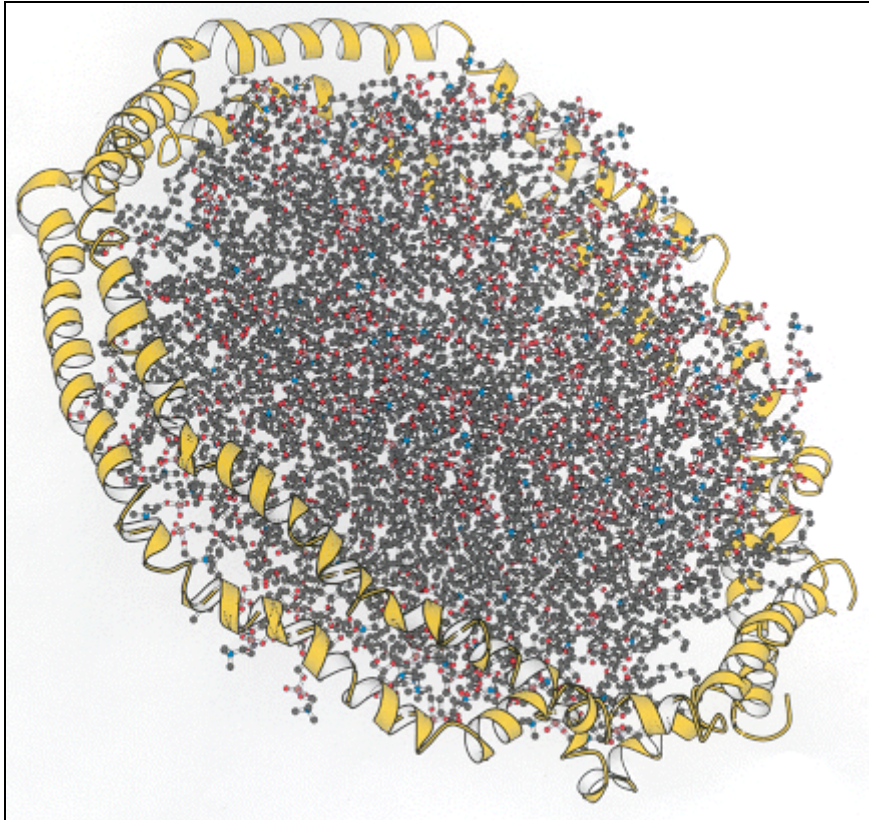


I. The Molecular Design of Life



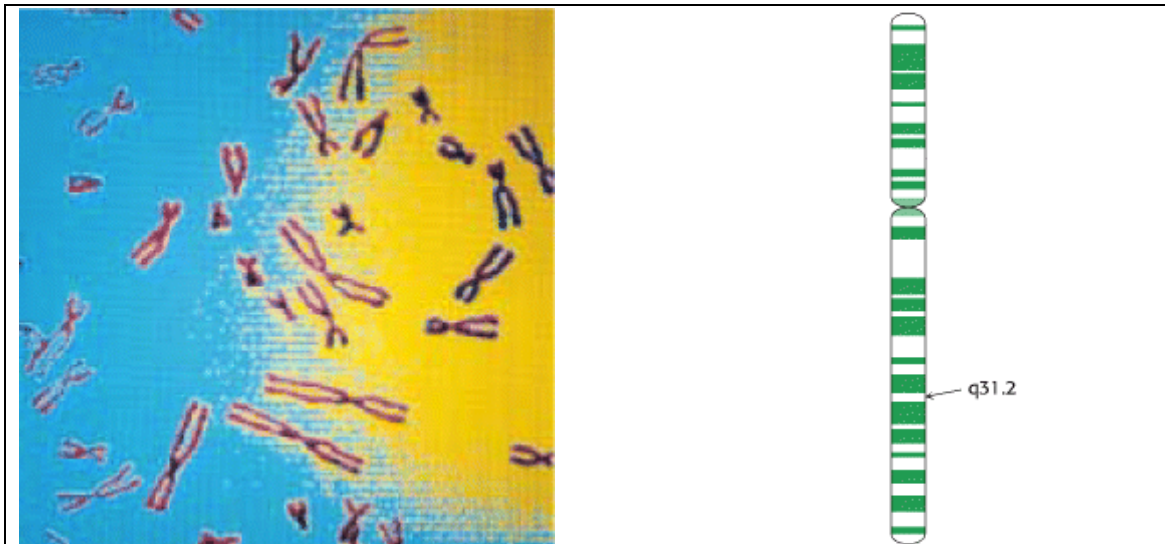
Part of a lipoprotein particle. A model of the structure of apolipoprotein A-I (yellow), shown surrounding sheets of lipids. The apolipoprotein is the major protein component of high-density lipoprotein particles in the blood. These particles are effective lipid transporters because the protein component provides an interface between the hydrophobic lipid chains and the aqueous environment of the bloodstream. [Based on coordinates provided by Stephen Harvey.]

1. Prelude: Biochemistry and the Genomic Revolution

GACTTCACTTCTAATGATGATTATGGGAGAAGCTGGAGCCTTCAGAGGGTAAAAATTAAGCAC
AGTGGGAAGAATTTTCATTCTGTTCTCAGTTTTCTGGATTATGCCTGGCACCATTAAAGAAAAT
ATCTTTGGTGTTCCTATGATGAATATAGATACAGAAGCGTCATCAAAGCATGCCAACTAGA
AGAG. . . This string of letters A, C, G, and T is a part of a DNA sequence. Since the biochemical techniques for DNA sequencing were first developed more than three decades ago, the genomes of dozens of organisms have been sequenced, and many more such sequences will be forthcoming. The information contained in these DNA sequences promises to shed light on many fascinating and important questions. What genes in *Vibrio cholera*, the bacterium that causes cholera, for example, distinguish it from its more benign relatives? How is the development of complex organisms controlled? What are the evolutionary relationships between organisms?

Sequencing studies have led us to a tremendous landmark in the history of biology and, indeed, humanity. A nearly complete sequence of the entire human genome has been determined. The string of As, Cs, Gs, and Ts with which we began this book is a tiny part of the human genome sequence, which is more than 3 billion letters long. If we included the entire sequence, our opening sentence would fill more than 500,000 pages.

The implications of this knowledge cannot be overestimated. By using this blueprint for much of what it means to be human, scientists can begin the identification and characterization of sequences that foretell the appearance of specific diseases and particular physical attributes. One consequence will be the development of better means of diagnosing and treating diseases. Ultimately, physicians will be able to devise plans for preventing or managing heart disease or cancer that take account of individual variations. Although the sequencing of the human genome is an enormous step toward a complete understanding of living systems, much work needs to be done. Where are the functional genes within the sequence, and how do they interact with one another? How is the information in genes converted into the functional characteristics of an organism? Some of our goals in the study of biochemistry are to learn the concepts, tools, and facts that will allow us to address these questions. It is indeed an exciting time, the beginning of a new era in biochemistry.



