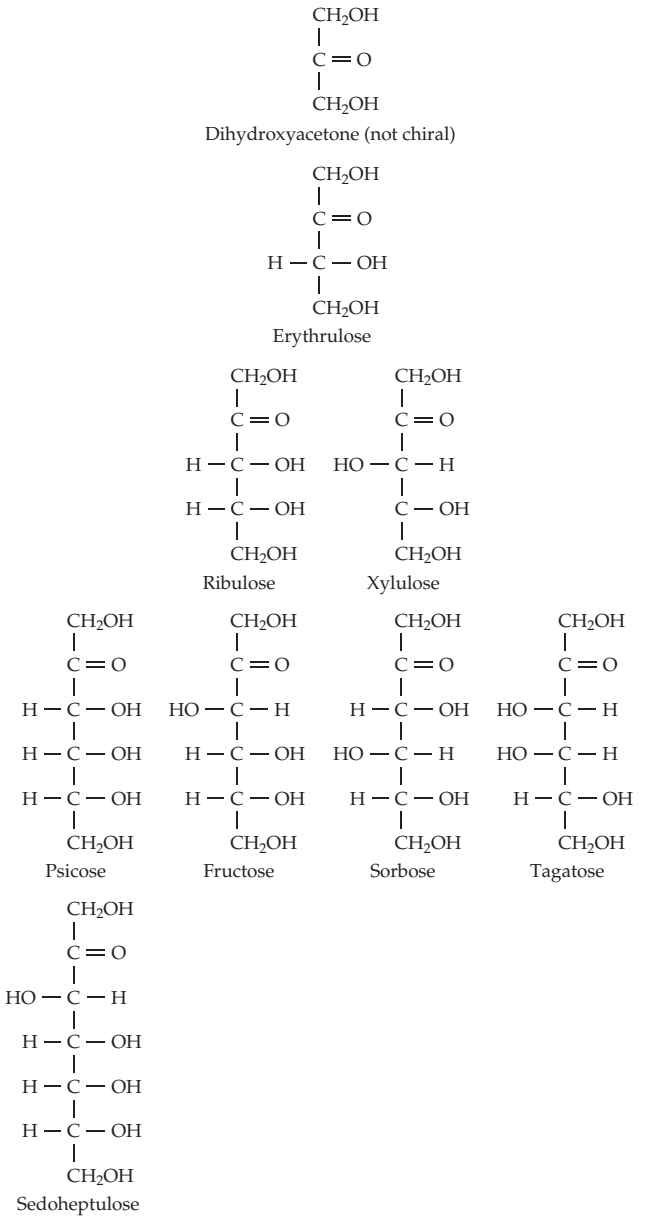
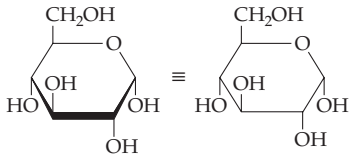



**Figure 4-3** Formulas for the open forms of the D-ketoses.

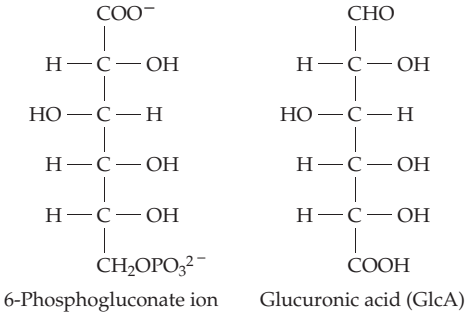
carbon atom is always on the right side of the molecules for α forms in the D series of sugars and at the left side for the β forms. For the L sugars the opposite is true. For example, since α-D-glucopyranose and α-L-glucopyranose are enantiomers, their Fischer formulas must be mirror images.

Simplified sugar rings are often drawn with **Haworth structural formulas**. The lower edge of the ring, which may be shown as a heavy line, is thought of as projecting out toward the reader and the other edge as projecting behind the plane of the paper.




Haworth structures are easy to draw and unambiguous in depicting configurations,[14] but they also do not show the spatial relationships of groups attached to other rings correctly. For this reason conformational formulas of the type described in Section 2 and shown in Fig. 4-4 are used most often in this book.

***Natural derivatives of sugars.*** The aldehyde group of an aldose can be oxidized readily to a carboxyl group to form an **aldonic acid**. Among the several aldonic acids that occur naturally is 6-phosphogluconic acid, which is pictured here as the 6-phosphogluconate ion:




The enzymatically formed **uronic acids** have –COOH in the terminal position. This is position 6 in **glucuronic acid**, whose structure is given here and as a ring in Fig. 4-4. Sugar chains with –COOH at both ends are called **aldaric acids**, e.g., glucaric acid. The –OH group in the 2 position of glucose may be replaced by $-NH_2$ to form 2-amino-2- deoxyglucose, commonly called **glucosamine** (GlcN), or by $-NH-CO-CH_3$ to form ***N*-acetylglucosamine** (GlcNAc). Similar derivatives of other sugars exist in nature. In many poly-saccharides, sulfate groups are attached in ester linkage to the sugar units. The sulfo ($-SO_3^-$) sugar **6-sulfo-α-D-quinovose** (Fig. 4-4) is found in lipids of photosynthetic membranes.[15,16]

The sugar alcohols, in which the carbonyl group has been reduced to –OH, also occur in nature. For example, **D-glucitol** (D-sorbitol), the sugar alcohol obtained by reducing either D-glucose or L-sorbose (Eq. 4-2), is a major product of photosynthesis and widely distributed in bacteria and throughout the eukaryotic kingdom. It is present in large amounts in berries of the mountain ash and in many other fruits. It exists in a high concentration in human semen and

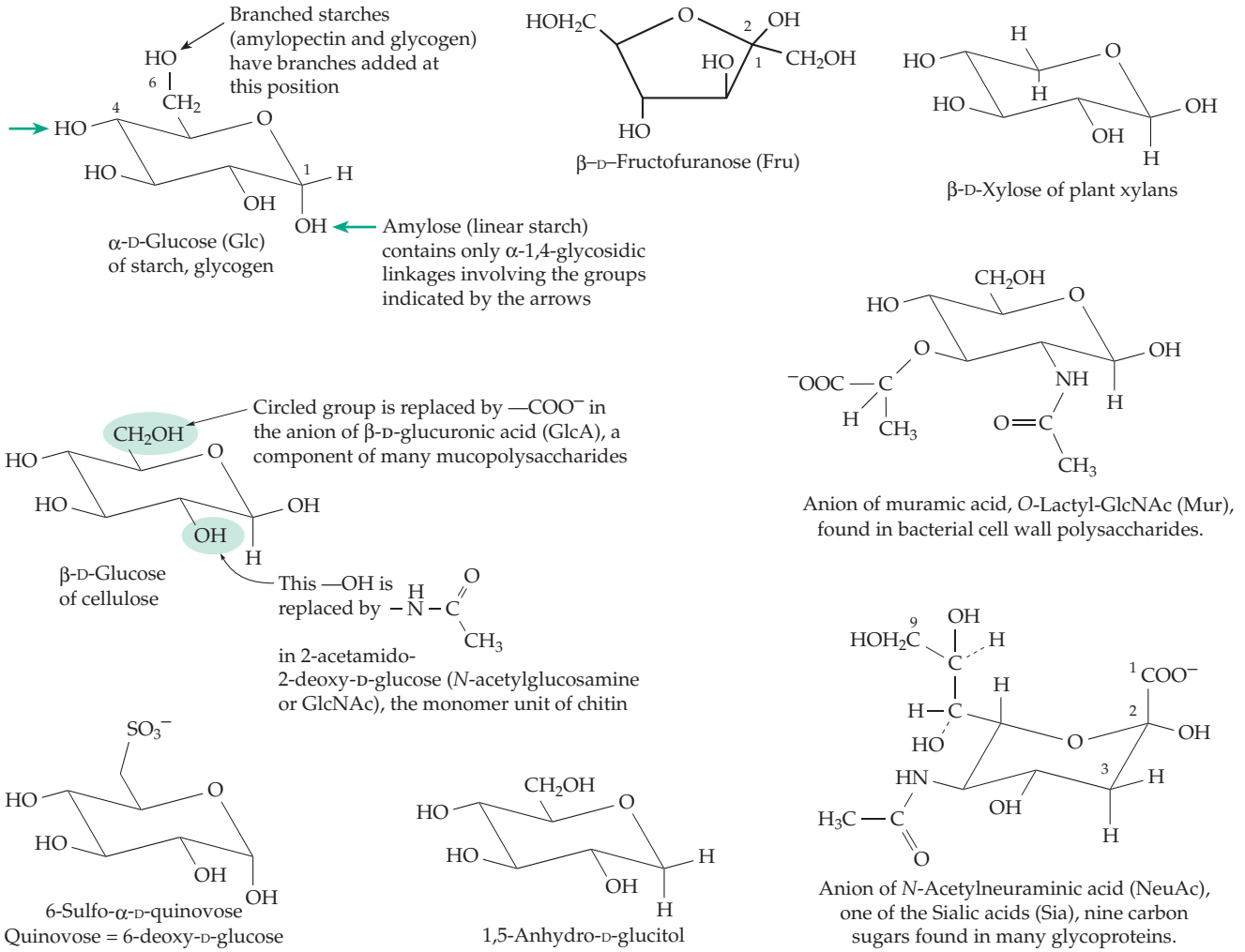
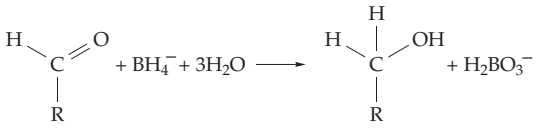
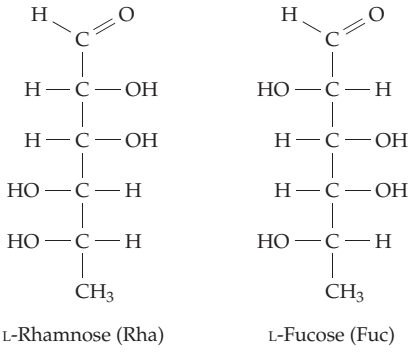



**Figure 4-4** Some simple sugars and sugar derivatives in ring forms. Most of these are present in polysaccharides.



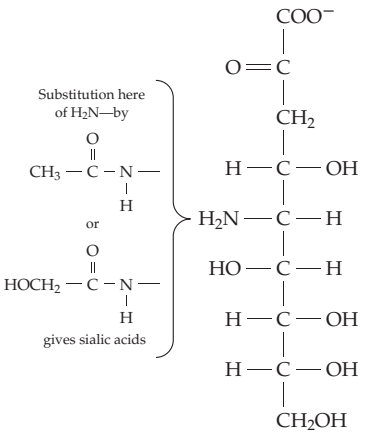

(4-2)




accumulates in lenses of diabetics. D-Glucitol and other sugar alcohols arise in some fungi during metabolism of the corresponding sugars.[17] Mannitol, another product of photosynthesis, is also present in many organisms.[18] Another polyol that is found in human blood in significant concentrations (0.06–0.25 mM) is 1,5-anhydro-D-glucitol (Fig. 4-4). It is largely of dietary origin.[19]

Two common 6-deoxy sugars which lack the hydroxyl group at C-6 are **rhamnose** and **fucose**. Both are of the "unnatural" L configuration but are derived metabolically from D-glucose and D-mannose, respectively.

Vitamin C (ascorbic acid, Box 18-D) is another important sugar derivative. **Neuraminic acid** is a 9-carbon sugar made by transferring a 3-carbon piece onto a hexosamine. Its *N*-acetyl and *N*-glycolyl derivatives are called **sialic acids**. Their names may be abbreviated NeuAc and NeuGl, respectively, or simply as Sia (see also Fig. 4-4).

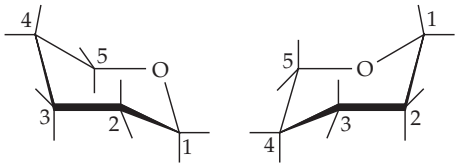



Neuraminic acid anion (Neu)

The sialic acids are prominent constituents of the glycoproteins of cell surfaces. More than 30 modified forms, for example, with added methyl or acetyl groups, are known.[20–24]

## 2. Conformations of Five- and Six-Membered Ring Forms

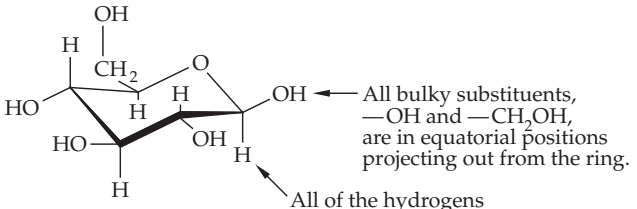
Single-bonded six-membered molecular rings, such as those in cyclohexane and in sugars, most often assume a **chair (C)** conformation. For sugars in pyranose ring forms, there are two possible chair conformations,[4,25,26] which are designated $^4C_1$(or C1) and $^1C_4$ (or 1C). The superscript and subscript numbers on the designations indicate which atoms are above and



$^4C_1$ (or C1) $^1C_4$ (or 1C)

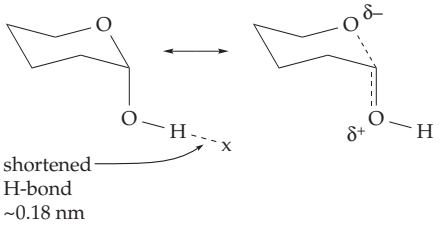
below the plane of the other four ring atoms. *These conformers are not easily interconvertible unless the ring is opened by a chemical reaction,* e.g., that of Eq. 4-1. However, manipulation with an atomic force microscope has shown that stretching of single polysaccharide chains can cause interconversion of the two chair conformations of pyranose rings, which are separated by an energy barrier of ~46 kJ/mol.[26a]

Most sugars occur in the chair conformation that places the largest number of substituents in equatorial positions and is therefore most stable thermodynamically. For D-aldoses this is usually the $^4C_1$ conformation:




β-D-Glucose, a pyranose ring form in a *chair* conformation.

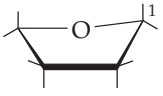
For L-aldoses it is the $^1C_4$ form. There are exceptions to this rule. For example, α-D-idopyranose as well as iduronate rings assume the $^1C_4$ conformation because this conformation places the maximum number of bulky groups in equatorial positions.[27,28] It is noteworthy that electronegative substituents on the anomeric carbon atom of a sugar ring often prefer an axial orientation. This **anomeric effect** is also reflected in a shortening by about 0.01 nm of hydrogen bonds involving the hydrogen atom of the anomeric axial OH group. These effects can be explained partially as a result of coulombic repulsion of the two C–O dipoles and of resonance of the following type:[29–31]




In addition to the chair conformations of six-membered rings the less stable **boat (B)** conformations are also possible. The six boat forms are smoothly interconvertible through intermediate **twist** (**T**) forms, which are also called **skew** (**S**) forms.[3,5] Since the

Boat Skew or twist

internal angle in a pentagon is 108° (close to the tetrahedral angle) we might anticipate a nearly planar five-membered ring. However, eclipsing of hydrogen atoms on adjacent carbons prevents formation of such a flat structure. One of the atoms may be buckled out of the plane of the other four about 0.05 nm, into an **envelope conformation** (e.g., $^1$E), or only three atoms may be in a plane, as in a twist conformation.

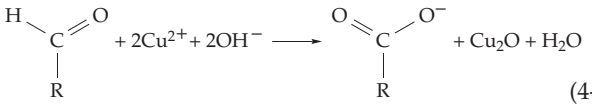


The $^1$E envelope conformation

Any one of the five atoms of the ring can be either above or below the plane defined by the other four in the envelope conformation. The energy barriers separating them are very low, and in cyclopentane or in proline all of the envelope conformations are freely interconvertible through intermediate skew forms.[32] Furanose sugar rings are very flexible but the presence of the bulky substituents reduces the number of possible conformations.[33–36a] See Chapter 5 for further discussion.

## 3. Characteristic Reactions of Monosaccharides

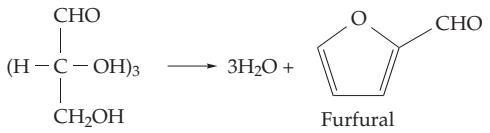
The aldehyde group of aldoses can either be oxidized or reduced and ketoses can be reduced. The best laboratory reagent for reduction is **sodium borohydride** which acts rapidly in neutral aqueous solutions (see Eq. 4-2). Since both $NaB^3H_4$ and $NaB^2H_4$ are available, radioactive or heavy isotope labels can be introduced in this way. The aldehyde groups can be oxidized by a variety of agents to the corresponding aldonic acids, a fact that accounts for the reducing properties of these sugars. In alkaline solution aldoses reduce $Cu^{2+}$ ions to cuprous oxide (Eq. 4-3), silver ions



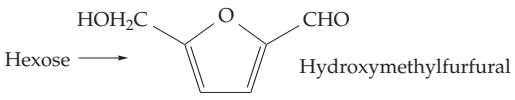

(4-3)

to the free metal, or hexacyanoferrate (III) to hexacyanoferrate (II). These reactions provide the basis for sensitive analytical procedures. Even though the aldoses tend to exist largely as hemiacetals (Eq. 4-1) the reducing property is strongly evident. Oxidation by metal-containing reagents is usually via the free aldehyde, but oxidation by hypobromite $BrO^-$ ($Br_2$ in alkaline solution) yields the lactone, as does enzymatic oxidation (Eq. 15-10).

Sugars are unstable in acid. Boiling with concentrated HCl or $H_2SO_4$ converts pentoses to furfural (Eq. 4-4a) and hexoses to hydroxymethylfurfural (Eq. 4-4b).




(4-4a)

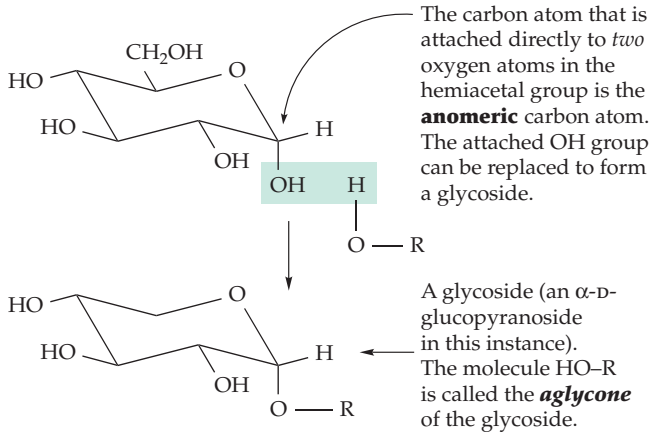



(4-4b)

The aldehydes produced in these reactions can be condensed with various phenols or quinones to give colored products useful in quantitative estimation of sugar content or in visualizing sugars on thin-layer chromatographic plates. Phenol and sulfuric acid yield a product whose absorbance at 470 nm can be used as a measure of the total carbohydrate content of most samples. Resorcinol (1,3-dihydroxybenzene) in 3 M HCl (Seliwanoff's reagent) gives a red precipitate with ketoses. Orcinol (5-methylresorcinol) reacts rapidly with pentoses and with ribonucleosides and ribonucleotides. Since 2-deoxy sugars react slowly this can be used as a test for RNA. Diphenylamine with $H_2SO_4$ gives a blue-green color specifically with 2-deoxy sugars and can be used to test for DNA if the sample is first hydrolyzed.[37]

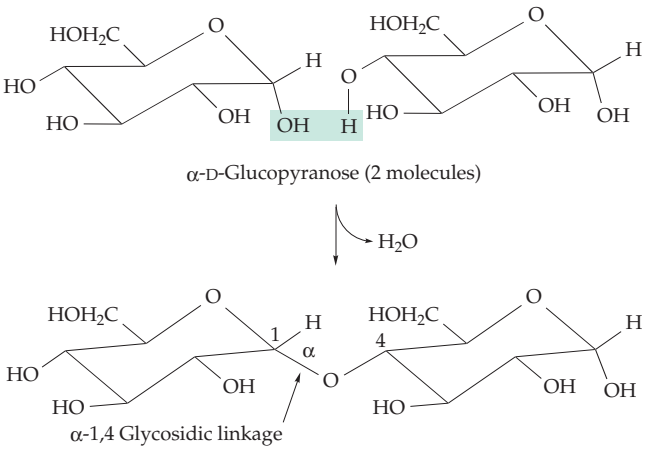
## B. Glycosides, Oligosaccharides, Glycosylamines, and Glycation

The hydroxyl group on the anomeric carbon atom of the ring forms of sugars is reactive and can be replaced by another nucleophilic group such as –O–R from an alcohol. The product is a glycoside (Eq. 4-5). The reaction with methanol occurs readily with acid catalysis under dehydrating conditions, e.g., in 100% methanol.



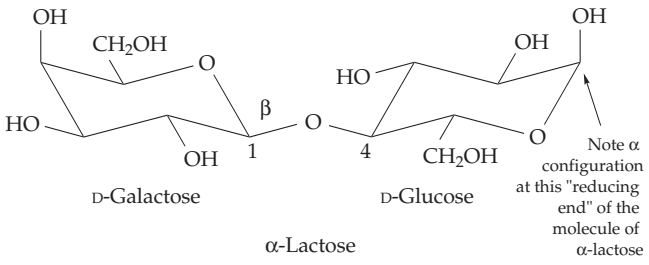

(4-5)

The alcohol in this equation can be a simple one such as methanol or it can be any of the –OH groups of another sugar molecule. For example, two molecules of α-D-glucopyranose can be joined, in an indirect synthesis, to form **maltose** (Eq. 4-6). Maltose is formed by the hydrolysis of starch and is otherwise not found in nature. There are only three abundant naturally occurring **disaccharides** important to the metabolism of plants and animals.[38] They are **lactose** (milk), **sucrose** (green plants), and **trehalose** (fungi and insects).




(4-6)

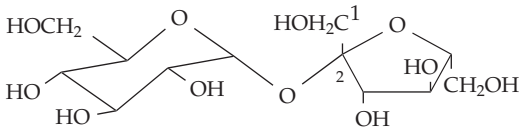
Disaccharides are linked by **glycosidic** (acetal) linkages. The symbol α-1,4, used in Eq. 4-6, refers to the fact that in maltose the glycosidic linkage connects carbon atom 1 (the anomeric carbon atom) of one ring with C-4 of the other and that the configuration about the anomeric carbon atom is α. While the α and β ring forms of free sugars can usually undergo ready interconversion, the configuration at the anomeric carbon atom is "frozen" when a glycosidic linkage is formed. To describe such a linkage, we must state this configuration together with the positions joined in the two rings (see Eq. 4-6). Lactose, whose structure follows, can be described as a disaccharide containing one galactose unit in a β-pyranose ring form and whose anomeric carbon atom (C-1) is joined to the 4 position of glucose, giving a β-1,4 linkage:



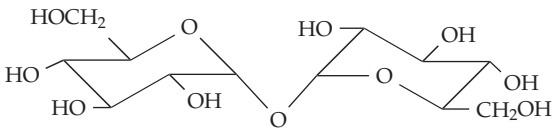

The systematic name for α-lactose, *O*-β-D-galactopyranosyl-(1→4)-α-D-glucopyranose, provides a complete description of the stereochemistry, ring sizes, and mode of linkages. This name may be abbreviated β-D-Gal*p*-(1→4)-α-D-Glc*p*. Since pyranose rings are so common and since most natural sugars belong to the D family, the designations D and *p* are often omitted. It may be assumed that in this book sugars are always D unless they are specifically designated as L. When linkages remain uncertain abbreviated formulas are given. Because the glucose ring in lactose is free to open to an aldehyde and to equilibrate (in solution) with other ring forms, the name lactose does not imply a fixed ring structure for the glucose half. Thus, the name lactose can be abbreviated as βGal-(1→4)-Glc or, more succinctly, as Galβ1→4Glc. However, in crystalline form the sugar exists either as α-lactose or β-lactose. The latter is more soluble and sweeter than α-lactose, which sometimes crystallizes in ice cream upon prolonged storage and produces a "sandy" texture. An isomer of lactose, Galβ1→6Glc or **allolactose**, is an important inducer of transcription in cells of *E. coli* (Chapter 28).

Notice that in the drawing of the lactose structure the glucose ring has been "flipped over" with respect to the orientation of the galactose ring, a consequence of the presence of the β-1,4 linkage. For maltose, where the linkage is α-1,4, the two rings are usually drawn with the same orientation (Eq. 4-6). Maltose can be described as α-D-Glc*p*-(1→4)-D-Glc*p* or more simply as Glcα1→4Glc.

In sucrose and in α,α-trehalose the reducing groups of two rings are joined. Each of these sugars exists in a single form. Sucrose serves as the major transport sugar in green plants, while trehalose plays a similar role in insects, as does D-glucose in our blood.



Sucrose: Glc*p*α1→β2Fru*f*



α, α-Trehalose: Glc*p*α1→α1Glc*p*

Trehalose, or "mushroom sugar," is found not only in fungi but also in many other organisms.[39–41]. It serves as the primary transport sugar in the hemolymph of insects and also acts as an "antifreeze" in many species. It accounts for up to 20% of the dry weight of **anhydrobiotic organisms**, which can survive complete dehydration. These include spores of some fungi, yeast cells, macrocysts of *Dictyostelium*, brine shrimp cysts (dried gastrulas of *Artemia salina*), some nematodes, and the resurrection plant. These organisms can remain for years in a dehydrated state. Hydrogen bonding between the trehalose and phosphatidylcholine may stabilize the dry cell membranes.[18,40,41] Although they can be desiccated, fungal spores remain dormant even when considerable water is present. One of the first detectable changes when the spores germinate is a rapid increase in the activity of the enzyme **trehalase**

which hydrolyzes trehalose to glucose.[42,43]

Disaccharides, as well as higher oligosaccharides and polysaccharides, are thermodynamically unstable with respect to hydrolysis, for example, for lactose in aqueous solution:

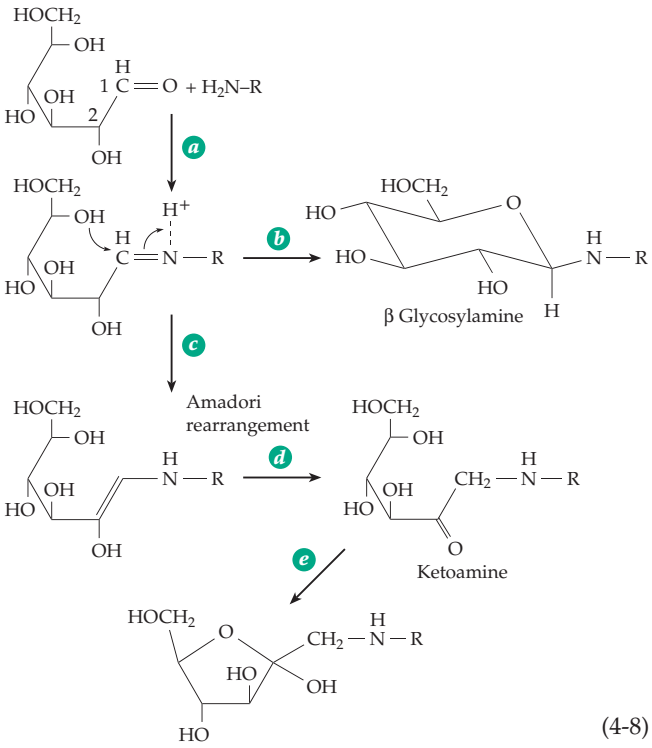
$$\text{Lactose} + H_2O \rightarrow \text{D-glucose} + \text{D-galactose}$$
$$\Delta G^\circ \approx -8.7 \pm 0.2 \text{ kJ mol}^{-1} \text{ at } 25^\circ\text{C} \qquad (4\text{-}7)$$

The corresponding equilibrium constant $K$ = [D-glucose][D-galactose]/[lactose] ≈ 34 M. For other oligosaccharides $K$ varies from 17 to 500 M.[44,45] However, sucrose is far less stable with $K = 4.4 \times 10^4$ M and $\Delta G^\circ = -26.5 \pm 0.3$ kJ mol$^{-1}$ for its hydrolysis.[46] Sucrose is also less stable kinetically and undergoes rapid acid-catalyzed hydrolysis. This fact is exploited when vinegar is used to convert sucrose to the less crystallizable mixture of glucose and fructose during candy-making. In marked contrast, trehalose is extremely resistant to acid-catalyzed hydrolysis.

The joining of additional sugar rings through glycosidic linkages to a disaccharide leads to the formation of **oligosaccharides**, which contain a few residues, and to **polysaccharides**, which contain many residues. Among the well-known oligosaccharides are the substituted sucroses **raffinose**, Gal*p*(1→6) Glc*p*(1→2) Fru*f*, and **stachyose**, Gal*p*(1→6) Gal*p*(1→6) Glc*p*(1→2) Fru*f*. Both sugars are found in many legumes and other green plants in which they are formed by attachment of the galactose rings to sucrose. Oligosaccharides have many functions. For example, gram-negative bacteria often synthesize oligosaccharides of 6–12 glucose units in β-1,2 linkage joined to *sn*-1-phosphoglyceryl groups. They are found in the periplasmic space between the inner and outer cell membranes and may serve to control osmotic pressure.[47] Oligosaccharides of 10–14 α-1,4-linked D-galacturonic acid residues serve as signals of cell wall damage to plants and trigger defensive reactions against bacteria in plants.[48]

Just as alcohols can be linked to sugars by glycoside formation (Eq. 4-5), amines can react similarly to give **glycosylamines** (*N*-glycosides). In this instance it is usually the free aldehyde that reacts via formation of a Schiff base (Eq. 4-8*a*). The latter can cyclize (Eq. 4-8*b*) to the glycosylamine with either the α or β configuration. Another important reaction of Schiff bases of sugars is the **Amadori rearrangement** (Eq. 4-8*c*,*d*), which produces a secondary ketoamine with a 1-oxo-2-deoxy structure.

This can cyclize as in step *e* of Eq. 4-8. In addition, epimerization at position 2 occurs through the reversal of step *d*. The overall reaction of Eq. 4-8 is often called **glycation** to distinguish it from **glycosylation,** the transfer of a glycosyl group. The Amadori rearrangement is important in nitrogen metabolism and in nonenzymatic reactions of sugars. For example, small




(4-8)

amounts of glycosylated hemoglobin and other proteins modified by glycation of the protein amino groups are normally present in the blood.[49–51] People with diabetes, who have a high concentration of blood glucose, have increased amounts of glycated protein. High concentration of either glucose or fructose[41,52] may cause serious problems. For example, modification of the protein **crystallins** of the lens of the eye may lead to cataracts. Similar problems with galactose may accompany galactosemia (Chapter 20).[53] Ketoamines formed by glycation of proteins may undergo crosslinking reactions with side chains of other protein molecules and this may be one cause of aging.[42,43] Other reactions of the ketoamines lead to formation of fluorescent and colored products.[41,43,50,54–56] Oxidation products of sugars also participate in these reactions,[57] and nucleic acid bases also react.[56]

## C. Polysaccharides (Glycans)

Polymers of sugars are present in all cells and serve a variety of functions.[58–60] The simple sugars commonly used in the assembly of polysaccharides include D-glucose, D-mannose, D-galactose, D-fructose, D-xylose, L-arabinose, related uronic acids, and amino sugars (Fig. 4-4). These monomer units can be put together in many ways, either as **homopolysaccharides** containing a single kind of monomer or as **heteropolysaccharides** containing two or more different monomers. Because there are many sugars and many ways in which they can be linked, there is a bewildering

**TABLE 4-1**
**Some of the Many Polysaccharides Found in Nature**

| Name | Source | Monomer | Main linkage | Branch linkages |
|---|---|---|---|---|
| Starch | Green plants | | | |
| Amylose | | D-Glucose | α1,4 | |
| Amylopectin | | D-Glucose | α1,4 | α1,6 |
| Glycogen | Animals, bacteria | D-Glucose | α1,4 | α1,6 |
| Cellulose | Green plants, some bacteria | D-Glucose | β1,4 | |
| Dextrans | Some bacteria | D-Glucose | α1,6 | α1,3 |
| Pullulan | Yeast | D-Glucose | α1,6 + α1,4 | |
| Callose | Green plants | D-Glucose | β1,3 | |
| Yeast glucan | Yeast | D-Glucose | β1,3 | |
| Schizophyllan, curdlan, paramylon | | D-Glucose | β1,3 | β1,6 on every third residue |
| Mannans | Algae | D-Mannose | 1,4 | |
| | Yeast | D-Mannose | α1,6 | |
| Xylans | Green plants<br>Brown seaweed | D-Xylose | β1,3 | |
| Inulin | Some plant tubers | D-Fructose | β2,6 | |
| Chitin | Fungi, anthropods | *N*-acetyl-D-Glucosamine | β1,4 | |
| **Alternating polysaccharides** | | | | |
| Hyaluronan | Animal connective tissue | Glucuronic acid + *N*-acetylglucosamine | β1,4 | |
| Chondroitin sulfate | | D-Glucosamine *N*-acetyl-D-Galactosamine | β1,3 + β1,4 | |
| Dermatan sulfate | | α-L-Iduronate + *N*-acetyl-D-Galactosamine | β1,3 + β1,4 | |
| Pectin | Higher plants | D-Galactunonate + others | β1,4 + others | |
| Alginate | Seaweed | D-Mannuronate + L-Guluronate | β1,4 + α1,4 | |
| Agar-agar | Red seaweed | Galactose | β1,4 and α1,3 | |
| Carageenan | Red seaweed | Galactose-4-sulfate + 3,6-anhydro-D-Galactose-2-sulfate | β1,4 + α1,3 | |
| Murein | Bacterial cell wall | *N*-acetyl-D-Glucosamine + *N*-acetyl-D-Muramic acid | β1,4 | |

variety of different polysaccharides. Their chains can be linear or helical. The most numerous functional groups present are the free hydroxyl groups, some of which may form additional glycosidic linkages to produce **branched chains**. Polysaccharides may also contain –COOH, $-NH_2$, $-NHCOCH_3$, and other groups. After polymerization, hydroxyl groups are sometimes methylated or converted to sulfate esters or to ketals formed with pyruvic acid. Structural characteristics of some of the major polysaccharides are listed in Table 4-1.

## 1. Conformations of Polysaccharide Chains

Despite the variety of different monomer units and kinds of linkage present, the conformational possibilities for carbohydrate chains are limited. The sugar ring is a rigid unit and the connection of one unit to the next can be specified by means of two torsion angles ϕ and ψ just as with peptides.[61–63] To specify a torsion angle four atoms must be selected– the two at the ends of the bond about which rotation is being considered and two others. There is more than one way to define the zero angle for ϕ and ψ. As illustrated in the drawing on p. 172, ϕ may be taken as the H1–C1–O–C4' dihedral angle and ψ as the H4'–C4'–O–C1 angle. The zero angle is when H1 and H4' are eclipsed. A related alternative is to take ϕ and ψ as 0° when the two midplanes of the sugar rings are coplanar.

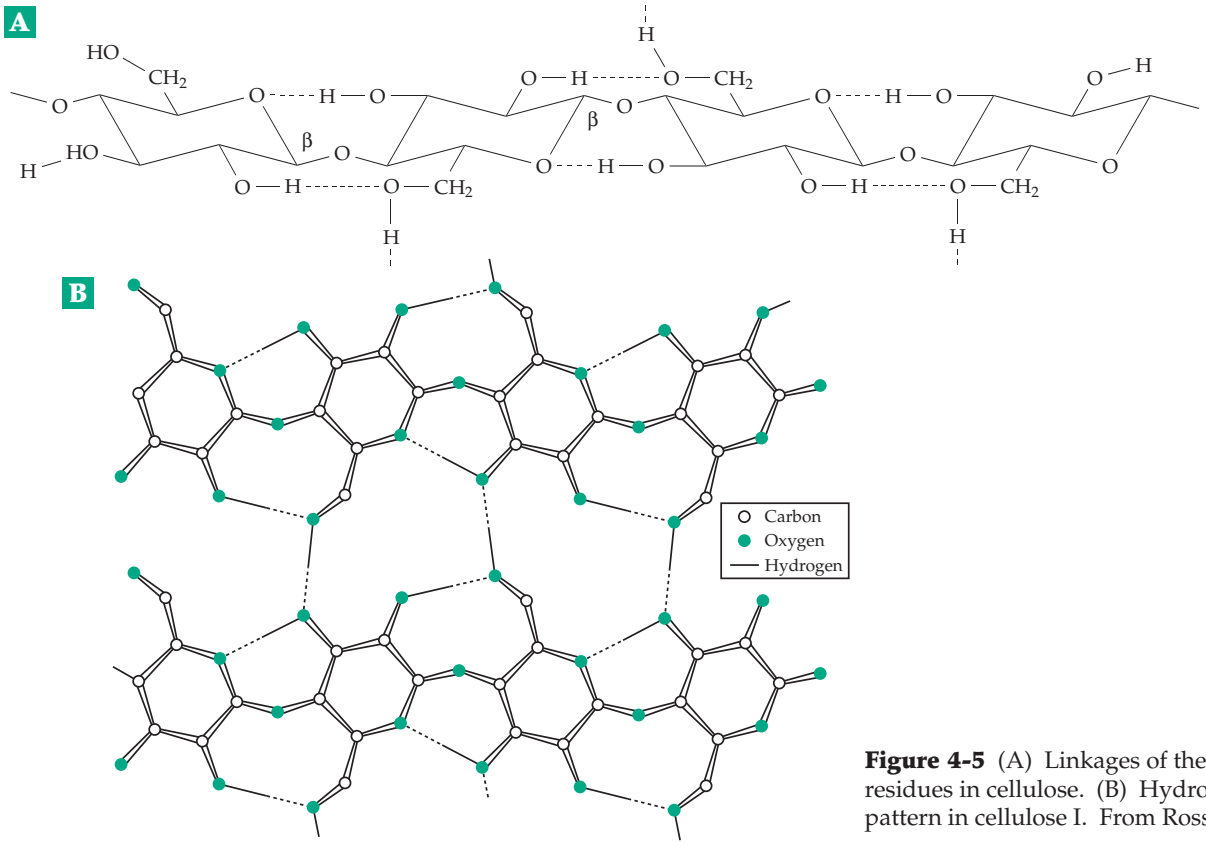
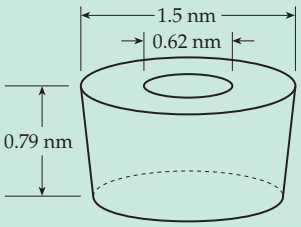



**Figure 4-5** (A) Linkages of the D-glucose residues in cellulose. (B) Hydrogen-bonding pattern in cellulose I. From Ross *et al.*[66]

## BOX 4-A CYCLODEXTRINS

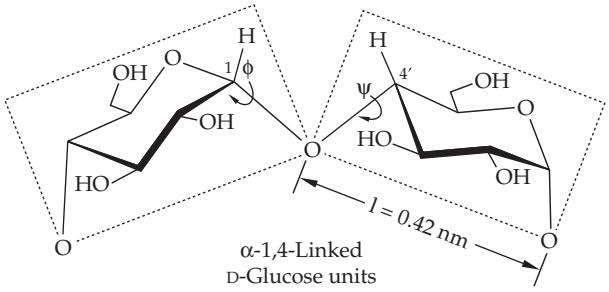
An enzyme produced by some bacteria of the genus *Bacillus* cuts chains of amylose and converts them into tiny rings consisting of six (α-cyclodextrin), seven (β-cyclodextrin), eight (γ-cyclodextrin), or more glucose units. The cyclodextrins, which were first isolated by F. Schardinger in 1903, have intrigued carbohydrate chemists for many years by their unusual properties.[a] They are surprisingly resistant to acid hydrolysis and to attack by amylases and cannot be fermented by yeast.

The torus-like cyclodextrin molecules have an outer polar surface and an inner nonpolar surface.




The small hydrophobic cavities within the cyclodextrins have diameters of 0.50, 0.62, and 0.79 nm, respectively, for the α, β, and γ dextrins.[b] Cavities are potential binding sites for a great variety of both inorganic and organic molecules.[a–g] Complexes of simple alcohols, polyiodides,[c] ferrocene,[d] and many other compounds have been observed. Cyclodextrins can be used to "encapsulate" food additives[b] and their complexes may be useful in separation of enantiomers of drugs.[f,g] They can be chemically modified by adding catalytic groups to serve as enzyme models[h,i] or as color-change indicators sensitive to binding of organic molecules.[j] They can be linked together to form molecular **nanotubes**. Polymer chains can even be threaded through the tubes.[k] What practical applications may yet come from this?


[a] French, D. (1957) *Adv. Carbohydr. Chem. Biochem.* **12**, 189–260
[b] Korpela, T., Mattsson, P., Hellman, J., Paavilainen, S., and Mäkelä, M. (1988–89) *Food Biotechnology* **2**, 199–210
[c] Noltemeyer, M., and Saenger, W. (1980) *J. Am. Chem. Soc.* **102**, 2710–2722
[d] Menger, F. M., and Sherrod, M. J. (1988) *J. Am. Chem. Soc.* **110**, 8606–8611
[e] Hamilton, J. A., and Chen, L. (1988) *J. Am. Chem. Soc.* **110**, 4379–4391
[f] Armstrong, D. W., Ward, T. J., Armstrong, R. D., and Beesley, T. E. (1986) *Science* **232**, 1132–1135
[g] Lipkowitz, K. B., Raghothama, S., and Yang, J. (1992) *J. Am. Chem. Soc.* **114**, 1154–1162
[h] Anslyn, E., and Breslow, R. (1989) *J. Am. Chem. Soc.* **111**, 8931–8932
[i] Granados, A., and de Rossi, R. H. (1995) *J. Am. Chem. Soc.* **117**, 3690–3696
[j] Ueno, A., Kuwabara, T., Nakamura, A., and Toda, F. (1992) *Nature* **356**, 136–137
[k] Harada, A., Kamachi, J. L., and Kamachi, M. (1993) *Nature* **364**, 516–518




The IUB has proposed another convention.[64]

## 2. The Glucans

From glucose alone various organisms synthesize a whole series of polymeric glucans with quite different properties. Of these, **cellulose,** an unbranched β-1,4-linked polyglucose (Fig. 4-5A), is probably the most abundant. It is the primary structural polysaccharide of the cell walls of most green plants.[65] For the whole Earth, plants produce ~$10^{14}$ kg of cellulose per year.

A systematic examination of the possible values for ϕ and ψ shows that for cellulose these angles are constrained to an extremely narrow range which places the monomer units in an almost completely extended conformation.[62] Each glucose unit is flipped over 180° from the previous one. The polymer has a twofold screw axis and there is a slight zigzag in the plane of the rings.[66–68] Remember that in the chair form of glucose all of the –OH groups lie in equatorial positions and are able to form hydrogen bonds with neighboring chains. This feature, together with the rigidity of conformation imposed by the β configuration of the monomer units, doubtless accounts for the ability of cellulose to form strong fibers.

It has been impossible to obtain large single crystals of cellulose. However, from 60 to 90% of native cellulose is thought to be clustered to form the needle-like crystalline microfibrils. These microcrystals of **cellulose I** can be separated from other plant materials by prolonged boiling with dilute NaOH and HCl. Their structure has been established by electron diffraction[67] and by comparison with high-resolution X-ray diffraction structure of the tetrasaccharide β-D-cellotetraose.[66,69] Two closely similar parallel-chained, hydrogen-bonded structures appear to be present. One is shown in Fig. 4-5B. The hydrogen bonds and van der Waals forces bind the chains into sheets which are stacked to form fibers. A typical fiber of plant cellulose has a diameter of 3.5–4 nm and contains 30–40 parallel chains, each made up of 2000–10,000 glucose units. The chain ends probably overlap to form essentially endless fibers that can extend for great distances through the cell wall. They interact with other polysaccharides as is illustrated in Fig. 4-14. A single cotton fiber may be 2–3 cm in length.

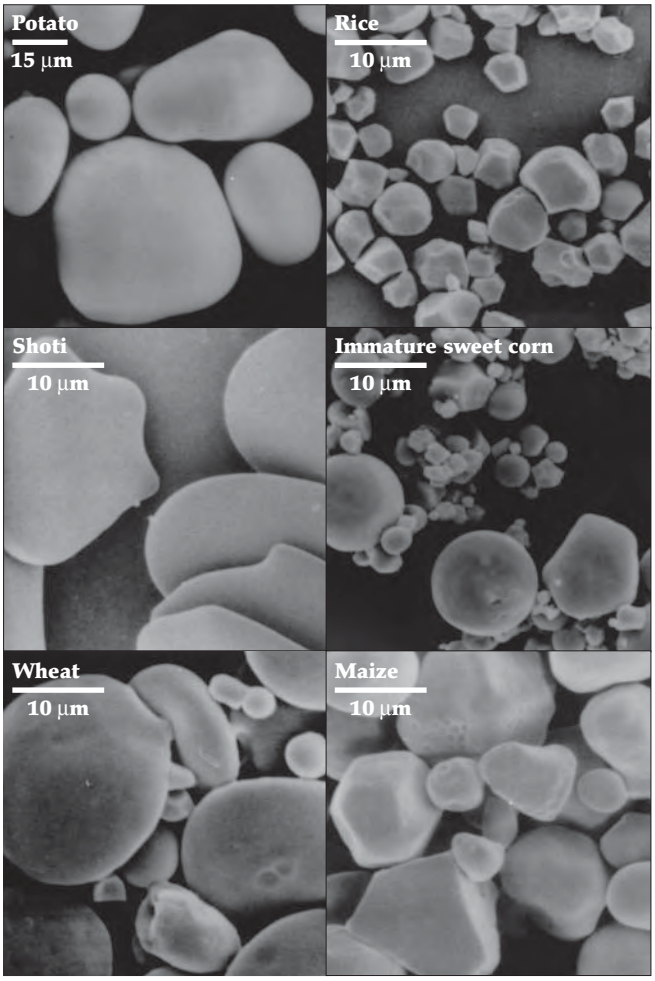



**Figure 4-6** Scanning electron micrographs of starch granules. Magnification 2500× and 3000×. From Jane *et al.*[73]

Cotton thread treated with concentrated NaOH shrinks and has an increased luster. The resulting "Mercerized" cellulose has changed into other crystalline forms. The major one is **cellulose II**, in which the chains in the sheets are antiparallel.[69a] Cellulose II may also occur to some extent in nature. Many other modified celluloses, e.g., **methylcellulose**, in which some –OH groups have been converted to methyl ethers[70] are important commercial products.

**Starch**, another of the most abundant polymers of glucose, is stored by most green plants in a semi-crystalline form in numerous small granules. These granules, which are usually formed within colorless membrane-bounded plastids, have characteristic shapes and appearances (Fig. 4-6) that vary from plant to plant. One component of starch, **amylose**, is a linear polymer of many α-D-glucopyranose units in 1,4 linkage (Fig. 4-7) as in maltose. Starch granules always contain a second kind of molecule known as **amylopectin**.[58]

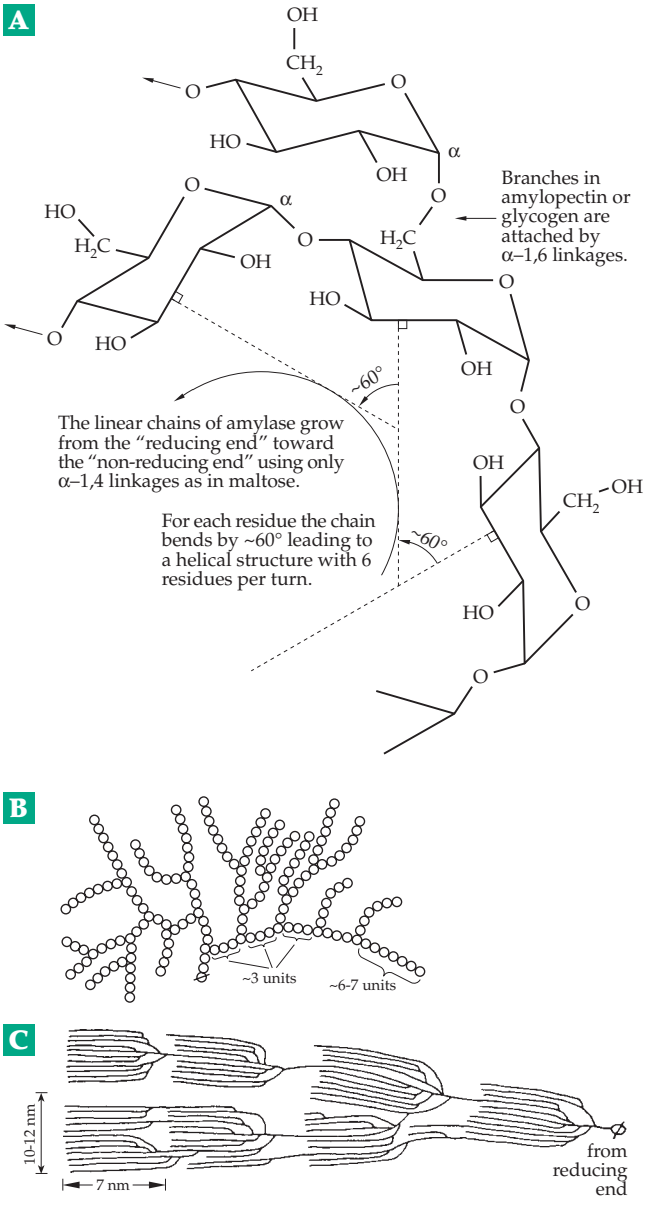


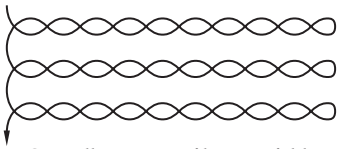

**Figure 4-7** (A) Linkages of the glucose residues in starches and in glycogen. (B) Schematic diagram of the glycogen molecule as proposed originally by K. H. Meyer.[74] The circles represent glucose residues which are connected by α-1,4 linkages and, at the branch points, by α-1,6 linkages. The symbol ϕ designates the reducing group. From D. French.[75] (C) Proposed broomlike clusters in amylopectin. After D. French.[71]

Both amylopectin and **glycogen** (animal starch) consist of highly branched bushlike molecules. Branches are attached to α-1,4-linked chains through α-1,6 linkages (Fig. 4-7). When glucose is being stored as amylopectin or glycogen the many "nonreducing" ends appear to grow like branches on a bush, but when energy is needed the tips are eaten back by enzymatic action (Fig. 11-2). There is only one **reducing end** of a starch molecule where one might expect to find a hemiacetal ring in equilibrium with the free aldehyde. However, this end may be attached to a protein.

Starch granules show a characteristic pattern of growth rings of 0.3- to 0.4-μm thickness with thinner dense layers about 7 nm apart. Study of this layered structure suggests that the individual amylopectin molecules have branches close together in broomlike clusters (Fig. 4-7C).[61,71,72] The amylopectin chains are 120–400 nm long and their relative molecular masses may reach 15–30 million. In addition, starch granules usually contain molecules of the straight-chain amylose, each containing several hundred glucose units and having molecular masses of ~100 kDa. Most starches contain 20–21% amylose but there are special varieties of plants that produce starch with 50–70% amylose. On the other hand, the "waxy" varieties of maize form only amylopectin and lack amylose.

In both starch and glycogen the glucose units of the main chains are linked with α-1,4 linkages. An extended conformation is not possible and the chains tend to undergo helical coiling. One of the first helical structures of a biopolymer to be discovered (in 1943)[76,77] was the left-handed helix of amylose wound around molecules of pentaiodide ($I_5^-$) in the well-known blue starch–iodine complex[78] (Fig. 4-8). The helix contains six residues per turn, with a pitch of 0.8 nm and a diameter of nearly 14 nm. Amylose forms complexes of similar structure with many other small molecules.[79]

Another more tightly coiled double-helical form of amylose has been proposed.[80] Each chain would contain six glucose units per turn and the two chains could be arranged in either parallel or antiparallel directions. The average amylose molecule contains 2000 glucose units and could be stretched to a slender chain over 1 μm long, longer than the crystalline regions observed in starch granules. Thus, the chains within the granules would have to fold back on themselves, possibly in hairpin fashion:



Crystalline array of hairpin folds proposed for linear starch molecules

X-ray diffraction studies support the double-helical structure but suggest a *parallel* orientation of the amylose chains.[81] Since amylose has not been obtained as single crystals the diffraction data do not give a definitive answer. However, if double helices are formed by adjacent branches in amylopectin and glycogen the two strands would be parallel. Starch granules also contain amorphous starch which appears to contain single helices, possibly wrapped around lipid materials.[82]

Glycogen is stored in the cytoplasm of animal cells and to some extent in the lysosomes as enormous 100- to 200-MDa particles. These appear in the electron microscope as aggregates of smaller particles of molecular masses up to 20–40 MDa. A laminated internal structure with surface bumps is suggested by STM microscopy (Fig. 4-9). Biosynthesis of glycogen may be initiated by a 37-kDa protein called **glycogenin**, which remains covalently attached to the reducing end of the glycogen (Chapter 20).[83] Despite the huge molecular masses of glycogen particles, both $^{1}H$ and $^{13}C$ NMR resonances are sharp, indicating a high degree of mobility of the glycosyl units.[84]

**Beta-1,3-linked glucans** occur widely in nature. When a new green plant cell is formed the first polysaccharide to be synthesized is not cellulose but the β-1,3-linked glucose polymer **callose**. Cellulose appears later. Callose is also produced in some specialized plant tissues, such as pollen tubes,[85] and is formed in massive amounts at the site of wounds or of attack by pathogens. The major structural component of the yeast cell wall is a β-1,3-linked glucan with some β-1,6 branches.[86] **Schizophyllan** is a β-1,3-linked glucan with a β-1,6-linked glucosyl group attached to every third residue. A glucan from the coleoptiles of oats contains 30% β-1,3 linkages in a linear chain that otherwise has the structure of cellulose.[87] Other β-1,3-linked glucans serve as energy storage molecules in lower plants and in fungi. Among these are β-1,3-linked glucans such as **paramylon**,[88,89] which is stored by the euglena. A similar polysaccharide, **curdlan**, is formed by certain bacteria.[90]

Some other bacteria, e.g., *Leuconostoc mesenteroides*, make 1,6-linked poly-D-glucose or **dextrans**.[91,92] These always contain some α-1,3-linked branches and may also have α-1,4 and α-1,2 linkages, the structures varying from species to species. Dextrans formed by bacteria such as *Streptococcus mutans* growing on the surfaces of teeth are an important component of dental plaque. Bacterial

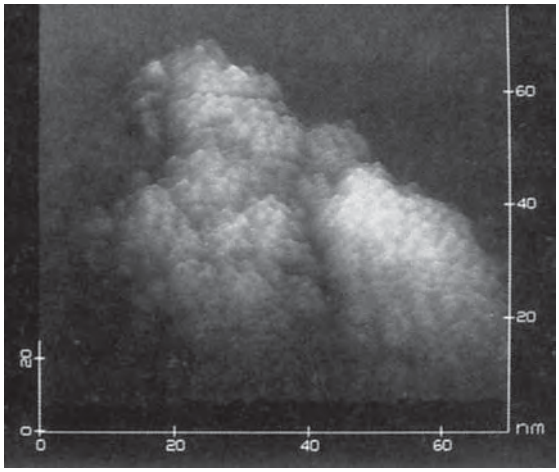



**Figure 4-9** A scanning tunneling microscopic (STM) image of three glycogen molecules. The corrugated surface suggests a laminated structure. These molecules have been purified from tissues by treatment with strong alkali, which breaks the larger aggregates into particles of $M_r = 1–10 \times 10^6$ and diameter 25–30 nm. Courtesy of Fennell Evans.[93]

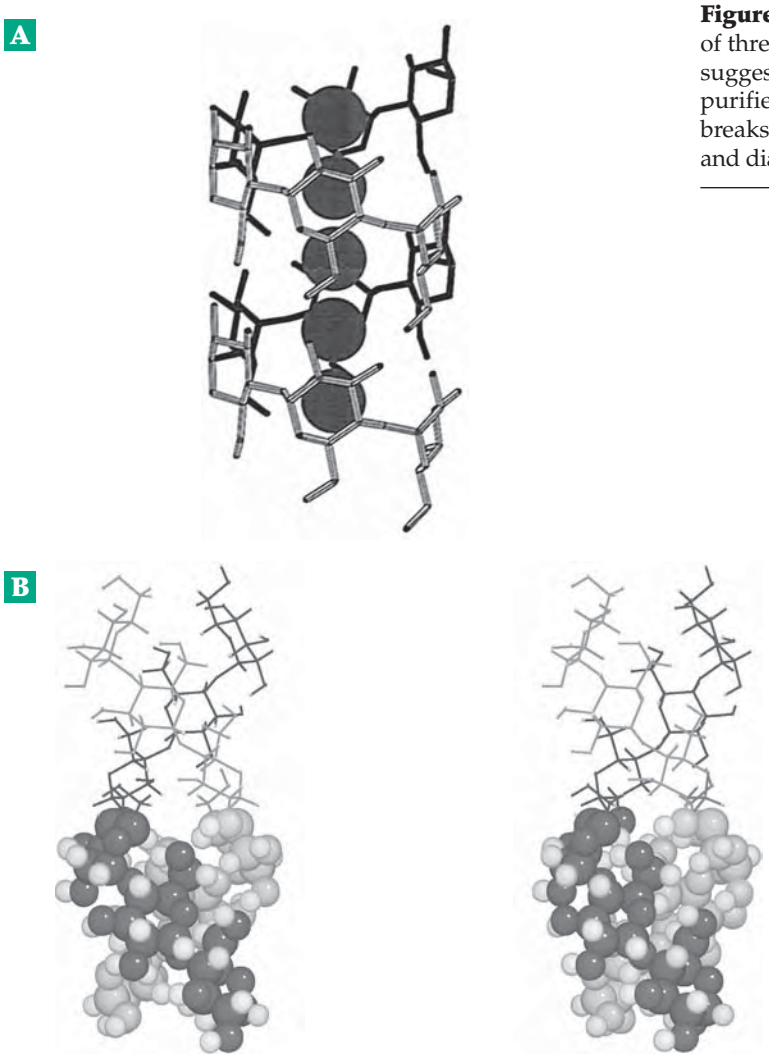



**Figure 4-8** (A) Structure of the helical complex of amylose with $I_3^-$ or $I_5^-$. The iodide complex is located in the interior of the helix having six glucose residues per turn. (B) Model of a parallel-stranded double helix. There are six glucose units per turn of each strand. The repeat period measured from the model is 0.35 nm per glucose unit. Courtesy of Alfred French.

