CHAPTER 2

Macromolecules of Life

2.1 Introduction

Water constitutes about 70 percent of the weight of a living cell; the rest is mostly composed of *macromolecules* that each contain thousands of atoms. These very large molecules are made of chains of smaller units called *monomers*. The monomers that make up a biological macromolecule need not be identical. For example, the information macromolecule deoxyribonucleic acid (DNA) is made of four distinct monomers (nucleotides), whereas proteins can have up to twenty different monomers (amino acids). Even in the simplest living systems, hundreds of thousands of such macromolecules interact with each other at any instant of time and undergo or catalyze chemical changes.

The macromolecules of living organisms are classified into three groups: *proteins*, *nucleic acids*, and *carbohydrates*. Proteins make up most of the molecular machinery of all organisms. The word "protein" is derived from the Greek word "proteios," which means "of the first rank." Proteins are linear chains of at most twenty different amino acids (Fig. 2.1). Proteins constitute the building blocks of our tissues, facilitate complex chemical reactions, and act as sensors, transducers, and energy transformers. As enzymes, proteins bring substrates to appropriate configurations for chemical reactions to proceed. Figure 2.2 shows the number of proteins synthesized by various multicellular organisms grouped into major functional categories. All proteins contain the elements carbon, hydrogen, oxygen, nitrogen, and sulfur. When temporarily associated with a phosphate group, many proteins rapidly undergo shape changes and gain or lose enzymatic activity.

Nucleic acids are the information macromolecules of living systems. Like proteins, all nucleic acids contain carbon, hydrogen, oxygen, and nitrogen. Nucleic acids do not contain sulfur, but do contain phosphorus. Nucleic acids can be grouped into two sets: DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). The chemical recipes of proteins are coded on long and fragile DNA molecules. Simple unicellular organisms such as bacteria store all their hereditary information

Number of Proteins

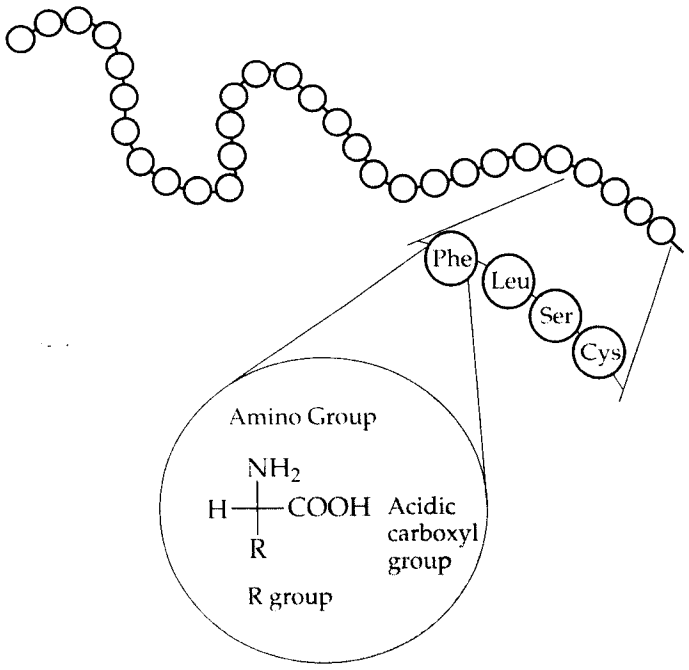


FIGURE 2.1 Polypeptides are long strings of molecular beads called amino acids. A protein is composed of one or more polypeptides. (Modified from http://www.nhgri.nih.gov/DIR/VIP/Glossary/Illustration/aminoacid.html.)

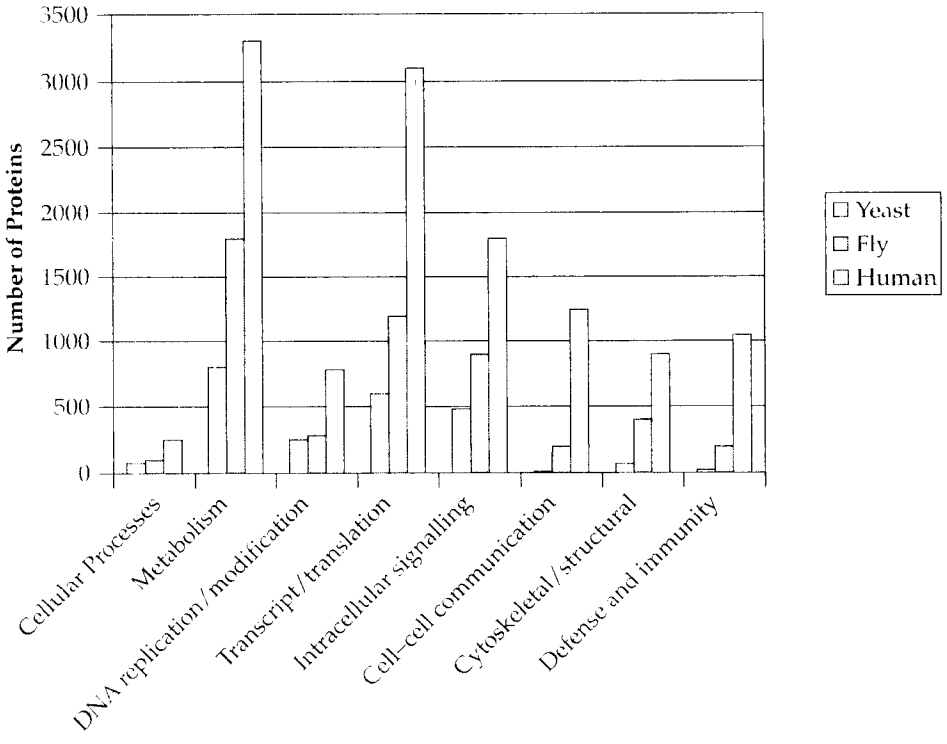


FIGURE 2.2 Number of different types of proteins generated by yeast, fruit fly, and human in major functional categories. This figure is based on that published by Eric S. Lander et al., *Nature* (2001, 409: 860–921) in an article entitled "Initial Sequencing and Analysis of the Human Genome."

in a single circular DNA, whereas animal and plant cells need multiple DNA molecules as their information database. In higher forms of life, from yeast cells to human, DNA is stored in a special cellular compartment called the nucleus. RNA molecules have chemical compositions that complement DNA and are involved in the synthesis of proteins. In a process called *transcription*, protein recipes encoded on DNA are copied onto messenger RNA (mRNA). The mRNA is then used as a template to build proteins from amino acids (Fig. 2.3). Transfer RNA molecules add amino acids to the growing chain of the protein according to the recipe encoded by mRNA. Therefore, in most modern life forms, information flows from DNA to RNA to protein. This complex process of information flow probably evolved gradually from simpler processes. Biochemical evidence points to short strands of RNA as the first information-carrying molecules. RNA molecules have the capacity to replicate without help from protein enzymes. If the first cells used RNA as the hereditary molecule, then RNA must have functioned as a template for the synthesis of DNA. Because DNA is double stranded and therefore stores two copies of its information, it probably evolved rapidly as the primary carrier of hereditary information in living cells. Meanwhile, RNA assumed its current role as a crucial intermediary in the translation of genetic recipes into proteins.

The third group of macromolecules found in biological systems is *carbohydrates*, which contain roughly equal amounts of carbon atoms and water molecules.

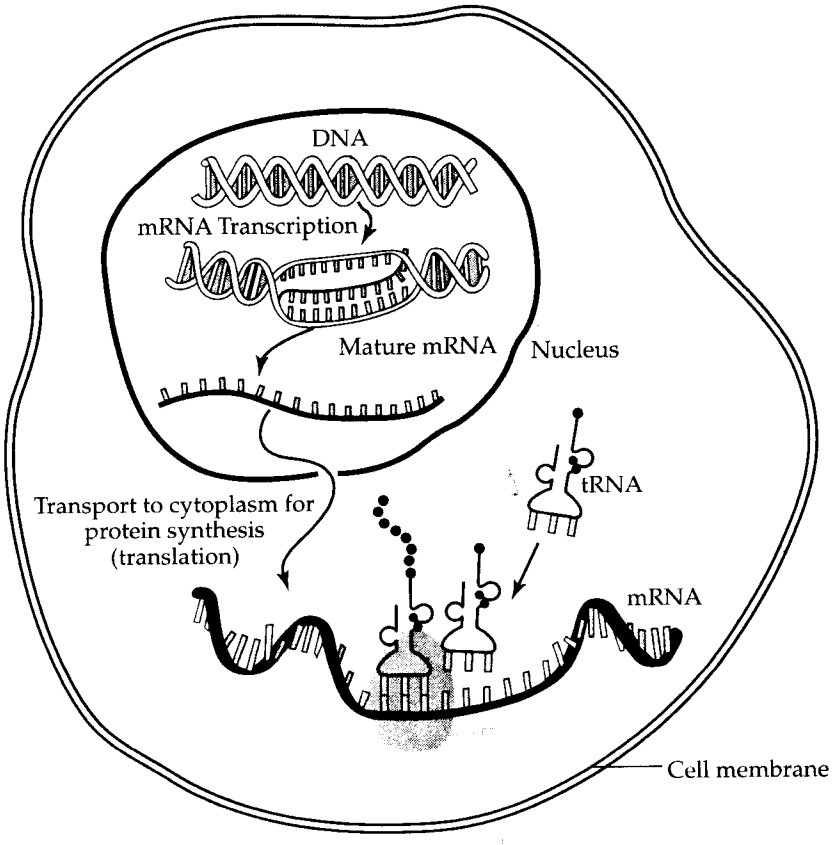


FIGURE 2.3 Information flow in a cell with a nucleus (eukaryotic cell). Hereditary information flows from DNA to RNA to protein. The protein recipes encoded by DNA are transcribed into mRNA and then these transcripts are used as templates in the synthesis of proteins. (Modified from http://www.nhgri.nih.gov/DIR/VIP/Glossary/Illustration/transcription.html.)

Carbohydrates such as glucose are an important source of energy for driving cellular processes. Carbohydrates are often present on proteins where they form covalent bonds with the free amino or hydroxyl groups on the amino acids asparagine and serine, respectively. There is considerable heterogeneity of these N- and O-linked carbohydrates among different proteins, and they can often form branched chains. Some carbohydrates such as sialic acid are negatively charged and when present on proteins exposed on the outside of cells form a charge barrier known as a glycocalyx. In some cases, the addition of a particular carbohydrate residue to a protein will determine its destination in the cell. For example, the presence of mannose-6-phosphate on a protein targets it to a subcellular compartment called the lysosome. Other carbohydrate-containing proteins are the blood group antigens A, B, A/B, and O. Differences in their carbohydrate makeup are responsible for the difference in the ability of a recipient to recognize the blood as "self." Carbohydrates on the surface of proteins are often the source of immune or allergic reactions.

Simple organisms such as bacteria and higher forms of life such as plants also use carbohydrates to build structures. The exoskeleton of all insects is made of a special carbohydrate called chitin (Fig. 2.4). Other polymeric carbohydrates include the plant carbohydrate, cellulose. The matrix surrounding most cellular structures is made up of carbohydrates and protein–carbohydrate complexes and is discussed in Chapter 6. Carbohydrates also play important roles in cell signaling and communication.

Another important group of molecules that form large structures is *lipids*. Lipids are mostly made up of carbon and hydrogen. Lipids are not true macromolecules, because they form large structures through associations other than covalent bonding. Lipid structures form membranes that separate cells from each other, create cellular compartments, and perform other complex tasks. Practically all biological membranes are made of lipid bilayers and associated proteins (Fig. 2.5).

Whereas each organism has its own unique set of deoxyribonucleic acids and proteins, all organisms use the same processes to build carbohydrates and lipids. A 65-kg adult male human is made of approximately 11 kg of protein, 9 kg of fat, 1 kg of carbohydrate, 4 kg of minerals and 40 kg of water. The weight of nucleic acids present in an organism is much less than the corresponding weights of other macromolecules.

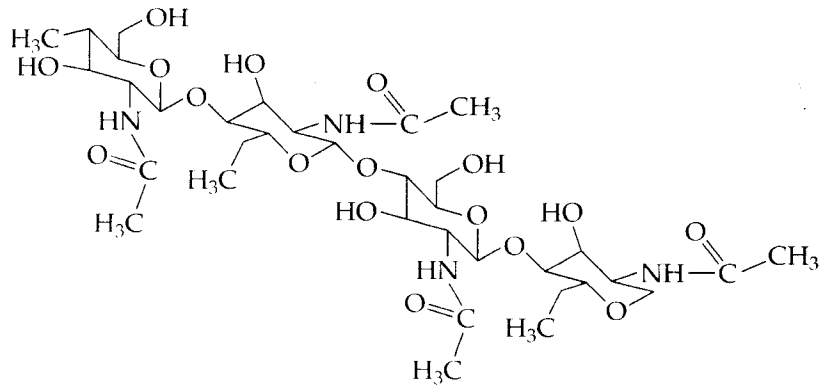


FIGURE 2.4 The exoskeleton of a scorpion is made of the carbohydrate chitin.

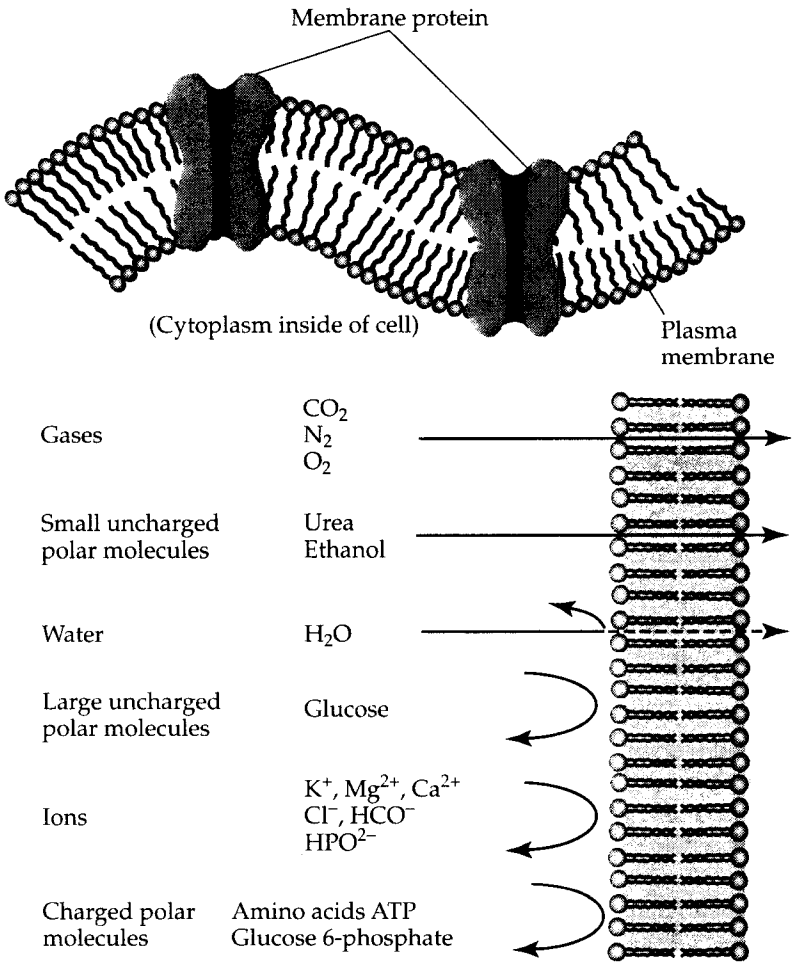


FIGURE 2.5 Schematic of a region of a cell membrane. Protein molecules embedded in a lipid bilayer carry out important functions in adhesion and signaling. Lipid bilayers and associated proteins control the entry of molecules into the cell.