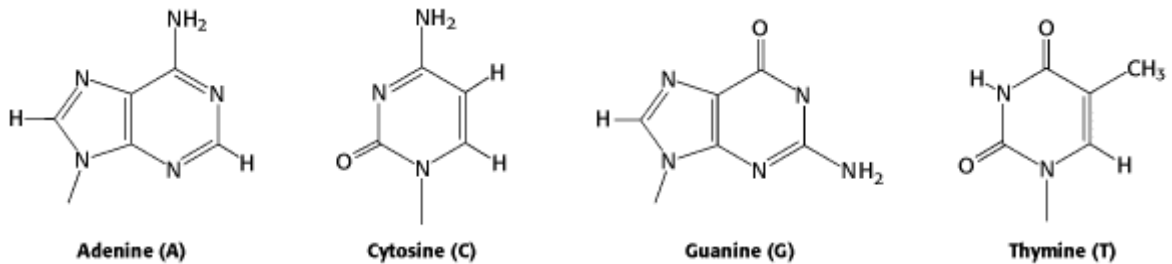
Disease and the genome. Studies of the human genome are revealing disease origins and other biochemical mysteries. Human chromosomes, left, contain the DNA molecules that constitute the human genome. The staining pattern serves to identify specific regions of a chromosome. On the right is a diagram of human chromosome 7, with band q31.2 indicated by an arrow. A gene in this region encodes a protein that, when malfunctioning, causes cystic fibrosis. [(Left) Alfred Pasieka/Peter Arnold.]

1.1. DNA Illustrates the Relation between Form and Function

The structure of DNA, an abbreviation for *deoxyribonucleic acid*, illustrates a basic principle common to all biomolecules: the intimate relation between structure and function. The remarkable properties of this chemical substance allow it to function as a very efficient and robust vehicle for storing information. We begin with an examination of the covalent structure of DNA and its extension into three dimensions.

1.1.1. DNA Is Constructed from Four Building Blocks

DNA is a *linear polymer* made up of four different monomers. It has a fixed backbone from which protrude variable substituents (Figure 1.1). The backbone is built of repeating sugar-phosphate units. The sugars are molecules of *deoxyribose* from which DNA receives its name. Joined to each deoxyribose is one of four possible bases: adenine (A), cytosine (C), guanine (G), and thymine (T).



All four bases are planar but differ significantly in other respects. Thus, the monomers of DNA consist of a sugar-phosphate unit, with one of four bases attached to the sugar. *These bases can be arranged in any order along a strand of DNA.* The order of these bases is what is displayed in the sequence that begins this chapter. For example, the first base in the sequence shown is G (guanine), the second is A (adenine), and so on. *The sequence of bases along a DNA strand constitutes the genetic information* - the instructions for assembling proteins, which themselves orchestrate the synthesis of a host of other biomolecules that form cells and ultimately organisms.

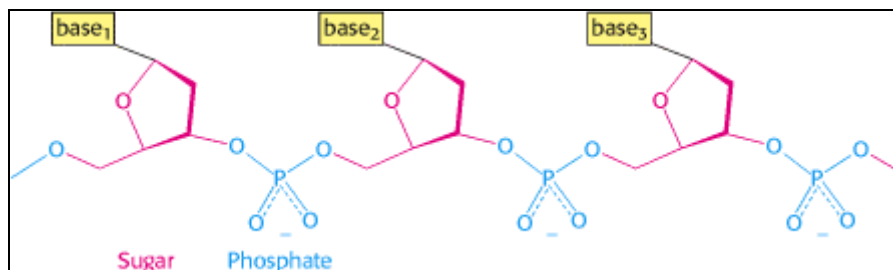


Figure 1.1. Covalent Structure of DNA. Each unit of the polymeric structure is composed of a sugar (deoxyribose), a phosphate, and a variable base that protrudes from the sugar-phosphate backbone.

1.1.2. Two Single Strands of DNA Combine to Form a Double Helix

Most DNA molecules consist of not one but two strands (Figure 1.2). How are these strands positioned with respect to one another? In 1953, James Watson and Francis Crick deduced the arrangement of these strands and proposed a three-dimensional structure for DNA molecules. This structure is a *double helix* composed of two intertwined strands arranged such that the sugar-phosphate backbone lies on the outside and the bases on the inside. The key to this structure is that the bases form *specific base pairs* (bp) held together by *hydrogen bonds* (Section 1.3.1): adenine pairs with thymine (A-T) and guanine pairs with cytosine (G-C), as shown in Figure 1.3. Hydrogen bonds are much weaker than covalent bonds such as the carbon-carbon or carbon-nitrogen bonds that define the structures of the bases themselves. Such weak bonds are crucial to biochemical systems; they are weak enough to be reversibly broken in biochemical processes, yet they are strong enough, when many form simultaneously, to help stabilize specific structures such as the double helix.



Figure 1.2. The Double Helix. The double-helical structure of DNA proposed by Watson and Crick. The sugar-phosphate backbones of the two chains are shown in red and blue and the bases are shown in green, purple, orange, and yellow.

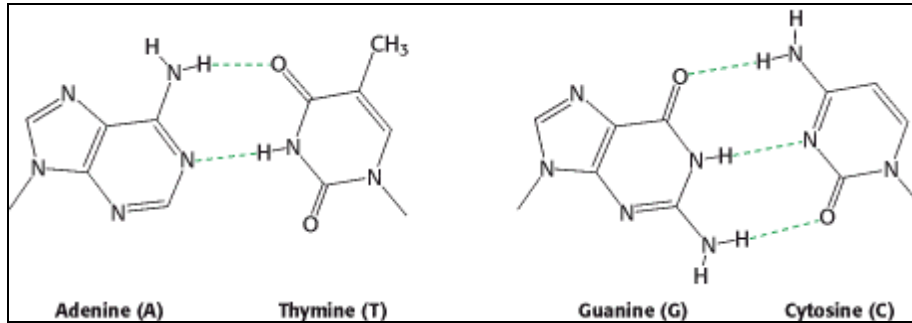


Figure 1.3. Watson-Crick Base Pairs. Adenine pairs with thymine (A-T), and guanine with cytosine (G-C). The dashed lines represent hydrogen bonds.

The structure proposed by Watson and Crick has two properties of central importance to the role of DNA as the hereditary material. First, the structure is compatible with *any sequence of bases*. The base pairs have essentially the same shape (Figure 1.4) and thus fit equally well into the center of the double-helical structure. Second, because of base-pairing, *the sequence of bases along one strand completely determines the sequence along the other strand*. As Watson and Crick so coyly wrote: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." Thus, if the DNA double helix is separated into two single strands, each strand can act as a template for the generation of its partner strand through specific base-pair formation (Figure 1.5). *The three-dimensional structure of DNA beautifully illustrates the close connection between molecular form and function.*

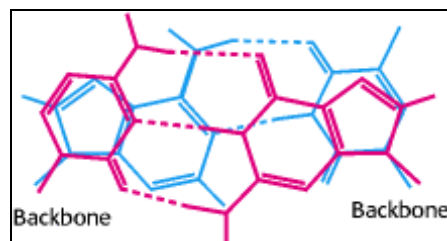


Figure 1.4. Base-Pairing in DNA. The base-pairs A-T (blue) and C-G (red) are shown overlaid. The Watson-Crick base-pairs have the same overall size and shape, allowing them to fit neatly within the double helix.

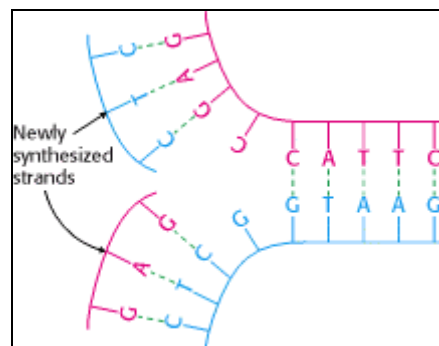
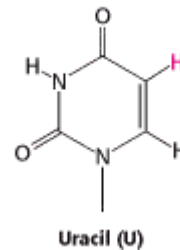
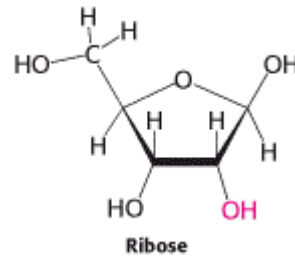


Figure 1.5. DNA Replication. If a DNA molecule is separated into two strands, each strand can act as the template for the generation of its partner strand.