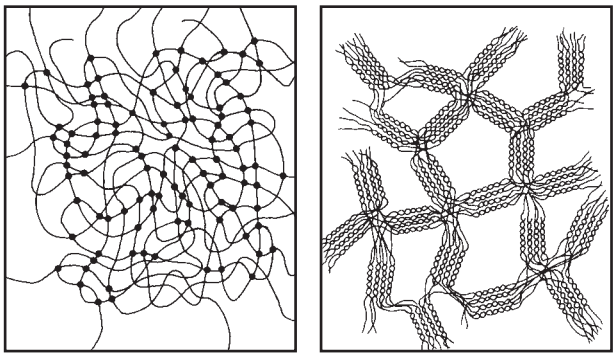


**Figure 4-10** A schematic representation of the gel networks of Sephadex (left) and agarose (right). Note that the aggregates in agarose gels may actually contain 10–$10^4$ helices rather than the smaller numbers shown here. From Arnott *et al*.[95]

dextrans are also produced commercially and are chemically crosslinked to form gels (Sephadex) which are widely used in biochemical separation procedures (Figs. 3-1 and 4-10). A yeast polysaccharide **pullulan** is a regular polymer of maltotriose units with α-1,4 linkages joined in a single chain by α-1,6 linkages.[94] One glucan, called **alternan** is a linear polymer with alternate α-1,6 and 1,3 linkages.

## 3. Other Homopolysaccharides

Cell walls of yeasts contain **mannans** in which the main α-1,6-linked chain carries short branches of one to three mannose units joined in α-1,2, α-1,3, and sometimes α-1,6 linkages.[96] These are covalently linked to proteins.[97] A β-1,4-linked mannan forms microfibrils in the cell wall of some algae such as *Acetabularia* (Fig. 1-11) which do not make cellulose.[98,99] The cell walls of some seaweeds contain a β-1,3-linked **xylan** instead of cellulose. This polysaccharide forms a three-stranded right-handed triple helix.[99,100] Even though xylose is a five-carbon sugar, the polymer contains the thermodynamically more stable six-membered pyranose rings. On the other hand, fructose, a 6-carbon sugar, is present as five-membered furanose rings in **inulin**, the storage polysaccharide of the Jerusalem artichoke and other Compositae, and also in sweet potatoes. The difference has to do with biosynthetic pathways. Furanose rings arise both in inulin and in sucrose because the biosynthesis occurs via the 6-phosphate ester of fructose, making it impossible for the phosphate derivative to form a 6-membered ring.

The major structural polysaccharide in the exoskeletons of arthropods and of other lower animal forms is **chitin**, a linear β-1,4-linked polymer of *N*-acetylglucosamine whose structure resembles that of cellulose. In β chitin the individual parallel chains are linked by hydrogen bonds to form sheets in which parallel chains are held together by NH - - - O = C hydrogen bonds between the carboxamide groups. In the more abundant α chitin the chains in alternate sheets have opposite orientations,[101,102] possibly a result of hairpin folds in the strands. Native chitin exists as microfibrils of 7.25 nm diameter. These contain a 2.8-nm core consisting of 15–30 chitin chains surrounded by a sheath of 27-kDa protein subunits. The microfibrils pack in a hexagonal array, but the structure is not completely regular. Several proteins are present; some of the glucosamine units of the polysaccharide are not acetylated and the chitin core is often calcified.[103] The commercial product **chitosan** is a product of alkaline deacetylation of chitin but it also occurs naturally in some fungi.[102] Chitin is also present in cell walls of yeasts and other fungi. It is covalently bonded to a β-1,3-linked glycan which may, in turn, be linked to a mannoprotein (see Section D,2).[97]

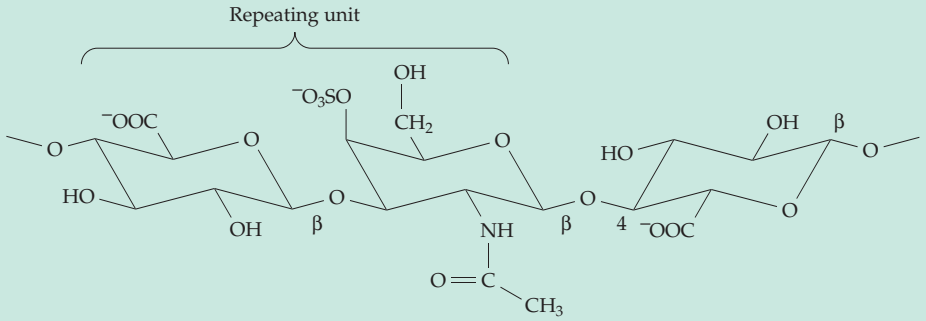
## 4. Heteropolysaccharides of the Animal Body

Many polysaccharides contain repeating units consisting of more than one different kind of monomer.[104,105] Some of these are composed of two sugars in a simple alternating sequence. Examples are **hyaluronan** (hyaluronic acid) and the **chondroitin**, **dermatan**, **keratan**, and **heparan sulfates**. They are important components of the "ground substance" or intracellular cement of connective tissue in animals. Hyaluronan,[106,107] which is abundant in synovial fluid and the vitreous humor of the eye, is a repeating polymer of glucuronic acid and *N*-acetylglucosamine with the structure shown in Fig. 4-11 and $M_r$ of several million. The chondroitin sulfates and dermatan sulfate are similar polymers but with substitution by *N*-acetylgalactosamine and α-L-iduronic acid, respectively, and with sulfate ester groups in the positions indicated in Fig. 4-11.

Hyaluronan solutions are remarkably viscous and at a concentration of only 0.1% can have over 80% of the typical viscosity of biological fluids. This property may result from the presence of hydrogen bonds between the carboxylate, carboxamide, and hydroxyl groups of adjacent sugar residues as in Fig. 4-12. The hydrogen bonds stiffen the chain to give a slender rod. The tetrasaccharide shown in the figure is the repeating unit in a threefold helix.[109] However, the charged molecules do not associate to form strong fibrils like those of cellulose. The chain can be bent easily with breakage of the H-bonds at various positions to give a random coil structure.

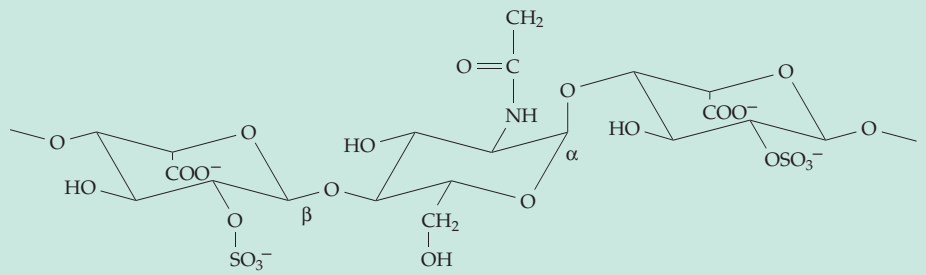
While hyaluronan is not covalently attached to proteins, it is usually anchored to cell surfaces and to

**Type I Chain. Alternating 1→3 and 1→4 linkages; all β, (equatorial).**




| Name | Repeating unit | Variations |
|---|---|---|
| Hyaluronan | $[-4GlcA\beta1\rightarrow3GlcNAc\beta1\rightarrow]_n$ | None. Homogeneous. |
| Chondroitin sulfate | $[-4GlcA\beta1\rightarrow3GalNAc(SO_3^-]\beta1\rightarrow)_n$ | Some GlcA-2-$SO_3^-$, GalNAc-4- or 6-$SO_3^-$or both. |
| Dermatan sulfate | $[-4IdoA\alpha1\rightarrow3GalNAc(4\text{-}SO_3^-]\beta1\rightarrow)_n$ | Some L-IdoA-2-$SO_3^-$. |
| Keratan sulfate | $[-3Gal\beta1\rightarrow4GlcNAc(6\text{-}SO_3^-]\beta1\rightarrow)_n$ | Some Gal-6-$SO_3$, some Sia, Man, Fuc. |

**Type II Chain. All 1→4 linkages, alternating β and α (axial).**




| Name | Repeating unit | Variations |
|---|---|---|
| Heparan sulfate | $[-4\text{-}GlcA\beta1\rightarrow4GlcNAc]_n$ | Some IdoA, sulfation. |
| Heparin | $[-4\text{-L-}IdoA(SO_3^-)\alpha1\rightarrow4GlcN(SO_3^-)\alpha1\rightarrow]_n$ | Some IdoA-2-$SO_3^-$, GlcNAc. 24 different disaccharides possible. |

**Figure 4-11** The repeating disaccharide units of hyaluronan and other glycosaminoglycans. See Fransson[108] and Hardingham and Fosang.[107]

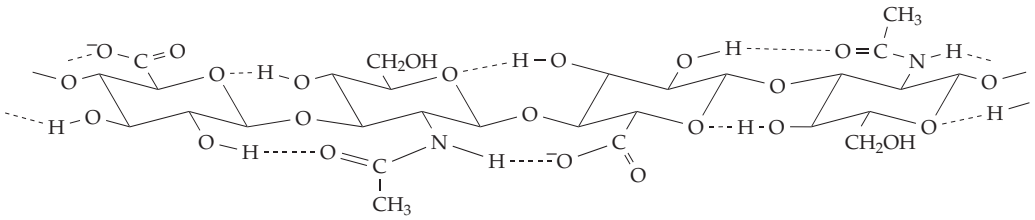


**Figure 4-12** Proposed hydrogen-bonding scheme for the "native" conformation of hyaluronan. See Morris *et al.*[109]

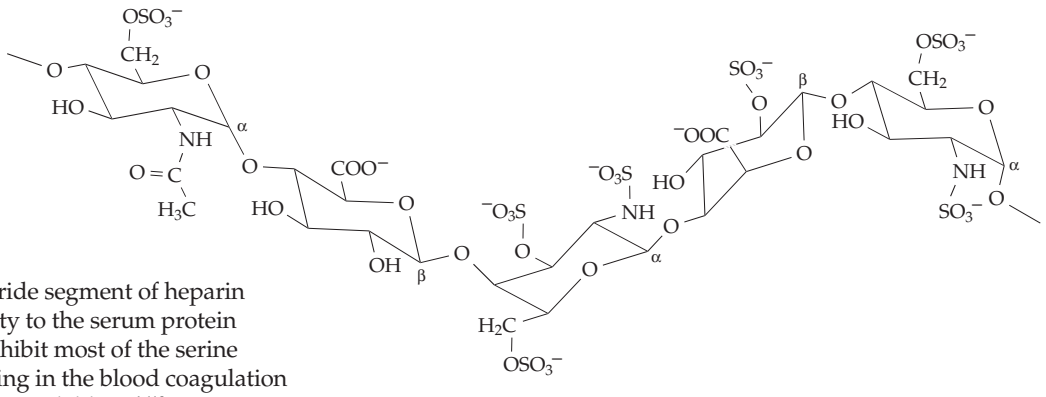


**Figure 4-13** A pentasaccharide segment of heparin which binds with high affinity to the serum protein antithrombin causing it to inhibit most of the serine protease enzymes participating in the blood coagulation process (see Chapter 12). See Lindahl *et al*.[118]

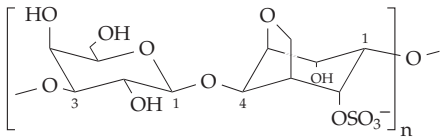
protein "receptors" within the extracellular matrix. In contrast, chondroitin, dermatan, keratan, and heparan sulfates are attached covalently to the proteins at the reducing ends of the polymer chains (see Section D). The attached polymers undergo enzyme-catalyzed chemical alteration. In dermatan most of the glucuronate residues found in chondroitin have been epimerized to iduronate and sulfate groups in ester linkages have been added. Chondroitin sulfate is especially abundant in cartilage; dermatan sulfate is concentrated in skin. Heparan sulfates are more heterogeneous than the other polymers of this group and have been described as "the most complex polysaccharides on the surface of mammalian cells."[110]

In the mast cells of lungs, liver, and other cells heparan sulfate chains are modified further and released into the bloodstream as **heparin**, a material with important anticoagulant properties. Both the amino groups and the 6-hydroxyls of the glucosamine residues of heparin carry sulfate groups. In some units D-glucuronic acid or glucuronate-2-sulfate[111] is present in α-1,4 linkage, but more often L-iduronic acid-2-sulfate is the first unit in the disaccharide.[112–114] The iduronate ring appears to have unusually high conformational flexibility which may influence the biological activity of heparin and related polysaccharides.[115] Because of its anticoagulant property, heparin is an important drug for prevention of blood clot formation.[116] In the United States, in 1976 six metric tons of heparin were used to treat 10 million patients.[117] The anticoagulant activity resides in large part in a nonrepeating pentasaccharide (Fig. 4-13).[112,118] This portion of the heparin, especially if part of a larger octasaccharide,[113] binds to several proteins including the enzyme inhibitors **heparin cofactor II** and **plasma antithrombin III** (Chapter 12). Heparin greatly accelerates the rate at which these proteins bind and inactivate blood clotting factors. See also Section D,1. Lower invertebrates, as well as marine brown algae, contain heavily sulfated fucans which are largely 1,3-linked.[119]

## 5. Plant Heteropolysaccharides

Fibers of cellulose, which run like rods through the amorphous matrix of plant cell walls, appear to be coated with a monolayer of **hemicelluloses**. Predominant among the latter is a **xyloglucan**, which has the basic cellulose structure but with α-1,6-linked xylose units attached to three-fourths of the glucose residues.[104,120,121] L-Fucose may also be present in trisaccharide side chains: L-Fucα1→2Galβ1→ 2Xylα1→.[122] **Pectins** of higher plants contain β-1,4-linked polygalacturonates interrupted by occasional 1,2-linked L-rhamnose residues. Some of the carboxyl groups of these **rhamnogalacturonan** chains are methylated.[123] **Arabinans** and **galactans** are also present in pectin. A possible arrangement of cellulose fibers, hemicelluloses, and pectic materials in a cell wall has been proposed (Fig. 4-14).

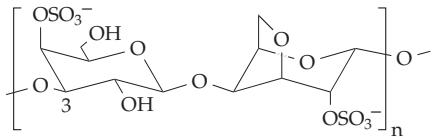
Agarose, an alternating carbohydrate polymer consisting of ~120-kDa chains, is the principal component of agar and the compound that accounts for most of the gelling properties of that remarkable substance. A solid agar gel containing 99.5% water can be formed. Agarose molecules form left-handed double helices with a threefold screw axis, a pitch of 1.90 nm, and a central cavity containing water molecules.[124,125]



Agarose:
[-3-D-Galβ1→4-(3,6-anhydro)-L-Galα1→]$_n$

A similar structure has been established for the gel-forming **carrageenans** from red seaweed. The X-ray data suggest that three of the disaccharide units form one turn of a right-handed helix with a pitch of 2.6 nm. A second chain with a parallel orientation, but displaced by half a turn, wraps around the first helix.[124] Such

double-helical regions provide "tie points" for the formation of gels (Fig. 4-10).[104,127,128]

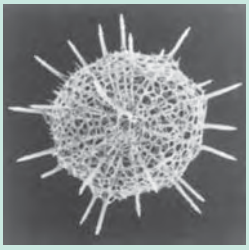



ι-Carrageenan:
[3Gal (4 sulfate) β1→4 (3,6-anhydro) Gal (2-sulfate) α1→]$_n$

Sulfate groups protrude from the structure in pairs and provide binding sites for calcium ions, which stabilize the gel. The presence of occasional extra sulfate groups in these polymers causes kinks in the chains because the derivatized pyranose rings reverse their conformation to the other chair form. This prevents the entire polysaccharide chain from assuming a regular helical structure.[100,104]

**Alginates**, found in cell walls of some marine algae and also formed by certain bacteria, consist in part of a linear β-1,4-linked polymer of D-mannuronate with a cellulose-like structure. Alginates also contain

## BOX 4-B SILICON: AN ESSENTIAL TRACE ELEMENT

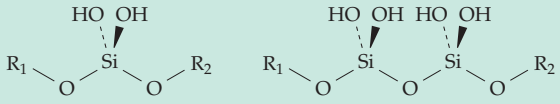


Silica skeleton of a radiolarian[e]

No one can doubt that diatoms, which make their skeletons from $SiO_2$, have an active metabolism of silicon. They can accumulate as much as 0.7 μM soluble silicon, possibly attached to proteins.[a-d]

The radiolaria and sponges often accumulate silicon and limpets make opal base plates for their teeth. Silicon may account for as much as 4% of the solids of certain grasses. Although silicon is usually not considered an essential nutrient for all plants, there is much evidence that it is essential to some and that it is often beneficial.[f] Silicon is found in soil primarily as silicic acid, $H_4SiO_4$, whose concentration ranges from 0.1–0.6 mM. Most of the silicon taken up by plants is deposited within cells, in cell walls, between cells, or in external layers as hydrated $SiO_2$. Presumably the organic components of the plant control the deposition. The $SiO_2$ in some plants takes the form of sharp particles which may have a defensive function. They abrade the enamel surfaces of the teeth of herbivores and can cause other illnesses.[g]

Silicon is essential for growth and development of higher animals,[h-l] and it has been suggested that humans may require 5–20 mg per day.[m] In the chick, silicon is found in active calcification sites of young bone.[i] Silicon-deficient animals have poorly calcified bones and also an elevated aluminum content in their brains.[m] Silicon is present in low amounts in the internal organs of mammals but makes up ~0.01% of the skin, cartilage, and ligaments, in which it is apparently bound to proteoglycans such as chondroitin-4-sulfate, dermatan sulfate, and heparan sulfate (Fig. 4-11).[m,n] These polymers contain ~0.04% silicon or one atom of silicon per 130–280 repeating units of the polysaccharides. Plant pectins contain about five times this amount. The silicon is apparently bound tightly in *ether* linkage. Perhaps orthosilicic acid, $Si(OH)_4$, reacts with hydroxyl groups of the carbohydrates to form bridges between two chains as follows:

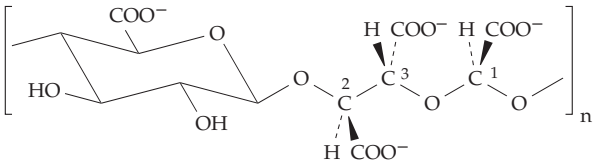



In each of these formulas additional free OH groups are available on the silicon so that it is possible to crosslink more than two polysaccharide chains. Silicon may function as a biological crosslinking agent in connective tissue. **Silaffins**, small polypeptides containing polyamine side chains of modified lysine residues, apparently initiate silica formation from silicic acid in diatoms.[o]


[a] Robinson, D. H., and Sullivan, C. W. (1987) *Trends Biochem. Sci.* **12**, 151–154
[b] Round, F. E. (1981) in *Silicon and Siliceous Structures in Biological Systems* (Simpson, T. L., and Volcani, B. E., eds), pp. 97–128, Springer, New York
[c] Evered, D., and O'Connor, M. (1986) *Silicon Biochemistry*, Wiley, New York
[d] Round, F. E., Crawford, R. M., and Mann, D. G. (1990) *The Diatoms*, Cambridge Univ. Press, Cambridge UK
[e] Buchsbaum, R., Buchsbaum, M., Pearse, J., and Pearse, V. (1987) *Animals Without Backbones*, 3rd ed., Univ. Chicago Press, Chicago
[f] Epstein, E. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11–17
[g] McNaughton, S. J., and Tarrants, J. L. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 790–791
[h] Schwarz, K. (1970) in *Trace Element Metabolism in Animals* (Mills, F., ed), pp. 25–38, Livingstone, Edinburgh, UK
[i] Schwarz, K., and Milne, D. B. (1972) *Nature* **239**, 333–334
[j] Carlisle, E. M. (1972) *Science* **278**, 619–621
[k] Hoekstra, W. H., Suttie, J. W., Ganther, H. E., and Mertz, W., eds. (1974) *Trace Element Metabolism in Animals-2*, University Park Press, Baltimore, Maryland
[l] Carlisle, E. M. (1988) *Science Total Environment* **73**, 95–106
[m] Nielsen, F. H. (1991) *FASEB J.* **5**, 2661–2667
[n] Schwarz, K. (1973) *Proc. Natl. Acad. Sci. U.S.A.* **70**, 1608–1612
[o] Kröger, N., Deutzmann, R., and Sumper, M. (1999) *Science* **286**, 1129–1132

α-L-guluronate, sometimes in homopolymeric "blocks" and sometimes alternating with the mannuronate residues. Groups of adjacent guluronate units are thought to impart calcium-binding properties to alginates.[129]

Polysaccharides with calcium-binding sites may also serve to initiate deposition of calcium carbonate. For example, the unicellular alga *Pleurochrysis carterae* contains an unusual polysaccharide with the following highly negatively charged repeating unit:



[-4GlcAβ1→2-*meso*-tartrate-3→1-glyoxylate-1→]$_n$

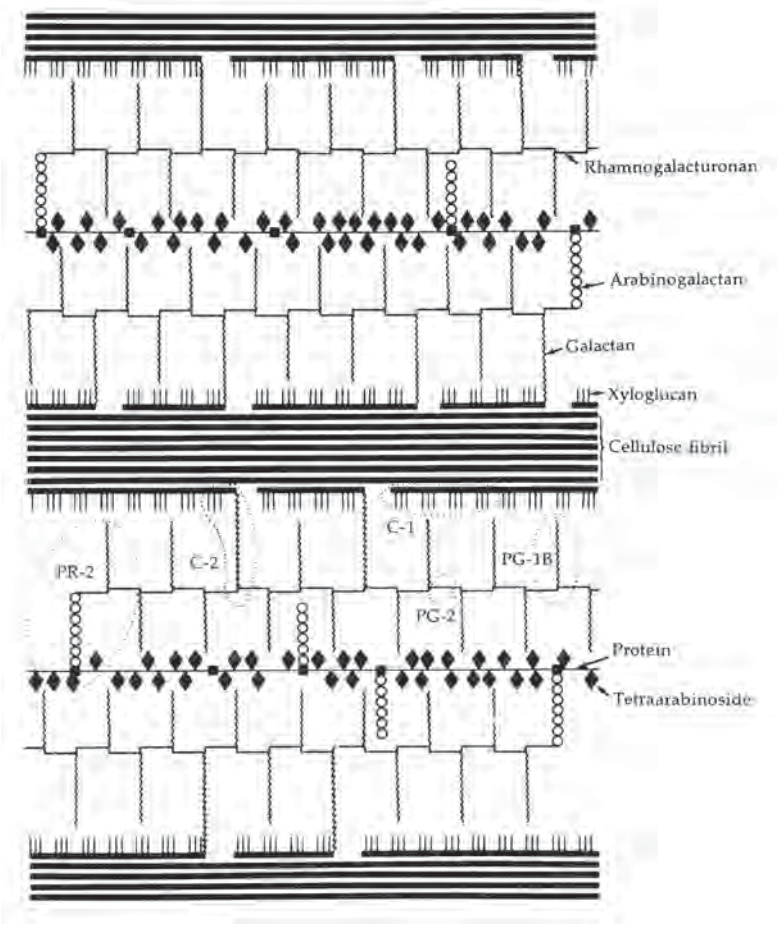
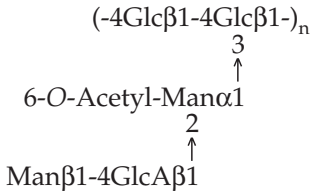


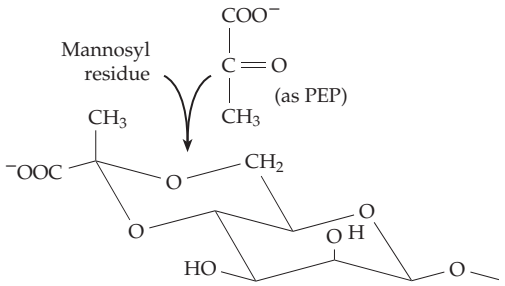

**Figure 4-14** Tentative structure of the walls of suspension-cultured sycamore cells. The wall components are in approximately proper proportions but the distance between cellulose elementary fibrils is expanded to allow room to present the interconnecting structure. There are probably between 10 and 100 cellulose elementary fibrils across a single primary cell wall. From Albersheim *et al*.[126]

*Meso*-tartrate is joined to glucuronic acid in glycosidic linkage and by acetal formation to the aldehyde **glyoxylic acid**, HOOC—CHO, which is also joined in an ether linkage to the next repeating unit.[130] Similar **open acetal linkages** join monosaccharide units in some bacterial lipopolysaccharides and may occur more widely.[130a]

Bacteria form and secrete a variety of heteropolysaccharides, several of which are of commercial value because of their useful gelling properties. **Xanthan gum** (formed by *Xanthomonas campestris*) has the basic cellulose structure but every second glucose residue carries an α-1,3-linked trisaccharide consisting of 6-*O*-acetylmannose, glucuronic acid, and mannose in the following repeating unit:[131,132]



The polymer is further modified by reaction of about half of the mannosyl residues with pyruvate to form ketals (Eq. 4-9).



(4-9)

**Acetan** of *Acetobacter xylinum* has pentasaccharide side chains that contain L-rhamnose.[122] A helical structure for the strands has been observed by atomic force microscopy.

## 6. Polysaccharides of Bacterial Surfaces

The complex structure of bacterial cell walls is discussed in Chapter 8. However, it is appropriate to mention a few bacterial polysaccharides here. The innermost layer of bacterial cell walls is a porous network of a highly crosslinked material known as **peptidoglycan** or **murein** (see Fig. 8-29). The backbone of the peptidoglycan is a β-1,4-linked

alternating polymer of *N*-acetyl-D-glucosamine and *N*-acetyl-D-muramic acid. Alternate units of the resulting chitin-like molecule carry unusual peptides that are attached to the lactyl groups of the *N*-acetyl-muramic acid units and crosslink the polysaccharide chains. In *E. coli* and other gram-negative bacteria the peptidoglycan forms a thin (2 nm) continuous network around the cell. This "baglike molecule" protects the organism from osmotic stress. In addition, gram-negative bacteria have an outer membrane and on its outer surface a complex lipopolysaccharide. The projecting ends of the lipopolysaccharide molecules consist of long carbohydrate chains with repeating units that have antigenic properties and are called **O antigens**. Specific antibodies can be prepared against these polysaccharides, and so varied are the structures that 1000 different "serotypes" of *Salmonella* are known. These are classified into 17 principal groups. For example, group E3 contains the following repeating unit, where *n* may be ~50 on the average. Rha = L-rhamnose.

$$\begin{array}{l} \text{Glc}\alpha 1 \\ \quad\downarrow \\ \quad 4 \\ (\text{Gal}\beta 1\rightarrow 6\text{Man}1\rightarrow 4\text{Rha}\rightarrow)_n \end{array}$$

Polysaccharides of groups A, B, and D contain the repeating unit

$$\begin{array}{l} \quad\text{X} \\ \quad\downarrow \\ (\text{Man}\rightarrow\text{Rha}\rightarrow\text{Gal}\rightarrow)_n \end{array}$$

where X is a 3,6-dideoxyhexose: **paratose** in type A, **abequose** in type B, and **tyvelose** in type D (Fig. 4-15). The existence of the many serotypes depends on the variety of components, on the many types of linkage (α and β, 1→2, 1→3, 1→4, and 1→6) in the repeating units, and on further structural variations at the chain ends.

At the inner end of the O antigen is a shorter polysaccharide "core" whose structure is less varied than that of the outer ends but which is remarkable in containing two sugars found only in bacterial cell walls: a seven-carbon heptose and an eight carbon α-oxo sugar acid, **ketodeoxyoctonate** (KDO). The structures are given in Fig. 4-15 and the arrangement of these sugars in the *Salmonella* lipopolysaccharide is shown in Fig. 8-30. That figure also shows the manner in which the oligosaccharide that bears the O antigen is attached to a lipid anchoring group that is embedded in the outer membrane of the bacteria.

A great variety of polysaccharides are present in the outer layers of other types of bacteria. For example, the mycobacteria have an alternating 5- and 6-linked β-D-Gal*f* polymer attached to their peptidoglycan.

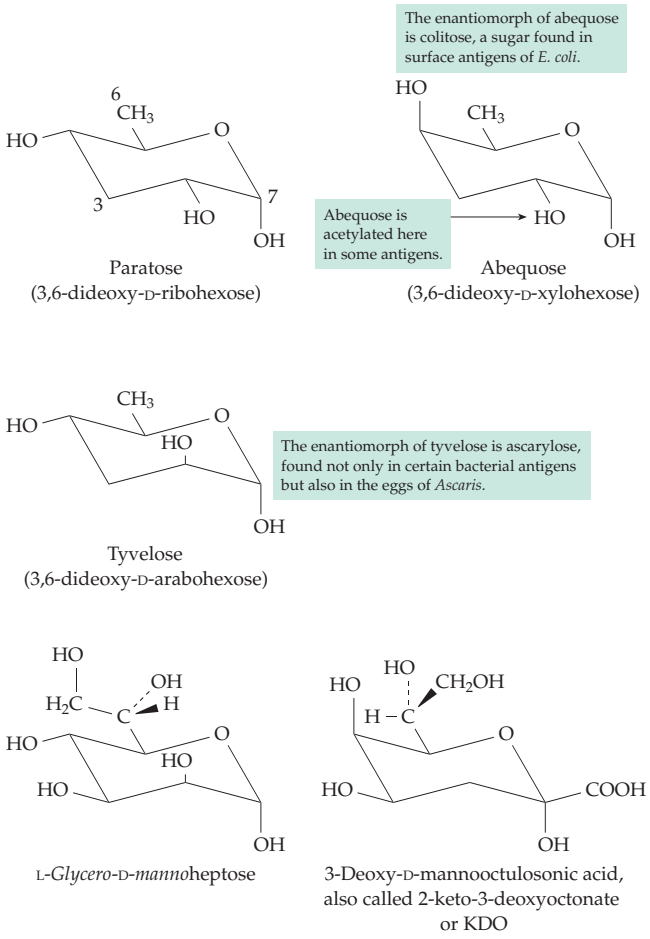



**Figure 4-15** Structures of special sugars found in the "antigens" of the outer cell walls of gram-negative bacteria.

Attached to this galactan are branched penta-D-arabinose units:

$$(\text{Ara}f\beta 1\rightarrow 2\text{Ara}f\alpha 1)_2\rightarrow 3,5\text{Ara}f\alpha\text{—}$$

These are further modified by esterification with mycolic acids.[133] Information on some other cell wall components is given in Chapter 20.

## D. Glycoproteins and Proteoglycans

Many proteins, including almost all of those that are secreted from cells and many that are components of cell surfaces, carry covalently attached oligosaccharides.[134–139] These **glycoproteins** may carry just one or a few, often highly branched, oligosaccharide chains. For example, ribonuclease B has a structure identical to that of ribonuclease A (Fig. 12-25) except for the presence of an oligosaccharide on asparagine 34.[140] In other instances proteins carry a large number of

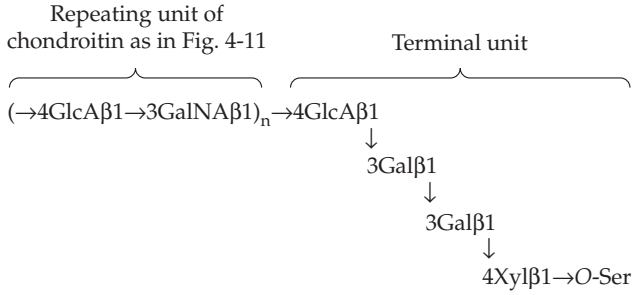
such chains and the carbohydrate may account for over half of the mass of a glycoprotein. Most carbohydrate chains are attached either as *O*-glycosides with the hydroxyl groups of the side chains of serine, threonine, or other hydroxyamino acid residues or as *N*-glycosyl groups through linkage to the amide groups of asparagine side chains. Both types of linkage may be present in a single protein. Here are some examples.

| | |
|---|---|
| Xylβ1→*O*-Ser(Thr) | Proteoglycans of connective tissue; thyroglobulins |
| ε<br>→N—Lys<br>H | Some dermatan sulfates |
| Galβ1→*O*-Hydroxylysine (Hydroxyproline) | Collagen, extension |
| L-Araα1→*O*-4-Hydroxy-proline | Plants |
| GalNAcα1→*O*-Ser(Thr) | Many glycoproteins |
| GlcNAcα1→*O*-Ser(Thr) | Glycoproteins of cytoplasmic surfaces |
| GlcNAcβ1→NH-β$CH_2$-Asn | Many glycoproteins<br>Some dermatan and heparan sulfates |

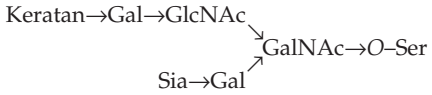
Linkage of a glycosyl group to a carbon atom of an indole ring of tryptophan has also been demonstrated.[141]

## 1. *O*-Linked Glycosyl Groups

In the *O*-linked glycoproteins the sugar that is attached directly to the protein is usually either xylose, galactose, or *N*-acetylgalactosamine, all in the pyranose ring form. Xylose is found only in the intercellular **proteoglycans** which carry the chondroitin, dermatan, and related sulfated polysaccharide chains of connective tissues.[141a,b] Since amino sugars are a major constituent, proteoglycans are often called **glycosaminoglycans**. Chondroitin, dermatan, and heparan sulfates are all attached to "core" proteins by the same linkage, which is illustrated here for chondroitin

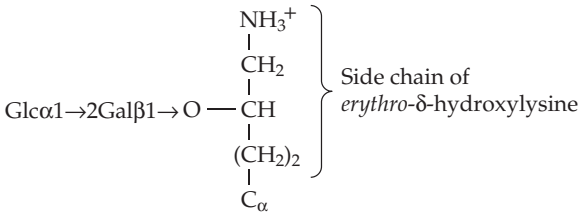



Some IdoA may be present in the terminal unit of dermatan sulfate.[142] Keratan sulfate has its own core proteins[143] and has different terminal units including the following:[107]




The large proteoglycan of human cartilage is built upon the 246 kDa protein **aggrecan**. In the central half of the peptide chain are many Ser-Gly sequences to which about one hundred 10- to 25-kDa chondroitin sulfate chains are attached. About 30 keratan sulfate chains as well as other oligosaccharide groups are also present. These proteoglycan subunits are joined with the aid of 44- to 49-kDa **link protein** to molecules of hyaluronan[144–146] (Fig. 4-16). Several types and sizes of proteoglycan are known.[143,145–147] Dermatan sulfates may be linked to these through either serine or asparagine, depending upon the tissue. The polysaccharide chains of the proteoglycans also bind to collagen fibrils to form a "fiber-reinforced composite material" between cells. Chondroitin and heparan sulfates may be attached at different Ser-Gly sites in a single peptide chain.[148] Degradation of heparan proteoglycans may lead to the shorter free carbohydrate chains found in the circulating heparin. Commercial heparin preparations used as anticoagulants are produced by oxidative destruction of the attached proteins.[113]

A quite different situation holds for **collagen** in which β-galactosyl units and glucosyl-β-galactosyl disaccharide units are attached to side chains of hydroxylysine formed by postsynthetic modification of the original procollagen chain.




A great deal of variation in the amount of glycosylation is observed from one species to another. The human α-1(II) chains of collagen usually carry four disaccharides and four monosaccharide units. In the related collagen-like **extensins**, which are found in plant cell walls, the hydroxyproline (Hyp) side chains are *O*-glycosylated, largely by short oligosaccharides of arabinose in furanose ring form,[150–152] e.g., Ara*f* β1→2Ara*f* β1→2Ara*f* β1→4Hyp. There are as many as 25 repeats of Ser-Pro-Pro-Pro-Pro encoded in an extensin gene. Most of the prolines are hydroxylated and glycosylated. The presence of two or more contiguous proline residues seems to be the signal for the hydroxylation reaction to take place.[152]

Another distinct family of *O*-linked glycoproteins are the **mucins**, which are present in saliva and other

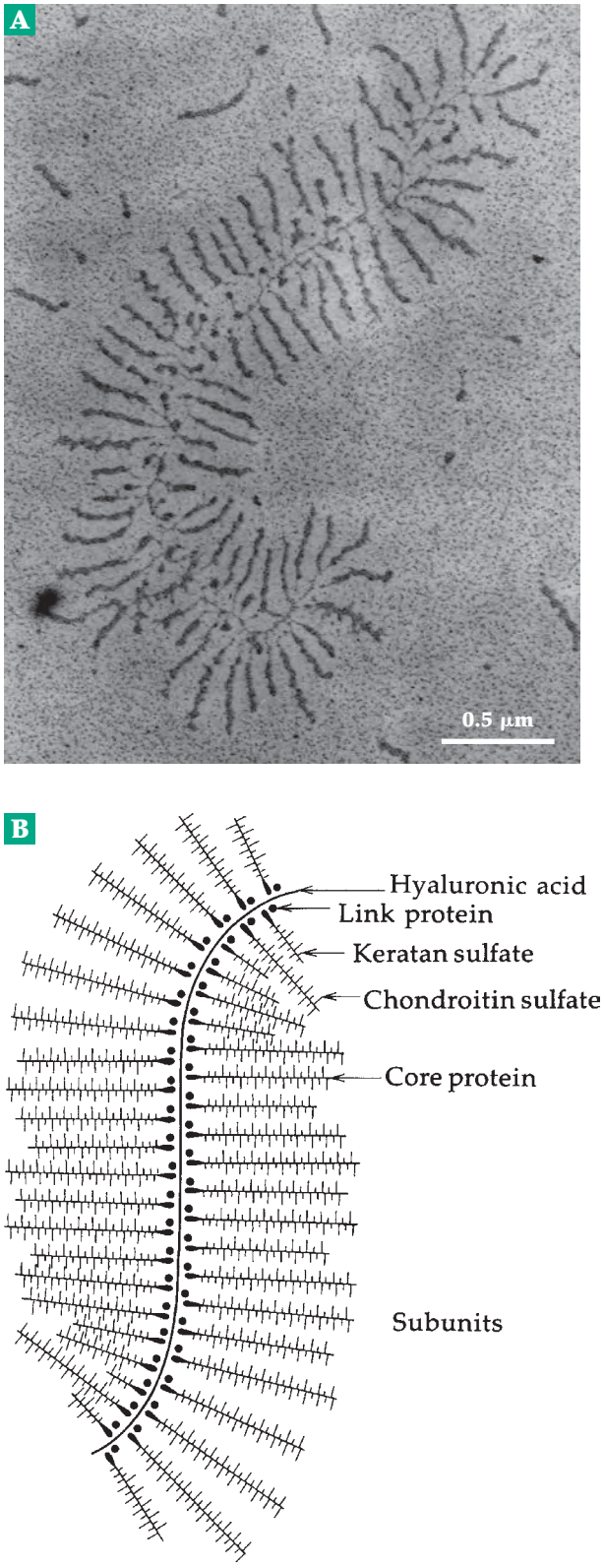



**Figure 4-16** (A) Dark field electron micrograph of a proteoglycan aggregate from bovine articular cartilage (from bearing surfaces of joints). Courtesy of Joseph A. Buckwalter. The filamentous backbone consists of hyaluronic acid, as in (B). The proteoglycan subunits extend from the backbone. From Rosenberg.[149]

mucous secretions. The polypeptides consist largely of serine, threonine, proline, and glycine, with up to one-third of the residues being Ser + Thr. Mucins may contain 70 to 85% carbohydrate linked through *N*-acetylgalactosamine residues to the serine and threonine hydroxyls.[153–160a] Salivary mucins of cows, sheep, and dogs contain largely the following disaccharide, which may contain either *N*-acetyl- or *N*-glycolyl-neuraminic acid:

NeuNAcα2→6GlcNAcβ→Ser(Thr)

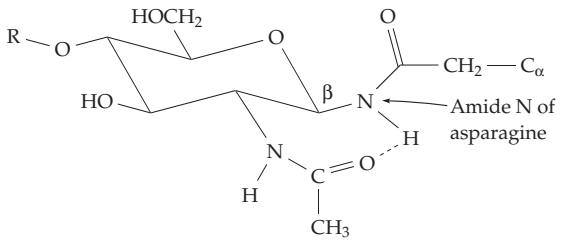
Up to 800 disaccharides are present on a single large protein which is composed of disulfide-linked subunits. The many negatively charged sialic acid (*N*-glycolyl-neuraminic acid) groups are thought to cause expansion and rigidity which increase the viscosity of the protein. Some other mucins contain predominantly large oligosaccharides, which often carry **blood group determinants**.[161,162]

The ABO(H) family of blood group determinants (Box 4-C) are oligosaccharide groups assembled from D-galactose, *N*-acetyl-D-galactosamine, and L-fucose. They are carried on the nonreducing ends of *O*-linked oligosaccharides which may be attached to cell surface proteins, to mucins, or to the sphingolipid known as ceramide (Chapter 8). Attachment to proteins is usually via GalNAc. An example of a blood type A determinant on an *O*-linked oligosaccharide is the following, where the linkage between the galactose and *N*-acetylglucosamine may be either β-1,3 or β-1,4:

GalNAcα1→3Galβ1→3(4)GlcNAcβ1→3GalNAc→*O*-Ser(Thr)
2
↑
Fucα1

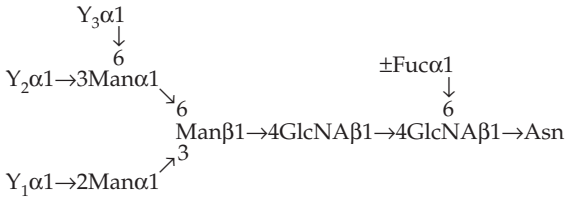
## 2. Asparagine-Linked Oligosaccharides

In many glycoproteins oligosaccharides are linked through *N*-acetylglucosamine to the side chain of asparagine.[134,163,164]




This structure also illustrates one of the hydrogen-bonding possibilities through which the sugar can

interact with the protein. Since the site of glycosylation is often at β bends in the surface of the protein, the amide groups of the *N*-acetylglucosamine may alternatively hydrogen bond to amide groups of the peptide.[165] This asparagine linkage is very common. For example, it is present in 97% of glycoproteins of blood plasma[166] and it is also predominent in the glycoproteins of tissue surfaces.[167] There are numerous structures for asparagine-linked oligosaccharides but many contain the following core to which additional glycosyl groups may be attached:




Here the ±Fuc indicates that this residue is present only on some of the chains.

Notice the three mannose residues on the left side. In the **high mannose type** oligosaccharide, $Y_1$, $Y_2$, and $Y_3$ are additional mannose units. In many instances $Y_1 = Y_2 = Y_3$ = Manα→Man. These "extra" mannose units are put onto the oligosaccharide during the original biosynthesis and before it is attached to the protein (Chapter 20). Some of the mannose units may then be removed during the "processing" of the oligosaccharide in the endoplasmic reticulum and residues of glucosamine, galactose, and sialic acid (Sia) may be added. Thus, $Y_1$ and $Y_2$ in the foregoing structure often become

$$\text{Sia}\alpha2\rightarrow3(6)\text{Gal}\beta1\rightarrow4\text{GlcNAc}\beta1\rightarrow$$

Here, the sialic acid may be linked either 2,3 or 2,6 and the GlcNAc either 1,2 or 1,3. Both Y groups may consist of trisaccharides of this type in "biantennary" oligosaccharides and a third trisaccharide ($Y_3$) may be added to form a "triantennary" molecule. An additional *N*-acetylglucosamine is often linked by β-1,4 linkage to the central mannose of the core and fucosyl residues in α-1,6 linkage are often linked to the *N*-acetylglucosamine next to the asparagine.[168] The Galβ1→4GlcNAc disaccharide unit in the above Y group is also called ***N*-acetyllactosamine** because of its relationship to lactose. It is often repeated in long Y groups, e.g., as $(\rightarrow3\text{Gal}\beta1\rightarrow4\text{Glc-NAc}\beta1\text{—})_n$, the oligosaccharides being called poly-*N*-acetyllactosamino-glycans. These structures are also principal carriers of the ABO blood type determinants on erythrocyte surfaces.[169]

Many *N*-linked oligosaccharides are highly branched. For example, in ovomucoid, a protease inhibitor of hen eggs, "pentaantennary" oligosaccharides have two and three *N*-acetylglucosamine rings, respectively, attached to the terminal mannose units of the oligosaccharide core in 1,2, 1,4 and 1,6 linkages. Another large

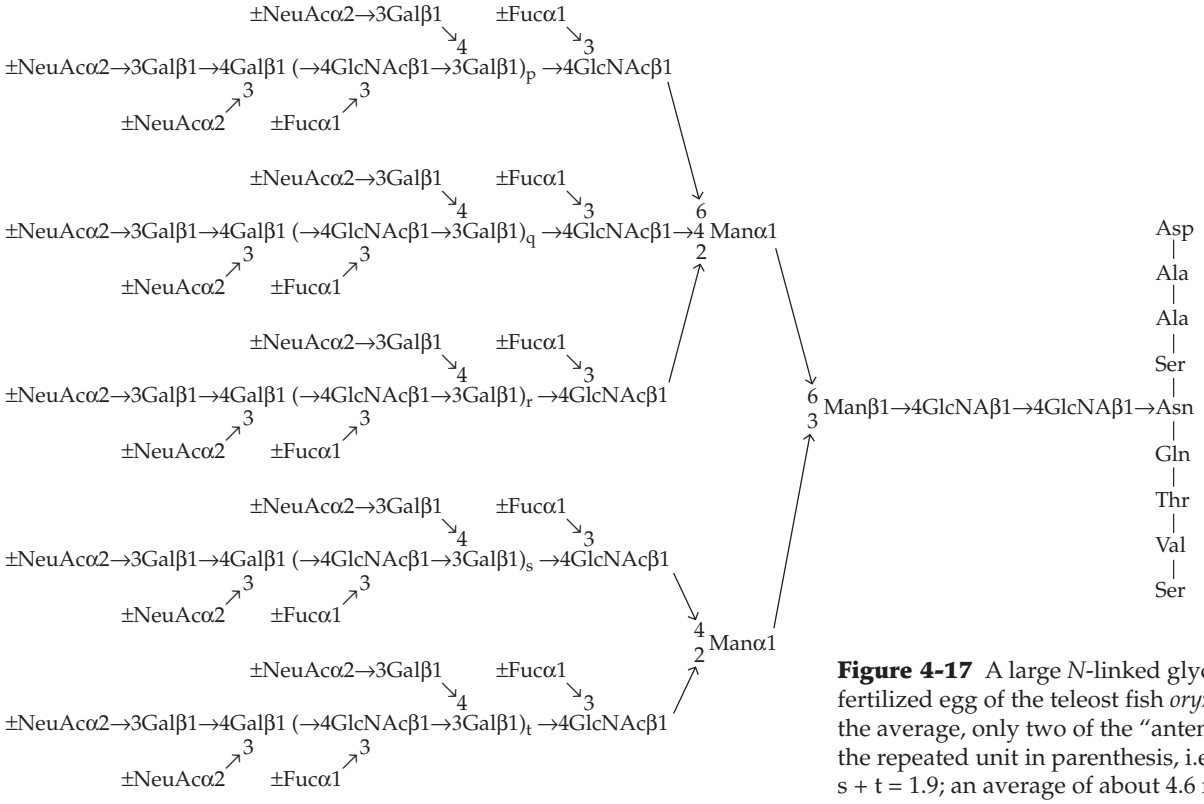



**Figure 4-17** A large *N*-linked glycan from the fertilized egg of the teleost fish *oryzias*.[171] On the average, only two of the "antenna" contain the repeated unit in parenthesis, i.e., p + q + r + s + t = 1.9; an average of about 4.6 residues of sialic acid are present.

## BOX 4-C THE BLOOD GROUP DETERMINANTS

The role of carbohydrates in biological communication is well illustrated by the human blood types.[a,b] According to the ABO system first described by Landsteiner in 1900, individuals are classified into types A, B, AB, and O. Blood of individuals of the same type can be mixed without clumping of cells, but serum from a type O individual contains antibodies that agglutinate erythrocytes of persons of types A and B. Serum of persons of type B causes type A cells to clump and vice versa. Individuals of none of the four types have antibodies against type O erythrocytes. For this reason, persons with type O blood are sometimes inaccurately described as "universal donors."

The ABO blood types are determined by specific **blood group determinants** which are attached to the nonreducing ends of *O*-linked oligosaccharides of surface glycoproteins, mucins, glycolipids, and, to a lesser extent, *N*-linked oligosaccharides. The blood group determinants are found on erythrocytes and all endothelial cells of the body. In about 80% of the population they are also present on glycoproteins of the saliva and other secretions.

The minimal determinant structures, attached to "carrier" R, are as follows:

GalNAcα1→3 Galβ2→R (Fucα1 at 2) — A determinant

Galα1→3 Galβ1→R (Fucα1 at 2) — B determinant

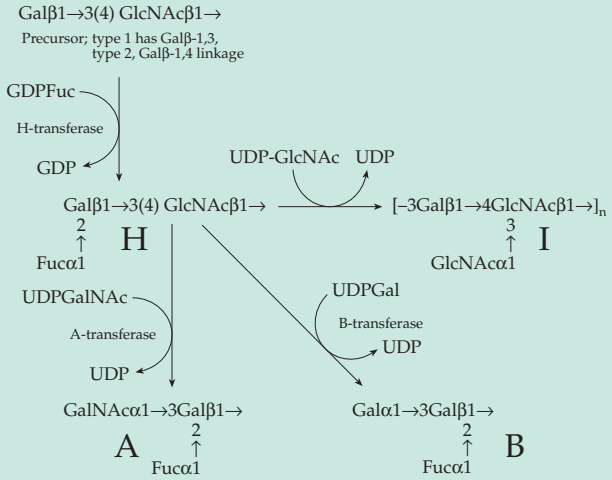
Galβ1→R (Fucα1 at 2) — H(O) determinant

Here R refers to a "carrier oligosaccharide" which can be as simple as

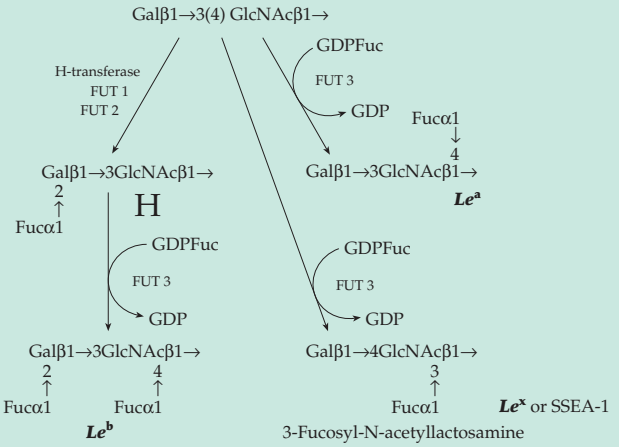
→3(4)GlcNAcβ1→3Galβ1→4Glcβ1—*O*—Ser (Thr, ceramide)

or may be a complex lactosaminoglycan or an oligosaccharide such as that in Fig. 4-17. The minimal determinants can be linked to the carrier oligosaccharide by either a β-1,3 (type I chain) or β-1,4 (type II chain) glycosidic bond.

The genetic basis for the ABO blood groups is well understood. There are three **alleles**, variants of a gene, that encodes a **glycosyltransferase**. In A type individuals, this enzyme transfers *N*-acetylgalactosamine from a carrier molecule, called UDP (Chapter 17), onto the terminal positions of the H(O) determinant. The enzyme specified by the *B* allele transfers galactose. The two enzymes differ in only four amino acid residues but the result is an altered substrate specificity.[a,c,d] The *O* allele produces inactive enzyme as a result of a single base deletion in the gene.[a] The H gene has been identified as that of an α-1,2 fucosyltransferase that transfers α-L-fucose from the carrier GDP to the galactose unit in the foregoing structure.[e] Persons with an inactive *H* transferase gene may have the rare type I, which results from addition of glucosamine branches to the repetitive H antigen structures of polylactosaminoglycans.[f] More often though, the H antigen

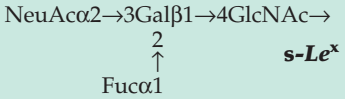



is acted upon by a different fucosyltransferase encoded by the *Le* gene, which determines the **Lewis blood group**. This enzyme places α-1,4-linked L-fucose onto the H antigen to give the *Le*$^b$ antigen. The same fucosyltransferase (which is



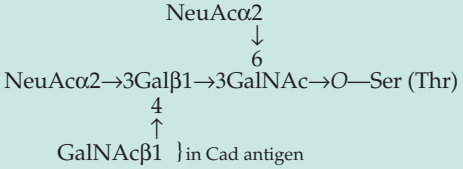
**BOX 4-C Continued**

designated FUT 3) acting on type 2 precursor chains forms the *Le*[a] antigen. The same transferase adds a β-1,3-linked fucose to type 2 chains to give the *Le*[x] antigen, which is also called SSEA-1.[g,h] Lactosamine-type precursors carrying α-2,3-linked sialic acid at their nonreducing ends on surfaces of granulocytes, monocytes, and natural killer cells can be converted to sialylated *Le*[x] (s-*Le*[x]).[i,j]




Individuals with active gene *Se* (for secretion) secrete glycoproteins bearing the blood group substance into saliva and other body fluids. The *Se* gene is another H-transferase present in epithelial cells and salivary glands. In individuals with active *Se* genes (~80% of most populations) most soluble H antigens are converted to *Le*[b] oligosaccharides.[k]

There are at least 12 other well established human blood groups, several of which involve oligosaccharides attached to glycoproteins or glycolipids. The MN antigens consist of a sequence of amino acids near the N terminus of the protein glycophorin (Chapter 8) with attached sialic acid-containing *O*-linked oligosaccharides,[l] e.g., as in the following structure:




The Cad antigen, also found on glycophorin as well as on gangliosides, has an additional β-1,4-linked GalNAc at the indicated position.[m] The Kell blood group antigens are carried on a 93 kDa glycoprotein of erythrocyte surfaces.[n] The P blood group depends upon surface carbohydrates such as the tetrasaccharide part of the ganglioside called globoside; see Fig. 20-9. Its characteristic antigenic activities are destroyed by treatment with dilute periodate.

Other antigens, including those of the Rh and LW groups, are represented by exposed parts of proteins on the erythrocyte surfaces.[o]


[a] Yamamoto, F., Clausen, H., White, T., Marken, J., and Hakomori, S. (1990) *Nature* **345**, 229–233
[b] Frevert, J., and Ballou, C. E. (1985) *Biochemistry* **24**, 753–759
[c] Feizi, T. (1990) *Trends Biochem. Sci.* **15**, 330–331
[d] Yamamoto, F., and Hakomori, S. (1990) *J. Biol. Chem.* **265**, 19257–19262
[e] Larsen, R. D., Ernst, L. K., Nair, R. P., and Lowe, J. B. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6674–6678
[f] van den Eijnden, D., Koenderman, A., and Schiphorst, W. (1988) *J. Biol. Chem.* **263**, 12461–12471
[g] Mollicone, R., Reguigne, I., Kelly, R. J., Fletcher, A., Watt, J., Chatfield, S., Aziz, A., Cameron, H. S., Weston, B. W., Lowe, J. B., and Oriol, R. (1994) *J. Biol. Chem.* **269**, 20987–20994
[h] Nishihara, S., Narimatsu, H., Iwasaki, H., Yazawa, S., Akamatsu, S., Ando, T., Seno, T., and Ikuyo, N. (1994) *J. Biol. Chem.* **269**, 29271–29278
[i] Natsuka, S., Gersten, K. M., Zenita, K., Kannagi, R., and Lowe, J. B. (1994) *J. Biol. Chem.* **269**, 16789–16794
[j] Murray, B. W., Wittmann, V., Burkart, M. D., Hung, S.-C., and Wong, C.-H. (1997) *Biochemistry* **36**, 823–831
[k] Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G., and Lowe, J. B. (1995) *J. Biol. Chem.* **270**, 4640–4649
[l] Adamany, A., Blumenfeld, O., Sabo, B., and McCreary, J. (1983) *J. Biol. Chem.* **258**, 11537–11545
[m] Gillard, B., Blanchard, D., Bouhours, J.-F., Cartron, J.-P., van Kuik, J. A., Kamerling, J., Vliegenthart, J., and Marcus, D. (1988) *Biochemistry* **27**, 4601–4606
[n] Redman, C., Avellino, G., Pfeffer, S., Mukherjee, T., Nichols, M., Rubinstein, P., and Marsh, W. L. (1986) *J. Biol. Chem.* **261**, 9521–9525
[o] Bloy, C., Hermand, P., Blanchard, D., Cherif-Zahar, B., Goossens, D., and Cartron, J.-P. (1990) *J. Biol. Chem.* **265**, 21482–21487


*N*-linked glycan is pictured in Fig. 4-17. Like many others, it has a number of sialic acid residues at the nonreducing ends and also contains *N*-acetyllactosamine units. The major component of the cell walls of yeast (*S. cerevisiae*) is a mannoprotein that carries long *N*-linked oligosaccharides with highly branched outer chains of over 100 mannose residues[170] (see also Section C,3).