The four groups of macromolecules found in living systems are not mutually exclusive. Often macromolecules of different classes interact with each other through formation of covalent bonds and weaker bonds. Proteins bind to carbohydrates to form glycoproteins; carbohydrate chains bind to lipids, forming glycolipids; enzymes bind to their substrates using hydrogen bonding, electrostatic interactions, and van der Waal forces. These interactions are essential for the survival of living organisms.

2.2 Fundamentals

2.2.1 Condensation Reactions

Macromolecules are constructed by a series of reactions identified as *condensation reactions* or *dehydration reactions*. Let *BABBA*—H be a polymer with reactive hydrogen

at one end, and B—OH another monomer with a hydroxyl group (—OH) at one end. The condensation reaction between them can be written as

$$BABBA—H + B—OH \rightarrow BABBAB + H_2O.$$

The products of the reaction are the polymer *BABBAB* and the water molecule H_2O. A macromolecule is formed by adding one monomer at a time to a growing chain of monomers. The monomers *A* and *B* are the molecular beads that make up the macromolecule, and they are chemically different than the building blocks *A*—H and *B*—OH. Monomers are identified by the name of the building block followed by the word "residue." For example, nucleotides are the building blocks of information molecules, and nucleotide residues are the monomers in these molecules.

The synthesis of macromolecules occurs only if energy is added to the system. Therefore, the synthesis of macromolecules from their building blocks must be coupled to energy-releasing (exergonic) reactions. The reverse reaction of condensation, *hydrolysis*, involves the use of water in the disassembly of an organic molecule. Although hydrolysis releases energy to the environment, due to its high activation energy, it does not occur in the absence of mediating enzymes. In living systems, both hydrolysis and condensation reactions require the catalytic action of enzymes.

2.3 Structural Representation

The four types of macromolecules that are essential for living systems exhibit distinctly different three-dimensional structures. As in many other fields of science and engineering, biochemists have struggled to capture the best three-dimensional representations of a structure in the two-dimensional space of a page or on a computer screen. In the following treatment, we use two distinct forms of structural representation for macromolecules. The representation based on the so-called *Haworth* projections is the standard in many biochemistry books; therefore, we will call this representation the *standard* representation. The second representation is used most often by institutions that create digital databases for molecules. One such database is the *Kyoto Encyclopedia of Genes and Genomes* (KEGG). The KEGG database was formed by the Institute for the Chemical Research, Kyoto University of Japan. For brevity, we call this second representation of molecular structure, the KEGG representation. These two ways of visualizing three-dimensional structures of biomolecules are illustrated in Fig. 2.6.

1. *Standard Representation*: (A) Simple molecules such as three-carbon sugars are presented as planar diagrams, with the understanding that atoms joining a carbon atom by horizontal bonds are in front of the page and those joined by vertical bonds are behind. This representation is called the *Fischer* representation and is illustrated in Fig. 2.6 for D-glyceraldehyde, a carbohydrate containing three carbon atoms. (B) In more complex organic molecules that form ring-like structures, the carbon atoms in the corners of the ring are not always explicitly shown. The approximate plane of the ring is considered to be perpendicular to the plane of the paper, with the thick line on the ring closest to the reader. This representation is called the Haworth representation and is illustrated in Fig. 2.6 for the biomolecule AMP.

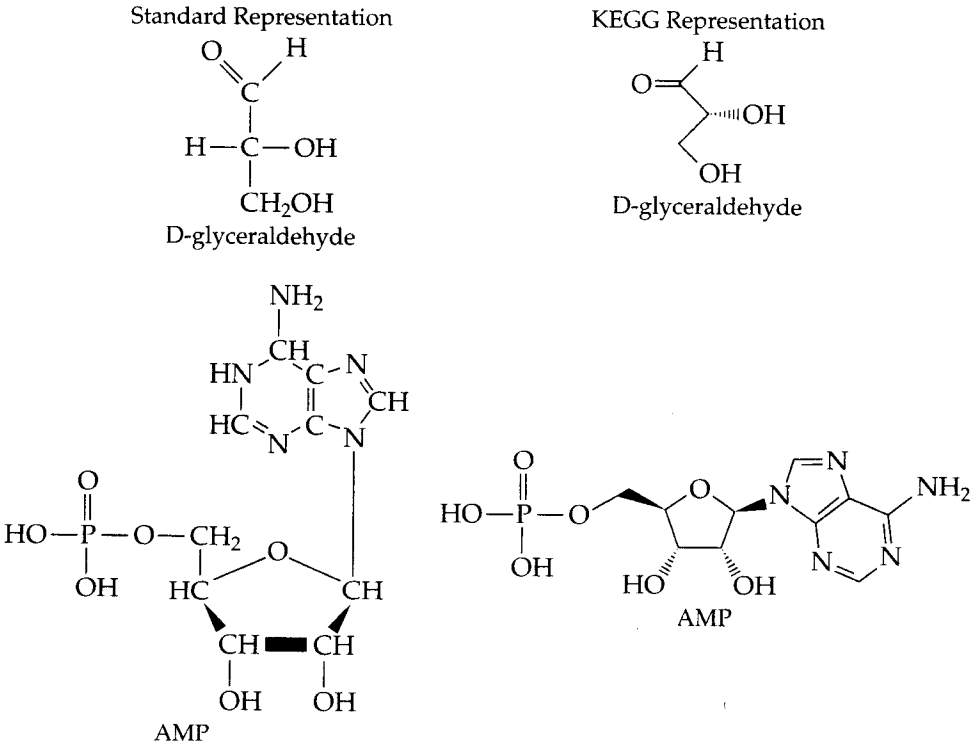


FIGURE 2.6 Standard (left column) and KEGG representation (right column) of the three-carbon sugar D-glyceraldehyde and AMP, a building block of DNA.

2. Compound formulas presented in the KEGG representation assume implicitly the presence of a carbon atom at each corner of a structural diagram if these corners are not explicitly marked for other atoms such as nitrogen or oxygen. Similarly, KEGG diagrams do not explicitly indicate the presence of hydrocarbon bonds. The presence of such bonds can be deduced from the fact that each carbon atom forms four bonds with neighboring atoms. The KEGG representations of the biomolecules D-glyceraldehyde and AMP are shown in Figure 2.6. The Web address for the KEGG compound database at the time of publication of this book was http://www.genome.ad.jp/kegg/catalog/compounds.html. The KEGG database is publicly accessible and contains the structural representations of more than 5000 biologically relevant molecules.

2.2.3 Stereoisomers and Biological Activity

A fundamental principle in chemistry is that molecules with identical chemical composition can have different spatial distributions of their atoms and therefore have significantly different physical and chemical properties. Molecules with the same chemical composition, but with different three-dimensional bond positioning are called *isomers*. Isomers that are mirror images of each other, like a left and a right hand, are called *stereoisomers*. Whenever a carbon atom binds to four different atoms or molecules, it gives rise to stereoisomers. Such a carbon atom is said to be

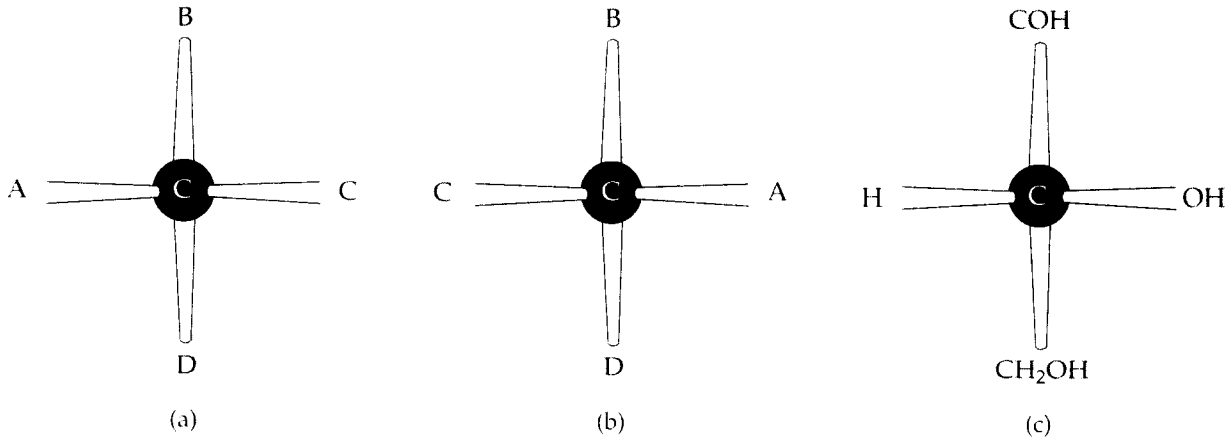
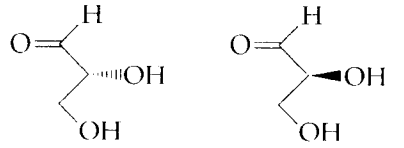


FIGURE 2.7 Molecules that are mirror images of each other (a, b). A specific example is that of D-glyceraldehyde, a three-carbon sugar (c).

an *asymmetric* or *chiral* carbon. The carbon atom shown in the black sphere in Fig. 2.7 is an asymmetric carbon. The molecular subunit with an asymmetric carbon may assume one of the two distinct structural configurations shown in Fig. 2.7a and Fig. 2.7b.

The macromolecule family of carbohydrates contains many examples of stereoisomers. For example, the three-carbon sugars D-glyceraldehyde and L-glyceraldehyde are stereoisomers. Both carbohydrates have the chemical formula $C_3H_6O_3$, but have three-dimensional shapes that are mirror images of each other, as shown by the following KEGG structural diagrams:



D-glyceraldehyde L-glyceraldehyde

The bond indicated by the dashed arrow points into the plane of the diagram, whereas the bond shown by the black arrow protrudes away from the plane of the diagram, toward the reader. The central carbon atom in each molecule is the asymmetric carbon. These carbon atoms are linked to four different subunits: H, OH, CH_2OH, and CHO (Fig. 2.7c). The bonds of the central asymmetric carbon in a three-carbon sugar can be arranged in space in two different ways, *D* (*dextro*) and *L* (*levo*). According to the standard convention, when the carbohydrate under consideration is positioned on the paper as shown earlier, the hydroxyl group attached to the asymmetric carbon in the dextro molecule points inward from the plane of the paper whereas that of levo points outward. In general, a molecule with n asymmetric carbon atoms has 2^n stereoisomeric forms. As discussed in the next section, different stereoisomers of a molecule might have completely different biological activities.

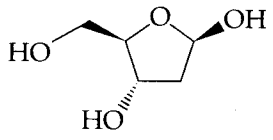
2.3 Carbohydrates

Carbohydrates are organic molecules constructed of carbon, hydrogen, and oxygen. Their chemical composition more or less follows the formula $C_n(H_2O)_m$, where n and m take the values of positive integers. As will be discussed in Chapter 3, carbohydrates provide chemical energy for powering biological processes. These macromolecules are metabolized in cells, and the energy contained in their covalent bonds is used in activities such as locomotion, cell division, and protein synthesis. Carbohydrates also have important functions as structural building blocks in organisms ranging from plants to bacteria. A subset of five-sugar carbohydrates forms part of the backbone of nucleic acids. Compounds that carbohydrates form with lipids and proteins play fundamental roles in protein trafficking and cell–cell recognition. Not all carbohydrates are sweet, but they are nevertheless called *saccharides* or sugars. Carbohydrates found in nature are mostly D-carbohydrates. Why the D-sugars are more abundant than L-sugars in living organisms is a mystery of molecular evolution.

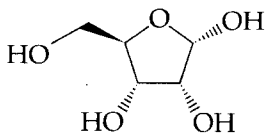
2.3.1 Monosaccharides

The simplest carbohydrates are monosaccharides, the monomers from which all carbohydrates are built. The smallest monosaccharide molecules contain three and the largest contain seven carbon atoms. The three-carbon monosaccharides such as D-glyceraldehyde are found in living cells as intermediaries in the chemical reaction pathways that harvest energy from food. The metabolic process of reducing the hydrocarbon bond energy into readily usable free energy is discussed in Chapter 3.

Ribose and *deoxyribose* are five-carbon sugars with important biological functions. The molecular formulas of these sugars are as follows:



D-deoxyribose



D-ribose

D-deoxyribose and D-ribose form parts of the backbones of the nucleic acids DNA and RNA, respectively. Both sugars are in the form of a five-membered ring. Because these sugars can undergo condensation reactions by linking at different molecular sites, a notation for identifying the carbon atoms in ring-like sugar molecules has been developed. The carbons in five-carbon sugars are identified using primed integer numbers 1′, 2′, 3′, 4′, and 5′. The values of these integers increase clockwise from the central oxygen atom as shown in Fig. 2.8.

The five-carbon sugars, ribose and deoxyribose differ from each other by one atom: An oxygen atom is missing from carbon 2′ in deoxyribose in relation to ribose. In other words, the hydroxyl group (—OH) attached to carbon 2′ of the ring structure in ribose is replaced by a hydrogen atom in deoxyribose. As we shall see

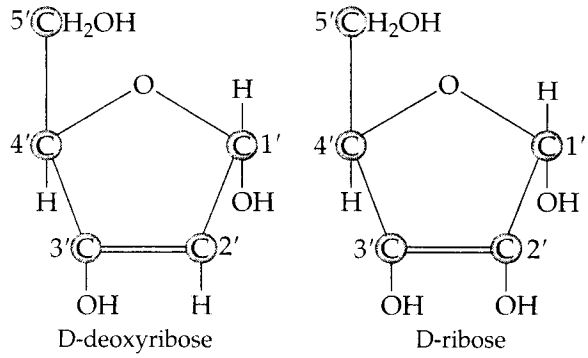
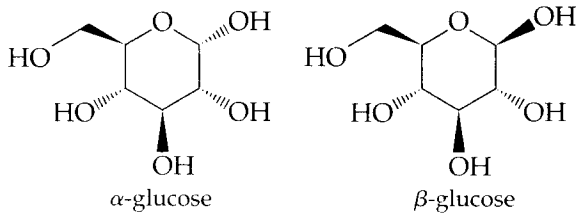


FIGURE 2.8 Notation used to identify carbon atoms in ring-like sugar molecules.

later in this chapter, this subtle difference in molecular formula is responsible for the significantly different functional roles RNA and DNA play in gene expression.