1.1.3. RNA Is an Intermediate in the Flow of Genetic Information

An important nucleic acid in addition to DNA is *ribonucleic acid (RNA)*. Some viruses use RNA as the genetic material, and even those organisms that employ DNA must first convert the genetic information into RNA for the information to be accessible or functional. Structurally, RNA is quite similar to DNA. It is a linear polymer made up of a limited number of repeating monomers, each composed of a sugar, a phosphate, and a base. The sugar is ribose instead of deoxyribose (hence, RNA) and one of the bases is uracil (U) instead of thymine (T). Unlike DNA, an RNA molecule usually exists as a single strand, although significant segments within an RNA molecule may be double stranded, with G pairing primarily with C and A pairing with U. This intrastrand base-pairing generates RNA molecules with complex structures and activities, including catalysis.



RNA has three basic roles in the cell. First, it serves as the intermediate in the flow of information from DNA to protein, the primary functional molecules of the cell. The DNA is copied, or *transcribed*, into messenger RNA (mRNA), and the mRNA is *translated* into protein. Second, RNA molecules serve as adaptors that translate the information in the nucleic acid sequence of mRNA into information designating the sequence of constituents that make up a protein. Finally, RNA molecules are important functional components of the molecular machinery, called ribosomes, that carries out the translation process. As will be discussed in [Chapter 2](#), the unique position of RNA between the storage of genetic information in DNA and the functional expression of this information as protein as well as its potential to combine genetic and catalytic capabilities are indications that RNA played an important role in the evolution of life.

1.1.4. Proteins, Encoded by Nucleic Acids, Perform Most Cell Functions

A major role for many sequences of DNA is to encode the sequences of *proteins*, the workhorses within cells, participating in essentially all processes. Some proteins are key structural components, whereas others are specific catalysts (termed *enzymes*) that promote chemical reactions. Like DNA and RNA, proteins are linear polymers. However, proteins are more complicated in that they are formed from a selection of 20 building blocks, called *amino acids*, rather than 4.

The functional properties of proteins, like those of other biomolecules, are determined by their three-dimensional structures. Proteins possess an extremely important property: a protein spontaneously folds into a welldefined and elaborate three-dimensional structure that is dictated entirely by the sequence of amino acids along its chain ([Figure 1.6](#)). *The self-folding nature of proteins constitutes the transition from the one-dimensional world of sequence information to the three-dimensional world of biological function.* This marvelous ability of proteins to self assemble into complex structures is responsible for their dominant role in biochemistry.

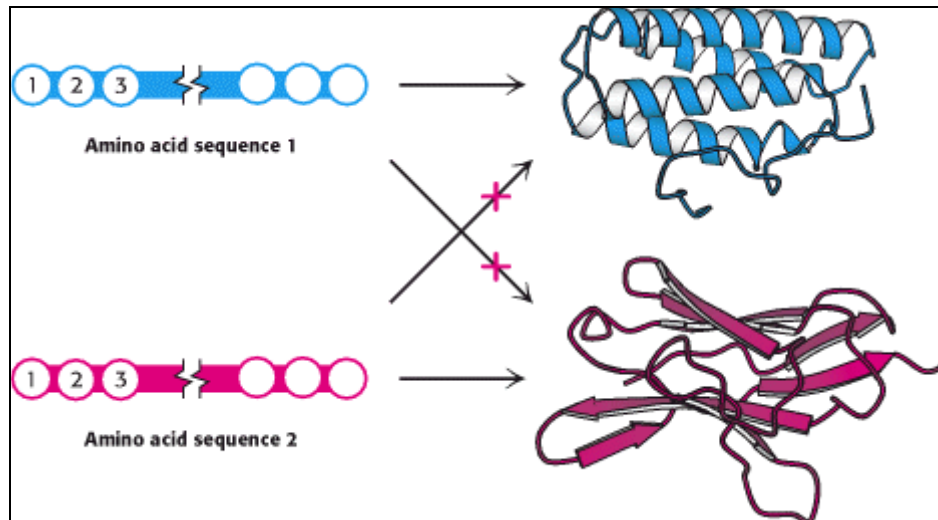


Figure 1.6. Folding of a Protein. The three-dimensional structure of a protein, a linear polymer of amino acids, is dictated by its amino acid sequence.

How is the sequence of bases along DNA translated into a sequence of amino acids along a protein chain? We will consider the details of this process in later chapters, but the important finding is that *three bases along a DNA chain encode a single amino acid*. The specific correspondence between a set of three bases and 1 of the 20 amino acids is called the *genetic code*. Like the use of DNA as the genetic material, the genetic code is essentially universal; the same sequences of three bases encode the same amino acids in all life forms from simple microorganisms to complex, multicellular organisms such as human beings.

Knowledge of the functional and structural properties of proteins is absolutely essential to understanding the significance of the human genome sequence. For example, the sequence at the beginning of this chapter corresponds to a region of the genome that differs in people who have the genetic disorder *cystic fibrosis*. The most common mutation causing cystic fibrosis, the loss of three consecutive Ts from the gene sequence, leads to the loss of a single amino acid within a protein chain of 1480 amino acids. This seemingly slight difference - a loss of 1 amino acid of nearly 1500 - creates a life-threatening condition. What is the normal function of the protein encoded by this gene? What properties of the encoded protein are compromised by this subtle defect? Can this knowledge be used to develop new treatments? These questions fall in the realm of biochemistry. Knowledge of the human genome sequence will greatly accelerate the pace at which connections are made between DNA sequences and disease as well as other human characteristics. However, these connections will be nearly meaningless without the knowledge of biochemistry necessary to interpret and exploit them.

Cystic fibrosis-

A disease that results from a decrease in fluid and salt secretion by a transport protein referred to as the cystic fibrosis transmembrane conductance regulator (CFTR). As a result of this defect, secretion from the pancreas is blocked, and heavy, dehydrated mucus accumulates in the lungs, leading to chronic lung infections.

1.2. Biochemical Unity Underlies Biological Diversity

The stunning variety of living systems (Figure 1.7) belies a striking similarity. The common use of DNA and the genetic code by all organisms underlies one of the most powerful discoveries of the past century - namely, that *organisms are remarkably uniform at the molecular level*. All organisms are built from similar molecular components distinguishable by relatively minor variations. *This uniformity reveals that all organisms on Earth have arisen from a common ancestor*. A core of essential biochemical processes, common to all organisms, appeared early in the evolution of life. The diversity of life in the modern world has been generated by evolutionary processes acting on these core processes through millions or even billions of years. As we will see repeatedly, the generation of diversity has very often resulted from the adaptation of existing biochemical components to new roles rather than the development of fundamentally new biochemical technology. The striking uniformity of life at the molecular level affords the student of biochemistry a particularly clear view into the essence of biological processes that applies to all organisms from human beings to the simplest microorganisms.