The special importance of sialic acids in glycoproteins may lie in the negative electrical charges and the resultant $Ca^{2+}$-binding properties that they impart to cell surfaces. Some glycoproteins carry sulfate or phosphate groups that have similar effects. The glycoproteins of the slime mold *Dictyostelium* contain both mannose-6-*P* and mannose-6-sulfate residues.[163] Oligosaccharides on the protein subunits of flagella of halobacteria contain sulfate esters of glucuronic acid.[172] Some *N*-linked oligosaccharides carry chains of keratan sulfate in place of the sialic acid in the $Y_1$ and $Y_2$ groups of the oligosaccharide.[173]

## 3. Glycoproteins in Biological Recognition

The clusters of sugar rings that form the oligosaccharides on glycoproteins play a vital role in many aspects of biological recognition.[139,174–176] A good example is provided by the human blood groups whose characteristics are determined largely by oligosaccharides (Box 4-C). The adhesion of viruses, bacteria, and eukaryotic parasites to cell surfaces and of one cell to another in multicellular organisms is also dependent on carbohydrates. Recently, it has become clear that the oligosaccharides of cell surfaces change during growth and development and provide an important mechanism by which cells can recognize each other and respond. Why are carbohydrates used for this purpose? It has been pointed out by Sharon and Lis[175] that four different nucleotides can make only 24 distinct tetranucleotides but that four different monosaccharides can make 35,560 unique tetrasaccharides.

***Lectins and other carbohydrate-binding proteins.*** Much of biological carbohydrate-dependent recognition is a result of interaction of individual glycosyl groups or of oligosaccharides on a glycoprotein with a second protein. In some cases that protein is referred to as a **receptor**; in other cases the glycosyl groups may be called the receptor. Carbohydrate-binding proteins include antibodies, enzymes, and carriers that help sugars to cross cell membranes. In addition, there is a large group of carbohydrate binding proteins called **lectins** (from the latin *lectus*; to select).[24,174,177,178]

The first lectins discovered were proteins of plant seeds with specific sugar-binding properties and the ability to agglutinate erythrocytes. **Ricin**, a very toxic protein from castor beans, was isolated in 1888.[178,179] Perhaps the best known lectin is **concanavalin A**, a protein crystallized by Sumner in 1919.[180] Concanavalin A makes up 2–3% of the protein of the jack bean. It is one of a family of legume lectins that resemble favin, whose structure is shown in Fig. 2-15.[177,181,182] Many lectins, including ricin, have quite different three-dimensional structures but share the common characteristic of having a selective binding site for one or more glycosyl rings. Concanavalin A binds to α-D-mannopyranose or α-D-glucopyranose with unmodified hydroxyl groups at C-3, C-4, and C-6.[183,184] Tighter binding is observed if additional mannose residues are present in an oligosaccharide.[178,183] The protein also has specific binding sites for $Ca^{2+}$ and for a transition metal ion such as $Mn^{2+}$. Soybean lectin binds D-*N*-acetylgalactosamine and D-galactose units, while wheat lectin is specific for D-*N*-acetylglucosamine.

Animal cells also produce lectins.[24,105,185–193] The amebas of the cellular slime mold *Dictyostelium* synthesize a classical lectin called **discoidin I** that binds GalNAc or Gal. It is absent from cells until they are ready to differentiate into an aggregating form (Chapter 1). Then it is produced in abundance.[178] Discoidin I has a second binding site specific for the peptide sequence Arg-Gly-Asp (RGD) which is known to be involved in cell adhesion and which binds to such surface proteins as fibronectin and laminin (Chapter 8). Many animal tissues contain soluble lactose-binding lectins known as **galectins** or S-Lac lectins. The best known member is a dimer of 14-kDa subunits;[187,188] many other related lectins have been found.[189,192] Another family are $Ca^{2+}$-dependent or C-type lectins which are specific for mannose, L-fucose, or other sugars.[105,193] **Lectin domains** are being discovered in many proteins.

***Carbohydrate-binding sites.*** The structures of the sites that recognize and bind sugar rings in lectins, enzymes, transporter proteins, and other carbohydrate-binding proteins vary greatly, as does the tightness of binding. However, there are certain common features: Sugar rings are bound by hydrogen bonds, which are often numerous. An example is the galactose chemoreceptor protein from *E. coli*. It binds both α and β anomers of either D-glucose or D-galactose and is utilized by the bacteria in searching for food (see Chapter 19). The structure of D-glucose bound to this protein is shown in Fig. 4-18. Notice the many hydrogen bonds. Two of the –OH groups have the maximum of three hydrogen bonds apiece. There are three negatively charged aspartate side chains and one positively charged guanidinium group. These provide strong ion–dipole interactions which add strength to the bonds. The presence of ionized groups in varying numbers and constellations is another common feature of protein–carbohydrate interactions. A third common feature is the presence of aromatic rings, which often lie against one face of the sugar. The stereoscopic drawing of Fig. 4-18A shows an indole ring of a tryptophan residue in front and a phenylalanine side chain behind the sugar.[194] Sugars bind to lectins,[181,184,195] to enzymes (Chapter 12), and to antibodies[196] through similar interactions.

***Binding of viruses and bacteria to cells.*** The cholera toxin and a related toxic protein from *E. coli* bind to Sia→Gal groups attached to glycolipids (gangliosides, Chapter 8) of erythrocytes and other cells.[197] The influenza virus gains access to our body cells by first binding through a viral surface **hemagglutinin**. This is a protein that binds specifically to NeuAcα2→6Gal or NeuAcα2→3Gal of cell surface oligosaccharides.[174,175,198] Removal of sialic acid from erythrocyte surfaces abolishes the ability of the influenza and some other viruses to bind. It seems somewhat surprising that a second surface protein on the influenza virus is a **neuraminidase** (or **sialidase**) which catalyzes the removal of sialic acid from cell surface proteins,

destroying unoccupied virus receptors.[199,200] This may facilitate movement of a virus particle through the mucin layer surrounding a cell. Bacteria and other invading parasites also produce neuraminidases.[201] *Trypanosamas cruzi,* the causative agent of chagas disease, employs a **transsialidase** to transfer sialic acid from a Galβ on the host cell onto a protein on the parasite surface. This is essential for successful invasion of the host.[202]

The adherence of cells of *E. coli* to mannosyl units of cell surface proteins may initiate the infections that sometimes occur with this bacterium.[174,203] However, cells of *E. coli* from strains that cause urinary infections bind to Galα1→4Gal on glycolipids that carry the blood group P antigens (Box 4C).[174,204] Neuropathogenic strains of *E. coli* or of *Neisseria meningitidis*, which may cause neonatal meningitis, bind to α-2,8-polysialic acid chains on nerve cells.[135,205,206] *Helicobacter pylori*, the stomach ulcer bacterium, binds to the human Lewis$^b$ blood group antigen (Box 4-C).[207] *Entamoeba histolytica*, which causes amebic dystentery, binds to Gal and GalNAc-containing oligosaccharides such as GalNAcα1→3Galα.[208]

***Aggregation and adherence of cells.*** Differences among cell surfaces are fundamental to the formation of multicellular organisms and to many physiological processes. Proteins, carbohydrates, and lipids all contribute material to exposed cell surfaces. The adhesion of one cell to another is mediated by a group of **adhesion proteins** together with oligosaccharide groups and sometimes polysaccharides. Families of adhesion proteins include **integrins** and **cadhedrins** (discussed in Chapter 8), various members of the **immunoglobulin superfamily**, the **cell differentiation antigens**, (often designated CD44, etc.), the C-type lectins known as **selectins**, and proteoglycans.

Here are two of many known examples of specific cell–cell adhesion. The species-specific reaggregation of dissociated cells of marine sponges (Chapter 1) depends upon a 20-kDa proteoglycan of unique structure[209–211] together with a cell surface receptor protein and calcium ions. The recognition of egg cell surfaces by sperm[212–214] is species specific and depends upon interaction of sperm receptors with *O*-linked oligosaccharides of the extracellular coat of the ovum.

***Growth and differentiation.*** The exact structures of the oligosaccharides and polysaccharides of cell surfaces vary not only with cell and tissue type but also with the position of a cell and with time. Actions of numerous glycosyl transferases alter these saccharide groups as an organism grows and develops. Other enzymes alter them by hydrolytic removal of sugars, by isomerizaion, oxidation, and addition of other components such as phospho, sulfo, and acetal groups. For example, the presence of the H-antigen determinant, whose structure is shown in Box 4-C, is strictly regulated, both temporally and spatially, during vertebrate development.[215,216] The relative amounts of the H determinant vs Le$^x$, sLe$^x$ (Box 4-C)

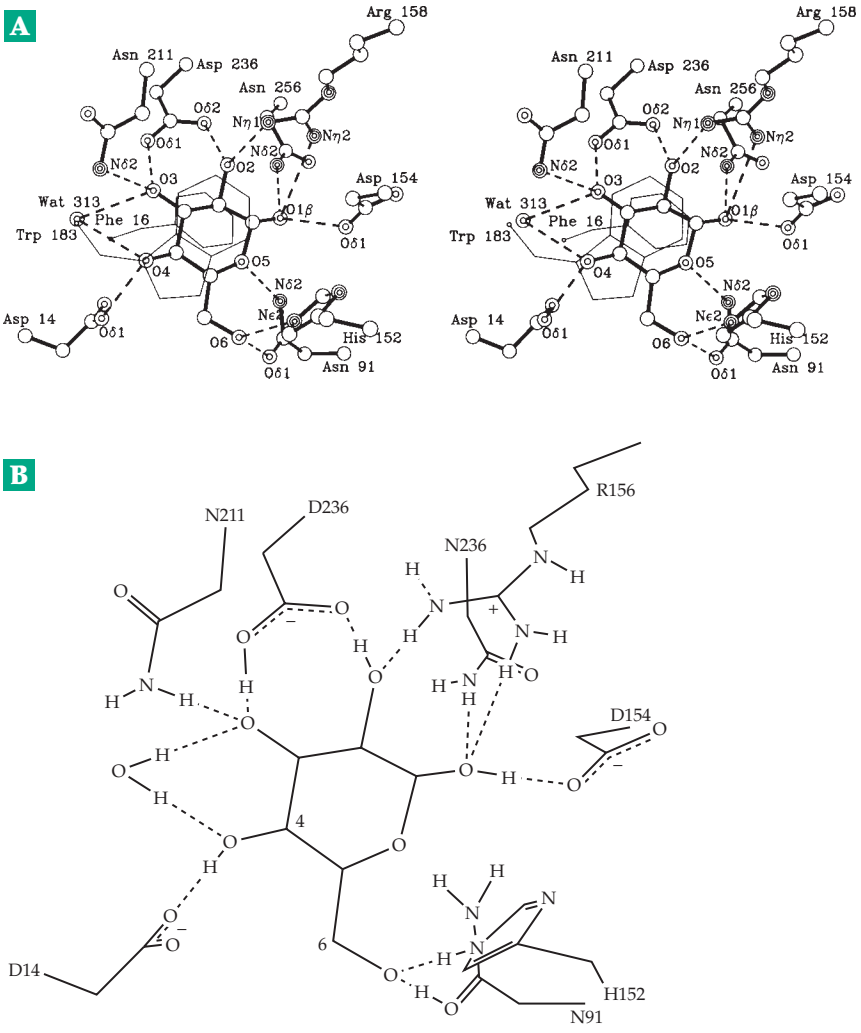
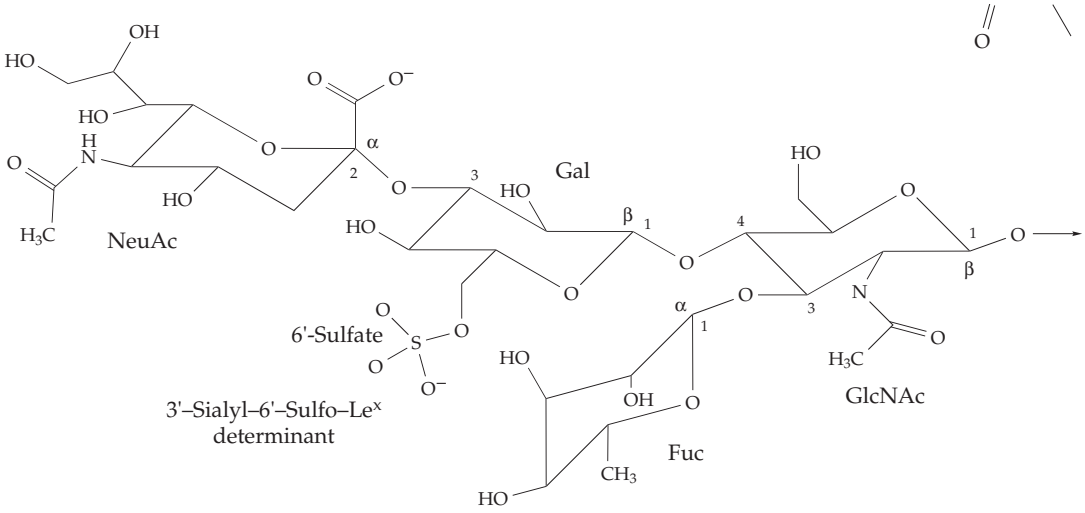



**Figure 4-18** (A) Stereoscopic view of the interactions between the *E. coli* galactose chemoreceptor protein and a bound molecule of D-glucose. (B) Schematic drawing showing many of these interactions and the state of ionization deduced for the aspartate and arginine side chains. From Vyas *et al.*[194]

and other surface groupings is controlled by fucosyltransferases, sialyltransferases, etc. Human cancers often accumulate large amounts of fucosylated glycoproteins and glycolipids carrying $Le^a$, $Le^x$, and sialyl-$Le^x$ antigens[175,216,217] and sialomucins.[218] Glycoproteins help to control many metabolic processes. For example, a protein **calnexin** (a chaperonin, Chapter 10) helps glycoproteins to fold correctly.[219] On the other hand, removal of terminal sialic acid residues from blood plasma proteins leads to rapid removal of the proteins from circulation and to catabolism by liver cells.[175] This process depends upon a receptor protein specific for oligosaccharides with terminal galactosyl residues (Chapter 20).

***Recognition and adhesion by leukocytes.*** A group of three calcium-dependent lectins known as **selectins** bind the sialyl Lewis x ($sLe^x$) antigen and play important roles in adhesion to cells of the vascular endothelium and leukocytes[175,220–224] and also to platelets. Although all of the selectins bind the $sLe^x$ antigen, the binding is weak and these multidomain proteins may simultaneously bind to other ligands such as heparan sulfate.[225] The leukocyte L-selectin (CD62L) binds very tightly to the 3'-sialyl-6'-sulfo-$Le^x$ determinant[220,226] which occurs on mucin-like glycoproteins. The interaction with P selectin helps leukocytes to bind to surfaces on endothelial cells in lymph nodes and sites of chronic or acute inflammation. In a similar way E selectin is synthesized in vascular endothelial cells that have been transiently "activated" by cytokines in response to injury and other inflammatory stimuli. The E selectin binds the sulfated $sLe^x$ antigen on surfaces of neutrophils, monocytes, eosinophils, and basophils.[220,222,227–230] P selectin is stored in secretory granules of platelets and endothelial cells and is released to the cell membrane upon activation by thrombin.[220]

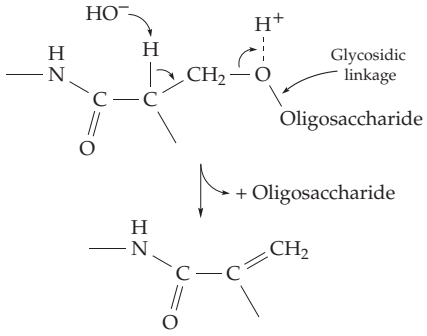



## E. Some Special Methods

Small monosaccharides and oligosaccharides can be separated readily from polymeric constituents and can be purified further by chromatographic procedures including gel filtration as is illustrated in Fig. 3-3. However, polysaccharides and complex oligosaccharides are harder to purify. A few of these, such as cellulose and glycogen, are sufficiently stable that other materials can be dissolved away from them by prolonged boiling in strongly basic solutions. Complex carbohydrates are usually cut into smaller oligosaccharides or glycopeptides. These may be separated by HPLC, capillary electrophoresis, or thin-layer chromatography[134,231–235] (Fig. 4-19) or by chromatography on immobilized lectins.[234] Quantities of less than 25 picomoles can be separated by use of mass spectrometry with liquid chromatography.[236,237] High resolution Fourier transform mass spectroscopy is very useful in the study of posttranslational glycosylation of proteins.[237a,b]

### 1. Release of Oligosaccharides from Glycoproteins

The *O*-linked oligosaccharides of glycoproteins or glycolipids can be split off from the proteins by β elimination (see also Chapter 13):




(4-10)

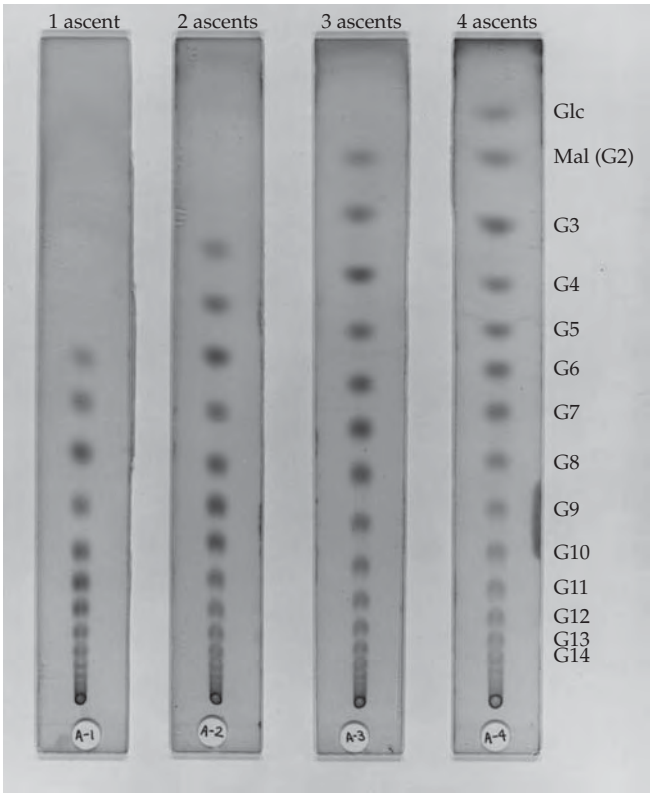
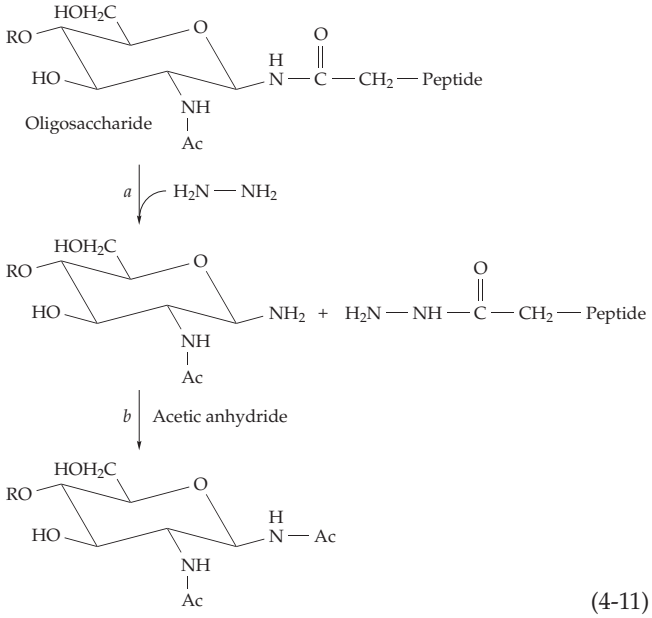



**Figure 4-19** Ascending thin-layer chromatography of a mixture of maltooligosaccharides obtained by the hydrolysis of linear starch. G2 (maltose), G3 (maltotriose), G4 (maltotetraose), G5 (maltopentaose) – G14 represent oligosaccharides with the indicated number of glucose residues, all in α-1,4 linkage. In the multiple ascent technique the chromatographic solvent, whose composition (by volume) is indicated, is allowed to ascend the thin-layer plate repeatedly with the plate allowed to dry between ascents. The separation of the higher oligosaccharides is distinctly improved by a larger number of ascents. Photograph courtesy of John Robyt.

Treatment with 0.1–0.5 M NaOH for several hours will completely liberate the oligosaccharides, whose released carbonyl groups may then be reduced with $NaBH_4$, $NaB^2H_4$, or $^3H$-labeled borohydride to form stable sugar alcohols (Eq. 4-2).[238] Asparagine-linked oligosaccharides are often recovered as glycopeptides prepared by complete proteolytic digestion of the denatured glycoprotein. These can be separated by high voltage electrophoresis in borate buffers.[239] The oligosaccharides can be released from the glycopeptides by enzymes such as endo-*N*-acetylglucosaminidase[240] or glycopeptidyl amylase of almonds.[241,242] These release the oligosaccharide as a 1-amino derivative.

When a dry glycoprotein is heated at 105° with anhydrous hydrazine for 8–12 h all of the asparagine-linked oligosaccharide chains are released (Eq. 4-11).[239,243–245] In addition to the glycosylamine product of step *a*, the corresponding 1–OH and –NH–$NH_2$ compounds are also formed. These are all converted to the stable acetyl derivative (step *b*), after which the oligosaccharides may be characterized by mass spectrometry and high-field NMR spectoscopy.[234,246–248] Crystallization of oligosaccharides in complex mixtures is difficult to impossible.




(4-11)

## 2. Hydrolysis

Most glycosidic linkages are hydrolyzed readily by heating with 1 N mineral acids. The mechanism of the hydrolytic reaction is similar to that employed by the enzyme lysozyme (Chapter 12). Some linkages are unusually sensitive to acid and a few are very resistant. Thus, a variety of conditions may be applied for partial acid hydrolysis as an aid to characterization.[249] Acetolysis, cleavage by acetic anhydride, is also of value.[250] A battery of hydrolytic enzymes specific for sugars that are joined in a given type of glycosidic linkage are available.[251] These are useful in determining sequences of oligosaccharides released from glycoproteins. Radioactive tracer techniques can also be applied.[252]

## 3. Methylation

An important general method in characterization of carbohydrates is the classical **exhaustive methylation** (permethylation). Repeated treatment with a methylating agent such as methyl iodide converts all free OH groups to $OCH_3$ groups. Then, complete acid

hydrolysis, followed by separation of the methylated sugars and their quantitative determination, reveals the relative amounts of **end units** (containing four methoxyl groups), straight **chain units** (containing three methoxyl groups), and **branch points** (containing two methoxyl groups). In addition, the structure of the methylated derivatives provides information on the positions of the linkages in the sugar rings.

After a methylated polysaccharide is subjected to partial hydrolysis, the newly exposed hydroxyl groups (or those created by borohydride reduction of carbonyl groups) can be labeled by ethylation or propylation. One procedure for sequencing complex carbohydrates makes use of high-resolution reversed-phase liquid chromatography to separate the many alkylated oligosaccharides produced by methylation followed by partial acid hydrolysis, reduction, and ethylation.[233] Reductive cleavage of the glycosidic linkages in methylated polysaccharides allows uniquivocal determination of ring size.[253] Branch points may be located by methanolysis of the permethylated polysaccharide followed by conversion of free –OH groups to *p*-bromobenzoate esters. The latter are separated and the circular dichroism (Chapter 23) is measured. Mass spectrometry has also been applied successfully.[246] A simple procedure that can be conducted in any laboratory using thin-layer chromatography is illustrated in Fig. 4-20.[254]

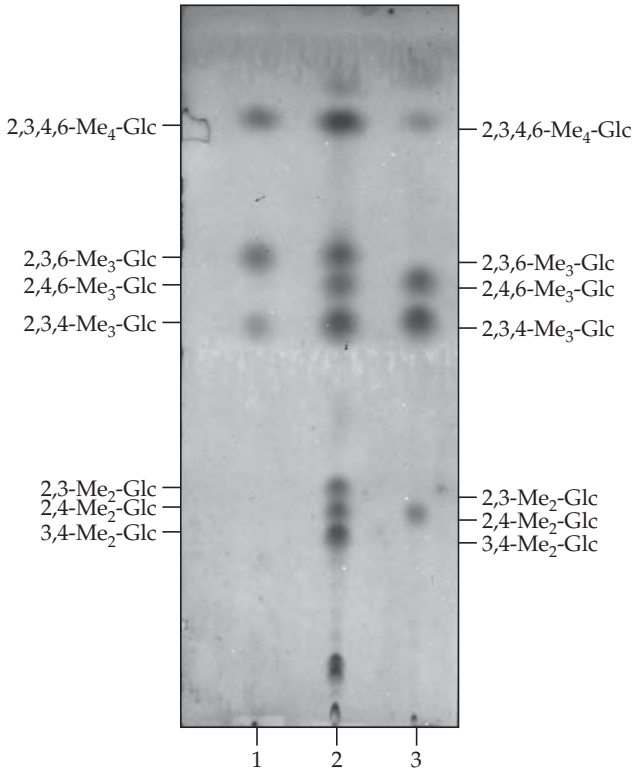
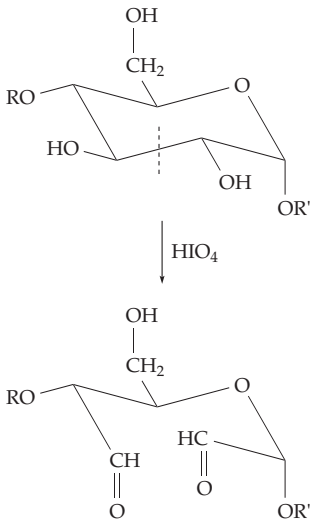



**Figure 4-20** Separation of a mixture of *O*-methylated glucoses by ascending thin-layer chromatography. Whatman K6 TLC plates were used with two ascents of the solvent acetonitrile/chloroform/methanol in the ratio 3/9/2, V/V/V. Courtesy of John Robyt.

## 4. Periodate Oxidation (Smith Degradation)

One of the most important reagents in investigations of carbohydrate structure is periodic acid (or sodium periodate).[255] This reagent oxidatively cleaves C–C bonds bearing adjacent OH or $NH_2$ groups to form dialdehydes (Eq. 4-12). The method is quantitative. After some hours of reaction, excess periodate not consumed in the oxidation can be determined. If three consecutive carbon atoms bear hydroxyl or amino groups, formic acid is liberated from the central atom and can also be measured quantitatively. After destruction of excess periodate the dialdehyde can be reduced by addition of solid sodium borohydride to form stable $CH_2OH$ groups. Following mild acid hydrolysis to split the acyclic acetal linkages, the fragments can be separated and identified. The sequence of reactions is known as the Smith degradation. If sodium borotritide ($NaB^3H_4$) is used for the reduction the fragments will be radioactive and can be located on chromatograms by fluorography. Periodate oxidation can also be used to alter all surface oligosaccharides.[99] Removal of *O*-linked oligosaccharides from glycoproteins can be accomplished by periodate oxidation and alkaline β elimination from the dialdehyde product of Eq. 4-12 if OR′ is part of a serine or other hydroxyamino acid side chain in a protein.[256]



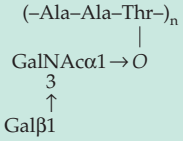

(4-12)

## 5. Nuclear Magnetic Resonance

As in other areas of biochemistry, NMR has become an extremely important tool in carbohydrate research. The mixtures of anomers of various ring

## BOX 4-D ANTIFREEZE AND ICE-NUCLEATION PROTEINS

Fish living in Arctic and Antarctic waters may encounter temperatures as low as –1.9°C. The freezing point depression provided by dissolved salts and proteins in the blood is insufficient to protect the fish from freezing. As winter approaches, they synthesize and accumulate in their blood serum a series of eight or more special antifreeze proteins.[a–d] One type of antifreeze glycoprotein from winter flounder contains the following unit repeated 17–50 times.




Destruction of the galactosyl residues by oxidation with periodate, acetylation of the free hydroxyl groups of the oligosaccharides, or their removal by β elimination all lead to loss of antifreeze activity.

The same fish contain a second series of alanine-rich antifreeze polypeptides that are *not* glycosylated but which exist as amphipathic helices. One of these (Type I) contains ~40 residues in a single helix.[e–i] A third family of antifreeze proteins (Type II), found in the sea raven are globular proteins, rich in cysteine and β structures. They are members of the lectin family.[d,j] A fourth type (Type III) found in the sea pout and some other fishes are 62- to 66-residue globular proteins containing an orthogonal β sandwich structure.[k,l] Messenger RNA molecules coding for the antifreeze proteins are found in the livers of flounder in the winter but are absent in the summer.[m]

Antifreeze proteins, that are 3 – 4 times as effective as those in fish, have been isolated from some insects and other arthropods.[m,n,o] They help beetle larvae to overwinter.[m] The insect proteins have a parallel β helix structure resembling that in Fig. 2-17 and stabilized by S — S bridges.[o,p] Some plants also synthesize antifreeze proteins.[n,q,r] One of these, isolated from carrots, is a member of the leucine-rich-repeat family.[q]

How do antifreeze proteins work? The major effect is to greatly slow the freezing rather than to decrease the freezing point. The proteins apparently accomplish this by binding to the surfaces of small ice crystals and preventing their growth.[d,f,h,i,k,l,s,t] This provides the fish with enough time for the blood to pass back into the liver, in which a high enough temperature is maintained to melt any microcrystals before the blood again circulates through the colder tissues. Some of the proteins have clusters of polar side chains that bind to specific faces of the ice crystals and inhibit growth.[s]

A few fishes tolerate a high internal osmotic pressure and accumulate glycerol in their blood up to a concentration of 0.4 M.[t] Insects may accumulate up to 3 M glycerol and some species utilize various other cryoprotectants, such as mannitol, sorbitol, erythritol, threitol, trehalose, glucose, fructose, proline, and alanine.[u] Some amphibians and reptiles can survive freezing and recover fully. For the most studied wood frog, rapid freezing is fatal, but slow freezing leaves the frog, whose heart ceases to function, with a 200-fold increased glucose concentration and a decreased water content in its organs. It resumes normal activities within 14–24 h of thawing.[t]

Having an effect opposite to that of the antifreeze proteins are surface proteins of some bacteria of the genera *Pseudomonas*, *Erwinia*, and *Xanthomonas*. These proteins provide nuclei for growth of ice crystals from supercooled water.[v,w]

---


[a] Feeney, R. E., Burcham, T. S., and Yeh, Y. (1986) *Ann. Rev. Biophys. Biophys. Chem.* **15**, 59–78
[b] Eastman, J. T., and DeVries, A. L. (1986) *Sci. Am.* **255** (Nov), 106–114
[c] Davies, P. L., and Hew, C. L. (1990) *FASEB J.* **4**, 2460–2468
[d] Gronwald, W., Loewen, M. C., Lix, B., Daugulis, A. J., Sönnichsen, F. D., Davies, P. L., and Sykes, B. D. (1998) *Biochemistry* **37**, 4712–4721
[e] Chakrabartty, A., Ananthanarayanan, V. S., and Hew, C. L. (1989) *J. Biol. Chem.* **264**, 11307–11312
[f] Yang, D. S. C., Sax, M., Chakrabartty, A., and Hew, C. L. (1988) *Nature (London)* **333**, 232–237
[g] Madura, J. D., Wierzbicki, A., Harrington, J. P., Maughon, R. H., Raymond, J. A., and Sikes, C. S. (1994) *J. Am. Chem. Soc.* **116**, 417–418
[h] Sicheri, F., and Yang, D. S. C. (1995) *Nature (London)* **375**, 427–431
[i] Wierzbicki, A., Taylor, M. S., Knight, C. A., Madura, J. D., Harrington, J. P., and Sikes, C. S. (1996) *Biophys. J.* **71**, 8–18
[j] Ng, N. F. L., and Hew, C. L. (1992) *J. Biol. Chem.* **267**, 16069–16075
[k] Chao, H., Sönnichsen, F. D., DeLuca, C. I., Sykes, B. D., and Davies, P. L. (1994) *Protein Sci.* **3**, 1760–1769
[l] Jia, Z., DeLuca, C. I., Chao, H., and Davies, P. L. (1996) *Nature (London)* **384**, 285–288
[m] Gourlie, B., Lin, Y., Price, J., DeVries, A. L., Powers, D., and Huang, R. C. C. (1984) *J. Biol. Chem.* **259**, 14960–14965
[n] Li, N., Chibber, B. A. K., Castellino, F. J., and Duman, J. G. (1998) *Biochemistry* **37**, 6343–6350
[o] Graether, S. P., Kuiper, M. J., Gagné, S. M., Walker, V. K., Jia, Z., Sykes, B. D., and Davies, P. L. (2000) *Nature (London)* **406**, 325 – 328
[p] Liou, Y.-C., Tocilj, A., Davies, P. L., and Jia, Z. (2000) *Nature (London)* **406**, 322–324
[q] Worrall, D., Elias, L., Ashford, D., Smallwood, M., Sidebottom, C., Lillford, P., Telford, J., Holt, C., and Bowles, D. (1998) *Science* **282**, 115–117
[r] Sidebottom, C., Buckley, S., Pudney, P., Twigg, S., Jarman, C., Holt, C., Telford, J., McArthur, A., Worrall, D., Hubbard, R., and Lillford, P. (2000) *Nature (London)* **406**, 256
[s] Knight, C. A., (2000) *Nature (London)* **406**, 249–251
[t] Costanzo, J. P., Lee, R. E., Jr., DeVries, A. L., Wang, T., and Layne, J. R., Jr. (1995) *FASEB J.* **9**, 351–352
[u] Storey, K. B., and Storey, J. M. (1990) *Sci. Am.* **263** (Dec), 92–97
[v] Wolber, P., and Warren, G. (1989) *Trends Biochem. Sci.* **14**, 179–182
[w] Gurian-Sherman, D., and Lindow, S. E. (1993) *FASEB J.* **7**, 1338–1343

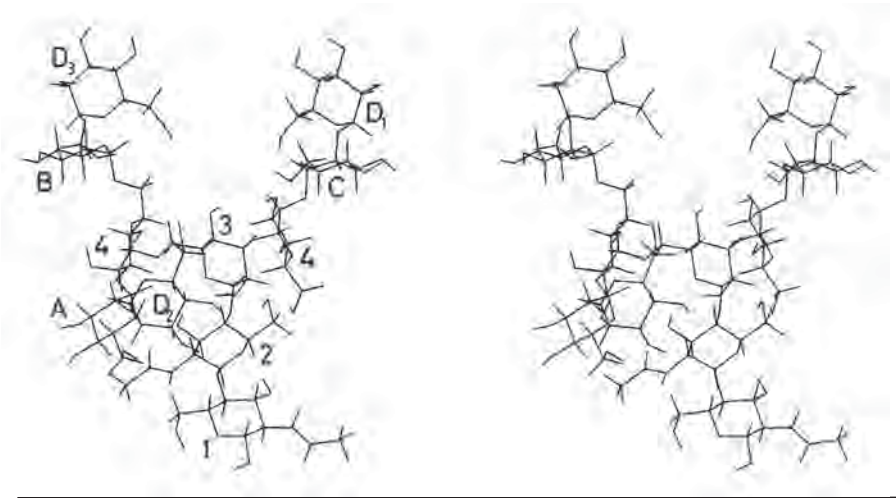



**Figure 4-21** Stereoscopic view of an N-linked oligosaccharide whose structure has been deduced by two-dimensional NMR spectroscopy and energy calculations. This is one of a range of allowed conformations. From Homans *et al.*[257]

sizes (Eq. 4-1) can be analyzed with proton or $^{13}C$ NMR.[10,257] A variety of newer NMR techniques, some of which have been described in Chapter 3, have been applied.[28,258–267] A problem with NMR spectroscopy of these compounds has been the lack of the large number of nuclear Overhauser enhancements (NOEs) that can be observed.[268,269] Nevertheless, when combined with energy calculations it is possible to use NMR measurements to deduce conformations of sugar rings, three-dimensional structures, and the degree of conformational flexibility in various parts of N-linked oligosaccharides (Fig. 4-21). Measurement of C–O–C –C spin-coulping constants is also of value.[270] Use of multidimensional NMR has permitted analysis of mixtures of cellulose oligosaccharides.[269]

## References


1. Pigman, W. W., and Horton, D., eds. (1972) *The Chemistry of Carbohydrates*, 2nd ed., Academic Press, New York
2. Guthrie, R. D. (1974) *Guthrie and Honeyman's Introduction to Carbohydrate Chemistry*, 4th ed., Oxford Univ. Press (Clarendon), London and New York (p. 17)
3. Binkley, R. W. (1988) *Modern Carbohydrate Chemistry*, Dekker, New York
4. Shalenberger, R. S. (1982) *Advanced Sugar Chemistry: Principles of Sugar Stereochemistry*, AVI Publ. Co., Westport, Connecticut
5. El-Khadem, H. S. (1988) *Carbohydrate Chemistry*, Academic Press, San Diego, California
6. Prince, R. C., Gunson, D. E., Leigh, J. S., and McDonald, G. G. (1982) *Trends Biochem. Sci.* **7**, 239–240
7. Wertz, P. W., Garver, J. C., and Anderson, L. (1981) *J. Am. Chem. Soc.* **103**, 3916–3922
8. Ha, S., Gao, J., Tidor, B., Brady, J. W., and Karplus, M. (1991) *J. Am. Chem. Soc.* **113**, 1553–1557
9. Angyal, S. J. (1979) in *Asymmetry of Carbohydrates* (Harmon, R. E., ed), p. 15, Decker, New York
10. Gray, G. R. (1976) *Acc. Chem. Res.* **9**, 418
11. Pierce, J., Serianni, A. S., and Barker, R. (1985) *J. Am. Chem. Soc.* **107**, 2448–2456
12. Maple, S. R., and Allerhand, A. (1987) *J. Am. Chem. Soc.* **109**, 3168–3169
13. Goux, W. J. (1985) *J. Am. Chem. Soc.* **107**, 4320–4327
14. Barker, S. A., and Baggett, N. (1978) *Trends Biochem. Sci.* **3**, 140–141
15. Benson, A. A. (1963) *Adv. Lipid Res.* **1**, 387–394
16. Benning, C., Beatty, J. T., Prince, R. C., and Somerville, C. R. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 1561–1565
17. Dijkema, C., Kester, H. C. M., and Visser, J. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 14–18
18. Tarczynski, M. C., Jensen, R. G., and Bohnert, H. J. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2600–2604
19. Yamanouchi, T., Tachibana, Y., Sekino, N., Akanuma, H., Akaoka, I., and Miyashita, H. (1994) *J. Biol. Chem.* **269**, 9664–9668
20. Nadano, D., Iwasaki, M., Endo, S., Kitajima, K., Inoue, S., and Inoue, Y. (1986) *J. Biol. Chem.* **261**, 11550–11557
21. Schauer, R. (1985) *Trends Biochem. Sci.* **10**, 357–360
22. Faillard, H. (1989) *Trends Biochem. Sci.* **14**, 237–241
23. Roth, J., Kempf, A., Reuter, G., Schauer, R., and Gehring, W. J. (1992) *Science* **256**, 673–675
24. Powell, L. D., and Varki, A. (1995) *J. Biol. Chem.* **270**, 14243–14246
25. Stoddart, J. F. (1971) *Stereochemistry of Carbohydrates*, Wiley-Interscience, New York
26. Brady, J. W. (1986) *J. Am. Chem. Soc.* **108**, 8152–8160

26a. Marszalek, P. E., Pang, Y.-P., Li, H., Yazal, J. E., Oberhauser, A. F., and Fernandez, J. M. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96**, 7894–7898

27. Venkataraman, G., Sasisekharan, V., Cooney, C. L., Langer, R., and Sasisekharan, R. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6171–6175
28. Hajduk, P. J., Horita, D. A., and Lerner, L. E. (1993) *J. Am. Chem. Soc.* **115**, 9196–9201
29. Wolfe, S., Whangbo, M.-H., and Mitchell, D. J. (1979) *Carbohydrate Res.* **69**, 1–26
30. Kirby, A. J. (1983) *The Anomeric Effect and Related Stereoelectronic Effects at Oxygen*, Springer, Berlin
31. Perrin, C. L., Armstrong, K. B., and Fabian, M. A. (1994) *J. Am. Chem. Soc.* **116**, 715–722
32. Cui, W., Li, F., and Allinger, N. L. (1993) *J. Am. Chem. Soc.* **115**, 2943–2951
33. Westhof, E., and Sundaralingam, M. (1980) *J. Am. Chem. Soc.* **102**, 1493–1500

33a. French, A. D., Dowd, M. K., and Reilly, P. J. (1997) *J. Mol. Struct. (Theochem)*, 395–396

34. Gabb, H. A., and Harvey, S. C. (1993) *J. Am. Chem. Soc.* **115**, 4218–4227
35. Plavec, J., Tong, W., and Chattopadhyaya, J. (1993) *J. Am. Chem. Soc.* **115**, 9734–9746
36. Ellervik, U., and Magnusson, G. (1994) *J. Am. Chem. Soc.* **116**, 2340–2347

36a. French, A. D., and Murphy, V. G. (1973) *Carbohydr. Res.* **27**, 391–406

37. Robyt, J. F., and White, B. J. (1996) *Biochemical Techniques, Theory and Practice*, 2nd ed., Waveland Press, Prospect Heights, Illinois
38. Lipták, A., Fügedi, P., Szurmai, Z., and Harangi, J. (1990) *CRC Handbook of Oligosaccharides. Vol. I: Disaccharides*, CRC Press, Boca Raton, Florida
39. Cerami, A. (1986) *Trends Biochem. Sci.* **11**, 311–314
40. Watkins, N. G., Neglia-Fisher, C. I., Dyer, D. G., Thorpe, S. R., and Baynes, J. W. (1987) *J. Biol. Chem.* **262**, 7207–7212
41. Súarez, G., Rajaram, R., Oronsky, A. L., and Gawinowicz, M. A. (1989) *J. Biol. Chem.* **264**, 3674–3679

## References


42. Takata, K., Horiuchi, S., Araki, N., Shiga, M., Saitoh, M., and Morino, Y. (1988) *J. Biol. Chem.* **263**, 14819–14825
43. Hayase, F., Nagaraj, R. H., Miyata, S., Njoroge, F. G., and Monnier, V. M. (1989) *J. Biol. Chem.* **264**, 3758–3764
44. Goldberg, R. N., and Tewari, Y. B. (1989) *J. Biol. Chem.* **264**, 9897–9900
45. Tewari, Y. B., and Goldberg, R. N. (1989) *J. Biol. Chem.* **264**, 3966–3971
46. Goldberg, R. N., Tewari, Y. B., and Ahluwalia, J. C. (1989) *J. Biol. Chem.* **264**, 9901–9904
47. Weissborn, A. C., Rumley, M. K., and Kennedy, E. P. (1991) *J. Biol. Chem.* **266**, 8062–8067
48. Reymond, P., Grünberger, S., Paul, K., Müller, M., and Farmer, E. E. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 4145–4149
49. Nacharaju, P., and Acharya, A. S. (1992) *Biochemistry* **31**, 12673–12679
50. Baynes, J. W., and Monnier, V. M., eds. (1989) *the Maillard Reaction in Aging, Diabetes and Nutrition*, Liss, New York
51. Larsen, M. L., Hørder, M., and Mogensen, E. F. (1990) *N. Engl. J. Med.* **323**, 1021–1025
52. Luthra, M., and Balasubramanian, D. (1993) *J. Biol. Chem.* **268**, 18119–18127
53. Urbanowski, J. C., Cohenford, M. A., and Dain, J. A. (1982) *J. Biol. Chem.* **257**, 111–115
54. Grandhee, S. K., and Monnier, V. M. (1991) *J. Biol. Chem.* **266**, 11649–11653
55. Brownlee, M., Cerami, A., and Vlassara, H. (1988) *N. Engl. J. Med.* **318**, 1315–1320
56. Papoulis, A., Al-Abed, Y., and Bucala, R. (1995) *Biochemistry* **34**, 648–655
57. Wells-Knecht, K. J., Zyzak, D. V., Litchfield, J. E., Thorpe, S. R., and Baynes, J. W. (1995) *Biochemistry* **34**, 3702–3709
58. French, D. (1975) in *International Review of Biochemistry*, Vol. 5 (Whelan, W. J., ed), p. 269, Univ. Park Press, Baltimore, Maryland (Series One)
59. Atkins, E. D. T., ed. (1986) *Polysaccharides, Topics in Structure and Morphology*, VCH Publisher, New York
60. Whistler, R. L., and BeMiller, J. N., eds. (1993) *Industrial Gums*, 3rd ed., Academic Press, San Diego, California
61. Rao, V. S. R., and Sathyanarayana, B. K. (1972) *Biopolymers* **11**, 1379–1394
62. French, A. D., and Dowd, M. K. (1993) in *Cellulostics: Chemical,Biochemical and Materials Aspects* (Kennedy, J. F., Phillips, G. O., and Williams, P. A., eds), pp. 51–56, Ellis Harwood, New York
63. Duda, C. A., and Stevens, E. S. (1993) *J. Am. Chem. Soc.* **115**, 8487–8488
64. Liébecq, C., ed. (1992) *Biochemical Nomenclature*, Portland Press, London and Chapel Hill, North Carolina (for the International Union of Biochemistry and Molecular Biology)
65. Preston, R. D. (1986) *Trends Biochem. Sci.* **11**, 377–380
66. Ross, P., Mayer, R., and Benziman, M. (1991) *Microbiol. Rev.* **55**, 35–58
67. Sugiyama, J., Vuong, R., and Chanzy, H. (1991) *Macromolecules* **24**, 4168–4175
68. Dudley, R. L., Fyfe, C. A., Stephenson, P. J., Deslandes, Y., Hamer, G. K., and Marchessault, R. H. (1983) *J. Am. Chem. Soc.* **105**, 2469–2472
69. Gebler, K., Kraub, N., Steiner, T., Betzel, C., Sandmann, C., and Saenger, W. (1994) *Science* **266**, 1027–1029

69a. Langan, P., Nishiyama, Y., and Chanzy, H. (1999) *J. Am. Chem. Soc.* **121**, 9940–9946

70. Grover, J. A. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 475–504, Academic Press, San Diego, California
71. French, D. (1973) *J. Animal Sci.* **37**, 1048–1061
72. Maddelein, M. L., Libessart, N., Bellanger, F., Delrue, B., D'Hulst, C., Van den Koornhuyse, N., Fontaine, T., Wieruszeski, J. M., Decq, A., and Ball, S. (1994) *J. Biol. Chem.* **269**, 25150–25157
73. Jane, J., Kasemsuwan, T., Leas, S., Zobel, H., and Robyt, J. F. (1994) *Starch/Stärke* **46**, 121–129
74. Meyer, K. H. (1943) *Adv. Enzymol.* **3**, 109–136
75. French, D. (1969) in *Symposium on Foods: Carbohydrates and their Roles* (Schultz, H. W., ed), pp. 26–54, AVI Publ. Co., Westport, Connecticut
76. Rundle, R. E., and French, D. (1943) *J. Am. Chem. Soc.* **65**, 1707–1710
77. Hybl, A., Rundle, R. E., and Williams, D. E. (1965) *J. Am. Chem. Soc.* **87**, 2779–2788
78. Teitelbaum, R. C., Ruby, S. L., and Tobin, J. M. (1980) *J. Am. Chem. Soc.* **102**, 3322–3328
79. Winter, W. T., and Sarko, A. (1974) *Biopolymers* **13**, 1461–1482
80. Kainuma, K., and French, D. (1972) *Biopolymers* **11**, 2241–2250
81. Wu, H.-C., H, and Sarko, A. (1978) *Carbohydr. Res.* **61**, 7–25
82. Gidley, M. J., and Bociek, S. M. (1988) *J. Am. Chem. Soc.* **110**, 3820–3829
83. Lomako, J., Lomako, W. M., Whelan, W. J., Dombro, R. S., Neary, J. T., and Norenberg, M. D. (1993) *FASEB J.* **7**, 1386–1393
84. Chen, W., Avison, M. J., Zhu, X. H., and Shulman, R. G. (1993) *Biochemistry* **32**, 11483–11487
85. Ohana, P., Delmer, D. P., Steffens, J. C., Matthews, D. E., Mayer, R., and Benziman, M. (1991) *J. Biol. Chem.* **266**, 13742–13745
86. Mol, P. C., Park, H., Mullins, J. T., and Cabib, E. (1994) *J. Biol. Chem.* **269**, 31267–31274
87. Yamamoto, R., and Nevins, D. J. (1978) *Carbohydrate Res.* **67**, 275–280
88. Marchessault, R. H., and Deslandes, Y. (1979) *Carbohydrate Res.* **75**, 231–242
89. Kang, M. S., and Cabib, E. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 5808–5812
90. Harada, T., Terasaki, M., and Harada, A. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 427–445, Academic Press, San Diego, California
91. Walker, G. J. (1978) in *International Review of Biochemistry*, Vol. 16 (Manners, D. J., ed), p. 75, Univ. Park Press, Baltimore, Maryland
92. de Belder, A. N. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 399–425, Academic Press, San Diego, California
93. Yang, X., Miller, M. A., Yang, R., Evans, D. F., and Edstrom, R. D. (1990) *FASEB J.* **4**, 3140–3143
94. Tsujisaka, Y., and Mitsuhashi, M. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 447–460, Academic Press, San Diego, California
95. Arnott, S., Fulmer, A., Scott, W. E., Dea, I. C. M., Moorehouse, R., and Rees, D. A. (1974) *J. Mol. Biol.* **90**, 269–284
96. Shibata, N., Ikuta, K., Imai, T., Satoh, Y., Satoh, R., Suzuki, A., Kojima, C., Kobayashi, H., Hisamichi, K., and Suzuki, S. (1995) *J. Biol. Chem.* **270**, 1113–1122
97. Kollár, R., Petráková, E., Ashwell, G., Robbins, P. W., and Cabib, E. (1995) *J. Biol. Chem.* **270**, 1170–1178
98. Preston, R. D. (1968) *Sci. Am.* **218**(Jun), 102–108
99. Turvey, J. R. (1978) in *International Review of Biochemistry*, Vol. 16 (Manners, D. J., ed), Univ. Park Press, Baltimore, Maryland
100. Kirkwood, S. (1974) *Ann. Rev. Biochem.* **43**, 401–417
101. Minke, R., and Blackwell, J. (1978) *J. Mol. Biol.* **120**, 167–181
102. Roberts, G. A. F. (1992) *Chitin Chemistry*, MacMillan, London
103. Blackwell, J., and Weih, M. A. (1980) *J. Mol. Biol.* **137**, 49–60
104. Rees, D. A., and Welsh, E. J. (1977) *Angew. Chem. Int. Ed. Engl.* **16**, 214–224
105. Lee, Y. C. (1992) *FASEB J.* **6**, 3193–3200
106. Laurent, T. C., and Fraser, J. R. E. (1992) *FASEB J.* **6**, 2397–2404
107. Hardingham, T. E., and Fosang, A. J. (1992) *FASEB J.* **6**, 861–870
108. Fransson, L. (1987) *Trends Biochem. Sci.* **12**, 406–411
109. Morris, E. R., Rees, D. A., and Welsh, E. J. (1980) *J. Mol. Biol.* **138**, 383–400
110. David, G. (1993) *FASEB J.* **7**, 1023–1030
111. Yamada, S., Murakami, T., Tsuda, H., Yoshida, K., and Sugahara, K. (1995) *J. Biol. Chem.* **270**, 8696–8705
112. Lane, D. A., and Björk, I., eds. (1992) *Heparin And Related Polysaccharides*, Plenum, New York
113. Edens, R. E., Fromm, J. R., Linhardt, R. J., and Weiler, J. M. (1995) *Biochemistry* **34**, 2400–2407
114. Linhardt, R. J., Ampofo, S. A., Fareed, J., Hoppensteadt, D., Mulliken, J. B., and Folkman, J. (1992) *Biochemistry* **31**, 12441–12445
115. Casu, B., Petitou, M., Provasoli, M., and Sinaÿ, P. (1988) *Trends Biochem. Sci.* **13**, 221–225
116. Oates, J. A., and Wood, A. J. J. (1991) *N. Engl. J. Med.* **324**, 1565–1574
117. Rosenberg, R. D., and Lam, L. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1218–1222
118. Lindahl, U., Kusche-Gullberg, M., and Kjelln, L. (1998) *J. Biol. Chem.* **273**, 24979–24982
119. Mulloy, B. A., Ribeiro, A.-C., Alves, A.-P., Vieira, R. P., and Mourao, P. A. S. (1994) *J. Biol. Chem.* **269**, 22113–22123
120. Gibeaut, D., and Carpita, N. C. (1994) *FASEB J.* **8**, 904–915
121. Hayashi, T., and Matsuda, K. (1981) *J. Biol. Chem.* **256**, 11117–11122
122. Millane, R. P. (1992) in *Frontiers in Carbohydrate Research-2* (Chandrasekaran, R., ed), pp. 168–190, Elsevier, London
123. Rolin, C. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 257–293, Academic Press, San Diego, California
124. Arnott, S., Scott, W. E., Rees, D. A., and McNab, C. G. A. (1974) *J. Mol. Biol.* **90**, 253–267
125. Selby, H. H., and Whistler, R. L. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 87–103, Academic Press, San Diego, California
126. Albersheim, P., Bauer, W. D., Keestra, K., Talmadge, K. W. (1973) *Biogenesis of Plant Cell Wall Polysaccharides*, (Loewus, F., ed.), Academic Press, New York
127. Rees, D. A. (1972) *Biochem. J.* **126**, 257–273
128. Therkelsen, G. H. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 145–180, Academic Press, San Diego, California
129. Clare, K. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 105–143, Academic Press, San Diego, California
130. Marsh, M. E., Chang, D. C., and King, G. C. (1992) *J. Biol. Chem.* **267**, 20507–20512

130a. Vinogradov, E., and Bock, K. (1999) *Angew. Chem. Ed. Engl.* **38**, 671–674

131. Sutherland, I. W. (1979) *Trends Biochem. Sci.* **4**, 55–59

## References


132. Kang, K. S., and Pettitt, D. J. (1993) in *Industrial Gums*, 3rd ed. (Whistler, R. L., and BeMiller, J. N., eds), pp. 341–397, Academic Press, San Diego, California
133. Wolucka, B. A., McNeil, M. R., de Hoffmann, E., Chonjnacki, T., and Brennan, P. J. (1994) *J. Biol. Chem.* **269**, 23328–23335
134. Kobata, A. (1984) in *Biology of Carbohydrates*, Vol. 2 (Ginsburg, V., and Robbins, P. W., eds), pp. 87–161, Wiley, New York
135. Lennarz, W. J., ed. (1980) *The Biochemistry of Glycoproteins and Proteoglycans*, Plenum, New York
136. Montreul, J. (1982) in *Comprehensive Biochemistry*, 1st ed., Vol. 19B, Part II (Florkin, M., and Stotz, E. H., eds), pp. 1–188, Elsevier, Amsterdam
137. Rademacher, T. W., Parekh, R. B., and Dwek, R. A. (1988) *Ann. Rev. Biochem.* **57**, 785–838
138. Roberts, D. D., and Mecham, R. P., eds. (1993) *Cell Surface and Extracellular Glycoconjugates: Structure and Function*, Academic Press, San Diego, California

138a. Varki, A., Cummings, R., Esko, J., Freeze, H., Hart, G., and Marth, J., eds. (1999) *Essentials of Glycobiology*, Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York

139. Fukuda, M., and Hindsgaul, O., eds. (1994) *Molecular Glycobiology*, Oxford Publishing, Oxford
140. Williams, R. L., Greene, S. M., and McPherson, A. (1987) *J. Biol. Chem.* **262**, 16020–16031
141. Hofsteenge, J., Müller, D. R., de Beer, T., Löffler, A., Richter, W. J., and Vliegenthart, J. F. G. (1994) *Biochemistry* **33**, 13524–13530