Let us next consider the six-carbon sugars. *D-glucose* is a six-carbon sugar with the molecular formula $C_6H_{12}O_6$. It is also called grape sugar. The common stereoisomers of glucose, α- and β-*glucose*, are given by the following structural formulas:



As indicated in the preceding diagram, these two forms differ from each other only in the spatial placement of the —H and —OH attached to carbon 1′. The ring structure forms α-glucose and β-glucose can be obtained from the chain form through the intermediate form shown in Fig. 2.9.

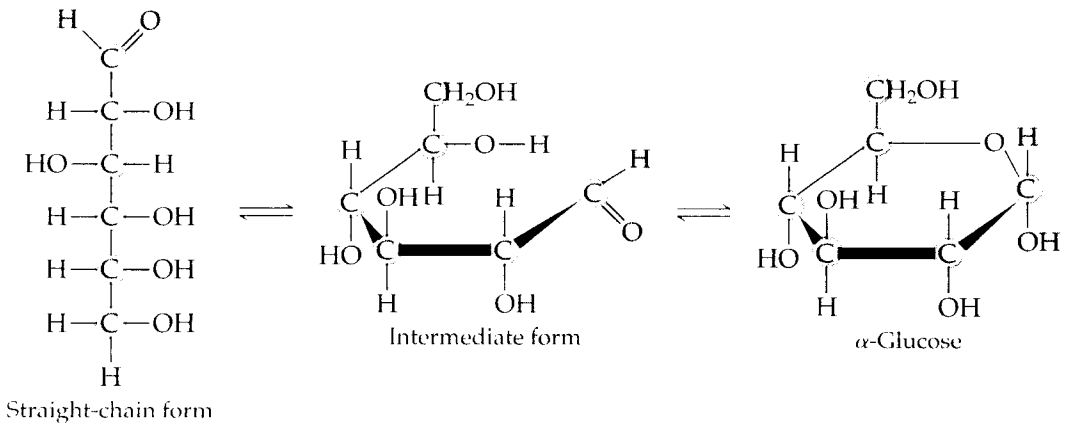
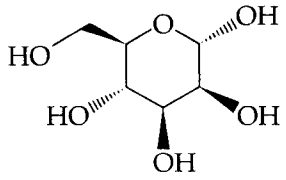
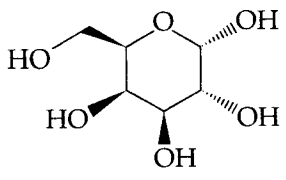


FIGURE 2.9 Various structural configurations of six-carbon sugar glucose. The ring form is the dominant form in an aqueous environment.

The predominant forms of sugars such as glucose in solution are the ring forms, not open chains. The six-carbon sugars have four asymmetric carbon atoms and therefore have 16 stereoisomers. These stereoisomers include the fruit sugar, *D-mannose*, a constituent of many glycoproteins, and *D-galactose*, the building block of the milk sugar, *lactose*:



D-mannose



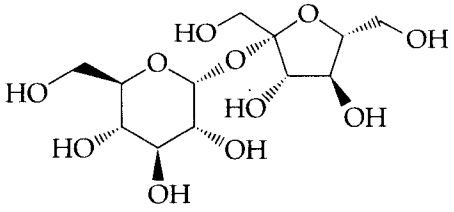
D-galactose

D-glucose is contained in all living cells. Green plants produce this sugar by photosynthesis, and other organisms acquire it from plants. Cells use glucose as a primary energy source and harvest its chemical energy through a series of energy-releasing reactions.

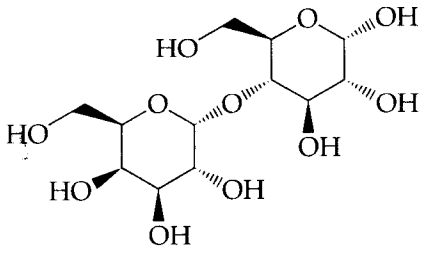
2.3.2 Disaccharides

Disaccharides are sugars composed of two monosaccharide residues united through a condensation reaction. Figure 2.10 illustrates the formation of maltose from α-glucose and β-glucose through a condensation reaction.

The many versions of disaccharides and their structural formulas are presented in the KEGG database. Among these sugars is *sucrose*, the most abundant disaccharide throughout the plant kingdom (also known as table sugar), and lactose, the milk sugar:



Sucrose



Lactose

The chemical properties of disaccharides depend on the nature of the linked monosaccharides, the carbon atoms involved in bonding, and the form of the linkage. Disaccharides comprise an important food source for all organisms. Enzymes

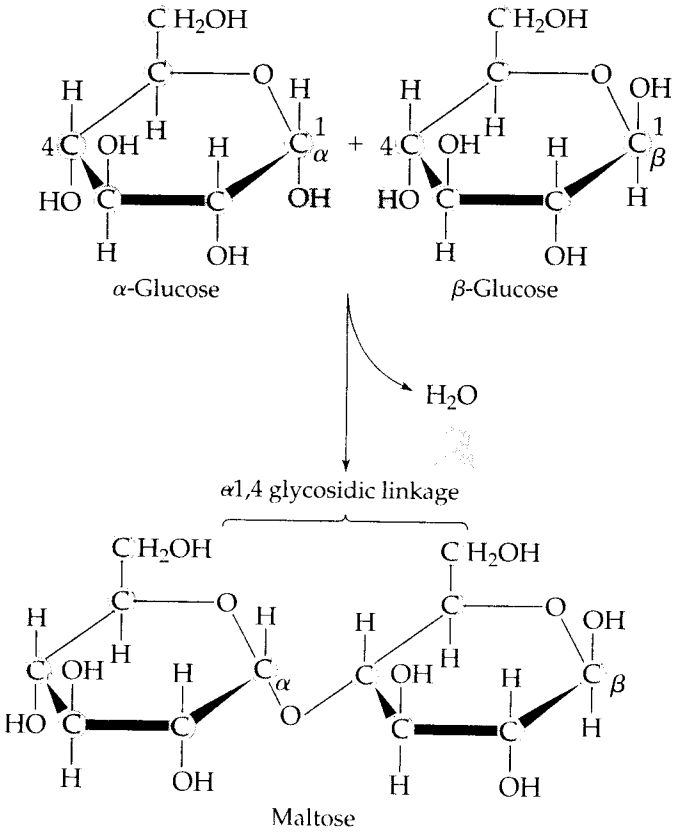


FIGURE 2.10
Condensation reaction for the formation of a disaccharide from two monosaccharides.

facilitate the degradation of these carbohydrates into smaller units to capture readily usable free energy.

2.3.3 Polysaccharides

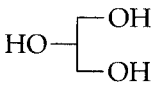
Polysaccharides are giant chains of monosaccharide residues. Because monosaccharides can bind to each other in a number of ways, they can form polysaccharides of incredible diversity. These large carbohydrates may be linear polymers or have many branches. *Starch* and *cellulose* are among the most abundant polysaccharides found in nature. Starch is a long chain formed by the linkage of α-glucose, whereas cellulose is made up of long chains of β-glucose. Although cellulose and starch have the same chemical composition, they have strikingly different chemical structures and physical properties. For example, starch is soluble in water, but cellulose is not. Humans use starch as a primary food for providing energy. In contrast, humans lack the enzymes that digest cellulose, the structural material of choice for plants. Cellulose comprises more than half of all organic carbon and therefore is the most abundant organic compound in the biosphere. A third important polysaccharide is *chitin*, the material from which insect skeletons are built. Chitin is an amino sugar (a sugar containing nitrogen) and forms long straight chains that serve structural roles very similar to the bone tissue of mammals. In general, polysaccharides are not limited

by a finite upper boundary. Although carbohydrates are not genetically prescribed, they are built and degraded in living organisms according to need through sequential actions of specific enzymes, which are regulated genetically.

2.4 Lipids

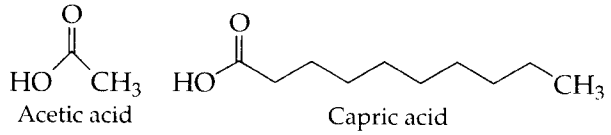
Lipids have the common property of being insoluble in water. Strictly speaking, lipids are not macromolecules, but compounds of molecules whose structure depends on hydrophobic interactions and van der Waals forces. Another common feature of lipids is their high hydrocarbon content. *Simple lipids* such as fats, oils, and waxes are well known in everyday life.

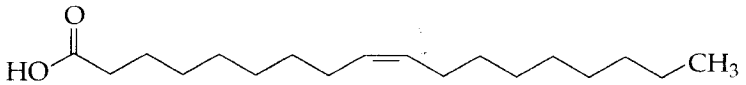
Simple lipids are composed of two types of building blocks: *glycerol*, a small three-carbon molecule with three hydroxyl groups, and three hydrocarbon chains called *fatty acids*. The molecular formula of glycerol in the KEGG compound notation is



Glycerol

Structural formulas of selected fatty acids in the KEGG compound notation are as follows:





Oleic acid

A molecule of glycerol and three fatty acid molecules react to form a lipid and three molecules of water. The formation of a lipid from glycerol and three fatty acids is illustrated in Fig. 2.11.

The physical properties of lipids are highly dependent on the type of fatty acids used as building blocks. The tail of a fatty acid can be of different lengths and be either *saturated* or *unsaturated*, depending on the presence of double carbon bonds.

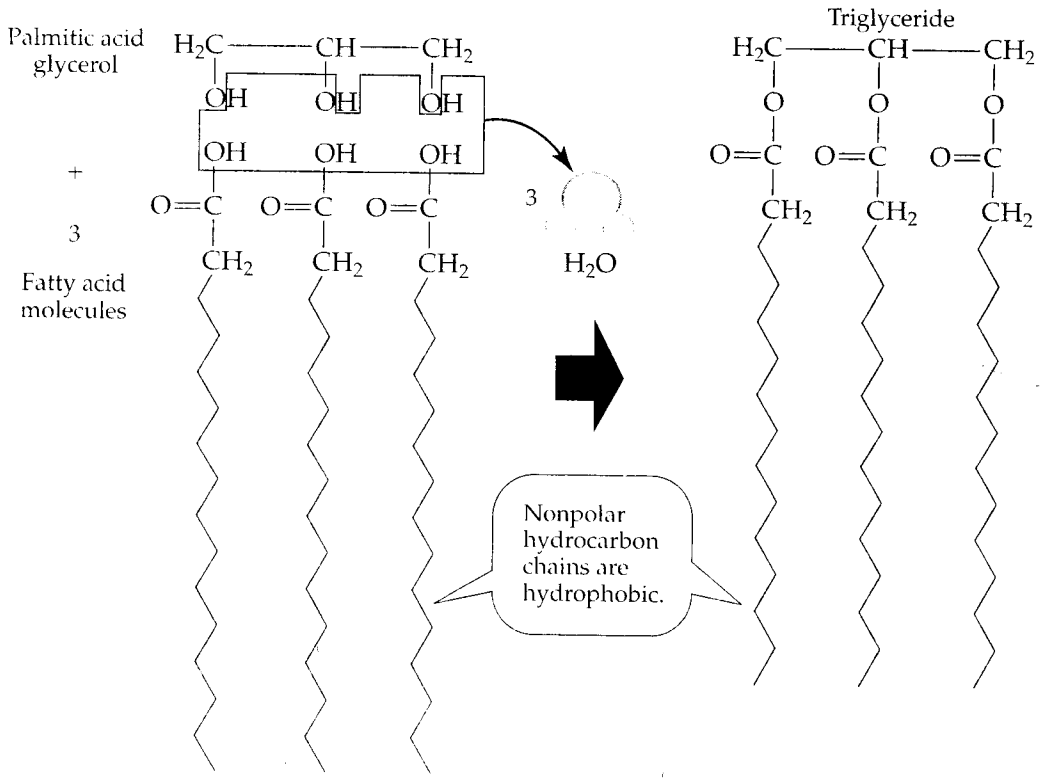


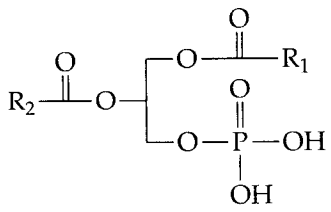
FIGURE 2.11 Condensation reaction leads to the formation of one lipid molecule from a glycerol molecule and three fatty acids.

The tail of a saturated fatty acid such as the capric acid shown earlier contains no double carbon bonds. The hydrocarbon chains of such saturated fatty acids are relatively rigid and straight, and they pack together tightly when forming fat. In *unsaturated fatty acids*, the hydrocarbon chain contains one or more double bonds. Each double bond causes a kink in the fatty acid, preventing the fatty acids from packing together tightly. The kinks associated with double bonds are important determinants of the fluidity and melting point of a lipid. Animal fats, which are solids at room temperature, tend to have long-chain, saturated fatty acids, whereas the liquid plant oils have short or unsaturated fatty acids. An example of unsaturated fatty acid is the oleic acid shown in the foregoing diagrams.

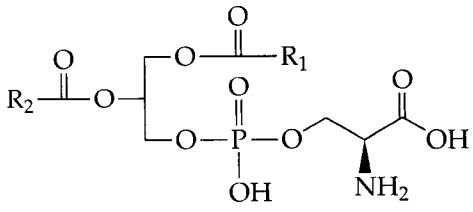
Phospholipids and Biomembranes

Phospholipids constitute the core elements of biological membranes. In the presence of low levels of detergents, phospholipids can also form other structures such as micelles and liposomes. Liposomes can be made in the presence of DNA and used as a vehicle for the delivery of genes in gene therapy. The lipid membrane can fuse with the plasma membrane of target cells and release the DNA into the cell. Like fats,

phospholipids have fatty acids bound to glycerol; however, one of the three fatty acids in the molecule is replaced by a phosphate group. The two fatty acids bound to glycerol form hydrophobic (nonpolar) chains, whereas the phosphorus-containing compound linked to the third —OH bond of the glycerol forms a hydrophilic head. Two examples of phospholipids are as follows:



Phosphatidic acid



Phosphatidylserine

Note that in these diagrams the terms R_1 and R_2 represent the fatty acids attached to glycerol. Structural diagrams of other phopholipids can be found in the KEGG compound database. Phospholipids have the important property of having one hydrophilic and one hydrophobic region. The phosphate group attached to the glycerol has a negative charge and is therefore hydrophilic. These globular subunits are called phospholipid heads. The fatty acid tails are hydrocarbon chains, have no charge, and are hydrophobic. Each phospholipid molecule can be visualized as consisting of a hydrophilic globular head and a long hydrophobic tail.