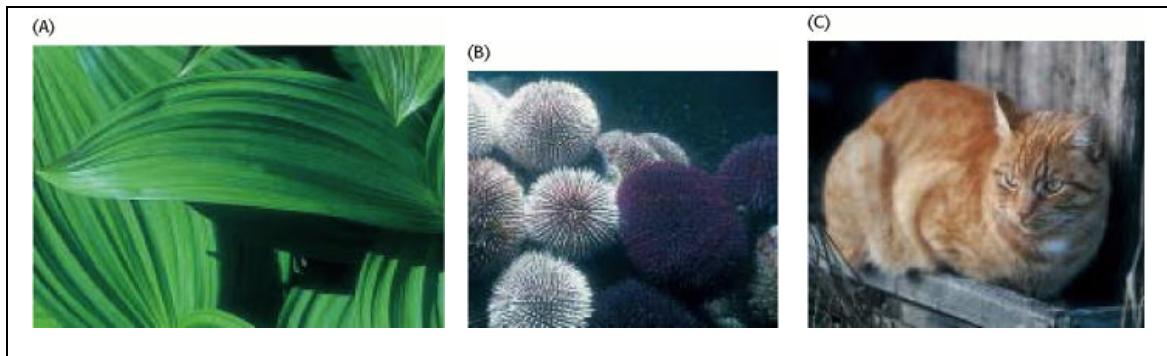


Figure 1.7. The Diversity of Living Systems. The distinct morphologies of the three organisms shown—a plant (the false hellebora, or Indian poke) and two animals (sea urchins and a common house cat)—might suggest that they have little in common. Yet biochemically they display a remarkable commonality that attests to a common ancestry. [(Left and right) John Dudak/Phototake. (Middle) Jeffrey L. Rotman/Peter Arnold.]

On the basis of their biochemical characteristics, the diverse organisms of the modern world can be divided into three fundamental groups called *domains*: *Eukarya* (eukaryotes), *Bacteria* (formerly Eubacteria), and *Archaea* (formerly Archaeobacteria). *Eukarya* comprise all macroscopic organisms, including human beings as well as many microscopic, unicellular organisms such as yeast. The defining characteristic of *eukaryotes* is the presence of a well-defined nucleus within each cell. Unicellular organisms such as bacteria, which lack a nucleus, are referred to as *prokaryotes*. The prokaryotes were reclassified as two separate domains in response to Carl Woese's discovery in 1977 that certain bacteria-like organisms are biochemically quite distinct from better-characterized bacterial species. These organisms, now recognized as having diverged from bacteria early in evolution, are archaea. Evolutionary paths from a common ancestor to modern organisms can be developed and analyzed on the basis of biochemical information. One such path is shown in Figure 1.8.

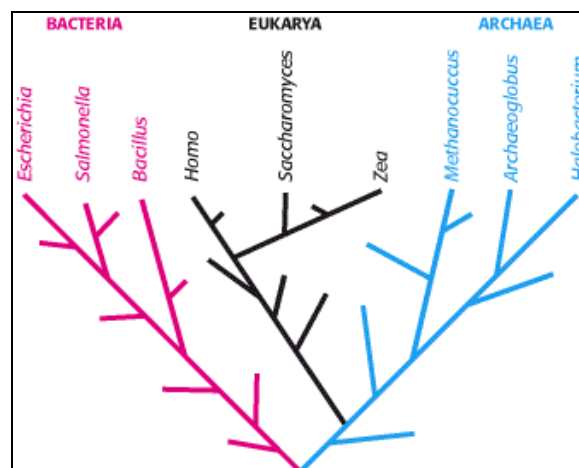


Figure 1.8. The Tree of Life. A possible evolutionary path from a common ancestral cell to the diverse species present in the modern world can be deduced from DNA sequence analysis.

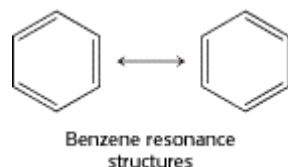
By examining biochemistry in the context of the tree of life, we can often understand how particular molecules or processes helped organisms adapt to specific environments or life styles. We can ask not only *what* biochemical processes take place, but also *why* particular strategies appeared in the course of evolution. In addition to being sources of historical insights, *the answers to such questions are often highly instructive with regard to the biochemistry of contemporary organisms.*

1.3. Chemical Bonds in Biochemistry

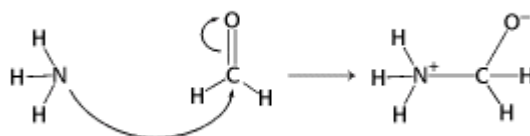
The essence of biological processes - the basis of the uniformity of living systems - is in its most fundamental sense molecular interactions; in other words, the chemistry that takes place between molecules. Biochemistry is the *chemistry* that takes place within living systems. To truly understand biochemistry, we need to understand chemical bonding. We review here the types of chemical bonds that are important for biochemicals and their transformations.

The strongest bonds that are present in biochemicals are *covalent bonds*, such as the bonds that hold the atoms together within the individual bases shown in [Figure 1.3](#). A covalent bond is formed by the sharing of a pair of electrons between adjacent atoms. A typical carbon-carbon (C-C) covalent bond has a bond length of 1.54 Å and bond energy of 85 kcal mol⁻¹ (356 kJ mol⁻¹). Because this energy is relatively high, considerable energy must be expended to break covalent bonds. More than one electron pair can be shared between two atoms to form a multiple covalent bond. For example, three of the bases in [Figure 1.4](#) include carbon-oxygen (C=O) double bonds. These bonds are even stronger than C-C single bonds, with energies near 175 kcal mol⁻¹ (732 kJ mol⁻¹).

For some molecules, more than one pattern of covalent bonding can be written. For example, benzene can be written in two equivalent ways called *resonance structures*. Benzene's true structure is a composite of its two resonance structures. A molecule that can be written as several resonance structures of approximately equal energies has greater stability than does a molecule without multiple resonance structures. Thus, because of its resonance structures, benzene is unusually stable.



Chemical reactions entail the breaking and forming of covalent bonds. The flow of electrons in the course of a reaction can be depicted by curved arrows, a method of representation called "arrow pushing." Each arrow represents an electron pair.



1.3.1. Reversible Interactions of Biomolecules Are Mediated by Three Kinds of Noncovalent Bonds

Readily reversible, noncovalent molecular interactions are key steps in the dance of life. Such weak, noncovalent forces play essential roles in the faithful replication of DNA, the folding of proteins into intricate three-dimensional forms, the specific recognition of substrates by enzymes, and the detection of molecular signals. Indeed, all biological structures and processes depend on the interplay of noncovalent interactions as well as covalent ones. The three fundamental noncovalent bonds are *electrostatic interactions*, *hydrogen bonds*, and *van der Waals interactions*. They differ in geometry, strength, and specificity. Furthermore, these bonds are greatly affected in different ways by the presence of water. Let us consider the characteristics of each:

1. Electrostatic interactions. An electrostatic interaction depends on the electric charges on atoms. The energy of an electrostatic interaction is given by *Coulomb's law*:

$$E = kq_1q_2/Dr$$

where E is the energy, q_1 and q_2 are the charges on the two atoms (in units of the electronic charge), r is the distance between the two atoms (in angstroms), D is the dielectric constant (which accounts for the effects of the intervening medium), and k is a proportionality constant ($k = 332$, to give energies in units of kilocalories per mole, or 1389, for energies in kilojoules per mole). Thus, the electrostatic interaction between two atoms bearing single opposite charges separated by 3 Å in water (which has a dielectric constant of 80) has an energy of 1.4 kcal mol⁻¹ (5.9 kJ mol⁻¹).