141a. Iozzo, R. V. (1998) *Ann. Rev. Biochem.* **67**, 609–652

141b. Bernfield, M., Götte, M., Park, P. W., Reizes, O., Fitzgerald, M. L., Lincecum, J., and Zako, M. (1999) *Ann. Rev. Biochem.* **68**, 729–777

142. Sugahara, K., Ohkita, Y., Shibata, Y., Yoshida, K., and Ikegami, A. (1995) *J. Biol. Chem.* **270**, 7204–7212
143. Funderburgh, J. L., Funderburgh, M. L., Brown, S., Vergnes, J.-P., Hassell, J. R., Mann, M. M., and Conrad, G. W. (1993) *J. Biol. Chem.* **268**, 11874–11880
144. Neame, P. J., Christner, J. E., and Baker, J. R. (1987) *J. Biol. Chem.* **262**, 17768–17778
145. Scott, J. E. (1987) *Trends Biochem. Sci.* **12**, 318–321
146. Rosenberg, L., Tang, L.-H., Pal, S., Johnson, T. L., and Choi, H. U. (1988) *J. Biol. Chem.* **263**, 18071–18077
147. Choi, H. U., Johnson, T. L., Pal, S., Tang, L.-H., Rosenberg, L., and Neame, P. J. (1989) *J. Biol. Chem.* **264**, 2876–2884
148. Kokenyesi, R., and Bernfield, M. (1994) *J. Biol. Chem.* **269**, 12304–12309
149. Rosenberg, L. (1975) in *Dynamics of Connective Tissue Macromolecules* (Burleigh, P. M. C., and Poole, A. R., eds), p. 107, North-Holland Publ., Amsterdam
150. Chen, J., and Varner, J. E. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4399–4403
151. Kieliszewski, M., deZacks, R., Leykam, J. F., and Lamport, D. T. A. (1992) *Plant Physiol.* **98**, 919–926
152. Kieliszewski, M. J., O'Neill, M., Leykam, J., and Orlando, R. (1995) *J. Biol. Chem.* **270**, 2541–2549
153. Eckhardt, A. E., Timpte, C. S., Abernethy, J. L., Toumadje, A., Johnson, W. C., and Hill, R. L. (1987) *J. Biol. Chem.* **262**, 11339–11344
154. Timpte, C. S., Eckhardt, A. E., Abernethy, J. L., and Hill, R. L. (1988) *J. Biol. Chem.* **263**, 1081–1088
155. Gupta, R., and Jentoft, N. (1989) *Biochemistry* **28**, 6114–6121
156. Carter, S. R., Slomiany, A., Gwozdzinski, K., Liau, Y. H., and Slomiany, B. L. (1988) *J. Biol. Chem.* **263**, 11977–11984
157. Jentoft, N. (1990) *Trends Biochem. Sci.* **15**, 291–294
158. Butenhof, K. J., and Gerken, T. A. (1993) *Biochemistry* **32**, 2650–2663
159. Toribara, N. W., Roberton, A. M., Ho, S. B., Kuo, W.-L., Gum, E., Hicks, J. W., Gum, J. R., Byrd, J. C., Siddiki, B., and Kim, Y. S. (1993) *J. Biol. Chem.* **268**, 5879–5885
160. Sadler, J. E. (1984) in *"Biology of Carbohydrates"*, Vol. 2 (Ginsburg, V., and Robbins, P. W., eds), pp. 199–288, Wiley, New York

160a. Perez-Vilar, J., and Hill, R. L. (1999) *J. Biol. Chem.* **274**, 31751–31754

161. Fukuda, M., Carlsson, S. R., Klock, J. C., and Dell, A. (1986) *J. Biol. Chem.* **261**, 12796–12806
162. Gerken, T. A., and Jentoft, N. (1987) *Biochemistry* **26**, 4689–4699
163. Freeze, H. H., and Wolgast, D. (1986) *J. Biol. Chem.* **261**, 127–134
164. Carver, J. P., and Brisson, J.-R. (1984) in *Biology of Carbohydrates*, Vol. 2 (Ginsburg, V., and Robbins, P. W., eds), pp. 289–329, Wiley, New York
165. Bush, C. A., Duben, A., and Ralapati, S. (1980) *Biochemistry* **19**, 501–504
166. Finne, J., and Krusius, T. (1979) *Eur. J. Biochem.* **102**, 583–588
167. Järnefelt, J., Finne, J., Krusius, T., and Rauvala, H. (1978) *Trends Biochem. Sci.* **3**, 110–114
168. Kornfeld, R., and Kornfeld, S. (1980) in *The Biochemistry of Glycoproteins and Proteoglycans* (Lennarz, W. J., ed), Plenum, New York
169. Fukuda, M., Dell, A., and Fukuda, M. N. (1984) *J. Biol. Chem.* **259**, 4782–4791
170. Hernández, L. M., Olivero, I., Alvarado, E., and Larriba, G. (1992) *Biochemistry* **31**, 9823–9831
171. Taguchi, T., Seko, A., Kitajima, K., Muto, Y., Inoue, S., Khoo, K.-H., Morris, H. R., Dell, A., and Inoue, Y. (1994) *J. Biol. Chem.* **269**, 8762–8771
172. Wieland, F., Paul, G., and Sumper, M. (1985) *J. Biol. Chem.* **260**, 15180–15185
173. Funderburgh, J. L., Funderburgh, M. L., Mann, M. M., and Conrad, G. W. (1991) *J. Biol. Chem.* **266**, 14226–14231
174. Sharon, N., and Lis, H. (1989) *Science* **246**, 227–229
175. Sharon, N., and Lis, H. (1993) *Sci. Am.* **268**(Jan), 82–89
176. Opdenakker, G., Rudd, P. M., Ponting, C. P., and Dwek, R. A. (1993) *FASEB J.* **7**, 1330–1337
177. Sharon, N. (1993) *Trends Biochem. Sci.* **18**, 221–226
178. Barondes, S. H. (1988) *Trends Biochem. Sci.* **13**, 480–482
179. Sphyris, N., Lord, J. M., Wales, R., and Roberts, L. M. (1995) *J. Biol. Chem.* **270**, 20292–20297
180. Sumner, J. B., and Howell, S. F. (1936) *J. Bacteriol.* **32**, 227–237
181. Bourne, Y., Rougés, P., and Cambillau, C. (1990) *J. Biol. Chem.* **265**, 18161–18165
182. Sharon, N., and Lis, H. (1990) *FASEB J.* **4**, 3198–3208
183. Mandal, D. K., and Brewer, C. F. (1993) *Biochemistry* **32**, 5116–5120
184. Derewenda, Z., Yariv, J., Helliwell, J. R., Kalb, A. J., Dodson, E. J., Papiz, M. Z., Wan, T., and Campbell, J. (1989) *EMBO J.* **8**, 2189–2193
185. Drickamer, K. (1988) *J. Biol. Chem.* **263**, 9557–9560
186. Krusius, T., Gehlsen, K. R., and Ruoslahti, E. (1987) *J. Biol. Chem.* **262**, 13120–13125
187. Gitt, M. A., and Barondes, S. H. (1991) *Biochemistry* **30**, 82–89
188. Barondes, S. H., Cooper, D. N. W., Gitt, M. A., and Leffler, H. (1994) *J. Biol. Chem.* **269**, 20807–20810

188a. Varela, P. F., Solís, D., Díaz-Maurino, T., Kaltner, H., Gabius, H.-J., and Romero, A. (1999) *J. Mol. Biol.* **294**, 537–549

189. Chiu, M. L., Parry, D. A. D., Feldman, S. R., Klapper, D. G., and O'Keefe, E. J. (1994) *J. Biol. Chem.* **269**, 31770–31776
190. Madsen, P., Rasmussen, H. H., Flint, T., Gromov, P., Kruse, T. A., Honoré, B., Vorum, H., and Celis, J. E. (1995) *J. Biol. Chem.* **270**, 5823–5829
191. Cho, M., and Cummings, R. D. (1995) *J. Biol. Chem.* **270**, 5207–5212
192. Gitt, M. A., Wiser, M. F., Leffler, H., Herrmann, J., Xia, Y.-R., Massa, S. M., Cooper, D. N. W., Lusis, A. J., and Barondes, S. H. (1995) *J. Biol. Chem.* **270**, 5032–5038
193. Weis, W. I., Kahn, R., Fourme, R., Drickamer, K., and Hendrickson, W. A. (1991) *Science* **254**, 1608–1615
194. Vyas, N. K., Vyas, M. N., and Quiocho, F. A. (1988) *Science* **242**, 1290–1295
195. Shaanan, B., Lis, H., and Sharon, N. (1991) *Science* **254**, 862–865
196. Cygler, M., Rose, D. R., and Bundle, D. R. (1991) *Science* **253**, 442–445
197. Ångström, J., Teneberg, S., and Karlsson, K.-A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11859–11863
198. Sauter, N. K., Hanson, J. E., Glick, G. D., Brown, J. H., Crowther, R. L., Park, S.-J., Skehel, J. J., and Wiley, D. C. (1992) *Biochemistry* **31**, 9609–9621
199. Janakiraman, M. N., White, C. L., Laver, W. G., Air, G. M., and Luo, M. (1994) *Biochemistry* **33**, 8172–8179
200. Bossart-Whitaker, P., Carson, M., Babu, Y. S., Smith, C. D., Laver, W. G., and Air, G. M. (1993) *J. Mol. Biol.* **232**, 1069–1083
201. Crennell, S. J., Garman, E. F., Laver, E. G., Vimr, E. R., and Taylor, G. L. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 9852–9856
202. Colli, W. (1993) *FASEB J.* **7**, 1257–1264
203. Ofek, I., Beachey, E. H., and Sharon, N. (1978) *Trends Biochem. Sci.* **3**, 159–160
204. Lindberg, F., Lund, B., and Normark, S. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1891–1895
205. Cho, J.-W., and Troy, F. A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11427–11431
206. Baumann, H., Brisson, J., Michon, F., Pon, R., and Jennings, H. J. (1993) *Biochemistry* **32**, 4007–4013
207. Falk, P. G., Bry, L., Holgersson, J., and Gordon, J. I. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 1515–1519
208. Adler, P., Wood, S. J., Lee, Y. C., Lee, R. T., Petri, W. A., and Schnaar, R. L. (1995) *J. Biol. Chem.* **270**, 5164–5171
209. Misevic, G. N., and Burger, M. M. (1990) *J. Biol. Chem.* **265**, 20577–20584
210. Spillmann, D., Hård, K., Thomas-Oates, J., Vliegenthart, J. F. G., Misevic, G., Burger, M. M., and Finne, J. (1993) *J. Biol. Chem.* **268**, 13378–13387
211. Spillmann, D., Thomas-Oates, J. E., van Kuik, J. A., Vliegenthart, J. F. G., Misevic, G., Burger, M. M., and Finne, J. (1995) *J. Biol. Chem.* **270**, 5089–5097
212. Rosati, F., and De Santis, R. (1980) *Nature (London)* **283**, 762–764
213. Seppo, A., Penttilä, L., Niemelä, R., Maaheimo, H., and Renkonen, O. (1995) *Biochemistry* **34**, 4655–4661
214. Litscher, E. S., Juntunen, K., Seppo, A., Penttilä, L., Niemelä, R., Renkonen, O., and Wassarman, P. M. (1995) *Biochemistry* **34**, 4662–4669

## References


215. Hitoshi, S., Kusunoki, S., Kanazawa, I., and Tsuji, S. (1995) *J. Biol. Chem.* **270**, 8844–8850
216. de Vries, T., Srnka, C. A., Palcic, M. M., Swiedler, S. J., van der Eijnden, D. H., and Macher, B. A. (1995) *J. Biol. Chem.* **270**, 8712–8722
217. Stroud, M. R., Levery, S. B., Mårtensson, S., Salyan, M. E. K., Clausen, H., and Hakomori, S. (1994) *Biochemistry* **33**, 10672–10680
218. Hull, S. R., Sugarman, E. D., Spielman, J., and Carraway, K. L. (1991) *J. Biol. Chem.* **266**, 13580 –13586
219. Ware, F. E., Vassilakos, A., Peterson, P. A., Jackson, M. R., Lehrman, M. A., and Williams, D. B. (1995) *J. Biol. Chem.* **270**, 4697–4704
220. Leppänen, A., Mehta, P., Ouyang, Y.-B., Ju, T., Helin, J., Moore, K. L., van Die, I., Canfield, W. M., McEver, R. P., and Cummings, R. D. (1999) *J. Biol. Chem.* **274**, 24838 – 24848
221. Chandrasekaran, E. V., Jain, R. K., Larsen, R. D., Wlasichuk, K., and Matta, K. L. (1995) *Biochemistry* **34**, 2925–2936
222. Kogan, T. P., Revelle, B. M., Tapp, S., Scott, D., and Beck, P. J. (1995) *J. Biol. Chem.* **270**, 14047–14055
223. Waddell, T. K., Fialkow, L., Chan, C. K., Kishimoto, T. K., and Downey, G. P. (1995) *J. Biol. Chem.* **270**, 15403–15411
224. Bertozzi, C. R., Fukuda, S., and Rosen, S. D. (1995) *Biochemistry* **34**, 14271–14277
225. Norgard-Sumnicht, K., and Varki, A. (1995) *J. Biol. Chem.* **270**, 12012–12024
226. Hemmerich, S., Leffler, H., and Rosen, S. D. (1995) *J. Biol. Chem.* **270**, 12035–12047
227. Yuen, C.-T., Bezouska, K., O'Brien, J., Stoll, M., Lemoine, R., Lubineau, A., Kiso, M., Hasegawa, A., Bockovich, N. J., Nicolaou, K. C., and Feiz, T. (1994) *J. Biol. Chem.* **269**, 1595–1598
228. Patel, T. P., Goelz, S. E., Lobb, R. R., and Parekh, R. B. (1994) *Biochemistry* **33**, 14815–14824
229. Graves, B. J., Crowther, R. L., Chandran, C., Rumberger, J. M., Li, S., Huang, K.-S., Presky, D. H., Familletti, P. C., Wolitzky, B. A., and Burns, D. K. (1994) *Nature (London)* **367**, 532–538
230. Lasky, L. A. (1995) *Ann. Rev. Biochem.* **64**, 113–139
231. Ginsburg, V. E. (1982) *Meth. Enzymol.* **83**, all
232. Churms, S. C. (1990) *CRC Handbook of Chromatography Carbohydrates*, Vol. 2, CRC Press, Boca Raton, Florida
233. Valent, B. S., Darvill, A. G., McNeil, M., Robertsen, B. K., and Albersheim, P. (1980) *Carbohydrate Res.* **79**, 165–192
234. Chaplin, M. F., and Kennedy, J. F., eds. (1994) *Carbohydrate Analysis A Practical Approach*, Oxford University, Oxford
235. Robyt, J. F., and Mukerjea, R. (1994) *Carbohydr. Res.* **251**, 187–202
236. Green, E. D., and Baenziger, J. U. (1989) *Trends Biochem. Sci.* **14**, 168–172
237. Carr, S. A., Huddleston, M. J., and Bean, M. F. (1993) *Protein Sci.* **2**, 183–196
237a. Fridriksson, E. K., Beavil, A., Holowka, D., Gould, H. J., Baird, B., and McLafferty, F. W. (2000) *Biochemistry* **39**, 3369–3376
237b. Burlingame, A. L., Carr, S. A., and Baldwin, M. A., eds. (2000) *Mass Spectrometry in Biology and Medicine*, Hamana Press, Totowa, New Jersey
238. Downs, F., and Pigman, W. (1969) *Biochemistry* **8**, 1760–1766
239. Narasimhan, S., Harpaz, N., Longmore, G., Carver, J. P., Grey, A. A., and Schacter, H. (1980) *J. Biol. Chem.* **255**, 4876–4884
240. Cohen, R. E., and Ballou, C. E. (1980) *Biochemistry* **19**, 4345–4358
241. Plummer, T. H., Jr., and Tarentino, A. L. (1981) *J. Biol. Chem.* **256**, 10243–10246
242. Risley, J. M., and Van Etten, R. L. (1985) *J. Biol. Chem.* **260**, 15488–15494
243. Takasaki, S., Mizuochi, T., and Kobata, A. (1982) *Meth. Enzymol.* **83**, 263–268
244. Parente, J. P., Wieruszeski, J.-M., Strecker, G., Montreuil, J., Fournet, B., van Halbeek, H., Dorland, L., and Vliegenthart, J. F. G. (1982) *J. Biol. Chem.* **257**, 13173–13176
245. Patel, T., Bruce, J., Merry, A., Bigge, C., Wormald, M., Jaques, A., and Parekh, R. (1993) *Biochemistry* **32**, 679–693
246. Sweeley, C. C., and Nunez, H. A. (1985) *Ann. Rev. Biochem.* **54**, 765–801
247. Dua, V. K., Roa, B. N. N., Wu, S.-S., Dube, V. E., and Bush, C. A. (1986) *J. Biol. Chem.* **261**, 1599–1608
248. Tsai, P.-K., Dell, A., and Ballou, C. E. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 4119–4123
249. Lindberg, B., Lönngren, L., and Svensson, S. (1975) *Adv. Carbohydr. Chem. Biochem.* **31**, 185–240
250. Guthrie, R. D., and McCarthy, J. F. (1967) *Adv. Carbohydr. Chem. Biochem.* **22**, 11–23
251. Edge, C., Parekh, R., Rademacher, T., Wormald, M., and Dwek, R. (1992) *Nature (London)* **385**, 693–694
252. Varki, A. (1991) *FASEB J.* **5**, 226–235
253. Rolf, D., and Gray, G. R. (1982) *J. Am. Chem. Soc.* **104**, 3539–3541
254. Mukerjea, R., Kim, D., and Robyt, J. F. (1996) *Carbohydrate Res.* **292**, 11–20
255. Bobbitt, J. M. (1956) *Adv. Carbohydr. Chem. Biochem.* **11**, 1–41
256. Rich, A., Nordheim, A., and Wang, A. H.-J. (1984) *Ann. Rev. Biochem.* **53**, 791–846
257. Homans, S. W., Dwek, R. A., and Rademacher, T. W. (1987) *Biochemistry* **26**, 6571–6578
258. Homans, S. W., Pastore, A., Dwek, R. A., and Rademacher, T. W. (1987) *Biochemistry* **26**, 6649 –6655
259. Cumming, D. A., Shah, R. N., Krepinsky, J. J., Grey, A. A., and Carver, J. P. (1987) *Biochemistry* **26**, 6655–6663
260. Reuben, J. (1985) *J. Am. Chem. Soc.* **107**, 1747–1755
261. Cumming, D. A., Dime, D. S., Grey, A. A., Krepinsky, J. J., and Carver, J. P. (1986) *J. Biol. Chem.* **261**, 3208–3213
262. Vuister, G. W., de Waard, P., Boelens, R., Vliegenthart, J. F. G., and Kaptein, R. (1989) *J. Am. Chem. Soc.* **111**, 772–774
263. Gronenborn, A. M., and Clore, G. M. (1989) *Biochemistry* **28**, 5978–5984
264. Abeygunawardana, C., and Bush, C. A. (1991) *Biochemistry* **30**, 8568–8577
265. Rivière, M., and Puzo, G. (1992) *Biochemistry* **31**, 3575–3580
266. van Duynhoven, J. P. M., Goudriaan, J., Hilbers, C. W., and Wijmenga, S. S. (1992) *J. Am. Chem. Soc.* **114**, 10055–10056
267. Wu, J., Bondo, P. B., Vuorinen, T., and Serianni, A. S. (1992) *J. Am. Chem. Soc.* **114**, 3499–3505
268. Homans, S. W. (1990) *Biochemistry* **29**, 9110–9118
269. Bose, B., Zhao, S., Stenutz, R., Cloran, F., Bondo, P. B., Bondo, G., Hertz, B., Carmichael, I., and Serianni, A. S. (1998) *J. Am. Chem. Soc.* **120**, 11158–11173
270. Flugge, L. A., Blank, J. T., and Petillo, P. A. (1999) *J. Am. Chem. Soc.* **121**, 7228 – 7238

## Study Questions

1. A nonreducing disaccharide gives an octamethyl derivative with dimethyl sulfate and alkali. On acid hydrolysis, this derivative yields 1 mol of 2,3,4,6-tetramethyl-D-glucose and 1 mol of 2,3,4,6-tetramethyl-D-galactose. The disaccharide is hydrolyzed rapidly by either maltase or lactase (a β-galactosidase).

   Give an adequately descriptive name of the disaccharide, and draw its Haworth projection formula.

2. An aldopentose (A) of the D-configuration on oxidation with concentrated nitric acid gives a 2,3,4-trihydroxypentanedioic acid (a trihydroxyglutaric acid) (B) which is optically inactive. (A) on addition of HCN, hydrolysis, lactonization, and reduction gives two stereoisomeric aldohexoses (C) and (D). (D) on oxidation affords a 2,3,4,5-tetrahydroxy-hexanedioic acid (a saccharic acid) (E) which is optically inactive. Give structures of compounds (A)-(E).

3. What products are formed when periodic acid reacts with sorbitol?
4. A 10.0 g sample of glycogen gave 6.0 millimol of 2,3-di-*O*-methylglucose on methylation and acid hydrolysis.

   a. What percent of the glucose residues in glycogen have chains substituted at the α–1→6 position?

   b. What is the average number of glucose residues per chain?

   c. How many millimols of 2,3,6-tri-*O*-methylglucose were formed?

   d. If the molecular weight of the polysaccharide is $2 \times 10^6$, how many glucose residues does it contain?

   e. How many nonreducing ends are there per molecule or equivalently how many chains are there per molecule?

5. D-Mannitol is a symmetric molecule, yet it is optically active. Explain.

6. When D-glucose is treated with acidic methanol, the first products which can be isolated are mainly methyl furanosides, but after extensive reaction the furanosides disappear and methyl glucopyranosides accumulate. Why?

7. Write the structural formulas for (a) 1,6 anhydro β-D-glucopyranose; (b) 1,6 anhydro β-D-altrose. Compound (b) is many times more stable than compound (a). Explain this on stereochemical grounds.

8. The disaccharide *nigerose* is α-D-Glu*p*-(1→3)-D-glu. Write out its structure. How would you prove this structure using methylation, periodate oxidation, and other methods.

9. Inositol is 1,2,3,4,5,6-hexahydroxycyclohexane. Draw configurational formulas for all possible stereoisomers and indicate which would be expected to be optically active.

10. Why do you suppose that the major form of D-fructose in solution is the pyranose form but D-fructose in sucrose is in the furanose form?

11. D-Xylose is an easily prepared sugar, potentially available in enormous quantity. What is a common source? How can it be obtained from this source?

12. How can xylitol be obtained from xylose? Discuss the stereochemical properties of xylitol.

13. The enzyme xylose isomerase is important industrially. Why?

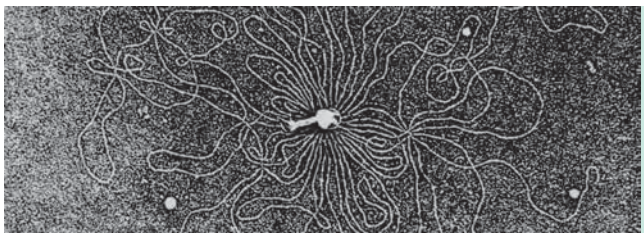
14. Glucose reacts non-enzymatically with amino acids and proteins, including hemoglobin, egg-white proteins and serum albumin. For example, if glucose is not removed prior to drying, dried egg whites slowly turn brown and develop off-flavors and odors. What do you propose as the most likely first step in the non-enzymatic glucose-protein chemical reaction? How can the first product transform spontaneously into a ketose derivative?

15. What characteristics would you expect in a binding site for a sugar ring in an enzyme, lectin, or other proteins?

16. Using structural formulas, descirbe the two major types of linkage of carbohydrate chains or clusters to proteins to form glycoproteins or proteoglycans.

17. What products would you expect from cellulose as a result of methylation analysis? Periodate oxidation? The Smith degradation? The action of an alpha amylase?

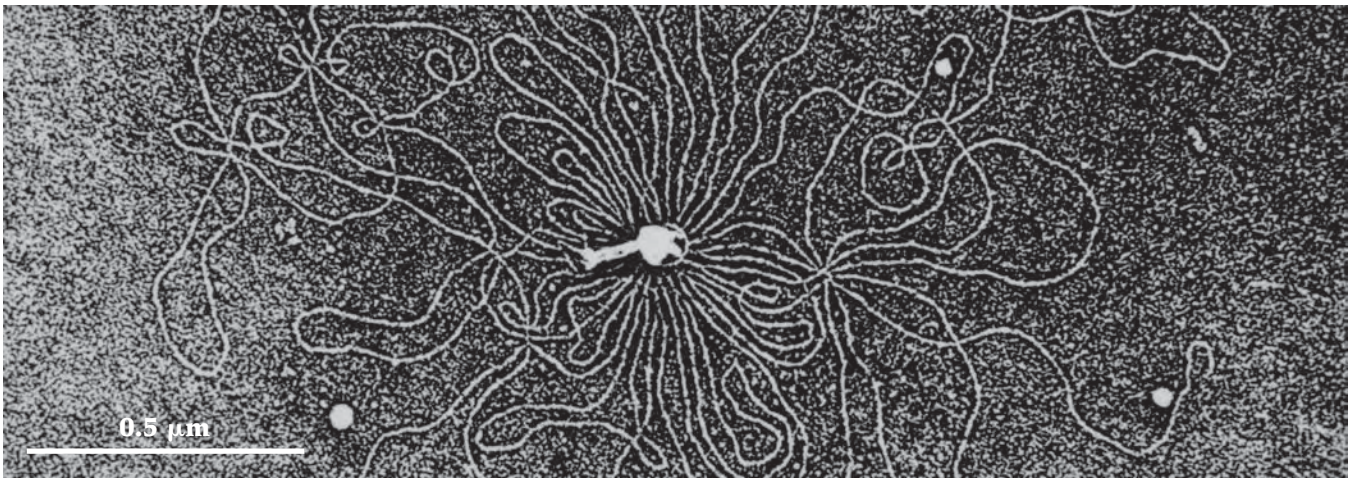
18. When glycoproteins are treated with alkaline brohydride, amino acid analysis often indicates a decrease in the amount of serine and a corresponding increase in the amount of alanine, or a decrease in threonine with the appearance of α-aminobutyric acid. Explain.



DNA spreading from the broken head of a bacteriophage T2 phage. This classic electron micrograph, published by A. K. Kleinschmidt and coworkers in 1962 (*Biochem. Biophys. Acta.* **61**, 857–864, 1962) was prepared by spreading the phage particles suspended in a protein–salt solution as a mixed monolayer on a water-air interface. The resultant osmotic shock burst the head and confined the DNA as a single thread near the phage ghost. After transfer to a suitable carbon surface, removal of water, and shadowing with platinum, the micrograph was obtained. Courtesy of Albrecht K. Kleinschmidt

# Contents

# 5 The Nucleic Acids




The phosphorus- and nitrogen-containing materials that came to be known as nucleic acids were first isolated from cells around 1870 by Friedrich Miescher but were long regarded as something of a curiosity.[1] Nevertheless, the structures of the monomer units, the **nucleotides**, were established by 1909 and the correct **polynucleotide** structure of the chains of DNA and RNA was proposed by Levene and Tipson in 1935.[2,3]

The nucleotides are made up of three parts:

1. One of the **pyrimidine** or **purine "bases"** uracil, cytosine, adenine, or guanine (Fig. 5-1). All four of these bases are present in RNA, while DNA contains thymine instead of uracil. Atoms in the bases are numbered 1–6 or 1–9.
2. A **sugar**, either D-ribose or D-2-deoxyribose. Carbon atoms in sugars are numbered 1′–5.′
3. **Phosphoric acid**

Although the biological synthesis is indirect, we can imagine that nucleotides are formed from these parts by elimination of two molecules of water as indicated in Eq. 5-1. In nucleic acids the nucleotides are combined through phosphodiester linkages between the 5'-hydroxyl of the sugar in one nucleotide and the 3'-hydroxyl of another. Again, we can imagine that these linkages were formed by the elimination of water (Eq. 5-2). The structures of a pair of short polynucleotide strands in DNA are shown in Fig. 5-2. That of a segment of double-helical DNA is shown in Fig. 5-3 and that of a transfer RNA in Figs. 5-30 and 5-31.

Despite the fact that Levene had deduced the correct structure for polynucleotides, he was thrown off the trail of a deeper understanding by the roughly equal amounts of the four bases found in either DNA or RNA. He assumed that nucleic acids must be regular repeating polymers for which there was no obvious biological function. It was not until 1944 that there

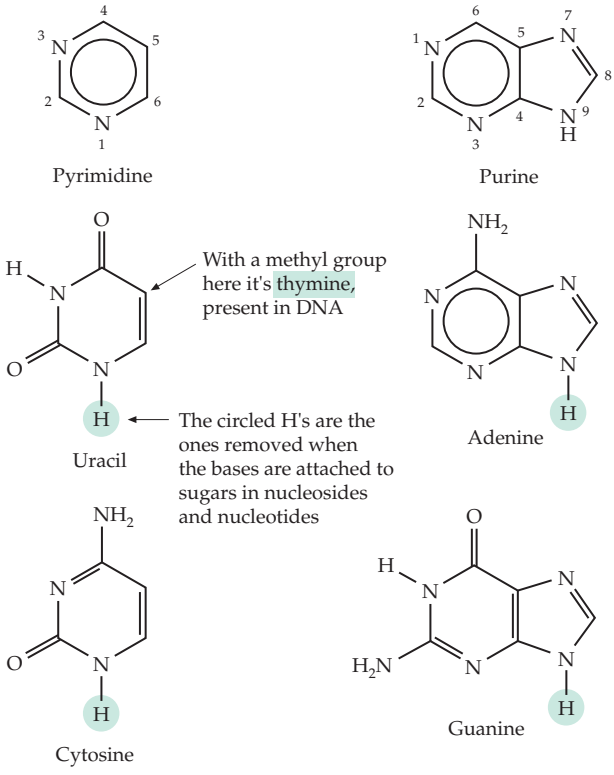
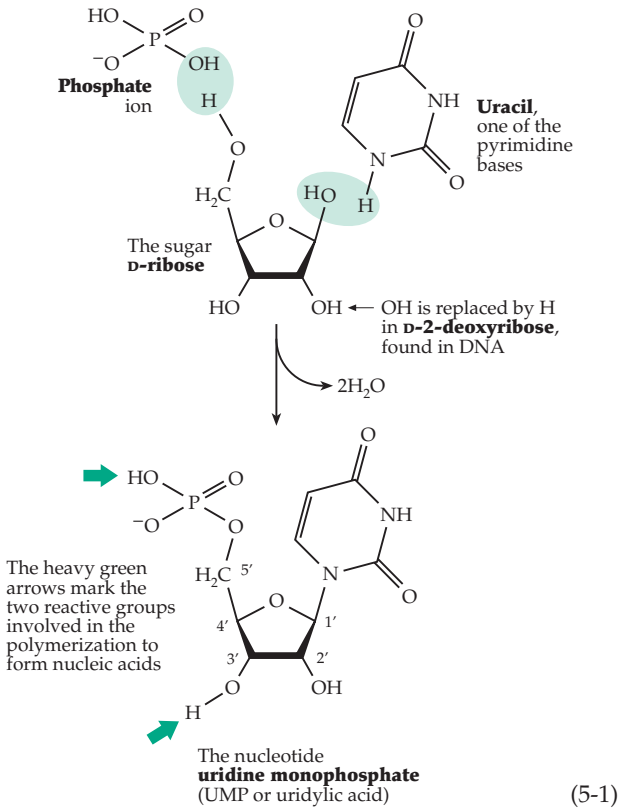
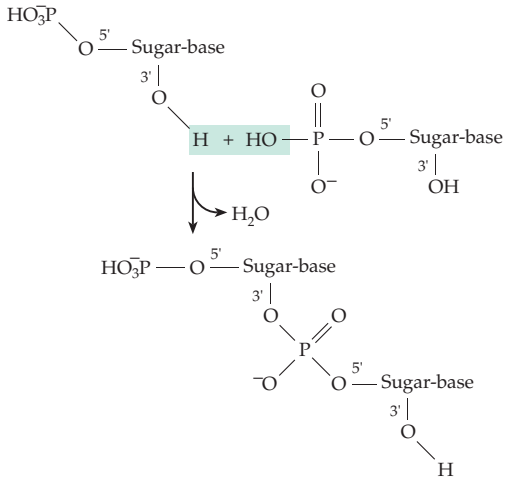



**Figure 5-1** Structures of the major pyrimidine and purine bases of DNA and RNA.




(5-1)




(5-2)

was concrete evidence that DNA carried genetic information (see Chapter 26). However, it was James Watson and Francis Crick's recognition of the double-helical structure of DNA[4-9] in 1953 and the mechanism of replication that this structure implied that captured the imagination of biologists and chemists alike and paved the way for the present-day explosion of knowledge of DNA, RNA, and of the encoded proteins.

Watson and Crick proposed that DNA is a double helix of two antiparallel polynucleotide chains (Figs. 5-2 and 5-3). The structure was deduced from model building together with knowledge of the X-ray diffraction data of Maurice F. Wilkins and Rosalind Franklin[9a] on artificially formed DNA fibers. An additional key piece of information was the discovery by Erwin Chargaff that *in all double-stranded DNA the content of adenine equals that of thymine and the content of guanine equals that of cytosine.*

The most significant feature of the proposed structure was the pairing of bases in the center through hydrogen bonding. The pairs and triplets of hydrogen bonds (Fig. 5-2) could form in the manner shown only if adenine (A) was paired with thymine (T) and cytosine (C) with guanine (G) at every point in the entire DNA structure. Thus, *the nucleotide sequence in one chain is complementary to but not identical to that in the other chain.* It was apparent almost immediately that the sequence of bases in a DNA chain must convey the encoded genetic information. The complementarity of the two strands suggested a simple mechanism for replication of genes during cell divisions. The two strands could separate and a complementary strand could be synthesized along each strand to give two molecules of the DNA, one for each of the two cells.

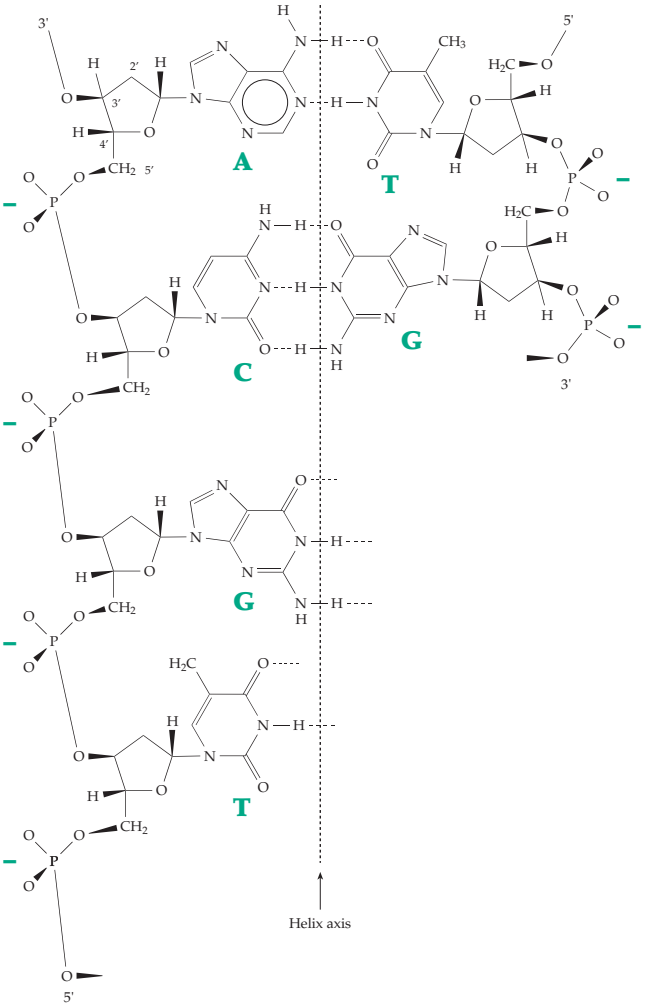



**Figure 5-2** A distorted (flattened) view of the Watson–Crick structure of DNA showing the hydrogen-bonded base pairs.

The essential correctness of the concept has been proved.

Two extremely important developments came in the 1970s: (1) Methods were found for cutting and rejoining DNA fragments and for **cloning** them in bacteria and (2) ways were devised for rapid determination of **nucleotide sequences**. The application of these techniques is now providing startling advances in biology and medicine. In 1970 we knew virtually nothing about the sequences of nucleotides in genes but today, we know the sequences for many thousands of genes of all types. By the 1980s sequences had been established for hemoglobin,[11] γ-globulins,[12] collagen,[13] and for many enzymes. An example is the gene sequence for mitochondrial aspartate aminotransferase (Fig. 5-4).[14] Its coding regions consist of 1299 pairs of mononucleotides, the bases being paired as in Fig. 5-2 in double helical form. Complete DNA sequences are known for numerous viruses including the 9740-base pair (bp) DNA provirus form of the RNA virus that causes AIDS[15] and for bacterial viruses such as T7 bacteriophage (39,936 bp).[16] Also determined in the 1980s were sequences of human mitochondrial DNA (16,598 bp),[17] and of chloroplast DNA from the tobacco plant (155,844 bp).[18]

In 1995 the first complete sequences of bacterial genomes were obtained (Table 1-3).[19,20] These were followed by sequences of many other bacterial genomes,[21] including the 4.2 x $10^6$ bp *E. coli* genome (Table 1-3). Sequencing of the 16 chromosomes of the 12.07 Mbp genome of yeast[22,23] containing ~6300 genes was completed in 1996[24] and that of the 97 Mbp genome of the nematode *Caenorhabditis elegans* in 1998. By 2000 the sequence of the 180 Mbp genome of the fruit fly *Drosophila melanogaster* was largely completed, and most sequences of the 3 x $10^9$ bp human genome were known.[24a]

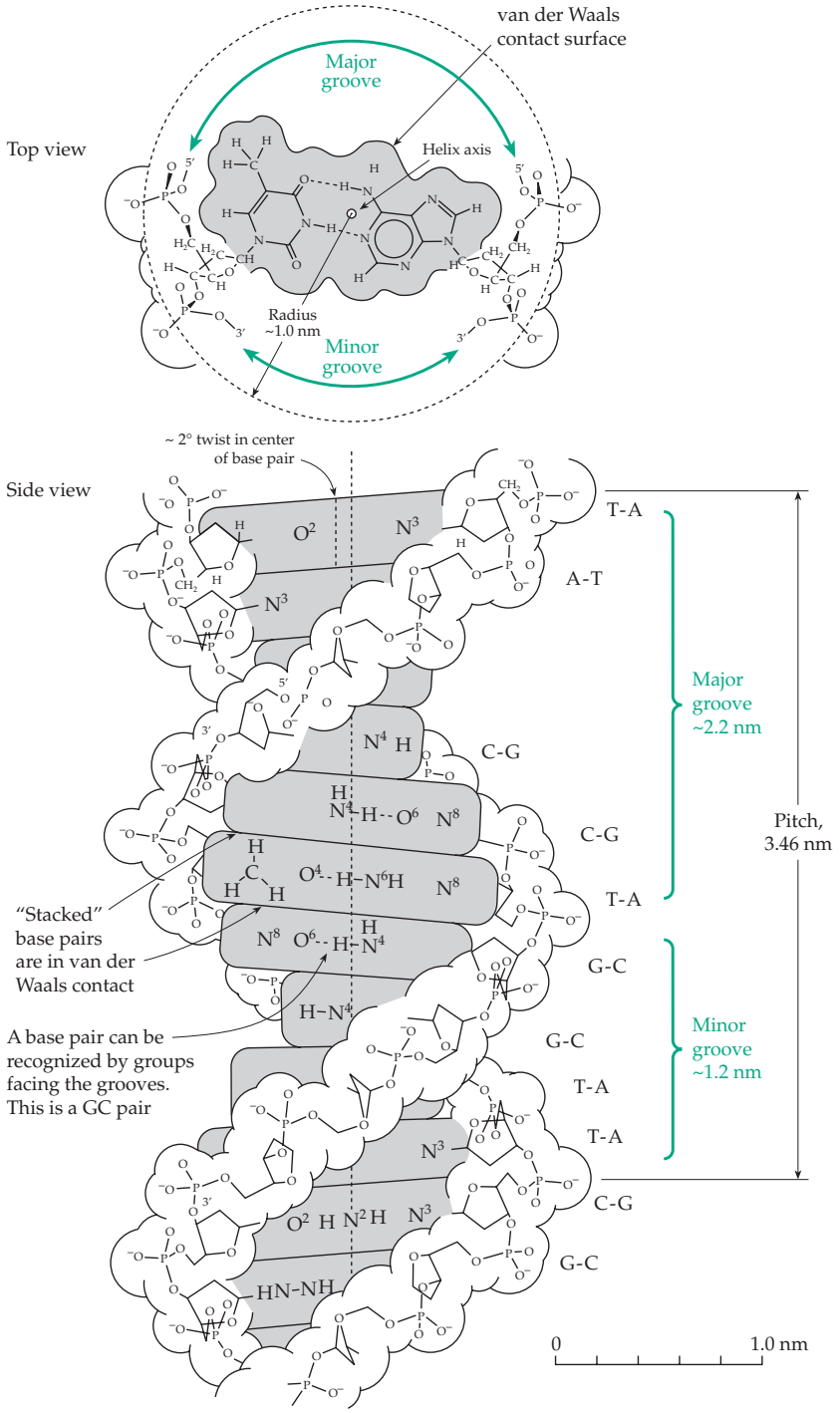



**Figure 5-3** The double-helical structure of DNA. The structure shown is that of the B form and is based on coordinates of Arnott and Hukins.[10] The major and minor grooves, discussed on p. 213, are marked.

Cloning of DNA has not only provided an essential step in sequence determination but also has given birth to a new industry devoted to producing proteins from genes cloned in bacteria, yeast, or other cells. Human insulin and the antiviral protein interferon were two of the first proteins produced in this way. Methods now in use permit us to introduce at will alterations at any point in a DNA sequence. We are able to locate and study the genes responsible for many genetic defects.

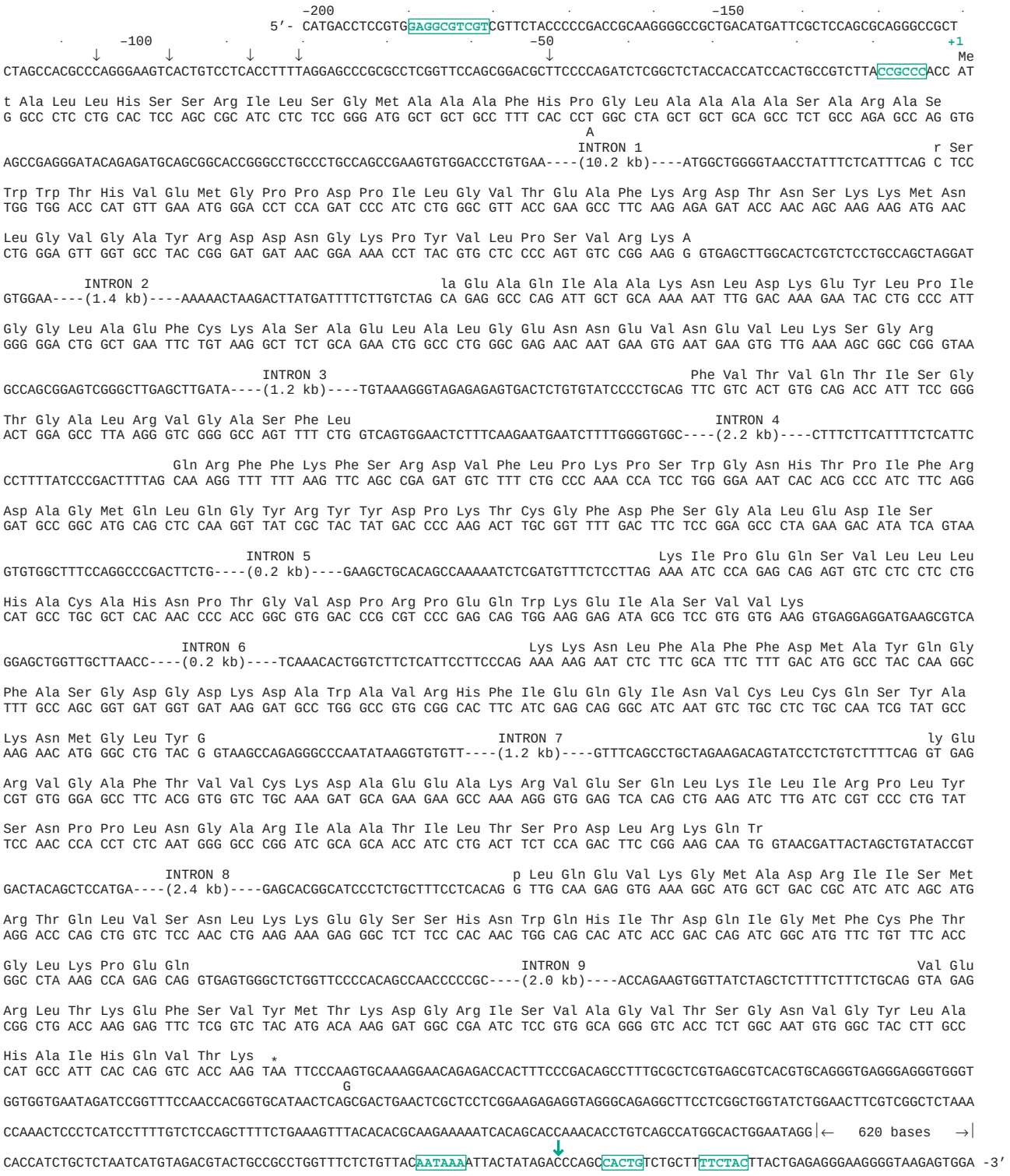


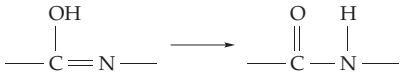

**Figure 5-4** The nucleotide sequence of the gene for mitochondrial aspartate aminotransferase from the mouse. From Tsuzuki *et al.*[14] The gene encodes a 433-residue protein requiring 1299 nucleotide pairs (1.3 kb). However, 29 residues are cut off from the N terminus to form the mature mitochondrial protein. In addition, the gene is split by nine introns which vary in length from 0.2–10.2 kb. The sequences at the ends of the introns are shown. There is also an "upstream" region (of which 200 nucleotides are shown) at the 5'-end of the gene. It contains two binding sites for **transcription factor Sp1** (boxed). At the 3' end the 993 additional nucleotides contain signals (boxed) surrounding the **polyadenylation site** (green arrow) for 3′ processing and termination of transcription. The mature messenger RNA is about 2400 nucleotides (2.4 kb) in length but the gene, with introns, occupies about 25 kb. The +1 marks the position of the first nucleotide of the initiation codon ATG (encoding methionine) and the asterisk (*) the termination codon TAA. These are AUG and UAA in the mRNA. Other codons are indicated by amino acid abbreviations.

The first successes in using virus-like particles to carry new pieces of DNA into the cells to help correct these defects have been reported. Our ability to breed new varieties of plants and microorganisms has been enormously enhanced. We can foresee the production of artificially designed enzymes to conduct many industrial chemical processes. These are among the many reasons for the excitement today in the fields of nucleic acid chemistry and molecular genetics.

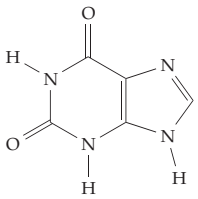
## A. Structure and Chemistry of Nucleotides

### 1. Names and Abbreviations

The purine and pyrimidine ring compounds found in nucleic acids are known as "bases," even though some of them have almost no basic character. **Nucleosides** are the *N*-glycosyl derivatives of the bases with ribose or 2-deoxyribose. The **nucleotides** are phosphate esters of nucleosides. Similar names are applied to related compounds such as adenosine triphosphate (ATP) that are not present in DNA or RNA. The names of the principal nucleotides from which the nucleic acids are formed are given in Table 5-1. The novice may find these confusing! Even worse than the names in the table is **hypoxanthine** (Hyp), which is derived from adenine by replacement of its $-NH_2$ group with –OH and tautomerization:



The nucleoside formed from hypoxanthine and ribose is known as **inosine** (Ino or I) and the corresponding nucleotide as **inosinic acid**. Further substitution at C-2 of –H by –OH and tautomerization yields **xanthine** (Xan). Its nucleoside is xanthosine (Xao, X). A similar hydroxylation at C-7 converts xanthine to **uric acid**, an important human urinary excretion product derived from nucleic acid bases.


Xanthine

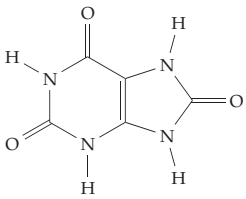

Uric acid

**TABLE 5-1**
**Names of Pyrimidine and Purine Bases, Nucleosides, and 5'-Nucleotides[a]**

**A Nucleotide units of RNA (abbreviations in parentheses)[b]**

| | | | | |
|---|---|---|---|---|
| Base: | Uracil (Ura) | Cytosine (Cyt) | Adenine (Ade) | Guanine (Gua) |
| Nucleoside: | Uridine (Urd or U) | Cytidine (Cyd or C) | Adenosine (Ado or A) | Guanosine (Guo or G) |
| 5'-Nucleotide: | Uridine 5'-phosphate or 5'-uridylic acid (Urd-5'-*P* or UMP) | Cytidine 5'-phosphate or 5'-cytidylic acid (Cyd-5'-*P* or CMP) | Adenosine 5'-phosphate or 5'-adenylic acid (Ado-5'-*P* or AMP) | Guanosine 5'-phosphate or 5'-guanylic acid (Guo-5'-*P* or GMP) |

**B Nucleotide units of DNA**

These contain 2-deoxyribose and the nucleosides and nucleotides are called deoxyadenosine (dAdo or dA), deoxyadenosine 5'-phosphate (dAMP), etc.

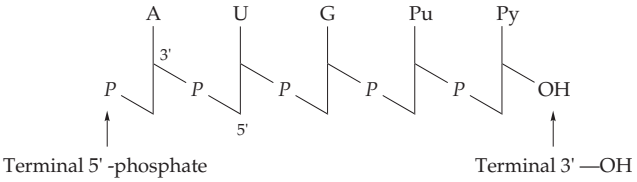
DNA contains thymine (Thy) rather than uracil. The deoxyribose derivatives are thymidine (dThd or dT) and thymidine 5'-phosphate. The ribose derivatives of thymine are the nucleoside ribosylthymidine (Thd) and ribosylthymidine 5'-phosphate (Thd-5'-*P*).

**C Abbreviations used in describing polynucleotide sequences**

| | |
|---|---|
| U,T,C,A,G | Uracil, thymine, cytosine, adenine, guanine |
| Y or Pyr | Pyrimidine (T or C) |
| R or Pur | Purine (A or G) |
| M | Amino base (A or C) |
| K | Keto base (G or T) |
| S | Strongly pairing (G or C) |
| W | Weakly pairing (A or T) |
| H | Not G (any other base) |
| B | Not A |
| V | Not T or U |
| D | Not C |
| N | Any base |

[a] From "Biochemical Nomenclature"Liébecq, C., ed.[25]
[b] Isomers of the 5'-nucleotides, in which the phosphate is attached to the oxygen on C-3', are the 3'-nucleotides. Care must be taken to avoid ambiguity. The simple abbreviations UMP, CMP, AMP, and GMP always refer to the 5'-nucleotides.

Nucleic acid structures are abbreviated in several ways. For example, the sugar rings may be portrayed by vertical lines. The abbreviations A, C, U, T, and G for the individual bases or Pu (purine) and Py (pyrimidine) are placed at the upper ends of the lines, and slanted lines with *P* in the centers represent the 3'–5' phosphodiester linkages in a polynucleotide.

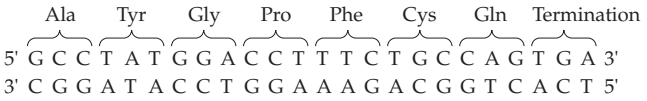



The same structure can be further abbreviated

| | | | |
|---|---|---|---|
| | 5' end | pApUpGpRpY | 3' end |
| or | 5' end | A U G R Y | 3' end |

Here purine is abbreviated R and pyrimidine Y. By convention the 5' end of a polynucleotide is ordinarily placed to the left in these formulas. Lengths of nucleic acid chains are usually given as a number of bases or kilobases (kb). For double-stranded DNA (dsDNA) the length is given as base pairs (bp), kilobase pairs (kbp), or megabase pairs (Mbp). However, in most places, including this book, the abbreviations kb and Mb will be used for a length of DNA whether single or double stranded.

For double-stranded DNA, one strand, usually the **coding strand,** from which the amino acid sequence can be read using the code in Tables 5-4 or 5-5 (Section C), has the 5' end at the left while the complementary strand has the 3' end at the left, e.g.,

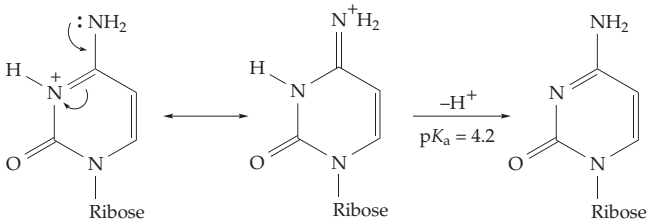



## 2. Acid–Base Chemistry and Tautomerism

The ionized phosphate groups of the polymer "backbone" give nucleic acid molecules a high negative charge. For this reason DNA in cells is usually associated with basic proteins such as the **histones** or **protamines** (in spermatozoa), with polycations of amines such as **spermidine** ($H_3{}^+NCH_2CH_2CH_2CH_2NH_2{}^+CH_2CH_2CH_2NH_3{}^+$), or with alkaline earth cations such as $Mg^{2+}$. If the pH of a solution containing double-helical DNA is either lowered to ~3 or raised to ~12, the two strands unravel and can be separated.

Over the entire range of pH the alternating sugar-phosphate backbone of the polymeric chains remains negatively charged. However, depending on the pH, the bases can be protonated or deprotonated with a resultant breaking of the hydrogen bonds that hold the pairs of bases together.[26,27]

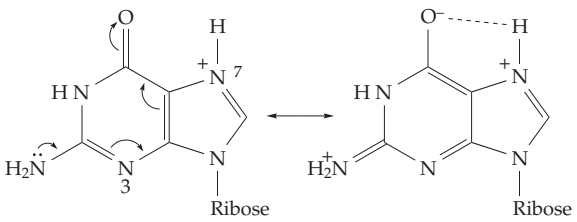
Pyrimidines and purines, which contain the $-NH_2$ group, are weakly basic. The cationic protonated conjugate acid forms of cytidine, adenosine, and guanosine have p$K_a$ values of 4.2, 3.5, and 2.7, respectively. Similar values are observed for the 5'-nucleotides. In these compounds it is not the $-NH_2$ group that binds the proton but an adjacent nitrogen atom in the ring (Eq. 5-3).



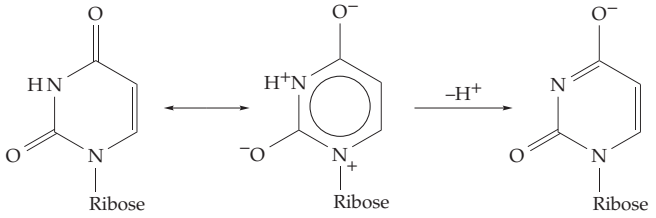

(5-3)

We can understand this if we recognize that the bases have substantial aromatic character.[26,28] In aniline (aminobenzene) electrons are withdrawn from the amino group into the aromatic ring with a strong decrease in basicity of this $-NH_2$ group (the p$K_a$ is 4.6). Similarly, electrons are withdrawn from the $NH_2$ groups of cytosine, adenosine, and guanine into the pyrimidine and purine rings as is indicated by the small curved arrows on the left-hand structure of Eq. 5-3. The effect is even stronger than in aniline, largely because of the presence of the nitrogen atoms in the rings. In cytosine it is primarily N-3 that serves as the electron acceptor. As a consequence this nitrogen becomes more basic than the $-NH_2$ group and is the major site of protonation. However, as is indicated in Eq. 5-3, the positive charge on the cation is shared by resonance with the exocyclic amino group.

Adenosine is similar to cytosine in its acid–base chemistry; N-1, adjacent to the $-NH_2$ group, is the principal site of protonation. A tautomer of the cation protonated at N-3 is formed in smaller amounts. Guanosine is electronically more complex, being protonated mainly at N-7 and to a lesser extent at N-3[29]. This can be understood in terms of electronic interaction with the adjacent oxygen as indicated in the resonance structure to the right in the following diagram:



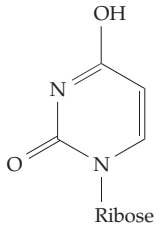
Under basic conditions the proton on N-3 of uridine or thymine or on N-1 of guanosine can dissociate with a p$K_a$ of ~9.2. These "bases" are actually weak acids!