In the presence of water, phospholipid molecules cluster together and form bilayer surfaces that shield the nonpolar fatty acid tails from water (Fig. 2.12). The hydrophilic heads face outward on both sides of the lipid bilayer, where they interact with the surrounding water molecules, and the hydrophobic tails remain in the interior of the bilayer. Phospholipids are needed to create a biochemical environment in the cell that is different from the external environment. Cellular structures such as nuclei and mitochondria are also enclosed in lipid bilayers.

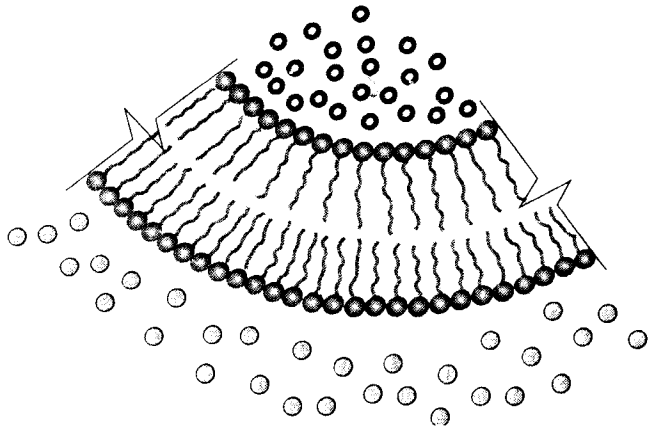
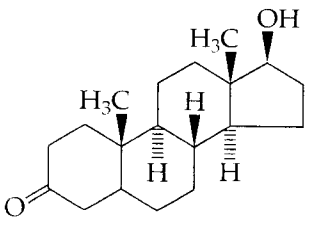


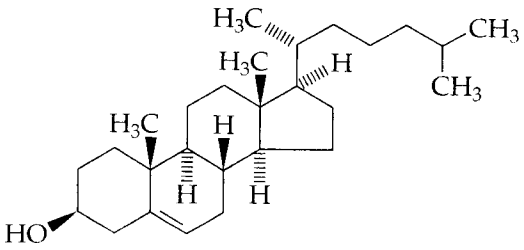
FIGURE 2.12 Formation of lipid bilayers from phospholipid molecules. Hydrophobic interactions between lipids and water lead to the organization of lipid molecules into bilayers. The globular heads of lipids face the water, whereas the hydrophobic tails that remain in the interior of the bilayer create a barrier against the passive diffusion of polar (water loving) molecules across the membrane. (Modified from http://www.tulane.edu/~biochem/faculty/facfigs/bilayer.htm.)

2.4.2 Other Subfamilies of Lipids

The lipid family of organic compounds has other biologically important subgroups: steroid hormones and steroids. These molecules are not composed of glycerol and fatty acids, but have ringlike structures similar to those observed in sugars. Unlike sugars, steroids consist mainly of hydrocarbons and are therefore hydrophobic. One important example of a steroid hormone is testosterone:



Testosterone



Cholesterol

This hormone is released into the blood stream from the testis and is involved in the development of male sexual characteristics. Testosterone is lipid soluble and therefore can diffuse across the plasma membranes of cells and enter the nucleus, where it can regulate gene expression.

Another biologically important lipid is cholesterol: Cholesterol can deposit on the inner surface of blood vessels, clogging the arteries and altering their mechanical properties. All steroids are synthesized from cholesterol and therefore have similar chemical structures.

Other types of lipids found in living systems include carotenoids (which trap light energy in plants), fat-soluble vitamins such as vitamins A and D, and glycolipids such as glucocerebroside. The structural formulas of these compounds as well as background information on them can be found in the KEGG database. Like carbohydrates, lipids are not directly specified genetically, but are formed from building blocks with the help of protein enzymes. However, the enzymes that catalyze the production of lipids are regulated genetically.

2.5 DNA and RNA

2.5.1 Deoxyribonucleic Acid (DNA)

DNA forms a double-stranded helix with a uniform radius and angle of twist (Fig. 2.13). Each strand is made of different combinations of the same four molecular beads, represented by the letters A, C, G, and T of the Latin alphabet. The beads in opposing strands complement each other according to base pairing combinations (A—T and C—G as shown in Fig. 2.13). The structure of DNA was discovered by

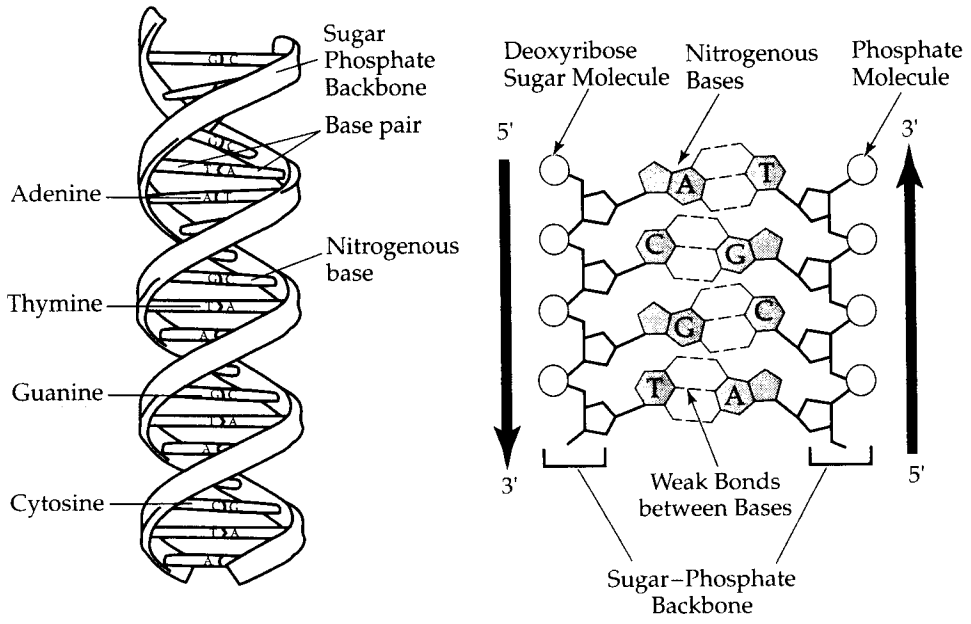
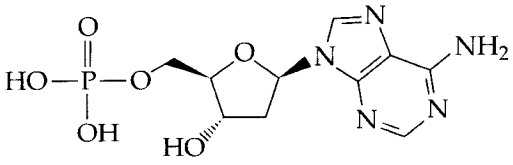


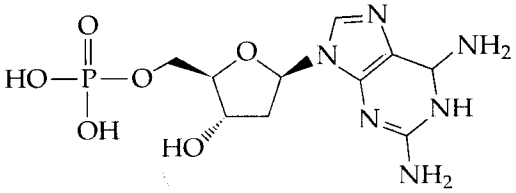
FIGURE 2.13 DNA forms a double-stranded helix with a uniform radius and angle of twist. The sugar–phosphate backbone forms the outer shell of the helix. The two strands of DNA run in opposite directions. Bases face toward each other and form hydrogen bonds. The types of base pairs found in DNA are restricted to those shown in the figure. (Modified from http://www.nhgri.nih.gov/DIR/VIP/Glossary/Illustration/basepair.html.)

James Watson and Francis Crick and was published in *Nature* in 1953. Their article is widely considered to signal the beginning of the modern molecular era in biology. Chemical properties of DNA and the resulting structural and functional features are discussed next.

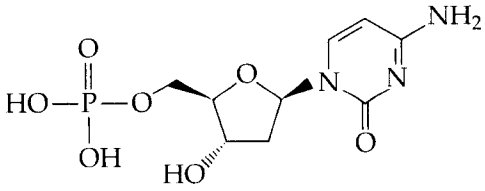
Building Blocks: Nucleotides of four different types constitute the building blocks of DNA. These nucleotides are called dAMP, dGMP, dTMP, and dCMP:



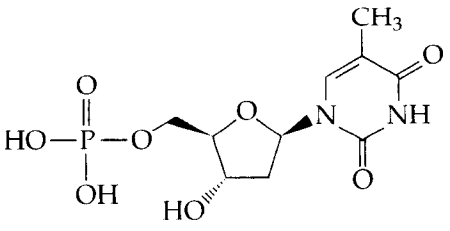
dAMP



dGMP

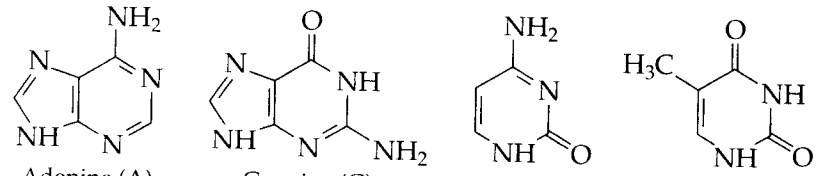


dCMP



dTMP

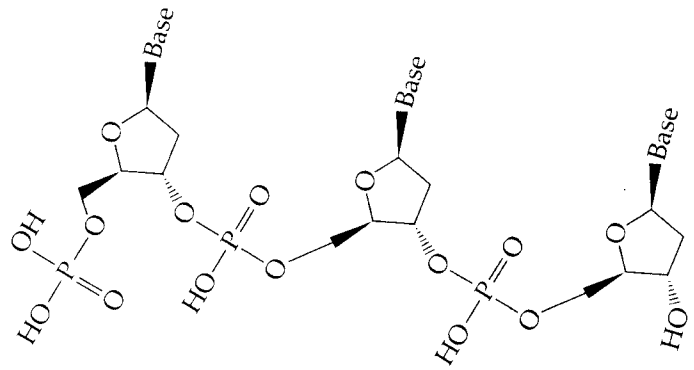
Each of these nucleotides is composed of three subunits: a phosphate group, a five-carbon sugar molecule (deoxyribose), and a nitrogenous base. The sugar and the phosphate group are identical in all four DNA nucleotides, but the nitrogen bases differ. The nitrogenous bases found in DNA are adenine (A) in dAMP, guanine (G) in dGMP, thymine (T) in dTMP, and cytosine (C) in dCMP:



Adenine (A) (purine) | Guanine (G) (purine) | Cytosine (C) (pyrimidine) | Thymine (T) (pyrimidine)

The nitrogenous bases adenine and guanine are composed of two rings. The corresponding nucleotides are called *purines*. The bases of the other two nucleotides, dCMP and dTMP, have single rings and are called *pyrimidines*. The name for the nitrogenous base of each DNA monomer identifies the monomer itself. Thus, a DNA strand is composed of monomers identified as adenine (A), guanine (G), thymine (T), and cytosine (C).

Backbone: Alternating subunits of sugars and phosphates form the backbone of a DNA strand. The sugar subunit of one nucleotide binds to the phosphate group of the adjacent nucleotide. This backbone is illustrated as follows:



Sense of Direction: The resulting DNA strand has a sense of direction. The phosphate group attached to the 5′ carbon of the sugar of one nucleotide attaches to the sugar subunit of the adjacent nucleotide at the 3′ carbon site as shown in the preceding diagram. As discussed previously in the section about carbohydrates, the symbols 3′ and 5′ refer to the third and fifth carbon atoms in the sugar ring, counted clockwise from the corner of the ring occupied by the O atom (Fig. 2.13). The resulting strand has at one end a bound OH, but no phosphate. This end is called the 3′ end. The other end of the DNA strand has a bound phosphate group. This end is

called the 5′ end and according to the convention used by biologists, the sequences of molecular beads that compose nucleic acids are listed from 5′ to 3′.

Complementary Base Pairs: DNA is a double-stranded helix with constant radius. The two strands of DNA twist like a screw, running in parallel, but opposite, directions. The nitrogen bases of DNA point toward the central axis of the helix. The two strands of the helical chain are held together by hydrogen bonding between monomers at the same positions in opposite strands. The same position refers to an identical number of monomers from one end of DNA. In particular, the monomer A on one strand always pairs with monomer T on the other strand at the same position. The monomers C and G are similarly paired. For this reason, the pairs (A, T) and (C, G) are called *complementary base pairs*. The constant diameter of the DNA helix is because these base pairs have identical physical dimensions in the direction normal to the axis of DNA. The complementary base pair rule assures that the information stored in DNA is in duplicate. The sequences of monomers on the opposing DNA strands are not identical, but complementary. For example, if one DNA strand has the sequence 5′ AACTTG 3′ at a certain location, the complementary strand will have the sequence 3′ TTGAAC 5′ at the same position. When a need arises to generate new copies of DNA (as occurs in cell division), each single-strand of DNA becomes a template to produce a complementary strand with a one-to-one correspondence between the two strands of DNA. The sequence of the DNA monomer is written either in capital or lowercase letters. For example, the sequences "ATTGCGGC" and "attgcggc" are identical. Capital letters are advantageous in distinguishing the DNA sequence from the rest of the text. Lowercase letters are used typically when large sequences from decoded DNA molecules are printed. For example, the following sequence is taken from one of the strands of a bacterial DNA molecule:

catctgtcggccataccacttcgcgcaacagatcgcccagcagtggggccgccagtgcagaaatccactgttcg
tcacgaaatccttcgcttaattgccgcactttgatggtcagtcgaaaactatcatcggaggggctacggcggaca
tatccctcttcctgcagcgtctccagcagtcgccgcacagtggtgcgatgcaggccgctgagttccgccagcagc
ccga

A survey of decoded DNA listed at the National Library of Medicine Gene Bank Web site shows that the content of the bases A, T, G, and C on DNA vary significantly from organism to organism. Because of base pairing, however, %A = %T and %G = %C when both strands of DNA are considered. In the human genome, G + C content is less than A + T content (Fig. 2.14). Although 5 percent of the genome has a G + C content of between 50 and 55 percent, this portion contains 15 percent of the genes.