2. Hydrogen bonds. Hydrogen bonds are relatively weak interactions, which nonetheless are crucial for biological macromolecules such as DNA and proteins. These interactions are also responsible for many of the properties of water that make it such a special solvent. The hydrogen atom in a hydrogen bond is partly shared between two relatively electronegative atoms such as nitrogen or oxygen. The *hydrogen-bond donor* is the group that includes both the atom to which the hydrogen is more tightly linked and the hydrogen atom itself, whereas the *hydrogen-bond acceptor* is the atom less tightly linked to the hydrogen atom (Figure 1.9). Hydrogen bonds are fundamentally electrostatic interactions. The relatively electronegative atom to which the hydrogen atom is covalently bonded pulls electron density away from the hydrogen atom so that it develops a partial positive charge (δ^+). Thus, it can interact with an atom having a partial negative charge (δ^-) through an electrostatic interaction.

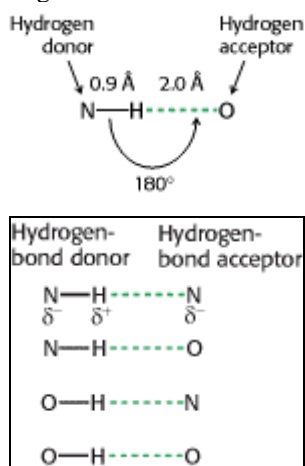


Figure 1.9. Hydrogen Bonds that Include Nitrogen and Oxygen Atoms. The positions of the partial charges (δ^+ and δ^-) are shown.

Hydrogen bonds are much weaker than covalent bonds. They have energies of 1-3 kcal mol⁻¹ (4-13 kJ mol⁻¹) compared with approximately 100 kcal mol⁻¹ (418 kJ mol⁻¹) for a carbon-hydrogen covalent bond. Hydrogen bonds are also somewhat longer than are covalent bonds; their bond distances (measured from the hydrogen atom) range from 1.5 to 2.6 Å; hence, distances ranging from 2.4 to 3.5 Å separate the two nonhydrogen atoms in a hydrogen bond. The strongest hydrogen bonds have a tendency to be approximately straight, such that the hydrogen-bond donor, the hydrogen atom, and the hydrogen-bond acceptor lie along a straight line.

3. van der Waals interactions. The basis of a van der Waals interaction is that the distribution of electronic charge around an atom changes with time. At any instant, the charge distribution is not perfectly symmetric. This transient asymmetry in the electronic charge around an atom acts through electrostatic interactions to induce a complementary asymmetry in the electron distribution around its neighboring atoms. The resulting attraction between two atoms increases as they come closer to each other, until they are separated by the van der Waals *contact distance* (Figure 1.10). At a shorter distance, very strong repulsive forces become dominant because the outer electron clouds overlap.

Energies associated with van der Waals interactions are quite small; typical interactions contribute from 0.5 to 1.0 kcal mol⁻¹ (from 2 to 4 kJ mol⁻¹) per atom pair. When the surfaces of two large molecules come together, however, a large number of atoms are in van der Waals contact, and the net effect, summed over many atom pairs, can be substantial.

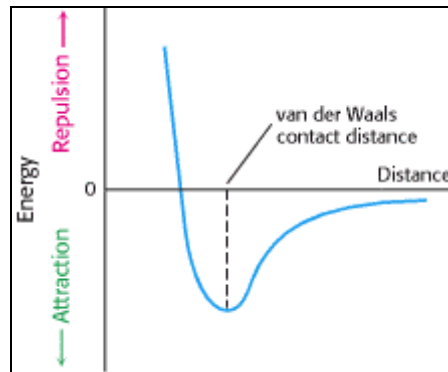
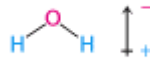


Figure 1.10. Energy of a van der Waals Interaction as Two Atoms Approach One Another. The energy is most favorable at the van der Waals contact distance. The energy rises rapidly owing to electron-electron repulsion as the atoms move closer together than this distance.

1.3.2. The Properties of Water Affect the Bonding Abilities of Biomolecules

Weak interactions are the key means by which molecules interact with one another - enzymes with their substrates, hormones with their receptors, antibodies with their antigens. The strength and specificity of weak interactions are highly dependent on the medium in which they take place, and the majority of biological interactions take place in water. Two properties of water are especially important biologically:



1. Water is a polar molecule. The water molecule is bent, not linear, and so the distribution of charge is asymmetric. The oxygen nucleus draws electrons away from the hydrogen nuclei, which leaves the region around the hydrogen nuclei with a net positive charge. The water molecule is thus an electrically polar structure.

2. Water is highly cohesive. Water molecules interact strongly with one another through hydrogen bonds. These interactions are apparent in the structure of ice (Figure 1.11). Networks of hydrogen bonds hold the structure together; similar interactions link molecules in liquid water and account for the cohesion of liquid water, although, in the liquid state, some of the hydrogen bonds are broken. The highly cohesive nature of water dramatically affects the interactions between molecules in aqueous solution.

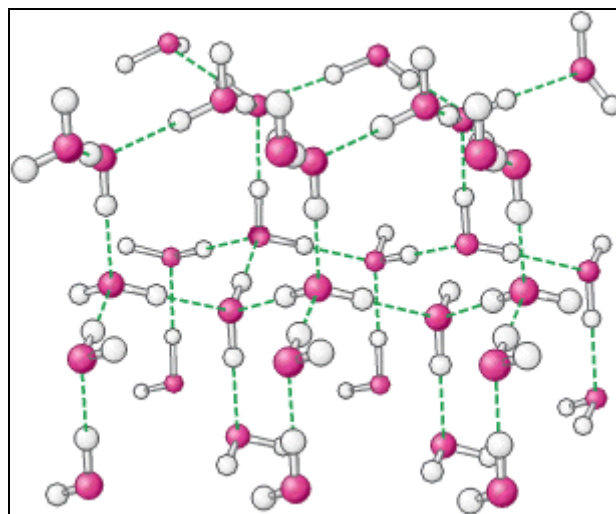
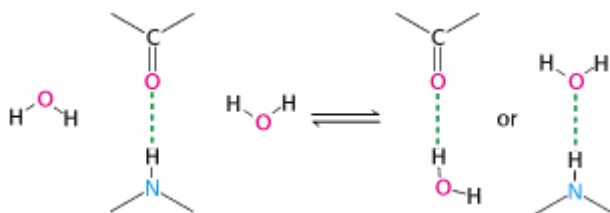


Figure 1.11. Structure of Ice. Hydrogen bonds (shown as dashed lines) are formed between water molecules.

What is the effect of the properties of water on the weak interactions discussed in [Section 1.3.1](#)? The polarity and hydrogen-bonding capability of water make it a highly interacting molecule. Water is an excellent solvent for polar molecules. The reason is that water greatly weakens electrostatic forces and hydrogen bonding between polar molecules by competing for their attractions. For example, consider the effect of water on hydrogen bonding between a carbonyl group and the NH group of an amide.



A hydrogen atom of water can replace the amide hydrogen atom as a hydrogen-bond donor, whereas the oxygen atom of water can replace the carbonyl oxygen atom as a hydrogen-bond acceptor. Hence, a strong hydrogen bond between a CO group and an NH group forms only if water is excluded.

The dielectric constant of water is 80, so water diminishes the strength of electrostatic attractions by a factor of 80 compared with the strength of those same interactions in a vacuum. The dielectric constant of water is unusually high because of its polarity and capacity to form oriented solvent shells around ions. These oriented solvent shells produce electric fields of their own, which oppose the fields produced by the ions. Consequently, the presence of water markedly weakens electrostatic interactions between ions.