(5-4)

Look at the "Kekule" resonance structure shown in the center of Eq. 5-4. The two negative charges flanking the dissociable proton are sufficient to hold it firmly to the ring at low pH. The proton is half dissociated only when the pH is raised to 9.2. At still higher values of pH (p$K_a$ ~12.4) a proton dissociates from a ribose hydroxyl group.

The tautomerism of pyrimidines was discussed in Chapter 2, Section A,6. The tautomeric forms shown in Fig. 5-1 predominate. However, it is possible that minor tautomers such as the following are sometimes preferentially bound into active sites of enzymes where the dielectric




A minor tautomer of uridine

constant may be low and where the geometrical arrangement of amino acid functional groups may favor protonation on oxygen rather than nitrogen.[30,31]

## 3. Absorption of Ultraviolet Light

Nucleic acids strongly absorb ultraviolet light of wavelengths below about 300 nm, with an absorption maximum at ~260 nm and a stronger one below 200 nm. This property is biologically important because of the resultant induction of mutations, a natural result of exposure to sunlight. In the laboratory the same property is useful in identification and quantitative analysis of nucleotides. The ultraviolet spectra of four nucleosides are shown in Fig. 5-5. Notice the changes that accompany protonation or deprotonation of the ring. The absorption bands, which are related to those of benzene[26,32] (see Chapter 23), provide another indication of the partial aromatic character of the bases. Chapter 23 also provides information on photochemical reactions of the pyrimidines.

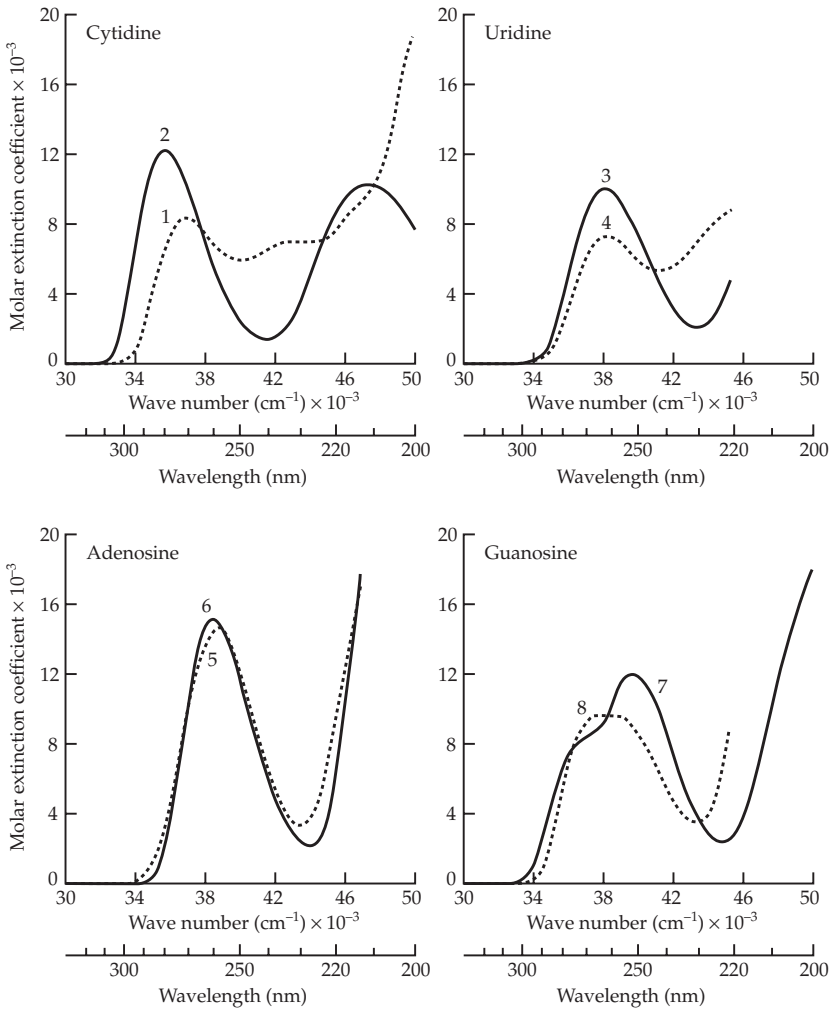



**Figure 5-5** Near ultraviolet absorption spectra of cytidine, uridine, adenosine, and guanosine. 1. Monoprotonated form of cytidine (for which p$K_a$ = 4.2). 2. Neutral form (pH ~ 7) of cytidine. 3. Neutral form of uridine (for which p$K_a$ = 9.2). 4. Monoanionic form of uridine. 5. Monoprotonated form of adenosine (p$K_a$ = 3.5). 6. Neutral form of adenosine. 7. Neutral form of guanosine (p$K_a$ = 9.2). 8. Monoanion of guanosine.

## BOX 5-A THE ALKALI METAL IONS

The many negative charges along a nucleic acid backbone interact with all of the cations in a cell. This box discusses some of these ions with emphasis on the group IA metal ions. Although sodium and potassium occur in similar amounts in the crust of the earth, living cells all accumulate potassium ions almost to the exclusion of sodium.[a-c] Sodium ions may be required only by certain marine organisms and by multicellular animals that regulate their internal body fluids. Most nonmarine plants have no demonstrable need for sodium.

The tendency to accumulate $K^+$ is even more remarkable since seawater is ~0.46 M in $Na^+$ and only 0.01 M in $K^+$. Other alkali metals occur in even smaller amounts, e.g., 0.026 mM $Li^+$, 0.001 mM $Rb^+$, and a trace of $Cs^+$. Soil water is ~0.1 mM in $K^+$ and 0.65 mM in $Na^+$. Again, strong discrimination in favor of potassium is observed in uptake by plants.

Intracellular concentrations of $K^+$ range from 200 mM in *E. coli* and 150 mM in mammalian muscle to ~30 mM in freshwater invertebrates such as clams, hydra, and some protozoa. While $K^+$ cannot be replaced by $Na^+$, a partial replacement by $Rb^+$ and to a lesser extent by $Cs^+$ is usually possible. In many microorganisms rubidium can almost completely replace potassium, and even a rat can survive for a *short* time with almost complete substitution of $K^+$ by $Rb^+$. Protons replace most $K^+$ in brown algae.[d] The human nutritional requirement for potassium is high, amounting to ~2 g/day. Present populations may suffer a chronic deficiency of potassium as a result of food processing and boiling of vegetables.[e]

Sodium is also essential to higher animals, and rats die on a sodium-free diet. The sodium content of cells varies among species, but it is usually no more than 0.1–0.2 times that of $K^+$. A measurement of the [$Na^+$] within heart cells gave a concentration of ~9 mM, which was increased by a factor of ~2.5 in a low $Ca^{2+}$ insulin-containing medium.[f] In this measurement the NMR resonance of the abundant external $Na^+$ was shifted by use of a paramagnetic reagent (e.g., a dysprosium (III) complex), that remained outside the cell. The signal from the internal $Na^+$ was then seen clearly. In blood, the relationship between $Na^+$ and $K^+$ concentrations is reversed from that within cells. Human plasma is 0.15 M in $Na^+$ and 0.005 M in $K^+$. Curiously, the taste for salt in the diet appears to be largely an acquired one.[g]

It is not immediately obvious why $K^+$ is the preferred counterion within tissues, but a fundamental reason may lie in the differences in hydration between $Na^+$ and $K^+$ (Chapter 6). On the other hand, the relationship of these ions to the excitability of membranes (Chapter 30) may be of paramount importance, even in bacteria. The concentration differences in the two ions across membranes represent a readily available source of Gibbs energy for a variety of membrane-associated activities. Cells actively pump $Na^+$ out and $K^+$ into cells (Chapter 8).

Many intracellular enzymes require $K^+$ for activity.[b,c] These include those promoting phosphorylation of carboxyl groups or enolate anions and elimination reactions yielding enols as well as some enzymes dependent upon the coenzyme pyridoxal phosphate.[h–j] In all of these enzymes $NH_4^+$, $Rb^+$, or $Tl^+$ can usually replace $K^+$. This permits study of the binding site for $Tl^+$ by the very sensitive $^{205}Tl$ NMR spectroscopy.[k] The discovery that $K^+$ is preferentially bound in some tetraplex DNA structures (see Fig. 5-8) further emphasizes the significant difference in biological properties of the alkali metal ions. Various synthetic macrocyclic compounds are also able to selectively bind specific alkali metal ions.[l]

The following tabulation gives concentrations not only of $K^+$ and $Na^+$ but also of the other principal ionic constituents in human blood plasma and within cells of skeletal muscle.[m] Units are mmol/kg $H_2O$.

| Ion | Blood plasma | Skeletal muscle (intracellular) |
|---|---|---|
| $Na^+$ | 150 | 14 |
| $K^+$ | 5 | 150 |
| $Mg^{2+}$ | 0.9 | 8 |
| $Ca^{2+}$ | 2.5 | 1 |
| $Cl^-$ | 105 | 16 |
| $HCO_3^-$ | 27 | 10 |
| Proteins$^-$ | 17* | 50* |
| Other anions† | 6 | 146 |

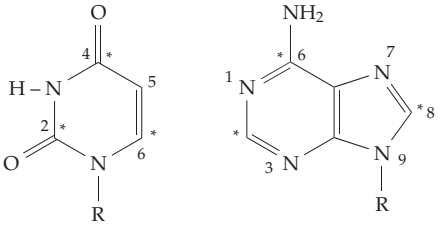
* Milliequivalents/kg $H_2O$
† Phosphates and other nonprotein anions.


[a] Kernan, R. P. (1965) *Cell K*, Butterworth, London
[b] Suelter, C. H. (1974) in *Metal Ions in Biological Systems*, Vol. 3 (Sigel, H., ed), pp. 201–251, Dekker, New York
[c] Suelter, C. H. (1970) *Science* **168**, 789–795
[d] Steinbach, H. B. (1962) *Comp. Biochem. Physiol.* **4**, 677–720
[e] Weber, C. E. (1970) *J. Theor. Biol.* **29**, 327–328
[f] Wittenberg, B. A., and Gupta, R. K. (1985) *J. Biol. Chem.* **260**, 2031–2034
[g] Kaunitz, H. (1956) *Nature (London)* **178**, 1141–1144
[h] Toney, M. D., Hohenester, E., Cowan, S. W., and Jansonius, J. N. (1993) *Science* **261**, 756–759
[i] Antson, A. A., Demidkina, T. V., Gollnick, P., Dauter, Z., Von Tersch, R. L., Long, J., Berezhnoy, S. N., Phillips, R. S., Harutyunyan, E. H., and Wilson, K. S. (1993) *Biochemistry* **32**, 4195–4206
[j] Metzler, C. M., Viswanath, R., and Metzler, D. E. (1991) *J. Biol. Chem.* **266**, 9374–9381
[k] Markham, G. D. (1986) *J. Biol. Chem.* **261**, 1507–1509
[l] Christensen, J. J., Hill, J. O., and Izatt, R. M. (1971) *Science* **174**, 459–467
[m] Composite data from Muntwyler, E. (1968) *Water and Electrolyte Metabolism and Acid–Base Balance*, p. 14. Mosby, St. Louis, Missouri; White, A., Handler, P., and Smith, E. L. (1973) *Principles of Biochemistry*, 5th ed., p. 802. McGraw-Hill, New York; Long, C. (1961) *Biochemist's Handbook*, p. 670. Van Nostrand, Princeton, New Jersey. Reported ranges for some constituents are very wide.

## 4. Chemical Reactions of the Bases

The purines and pyrimidines are relatively stable compounds with considerable aromatic character. Nevertheless, they react with many different reagents and, under some relatively mild conditions, can be completely degraded to smaller molecules. The chemistry of these reactions is complex and is made more so by the fact that a reaction at one site on the ring may enhance the reactivity at other sites. The reactions of nucleic acids are largely the same as those of the individual nucleosides or nucleotides, the rates of reaction are often influenced by the position in the polynucleotide chain and by whether the nucleic acid is single or double stranded. The reactions of nucleosides and nucleotides are best understood in terms of the electronic properties of the various positions in the bases.[26,33] Most of the chemical reactions are nucleophilic addition or displacement reactions of types that are discussed in Chapters 12 and 13.

Positions 2, 4, and 6 of pyrimidine bases are deficient in electrons and are therefore able to react with nucleophilic reagents. The 6 position is especially reactive toward additions, while the 2 position is the least reactive. The corresponding electron-deficient positions in the purine bases are 2, 6, and 8. These positions, which are marked by asterisks on the following structures, have electrophilic character in all of the commonly occurring pyrimidines and purines.



All of the oxygen and nitrogen atoms in the pyrimidines, as well as the 5 position of the ring, have nucleophilic character and can therefore react with electrophilic centers of various reagents. A number of specific reactions that have been found useful to biochemists are described in Section H,3.

## 5. Base Pairs, Triplets, and Quartets

The purine and pyrimidine bases are the "side chains" of the nucleic acids. The polar groups that are

present in the bases can form hydrogen bonds to other nucleic acid chains, e.g., in the base pairs of the DNA double helix, and also to proteins. Figure 5-6 shows the shapes and the hydrogen bonding groups available in the bases. The number of both electron donor groups and proton donors available for hydrogen bonding is large and more than one mode of base pairing is possible.

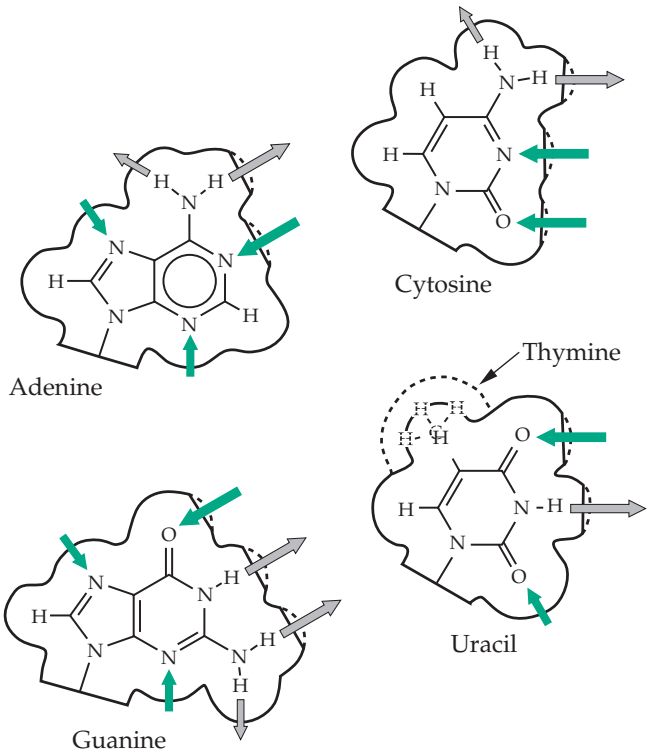
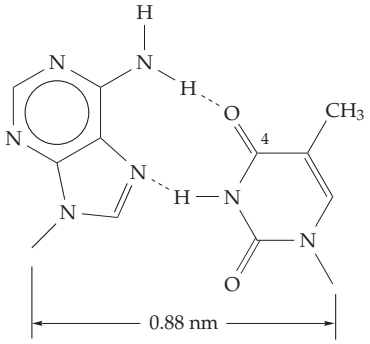



**Figure 5-6** Outlines of the purine and pyrimidine bases of nucleic acids showing van der Waals contact surfaces and some of the possible directions in which hydrogen bonds may be formed. Large arrows indicate the hydrogen bonds present in the Watson–Crick base pairs. Smaller arrows indicate other hydrogen bonding possibilities. The directions of the green arrows are from a suitable hydrogen atom in the base toward an electron pair that serves as a hydrogen acceptor. This direction is *opposite* to that in the first edition of this book to reflect current usage.

The base pairs proposed by Watson and Crick are shown in Fig. 5-2 and again in Fig. 5-7. While X-ray diffraction studies indicate that it is these pairs that usually exist in DNA, other possibilities must be considered. For example, Hoogsteen proposed an alternative A-T pairing using the 6-$NH_2$ and N-7 of adenine.[34] Here the distance spanned by the base pair, between the C-1' sugar carbons, is 0.88 nm, less than the 1.08 nm of the Watson–Crick pairs. Duplexes of certain substituted poly (A) and poly (U) chains contain only Hoogsteen base pairs[35] and numerous X-ray structure determinations have established that Hoogsteen pairs




Hoogsteen A*T base pair

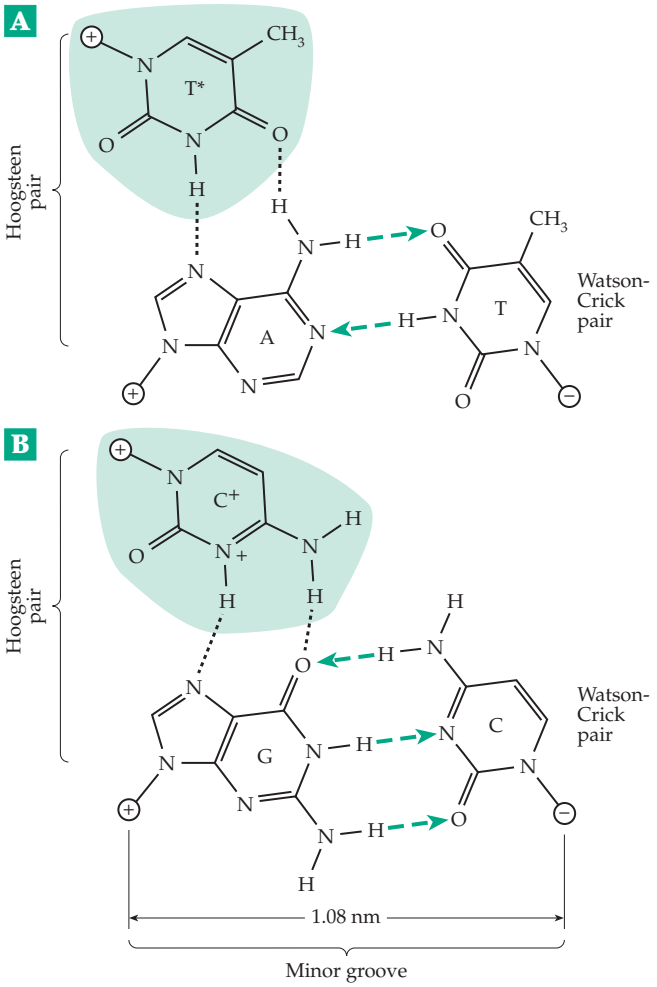



**Figure 5-7** Two base triplets that form in triple-stranded DNA and involve both Watson–Crick and Hoogsteen base pairing. (A) The triplet T*A•T, where the T (marked T*) of the third strand is hydrogen bonded as a Hoogsteen pair (·····) to an adenine of a Watson–Crick AT pair (whose hydrogen bonds are indicated (– – –). (B) The triplet $C^+$*G•C, where $C^+$ is cytosine in its N-1 protonated (low pH) form. The Watson–Crick strands are antiparallel, as indicated by the ⊕ and ⊖ signs. The third strand may have either orientation, but when it contains largely pyrimidines it is parallel to a purine-rich strand. An example is shown in Fig. 5-24.

do occur in true nucleic acids.[36] Figure 5-7 shows the structures of two Hoogsteen pairs. In each case the purine component has formed a Watson–Crick pair with a third base to give a **base triplet**. The first of these triplets may be designated T*A•T, where * represents the Hoogsteen hydrogen bonding and • represents the Watson–Crick bonding. The second triplet in Fig. 5-7, $C^+$*G•C, can form only with the N-1 protonated form (low pH form) of cytosine.

**Reversed** Watson–Crick or Hoogsteen AU or AT pairs are formed if the 2-carbonyl rather than the 4-carbonyl of the U or T makes a hydrogen bond with the amino group of adenine.[27] Because of the resulting arrangement of the ribose rings the base pairs cannot fit into the ordinary Watson–Crick double-stranded DNA structure. A reversed Watson–Crick pair can also be formed between G and C but with only two hydrogen bonds, while a reversed Hoogsteen pair can form only if a minor tautomer of cytosine is used.

Hoogsteen pairs were first observed in nature in transfer RNA molecules (Fig. 5-31). These molecules contain mostly Watson–Crick base pairs but there are also two reversed Hoogsteen pairs. One of them, between U8 and A14, is invariant in all tRNAs studied. Hoogsteen pairing also occurs in four-stranded DNA, which has important biological functions. A **G quartet** from a DNA tetraplex held together by Hoogsteen base pairs is shown in Fig. 5-8.

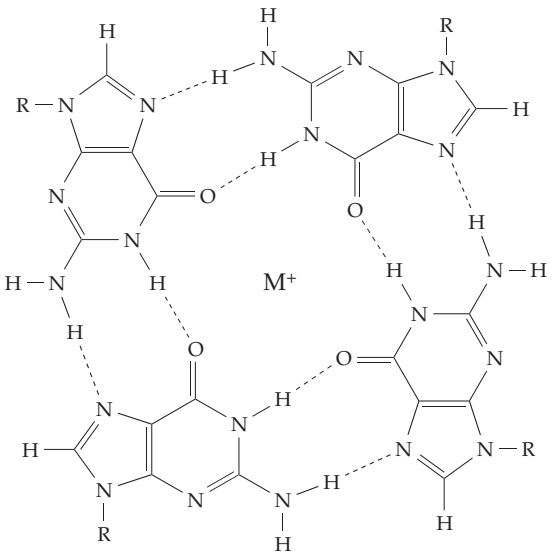
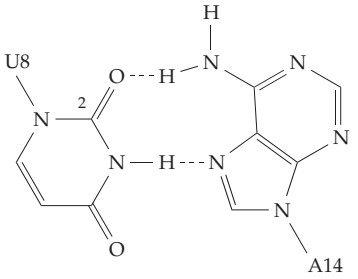
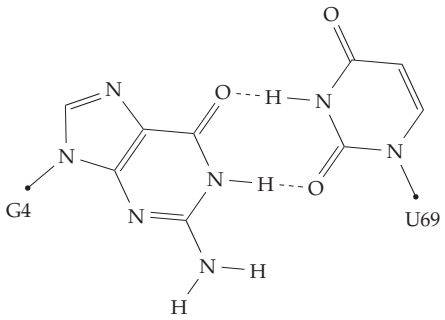



**Figure 5-8** A guanine quartet held together by Hoogsteen base pairing. This structure is found in the telomeres at the ends of linear chromosomes. Four segments of DNA, each of which may be part of a single folded strand, (see p. 227) give structures in which four or more of these G quartets are stacked one above the other. Monovalent ions, usually $K^+$ or $Na^+$, are bound in the center, although not always in the plane of the bases. See Fang and Cech[38] and Gellert *et al.*[39]




A reversed Hoogsteen base pair present in tRNA molecules

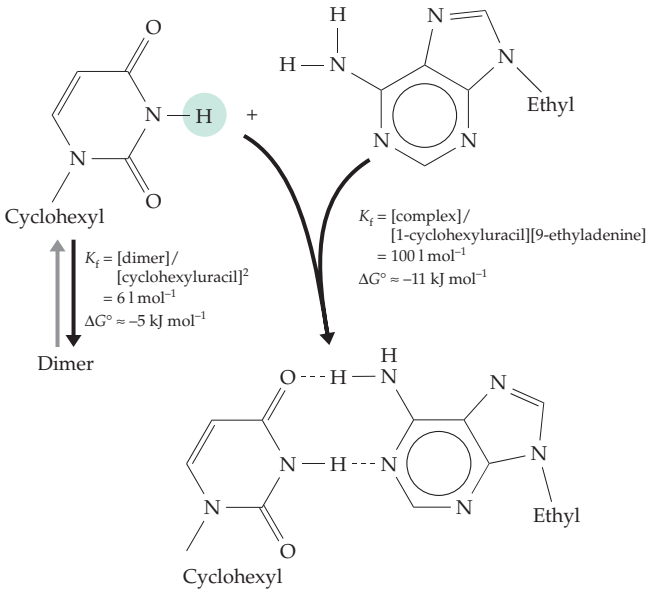
Another pairing that occurs in tRNAs allows guanine to pair with uracil, e.g., G4 with U69. This was originally proposed to account for codon–anticodon interactions betweentRNA molecules and messenger RNA (Chapter 29). It is commonly called **wobble pairing** because the uracil must wobble away from its orientation in the normal Watson–Crick pair.[27,37]




A "wobble" base pair

***Strengths of base pairs.*** How strong are the bonds between pairs of bases in DNA? The question is hard to answer because of the strong interaction of the molecules with polar solvents through hydrogen bonding and hydrophobic effects. Some insight has come from studies of the association of bases in nonpolar solvents. Thus, 1-cyclohexyluracil forms a dimer involving either hydrogen bonding or stacking, but the association is weak with the Gibbs energy of formation $\Delta G_f^\circ = 5$ kJ mol$^{-1}$. When the same compound was mixed with 9-ethyladenine, a base-paired complex formed between the two compounds with a formation constant over ten-fold greater than that for the dimer (Eq. 5-5).[40] When the circled hydrogen atom in Eq. 5-5 was replaced by $-CH_3$, which blocked the pairing, $K_f$ fell below 1 kJ/mol ($\Delta G^\circ > 0$). The difference in $\Delta G_f$ in the two cases was only 7 kJ mol$^{-1}$. Many other estimates have been made.[40a,40b] The small energies summed over the many base pairs present in the DNA molecule help provide stability to the structure.

***Stacking of bases.*** The purines and pyrimidines of nucleic acids, as well as many other compounds with flat ring structures and containing both polar and nonpolar regions, are sparingly soluble in either water




Complex (base pair) (5-5)

or organic solvents. Molecules of these substances prefer neither type of solvent but adhere tightly to each other in solid crystals. Both experimental measurements and theoretical computations[41] suggest that hydrogen bonding is the predominant force in the pairing of bases in a vacuum or in nonpolar solvents. However, in water stacking becomes important.[42] In a fully extended polynucleotide chain consecutive bases are 0.7 nm apart, twice the van der Waals thickness of a pyrimidine or purine ring, but in double-helical DNA of the B type (Figs. 5-3, 5-12) the distance between consecutive base pairs is only 0.34 nm. They are touching.

One effect of stacking is a decrease in the expected intensity of light absorption. The molar extinction coefficient of a solution of double-helical DNA or RNA is always less by up to 20–30% than that predicted from the spectra of the individual nucleosides (Fig. 5-5). This **hypochromic effect** is considered further in Chapter 23.

Because both hydrogen bonding and stacking are involved, the thermodynamics of base pairing in nucleic acids is complicated.[43] The hydrophobic parts of exposed bases tend to induce an ordering of the surrounding water molecules and therefore a decrease in their entropy. However, hydrogen bonding of the polar groups of the bases to the solvent causes a decrease in water structure. This is greater than the increase in structure around the hydrophobic regions and the stacking of bases leads to a net decrease in entropy. The entropy change $\Delta S$ for addition of a base pair to the end of a double-stranded RNA helix in a hairpin loop such as that displayed in Fig. 5-9 ranges from −0.05 to −0.15 kJ/degree per base pair.[44,45]

The enthalpy change $\Delta H$ tends to be small and positive for association of alkyl groups in water and nearly zero for association of aromatic hydrocarbons

(Chapter 2). However, $\Delta H$ is distinctly negative for association of heterocyclic bases. This has also been attributed to a decrease in the ordering of solvent around the bases as a result of exclusion of water. Attraction or repulsion of partial charges on the polar groups comprising the purine and pyrimidine bases may also be an important factor.[43,46–48] For addition of a base pair to an RNA helix, the change in enthalpy, $\Delta H$, varies from about –24 to –60 kJ/mol.[44,45]

Since $\Delta G = \Delta H - T\Delta S$, the net result is a negative value of $\Delta G$, a "hydrophobic effect" that favors association of bases. Substantial efforts have been made to estimate quantitatively the Gibbs energies of formation of helical regions of RNA molecules in hairpin stem-loops such as that of Fig. 5-9.[44,45,49–51] Table 5-2 shows the observed increments in $\Delta G_f°$ of such a helix upon addition of one base pair at the end of an existing helix. Addition of an AU pair supplies only –4 to –5 kJ to $\Delta G_f°$. The exact amount depends upon whether an A or a U is at the 5' end in the existing helix. If an AU pair is added to the helix terminating in CG or GC, about –9 kJ/mol is added. Larger increases in $-\Delta G_f°$ result from addition of GC pairs, which contain three hydrogen bonds between the bases versus the two in AU pairs.[27,41,52]

UG pairs provide a very small amount of stabilization to an RNA double helix, while the presence of unpaired bases has a destabilizing effect. The most stable hairpin loops contain four or five bases. Depending upon whether the loop is "closed" by CG or AU, the helix is destabilized by 20–30 kJ/mol. "Bulge loops," which protrude from one side of a helix, have a smaller destabilizing effect. An example of the way in which Table 5-2 can be used to estimate the energies of formation of a loop in a straight-chain RNA is illustrated in Fig. 5-9. Similar analysis of base pairing in DNA can also be done.[53–55]

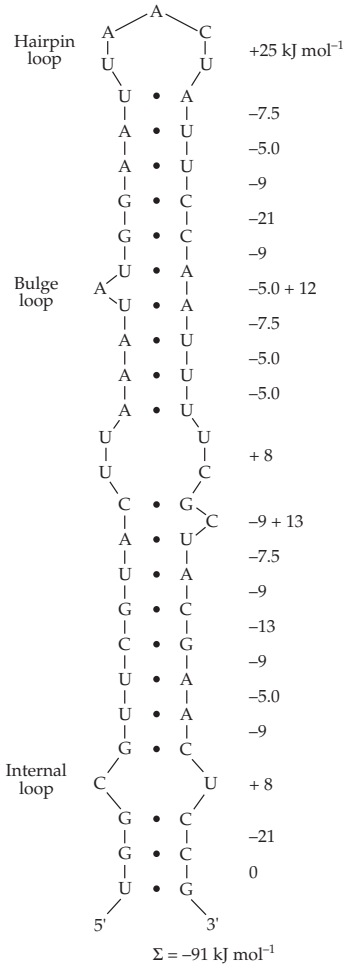



**Figure 5-9** The contribution of base-paired regions and loops to the Gibbs energy of a possible secondary structure for a 55 base fragment from R17 virus. The stem-loop structure shown here is part of a larger one considered by Tinoco *et al.*[58]

**TABLE 5-2**
**Gibbs Energies of Formation $\Delta G_f°$ at 25°C for Addition of One Base Pair to an Existing RNA Helix[a,b]**

| Base pair at end of existing helix | Base pair added | $\Delta G_f°$ (kcal mol$^{-1}$ ± 10%) | $\Delta G_f°$ (kJ mol$^{-1}$ ± 10%) |
|---|---|---|---|
| A • U | A • U | –1.2 | –5.0 |
| U[c] • A | A • U | –1.8 | –7.5 |
| C • G or G[d] • C | A • U | –2.2 | –9 |
| C • G | G • C | –3.2 | –13 |
| G • C | C • G or G • C | –5.0 | –21 |
| G • U | U • G | –0.3 | –1 |
| Hairpin loops of 4 or 5 bases | | | |
| Closed by GC | | +5 | +21 |
| Closed by AU | | +7 | +29 |
| Bulge loops | | | |
| 1 base | | +3 | +12 |
| 4–7 bases | | +5 | +21 |

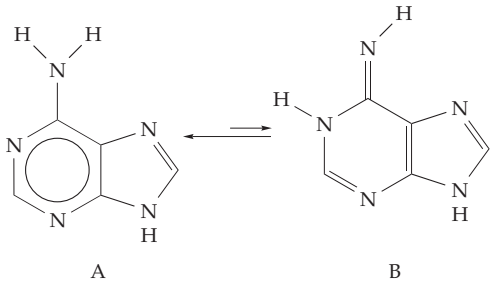
[a] Table modified from that of Tinoco *et. al.*[58]

[b] All base pairs in the table are oriented as follows: 5'—A→3' / 3'←U—5'

[c] $\Delta G_f°$ is the same for U • A added to an A • U end.

[d] $\Delta G_f°$ is the same for C • G or G • C added to an A • U end.

***Tautomerism and base pairing.*** Tautomerism has an interesting relationship to the formation of the pairs and triplets of hydrogen bonds in DNA or RNA. Each base exists predominately as one preferred




(5-6)

tautomer, but at any moment a very small fraction is present as less stable tautomers. Watson and Crick suggested that this fact may be responsible for the occurrence of some mutations.[56,57] Thus, tautomer B of Eq. 5-6 would not be able to pair with thymine, its proper pairing partner, but could pair with cytosine.

Similarly, tautomer B of uracil in Eq. 2-4 could pair with guanine instead of its proper partner adenine. If a similar event occurred to an AT pair during gene replication an incorrect copy of the gene, differing in a single "code letter," would be formed. However, because the tautomerism is affected so strongly by the environment (Chapter 2), the extent to which it may cause mispairing while replication enzymes act is uncertain.[59] Both bases in a pair could be tautomerized by synchronous transfer of protons in two parallel hydrogen bonds. However, theoretical calculations predict a high energy barrier to this process.[60] Proton transfer, which can also be induced by light, has been studied on a femtosecond scale.[61] After photochemical transfer of one proton a second is transferred within a few picoseconds.

## 6. Conformations of Nucleotides

The furanose ring of ribose or deoxyribose is flexible and can be interconverted smoothly among an infinite number of envelope (E) and skew or twist (T) conformations. See Chapter 4, Section A,2. However, there are limits set by steric and anomeric effects.[27,62–64] Conformations are often described as in Fig. 5-10 by stating which atom in an envelope conformation lies mostly out of the plane of the other four atoms. If this atom lies above the ring, i.e., toward the base, the ring

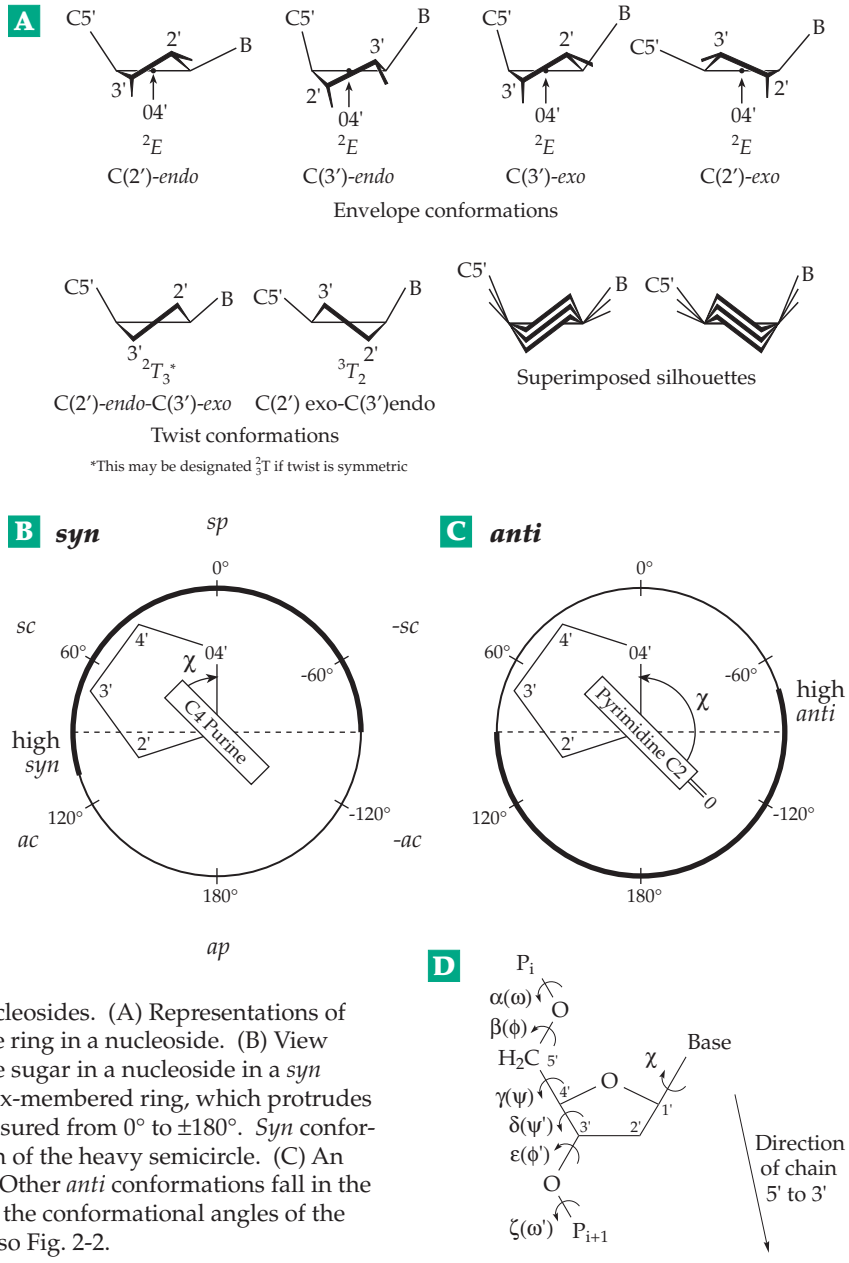
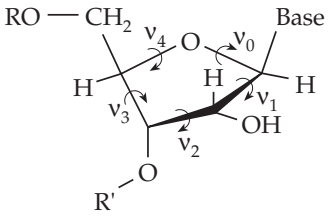



**Figure 5-10** Conformational properties of nucleosides. (A) Representations of several conformations of a ribose or deoxyribose ring in a nucleoside. (B) View down the N–C axis joining a purine base to the sugar in a nucleoside in a *syn* conformation. The atom marked C4 is in the six-membered ring, which protrudes further over the sugar ring. The angle χ is measured from 0° to ±180°. *Syn* conformations are those for which χ falls in the region of the heavy semicircle. (C) An *anti* conformation of a pyrimidine nucleoside. Other *anti* conformations fall in the region of the heavy semicircle. (D) Labeling of the conformational angles of the main chain in a polynucleotide.[25,27,37,37a] See also Fig. 2-2.

conformation is known as ***endo***; when below the ring, it is known as ***exo***. The C(2')-*endo* ($^2$E) and C(3')-*endo* ($^3$E) conformations are most commonly approximated in nucleotides and nucleic acids.[65] The C(3')-*exo* conformation is designated $E_3$, the twist conformation C(2')-*endo*-C(3')-*exo* as $^2T_3$, etc. A conformation can be specified more precisely by the five torsion angles $\nu_0$ to $\nu_4$ (which have also been designated $\tau_0$ to $\tau_4$). All of the envelope and twist conformers can be interconverted readily. The interconversions can be imagined to occur in a systematic way by **pseudorotation**, a rotation of the pucker around the sugar ring. In a commonly used convention a pseudorotation phase angle *P* ranges from 0° to 360° as the pucker moves *twice* around the ribose ring to restore the original conformation. *P* is taken as 0° for the symmetric $^3T_2$ twist, −18° for $E_2$, +18° for $^3$E, +54°




for $^4$E, +198° for $E_3$, etc. The five torsion angles $\nu_j$, for j = 0 – 4 as defined in the foregoing diagram, are related to *P* as follows:

$$\nu_j = \nu_{max} \cdot \cos[P + j \cdot \phi] \qquad (5\text{-}7)$$

where $\phi = 720°/5 = 144°$. The maximum torsion angle, $\nu_{max}$, is about 40°. See Saenger[27] for details.

Conformational alterations of ribose and deoxyribose rings can occur within polynucleotides and are of biochemical importance. An interesting consequence of changes in ring conformation is that the distance between the C(5') and N atoms attached to the sugar ring of a nucleoside may vary by as much as 0.05 nm (see Fig. 5-10A). The orientation of a base with respect to the sugar is specified by the angle χ. In one convention (Fig. 5-10B,C) its zero value is taken as that in which the N(1)–C(2) bond of a pyrimidine or the N(9)–C(4) bond of a purine is *cis* to the C(1')–O(4') bond, [sometimes called the C(1')–O(1') bond]. However, other definitions have been used.[27] Typical values of χ for nucleotides and nucleosides vary between −75° and −165°. In these ***anti*** conformations the CO and NH groups in the 2 and 3 positions of the pyrimidine ring (or in positions 1, 2, and 6 of the purine ring) are *away* from the sugar ring, while in the ***syn*** conformations they lie over the ring. *Anti* conformations are more often present than *syn* in nucleic acids. A different view of *syn* and *anti* pairs of nucleotides is shown in Fig. 5-11.

An additional five torsion angles are needed to specify the backbone conformation of a polynucleotide. According to the convention adopted by the IUB the six angles are desig-

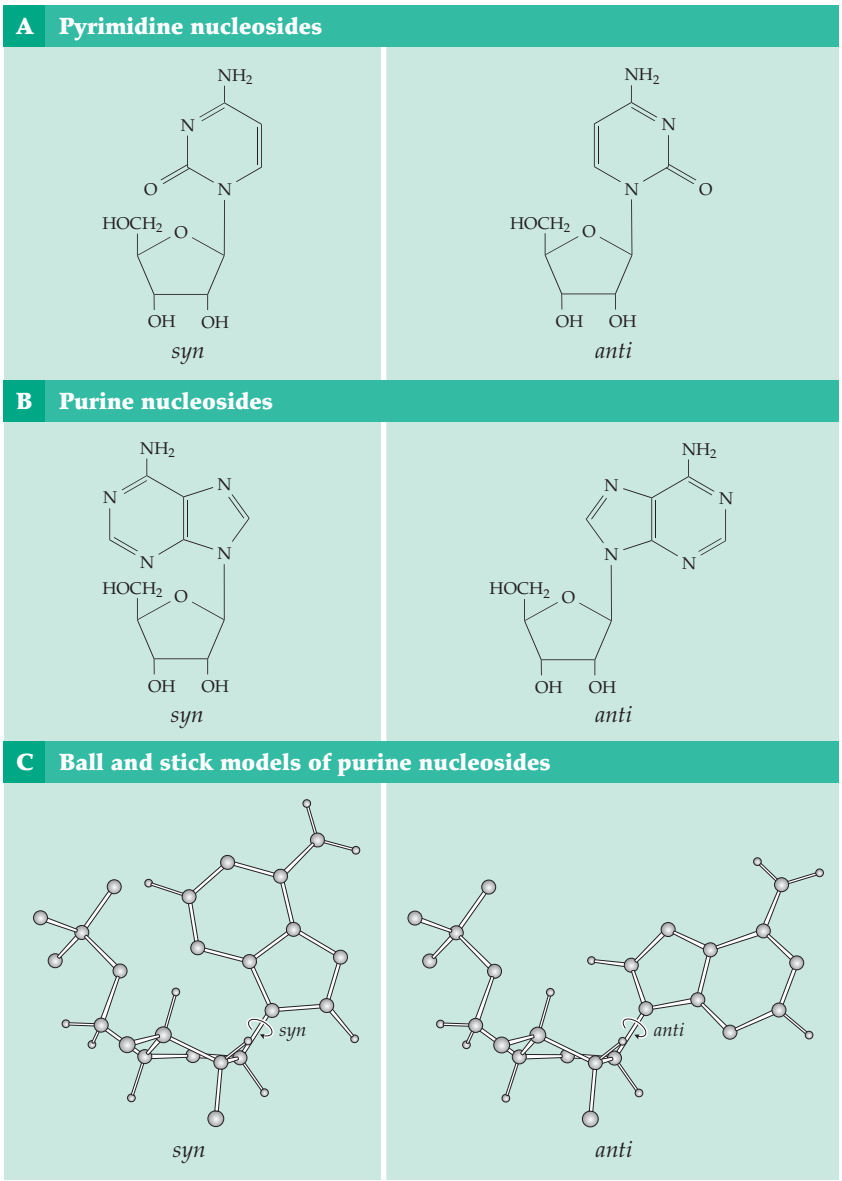



**Figure 5-11** Views of a nucleoside in *syn* and *anti* conformations.[66] (C) Courtesy of Dr. Muttaiya Sundaralingam.

nated α, β, γ, δ, ε, and ζ, as is indicated in Fig. 5-10D.[25,27] In an older but much used convention,[66] starting at any phosphorus atom the angles ω, ϕ, and ψ specify the next three torsion angles as one moves toward the 3' end of the chain, while ω′, ϕ′, and ψ′ specify the angles lying toward the 5' end (these are shown in parentheses in Fig. 5-10D). Notice that $\delta(\psi') = \nu_3 + 120°$.

## B. Double Helices

### 1. The B Form of DNA

This form of DNA, whose structure is depicted in Figs. 5-3 and 5-12, is stable at high humidity and is thought to approximate that of most DNA in cells.[27,37,67] If we look directly at the axis of the double helix and perpendicular to one of the base pairs, and ignore the fact that the base pair is asymmetric, we see that the nucleotide unit in one chain is related to the nucleotide unit lying across from it in the opposite plane by a two-fold axis of rotation (**dyad axis**). This symmetry element, which arises from the antiparallel arrangement of the chains, makes the DNA molecules from the outside look nearly identical whether viewed from one end or the other – and whether viewed as a model by the human eye or through contact with an enzyme which might act on the molecule. Actually, *the two chains are not identical*, and the genetic information can be read off from the functional groups exposed in two **grooves** in the surface of the helix (Figs. 5-3 and 5-7). The broader groove in the B form, which is referred to as the **major groove,** is about 0.85 nm deep and 1.1–1.2 nm wide when allowance is made for the van der Waals radii of the atoms. In some other forms of DNA the major groove is narrow, but it can always be identified by the larger of the arcs that can be drawn between the two N– C bonds of the nucleosidic linkages in a nucleotide pair. The **minor groove** or narrow groove, which is defined by the smaller arc between the two N–C bonds, is ~ 0.75 nm deep and 0.6 nm wide in B-DNA.

The diameter of the double helix of B-DNA, measured between phosphorus atoms, is just 2.0 nm. The rise per turn, the **pitch**, is 3.4 nm. There are about ten base pairs per turn (9.7 and 10.6 in two different crystal forms).[68,69] Thus, the rise per base pair is 0.34 nm, just the van der Waals thickness of an aromatic ring (Table 2-1). It is clear that the bases are stacked in the center of the helix. A 1000-bp (1-kb) gene would be a segment of DNA rod about 340 nm long, about 1/40 the length of the molecule in the electron micrograph of Fig. 5-13.

As is appropriate for the cell's master blueprint, DNA in the double helix is stable. Factors contributing to this stability are (1) the pairs and triplets of hydrogen bonds between the bases; (2) the van der Waals attraction between the flat bases which stack together; (3) the fact that on the outside of the molecule are many oxygen atoms, some negatively charged, which are able to form strong hydrogen bonds with water, with small ions, or with proteins that surround the DNA; and (4) the ability to form superhelices (see Section C,3). Nevertheless, the long DNA chains present in our chromosomes are frequently broken and an elaborate system of repair enzymes is needed to preserve the reliability of this master code for the cell.[73,74]

Today it is generally accepted that most DNA exists in nature as a double helix resembling the B form. However, doubts were expressed as recently as 1980. The reason for the uncertainty lay in the fact that X-ray data from the stretched "paracrystalline" DNA fibers used in earlier studies are much less precise than those from true crystals.[10] The X-ray data for B-DNA fibers could also be interpreted in terms of an alternative "side-by-side" structure which would permit easier separation of strands during replication.[75] However, numerous high-resolution X-ray structure determinations on single crystals of synthetic DNA fragments have confirmed the double-helical structure and the presence of Watson–Crick base pairs.[68,69,76–84] Similarly, fragments related to the double helices of RNA have been crystallized and the structures determined to atomic resolution.[85–87] The right-handed helical structure of DNA in solution has also been confirmed by independent methods based on electron microscopy,[88] scanning probe microscopy,[89,90] fluorescence resonance energy transfer,[91] and computation.[92]

Dickerson and associates discovered[78,79] that B-DNA has a "spine" of water in the minor groove of regions rich in A-T base pairs. Two water molecules per base pair are hydrogen bonded to form a long chain. Half of the water molecules in the chain also form hydrogen bonds to oxygen and nitrogen atoms that are exposed in the minor groove while the others hydrogen-bond to a second chain of water molecules, forming a ribbon of hydration (Fig. 5-14A).[81,93-95] Additional water binds to polar groups in the major groove. The hydration pattern is largely *local*, i.e., each base has characteristic hydration sites (Fig. 5-14B).[94] Some of the hydration sites may be occupied partially by the nomovalent ions, $Na^+$ or $K^+$, depending upon the medium. Bound divalent metal ions are also sometimes seen in X-ray structures.[81,83a] However, most cations are thought to remain mobile.

### 2. Other Double-Helical Forms of DNA

The B form of fibrous DNA is stable under conditions of high (~ 93%) humidity but at 75% humidity it is converted into **A-DNA** in which the base pairs are inclined to the helix axis by about 13° and in which the ribose rings are primarily C3'-*endo* rather than C2'-*endo*.

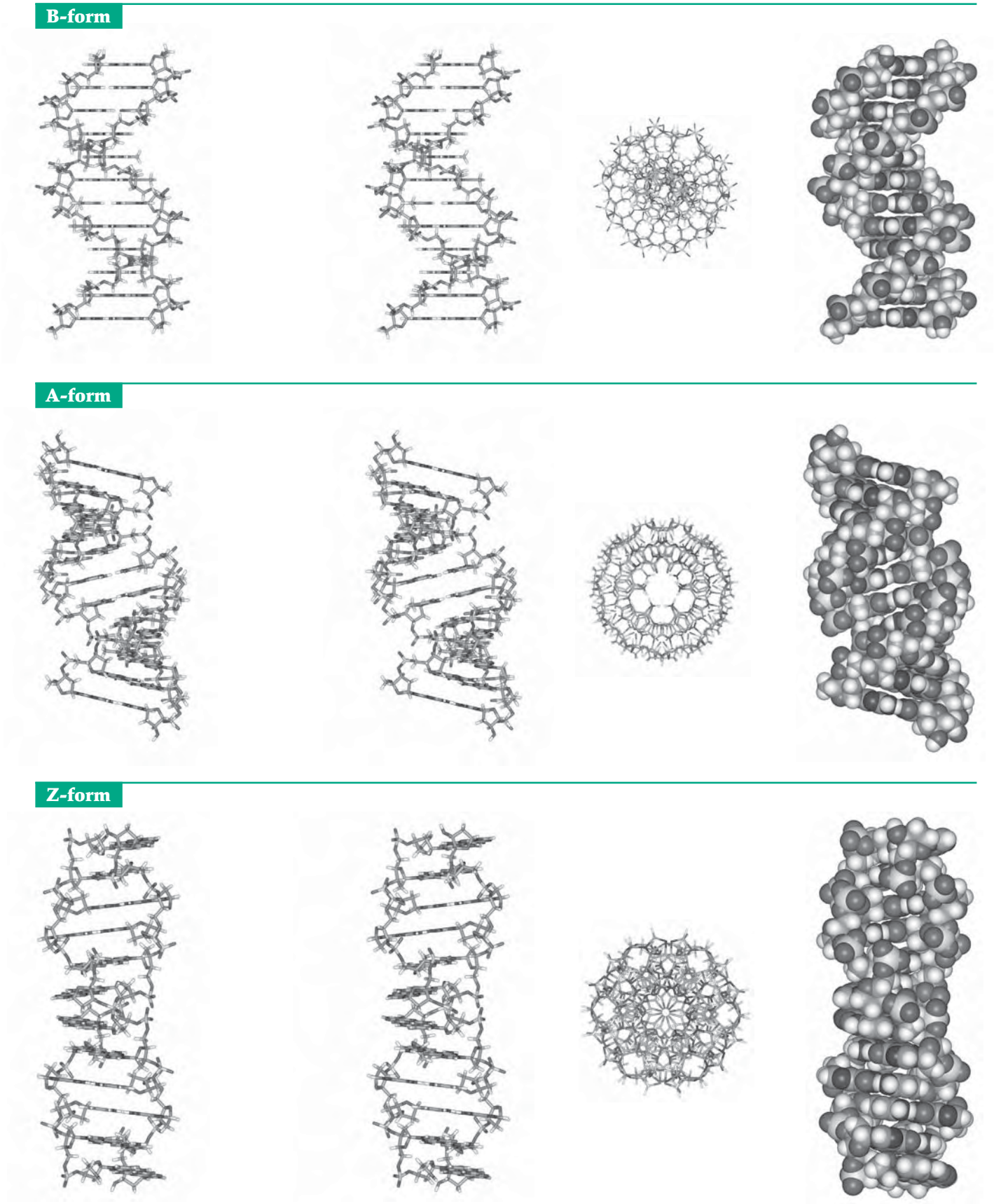



**Figure 5-12** (Left) Stereoscopic skeleton models of the B, A, and Z forms of double helical DNA. See Schlick.[70] (Center) End views (Right) Space-filling models of the same three DNA forms: B, A, and Z. Courtesy of Tamar Schlick.

There are 10.9 base pairs per turn[93,96–98] (Fig 5-12; Table 5-3). In the A form the major groove is very deep (~ 1.35 nm) and narrower (~ 0.27 nm) than in the B form and extensively hydrated.[99] The minor groove is wider (~ 1.1 nm) and shallower (~ 0.28 nm) than in the B form. The crystal structure for a form intermediate between A and B has also been reported.[99a] Paracrystalline forms of DNA known as B′, C, and D have also been observed[100,101] and others have been proposed.[102,103]

The most interesting additional form, **Z-DNA**, was discovered by X-ray studies of the alternating oligodeoxyribonucleotides d(CpGp-CpGpCpG) and d(CpGpCpG).[104] The helix of Z-DNA is *left-handed*.[71,72,105–112] The repeating unit consists of two Watson–Crick base pairs, the backbone following a zigzag pattern (Fig. 5-12). There are 12 base pairs per turn. The cytosine groups have the usual *anti* conformation but the guanosine groups are *syn*. Some of the ribose rings have the C2'-*endo* conformation characteristic of B-DNA but some are C3'-*endo*. The alternating CpG sequence is not essential for Z-DNA formation. However, alternating *syn* and *anti* conformations are important and purines assume the *syn* conformation more readily than do pyrimidines. The major groove of Z-DNA is shallow (Fig. 5-12), allowing it to accomodate bulky substituents at C8 of purines or C5 of pyrimidines. Such substituents favor Z-DNA.

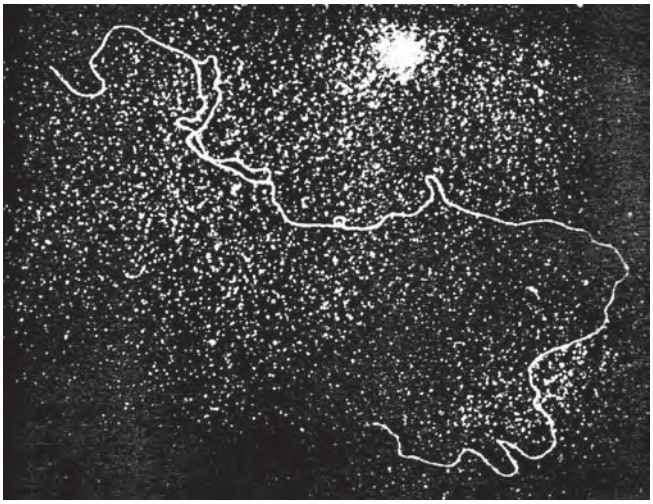


**Figure 5-13** Electron micrograph of a DNA molecule (from a bacterial virus bacteriophage T7) undergoing replication. The viral DNA is a long (~14 μm) duplex rod containing about 40,000 base pairs. In this view of a replicating molecule an internal "eye" in which DNA has been duplicated is present. The DNA synthesis was initiated at a special site (origin) about 17% of the total length from one end of the duplex. The DNA was stained with uranyl acetate and viewed by dark field electron microscopy. Micrograph courtesy J. Wolfson and D. Dressler.