2.5.2 Orientation of Genes along DNA

The hereditary information of an organism is distributed onto both strands of a DNA molecule. The molecular formulas of proteins synthesized by the organism constitute an important component of hereditary information. The term *gene* generally refers to those segments of one strand of DNA that are essential for synthesis of

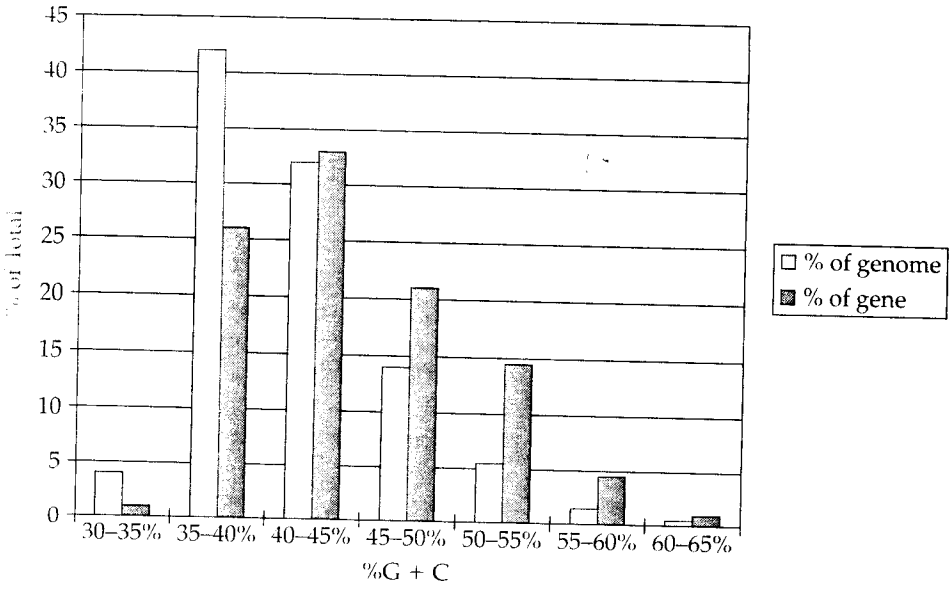


FIGURE 2.14 Distribution of G + C content in the human genome. (Modified from Venter et al., *Science* 2001, 291: 1304–1351.)

a functional protein or one of its major subunits (polypeptides). The term "gene" also refers to those segments of a DNA strand that encode a tRNA or rRNA molecule. As discussed later in this section, these two types of RNA molecules play pivotal roles in protein synthesis.

Figure 2.15 shows a schematic diagram of gene organization along both strands of DNA. All segments of a gene lie on the same strand of DNA. The strand on which the gene exists is sometimes called the sense strand and the noncoding strand the antisense strand.

The main features of gene position and orientation along DNA are discussed in the following list (a topic that is further expanded in Chapter 4, "Gene Circuits"):

1. In all organisms, the beginning of any gene faces the 3′ end of the DNA strand encoding it. Thus, when the DNA sequence of a strand is listed in the standard 5′ to 3′ direction, the sequence of a gene coded on the strand reads as in Arabic, from right to the left on the page of the text. In most instances, coding will be given of the complementary DNA strand (strand not containing the gene). Since the two strands have the opposite sense of direction, a regular 5′ to 3′ reading from the left to the right will give the complementary sequence of the actual gene.
2. The DNA strand segments that code for proteins are flanked by sequences that control the rate of their transcription. Depending on the type of cell, the production rates of gene products vary widely. The rate of *gene expression*, a term commonly used by biologists, refers to the rate at which a gene product is made. A gene is highly expressed when the rate at which it is transcribed and translated is high. Some genes are present in multiple copies in the genome further increasing the rate at which they can be made. Chapters 3–5 discuss the control mechanisms involved in the transcription of genes.

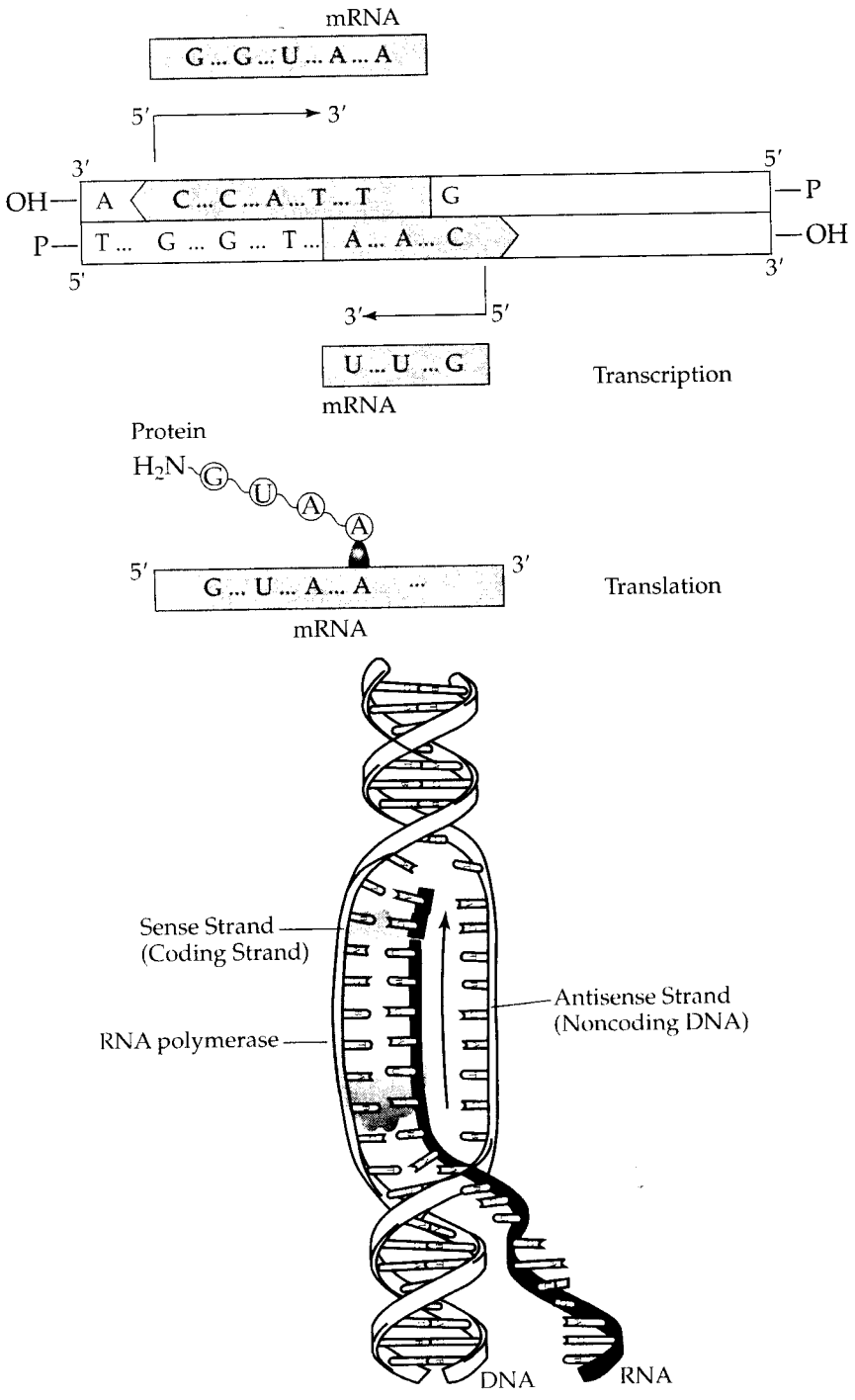


FIGURE 2.15 Schematic of the direction of genes along the two strands of a DNA molecule. The dark regions represent coding sections of DNA. The diagram also shows the two important steps in protein synthesis: transcription and translation. The figure on the bottom, depicting the event of transcription, is modified from http://www.nhgri.nih.gov/DIR/VIP/Glossary/Illustration/antisense.html.

DNA molecules from members of the same species vary slightly in their sequences of nucleotides. These variations often do not markedly affect the gene product and are known as polymorphisms. DNA differs more substantially across species. DNA harvested from members of hundreds of different organisms has been decoded. The sequence of DNA from a few individuals of the human race is also complete. Decoding the entire sequence of a human DNA is only the first step in genomic research. Among the billions of letters, scientists must identify those portions of the human DNA that actually encode protein-producing genes. Protein-coding regions make up about 1 percent of the human DNA; therefore, looking for the protein coding genes in the human genome is like looking for needles in a haystack. Scientists must not only identify the protein-coding sequences on DNA, but they must also sort out what these individual genes do. The genomic knowledge accumulated on simpler organisms can help identify the functions of many proteins found in higher organisms. Organisms such as yeast, fruit flies, and mice contain genes that are similar in sequence to those in humans. About 40 percent of all yeast genes have a human counterpart with interchangeable function. For example, many cancer-causing variants of human genes lead to uncontrolled growth of yeast colonies.

2.5.3 Ribonucleic Acid (RNA)

Like DNA, RNA molecules are chains of four molecular beads. These macromolecules play a fundamental role in protein synthesis. In the *transcription* phase of gene expression, the DNA sequence that ultimately encodes a protein is copied (transcribed) into RNA (Fig. 2.15). As noted previously, the resulting RNA molecule is called *messenger* RNA (mRNA). A single mRNA molecule may contain the instructions for a single protein, a protein subgroup, or a cluster of a few proteins. Transcription is tightly regulated to adjust the protein content of an organism to its needs.

The process of building proteins using mRNA molecules as templates is called *translation* (Fig. 2.15). The most critical step in translation is conducted by transfer RNA (tRNA). These molecules function as molecular adapters. The tRNA molecules match the sequences of beads on mRNA with the corresponding amino acids that make up the protein. Molecules in the third subgroup of RNA, ribosomal (r) RNA, associate with proteins to form ribosomes, which are factories for protein synthesis. Recent research has uncovered other types of RNA that play key roles in chromosome packing and gene expression. In the following, we present some of the important structural features of RNA and relate these features to their functions in living systems.

The structural and functional properties of RNA molecules are similar to those of DNA (Fig. 2.16). One chemical difference is that the deoxyribose sugar in DNA nucleotides is replaced by ribose in RNA, another five-carbon sugar. A second difference is that the DNA base thymine (T) is replaced by the base uracil (U) in RNA.

The building blocks of RNA are the nucleotides AMP, GMP, UMP, and CMP. Each of these nucleotides is composed of two nucleotide-invariant subunits (the phosphate group and ribose) and a nucleotide-dependent nitrogenous base. The bases found in RNA nucleotides are adenine (A), guanine (G), uracil (U), and cytosine (C).

Alternating sugars and phosphates form the backbone of RNA. This molecule is made of a single strand, but can bind in a complementary way to a DNA strand. In a DNA–RNA pair, the RNA strand runs in the opposite direction to the DNA

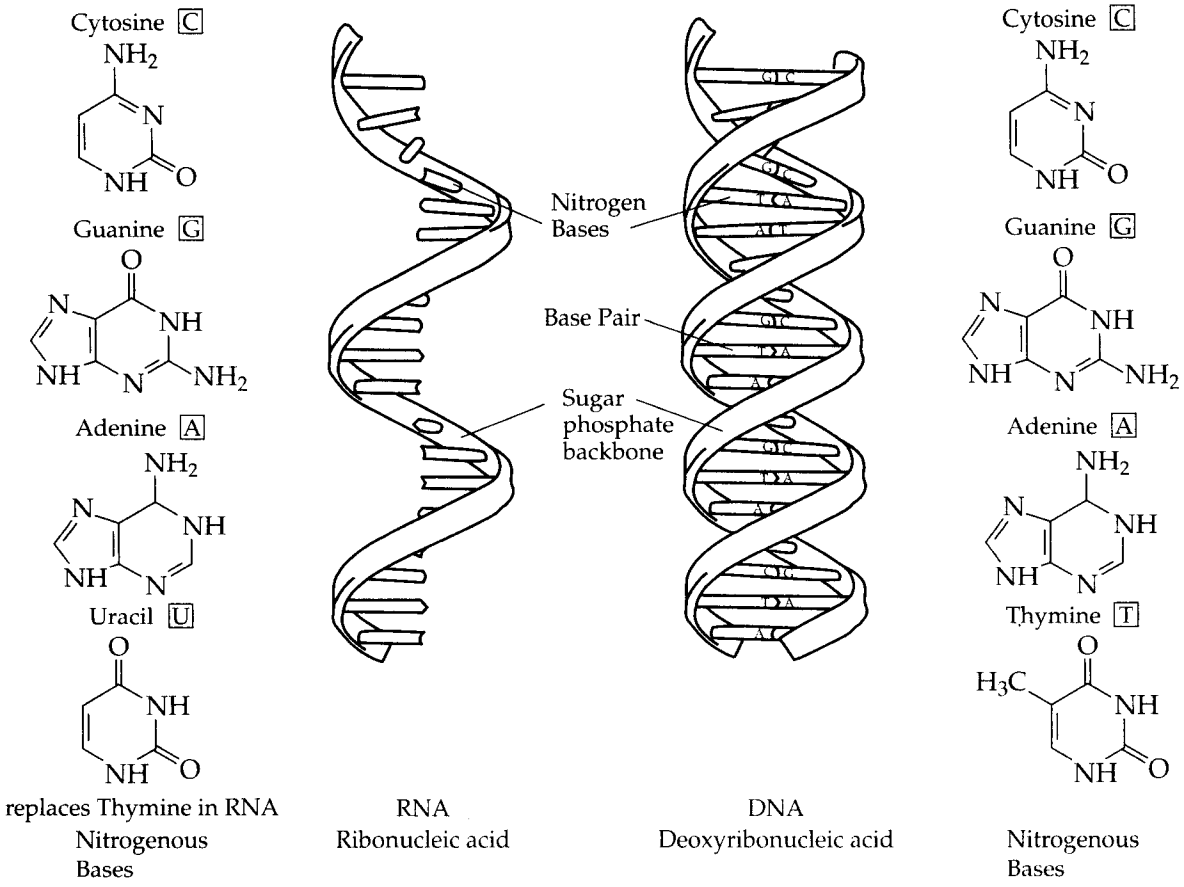


FIGURE 2.16 Comparison of RNA and DNA structures. (Modified from http://www.nhgri.nih.gov/DIR/VIP/Glossary/Illustration/RNA.html.)

strand such that the complementary base-pairing rule holds. The possible base pairs are (A, U), (T, A), (C, G), and (G, C). The complementary base pair rule between DNA and RNA assures that the information stored in DNA is accurately transcribed into RNA.

2.5.4 Types of RNA Molecules

Messenger RNA (mRNA) molecules are complementary copies of the instructions encoded on DNA (Fig. 2.15). As discussed in Chapter 3, mRNA is built by molecular machines with the use of a DNA strand as a template. The complementary-pair rule is illustrated by the following sequence:

3′ TACCGAAATGAGTGCGCTTA 5′ DNA

5′ AUGGCUUUACUCACGCGAAU 3′ mRNA

Whereas DNA contains the master copy of the genetic information, kept permanently on file, mRNA stores the working copy of this information for a brief period. After

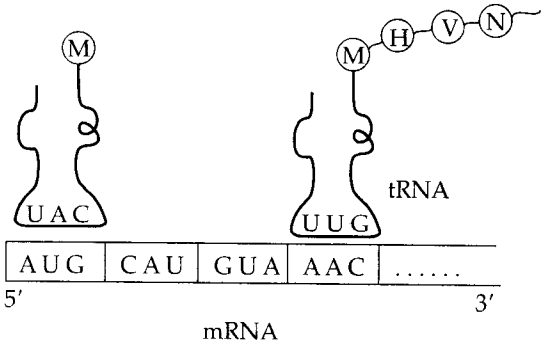


FIGURE 2.17 Schematic diagram of tRNA as a molecular adapter. The sequence of three nucleotides of tRNA that pair with a codon on mRNA is called an anticodon.

the instructions are implemented by the cell, mRNA is degraded into its constituent nucleotides, which are then available for use in the synthesis of other molecules.