The existence of life on Earth depends critically on the capacity of water to dissolve a remarkable array of polar molecules that serve as fuels, building blocks, catalysts, and information carriers. High concentrations of these polar molecules can coexist in water, where they are free to diffuse and interact with one another. However, the excellence of water as a solvent poses a problem, because it also weakens interactions between polar molecules. *The presence of water-free microenvironments within biological systems largely circumvents this problem.* We will see many examples of these specially constructed niches in protein molecules. Moreover, the presence of water with its polar nature permits another kind of weak interaction to take place, one that drives the folding of proteins ([Section 1.3.4](#)) and the formation of cell boundaries ([Section 12.4](#)).

The essence of these interactions, like that of all interactions in biochemistry, is energy. To understand much of biochemistry - bond formation, molecular structure, enzyme catalysis - we need to understand energy. Thermodynamics provides a valuable tool for approaching this topic. We will revisit this topic in more detail when we consider enzymes ([Chapter 8](#)) and the basic concepts of metabolism ([Chapter 14](#)).

1.3.3. Entropy and the Laws of Thermodynamics

The highly structured, organized nature of living organisms is apparent and astonishing. This organization extends from the organismal through the cellular to the molecular level. Indeed, biological processes can seem magical in that the well-ordered structures and patterns emerge from the chaotic and disordered world of inanimate objects. However, the organization visible in a cell or a molecule arises from biological events that are subject to the same physical laws that govern all processes - in particular, the *laws of thermodynamics*.

How can we understand the creation of order out of chaos? We begin by noting that the laws of thermodynamics make a distinction between a system and its surroundings. A *system* is defined as the matter within a defined region of space. The matter in the rest of the universe is called the *surroundings*. *The First Law of Thermodynamics states that the total energy of a system and its surroundings is constant.* In other words, the energy content of the universe is constant; energy can be neither created nor destroyed. Energy can take different forms, however. Heat, for example, is one form of energy. Heat is a manifestation of the *kinetic energy* associated with the random motion of molecules. Alternatively, energy can be present as *potential energy*, referring to the ability of energy to be released on the occurrence of some process. Consider, for example, a ball held at the top of a tower. The ball has considerable potential energy because, when it is released, the ball will develop kinetic energy associated with its motion as it falls. Within chemical systems, potential energy is related to the likelihood that atoms can react with one another. For instance, a mixture of gasoline and oxygen has much potential energy because these molecules may react to form carbon dioxide and release energy as heat. The First Law requires that any

energy released in the formation of chemical bonds be used to break other bonds, be released as heat, or be stored in some other form.

Another important thermodynamic concept is that of *entropy*. Entropy is a measure of the level of randomness or disorder in a system. *The Second Law of Thermodynamics states that the total entropy of a system and its surroundings always increases for a spontaneous process.* At first glance, this law appears to contradict much common experience, particularly about biological systems. Many biological processes, such as the generation of a well-defined structure such as a leaf from carbon dioxide gas and other nutrients, clearly increase the level of order and hence decrease entropy. Entropy may be decreased locally in the formation of such ordered structures only if the entropy of other parts of the universe is increased by an equal or greater amount.

An example may help clarify the application of the laws of thermodynamics to a chemical system. Consider a container with 2 moles of hydrogen gas on one side of a divider and 1 mole of oxygen gas on the other (Figure 1.12). If the divider is removed, the gases will intermingle spontaneously to form a uniform mixture. The process of mixing increases entropy as an ordered arrangement is replaced by a randomly distributed mixture.

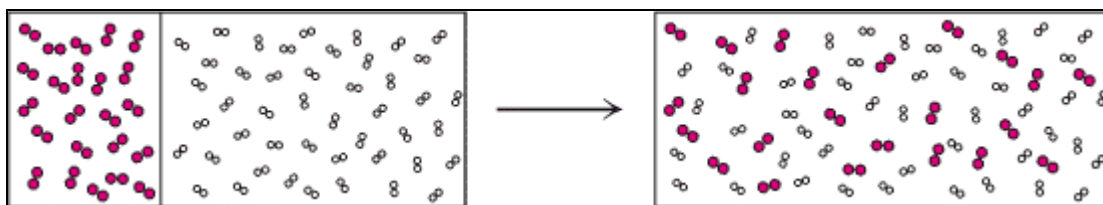
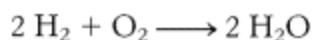


Figure 1.12. From Order to Disorder. The spontaneous mixing of gases is driven by an increase in entropy.

Other processes within this system can decrease the entropy locally while increasing the entropy of the universe. A spark applied to the mixture initiates a chemical reaction in which hydrogen and oxygen combine to form water:



If the temperature of the system is held constant, the entropy of the system decreases because 3 moles of two differing reactants have been combined to form 2 moles of a single product. The gas now consists of a uniform set of indistinguishable molecules. However, the reaction releases a significant amount of heat into the surroundings, and this heat will increase the entropy of the surrounding molecules by increasing their random movement. The entropy increase in the surroundings is enough to allow water to form spontaneously from hydrogen and oxygen (Figure 1.13).

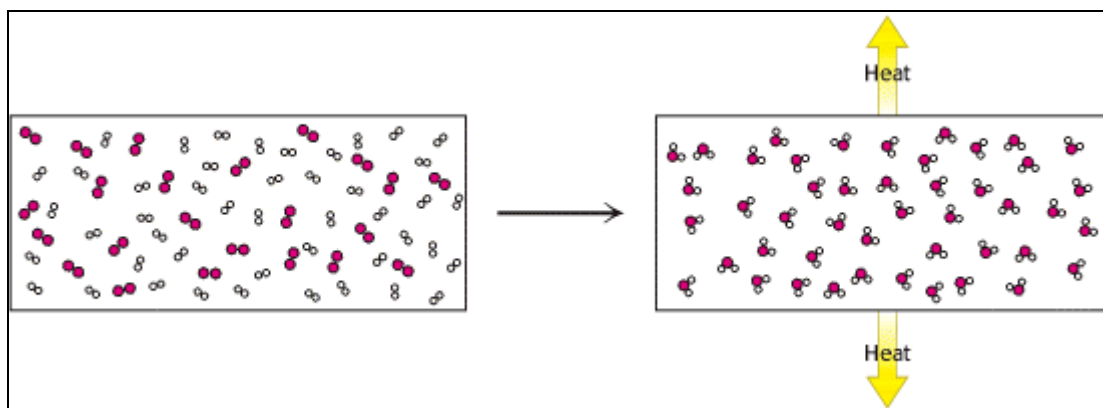


Figure 1.13. Entropy Changes. When hydrogen and oxygen combine to form water, the entropy of the system is reduced, but the entropy of the universe is increased owing to the release of heat to the surroundings.

The change in the entropy of the surroundings will be proportional to the amount of heat transferred from the system and inversely proportional to the temperature of the surroundings, because an input of heat leads to a greater increase in entropy at lower temperatures than at higher temperatures. In biological systems, T [in kelvin (K), absolute temperature] is assumed to be constant. If we define the heat content of a system as *enthalpy* (H), then we can express the relation linking the entropy (S) of the surroundings to the transferred heat and temperature as a simple equation:

$$\Delta S_{\text{surroundings}} = -\Delta H_{\text{system}}/T \quad (1)$$

The total entropy change is given by the expression

$$\Delta S_{\text{total}} = \Delta S_{\text{system}} + \Delta S_{\text{surroundings}} \quad (2)$$

Substituting equation 1 into equation 2 yields

$$\Delta S_{\text{total}} = \Delta S_{\text{system}} - \Delta H_{\text{system}}/T \quad (3)$$

Multiplying by $-T$ gives

$$-T\Delta S_{\text{total}} = \Delta H_{\text{system}} - T\Delta S_{\text{system}} \quad (4)$$

The function $-T\Delta S$ has units of energy and is referred to as *free energy* or *Gibbs free energy*, after Josiah Willard Gibbs, who developed this function in 1878:

$$\Delta G = \Delta H_{\text{system}} - T\Delta S_{\text{system}} \quad (5)$$

The free-energy change, ΔG , will be used throughout this book to describe the energetics of biochemical reactions.