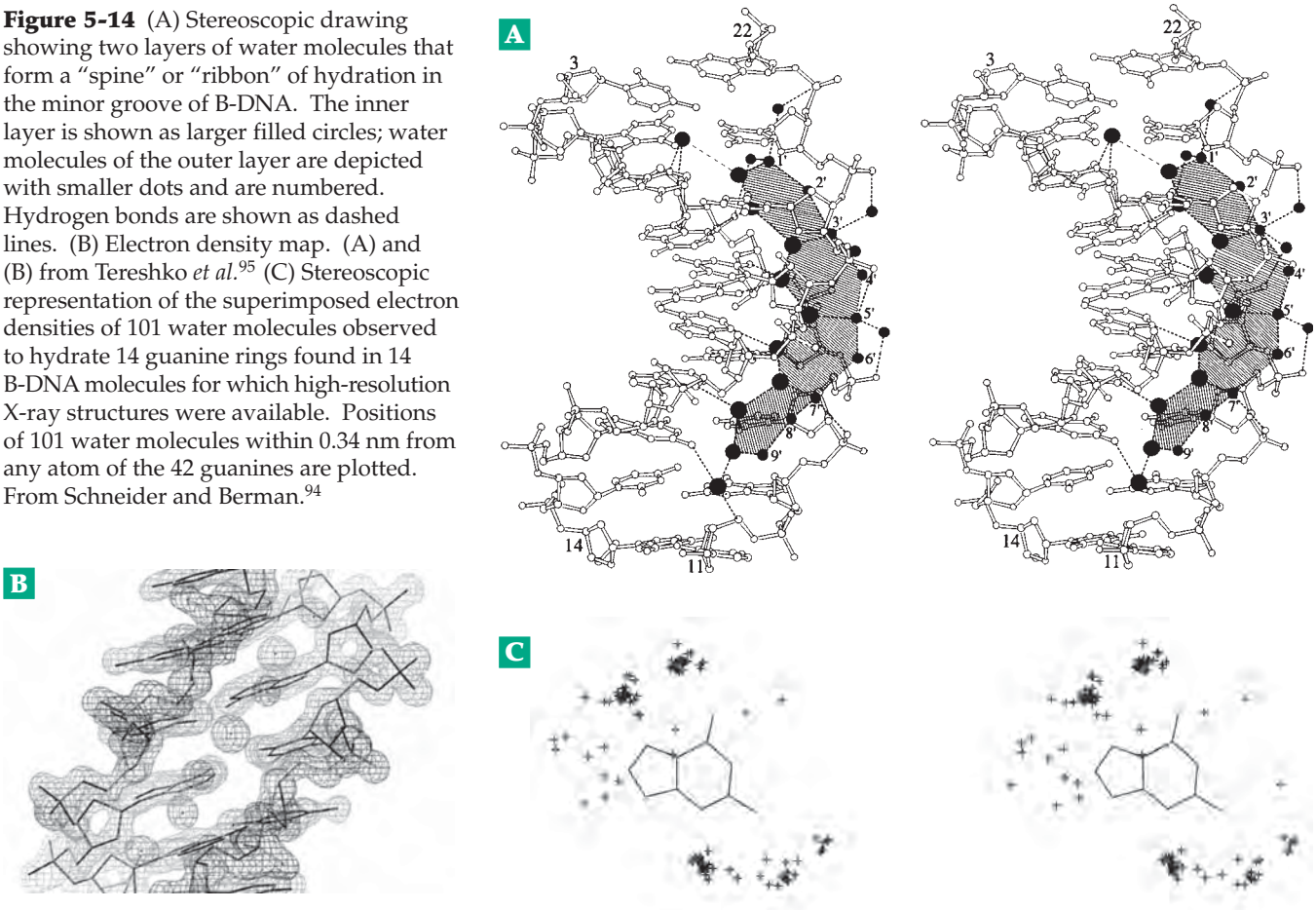



**Figure 5-14** (A) Stereoscopic drawing showing two layers of water molecules that form a "spine" or "ribbon" of hydration in the minor groove of B-DNA. The inner layer is shown as larger filled circles; water molecules of the outer layer are depicted with smaller dots and are numbered. Hydrogen bonds are shown as dashed lines. (B) Electron density map. (A) and (B) from Tereshko *et al.*[95] (C) Stereoscopic representation of the superimposed electron densities of 101 water molecules observed to hydrate 14 guanine rings found in 14 B-DNA molecules for which high-resolution X-ray structures were available. Positions of 101 water molecules within 0.34 nm from any atom of the 42 guanines are plotted. From Schneider and Berman.[94]

The B form is the most hydrated and most stable form of DNA under conditions of high humidity but even in solution it can be converted to A-DNA and Z-DNA by a high concentration of NaCl.[113] This is presumably because the salt dehydrates the DNA. Saenger and coworkers pointed out that the oxygen atoms of successive phosphate groups in the polynucleotide backbone of B-DNA are at least 0.66 nm apart, too far apart to be bridged by a water molecule.[114] However, the phosphates are individually hydrated. On the other hand, in A-DNA and Z-DNA the oxygen atoms on successive phosphates are as close as 0.53 and 0.44 nm, respectively. This allows one $H_2O$ molecule to bridge between two phosphates, stabilizing these forms in environments of low humidity. This may be one factor that affects the B to A and Z transitions.[115–117] In the narrow Z-DNA helix repulsion between negative charges on the phospho groups of opposite strands is strong. By shielding these charges high salt concentrations also help to stabilize Z-DNA.

Since the Z form of DNA is favored in regions rich in G-C pairs,[118,119] it is reasonable to expect that it may occur in nature. Antibodies have been prepared which bind specifically to regions of DNA in the Z form and have been used to identify many such regions.[111,120] Genetic studies in *E. coli* have also provided strong evidence that left-handed DNA sequences are formed *in vivo* in that bacterium.[121,122] Z-DNA-forming sequences have also been found in the *Halobacterium* genome.[119] Segments of Z-DNA may occur in control regions called **enhancers** (Chapter 28)[121–124] and Z-DNA may also form behind RNA polymerase molecules that are moving along a gene while synthesizing mRNA. The associated negative supercoiling of the DNA could cause it to assume the Z form. This region of Z-DNA can, in turn, be a site for interaction with specific proteins.

In the Watson–Crick structure the two strands are *antiparallel*, an essential for replication. However, stable segments of double-stranded DNA with *parallel* strands can also be formed and may occur in specialized regions of the genome.[125–128]

**TABLE 5-3**
**Helix Parameters for Three Types of DNA[a]**

| | Form of DNA or RNA (A-form) | | |
|---|---|---|---|
| Parameter | B | A | Z |
| Helical twist, degrees<br>mean | 28–42<br>36 ± 4 | 16–44<br>33 ± 6 | GC –51 ± 2<br>CG –8.5 ± 1 |
| Base pairs per turn | 10.0<br>(9.7–10.6) | 11-12 | 12 |
| Helix rise per base pair, nm | 0.34 ± .04 | 0.29 ± .04 | GC 0.35 ± .02<br>CG 0.41 ± .02 |
| Base inclination, degrees | –2.0 ± 5 | 13 ± 2 | 8.8 ± .7 |
| Propeller twist, degrees | 12 ± 5 | 15 ± 6 | 4.4 ± 3 |
| Base roll degrees | –1.0 ± 5 | 6 ± 5 | 3.4 ± 2 |
| Predominant conformation of deoxyribose | C2′-*endo* | C3′-*endo* | C C2′-*endo*<br>G C3′-*endo* |
| Depth of grooves (nm)<br>Major<br>Minor | <br>0.85<br>~0.75 | <br>1.35<br>~0.28 | <br>very shallow<br>very deep |
| Width of grooves (nm)<br>Major<br>Minor | <br>1.1–1.2<br>0.6 | <br>0.27<br>~1.1 | <br>broad<br>narrow |

[a] See R. E. Dickerson[78,78a]; based on single-crystal X-ray analysis.

## 3. The Conformational Flexibility of Double Helices

Local variations in the sequence of nucleotides affect the conformation of a DNA molecule and it is clear that the helix is not uniformly coiled throughout the entire length.[80,103,124,129–134] While most helix segments probably have a right-handed twist others may be left-handed. Most DNA is probably in the B form but there are segments in the A form. These may arise from formation of hybrid duplexes with RNA, which assume the A conformation and are also favored by certain base sequences. Rules for predicting the DNA conformation from the nucleotide sequence have been proposed.[116,135] In the simplest case[135] we consider each pair of adjacent nucleotides in the double helix. There are 16 possible pairs in one chain. These can be designated as in the following examples: (AA,TT), (CG,CG), and (AG,CT). The first two letters within the parentheses represent the sequence (from 5' to 3') in one chain while the second pair of letters represent the sequence (again from 5' to 3') in the complementary chain. The rules state that (AA,TT) or (TT,AA) repeated in a sequence will stabilize the B form of DNA. Repetition of (CC,GG) or (GG,CC) will favor conversion to the A form. Repetitions of (CG,CG) favor the Z form, especially if an alternating sequence of purines and pyrimidines is present throughout the (G + C)-rich region.

Even within the regular B structure the

torsion angles χ and δ (Fig. 5-10) are variable. Their changes are highly correlated.[136,136a] As χ ranges from – 140° to – 90°, ζ varies between 80° and 160°. Other pairs of torsion angles are also correlated. Besides allowing for changes between B, A, and Z conformations, this flexibility of the DNA helix together with cooperativity with adjacent base pairs may allow transmission of conformational effects for some distances along a DNA helix.[137,138] Supercoiling, discussed in Section C,3, also affects the helical conformation.

DNA can be stretched into yet another form or forms. Application of a force of 65 – 70 piconewtons (pN) stretches B-DNA by 70%. This "overstretched DNA" may also be important biologically.[130,139–141a]

***Rotational and translational movements of bases.*** In considering the conformational flexibility of a polynucleotide it is useful to define the parameters associated with movement of base pairs or individual bases. The possible movements are indicated in Fig. 5-15 and in Table 5-3. A **long axis** is established for each base pair in a structure for which X-ray data are available. Making use of these axes, we can move along the helix and measure the **angle of twist** from one base pair to the next.[142–145] While the mean value of the twist for B-DNA is 36°, values of 28–42° have been observed in structures of oligonucleotides determined by X-ray analysis. The **base tilt**, which is nearly zero for B-DNA, is less variable. Within each base pair there is also a **propeller twist** around the long axis.[146] It averages about 12° for B-DNA. In addition, the whole base pair may **roll**, i.e., be rotated around its long axis by several degrees. These motions relieve steric interferences in the center of the helix, for example, those that would arise between purines in adjacent base pairs if the same helix parameters were imposed on each base pair.[142,147] A more detailed analysis reveals the large range of motions, mostly small, which are described in Fig. 5-15.[145,148] The mathematics needed to deal with computer-based modeling of polynucleotide structures has been developed.[80,149,150] Its application showed that there is a strong correlation between twist and roll. Typical average values of base tilt, propeller twist, and roll are given in Table 5-3 for B-, A-, and Z-DNA.

***Bends and bulges.*** If we overlook the ridges and grooves on their surfaces the DNA structures shown in Figs. 5-3 and 5-12 are straight rods. However, real DNA rods are crooked and may contain distinct bends. The

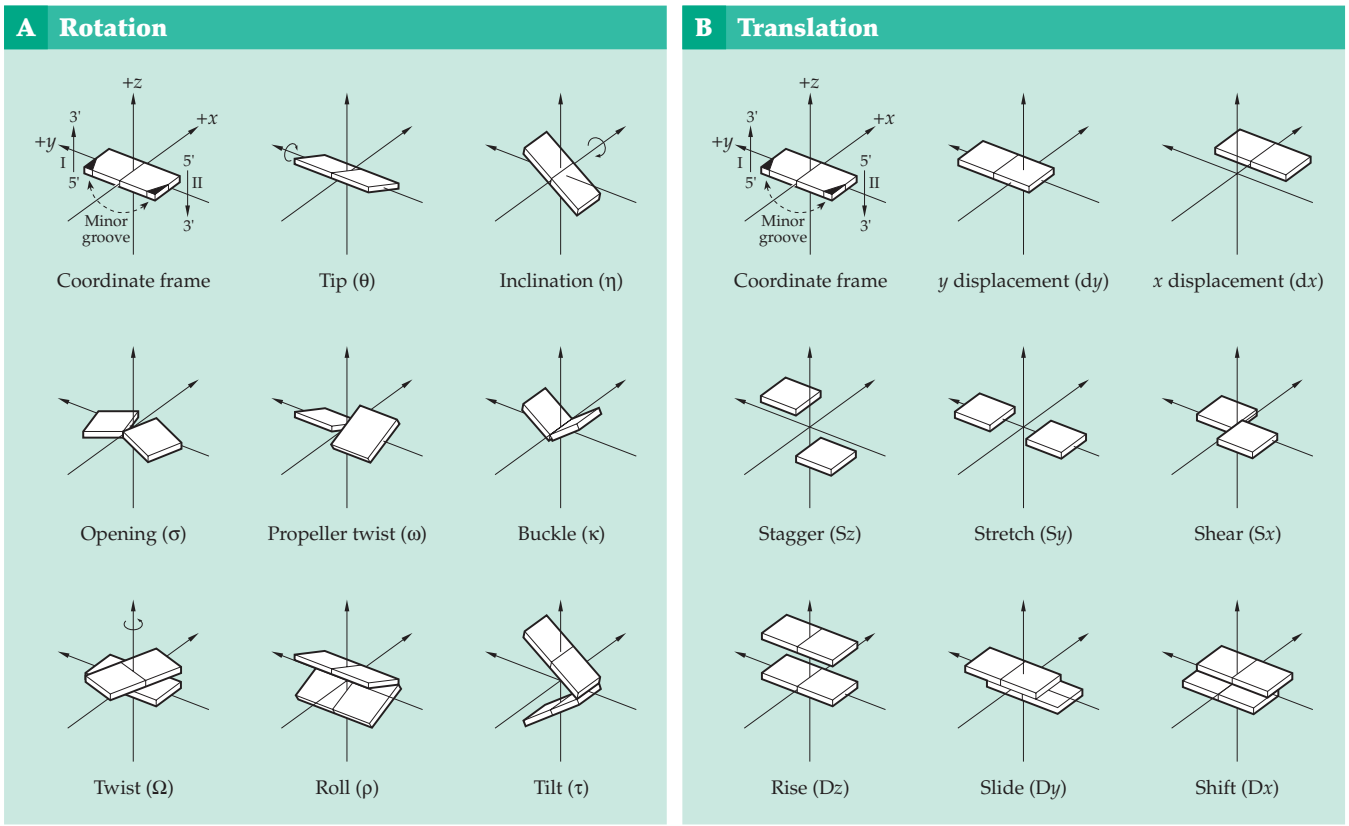



**Figure 5-15** (A) Drawing illustrating various rotational movements of bases in polynucleotides. *Upper two rows*: rotations of two bases of a pair. *Lower row*: rotations involving two successive base pairs. (B) Translational movements. *Upper two rows*: involving two bases of a pair. *Lower row*: two successive base pairs. From Diekmann.[145]

existence of **bent DNA** in nature was first discovered from study of fragments enzymatically cut from kinetoplast DNA of tropical parasites.[151,152] The bent fragments moved anomolously slowly during electrophoresis in polyacrylamide gels of small pore size. Evidently, the bent shape impedes movement through the pores. Since the initial discovery, regions of bent DNA have been found in origins of replication[152–154] and other specialized sites. Bending of DNA because of local sequence variations may also be important to the positioning of nucleosomes (Fig. 27-3) on DNA.[155]

Several causes for DNA bending have been identified.[124,138,142,156–164] For example, a sharp 26° bend is expected at the junction between B- and A-DNA segments.[158] The presence of a thymine photodimer (Eq. 23-26) may cause a 30° bend.[165] Some naturally occurring bent DNAs contain repeated (A+T)-rich sequences. This suggested that the sequence (AA,TT), containing adjacent thymine methyl groups, can be thought of as a wedge, consisting of both roll and tilt components. Such a wedge repeated at intervals averaging 10.5 bp along the helix might cause the helix to bend.[129,144] However, computer modeling experiments suggest that thymine methyl groups do not distort the DNA helix. The bending may be a result of the greater tendency for AT pairs to roll and to enhance the van der Waals and electrostatic attractions between the 2'-O of cytosine and the 2'-$NH_2$ of guanine in the CG base pairs of bent DNA.[157,166] Bending can occur in both (A+T)- and (G+C)-rich regions and is often induced by binding to proteins or protein complexes that act in replication, transcription, and recombination.[167]

Various errors are made during replication or recombination of DNA. Incorporation of an incorrect nucleotide will cause a mismatched base pair in which proper hydrogen bonds cannot be formed. Most of the mispaired bases that result from mistakes in replication are removed by repair processes (Chapter 27). Those that remain can often assume alternative pairings that distort the helix only slightly. For example, a G-T wobble pair fits readily into an oligodeoxyribonucleotide double helix. Even a bulky G-A base pair causes little perturbation of the helix.[168] Incorporation of an extra nucleotide into one strand of the DNA will create a bulge in the helix.[169] NMR spectroscopic studies on bulged oligonucleotides have shown that the extra base can be stacked into the helix causing a sharp bend.[170,171] However, some oligonucleotides with mismatched bases crystallize as straight helices with the extra nucleotide looped out.[171–173] Mismatched base pairs tend to destabilize helices.[174]

***Interactions with ions.*** Because each linking phospho group carries a negative charge (two charges per base pair, Fig. 5-2) the behavior of polynucleotides is strongly affected by cations of all kinds. The predominant small counterions within cells are $K^+$ and $Mg^{2+}$. They are attracted to the negative charges on the polynucleotide backbone and, although they remain mobile, they tend to occupy a restricted volume.[175–177] Some may bind in well-defined locations as in Fig. 5-8. Because of the presence of these positive ions the interactions of nucleic acids with cationic groups of proteins are strongly affected by the salt concentration.

Organic cations compete with the simple counterions $K^+$ and $Mg^{2+}$. Among these, the polyamines are predominant.[178] Crystal structures have revealed that spermine binds across the deep grooves of tRNA and the major groove of B-DNA. Spermine also binds into the deep groove of A-DNA interacting by hydrogen bonding with bases in GTG sequence in both strands.[179] It binds tightly to CG-rich sequences in the minor groove of Z-DNA, where it tightens the structure and shortens the helix.[110] At higher concentrations of DNA, as occur in cell nuclei, spermidine induces the conversion of the DNA into liquid crystalline phases.[180]

Heavier metal ions and metal complexes can find sites on nitrogen atoms of the nucleic acid bases. Examples are the platinum complex **cisplatin** and the DNA-cleaving antibiotic **neocarzinostatin** (Box 5-B). Can metals interact with the π electrons of stacked DNA bases? A surprising result has been reported for intercalating complexes of ruthenium (Ru) and rhodium (Rh). Apparent transfer of electrons between Ru (II) and Rh (III) over distances in excess of 4.0 nm, presumably through the stacked bases, has been observed,[181] as has electron transfer from other ions.[181a] Stacked bases are apparently semiconductors.[182]

## C. The Topology and Dynamics of Nucleic Acids

### 1. Rings, Catenanes, and Knots

While a DNA molecule may exist as a straight rod, the two ends are often covalently joined. Thus, the chromosomes of *E. coli* and of other bacteria are single closed circles. Circular DNA molecules are also found in mitochondria, chloroplasts, and many viruses. Further complexity arises from the fact that the circles of DNA are sometimes interlocked in chainlike fashion (**catenated**). An unusual example of this phenomenon is the presence of thousands of small catenated DNA circles in the single mitochondrion of a trypanosome (Fig. 5-16).[183] Sometimes circular DNA is **knotted** as in Fig. 5-17.[184–186] Knots and catenanes often appear as intermediate forms during replication and recombination, especially involving circular DNA.[187,188]

Methods have been devised for synthesis of even very complex DNA knots.[185,186] Let's look briefly at the topology of knots. The three simple knots shown here have a chirality beyond that of the nucleotide

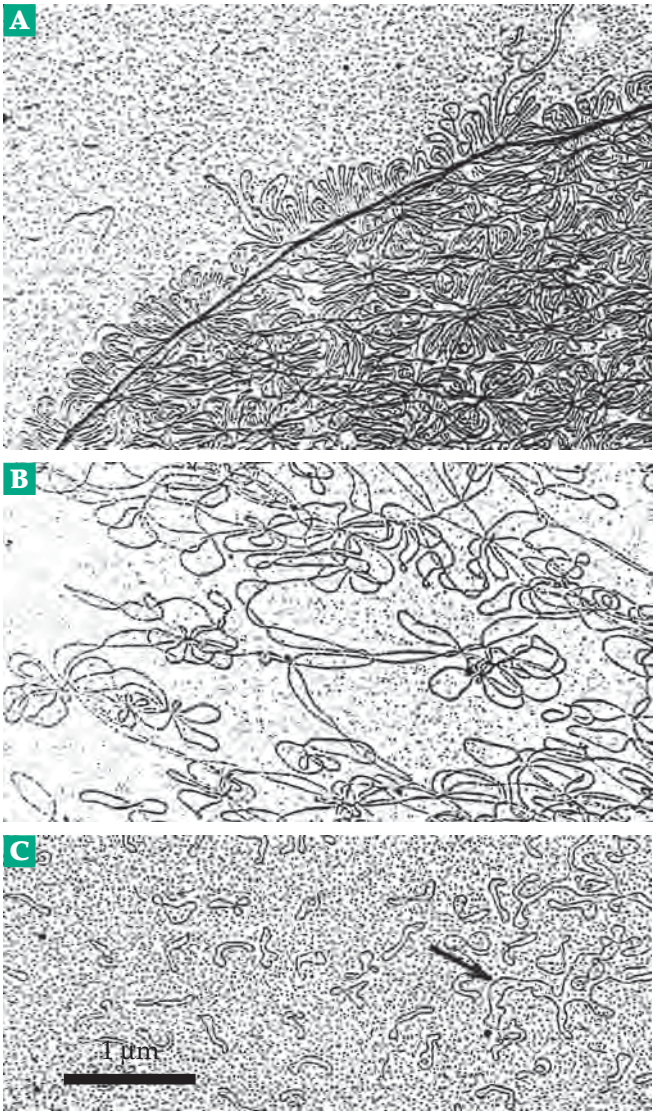



**Figure 5-16** (A) Electron micrograph of the network of catenated DNA circles in the mitochondrion of the trypanosome *Crithidia fasciculata*. (B) and (C) The same network after treatment with a **topoisomerase** from bacteriophage T4 that catalyzes a decatenation to form individual covalently closed circles (Chapter 27). Five times as much enzyme was added in (C) as in (B). Two sizes of circles are present. Most are "minicircles", each containing about 2300 bp but a smaller number of larger ~35-kb "maxicircles" are also present. One of these is marked by the arrow. From Marini, Miller, and Englund.[183]

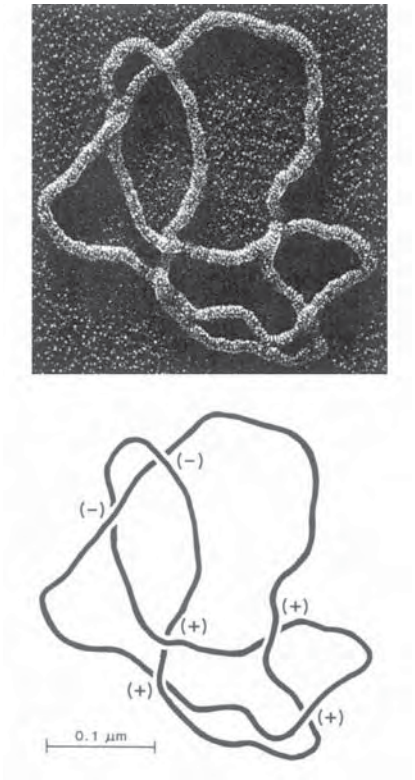
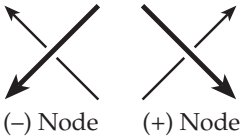



**Figure 5-17** Electron micrograph of a six-noded knot made by the Tn3 resolvase which is involved in movement of the Tn3 transposon (Chapter 27) from one location to another within the genome. Putative six-noded knot DNA was isolated by electroelution from an agarose gel. The knots, which are nicked in one strand, were denatured to allow the nicked strand to slide away and leave a ssDNA knot. This was coated with *E. coli* recA protein (Fig. 27-24) to greatly thicken the strand and to permit the sign of each node (designated in the tracing) to be seen. From Wasserman *et al.*[184]

units.[189,190] This can be expressed by indicating the sign of each node in the knot:


The node is negative if the crossing is like that in a left-handed supercoil and positive if like that in a right-handed supercoil:[191,192]




## 2. Supercoiled DNA

Double-stranded DNA in solutions of low salt content usually assume the B-DNA conformation with 10.4–10.5 base pairs per turn. If the two ends are joined the resulting covalently closed circular DNA will be "relaxed." However, there are topoisomerases that act on this form of DNA by cutting both strands, holding the ends, and twisting the two chains (Chapter 27).

The energy source for the process is provided by cleavage of ATP. **DNA gyrase** untwists relaxed circular dsDNA one turn at a time and reseals the cut ends. A **reverse gyrase** from certain bacteria twists the relaxed DNA more tightly. In both cases the change causes the DNA to form **superhelical turns**.[67,193–197] These may be either **solenoidal** or **plectonemically interwound** (as a twisted thread; Fig. 5-18).

The geometric and topological properties of closed supercoiled DNA molecues may be described by three quantities: The **linking number** (***Lk*** also called the winding number, α), the **twist** (***Tw)***, and the **writhe** (***Wr***). If a segment of double helical DNA were laid on a flat surface and the ends were joined to form a relaxed circle both *Lk* and *Tw* would equal the number of helical turns in the DNA. The writhe *Wr* would be zero. The linking number is a topological property. It has an integral value which is unchanged if the DNA molecule is distorted. It can be changed only for DNA with open ends. When the ends are joined the linking (winding) number is constant unless one or both chains are cleaved. However, twist and writhe are geometric properties, which can change according to Eq. 5-8 while the *Lk* remains constant. The twist is related to the number of helical turns while the writhe is related to the number of superhelical turns.[27,67,198–200]

$$Lk = Tw + Wr \qquad (5\text{-}8)$$

However, *Tw* and *Wr* do not usually have integral values. It is hard to define the relaxed state for a circular DNA. For example, changing the ionic composition will alter *Tw* and *Wr* according to equation 5-8. Both *Lk* and *Tw* are taken as positive for a right-handed **toroidal** (solenoidal) supercoil or a left-handed interwound twist.

For relaxed B-DNA, *Lk* is equal to the total number of base pairs in the circle divided by 10.5, $Wr = 0$, and $Tw = Lk$. Both *Lk* and *Tw* are positive in the right-handed B- and A-DNA forms, but *Tw* is negative for Z-DNA. In closed circular DNA the value of *Wr* is usually negative, the secondary structure being a fully formed Watson–Crick helix but with right-handed interwound superhelical turns or left-handed toroidal superhelical turns. The helix is said to be **underwound** ($Lk < Tw$).

Some of the topological properties of double-stranded DNA can be demonstrated by twisting together two pieces of flexible rubber tubing whose ends can be joined with short rods to form a closed circle.[196] Twist the tubing in a right-handed fashion as tightly as possible without causing supercoiling ($Lk = Tw$; $Wr = 0$). If the ends are now joined the circle will be relaxed. Now twist one turn tighter before joining the ends. A right-handed toroidal supercoil will be formed ($\Delta Wr = 1$; $\Delta Lk = 1 + Tw$; *Tw* is the same as before). If twisting is continued until several supercoils appear before the ends are joined, two interconvertible forms result. One form has right-handed toroidal supercoils and the other left-handed interwound supercoils as in Fig. 5-19. If relaxed circular DNA is unwound by one turn by cutting and resealing one chain, a single right-handed interwound supercoil will be formed ($\Delta Wr = -1$, $Lk = Tw - 1$). On the other hand, if the two chains in a closed relaxed helix are pried apart, as happens during intercalation (Section 3), *Lk* must remain constant, *Tw* will decrease, and *Wr* will increase with appearance of left-handed supercoiling.

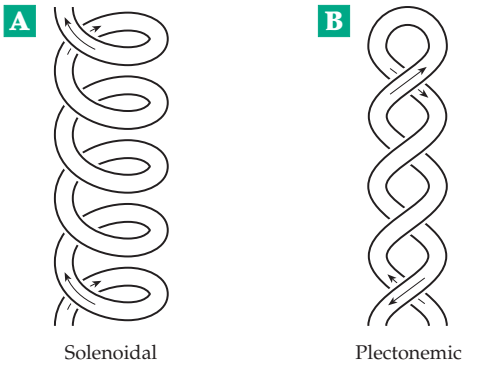



**Figure 5-18** Two forms of supercoiling of a DNA duplex: (A) Solenoidal. (B) Plectonemically interwound. These are both negatively supercoiled, as can be deduced from the arrows at the nodes, but the solenoid is left-handed and the plectonemic form right-handed. From Wasserman and Cozzarelli.[187]

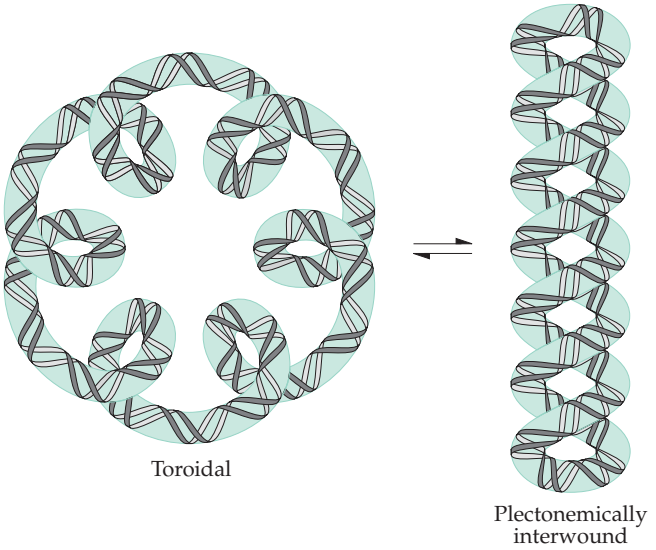



**Figure 5-19** Topological equivalence of toroidal (solenoidal) and plectonemically interwound forms of a circular DNA. These two forms have a constant value of the linking number *Lk* (or α), the twist *Tw*, and writhing number *Wr*.

During replication of DNA (Chapter 27) pairing of some bases associated with the replication apparatus is prevented. Upon release of the constraint the newly replicated DNA forms base pairs and becomes supercoiled.

Since *Lk* is constant in circularly closed DNA a change of one turn of B-DNA (*Tw* = 1) into a turn of Z-DNA (*Tw* = −1) will cause the writhing number *Wr* to change by −2. Conversely, if the writhing number is forced to change by −2, a turn of Z-DNA may develop somewhere in a suitable (G + C)-rich region of the DNA.

Some of the information about supercoiled DNA can be summarized as follows:

| | |
|---|---|
| Relaxed circular DNA | $Lk = Tw$, $Wr = 0$ |
| Underwound circular DNA | $Lk < Tw$; *Wr* is negative; right-handed interwound supercoils or left-handed toroidal supercoils may be present. Alternatively $Lk = Tw$ and left-handed Z-DNA regions may appear. |
| DNA with intercalated molecules | DNA is partially untwisted; *Tw* is lowered; *Wr* is increased with decrease in number of negative supercoils. As *Tw* approaches *Lk* the DNA becomes relaxed. |

The **superhelix density** of a DNA molecule is often expressed as $\sigma = Wr/Tw \approx$ number of superhelical turns per 10 bp.[31,198,199a,199b,201] In most naturally occurring circular DNA molecules σ is negative, a typical value being −0.05 (~5 negative superhelical turns per 1000 bp). The presence of superhelices in circular DNA molecules can be recognized readily by its effect upon the sedimentation constant of the DNA. Naturally occurring supercoiled DNA from polyoma virus sediments rapidly but after nicking of one of the strands of the double helix by brief exposure to a DNA-hydrolyzing enzyme the resulting relaxed form of the molecule sediments more slowly. Supercoiling lowers the viscosity of solutions of DNA and increases the electrophoretic mobility (Fig. 5-20) and may also be recognized by electron microscopy.

Naturally occurring or artificially prepared supercoiled DNA molecules can often be separated by electrophoresis into about ten forms, each differing from the other by one supercoiled turn and by $\Delta Lk = \pm 1$ (Fig. 5-20). The relative amounts of these **topological isomers** form an approximately Gaussian distribution. The isomers apparently arise as a result of thermal fluctuations in the degree of supercoiling at the time that the circles were enzymatically closed.[202]

Why is DNA in cells supercoiled? One effect of supercoiling is to contract the very long, slender double helices into more compact forms. In eukaryotic cells much of the DNA exists in **nucleosomes**. Each bead-like nucleosome consists of a core of eight subunits of proteins called **histones** around which an ~140 bp length of DNA is coiled into two negative, left-handed toroidal superhelical turns (Fig. 5-21).[203,204] There is some spacer DNA between nucleosomes. Otherwise a nucleosome providing two superhelical turns per 200 bp would produce a superhelix density of −0.1, twice that observed. A detailed analysis of the geometric properties of DNA wrapped around nucleosomes or other protein particles has been developed.[199–199c] Nucleosome formation is thought to protect DNA and to keep it in a more compact state than when it is fully active and involved in transcription. Nucleosomes, discussed further in Chapter 27, are also important in the regulation of transcription. Even though nucleosomes as such are absent, bacterial DNA also has a superhelix density of − 0.05, apparently a result of interaction with other proteins such as the histonelike HU.[205] Interaction with smaller molecules can also affect supercoiling.[206]

The presence of naturally supercoiled DNA and a variety of topoisomerases suggests that the control of DNA supercoiling is biologically important. In fact,

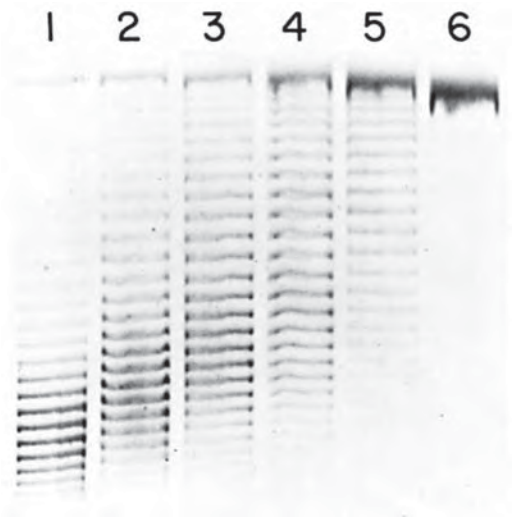



**Figure 5-20** Electrophoresis of DNA from the SV40 virus with varying numbers of superhelical turns. Molecules of the native DNA (lane 1) move rapidly toward the anode as a series of bands, each differing from its neighbors by one superhelical turn. The average number of superhelical turns is about 25. Incubation with a topoisomerase from human cells causes a stepwise removal of the superhelical turns by a cutting and resealing of one DNA strand. The DNA incubated with this enzyme at 0°C for periods of 1, 3, 6, 10, and 30 min (lanes 2–6) is gradually converted to a form with an average of zero supercoils. For details see Keller.[202,212] Electrophoresis was carried out in an 0.5% agarose –1.9% polyacrylamide slab gel (17 x 18 x 0.3 cm). The bands of DNA were visualized by staining with the fluorescent intercalating dye ethidium bromide. Photograph courtesy of Walter Keller.

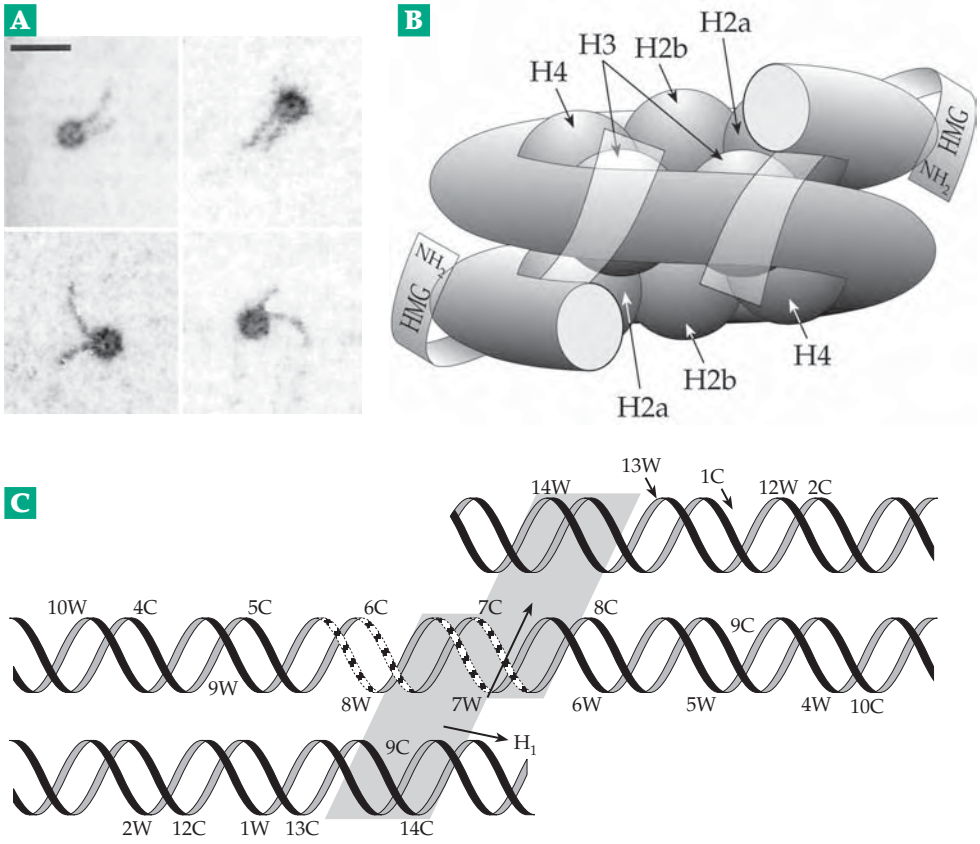



**Figure 5-21** Nucleosomes. (A) Electron micrographs of individual nucleosomes reconstituted from 256-bp DNA fragments and separated proteins. From Hamiche *et al.*[213] Courtesy of Ariel Prunell. (B) Model of a nucleosome core. The 1.75-turn (145-bp) DNA superhelix winds around the histone octomer which consists of two subunits apiece of histones H2A, H2B, H3, and H4. In addition, two elongated molecules of proteins HMG-14 or HMG-17 are indicated (see also Chapter 27). (C) Schematic radial projection of the double-helical DNA showing areas protected from cleavage by hydroxyl radicals (see Fig. 5-50) by the bound proteins. The shaded areas are those protected by HMGs. The zigzag lines near the dyad axis indicate the most prominent regions of protection. (B) and (C) are from Alfonso *et al.*[214]

topoisomerases play essential roles in both replication of DNA and in transcription of genes. Supercoiling requires energy and topoisomerases that induce supercoiling must provide energy, e.g., by cleavage of ATP. Conversely, supercoiled DNA can be a source of energy for biological processes. It has been estimated that the $\Delta H$, $\Delta S$, and $\Delta G$ per mole for formation of a single superhelical turn are 35.9 kJ, 68 J/ °K, and 14.6 kJ, respectively.[207] A reduction in superhelix density is accompanied by a decrease in Gibbs energy and can therefore be coupled to other processes that have positive values of $\Delta G$. An example is conversion of (G + C)-rich regions into the Z form of DNA, which is favored by negative supercoiling.[207,208] Supercoiling affects binding of various proteins to DNA[197,209,210] as well as intercalation (discussed in the next section) and formation of cruciform structures (Section D,3).

The folding of DNA into compact forms, such as that in chromosomes, is also influenced by supercoiling. In the absence of nucleosomes supercoiled DNA may assume the plectonemic form (Figs. 5-18, 5-19). Segments of the resulting rods may aggregate side-by-side in a liquid crystalline state.[200,211] The relatively high cation concentration within cells favors this transformation.[215–217] Polyamines such as spermidine are especially effective in promoting aggregation of DNA, in formation of Z-DNA,[110] and possibly in facilitating cooperative processes that require that two DNA molecules interact with each other.[218,219]

### 3. Intercalation

Flat, aromatic, hydrophobic rings are often able to insert themselves between the base pairs of a DNA duplex. Such **intercalation** is observed for many antibiotics, drugs, dyes, and environmental pollutants. Among them are proflavine, ethidium bromide, actinomycin (Box 28-A), hycanthone (Fig. 5-22), and daunomycin (Fig. 5-23). Hycanthone, employed in the treatment of schistosomiasis, is one of the most widely used drugs in the world. Since intercalating agents can be mutagenic, such drugs are not without their hazards.

Intercalation is often used to estimate the amount of negative supercoiling of DNA molecules. Varying amounts of the intercalating agent are added, and the sedimentation constant or other hydrodynamic property of the DNA is observed. As increasing intercalation occurs, the secondary turns of DNA are unwound (the value of *Tw* in Eq. 5-8 decreases). Each intercalated ring causes an unwinding of the helix of ~26°. Since for a closed covalent duplex the value of *Lk* in Eq. 5-8 is constant, the decrease in *Tw* caused by increased intercalation leads to an increase in the value of *Wr*,

which is usually negative for natural DNA. When sufficient intercalation has occurred to raise *Wr* to zero, a minimum sedimentation rate is observed. Addition of further intercalating agent causes a positive supercoiling.

When the "replicative form" of DNA of the virus ϕX174 (Chapter 27), a small circular molecule containing ~5000 bp, was treated with proflavine,[221] the binding of 0.06 mol of proflavine per mole of nucleotides reduced *Wr* to zero. From this it was estimated that $\sigma = 0.055$, corresponding to −27 superhelical turns at 25°, pH 6.8, ionic strength ~0.2. Changes in temperature, pH, and ionic environment strongly influence supercoiling. In general σ becomes less negative by ~$3.3 \times 10^{-4}$ per degree of temperature increase.[222] For example, the observed value[221] of σ for ϕX174 DNA was −0.059 at 15°C and −0.040 (−20 superhelical turns) at 75°C at an ionic strength of ~0.2.

The exact ways in which intercalating substances can fit between the base pairs of nucleic acids are being revealed by X-ray diffraction studies of complexes with nucleosides, dinucleotides, and other oligonucleotides.[220,223–225] The structure of a complex in which daunomycin is intercalated between two GC base pairs in DNA is shown as in Fig. 5-23. Daunomycin and other related anthracycline antibiotics also have an amino sugar ring that binds into the minor groove of the DNA, providing both electrostatic stabilization and hydrogen bonding. Substituents on this aliphatic ring also hydrogen bond to DNA bases.[226]

Does intercalation of flat molecules into nucleic acid chains have a biochemical function? Aromatic rings of amino acid side chains in proteins designed to interact with nucleic acids may sometimes intercalate into nucleic acid helices serving a kind of "bookmark" function.[227] Changes in superhelix density caused by such intercalation may be important in the orderly handling of DNA by enzymes within cells.

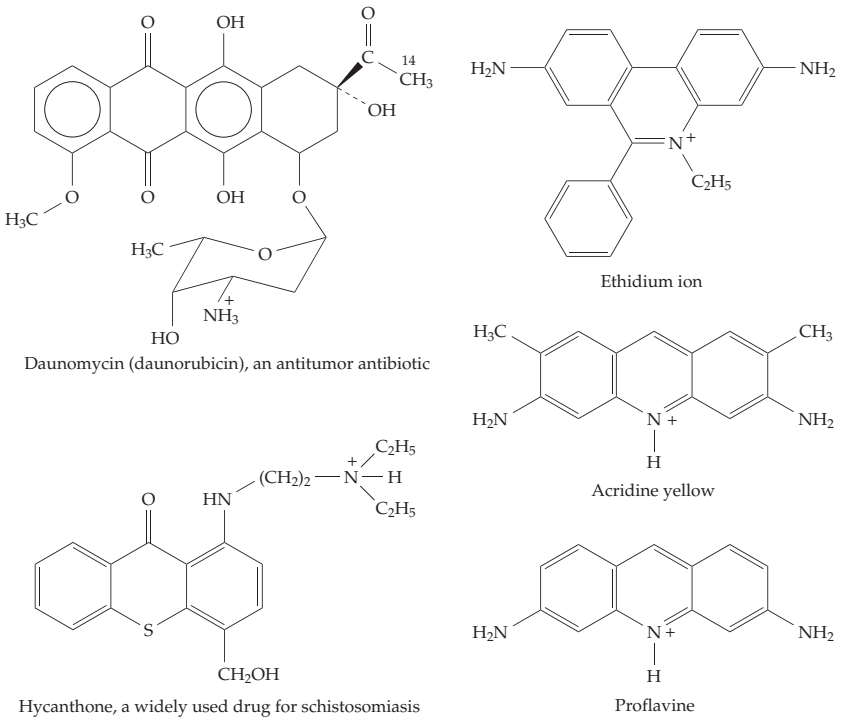



**Figure 5-22** Structures of some substances that tend to "intercalate" into DNA structures. See also Fig. 5-23.

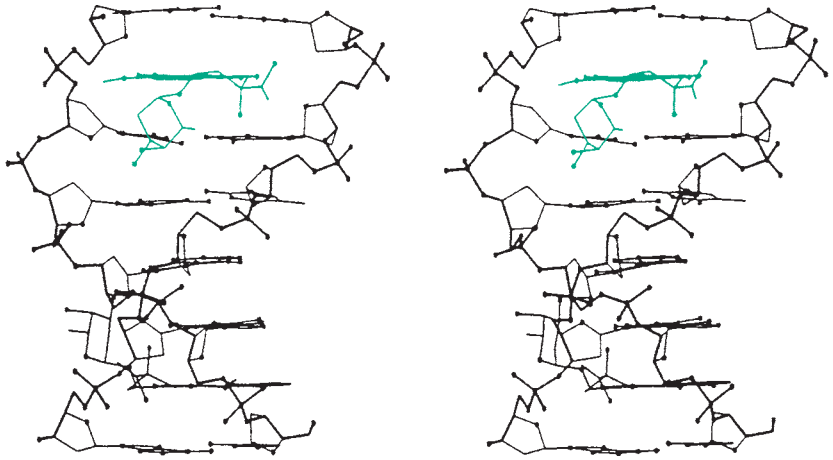


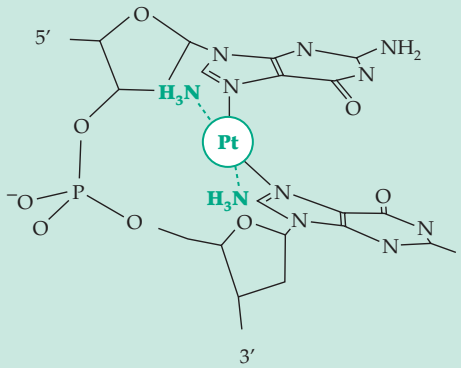
**Figure 5-23** Stereoscopic drawing showing a molecule of daunomycin (Fig. 5-22) intercalated between two base pairs in a molecule of double-helical DNA, d(CGTACG). Nitrogen and oxygen atoms are shown as dots. From Quigley *et al.*[220] Both daunomycin and adriamycin (doxorubicin; 14-hydroxydaunomycin) are important but seriously toxic anticancer drugs.

## BOX 5-B ANTITUMOR DNA DRUGS

Chemotherapy of cancer at present involves simultaneous use of two or more drugs. For example, antifolates (Chapter 15) or nucleoside analogs such as 5-fluorouridine may be used together with a drug that binds directly to DNA and inhibits the replication of cancer cells. In 1963, it was discovered accidentally that platinum ions released from supposedly inert platinum electrodes inhibited the division of *E. coli* cells. This led Rosenberg and associates to test platinum compounds against animal cells.[a] Among many compounds tested *cis*-dichlorodiammineplatinum(II) (*cis*-DDP or **cisplatin**) emerged as an important anticancer drug that is especially effective against testicular and ovarian cancers.[b–d] The *trans* isomer, however, is inactive.

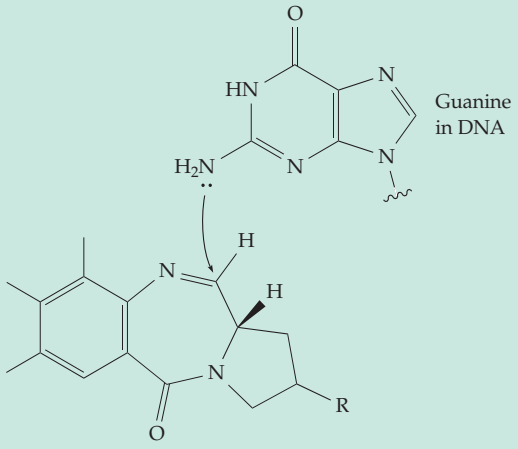

*cis*-DDP (Cisplatin) *trans*-DDP (inactive)

*cis*-DDP binds to adjacent deoxyguanosines within one strand of ds or ssDNA. The stacking of adjacent bases is disrupted as the platinum binds to N-7 nitrogen atoms of the two guanine rings by replacement of the two chloride ions. The product has the following structure,[c,e] with the Pt lying in the minor groove of dsDNA:



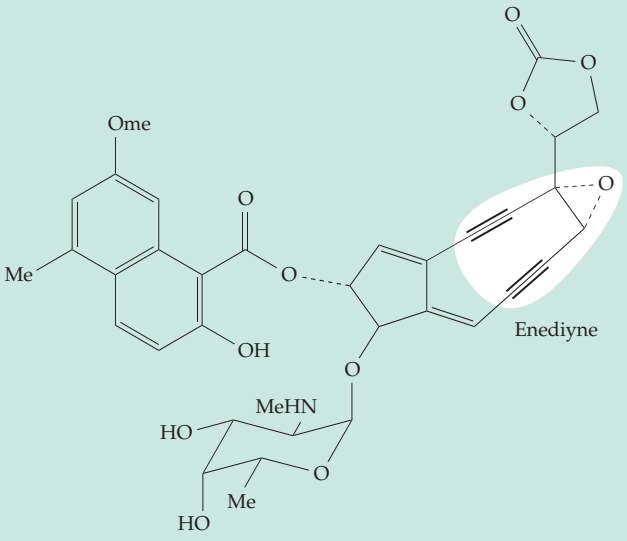

The two guanines are no longer stacked, but the structure is impossible for the *trans* isomer. The cisplatin adducts appear to prevent proper DNA repair and to induce programmed cell death (apoptosis).[f]

Since the 1950s, using a different approach, the U.S. National Cancer Institute, as well as agencies in other countries, has sought to find natural anticancer compounds in plants, fungi, microorganisms, and marine invertebrates.[g] Among these are many antibiotics that intercalate into DNA helices, e.g., **daunomycin** (Figs. 5-22, 5-23), **menogaril**,[h] **triostin A**,[i] and the antitrypanosomal drug **berenil.**[j] Some of these are also alkylating agents that contain double bonds to which such groups as the 2-$NH_2$ of guanine may add:



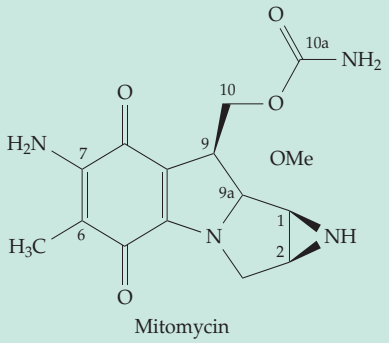

Tomaymycin and related antibiotics[k]

Diol epoxides and cyclic imines such as **mitomycin** also form adducts specifically with guanine 2-$NH_2$ groups:[l] **Neocarzinostatin** is an antitumor protein with a nonprotein "chromophore." After intercalation and binding into the minor groove of bulged DNA, it undergoes "activation" by addition of a thiol group. The enediyne structure undergoes rearrangement with formation of a reactive diradical that attacks the DNA.[m] A family of related antitumor enediynes has also been discovered.[n]

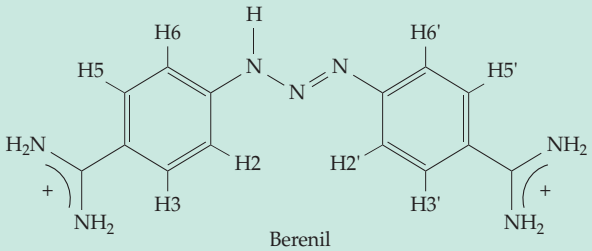



Neocarzinostatin chromophore

**BOX 5-B (continued)**



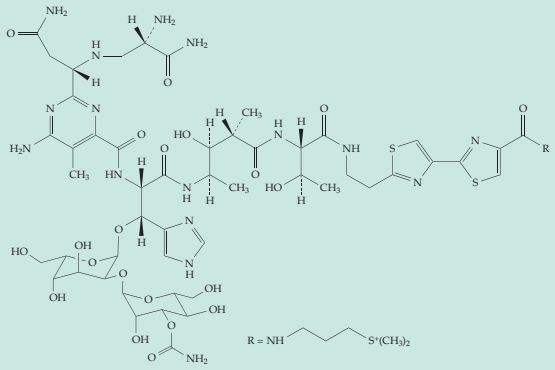

Mitomycin




Berenil

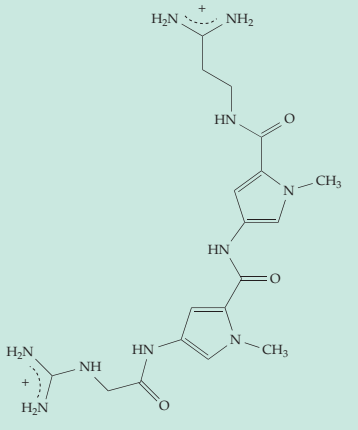
Another group of drugs occupy extended binding sites in the minor grooves of DNA double helices, often with specificity for a particular base sequence. Examples are the antitrypanosomal drug **berenil**,[j] toxic *Streptomyces* antibiotics **netropsin, distamycin**, and related synthetic compounds.[o–q] Netropsin lies within the minor groove in regions with two or more consecutive AT pairs, displacing the spine of hydration as shown in the following stereoscopic drawing. Binding depends upon both electrostatic interactions and formation of specific hydrogen bonds involving the amide groups of the antibiotics.

The antibiotic **bleomycin**, which also binds in the minor groove of B-DNA with some specificity for G-C Sites, forms an iron (II) complex. It can be

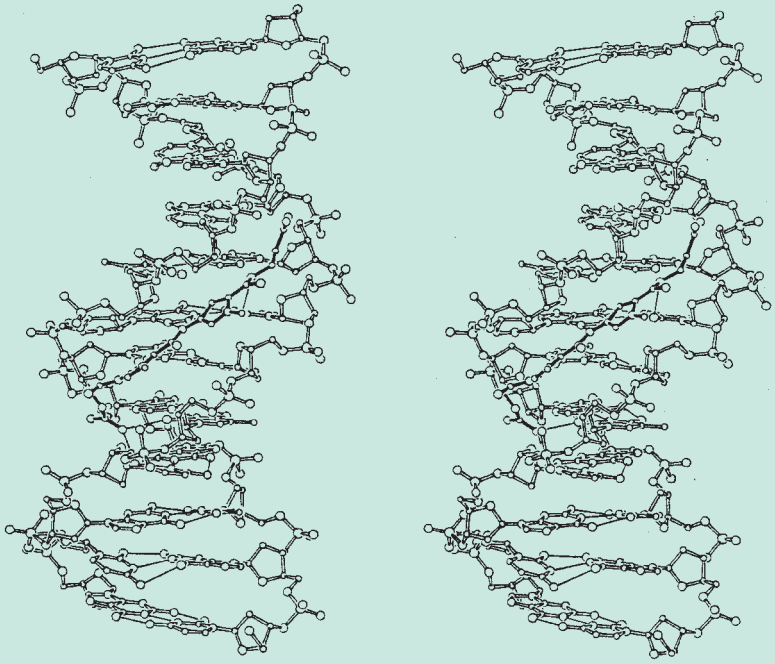


Bleomycin $A_2$

oxygenated to form an Fe(II)–$O_2$ complex (see Chapter 16) which cleaves the DNA chain.[s,t] Synthetic compounds that do the same thing have been made by connecting an EDTA–iron, or other iron chelate complex covalently to the DNA-binding compound.[n,u–x] A goal is to direct drugs to selected target sites in DNA and to induce bond cleavage at those sites in a manner analogous to that observed with restriction endonucleases. Most current chemotherapeutic agents are very toxic. Present research is designed to target these drugs more precisely to specific DNA sequences and to identify target sequences peculiar to cancers. See also **Designed third strands** in main text.