Transfer RNA (tRNA) functions as a structural adapter that matches each three-nucleotide amino acid codon sequence of the mRNA, with the amino acid specified by this codon (Fig. 2.17). A tRNA molecule consists of 75 to 80 nucleotides. The 3′ end of a tRNA molecule attaches to the specified amino acid, and in the middle of the tRNA are three bases (anticodon) that constitute the point of contact with mRNA. This binding site interacts with mRNA through complementary-base pairing. As illustrated in Fig. 2.17, each amino acid-specific tRNA binds to its amino acid in the cytoplasm and attaches it to the growing chain of amino acids (polypeptide) at the location prescribed by the mRNA.

Unlike mRNA, tRNA genes are directly encoded by DNA. The number of distinct tRNA molecules in many organisms is less than the $4^3 = 64$ possible combinations of three letters chosen from an alphabet of four letters. This is because some tRNA molecules attach the same amino acid to the growing chain of a polypeptide when they bind to different codons. The use of more than one codon to specify an amino acid means that there is some redundancy in the genetic code. However, the codons they bind to all share a common property, namely, that the first two letters in the codons must be the same. As discussed later in the chapter, it is possible that the very first tRNA molecules to emerge had specific affinity to nucleotide couplets, not triplets. However, adapters designed for nucleotide couplets, recognize and bind to at most $4^2 = 16$ different types of amino acids, apparently a subset of not enough diversity to sustain life.

The three-dimensional shapes of tRNA molecules enable them to bind not only amino acids and mRNA, but also enable them to interact with ribosomes, the miniplants for protein synthesis. Information about the structures of macromolecules such as RNA comes from X-ray crystallography, which allows investigators to determine the positions of atoms in a crystalline substance by the diffraction patterns of X rays passed through a crystal. Figure 2.18 shows the backbone of the tRNA that acts as an adapter for the amino acid phenylalanine. Similar diagrams for many different tRNA molecules can be obtained by going to the National Library of Medicine Web site http://www.ncbi.nlm.nih.gov/ and pressing *Entrez*. Press *Structure*, type the keyword tRNA, and press *go*. The reader will then have access to the Web page for structure. The entry for the specific tRNA shown in Fig. 2.18 is represented by the PDB identification number "1EHZ." Pressing the PDB id number will take the reader to the structure summary page where the structure can be viewed by pressing *view*. The

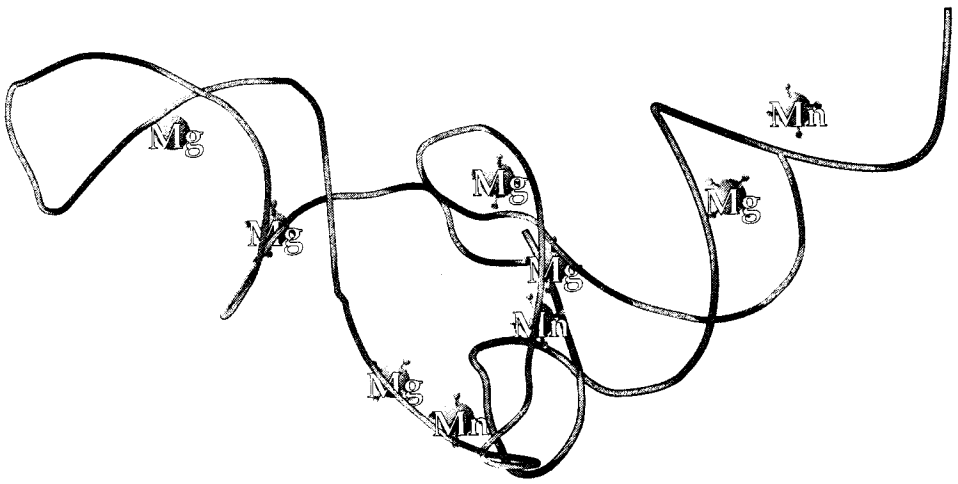


FIGURE 2.18 Crystal structure of yeast phenylalanine tRNA at 1.93 A Resolution. The tube-like curve in the figure represents the sugar-phosphate backbone of tRNA. This molecule is associated with mineral ions. (From http://www.ncbi.nlm.nih.gov/Structure/.)

program offers different modes of viewing the structure and the reader can investigate these various modes by clicking *style* and going through the options. The National Library of Medicine Web page also has a "*help*" section that provides information for structural investigations.

Ribomosal RNA (rRNA) molecules are also encoded by DNA. These molecules bind to specific proteins and form *ribosomes*. As discussed in detail in Chapter 3, several different types of rRNA molecules (ones with different lengths and sequences) are needed in the synthesis of ribosomes. Ribosomal RNA molecules contain thousands of nucleotides. Figure 2.19 shows the three-dimensional configuration of a 2900-nucleotides rRNA molecule found in the bacteria *Escherichia coli* (*E. coli*).

Some RNA molecules not only have complex three-dimensional shapes, but also have enzymatic activity. The existence of these "ribozymes" and other experimental data suggest that RNA emerged before proteins and DNA in evolution. The earliest catalysts of biochemical reactions might have been RNA molecules with complex three-dimensional shapes. Such molecules may have acted as templates both to produce DNA and to synthesize proteins. However, proteins, with their incredibly diverse shapes and physicochemical properties, eventually replaced RNA as enzymes.

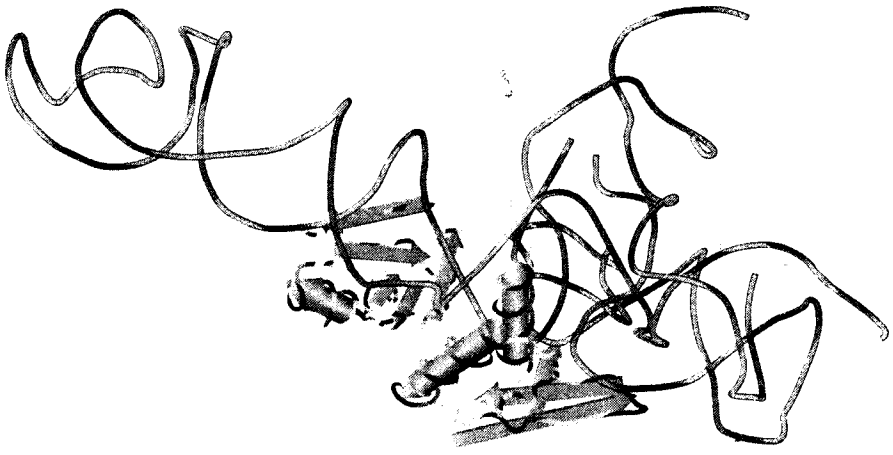


FIGURE 2.19 Structure of the RNA molecule 16s rRNA in the region around ribosomal protein S8 in *Escherichia coli*. A ribosomal RNA molecule (shown as a thick tube-like curve) interacts with protein subunits (helical cylinders, flat arrows, and connecting wire-like regions) in forming the ribosome. (Modified from http://www.ncbi.nlm.nih.gov/Structure/.)

The capacity of DNA to form double helices may have made it a more stable and attractive molecule than RNA for storing hereditary information.

2.6 Proteins

Protein networks guide the biochemistry of living cells. Metabolic and signaling pathways discussed later in the book are composed of networks of interacting proteins. As discussed in Chapter 9, elements of such networks have largely been identified in the yeast. The structure of protein networks, including connectivity of nodes as well as network logic, are currently the subject of intense research. In the following discussion, we focus on the common features of proteins, their composition, structure, and function.

All proteins in living systems follow a simple and universal blueprint: They are linear chains of amino acid residues (Fig. 2.20). Amino acids are a subset of organic molecules that serve as chemical messengers between cells or function as important intermediates in metabolic processes. Amino acids are composed of a central carbon atom with four subgroups attached to it: an amino group ($-NH_2$), a carboxyl group (—COOH), a hydrogen atom, and a distinctive side chain, represented by the letter R as shown in the following diagram:

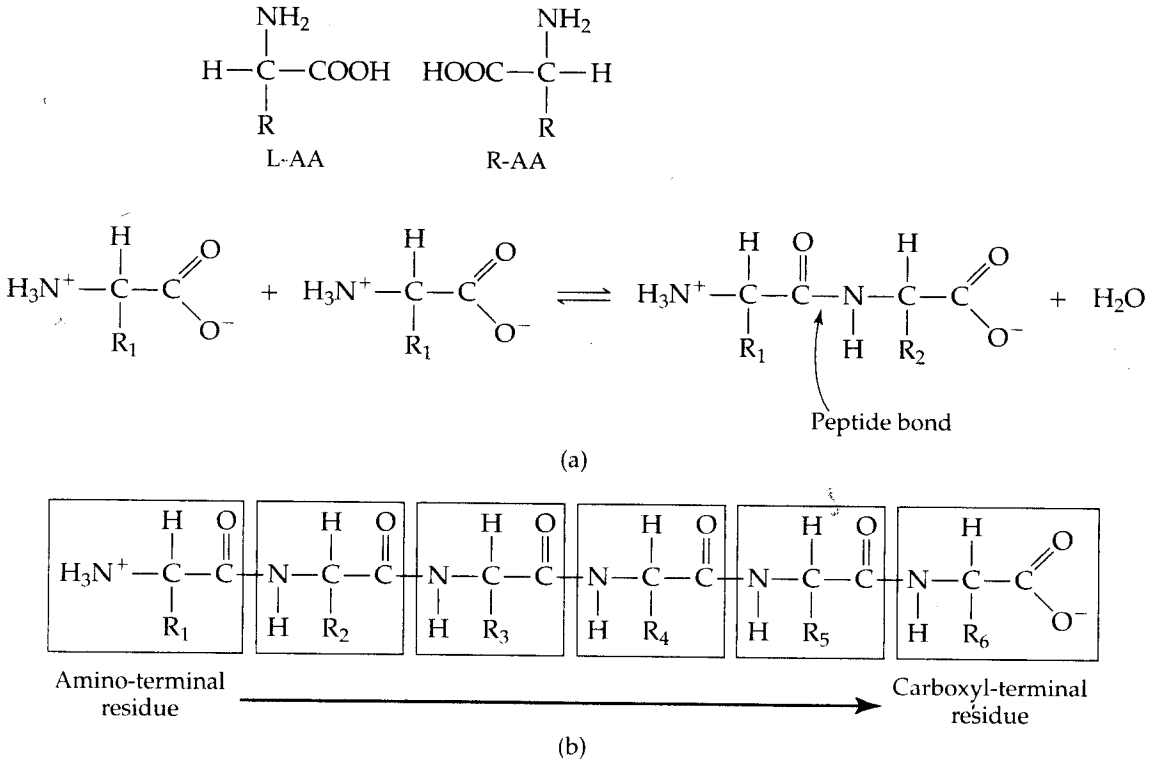


FIGURE 2.20 Synthesis of polypeptides from amino acids. Amino acids are composed of a central carbon atom with four subgroups attached to it: an amino group ($-NH_2$), a carboxyl group (—COOH), a hydrogen atom, and a distinctive side chain, represented by the symbol R.

The amino acids differ from each other only by their side chains (R). The tetrahedral array of four different groups about the central carbon atom yields two mirror-image forms: the L-isomer and the D-isomer. Only L-amino acids are found in proteins. D-amino acids are most widely found in bacterial cell walls.

About 300 amino acids are found in nature, but only 20 of these occur in proteins. Not every protein contains all of the 20 amino-acid types. Common elements in amino acids are carbon, hydrogen, oxygen, and nitrogen. In addition, all proteins have an amino acid that contains sulfur.

Serial linkage of amino acids in a condensation reaction results in proteins with intricate three-dimensional structures and a remarkable range of functions. The amino group of one amino acid is joined to the carboxyl group of the adjacent amino acid through the loss of one water molecule. This energy-requiring reaction is repeated over and over again in a growing polymer (polypeptide). An amino acid unit in a polypeptide is called a *residue*, and the covalent bond between two amino acid residues is called a peptide bond.

The peptide bonds formed by the amino and the carboxyl groups of amino-acid residues form the backbone of polypeptides and the distinctive *side chains* are exposed. Therefore, the side chains give the polypeptides their distinguishing properties. Polypeptides are linear polymers; that is, each amino acid is linked to its neighbors in a head-to-tail fashion rather than forming branched chains.

A polypeptide chain has a sense of direction such that one end has a terminal amino group and the other a terminal carboxyl group. Polypeptides grow from the terminal amino group toward the carboxyl terminal group. Thus, the sequence of amino acids in a polypeptide chain is written starting with the amino terminal residue. The first amino acid in most polypeptides is the sulfur-containing amino acid, methionine, which is identified with the letter M. Each of the 20 amino acids is represented by a letter in the Latin alphabet. A typical amino acid sequence for a protein composed of a single polypeptide reads as follows:

MLLVLVLIGLNMRPLLTSVGPLLPQLRQASGMSFSVAALLTALPVVTMGG
LALAGSWLHQHVSERRSVAISLLLIAVGALMRELYPQSALLLSSALLGGVG
IGIIQAVMPSVIKRRFQQRTPLVMGLWSAALMGGGGLGAAITPWLVQHS
ETWYQTLAWWALPAVVALFAWWWQSAREVASSHKTTTTPVRVVFTPRA
WTLGVYFGLINGGYASLIAWLPAFYIEIGASAQYSGSLLALMTLGQAAGA
LLMPAMARHQDRRKLLMLALVLQLVGFCGFIWLPMQLPVLWAMVCGL
GLGGAFPLCLLLALDHSVQPAIAGKLVAFMQGIGFIIAGLAPWFSGVLRSI
SGNYLMDWAFHALCVVGLMIITLRFAPVRFPQLWVKEA

Many proteins are single polypeptides. The shape and function of such proteins is then directly related to the primary sequence of amino acids present in the polypeptide. Other proteins are composed of multiple polypeptides that come together to form a complex. In some cases, each of these polypeptides is encoded by a single gene; in others, multiple genes may be involved.