Recall that the Second Law of Thermodynamics states that, for a reaction to be spontaneous, the entropy of the universe must increase. Examination of equation 3 shows that the total entropy will increase if and only if

$$\Delta S_{\text{system}} > \Delta H_{\text{system}}/T \quad (6)$$

Rearranging gives $T\Delta S_{\text{system}} > \Delta H$, or entropy will increase if and only if

$$\Delta G = \Delta H_{\text{system}} - T\Delta S_{\text{system}} < 0 \quad (7)$$

In other words, *the free-energy change must be negative for a reaction to be spontaneous*. A negative free-energy change occurs with an increase in the overall entropy of the universe. Thus, we need to consider only one term, the free energy of the system, to decide whether a reaction can occur spontaneously; any effects of the changes within the system on the rest of the universe are automatically taken into account.

1.3.4. Protein Folding Can Be Understood in Terms of Free-Energy Changes

The problem of protein folding illustrates the utility of the concept of free energy. Consider a system consisting of a solution of unfolded protein molecules in aqueous solution (Figure 1.14). Each unfolded protein molecule can adopt a unique conformation, so the system is quite disordered and the entropy of the collection of molecules is relatively high. Yet, protein folding proceeds spontaneously under appropriate conditions. Thus, entropy must be increasing elsewhere in the system or in the surroundings. How can we reconcile the apparent contradiction that proteins spontaneously assume an ordered structure, and yet entropy increases? The entropy decrease in the system on folding is not as large as it appears to be, because of the properties of water. Molecules in aqueous solution interact with water molecules through the formation of hydrogen and ionic interactions. However, some molecules (termed *nonpolar molecules*) cannot participate in hydrogen or ionic interactions. The interactions of nonpolar molecules with water are not as favorable as are interactions between the water molecules themselves. The water molecules in contact with these nonpolar surfaces form "cages" around the nonpolar molecule, becoming more well ordered (and, hence, lower in entropy) than water molecules free in solution. As two such nonpolar molecules come together, some of the water molecules are released, and so they can interact freely with bulk water (Figure 1.15). Hence, nonpolar molecules have a tendency to aggregate in water because the entropy of the water is increased through the release of water molecules. This phenomenon, termed the *hydrophobic effect*, helps promote many biochemical processes.

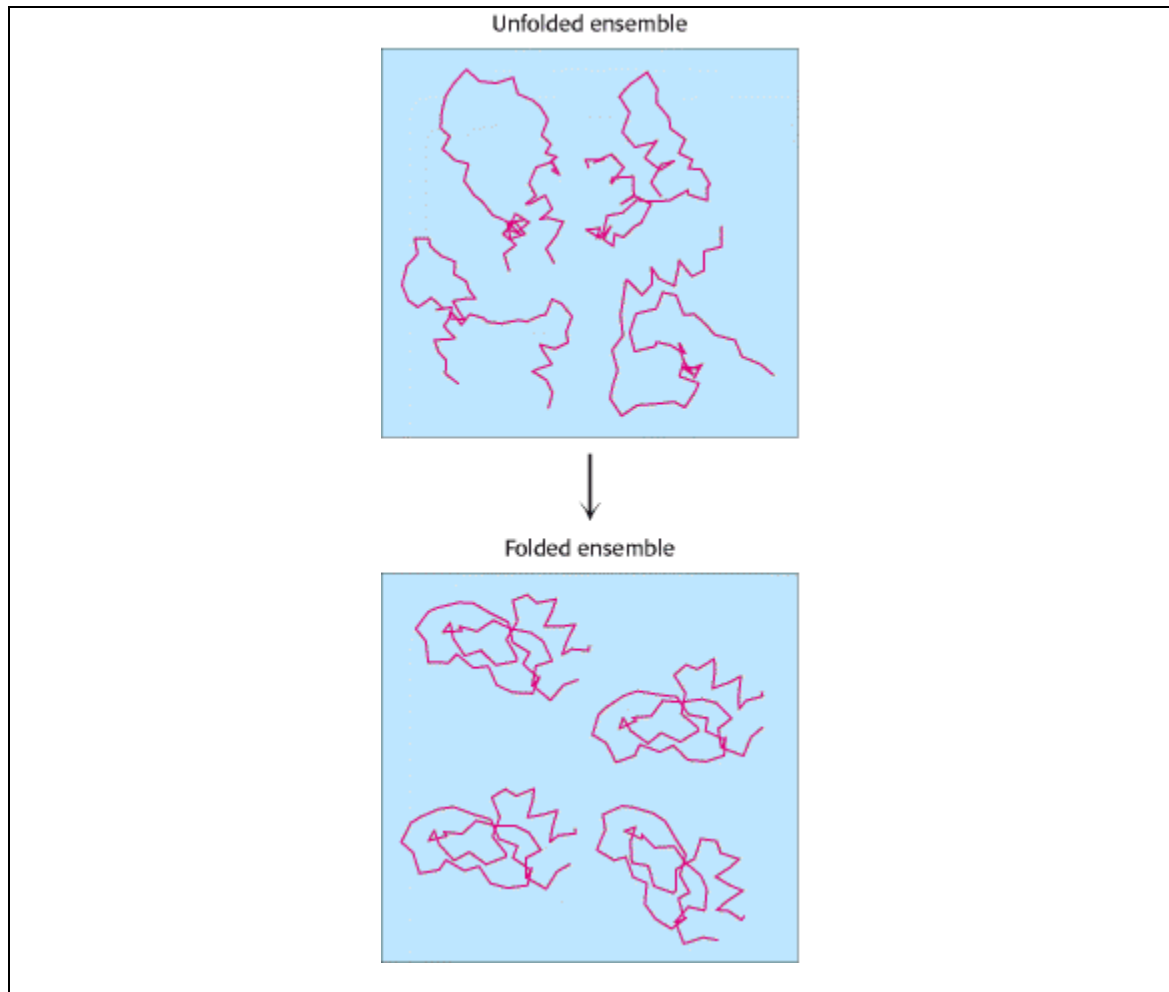


Figure 1.14. Protein Folding. Protein folding entails the transition from a disordered mixture of unfolded molecules to a relatively uniform solution of folded protein molecules.

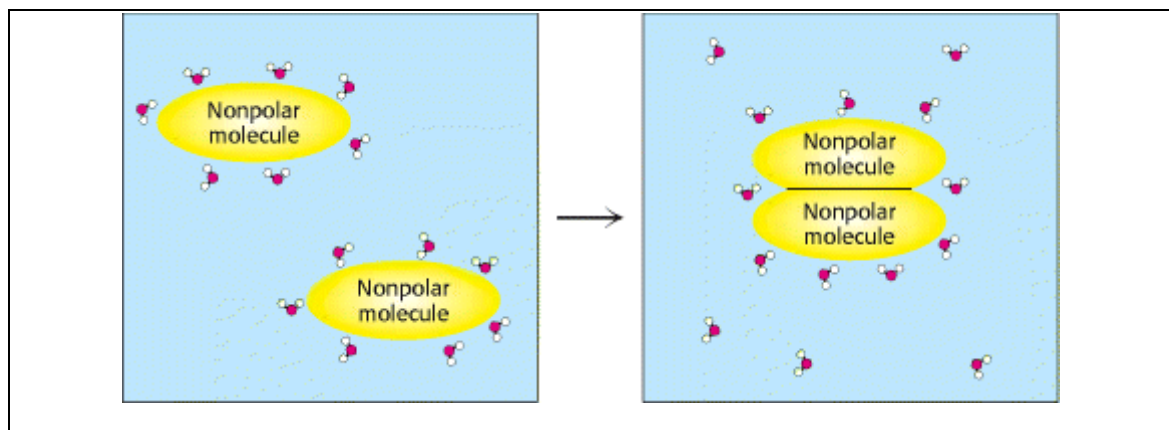


Figure 1.15. The Hydrophobic Effect. The aggregation of nonpolar groups in water leads to an increase in entropy owing to the release of water molecules into bulk water.