Netropsin



Netropsin lying in the minor groove of B-DNA hydrogen-bonded to bases in the central ATAT tetranucleotide. From Coll *et al.*[r]

**BOX 5-B Continued**


[a] Rosenberg, B., Van Camp, L., and Krigas, T. (1965) *Nature (London)* **205**, 698–699
[b] Zamble, D. B., and Lippard, S. J. (1995) *Trends Biochem. Sci.* **20**, 435–439
[c] Pilch, D. S., Dunham, S. U., Jamieson, E. R., Lippard, S. J., and Breslauer, K. J. (2000) *J. Mol. Biol.* **296**, 803–812
[d] Temple, M. D., McFadyen, W. D., Holmes, R. J., Denny, W. A., and Murray, V. (2000) *Biochemistry* **39**, 5593–5599
[e] Huang, H., Zhu, L., Reid, B. R., Drobny, G. P., and Hopkins, P. B. (1995) *Science* **270**, 1842–1845
[f] Zlatanova, J., Yaneva, J., and Leuba, S. H. (1998) *FASEB J.* **12**, 791–799
[g] Booth, W. (1987) *Science* **237**, 969–970
[h] Chen, H., and Patel, D. J. (1995) *J. Am. Chem. Soc.* **117**, 5901–5913
[i] Wang, A. H.-J., Ughetto, G., Quigley, G. J., Hakoshima, T., van der Marel, G. A., van Boom, J. H., and Rich, A. (1984) *Science* **225**, 1115–1121
[j] Pilch, D. S., Kirolos, M. A., Liu, X., Plum, G. E., and Breslauer, K. J. (1995) *Biochemistry* **34**, 9962–9976
[k] Barkley, M. D., Cheatham, S., Thurston, D. E., and Hurley, L. H. (1986) *Biochemistry* **25**, 3021–3031
[l] Kumar, G. S., Lipman, R., Cummings, J., and Tomasz, M. (1997) *Biochemistry* **36**, 14128 –14136
[m] Yang, C. F., Stassinopoulos, A., and Goldberg, I. H. (1995) *Biochemistry* **34**, 2267–2275
[n] Nicolaou, K. C., Dai, W.-M., Tsay, S.-C., Estevez, V. A., and Wrasidlo, W. (1992) *Science* **256**, 1172–1178
[o] Goodsell, D. S., Ng, H. L., Kopka, M. L., Lown, J. W., and Dickerson, R. E. (1995) *Biochemistry* **34**, 16654–16661
[p] Rentzeperis, D., Marky, L. A., Dwyer, T. J., Geierstanger, B. H., Pelton, J. G., and Wemmer, D. E. (1995) *Biochemistry* **34**, 2937–2945
[q] Tanious, F. A., Ding, D., Patrick, D. A., Tidwell, R. R., and Wilson, W. D. (1997) *Biochemistry* **36**, 15315–15325
[r] Coll, M., Aymami, J., van der Marel, G. A., van Boom, J. H., Rich, A., and Wang, A. H.-J. (1989) *Biochemistry* **28**, 310–320
[s] Burger, R. M., Drlica, K., and Birdsall, B. (1994) *J. Biol. Chem.* **269**, 25978–25985
[t] Kane, S. A., Hecht, S. M., Sun, J.-S., Garestier, T., and Hélène, C. (1995) *Biochemistry* **34**, 16715–16724
[u] Veal, J. M., and Rill, R. L. (1988) *Biochemistry* **27**, 1822 –1827
[v] Campisi, D., Morii, T., and Barton, J. K. (1994) *Biochemistry* **33**, 4130–4139
[w] Mack, D. P., and Dervan, P. B. (1992) *Biochemistry* **31**, 9399–9405
[x] Han, H., Schepartz, A., Pellegrini, M., and Dervan, P. B. (1994) *Biochemistry* **33**, 9831–9844


## 4. Polynucleotides with Three or Four Strands

Some nucleotide sequences in DNA favor the formation of a regular triple-helical (**triplex**) structure. This is possible when there are long stretches of adjacent pyrimidines having any sequence of C and T in one strand of double-helical DNA. The other strand of the DNA will contain the correct purines for formation of Watson–Crick base pairs. With such a structure it is always possible to add a third strand of a polypyrimidine using Hoogsteen base pairing. Triads formed in this way contain either two T's and one A or two C's and one G, as is shown in Fig. 5-7. In the latter case one of the C's must be protonated to allow formation of the pair of hydrogen bonds.[227a] These triplex structures can be formed only for stretches of DNA containing all pyrimidines in one strand (a **homopyrimidine** strand) and all purines in the other (a **homopurine** strand). A poly (AAU) triplex can also be formed, using the hydrogen bond pattern of the first triplet in Fig. 5-7. Triple-stranded synthetic polynucleotides of these types were prepared by Felsenfeld and others as early as 1957.[228–230] The third strand in a triplex can be either parallel or antiparallel to the homopurine strand. The third strand is either homopurine or a mixture of purine and thymine. The triplets are G•C*C, A*A•T, or T*A•T and are formed by Watson–Crick (•) and **reversed Hoogsteen** (*) pairing.[231] Recently, there has been a renewed interest in DNA triplets because of their occurrence in natural DNA, their possible importance in genetic recombination, and the potential for design of powerful inhibitors of replication and transcription that function via triplex formation.[35,232–238]

***H-DNA.*** Strands of DNA contain many homopyrimidine "tracts" consisting of repeated sequences of pyrimidines, e.g., $d(T\text{-}C)_n$, which may be abbreviated more simply as $(TC)_n$. In this example the complementary strand would contain the two purines in the repeated sequence $(GA)_n$. At low pH, where protonation of the cytosine rings occurs, or in negatively supercoiled DNA, the two strands of the repeating sequence may separate, with the $(TC)_n$ strand folding back to form a triple helix in which the base triplets have the hydrogen bonding pattern of Fig. 5-7. The resulting structure, which is shown in Fig. 5-24, is known as **H-DNA**.[239–241] A variety of related "nodule" and looped forms of DNA can also be formed.[35,242]

***R-DNA.*** A different type of triplex DNA may be formed during genetic recombination. A Watson–Crick duplex is brought together by one or more proteins with a single strand that is, for at least a considerable distance, an exact copy of one of the strands of the duplex. It is within such a triplex that cutting of a strand of the duplex takes place to initiate recombination (see Chapter 27 for a detailed discussion). Can a triplex structure containing two identical chains be formed? Possible base triplets include the following C•G*G triplet which occurs in a crystalline oligonucleotide structure.[243] This triplet contains a variation on Hoogsteen pairing and is related to the first triplet in Fig. 5-7. The other triplets needed for the proposed **R-DNA** recombination intermediates are G•C*C, T•A*A, and A•T*T. While there is keen interest in R-DNA[35,244–247] the formation of a stable intermediate triplex is still uncertain.[35,247]

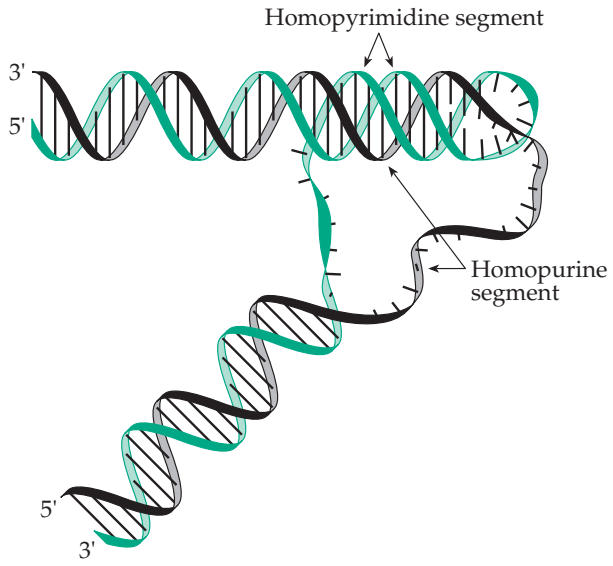
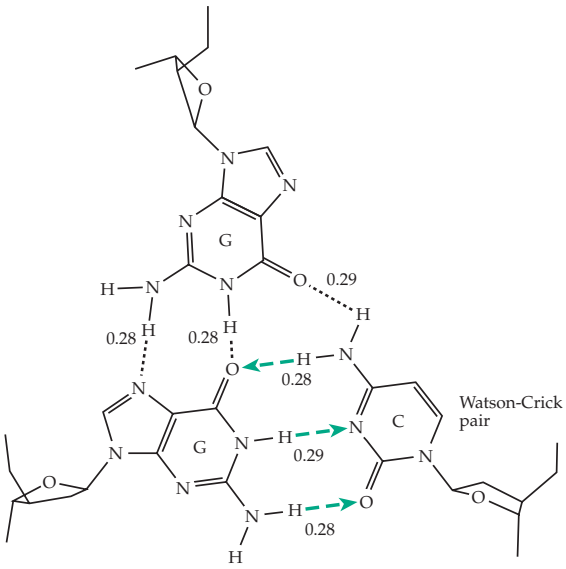
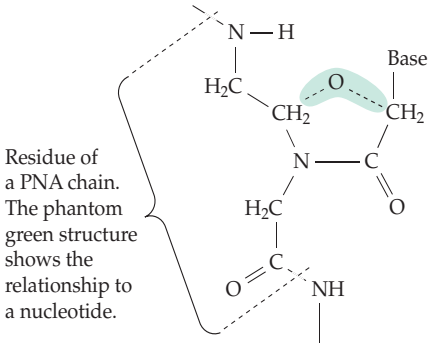



**Figure 5-24** Proposed structure of H-DNA which can be formed when a homopyrimidine segment and its complementary homopurine segment separate as a result of protonation of the cytosine rings or of negative supercoiling stress.[239,240] The triple-helical portion contains base triplets of the kind shown in Fig. 5-7.



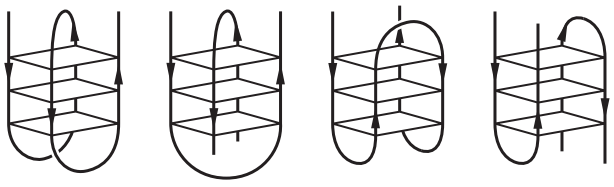

***Designed third strands.*** If triplex DNA segments can form naturally it should be possible to design oligonucleotides that will bind into the major groove of a DNA duplex to form a triplex at a specific location or locations in the genome. Dervan and associates are studying this approach systematically.[248–250] Such oligonucleotides may be modified chemically to provide stronger binding to targeted locations and may prove to be useful therapeutic agents. Among these are synthetic phosphoramidates in which the phosphodiester linkage between nucleosides is replaced by 3'-NH–$PO_3^-$–O-5'. The resulting oligonucleotide is resistant to digestion by phosphodiesterases present in cells. Its geometry favors binding to A-DNA or RNA.[251] Another oligonucleotide mimic, dubbed PNA (peptide nucleic acid), contains monomer units of the following type[252–253b] and is able to form a triple helix:




Yet another approach is to synthesize a "hairpin polyamide" that contains pyrrole, hydroxypyrrole, and imidazole groups in a sequence that favors tight binding to a specific dsDNA sequence.[254] Some possible geometries for triplex DNA are illustrated in Fig. 5-25. These are based on computer-assisted modeling.[238]

***Tetraplex (quadruplex) structures.*** The ends of linear chromosomes, the **telomeres**, have unusual nucleotide sequences repeated hundreds or thousands of times.[255–258] There is usually a **guanine-rich strand** running 5' to 3' toward the end of the chromosome and consisting of sequences such as TTAGGG in vertebrates, TTTTGGGG in the ciliate *Oxytricha*, and $TG_{1-3}$ in *Saccharomyces cerevisiae*. The complementary strand is cytosine rich. The guanine-rich strand is longer than its complement, "overhanging" by about 2 repeat units. The significance for the replication of chromosomes is discussed in Chapter 27.

The thing that has attracted most attention to telomeric DNA is the unusual structure of the G-rich strand that was signaled by the first NMR studies.[259] Subsequent investigation[260,261] revealed the presence of G quartets (Fig. 5-8) which are apparently stacked in folding patterns such as the following:[255,262–268]



These structures are stabilized by the presence of univalent cations, $K^+$ being more effective than $Na^+$,

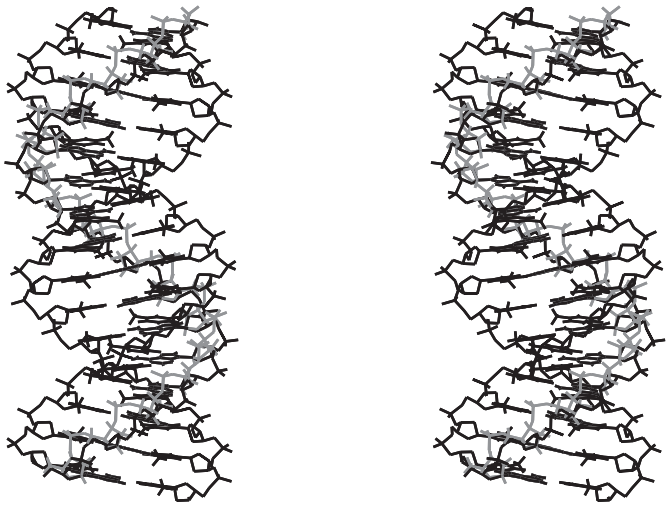
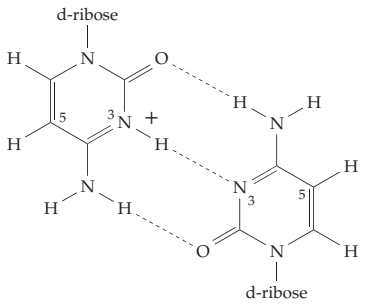


**Figure 5-25** Two superposed stereoscopic diagrams illustrating the lowest energy triple helical conformations of one pyrimidine, one purine, and one Hoogsteen DNA strand (in black). Superposed is an RNA purine strand in gray. See Srinivasan and Olson.[238] Courtesy of Wilma K. Olson.

$Li^+$, $Rb^+$, or $Cs^+$.[260,269–271] A protein isolated from *Tetrahymena* binds specifically to the G4-DNA regions.[272] The presence of 5-methylcytosine ($m^5c$) also stabilizes quadruplex structures.[273] The oligoribonucleotide $UG_4U$ forms $G_4$ quartets so stable that it takes days at 40°C in $D_2O$ for the hydrogen-bonded NH protons forming the quartet to be exchanged for $^2H$.[274]

Tetraplex structures have also been observed for G-rich repeating sequences associated with the human **fragile X syndrome**.[275,276] This is the most common cause of inherited mental retardation and appears to arise as a result of the presence of an excessive number of repeats of the trinucleotide sequence $(CGG)_n$. For normal persons n = 60 or less; for healthy carriers n may be as high as 200 but for sick individuals it may be much higher.[275] The structure in solution, as determined by NMR spectroscopy, is shown in Fig. 5-26. Another variant of four-stranded DNA, which arises from cytosine-rich DNA, contains C•$CH^+$ pairs such as the following at low pH.[277–279]




Hemiprotonated d($CH^+$•C)

In sequences such as d($TC_5$) and d($C_3T$) these C•$CH^+$ pairs are intercalated as is shown in Fig. 5-27. This intercalated DNA (**I-DNA**) may provide an alternative conformation for some telomeric sequences.[281,282] An I-DNA motif has also been identified in oligonucleotides from the DNA of human centromeres.[283] The seemingly unusual forms of DNA described in this section may represent only a fraction of the naturally occurring DNA structures of biological significance.

## 5. Junctions

Special structural features may be found at junctions between different types of DNA, e.g., between A-DNA and B-DNA.[284–286] However, the most interesting junctions are *branched*.[287–290] For example, Fig. 5-28 shows a four-way junction in which all of the bases form Watson–Crick pairs. This junction is better known as a **Holliday junction** because it was proposed by Holliday in 1964 as an intermediate in genetic recombination.[291] As shown at the top of Fig. 5-28A the junction is formed from *two homologous DNA duplexes*. These are identical except for the boxed and shaded base pairs. The ends of the first duplex are marked I and II and those of the second III and IV. The Holliday junction appears to arise by cleavage of one strand of each duplex with rejoining of the strands as indicated by the green arrows. Rotation gives the untwisted Holliday junction structure

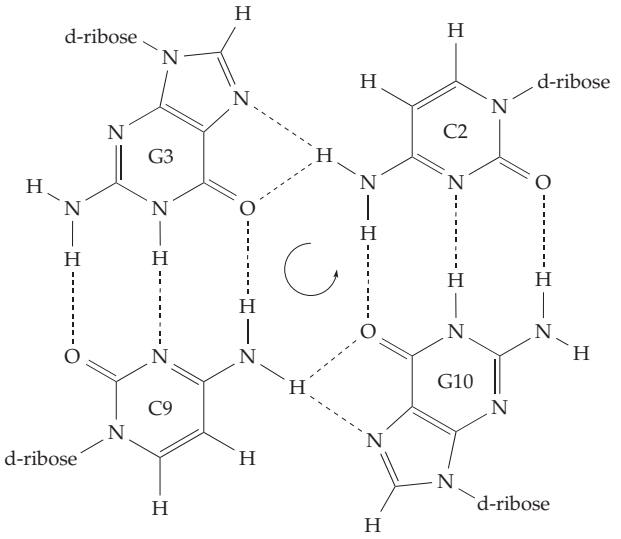



**Figure 5-26** Structure of a G•C*G•C* tetrad present in a quadruplex structure formed by the oligonucleotide d(GCGCTTTGCGC) in $Na^+$-containing solution. See Kettani *et al.*[276]

shown. The postulated three-dimensional structure of the junction is indicated in Fig. 5-28B.[292–294]

An important characteristic of Holliday junctions formed from homologous duplexes is that they can move by a process called **branch migration**.[295] Because of the twofold symmetry of the branched structure the hydrogen bonds of one base pair can be broken while those of a new base pair are formed, the branch moving as shown in Fig. 5-28. Notice that, in this example, the nonhomologous (boxed) base pairs TA and GC have become *mispaired* as TG and AC after branch migration. More significantly, the junction may be cut by a **resolvase** at the points marked

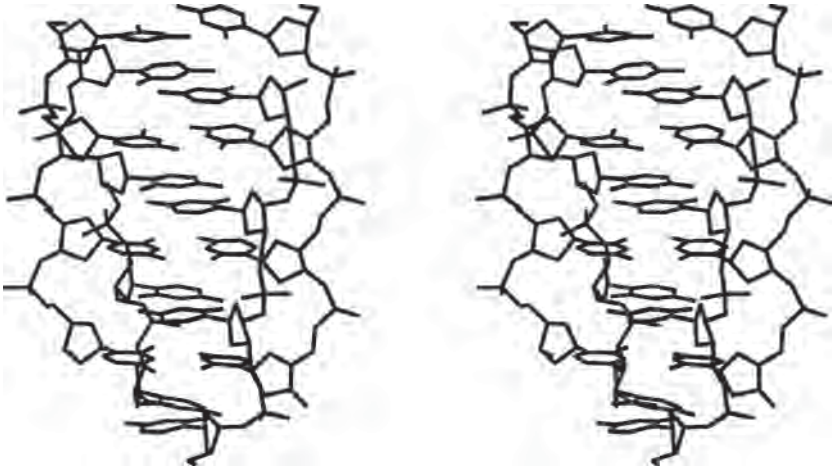
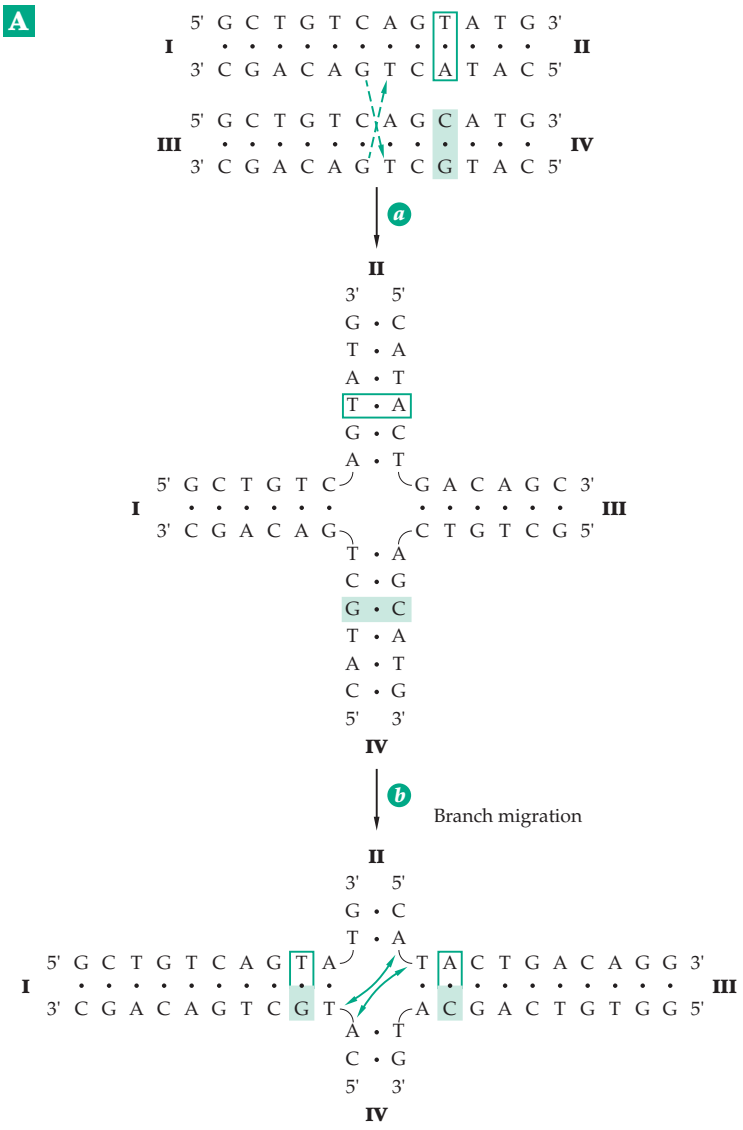


**Figure 5-27** Stereoscopic view of a four-stranded intercalated DNA or I-DNA formed from d($C_4$). Two parallel duplexes with C·$CH^+$ pairs are intercalated into each other. From Chen *et al*.[280]



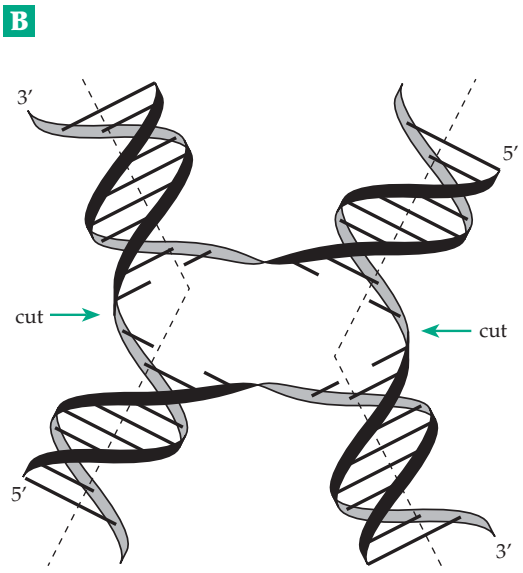




**Figure 5-28** (A) Abbreviated reaction sequence for formation of a four-way Holliday junction between two homologous DNA duplexes. In step *a* strands are cut and rejoined with movement of the strands to a roughly antiparallel orientation. The resulting structure is thought to resemble that shown below the four-stranded representation. In step *b* branch migration takes place, separating the nonhomologous base pairs TA and CG and causing mismatched pairs which will be subject to repair. (B) Proposed three-dimensional structure (after drawing by Bennett and West).[292]

by the green double-headed arrow. If the strand break is then resealed, and the DNA strands are replicated, the boxed base pair in duplex I–II will have been transferred into a strand replicated from III–IV and genetic recombination will have been accomplished. Branch migration can occur over much longer distances than are indicated in this figure, so the alteration transferred may be far from the site of the initial cleavage and whole genes or groups of genes can be transferred. Recombination is considered in more detail in Chapter 27.

The Holliday junctions formed during recombination are mobile, but synthetic **immobile Holliday junctions** can be synthesized by using nonhomologous base sequences or by locking the junctions.[288] This has permitted careful physical study of these and other more elaborate synthetic junctions. With suitable choices of base sequences for the oligonucleotides from which they are made, such junctions will assemble spontaneously. Double-stranded DNA with immobile junctions is a very suitable construction material on a "nanochemical" scale. It has been assembled into knots, rings, cubes, and more complex polyhedra.[296–297b]

## D. Ribonucleic Acids (RNA)

The best known forms of RNA are: (1) the long chains of messenger RNA (mRNA), which carry genetic messages copied from DNA to the ribosomes where proteins are made; (2) the much shorter transfer RNAs (tRNAs) which participate in reading the genetic code, correctly placing each amino acid in its sequence in the proteins; and (3) ribosomal RNAs (rRNA), which provide both structural material and a catalytic center for peptide bond formation. In addition there are numerous small RNAs that function in the splicing and editing[298–300] of mRNA, processing of tRNA precursors, methylation of ribosomal RNA,[301] transfer of proteins across membranes, and replication of DNA.[302] The genomes of many viruses consist of RNA. There are doubtless additional as yet undiscovered types of RNA.

Unlike DNA, which exists largely as double helices, the single chains of RNA can fold into complex forms containing many bulges and loops of the sort depicted in Fig. 5-9.[300,303,304] These loops are closed by double-stranded **stems** which have the A conformation. The B conformation is impossible because of the presence of the 2'-hydroxyl groups on the ribose rings in RNA. Even one ribonucleotide in a 10-nucleotide oligomer prevents formation of the B structure.[55,83] The 2'-OH groups not only keep the RNA in the A form but also engage in hydrogen-bond formation. Hydrogen bonds may form between the 2'-OH and the oxygen atom in the next ribose ring in the 3' → 5' direction. The –OH groups also hydrogen bond to water molecules which form a network within the minor groove.[305,306] These bound water molecules, in turn, can bond to associated protein and to other atoms of the complex loops found in RNA molecules. The 2'-OH groups also act as ligands for divalent metal ions in some tRNAs and in some RNA catalytic sites. Transient **hybrid DNA–RNA double helices** also exist within cells and they too usually have the overall shape of A-DNA.[55,307–309] However, the minor groove is intermediate in width between that expected for the A and B forms.

### 1. RNA Loops and Turns

Like polypeptides, polynucleotide chains have preferred ways of bending or turning. The loops at the ends of the hairpin turns of RNA molecules sometimes consist of a trinucleotide such as UUU,[304] but are usually larger. In tRNA there are typically seven bases that do not participate in regular Watson–Crick pairing (see Figs. 5-30, 5-31). The tetranucleotide **UUCG** is frequently present, and the sequence 5'-GGAC**UUCG**GUCC forms an unusually stable hairpin.[85,310] Other tetranucleotides, such as UGAA,[311] CCCG (also found in DNA loops),[312] GCAA,[313] and GAAA,[314] occur often. The latter are members of a larger group of loop structures with the consensus sequence GNRA, where N is any nucleotide and R is a purine.[315,315a] These sequences are very common in highly folded structures of ribosomal RNAs. Until recently high-resolution X-ray structures were available for only a few tRNAs and ribozymes. To help remedy this deficiency the structures of a great variety of oligonucleotide stem/loop (hairpin) structures are being determined, most by NMR spectroscopy.[304,316,317] The sharp turns in the loops involve mostly rotation about the two torsion angles around the phosphorus atom of the third, from the 5' end, of the seven nucleotides. Base bulges on stems (Fig. 5-9) not only introduce kinks and bends in RNA stems[169] but also provide well-defined hydrogen-bonded binding sites for proteins. Numerous branched three-way and more complex junctions provide other important motifs in folded RNA.[318]

Among the new RNA structures are those of RNA–antisense RNA pairs in "kissing" hairpin complexes.[87,319] Another interesting complex folding pattern in RNA is the **pseudoknot**, a structural feature that has been identified in many RNA sequences.[85,320–327] A pseudoknot can be formed if nucleotide sequences favorable to formation of two short RNA stems are overlapped as shown in Fig. 5-29. After stem 2 in this drawing is formed (step *a*) additional base pairing can lead to formation of stem 1 (step *b*). The base pairs of the two stems can stack coaxially to form the pseudoknot (step *c*).

For the formation of base-paired stems the RNA must contain antiparallel sequences that allow Watson

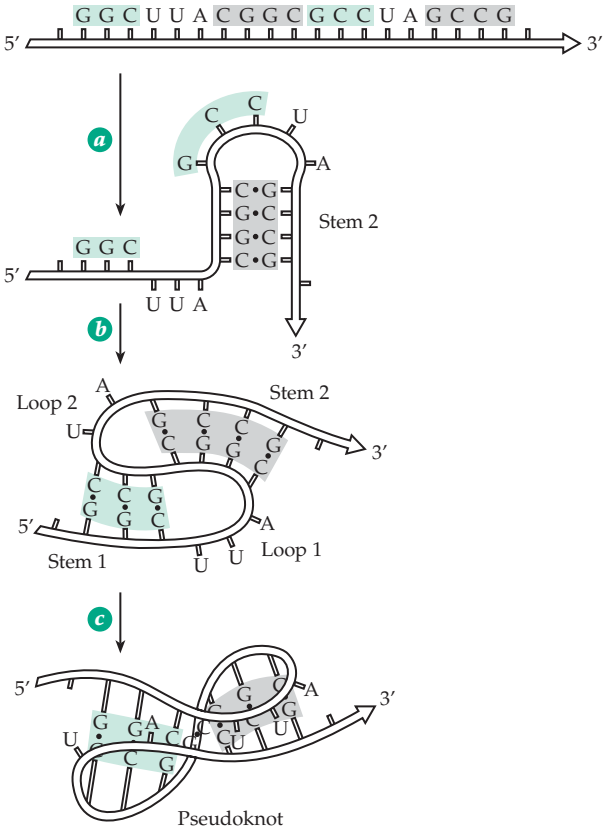


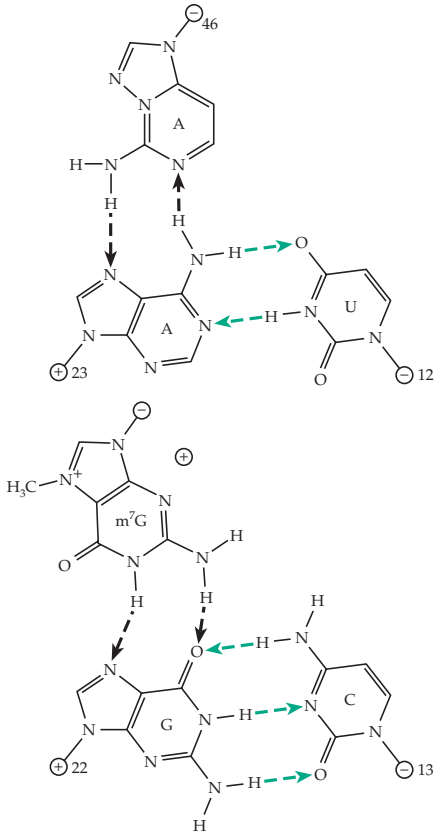

**Figure 5-29** Formation of a pseudoknot in an RNA chain. After Puglisi *et al.*[325]

–Crick or wobble (GU) base pairing in the stems. This requires certain relationships in the sequences of the DNA in the genes that encode these molecules as discussed in Section E,3. Because of the base-pairing requirements, some of the bases in the stems protrude as bulges and fail to form pairs. Often, it is possible to find more than one reasonable structure, each having some bases unpaired.[328]

## 2. Transfer RNAs

In all tRNAs the bases can be paired to form "clover-leaf" structures with three hairpin loops and sometimes a fourth as is indicated in Fig. 5-30.[329–331] This structure can be folded into the L-shape shown in Fig. 5-31. The structure of a phenylalanine-carrying tRNA of yeast, the first tRNA whose structure was determined to atomic resolution by X-ray diffraction, is shown.[170,332–334] An aspartic acid-specific tRNA from yeast,[335] and an *E. coli* chain-initiating tRNA, which places *N*-formyl-methionine into the N-terminal position of proteins,[336,337] have similar structures. These molecules are irregular bodies as complex in conformation as globular proteins. Numerous NMR studies show that the basic structure is conserved in all tRNAs. However, animal mitochondrial tRNAs often lack some of the usual stem-loop "arms" as well as the invariant nucleotides in the dihydrouridine and TψC loops (Fig. 5-30).[307,338,339] At the bottom of the structure as shown in Figs. 5-30 and 5-31 is the **anticodon**, a triplet of bases having the correct structures to permit pairing with the three bases of the codon specifying a particular amino acid (see Table 5-5), in this case phenylalanine.

While tRNAs consist largely of loops and stems containing Watson–Crick base pairs, they also contain Hoogsteen pairs, wobble pairs, and triplets such as the following.




The first of these contains a Watson–Crick A•U pair with a second A bound to form a reversed Hoogsteen A∗A pair. The second contains a G•C Watson–Crick pair with a Hoogsteen N-7-guanine∗G ($m^7G$∗G) pair. Among the complex base associations present in tRNAs are some that also involve hydrogen bonding to the 2'-hydroxyl groups of ribose rings and to at least one of the phosphate groups. There are over 100 internal hydrogen bonds, a large proportion of which are relatively invariant among the known tRNAs.

## 3. Properties and Structures of Some Other RNA Molecules

A few specialized RNA molecules are listed in

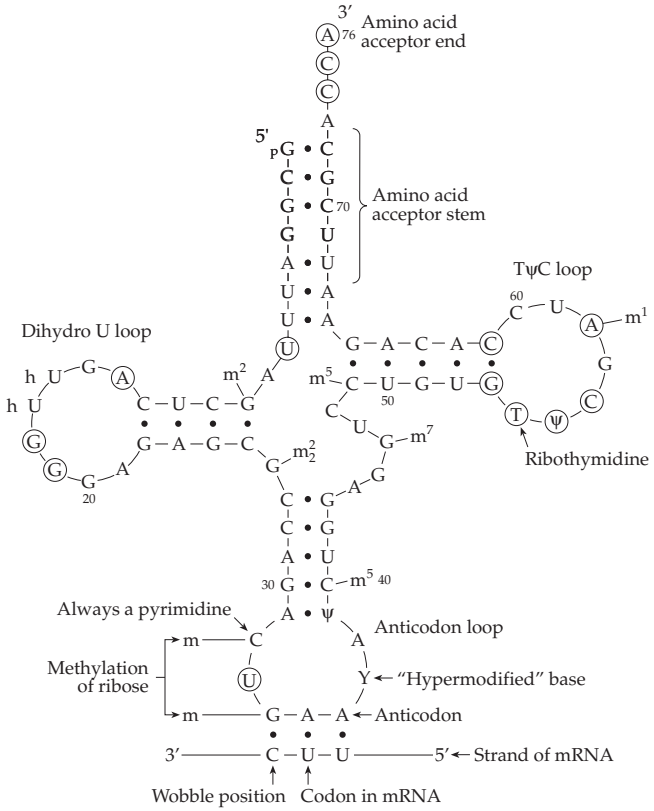



**Figure 5-30** Schematic cloverleaf structure of a phenylalanine-specific transfer RNA ($tRNA^{Phe}$) of yeast. The dots represent pairs or triplets of hydrogen bonds. Nucleosides common to almost all tRNA molecules are circled. Other features common to most tRNA molecules are also marked. The manner in which the anticodon may be matched to a codon of mRNA is indicated at the bottom.

Table 5-4. There are many others.[302] In addition, there are thousands of different mRNAs within a cell. The most abundant RNA molecules in cells are those of the ribosomes. Ribosomes consist of two elaborate RNA-protein subunits, a large subunit with sedimentation constant ~ 30S in bacteria and ~ 40S in eukaryotes. The small subunit contains 16S or 18S RNA and the large subunit 23S or 28S as well as smaller 5S and 5.8S RNAs (Table 5-4; Table 28-1). A proposed three-dimensional structure[341,342] of a bacterial 16S ribosomal RNA and the corresponding ribosomal subunit with its 21 proteins are shown in Fig. 5-32. It might seem impossible that the folding pattern of the RNA was deduced correctly before an X-ray structure was available. However, a **phylogenetic approach**, the comparison of nucleotide sequences among several species, suggested that the stem structures of rRNAs are highly conserved (see discussion in Chapter 29). This fact, together with a variety of other chemical

## TABLE 5-4
## Some Specialized RNA Molecules

| Kind of RNA | Number of nucleotides |
|---|---|
| Transfer RNAs (Figs. 5-30, 5-31) | 60 – 85 |
| Ribosomal RNAs | |
| 5S | ~ 120 |
| 5.8S (rat) | 158 |
| 16S (*E. coli*; Fig. 5-32) | 1542 |
| 18S (rat) | 1874 |
| 23S (*E. coli*) | 2904 |
| 28S (rat) | 4718 |
| Telomerase guide RNA[a-d] | 159 (*Tetrahymena*) |
| M1 RNA of Ribonuclease P[e,f] | 350 – 410 |
| | 377 (*E. coli*) |
| *Tetrahymena* Intron ribozyme (Fig. 12-26)[g] | 413 |
| Viroid hammerhead ribozyme (Fig. 12-27)[h] | ~ 55 |
| Signal recognition particle 7S RNA[i] | 295 (human) |
| Small nuclear RNAs[j,k] | 65 – 1200 |
| RNA-editing guide RNA[l] | ~ 60 |
| Thermotolerance factor (G8 RNA)[m] | ~ 300 |
| Viroid RNA (Fig. 28-19) | 240 – 380 |
| Virus MS2 genome (Chapter 29) | 3569 |


[a] Greider, C. W., and Blackburn, E. H. (1989) *Nature (London)* **337**, 331–337
[b] Bhattacharyya, A., and Blackburn, E. H. (1994) *EMBO J.* **13**, 5721–5731
[c] Singer, M. S., and Gottschling, D. E. (1994) *Science* **266**, 404–409
[d] Feng, J., Funk, W. D., Wang, S.-S., Weinrich, S. L., Avilion, A. A., Chiu, C.-P., Adams, R. R., Chang, E., Allsopp, R. C., Yu, J., Le, S., West, M. D., Harley, C. B., Andrews, W. H., Greider, C. W., and Villeponteau, B. (1995) *Science* **269**, 1236–1241
[e] Stark, B. C., Kole, R., Bowman, E. J., and Altman, S. (1978) *Proc. Natl. Acad. Sci. U.S.A.* **75**, 3717–3721
[f] Mattsson, J. G., Svärd, S. G., and Kirsebom, L. A. (1994) *J. Mol. Biol.* **241**, 1–6
[g] Cech, T. R. (1987) *Science* **236**, 1532–1539
[h] Hertel, K. J., Herschlag, D., and Uhlenbeck, O. C. (1994) *Biochemistry* **33**, 3374–3385
[i] Li, W.-Y., Reddy, R., Henning, D., Epstein, P., and Busch, H. (1982) *J. Biol. Chem.* **257**, 5136–5142
[j] Maxwell, E. S., and Fournier, M. J. (1995) *Ann. Rev. Biochem.* **35**, 897–934
[k] Nicoloso, M., Qu, L.-H., Michot, B., and Bachellerie, J.-P. (1996) *J. Mol. Biol.* **260**, 178–195
[l] Kable, M. L., Seiwert, S. D., Heidmann, S., and Stuart, K. (1996) *Science* **273**, 1189–1195
[m] Fung, P. A., Gaertig, J., Gorovsky, M. A., and Hallberg, R. L. (1995) *Science* **268**, 1036–1039

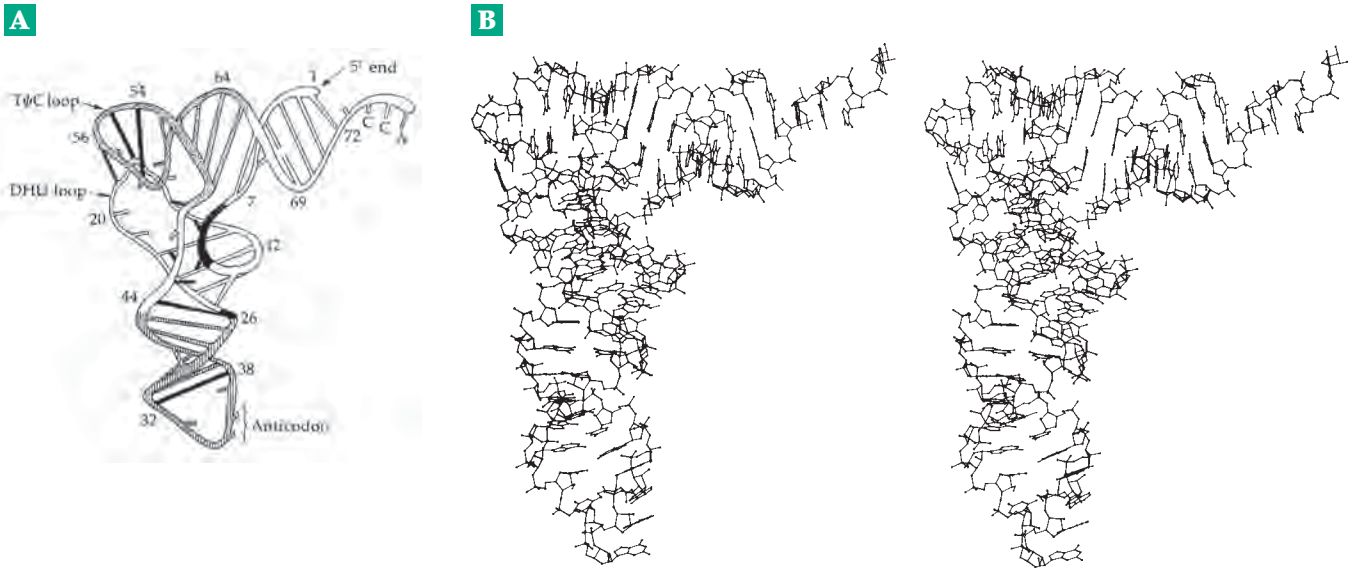



**Figure 5-31** The three-dimensional structure of a phenylalanine-specific transfer RNA of yeast. (A) Perspective diagram of folding of polynucleotide chain. The ribose phosphate backbone is drawn as a continuous cylinder with bars to indicate hydrogen-bonded base pairs. The positions of single bases are indicated by rods which are intentionally shortened. The TψC arm is heavily stippled, and the anticodon arm is marked by vertical lines. Tertiary structure interactions are illustrated by black rods. Redrawn from Quigley and Rich.[340] (B) Stereoscopic view of the structure of yeast $tRNA^{Phe}$ as revealed by X-ray crystallography. The acceptor stem with the protruding ACCA sequence at the 3' end is to the right. The anticodon GAA is at the bottom right side of the drawing. The guanine ring is clearly visible at the very bottom. The middle adenine of the anticodon is seen exactly edge-on as is the "hypermodified" base Y (see Fig. 5-33) which lies just above the anticodon. Its side chain is visible at the back of the drawing. Preceding the anticodon on the left (5' side) are two unpaired bases (C and U). The 2'-hydroxyl groups of the cytosine and of the guanosine in the anticodon are methylated. Moving far up the anticodon stem one can see two groups of base triplets which utilize both Watson–Crick and Hoogsteen types of base pairing. Drawing courtesy of Alexander Rich.

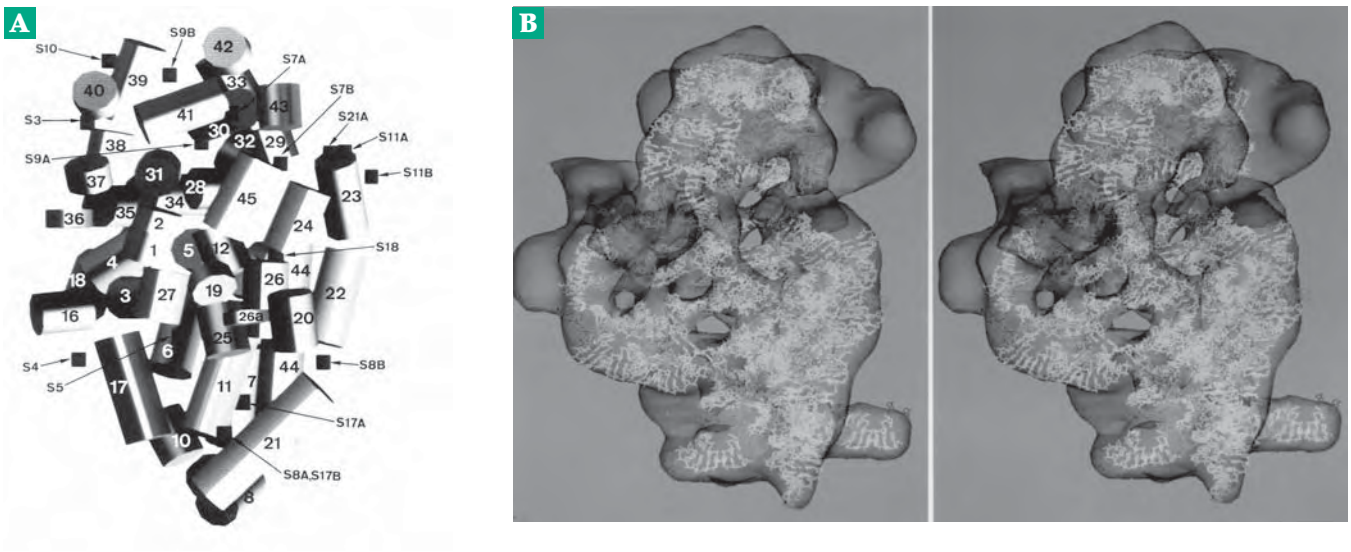



**Figure 5-32** (A) A three-dimensional computer graphics model proposed by Brimacombe *et al.*[341] for the single chain of *E. coli* 16S ribosomal RNA. The helices are depicted as cylinders, which are all connected. The small dark squares denote the positions of artificially formed RNA–protein crosslinks, marked with the appropriate protein number. For proteins exhibiting more than one crosslink site (e.g., S17), the sites are denoted A or B, in each case A being the site nearer to the 5' terminus of the 16S RNA. (B) Stereoscopic view of tentative atomic model of 16S RNA in the 30S ribosomal subunit. The viewing direction is different from that in (A). From Mueller and Brimacombe.[342] Courtesy of Richard Brimacombe.

evidence,[341,343,344] allowed the prediction of the structure shown as well as many characteristics of the rRNA of the large subunit.[345] Culminating decades of effort[346–350] (Chapter 29), complete structures of bacterial ribosomes were established by 2000[342,351–354] and the peptidyl transferase center was identified as a ribozyme[355–357] (discussed in Chapter 12).

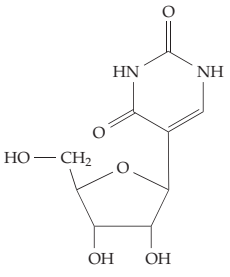
## 4. Modified Nucleotides

The picture of DNA or RNA as chains of only four kinds of nucleotides is not quite accurate. DNA contains a significant number of methylated bases among which are **5-methylcytosine** (5mC or $m^5C$) and 6-methyladenine (6mA). The former is regularly present in the nuclear DNA of higher animals and plants. In mammals from two to seven percent of the cytosine is methylated. It is likely that these methylated bases mark special points in the genetic blueprint. As is discussed in Chapter 27, methylation, which is accomplished after the synthesis of the polynucleotide, may block the expression of certain genes.[358] This appears to happen when one of the two X-chromosomes becomes inactivated in cells of females. Methylation may also be used to **imprint** certain genes, i.e. to mark them as coming from a specific one of the two parents.[359]

Another function of methylation is to protect DNA from attack by enzymes (restriction endonucleases) formed in response to invading viruses (Chapter 26). Some viruses, notably the bacteriophage of the T-even series that attack *E. coli* (Box 7-C), have developed their own protective devices. They contain **5-hydroxymethylcytosine** (HOMeC) in place of cytosine. The extra hydroxyl groups provided in this fashion often carry one or two glucose units in glycosidic linkage.[360] A bacteriophage attacking *Bacillus subtilis* substitutes hydroxymethyluracil for uracil and 5-dihydroxypentyluracil for thymine, and phage W14 of *Pseudomonas acidovorans* substitutes the 5-methyl of thymine with $-CH_2-NH-(CH_2)_4-NH_3^+$.[360,361]

The modifications carried out on RNA molecules are more varied and more extensive than those of DNA. Sixty or more modification reactions are known for tRNA, with the number and the extent of modification depending upon the species. Structures of some of the modified bases are indicated in Fig. 5-33. Uridine can be methylated either on the base or on the 2'-hydroxyl of the sugar. Methylation at the 5 position of uridine yields **ribothymidine**. Cytidine can be modified in the same positions. Reduction of the 5,6 double bond of uridine gives **dihydrouridine** (hU). Replacement of the oxygen at position 4 by sulfur gives **4-thiouridine** ($^4$sU). Positions in the guanosine structure that can be methylated are also indicated in Fig. 5-33. The symbol m is commonly used to designate methylation in nucleic acid bases; $m_2$ indicates dimethylation, e.g., 6,6-dimethyladenine is abbreviated $m_2^6$A.

A remarkable transformation is that of uridine into pseudouridine (ψ).



Pseudouridine (ψ)

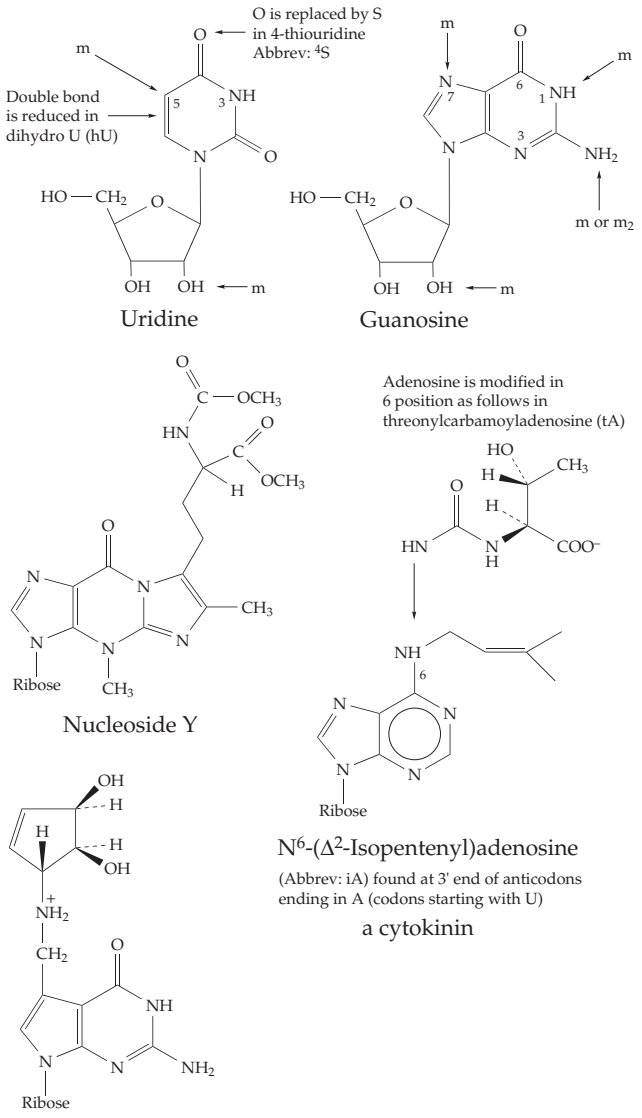



**Figure 5-33** Structures of some nucleosides containing modified bases and found in tRNA molecules. Positions where methylation may occur are designated m. Some other abbreviations are e (ethyl), ac (acetyl), and f (formyl).[25]

## BOX 5-C THE RNA WORLD

The discovery in the 1980s that RNA molecules often have catalytic properties and may serve as true enzymes (ribozymes; Chapter 12) stimulated new thinking about evolution. Although RNA catalysts are not as fast as the best enzymes they are able to catalyze a wide variety of different reactions. Could it be that in the early evolution of organisms RNA provided both the genetic material and catalysts? The "RNA world" would have been independent of both DNA and protein.[a,b] Later DNA could have been developed as a more stable coding molecule and proteins could have evolved as more efficient catalysts. Plausible reactions by which both cytosine and uracil could have arisen in drying ponds on early Earth have been demonstrated.[c]

A major objection to the RNA world is the lack of stability of ribose and the inability to demonstrate the nonenzymatic synthesis of ribose in significant amounts. Even if ribose were present, it would be largely in the pyranose ring forms. Initial formation of the 5-phosphate would be required to allow formation of a nucleotide with a furanose ring. These and other obstacles to the RNA world have led to the suggestion that some *other* genetic material preceded RNA and DNA.[d,e] One possibility is a peptide-like RNA analog.[e] A simple coding system could also have been used, e.g. one based on only two bases, such as C and G, instead of four.[f,g]

Perhaps it is more probable that formation of proteins *and* the present coding system evolved simultaneously? The major metabolic cycles (Chapter 10) could also have developed at the same time. RNAs could not have been the *first catalysts*. Hydrogen ions, hydroxyl ions, ammonium, cyanide, and other simple ions as well as amines and peptides could all have played a role in prebiotic chemistry. Another speculation suggests an "iron-sulfur world" in which organic materials would be formed on mineral surfaces through reactions involving reduction of bicarbonate by iron sulfide and $H_2S$.[h]

If the RNA world did exist, has it left us with any real clues? Benner *et al.* suggest that modern metabolism is a palimpsest of the RNA world, a parchment that has been inscribed two or more times, with previous texts imperfectly erased and therefore still partially legible.[i] If so, can we find a way to read the text of the ancient RNAs?

[a] Gesteland, R. F., Cech, T. R., and Atkins, J. F., eds. (1999) *The RNA World*, 2nd ed. Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York
[b] Orgel, L. E. (1994) *Sci. Am.* **271** (Oct), 77 – 83
[c] Robertson, M. P., and Miller, S. L. (1995) *Nature (London)* **375**, 772 – 774
[d] Larralde, R., Robertson, M. P., and Miller, S. L. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8158 – 8160
[e] Böhler, C., Nielsen, P. E., and Orgel, L. E. (1995) *Nature (London)* **376**, 578 – 581
[f] Sievers, D., and von Kiedrowski, G. (1994) *Nature (London)* **369**, 221 – 224
[g] Piccirilli, J. A. (1995) *Nature (London)* **376**, 548 – 549
[h] Maden, E. H. (1995) *Trends Biochem. Sci.* **20**, 337 – 341
[i] Benner, S. A., Ellington, A. D., and Tauer, A. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 7054 – 7058

Pseudouridine is formed by enzymatic rearrangement of uridine in the original transcript (Eq. 28-3). It can form a base pair with adenine in the same manner as does uracil. Pseudouridine is found not only in tRNA but also in several places in both large and small ribosomal RNA subunits. For example, it is present at position 516 in the *E. coli* 16S RNA,[364] at a specific position in the 23S RNA, and at many more locations in eukaryotic rRNA.

The bases called "Y" and "Q" are highly modified guanines (Fig. 5-33). Q is found at the 5' end of some anticodons in the "wobble position" (see Fig. 5-30). Two **hypermodified** adenosines are also shown in Fig. 5-33. The $N^6$-isopentenyladenosine is found at the 3' end of the anticodons that pair with codons starting with U. This compound is also a plant hormone, a **cytokinin** (Chapter 30). Another highly modified purine, threonylcarbamoyladenine, occurs adjacent to the end of anticodons pairing with codons starting with A. The function of these hypermodified bases is uncertain, but they appear to promote proper binding to ribosomes. The modifications are often not absolutely essential for function.

Another source of modified bases in both DNA and RNA is spontaneous or "accidental" alteration. Nucleic acids encounter many highly reactive and mutagenic materials including hydroxyl radicals, formed from $O_2$, and are able to convert guanine rings into 7,8-dihydro-8-oxoguanine.[362] Other reactive and carcinogenic compounds can form adducts with nucleic acid bases.[363] See Eq. 5-18 and also Chapter 27.

### 5. RNA Aptamers

Ellington and Szostak[365] synthesized a random "pool" of ~$10^{15}$ different oligodeoxyribonucleotides, each ~100 nucleotides in length. They "amplified" these using the polymerase chain reaction (PCR; Section H,6) and prepared a mixture of the corresponding RNAs by *in vitro* transcription. From the ~$10^{13}$ different sequences still present they selected individual

oligonucleotides by affinity chromatography on columns that contained well-defined immobilized ligands such as organic dyes. They called the selected RNAs **aptamers**. Their approach is being used to find RNA sequences that bind to such ligands as ATP, FMN,[366] the bronchodilator theophylline,[367] aminoglycoside antibiotics,[368] arginine,[369] etc.[370] Many of the selected aptamers bind their ligands very tightly and studying them may shed light on interactions of RNA with proteins and on the catalytic activities of RNA, which are discussed in Chapter 12.

## E. The Genetic Code

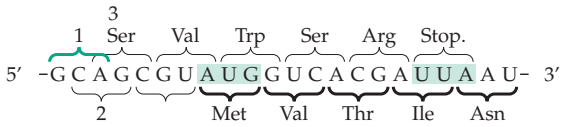
The general nature of the genetic code was suggested by the structure of DNA. Both DNA and proteins are linear polymers. Thus, it was logical to suppose that the sequence of the bases in DNA codes for the sequence of amino acids. There are only four bases in DNA but 20 different amino acids in proteins at the time of their synthesis. It is obvious that each amino acid must be specified by some combination of more than one base. While 16 pairs of bases are possible, this is still too few to specify 20 different amino acids. Therefore, it appeared that at least a triplet group of three nucleotides would be required to code for one amino acid.[371] Sixty-four ($4^3$) such triplet **codons** exist, as is indicated in Tables 5-5 and 5-6.

Simplicity argues that the genetic blueprint specifying amino acid sequences in proteins should consist of consecutive, nonoverlapping triplets. This assumption turned out to be correct, as is illustrated by the DNA sequence for a gene shown in Fig. 5-5. In addition to the codons that determine the sequence of amino acids in the protein, there are **stop codons** that tell the ribosomal machinery when to terminate the polypeptide chain. One methionine codon serves as an **initiation codon** that marks the beginning of a polypeptide sequence. One of the valine codons sometimes functions in the same way.

How does a cell read the code? This question is dealt with in detail in Chapters 28 and 29. A key step is the positioning of each amino acid on the ribosome in proper sequence. This is accomplished by the pairing of codons of messenger RNA with the anticodons of the appropriate transfer RNA molecules as is indicated at the bottom of Fig. 5-30. Each tRNA carries the appropriate "activated" amino acid at its 3' end ready to be inserted into the growing peptide.

### 1. The "Reading Frames"

It is immediately obvious that there are three ways of reading the genetic code in mRNA depending upon which nucleotide is used to start each codon. For example, in the following mRNA sequence either codons GCA, CAG, or AGC could be selected as first.