2.6.1 Chemical Properties of the Twenty Amino Acids Found in Proteins

The 20 amino acids found in proteins vary in size, charge, and their capacity to form hydrogen bonds with other molecules. This variation is an important determinant

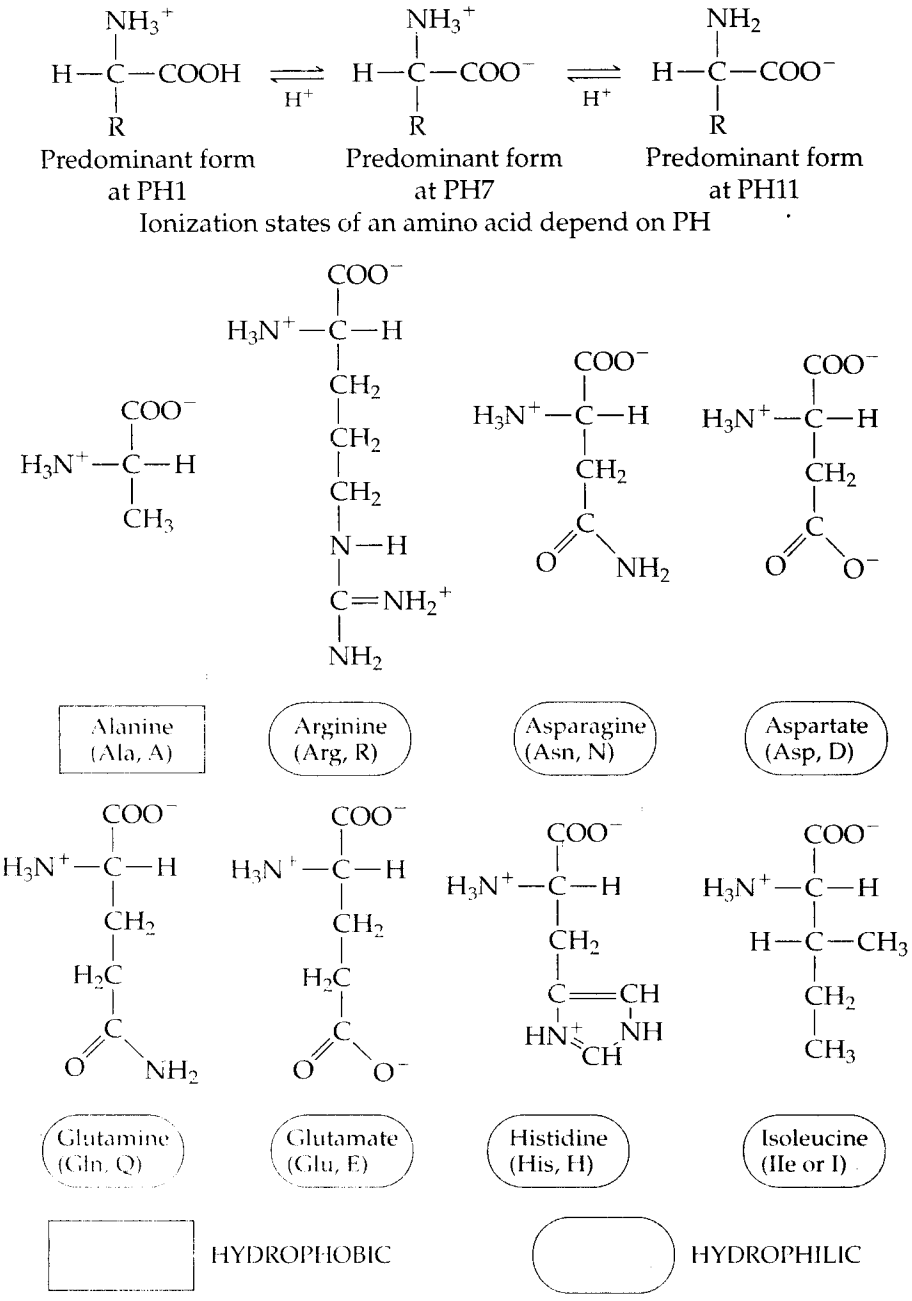


FIGURE 2.21 Structural diagrams and chemical classification of the 20 amino acids found in proteins.

of the diversity found in proteins. Figure 2.21 shows the structural diagrams of the 20 amino acids found in proteins.

Amino Acids with Hydrocarbon Chains: The simplest amino acid of the 20 found in proteins is *glycine* in which the side chain is a single hydrogen atom. As with all amino acids, glycine has a three-letter and single-letter representation (gly, G).

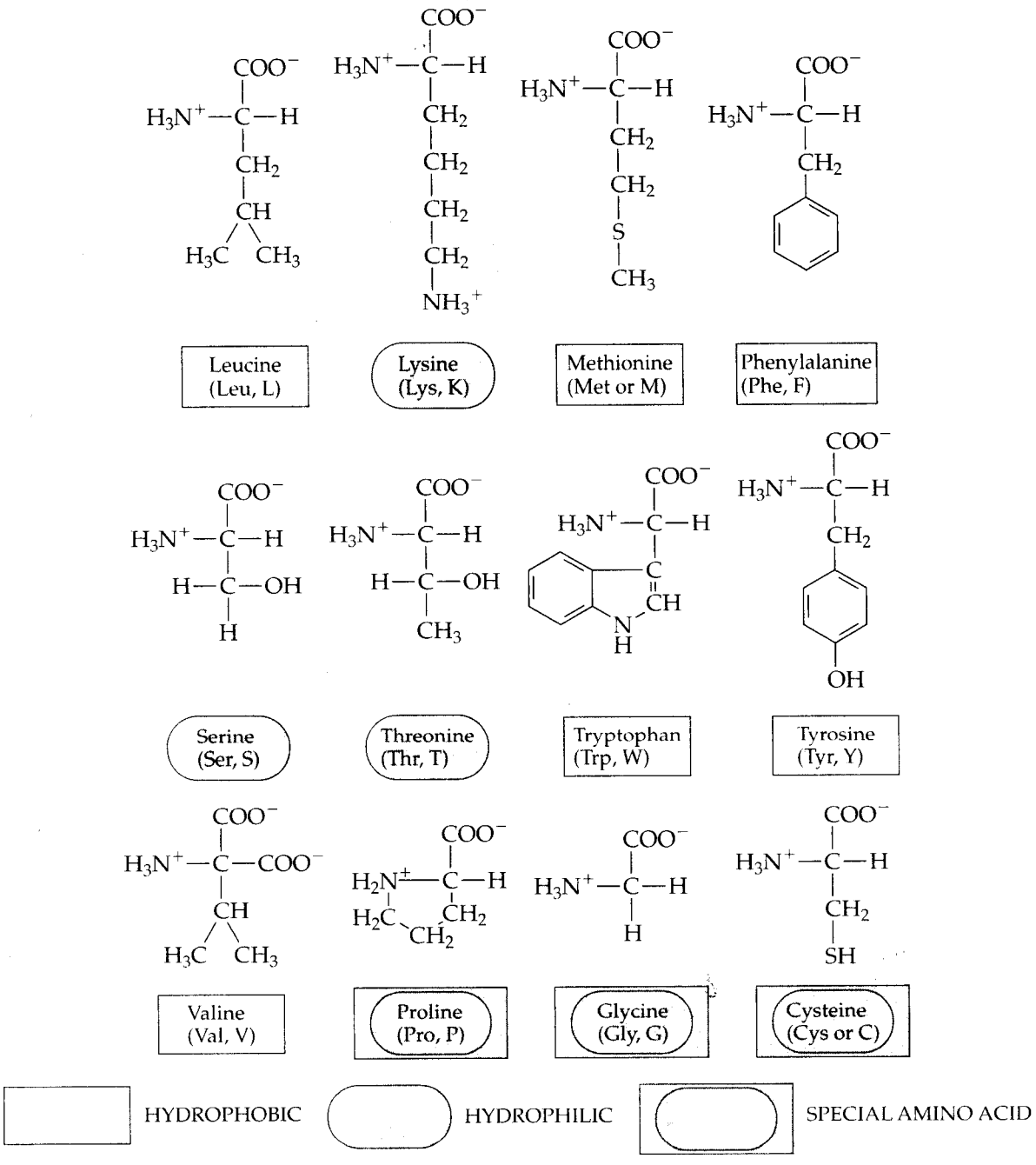


FIGURE 2.21 *(Continued)*

Glycine fits into tight corners in the interior of a protein molecule. Next in simplicity is *alanine* (ala, A), with a methyl group (CH_3) as its side chain. Larger hydrocarbon side chains, three or four carbons long, are found in *valine* (val, V), *leucine* (leu, L), and *isoleucine* (ile, I). These hydrocarbon chains are all hydrophobic. The different sizes and shapes of these hydrocarbon side chains enable them to pack together to form compact structures with few holes exposed to water. Regions of proteins rich in these hydrophobic residues often interact with lipid-containing membranes. *Proline* (pro, P) is often found in the bends of folded protein chains. Its hydrocarbon chain contains three carbon atoms, but unlike alanine, valine, leucine, and isoleucine, the side chain is bound to both the central carbon atom and the nitrogen atom. This configuration makes proline very rigid and its presence often creates a kink in a polypeptide chain.

Aromatic Amino Acids: Amino acids *phenylalanine* (phe, F), *tyrosine* (tyr, Y), and *tryptophan* (trp, W) have *aromatic side chains* in the form of rings. Tryptophan contains a nitrogen atom in its side chain. Phenylalanine and tryptophan are strongly hydrophobic. The side chain of tyrosine contains a hydroxyl group, which makes it less hydrophobic. This hydroxyl group is a potential site of addition of a phosphate group, a common post-translational modification of certain proteins.

Amino Acids Containing Sulfur: A sulfur atom is present in the side chains of *cysteine* (cys, C) and *methionine* (met, M). Both of these sulfur-containing side chains are hydrophobic. The side chain of cysteine is highly reactive, a property that enables cysteine to play a special role in shaping some proteins through disulfide links. Cysteine residues at different regions of a protein bind to each other and thus create folds and domains in the geometry of the protein.

Water-Loving Amino Acids: Amino acids *serine* (ser, S) and *threonine* (thr, T) are a hydroxylated version of alanine and valine, respectively. Replacement of a hydrogen atom in the side chain with a hydroxyl group leads to more reactive and more water-loving (hydrophilic) amino acids. Like tyrosine, these hydroxyl residues are potential sites of phosphate addition. *Lysine* (lys, K), *arginine* (arg, R), and *histidine* (his, H) possess polar side chains that contain nitrogen and are highly hydrophilic. The side chains of arginine and lysine are the longest of the 20 amino acids and are normally positively charged. Histidine can be uncharged or positively charged, depending on its environment. This amino acid is often found in the active sites of enzymes, where it can readily switch between these states to catalyze the making and breaking of bonds.

Aspartate (asp, D) and *glutamate* (glu, E) are polar and have negatively charged acidic side chains, carboxyl groups, at physiological pH. Uncharged derivatives of glutamate and aspartate are *glutamine* (gln, Q) and *asparagine* (asn, N), which contain a terminal amide group in place of a carboxylate. These amino acids are also polar molecules and the amide group of asn is a potential site of addition of sugar residues.

Bacteria can synthesize each of the 20 amino acids found in proteins using a carbon source and ammonium ions that exist in water. Plants use various simple nitrogen compounds and carbohydrates to make amino acids. In contrast, animals can synthesize only some of their amino acids using sugars and ammonia as starting materials. One could speculate that as higher organisms became more complex, they became more and more dependent on organic food. The enzymes used in synthesis

of some amino acids were used more and more infrequently because of their availability in foodstuff. As a result, the genes for these enzymes became nonfunctional over the course of time. Amino acids that humans cannot synthesize are called *essential amino acids*. Out of the eight essential amino acids, six are hydrophobic. The amino acids with large hydrocarbon side chains (valine, leucine, and isoleucine), amino acids with aromatic side chains (phenylalanine and tryptophan), and the sulfur-containing methionine are in this group. The two essential hydrophilic amino acids are threonine and lysine. Essential amino acids must be obtained through the diet. Meat, fish, milk, and eggs contain all of the amino acids needed in making human proteins. Legumes, grains, and other plant sources typically contain only a partial set of essential amino acids. The needed amino acids can be obtained from plant sources by combining certain foods. Beans, for example, provide isoleucine and lysine, whereas rice contains adequate amounts of other essential amino acids.

2.7 The Genetic Code

The nucleotide sequence of mRNA determines the sequence of amino acid residues of the protein it encodes. The set of relations that map mRNA coding onto amino-acid sequences of proteins is called the *genetic code*. The code provides one specific answer to the following question:

How do sequences of four letters (A,T,G,C) specify 20 distinct entities?

Clearly, single-letter and two-letter combinations do not provide enough choices to specify 20 amino acids. There are only $4^2 = 16$ distinct sequences of two letters in a four-letter language. Three-letter combinations, with each letter chosen from a pool of four, on the other hand, lead to $4^3 = 64$ distinct words, which is more than sufficient to code each of the twenty amino acids. Indeed, mRNA consists of a linear sequence of such three-letter words called *codons*. Table 2.1 shows the 64 mRNA codons and the entities they specify. Protein-coding genes all begin with a START codon and terminate with a STOP codon. The START codon specifies the sulfur-containing amino acid methionine (M).

Multiple nucleic acid codons correspond to the same amino acid. Amino acids represented by six codons are arginine (R), leucine (L), and serine (S). On the other hand, amino acids methionine (M) and tryptophan (W) are represented by single codons each. Codons specifying the same amino acid are said to be *synonymous*. Because the genetic code is not a one-to-one mapping between RNA and amino acids, it is called *degenerate*.

Table 2.1 indicates the following properties for the genetic code:

1. The first two letters in a codon are primary determinants of amino-acid identity. For example, all codons that begin with GU specify the amino acid valine (V) and are therefore synonymous. Similarly, codons that begin with GG specify glycine (G).
2. Codons that have U or C as the second nucleotide tend to specify the more hydrophobic amino acids. For example, codons beginning with GU and GC specify valine and alanine, respectively.

TABLE 2.1

Standard Genetic Code: Mapping between mRNA Codons and Amino Acids

UUU F Phe	UCU S Ser	UAU Y Tyr	UGU C Cys
UUC F Phe	UCC S Ser	UAC Y Tyr	UGC C Cys
UUA L Leu	UCA S Ser	UAA STOP	UGA STOP
UUG L Leu	UCG S Ser	UAG STOP	UGG W Trp
CUU L Leu	CCU P Pro	CAU H His	CGU R Arg
CUC L Leu	CCC P Pro	CAC H His	CGC R Arg
CUA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CUG L Leu	CCG P Pro	CAG Q Gln	CGG R Arg
AUU I Ile	ACU T Thr	AAU N Asn	AGU S Ser
AUC I Ile	ACC T Thr	AAC N Asn	AGC S Ser
AUA I Ile	ACA T Thr	AAA K Lys	AGA R Arg
AUG M, START	ACG T Thr	AAG K Lys	AGG R Arg
GUU V Val	GCU A Ala	GAU D Asp	GGU G Gly
GUC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GUA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GUG V Val	GCG A Ala	GAG E Glu	GGG G Gly

3. Codons that differ in the third nucleotide specify the same amino acid if the third nucleotide is either U or C. For example, histidine (H) is specified by codons CAU and CAC.