These codons define the three reading frames or phases in which the code may be read. Here, the term *frame*

**TABLE 5-5**
**The Genetic Code[a]**

| Amino acid | Codons | Total number of codons |
|---|---|---|
| Alanine | GCX | 4 |
| Arginine | CGX, AGA, AGG | 6 |
| Asparagine | AAU, AAC | 2 |
| Aspartic acid | GAU, GAC | 2 |
| Cysteine | UGU, UGC | 2 |
| Glutamic acid | GAA, GAG | 2 |
| Glutamine | CAA, CAG | 2 |
| Glycine | GGX | 4 |
| Histidine | CAU, CAG | 2 |
| Isoleucine | AAU, AUC, AUA | 3 |
| Leucine | UUA, UUG, CUX | 6 |
| Lysine | AAA, AUG | 2 |
| Methionine (also initiation codon) | AUG | 1 |
| Phenylalanine | UUU, UUC | 2 |
| Proline | CCX | 4 |
| Serine | UCX, AGU, AGC | 6 |
| Threonine | ACX | 4 |
| Tryptophan | UGG | 1 |
| Tyrosine | UAU, UAC | 2 |
| Valine (GUG is sometimes an initiation codon) | GUX | 4 |
| Termination | UAA (*ochre*) UAG (*amber*) UGA | 3 |
| Total | | 64 |

[a] The codons for each amino acid are given in terms of the sequence of bases in messenger RNA. From left to right, the sequence is from the 5' end to the 3' end. The symbol X stands for any one of the four RNA bases. Thus, each codon symbol containing X represents a group of four codons.

does not designate a single codon, although frame does designate a single exposure in a motion picture film. **Reading frame** designates which of the three possible sets of codons we are using. In the foregoing sequence the codons in reading frames 2 and 3 are labeled. Reading frame 2 contains the initiation codon AUG (shaded) which could mark the beginning of an encoded protein sequence. Reading frame 3 contains a termination (stop) codon which, when the mRNA transcript is read by ribosomes, will terminate polypeptide synthesis. It may be in the position shown but not have any real function. However, it could represent the end of the coding sequence that is marked if genes for the two proteins overlap a little at the ends, a situation that actually occurs in nature.

A reading frame in a specified part of a DNA sequence is said to be **"open"** if there is an initiation codon preceded by suitable regulatory signals (an **operator** region). This means that it *could* encode a protein. The reading frame of a sequence is open until the next termination codon. Recently another usage has appeared. Many writers refer to an **open reading frame** as a segment of DNA in which any one of the three reading frames is open. Another complexity in the reading of genetic messages arises because splicing of RNA may sometimes cause a shift in the reading frame. For example, a mRNA being transcribed from the sequence in reading frame 2 in the foregoing example may skip over a nucleotide part of the time to form an RNA in which the first part is encoded by reading frame 2 and the second by reading frame 3 (a + 1 frameshift). Alternatively, a nucleotide could be read twice with a – 1 frameshift with the sequence of reading frame 1 for the latter part of the mRNA. Frameshifts can also occur during protein synthesis as the mRNA is being read.

In the present example we have examined the sequence in mRNA. In the DNA there are two strands. One is the **coding strand** (also called the nontranscribing or nontranscribed strand), which has a sequence that corresponds to that in the mRNA and the one that is given in Fig. 5-4. The second antiparallel and complementary strand can be called the **template strand** or the noncoding, transcribing, or transcribed strand.[372] The mRNA that is formed is sometimes referred to as a **sense strand**. The complementary mRNA, which corresponds in sequence to the noncoding strand of DNA, is usually called **antisense RNA**.

**TABLE 5-6**
**The Sixty-Four Codons of the Genetic Code**

| 5′–OH Terminal base | Middle base | | | | 3′–OH Terminal base |
|---|---|---|---|---|---|
| | U(T) | C | A | G | |
| U(T) | Phe | Ser | Tyr | Cys | U(T) |
| | Phe | Ser | Tyr | Cys | C |
| | Leu | Ser | Term[c] | Term[d] | A |
| | Leu | Ser | Term | Trp | G |
| C | Leu | Pro | His | Arg | U(T) |
| | Leu | Pro | His | Arg | C |
| | Leu[a] | Pro | Gln | Arg | A |
| | Leu | Pro | Gln | Arg | G |
| A | Ile | Thr | Asn | Ser | U(T) |
| | Ile | Thr | Asn | Ser | C |
| | Ile | Thr | Lys | Arg | A |
| | Met[b] | Thr | Lys | Arg | G |
| G | Val | Ala | Asp | Gly | U(T) |
| | Val | Ala | Asp | Gly | C |
| | Val | Ala | Glu | Gly | A |
| | Val[b] | Ala | Glu | Gly | G |

[a] The codon CUA (CTA) encodes threonine and the codon AUA (ATA) methionine in mammalian mitochondria.
[b] Initiation codons. The methionine codon AUG is the most common starting point for translation of a genetic message but GUG can also serve. In such cases it codes for methionine rather than valine.
[c] The "termination codon" UAA (TAA) encodes glutamine in *Tetrahymena*.
[d] The termination codon UGA (TGA) encodes tryptophan in mitochondria and selenocysteine in some contexts in nuclear genes.

## 2. Variations

Is the genetic code "universal" or does it vary from one organism to another? Studies with bacteria, viruses, and higher organisms including humans have convinced us that the code is basically the same for all oganisms. However, there are some variations. For example, in mitochondria of both humans and yeast the codon TGA is not a termination codon but represents tryptophan. In mammalian mitochondria CTA represents threonine rather than leucine, and ATA encodes methionine instead of isoleucine. These differences in the code are related to the fact that mitochondria contain their own piece of DNA. It encodes not only several proteins but also tRNA molecules whose anticodon structures are altered to accommodate the changed meanings of the codons of the mitochondrial DNA and mRNA.[373,374]

Variations in the code for cytoplasmic proteins have been found. In *Tetrahymena* and other ciliates the codon TAA represents glutamine rather than being a termination codon.[375] A few proteins, including some in the human body, contain **selenocysteine**, the selenium-containing analog of cysteine. Selenocysteine is encoded by termination codon TGA. See Chapter 29 for details. However, even though TGA is occasionally used in this way, it serves as a termination codon for most proteins within the same cells.[376] Thus, the *context* in which the codon TGA occurs determines how it is read by the ribosomal machinery.

## 3. Palindromes and Other Hidden Messages

The sentence "Madam, I'm Adam" reads the same either forward or backward. Such sentences, known as **palindromes**,[377] are infrequent in the English language. However, most DNA contains many palindromes, sequences of base pairs that read the same in forward and reverse directions. Consider, for example, the gene that specifies the sequence of nucleotides for the tRNA molecules of Fig. 5-30. It is a double-stranded DNA segment in which one strand has a sequence identical to that of Fig. 5-30 except for the substitution of T for U and for ψ (pseudouridine) and for the lack of methylation and other base modifications. The second strand is the exact complement. Figure 5-34 shows the part of this gene (residues 49–76) that corresponds to the 3' end of the tRNA molecule. This DNA segment could exist in a second conformation having a loop on each side of the molecule (Fig. 5-34). The stems of the two loops in this **cruciform** conformation are identical and symmetrically disposed around the center of the molecule. If we overlook the seven nucleotides in the center of the loop, the message in the stems reads the same in both directions, as is indicated by the green arrows.

Palindromes are often imperfect as is the one shown in Fig. 5-34. Here the two stems in the cruciform structure are related by an exact twofold rotational symmetry but the loops at the ends of the stems are not. Unpaired bases may bulge at various points in double-stranded stems of longer palindromes. These imperfect palindromes in the DNA are responsible for much of the tertiary structure of the various kinds of RNA. The tertiary structure, in turn, often determines the interaction of the RNA with enzymes and other proteins.

Special properties may be observed for palindromes containing homopurine tracts in one strand, and therefore homopyrimidine tracts in the other. If two identical palindromes of this type occur close together it is possible that the pyrimidine-containing strand of one can join with a hairpin loop of the other to form a triplex base structure. A related triplex structure may be formed when inverted repeat sequences occur within a homopyrimidine tract in one chain of DNA (Fig. 5-34B).[378] These have been called **mirror-repeats** or **H-palindromes** to distinguish them from true palindromes.[237,379] Each base triplex structure contains one set of Watson–Crick hydrogen bonds and one set of Hoogsteen hydrogen bonds. The triplexes of H-DNA are all either TAT or $CGC^+$, where one C is protonated (see also Fig. 5-24).

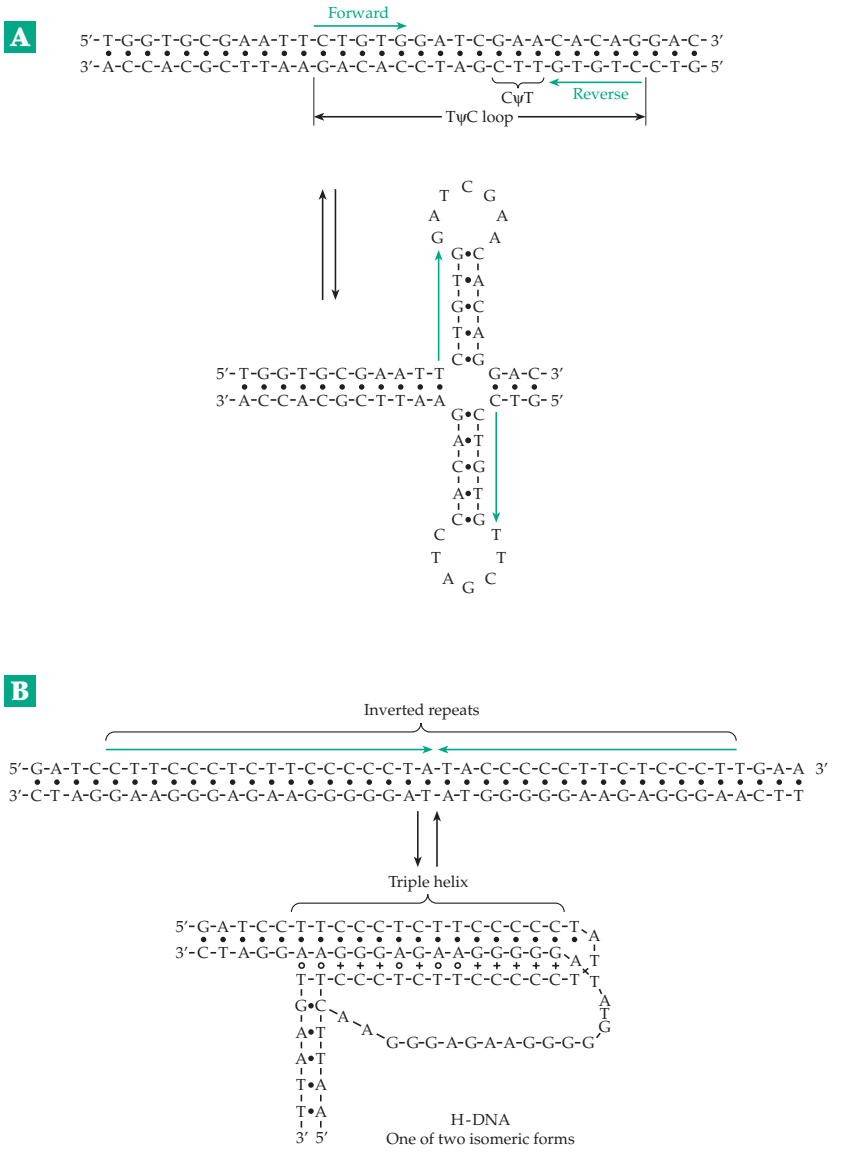



**Figure 5-34** (A) Two conformations of a segment of the yeast phenylalanine tRNA gene. The segment shown codes for the 3' end of the tRNA molecule shown in Fig. 5-30, including the TψC loop. (B) Formation of H-DNA (Fig. 5-24) proposed for a sequence in plasmid pGG32. The major element of the structure is the triplex, which is formed from the Watson–Crick duplex (•) associated with the homopyrimidine loop through Hoogsteen base pairing (∘, +). One of the two possible "isomeric" forms is shown. See Mirkin *et al.*[378]

Segments of H-DNA might block transcription, and proteins that bind to the triplex H-DNA may be involved in transcriptional regulation.[379]

Until rather recently there had been little to indicate that DNA actually assumes cruciform conformations in cells. However, strong experimental evidence suggests that some cruciform structures do form naturally.[380] Their formation from palindromic DNA [like the formation of Z-DNA from (G + C)-rich sequences] is a way of relieving torsional strain induced by supercoiling. Whether or not cruciform structures occur frequently within cells, there is no doubt that palindromic sequences are of great importance in the interaction of nucleic acids with symmetric dimeric and tetrameric protein molecules such as the gene repressor protein shown in Fig. 5-35.[381–383]

DNA contains numerous other protein binding sites which are not palindromes but whose sequences represent additional encoded information. The RNA transcripts likewise contain sequences that direct the catalytic machinery involved in splicing, that bind to ribosomal proteins, that control rates of transcription, and that cause termination of transcription.

### 4. The Base Composition of DNA and RNA

The nucleotide composition of DNA is surprisingly variable. The sum of the percentage cytosine plus the percentage guanine (C + G) for bacteria varies from 22 to 74%. That for *E. coli* is 51.7%. Among eukaryotic organisms, the range is somewhat narrower (28 to 58%; for humans, 39.7%). The fact that bacterial DNA molecules are more varied than those of higher organisms is not surprising. The prokaryotes have evolved for just as many more million of years as have we. Perhaps because of their simpler structure and rapid rate of division, nature has done more experimentation with genetic changes in bacteria than in people.

Comparisons of the C + G content of organisms have been used as a basis for establishing genetic relatedness.[384] However, since thymine is especially susceptible to photochemical alteration by ultraviolet light, bacteria with a high (C + G) content may have evolved in environments subject to strong sunlight or high temperatures, whereas those with a low (C + G) content have developed in more protected locations.[385,386]

## F. Interaction of Nucleic Acids with Proteins

Most functions of DNA and RNA are dependent on proteins. Cells contain a vast array of polypeptides that bind to nucleic acids in many specific ways.[387–390] We have seen (Fig. 5-21) that the histones hold supercoiled DNA in the form of nucleosomes[391] in eukaryotic nuclei and similar proteins bind to and protect the double-stranded DNA of bacteria.[205,392] Sperm cells "package" a large amount of DNA into a small space with the help of small arginine-rich proteins called **protamines.**[393] Cells always have some single-stranded DNA segments as well as single-stranded DNA binding (**SSB**) proteins. Among the latter are proteins from *E. coli*[394,395] and from viruses.[396–399] Specialized proteins bind to DNA sequences in telomeres[38,272] and centromeres.[400] A large number of proteins interact with RNA in ribosomes, spliceosomes, and other complexes.

A host of enzymes, which are described elsewhere in the book, act on DNA and RNA. They include hydrolytic nucleases, methyltransferases, polymerases, topoisomerases, and enzymes involved in repair of damaged DNA and in modifications of either DNA or RNA. While most of these enzymes are apparently proteins, a surprising number are **ribozymes**, which consist of RNA or are RNA–protein complexes in which the RNA has catalytic activity.

### 1. The Helix–Turn–Helix Motif

Much current interest in DNA–protein interactions is focused on regulatory processes. In prokaryotes the initiation of transcription of a large fraction of the genes is blocked by the binding of proteins known as **repressors**. While their structures are varied, one large group of repressors have DNA-binding domains with a similar helix–turn–helix architectural motif. They bind with high affinity to specific control regions which contain palindromic DNA sequences such as the following one, which defines a binding site for the *E. coli* **trp** (tryptophan) **repressor**:


These repressor proteins form dimers joined through a rigid central domain with a pair of arms, each containing two helices that form the helix–turn–helix motif. One helix of each pair fits into the major groove of the DNA helix and forms a "reading head" that carries a specific arrangement of amino acid side chains able to locate the symmetric nucleotide sequences flanking the twofold axis of the palindrome. In the case of the *trp* repressor, these are the ACTAGT hexanucleotides (marked by the arrows in the preceding structure). The interaction is depicted in Fig. 5-35.[401–405] The repressor has a high affinity for the DNA only if one molecule of L-tryptophan is bound to each subunit at a specific site near the DNA helix. Binding of the tryptophan causes a conformational change, which is pictured in Fig. 5-35. If tryptophan is absent the

repressor binds to the palindromic DNA only weakly and transcription of genes needed for tryptophan biosynthesis occurs freely. However, if tryptophan accumulates within the cell it binds to the repressor molecules causing them to bind firmly to DNA and prevent transcription. There are at least three of these palindromic sequences in *E. coli*, each one regulating a set of genes (**operons**) involved in tryptophan synthesis. In contrast to the effect of tryptophan, the tryptophan analog indole-3-propionate, which lacks the amino group of tryptophan, *derepresses* the same operons. It also binds to the *trp* repressor, but with the indole ring flipped over by 180° so that its carboxylate group contacts the phosphate groups of the DNA repelling them through both electrostatic and steric effects.[406,407] The picture in Fig. 5-35 is an oversimplification. In fact, at some binding sites (operator sites) more than one dimeric repressor binds in a tandem fashion.[405]

The helix–turn–helix motif is also found in many other proteins. One of these is the bacterial **lac** (lactose) **repressor** which controls the *lac* operon and for which

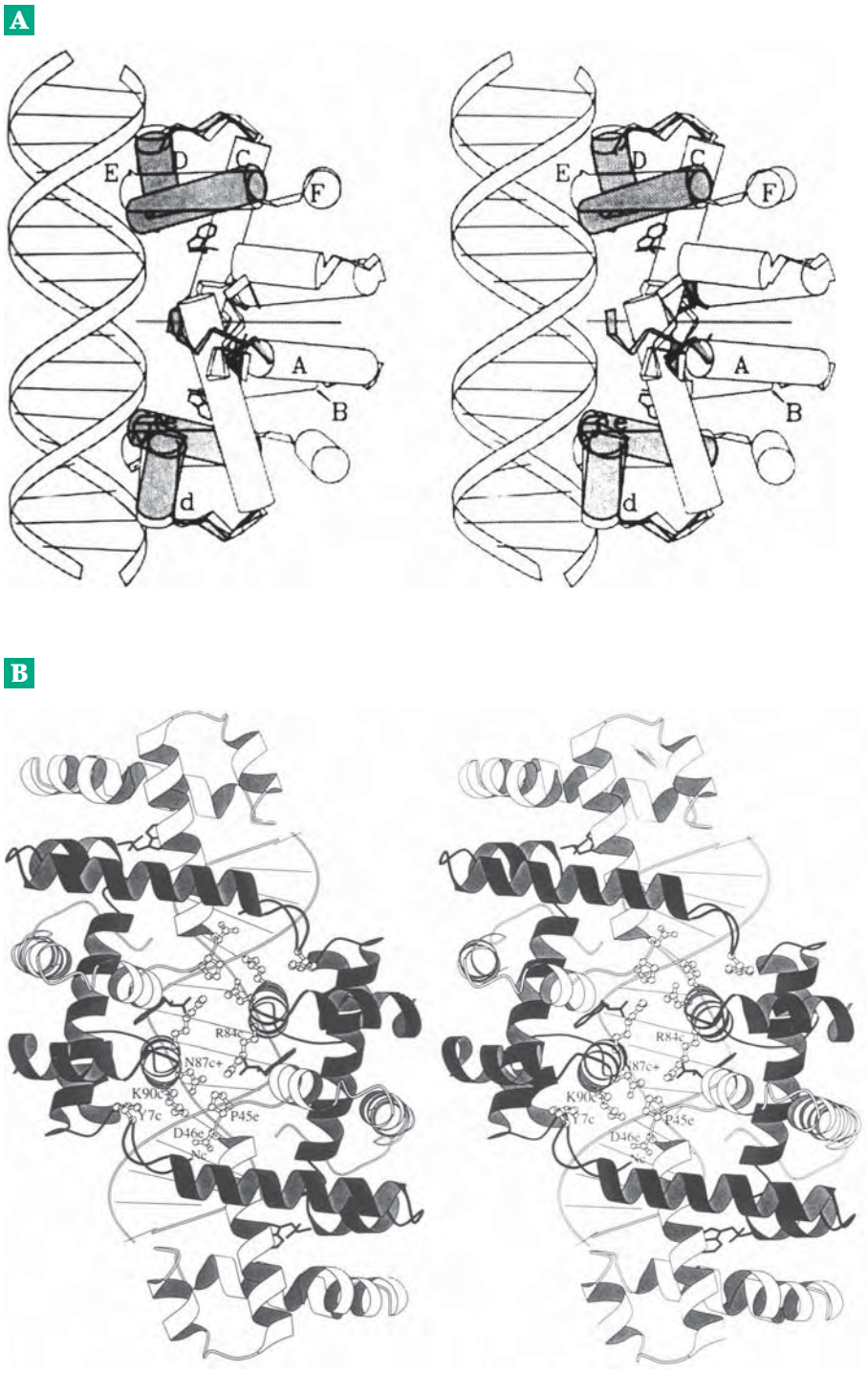



**Figure 5-35** Stereoscopic drawings illustrating the binding of a dimeric molecule of the Trp repressor protein to a palindromic sequence in DNA. (A) Schematic view showing structures of the aporepressor (partly shaded gray) and the holorepressor with bound tryptophan (unshaded) are superimposed. Cylinders represent the α helices in (B). From Zhang *et al.*[402] (B) MolScript ribbon diagram with a few side chains that interact with the DNA shown. Two tandemly bound dimeric repressor molecules are shown. Two bound molecules of tryptophan are visible in each dimer. The DNA is drawn as a double helix with lines representing the base pairs. From Lawson and Carey.[405]

the terms operator, promoter, repressor, and operon were first introduced (Chapter 28).[408–410] Some bacterial viruses, such as **phage lambda (λ)** of *E. coli*, encode repressors that allow the virus to reside in the bacteria without immediately destroying them. The λ repressor and other closely related proteins also utilize the helix–turn–helix motif[411–415] as do some proteins that *activate* transcription (see Chapter 28).

## 2. Other DNA-Binding Motifs

Nobody knows how many different DNA-binding structures may be discovered. However, most of those that are designed to recognize specific DNA sequences have some part that fits into the major groove of B-DNA.[416] Usually the DNA structure must be in a specific form: B-, A- or Z-, but it may sometimes be bent or distorted. The DNA recognition motifs in the proteins may consist of helices, β strands, or loops.

***Leucine zipper proteins.*** Several transcription factors have the leucine zipper structure, which has been described in Chapter 2 and was illustrated there by the structure of transcription factor Max (Fig. 2-21).[417] Related structures include those of the transcription activators c-Jun and c-Fos[418] and of the yeast transcription factor GCN4.[419,420] The latter is a dimeric protein in which the C-terminal halves of the two monomers form the helices of the leucine zipper. The helices, which are continuous for over 60 residues, fan out to interact with the DNA double helix as in Fig. 2-21. A 19-residue basic domain of each protein helix crosses the major groove of the DNA with side chain groups interacting with the DNA as is illustrated in Fig. 5-36.

***Zinc fingers.*** Another large group of transcription factors contain a bound zinc ion in a "finger" motif. The $Zn^{2+}$ is held by two cysteine $-S^-$ groups present on a loop of β structure and by two imidazole groups present on an α helix (Fig. 5-37).[421–424] Amino acid sequences found in many proteins that regulate transcription tend to form zinc fingers and to interact with DNA. These were recognized first in the transcription factor TFIIIA from *Xenopus laevis* in 1985.[421,425,426] This protein contains a zinc finger motif repeated nine times. X-ray structures, one of which is shown in Fig. 5-38, are known for proteins with three[423] and five[427] zinc fingers. (See Chapter 28.)

***Beta ribbons.*** An antiparallel double-stranded β ribbon can fit into the major groove of DNA and form

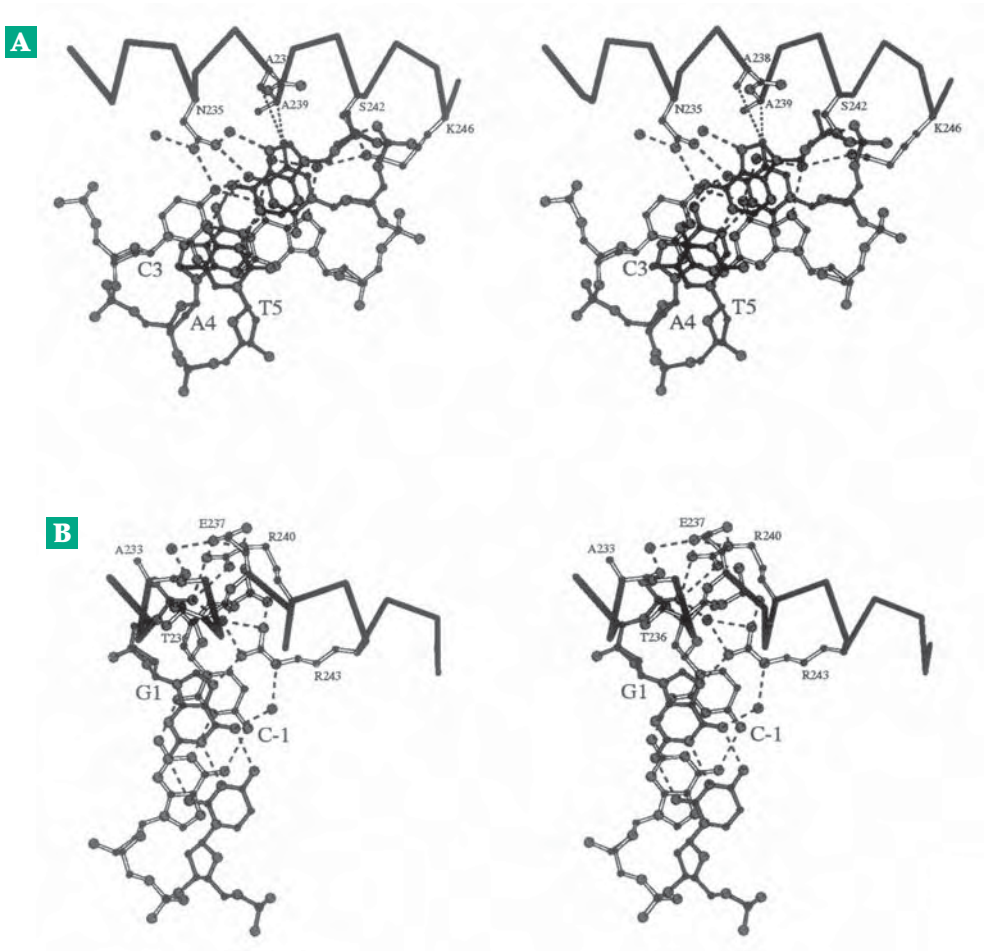



**Figure 5-36** Stereoscopic diagrams showing some of the interactions between an N-terminal helical domain of the yeast transcription factor GCN4-bZIP, a leucine zipper protein, and a specific palindromic DNA binding site:

| | –5 | | –1 | 1 | | 5 |
|---|---|---|---|---|---|---|
| 5′ | –A | T | G | A | C | G | T | C | A | T |

(The bases are numbered outward from the central C and G.) The small solid spheres are water molecules. Notice the water mediated interactions of the basic arginine and lysine side chains with the nucleic acid bases and also the interaction of R240 and R243 (in B) with a backbone phosphate. The overall structure of the protein is similar to that of another leucine zipper shown in Fig. 2-21. From Keller *et al.*[419] Drawings courtesy of Timothy J. Richmond.

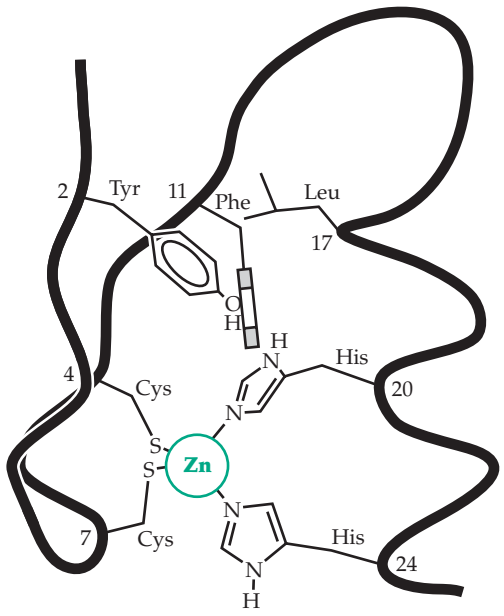
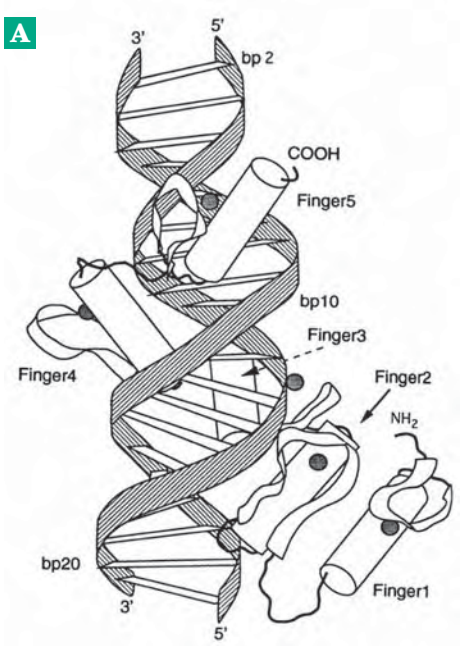
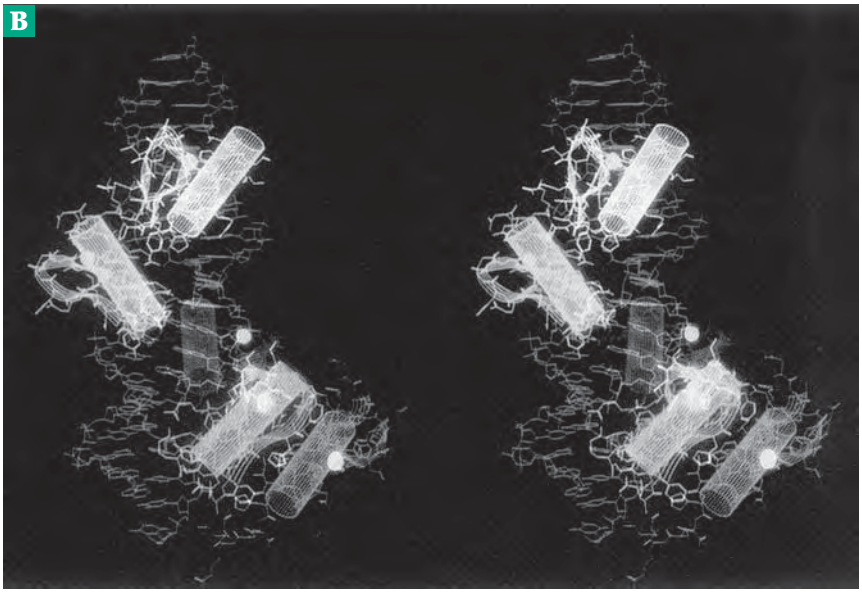



**Figure 5-37** Three-dimensional structure of a zinc finger. This is formed by the binding of $Zn^{2+}$ to the following sequence in a protein:

$$\underset{\beta}{\phi\ X\ C\ X_{2-4}\ C\ X\ X\ X\ F\ X_5\ L\ X\ X\ H\ X_{3,4}\ H}$$
$$\underset{\beta}{\longrightarrow}\ \text{turn}\ \underset{\beta}{\longrightarrow}\ \text{turn}\ \underset{\alpha\ \text{helix}}{\longrightarrow}$$

Here $\phi$ is a hydrophobic amino acid and X may be any amino acid. After Krizek *et al.*[428]



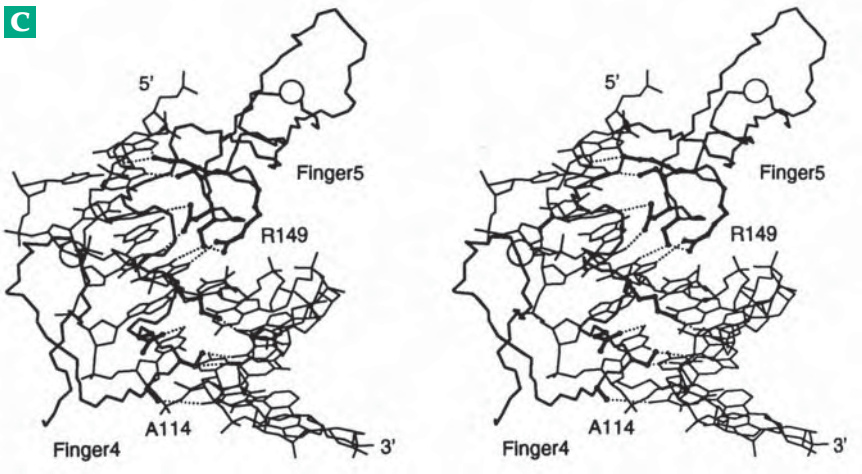




**Figure 5-38** (A) Sketch of a complex of a five-zinc finger protein bound to a high-affinity site in DNA. This protein is a human oncogene called GL1. Its gene is amplified in many tumors, but it is similar to many other transcription factors. The cylinders and ribbons highlight the $\alpha$ helices and $\beta$ sheets. Cobalt ions, which replaced the $Zn^{2+}$ in the crystals, are shown as spheres. (B) Stereoscopic view of the complex in a similar orientation. (C) Stereoscopic view emphasizing the contacts of side chains from fingers 5 and 6 with the DNA. The view is similar to that in (A) and (B) but the structure has been tilted back. Drawings courtesy of Nikola Pavletich.[427]

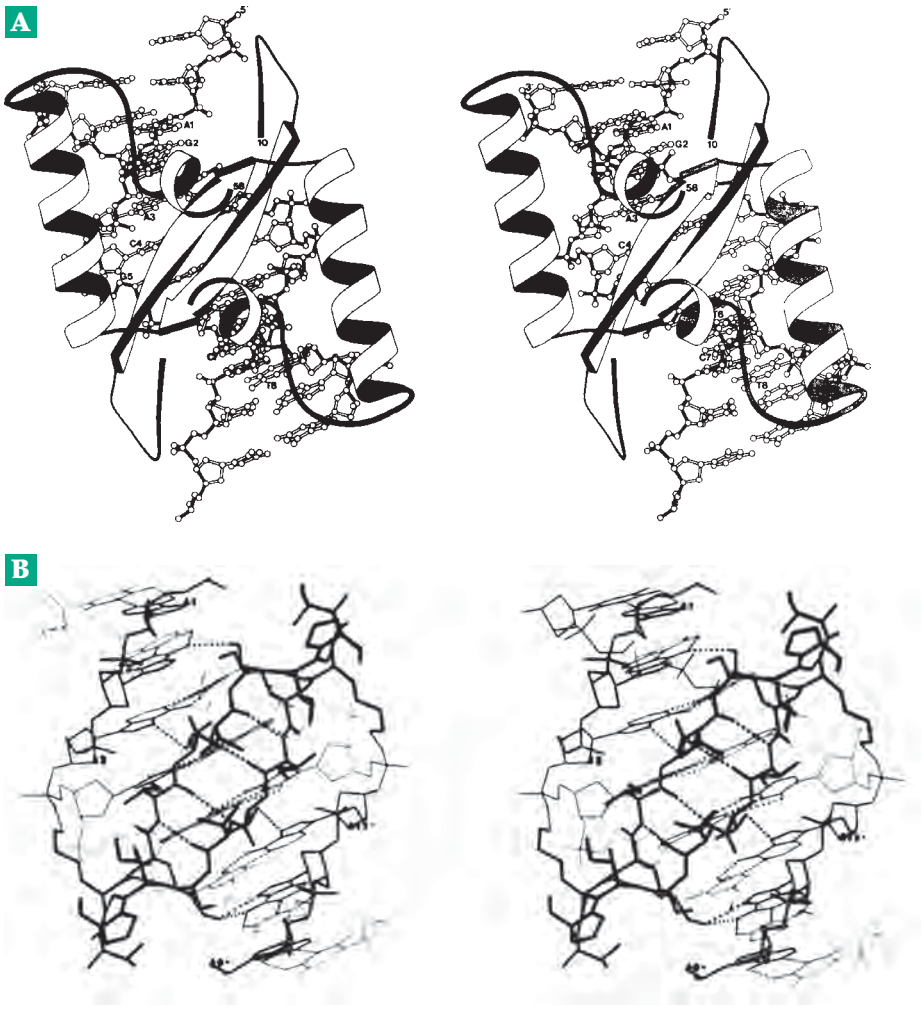



**Figure 5-39** (A) Stereoscopic ribbon representation of the *E. coli* methionine repressor–operator complex. Two subunits form a dimer with a double-stranded antiparallel β ribbon that fits into the major groove of the DNA in the B form. One strand is shaded more darkly than the other. (B) View of the β ribbon and its interactions with the DNA. Notice the direct hydrogen bonds from amino acid side chains of lysine and threonine residues to bases in the specific palindromic sequence AGACGTCT. From Somers and Phillips.[429] Drawings courtesy of S. E. V. Phillips.

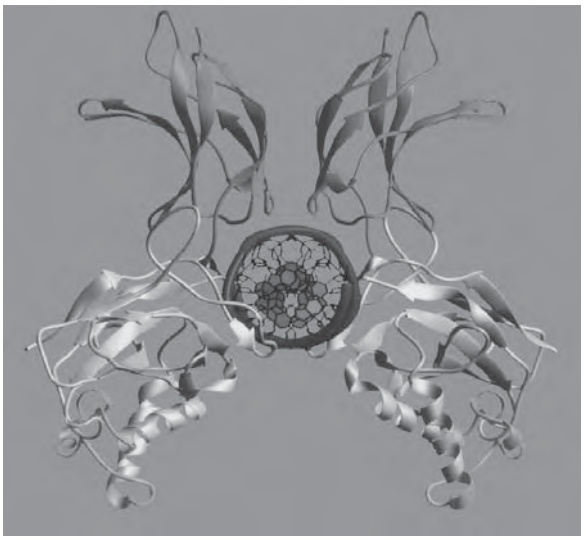


**Figure 5-40** Structure of a protein known as transcription factor NF-κB bound to its DNA target. Each subunit of the dimeric protein contains two β barrel domains. The loops at the ends of the barrels interact with the DNA in the center. From Müller *et al.*[433] Courtesy of Stephen C. Harrison.

hydrogen bonds and other interactions with the DNA. This is the basis for specific recognition of the operator sequence for the *E. coli* **met** (methionine) **repressor** as is illustrated in Fig. 5-39.[429] The same binding motif is employed by some other repressors[430,431] and also by the abundant bacterial DNA-binding protein **HU**.[432] It has also been utilized in designing the previously mentioned hairpin polyamide DNA-binding compounds.

***The winged helix family.*** A group of large protein transcription factors contain an N-terminal DNA-binding domain with the striking winged helix motif shown in Fig. 5-40.[433–436] It occurs in proteins from a wide range of organisms from yeast to human.

### 3. RNA-Binding Proteins

Among the many proteins that bind to RNA molecules[437–439] are the aminoacyl-tRNA synthetases, a variety of other well known enzymes,[440] the ribosomal proteins discussed in Chapter 29, and various proteins with dual functions of catalysis and regulation of

translation. A widely found RNA-binding **ribonucleoprotein domain** (RNP domain), occurs in hundreds of proteins including many of the RNA processing proteins considered in Chapter 28.[437,441,442] This 70–90 residue αβ module binds an RNA strand against a β sheet surface. Another ~ 70-residue protein motif binds ds RNA.[443]

## G. Viruses and Plasmids

Attacking every living thing from the smallest mycoplasma to human beings the nucleoprotein particles known as viruses have no metabolism of their own. However, they "come alive" when the nucleic acid that they contain enters a living cell. Viruses are significant to us not only because of the serious disease problems that result from their activities but also as tools in the study of molecular biology. A mature virus particle or **virion** consists of one or more nucleic acid molecules in a protein coat or **capsid**, usually of helical or icosahedral form. The capsid is made up of morphological subunits called **capsomers**. They can sometimes be seen clearly with the electron microscope. The capsomers in turn are usually composed of a number of smaller protein subunits. Some of the larger viruses are surrounded by membranous envelopes. Others, such as the T-even **bacteriophage** which attack *E. coli*, have extraordinarily complex structures (Box 7-C).

Most viruses contain a genome of either double-stranded DNA or single-stranded RNA, but some small viruses have single-stranded DNA and others have double-stranded RNA. The number of nucleotides in a virus genome may vary from a few thousand to several hundred thousand and the number of genes from 3 to over 200. Sometimes the nucleic acid molecules within the virion are circular, but in other cases they are linear. Table 5-7 lists a few of the known types of viruses as well as some individual viruses.[72,444] The size of the genome, in kilobases (kb), or kilobase pairs (kbp) is indicated. The number of genes in a virus is often somewhat more than one per kbp of DNA. A vast amount of information is available about these infectious particles. Only a few reference sources are cited here.[72,444–446] The architectures of some helical and icosahedral protein coats that surround genomic DNA or RNA are described further in Chapter 7.

### 1. Viruses with Single-Stranded DNA

The very small **helical bacteriophages** of the Ff family, such as fd, f1, and M13 (see Fig. 7-7) resemble thin bacterial pili but each virus particle contains a molecule of single-stranded circular ~6400 nucleotide DNA of $M_r$ ~2 x $10^6$ which encodes only ten proteins.[447–449] Bacteriophage ϕX174 (Fig. 5-41), an **icosahedral** DNA-containing virus 25 nm in diameter, is only three times as thick as the thinnest cell membrane. Its DNA contains just 5386 nucleotides.[450–452] The similar bacteriophage G4 contains 5577.[453,454]

It is remarkable that such tiny viruses are able to seize control of the metabolic machinery of the cell and turn it all in the direction of synthesis of more virus particles. There are only 11 known proteins encoded by the genes of ϕX or G4. Three genes encode the three kinds of protein subunits of the virus coat. Sixty copies of each are needed, as are 12 copies of a "pilot" protein.[451] Eight of the genes are spaced closely together, occupying most of the DNA. The other three genes are embedded within some of the first eight but in different reading frames. In one short region of the G4 chromosome the same nucleotides are part of three different genes, using all three possible reading frames. In addition to their own genes, these small viruses make use of many components of the cell that they infect. A large group of animal viruses, the **parvoviruses**, are similar in size and architecture to bacteriophage ϕX174.

Canine parvovirus, first identified in 1978, is now endemic.[455,456] Childhood **fifth disease** is also caused by a parvovirus.[456,457] When these single-stranded DNA viruses infect cells a double-stranded **replicative form** of DNA arises by synthesis of the complementary **negative strand** alongside the original **positive** DNA strand. Many copies of the replicative form are then synthesized. The negative strands of the replicative forms serve as templates for synthesis of numerous new positive strands that are incorporated into the progeny viruses. The whole process may take only 20 minutes. Some parvoviruses are unable to reproduce unless the cell is also infected by a larger adenovirus.

While most plant viruses contain dsRNA, the **geminiviruses**,[458] which cause a number of plant diseases, contain single-stranded DNA. The virus particle consists of a fused pair of incomplete icosahedra, evidently containing a single ~2500-bp DNA strand. Replication may require coinfection with two virus particles of differing sequence.

### 2. Viruses with Double-Stranded DNA

One of the smallest viruses containing double-stranded DNA is the 3180-nucleotide human **hepatitis B virus**. It infects millions of people throughout the world causing chronic hepatitis and often liver cancer.[459–461] The circular DNA is surrounded by an envelope consisting of proteins, carbohydrate, and a lipid bilayer. Many icosahedral viruses also contain dsDNA. Among them are the **papovaviruses**, some of which cause warts and others malignant tumors. Much studied by biochemists is the 5386-nucleotide **simian virus 40 (SV40)**, a monkey virus capable of inducing tumors in other species.[462,463] Closely related

## TABLE 5-7
## Characteristics of Some Individual Viruses and Groups of Viruses

| Type of genome[a] and group or individual virus name | Shape[b] | Diameter[c] (nm) | Masses in daltons x $10^{-6}$ | | Thousands of bases or base pairs (kilobases) |
|---|---|---|---|---|---|
| | | | Total | DNA or RNA | |
| **DNA, single-stranded** | | | | | |
| Bacteriophages *fd*, *f1*, *M13* | H | ~6 x ~880 (length) | 17.6 | 2.1 | 6.4 |
| Bacteriophages ϕ174, G4 | I | 25 | 6.2 | 1.8 | 5.4–5.6 |
| Parvoviruses | I | 18–25 | | 1.8 | 5.5 |
| Geminiviruses | I (fused pairs) | | | | |
| **DNA, double-stranded** | | | | | |
| Hepatitis B virus | I (enveloped) | | | 2.1 | 3.18 |
| Papoviruses | I | 45–55 | | | |
| SV40 (monkey) | I | | 17.6 | 3.5 | 5.22 |
| Polyoma (mouse) | I | 45 | 23.6 | 3.3 | 4.96 |
| BK virus (human) | I | | | 3.4 | 4.96 |
| Papilloma (human wart) | I | 56 | | 5.3 | 8.0 |
| Bacteriophage ϕ29 | T | | | 12 | 19.3 |
| Adenoviruses | I | 70 | | 20–30 | 30–45 |
| Bacteriophage Mu | | | | 25 | 38 |
| Bacteriophage T7 | T(short) | | | | 39.9 |
| Bacteriophage P22 | T | | | 28.5 | 43.2 |
| Bacteriophage λ | T | | | 32 | 48.6 |
| T-even bacteriophage | T | 100 x 80 (head) | 215 | 130 | 166 |
| Baculoviruses of insects | I | 70–130 | | | |
| Herpesviruses | I | 100 | | | |
| | core | 78 | ~1000 | 80–120 | 80–140 |
| | envelope | 150–200 | | | |
| Pox viruses, e.g., | | | | | |
| Smallpox, vaccinia (cowpox) | C | 160 x 250 | ~4000 | 150–240 | 240–300 |
| Cauliflower mosaic virus | I | | | | 8.0 |
| **RNA, single-stranded** | | | | | |
| Unsheathed | | | | | |
| Potato spindle tuber viroid | | | | 0.116 | 0.30 |
| Hepatitis delta virus | | | | | 1.7 |
| Sheathed, *plus-strand* | | | | | |
| Tobacco necrosis satellite | I | 18 | 1.7 | 0.4 | 1.20 |
| Small bacteriophages | | | | | |
| R17, MS2, Qβ | I | 23–26 | 3.6–4.0 | 1.2–1.5 | 3.5–4.5 |
| Picornaviruses | I | | 8.4 | 2.6 | 7.9 |
| Polioviruses | I | 27 | 6.4 | 2 | 6.1 |
| Rhinoviruses | I | 27–30 | 7–8 | 2.2–2.8 | 6.7–8.5 |
| Turnip yellow mosaic virus | I | 28 | 5.0–6.0 | 2.0 | 6.1 |
| Tobacco mosaic virus | H | 18 x 300 | 40 | 2.2 | 6.7 |
| Togaviruses | I | 20–40 | | 4 | 11 |
| *Negative-strand* viruses | | | | | |
| Influenza virus | I | 80–100 | 200 | 2.0 | 6.1 |
| Bullet-shaped viruses | | | | | |
| Rhabdoviruses | C | 20 x 130 | | | |
| Retroviruses | I | 80–100 | | 7–10 | 20–30 |
| **RNA, double-stranded** | | | | | |
| Reoviruses | I | 55–60 | | 11–12 | 16–18 |
| Mobillivirus | I (enveloped) | 38 | | | |
| Rotavirus | Wheel-shaped | 70 | | | |
| *Leishmania* RNA virus | | | | | 5.28 |

[a] Complete nucleotide sequences are known for most of these viruses.
[b] Shapes are indicated as I, icosahedral; H, helical; T, a tailed phage; C, complex.
[c] The second dimension given for some helical and complex viruses is the length (nm).

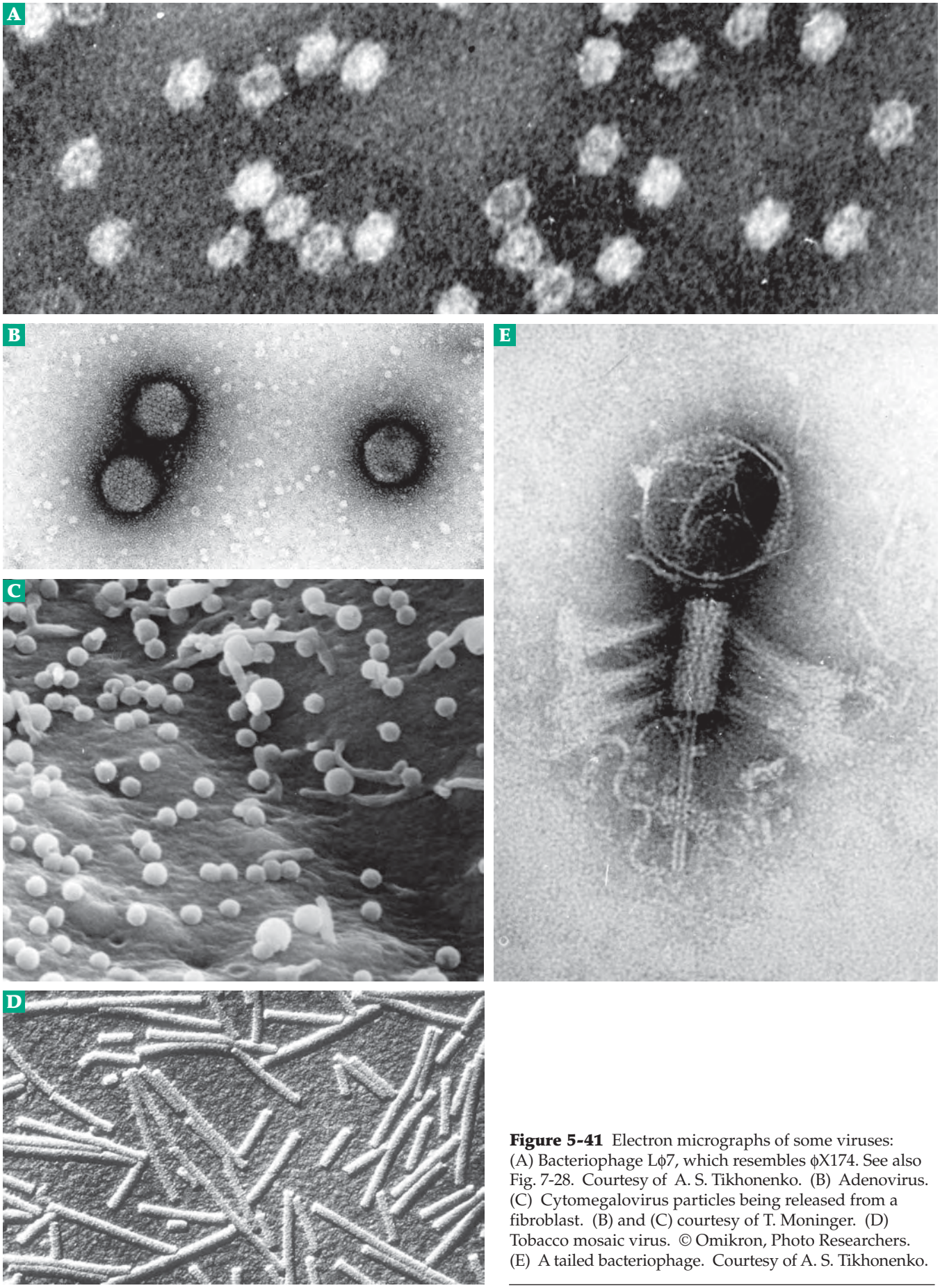



**Figure 5-41** Electron micrographs of some viruses: (A) Bacteriophage Lϕ7, which resembles ϕX174. See also Fig. 7-28. Courtesy of A. S. Tikhonenko. (B) Adenovirus. (C) Cytomegalovirus particles being released from a fibroblast. (B) and (C) courtesy of T. Moninger. (D) Tobacco mosaic virus. © Omikron, Photo Researchers. (E) A tailed bacteriophage. Courtesy of A. S. Tikhonenko.

is the **polyomavirus** of the mouse[464] and the **BK virus** of humans,[465] also suspected of causing cancer. The **papillomaviruses** cause warts and perhaps cancer[466] and the larger (70 nm diameter) **adenoviruses** (Fig. 5-41),[467,468] cause respiratory infections. Some 32 types infect humans. Many of the details of eukaryotic transcription were first studied using the adenovirus.[469] The very large **herpesviruses** are enveloped by a lipid-containing membrane.[470–472] Among them are herpes simplex viruses, which infect human mucous membranes and varicella-zoster virus, which causes chicken pox and shingles. Other herpesviruses include cytomegaloviruses (Fig. 5-41), another common human pathogen, and the Epstein–Barr virus, which causes mononucleosis and is suspected of causing cancer.[473] Another herpesvirus has been associated with multiple sclerosis.[474] The very large icosahedral **baculoviruses** cause polyhedroses in insects. One that infects the fly *Tipula* measures 130 nm in diameter. **Poxviruses** are also large and complex.[475] The tailed bacteriophages (Box 7-D) range in size from the small ~29-kDa phage P22 to the very complex T-even phage. Some plant viruses also contain dsDNA. One of these, the **cauliflower mosaic virus**,[476,477] is transmitted by aphids. It has proved useful as a gene-transfer vehicle for genetic engineering.

### 3. Viruses Containing RNA

Several plant diseases including the **potato spindle tuber disease** are caused by **viroids**, molecules of single-stranded RNA only 240–380 nucleotides in length with a folded structure.[478–481] Such an RNA could code for a protein containing only about 100 amino acids. However, the known 359-nucleotide sequence of the potato spindle tuber virus contains no AUG initiation codon. It seems impossible that the virus carries any gene for a protein. Conserved features of viroid sequences suggest a close relationship to the intervening DNA sequences known as type I introns (see Fig 28-19). Whatever its genetic message, a plant viroid causes the plant cell to replicate many copies of the viroid molecule, which may then be transmitted to other plants by aphids, or on the surface of tools, by humans. A larger 1678-nucleotide viroid-like RNA (**hepatitis delta virus)** has been identified in human patients with severe chronic hepatitis and who were also infected with hepatitis B virus. The ssRNA in this virus does encode at least one protein.[482–484]

One of the smallest of the encapsulated RNA-containing viruses is the **satellite tobacco necrosis** virus. It replicates only when the plant is also infected with the larger tobacco necrosis virus. The satellite virus, whose three-dimensional structure is known from X-ray diffraction studies,[485] contains a 1200-nucleotide strand of RNA which encodes a 195-residue protein. Sixty copies of the latter are assembled around the RNA in an icosahedral array (Fig. 7-14) to form the virion. The structure of the similar satellite tobacco mosaic virus has also been described in detail.[486,487]

RNAs of the small bacteriophages f2, R17, MS2 and Qβ contain 3500 – 4500 nucleotides and are enclosed in icosahedral shells made up of 90 identical subunits.[488–490] Initially only three genes were evident but a fourth small gene in a different reading frame has since been found in at least one of them.[491] The RNA molecules in these and other **positive-strand** viruses serve as mRNA within the host cells. Another group of small RNA viruses are the **picornaviruses** (picoRNA, meaning very little RNA). Many of these icosahedral viruses of 15–30 nm diameter attack humans. Among them are the **enteroviruses** including the **polioviruses**,[492,493] the **hepatitis A virus**,[494] the **coxsackieviruses**, and some of the **echoviruses**. A second class of picornaviruses include **rhinoviruses** which cause the common cold. More than 100 types are known.[495–497] The **foot–and–mouth disease** virus,[498–500] which attacks the cloven-footed animals, and the **Mengo virus**,[501] which can cause a fatal encephalitis in mice, are also picornaviruses. The three-dimensional structures of polio virus, human rhinoviruses, Mengo virus, and many other eicosahedral viruses are known. Their architecture is discussed in Chapter 7.

The **togaviruses**, which are a little larger than the picornaviruses, have an icosahedral core surrounded by a lipid membrane. Yellow fever and rubella (German measles) are both caused by togaviruses. Other togaviruses, such as **Sindbis virus**[502] and **Semliki Forest virus**,[503] have become important in biological research.

A large number of icosahedral RNA viruses of diameter 28–30 nm (Fig. 7-14) attack plants, causing diseases such as tomato bushy stunt,[504] southern bean mosaic,[505] or turnip yellow mosaic. Best known of the helical RNA viruses is the **tobacco mosaic virus** (Figs. 5-41, 7-8).[506-507a] Its genome contains 6395 nucleotides as linear ssRNA. Many strains are known. Related viruses cause cucumber green mottle[508] and other plant diseases.

Large viruses of 80 – 100 nm diameter bearing 8–10 spikes at the vertices of the icosahedra cause influenza,[509,510] mumps, measles, and related diseases. The internal structure must be complex. Only 1% of the virus is RNA, and that consists of several relatively small pieces. These are **negative strand viruses** whose RNA is of the opposite polarity to the mRNA. The latter must be formed by transcription from the negative strand. The viruses carry their own RNA polymerase for this purpose. Of even more complex structure are the bullet-shaped **rhabdoviruses** which cause rabies and vesicular stomatitis.[511] The diameter of these viruses is 65–90 nm and the length 120–500