Slight differences from the standard genetic code just described occur in a few organisms. These differences tend toward even less dependence on the third letter of the codon to specify the identity of an amino acid. Let us next illustrate the flow of information from DNA to protein by considering the following DNA sequence representing the beginning segment of a protein-coding gene:

3′ TACTTGCAAATG 5′ DNA template

The complement mRNA sequence is as follows:

5′ AUGAACGUUUAC 3′ mRNA

Table 2.1 relates the mRNA codons AUG, AAC, GUU, and UAC to their corresponding amino acids, and the resulting peptide sequence is MNVY. The fact that both DNA and mRNA are specified in Gene Banks in the 5′ to 3′ direction might cause confusion in translating the codons on DNA to mRNA. DNA coding, when written according to the standard 5′ to 3′ direction, will be different from that in 3′ to 5′ direction:

3′ TACTTGCAAATG 5′	DNA template
5′ GTAAACGTTCAT 3′	DNA as read from 5′ to 3′
5′AUGAACGUUUAC 3′	mRNA read from 5′ to 3′
MNVY	encoded amino acids

2.8 Protein Structure and Function

Protein folding is one of the most intriguing subjects of computational and experimental biology because of its significance in the design and discovery of new drugs. All proteins adopt a single unique confirmation (folded state). This folded state is encoded in the amino-acid sequence of the protein and is called the native state. The number of possible three-dimensional configurations of polypeptide chains increases geometrically with the length of the polypeptide. However, because protein folding occurs through a pathway that favors only a few intermediate steps, the native state is stable and the misfolding of proteins is rare. Furthermore, as discussed in Chapter 3, cellular systems prevent the formation of misfolded proteins.

The structure of the native state of a protein largely determines its function in a living system. Proteins are described and compared by depicting four levels of structure. The levels of protein structure that affect function are illustrated in Fig. 2.22.

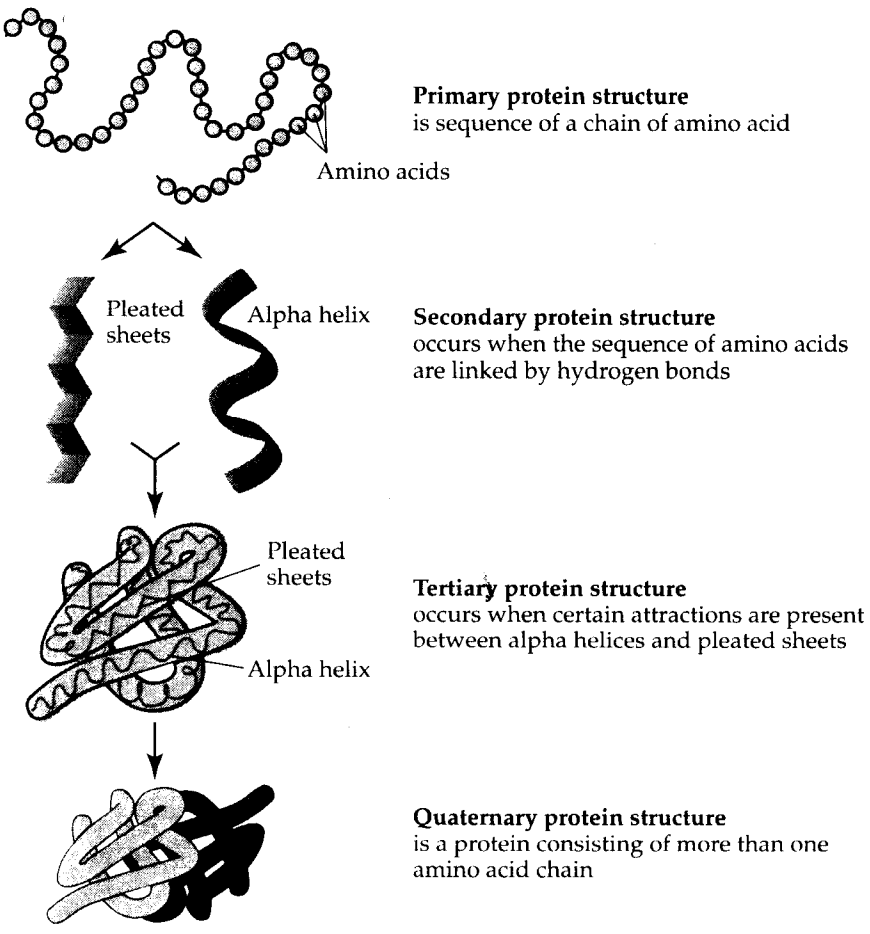


FIGURE 2.22 Structural levels of proteins. (Modified from http://www.nhgri.nih.gov/DIR/VIP/Glossary/Illustration/protein.html.)

The *primary structure* of a protein is its amino-acid sequence. It provides a complete description of the covalent connections of a protein. The primary structure is encoded on DNA, and the structural similarity of proteins can be assessed effectively by comparing amino-acid sequences. The *secondary structure* is a set of patterns concerning the spatial arrangement of amino acids that are near one another in the linear sequence. Secondary structures provide clues regarding the location of protein sites for interaction with other proteins and ligands. Figure 2.23 shows typical secondary structures such as *α-helix* and *β-pleated sheet* on the scallop myosin protein. The reader can obtain the structural diagram shown in the figure by going to the National Library of Medicine Web site http://www.ncbi.nlm.nih.gov/, pressing *Entrez*, pressing *Structure*, typing the keyword "myosin," and then pressing *go*. The particular myosin molecule shown in the figure has the PDB Id *1DFL*. Myosin is a very important contractile protein responsible for muscle contraction and is composed of six polypeptides: two identical polypeptides called heavy chains (A and B) and four shorter polypeptides called light chains (Y, Z, W, X). The *α-helical structure* is a corkscrew-like right-handed coil, with side chains extending outward from the peptide backbone of the helix. The α-helical structure is represented in protein graphics using a circular cylinder. An α-helical structure can be stretched, as the stretching requires only the breaking and rearrangement of hydrogen bonds; no covalent bonds are affected. When the tension in the helix is released, both the helix and the hydrogen bonds reform. Such behavior is called elastic in the field of mechanics. Fibrous structural proteins such as keratins are organized in α-helices. Keratin is found in fingernails, skin, and hair in humans. *β-pleated sheets* form when the protein chains extend and lie next to one another, forming flat sheets. This formation is also due to hydrogen bonding between the elements of the peptide linkages. A β-pleated sheet is shown with a flat arrow in protein graphics. The sense of direction of the arrow is from the beginning of the protein (the N-terminal) to its carboxyl terminal. Many proteins contain regions of α-helix and β-pleated sheet in the same polypeptide chain.

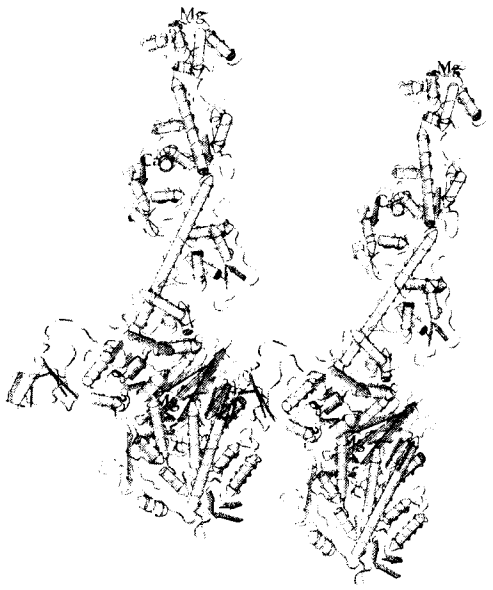


FIGURE 2.23 The three-dimensional molecular graphics of scallop myosin I. Myosin is the molecular machine supporting muscle contraction. Each head is composed of two different light chains as well as part of a heavy chain. The diagram indicates the presence of secondary structures such as α-helix and β-pleated sheets. The α-helix is shown as a circular cylinder decorated with helix and β-pleated sheet as a flat arrow pointing toward the carboxyl end of the polypeptide. The structure of the protein can be further studied by rotating the figure provided by the National Library of Medicine's Web page http://www.ncbi.nlm.nih.gov/Structure/.

Tertiary structure refers to the spatial arrangement of amino acids that are far apart in the linear sequence. The dividing line between secondary and tertiary structure is not precise. What is close and what is distant have not been defined in mathematical terms. However, in most proteins, a polypeptide chain is bent at specific sites and folded back and forth, enabling the interaction of amino acid residues that are far apart in the linear sequence. When cysteine residues from distant regions interact, they form disulfide bonds. A complete description of the tertiary structure of a protein requires the spatial location of every atom in the molecule in three-dimensional space. The identification of secondary and tertiary structures in a protein provides clues regarding its active (binding) sites and enable comparison with the reactive sites of other proteins. The MMDB structure summary for the scallop myosin shown in Fig. 2.23 identifies the regions of tertiary structures with symbols such as A.1, A.2, Y.6, and X.18. Pressing one of these symbols leads to the amino acid sequence of the tertiary structure.

Proteins containing more than one polypeptide chain exhibit an additional structure. *Quaternary structure* refers to the spatial arrangement of such subunits and the nature of their contacts. Hydrophobic interactions, hydrogen bonds, and ionic bonds all help hold the multiple polypeptide chains together to form a functional protein complex. In the scallop myosin example given earlier, the quaternary structure refers to the six polypeptide chains that compose the protein: A, Y, Z, B, W, and X. Pressing any of these symbols, one can obtain the corresponding amino-acid sequence. The National Library of Medicine has developed a database called VAST for assessing the structural similarities of various proteins and it is quite an effective tool in biotechnology because structural similarity often implies functional similarity. Manipulation of protein structure through changes in amino-acid sequence is a tool in modern drug design.

To accomplish all the fundamental functions that proteins carry out, many different types of protein are needed. A typical human cell contains about 100 million proteins of about 10,000 types. These cells all possess the same set of protein-coding genes (about 30,000), but different cell types (muscle cell, liver cell, neuron, and so on) express different subsets of these genes. The number of different types of proteins manufactured by an organism roughly correlates with the complexity of the organism itself. A typical gene in a vertebrate is composed of short sequences (exons) separated by long noncoding sequences (introns). Various spatial combinations of these genes correspond to different proteins. For example, a protein could be coded by exons 1 to 7, and another protein by exons 1 to 6 and 8 to 10 of the same gene (Fig. 2.24). Thus, a gene can code for multiple proteins in higher forms of life.

Many of the proteins made by humans and other vertebrates are similar in composition to those in simpler forms of life. Some of these proteins have additional domains that enable them to fulfill additional tasks (Fig. 2.23b). The enormous diversity in the types of protein arises primarily from the varying sequence of amino acids in the polypeptide. Permutations in the primary sequence allow for 20^r different protein types with r number of amino acid residues in the polypeptide. Protein lengths in most organisms range from 50 amino acids to tens of thousands. A typical protein molecule consists of a single polypeptide chain of about 100 residues. As discussed, some proteins have one or more polypeptide chains. These polypeptides can contain as many as 14,000 amino acids. The largest protein complex known has more than 40 separate polypeptide chains. Complicating the portrait of proteins is

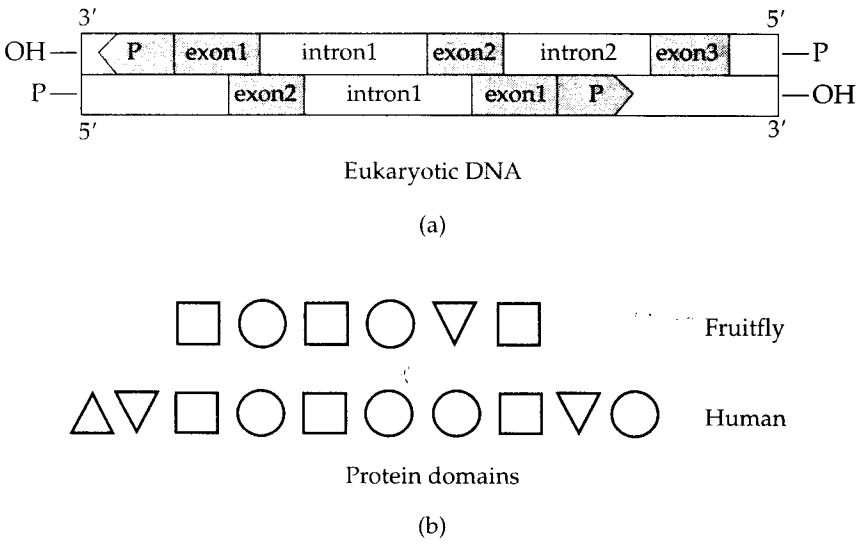


FIGURE 2.24 Schematic drawing of a human gene (a). The noncoding sections (introns) are much longer than the coding sections (exons). The existence of multiple exons for a gene leads to multidomain proteins in human and other mammals (b).

the presence of some proteins with carbohydrate, lipid, phosphate, and other types of attachments. These attachments and modifications occur not casually, but in controlled pathways after the formation of polypeptide chains. Such modifications lead to changes in shape that are necessary for specific functions. The diversity found in proteins is an example of how living organisms are able to create complex systems using the simple rules of permutation.

2.9 ASSIGNMENTS

2.1 Using the KEGG database, determine the structural formulas of two of the most common five-sugar carbons. Using the same database, explain the differences among ATP, ADP, AMP, and dAMP.

2.2 The primary carbohydrate found in milk is the disaccharide lactose. What are the monosaccharides that make up lactose? Conduct a literature search to determine the name of the enzyme vital for the degradation of lactose.

2.3 Use the KEGG database to elucidate the structural formulas of starch and cellulose?

2.4 Due to complementary base-pair formation, the number of purine nucleotides matches the number of pyrimidines in DNA. Pyrimidine nitrogen bases in DNA are thymine (T) and cytosine (C), and the purine bases are guanine (G) and adenine (A). These nucleotides are contained within the backbone of DNA. The length of A-T and G-C bonds are equal so that the cylindrical helix formed by double-stranded DNA has a uniform radius. Using the genetic code of *Escherichia coli* presented at the National Library of Medicine Web site www.ncbi.nlm.nih.gov/, estimate the number of AT and CG base pairs in these bacterial cells.

Ans: E. coli genome is 48 percent adenine and thymine. Guanine and cytosine base pairing makes up 52 percent of the genome. The weight percentages are as follows: 26.8 percent guanine, 26.3 percent cytosine, 23.8 percent adenine, and 23.1 percent thymine.

2.5 Consider the hypothetical case in which a nucleotide can appear only once in a DNA codon. Determine the number of nucleotides such codons must have in order to express the 20 amino acids found in proteins. If codons were made of four nucleotides, how many different codons would represent the 20 amino acids?