How does the hydrophobic effect favor protein folding? Some of the amino acids that make up proteins have nonpolar groups. These nonpolar amino acids have a strong tendency to associate with one another inside the interior of the folded protein. The increased entropy of water resulting from the interaction of these hydrophobic amino acids helps to compensate for the entropy losses inherent in the folding process.

Hydrophobic interactions are not the only means of stabilizing protein structure. Many weak bonds, including hydrogen bonds and van der Waals interactions, are formed in the protein-folding process, and heat is released into the surroundings as a consequence. Although these interactions replace interactions with water that take place in the unfolded protein, the net result is the release of heat to the surroundings and thus a negative (favorable) change in enthalpy for the system.

The folding process can occur when the combination of the entropy associated with the hydrophobic effect and the enthalpy change associated with hydrogen bonds and van der Waals interactions makes the overall free energy negative.

1.4. Biochemistry and Human Biology

Our understanding of biochemistry has had and will continue to have extensive effects on many aspects of human endeavor. *First, biochemistry is an intrinsically beautiful and fascinating body of knowledge.* We now know the essence and many of the details of the most fundamental processes in biochemistry, such as how a single molecule of DNA replicates to generate two identical copies of itself and how the sequence of bases in a DNA molecule determines the sequence of amino acids in an encoded protein. Our ability to describe these processes in detailed, mechanistic terms places a firm chemical foundation under other biological sciences. Moreover, the realization that we can understand essential life processes, such as the transmission of hereditary information, as chemical structures and their reactions has significant philosophical implications. What does it mean, biochemically, to be human? What are the biochemical differences between a human being, a chimpanzee, a mouse, and a fruit fly? Are we more similar than we are different?

Second, biochemistry is greatly influencing medicine and other fields. The molecular lesions causing sickle-cell anemia, cystic fibrosis, hemophilia, and many other genetic diseases have been elucidated at the biochemical level. Some of the molecular events that contribute to cancer development have been identified. An understanding of the underlying defects opens the door to the discovery of effective therapies. Biochemistry makes possible the rational design of new drugs, including specific inhibitors of enzymes required for the replication of viruses such as human immunodeficiency virus (HIV). Genetically engineered bacteria or other organisms can be used as "factories" to produce valuable proteins such as insulin and stimulators of blood-cell development. Biochemistry is also contributing richly to clinical diagnostics. For example, elevated levels of telltale enzymes in the blood reveal whether a patient has recently had a myocardial infarction (heart attack). DNA probes are coming into play in the precise diagnosis of inherited disorders, infectious diseases, and cancers. Agriculture, too, is benefiting from advances in biochemistry with the development of more effective, environmentally safer herbicides and pesticides and the creation of genetically engineered plants that are, for example, more resistant to insects. All of these endeavors are being accelerated by the advances in genomic sequencing.

Third, advances in biochemistry are enabling researchers to tackle some of the most exciting questions in biology and medicine. How does a fertilized egg give rise to cells as different as those in muscle, brain, and liver? How do the senses work? What are the molecular bases for mental disorders such as Alzheimer disease and schizophrenia? How does the immune system distinguish between self and nonself? What are the molecular mechanisms of short-term and long-term memory? The answers to such questions, which once seemed remote, have been partly uncovered and are likely to be more thoroughly revealed in the near future.

Because all living organisms on Earth are linked by a common origin, evolution provides a powerful organizing theme for biochemistry. This book is organized to emphasize the unifying principles revealed by evolutionary considerations. We begin in the next chapter with a brief tour along a plausible evolutionary path from the formation of some of the chemicals that we now associate with living organisms through the evolution of the processes essential for the development of complex, multicellular organisms. The remainder of Part I of the book more fully introduces the most important classes of biochemicals as well as catalysis and regulation. Part II, Transducing and Storing Energy, describes how energy from chemicals or from sunlight is converted into usable forms and how this conversion is regulated. As we will see, a small set of molecules such as adenosine triphosphate (ATP) act as energy currencies that allow energy, however captured, to be utilized in a variety of biochemical processes. This part of the text examines the important pathways for the conversion of environmental energy into molecules such as ATP and uncovers many unifying principles. Part III, Synthesizing the Molecules of Life, illustrates the use of the molecules discussed in Part II to synthesize key molecular building blocks, such as the bases of DNA and amino acids, and then shows how these precursors are assembled into DNA, RNA, and proteins. In Parts II and III, we will highlight the relation between the reactions within each pathway and between those in different pathways so as to suggest how these individual reactions may have combined early in evolutionary history to produce the necessary molecules. From the student's perspective, the existence of features common to several pathways enables material mastered in one context to be readily applied to new contexts. Part IV, Responding to Environmental Changes, explores some of the mechanisms that cells and multicellular organisms have evolved to detect and respond to changes in the environment. The topics range from general mechanisms, common to all organisms, for regulating the expression of genes to the sensory systems used by human beings and other complex organisms. In many cases, we can now see how these elaborate systems evolved from pathways that existed earlier in evolutionary history. Many of the sections in Part IV link biochemistry with other fields

such as cell biology, immunology, and neuroscience. We are now ready to begin our journey into biochemistry with events that took place more than 3 billion years ago.

Appendix: Depicting Molecular Structures

The authors of a biochemistry text face the problem of trying to present three-dimensional molecules in the two dimensions available on the printed page. The interplay between the three-dimensional structures of biomolecules and their biological functions will be discussed extensively throughout this book. Toward this end, we will frequently use representations that, although of necessity are rendered in two dimensions, emphasize the three-dimensional structures of molecules.

Stereochemical Renderings

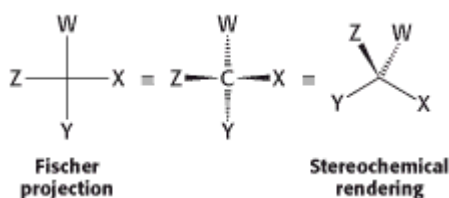
Most of the chemical formulas in this text are drawn to depict the geometric arrangement of atoms, crucial to chemical bonding and reactivity, as accurately as possible. For example, the carbon atom of methane is sp^3 hybridized and tetrahedral, with H-C-H angles of 109.5 degrees while the carbon atom in formaldehyde is sp^2 hybridized with bond angles of 120 degrees.



To illustrate the correct *stereochemistry* about carbon atoms, wedges will be used to depict the direction of a bond into or out of the plane of the page. A solid wedge with the broad end away from the carbon denotes a bond coming toward the viewer out of the plane. A dashed wedge, with the broad end of the bond at the carbon represents a bond going away from the viewer into the plane of the page. The remaining two bonds are depicted as straight lines.

Fischer Projections

Although more representative of the actual structure of a compound, stereochemical structures are often difficult to draw quickly. An alternative method of depicting structures with tetrahedral carbon centers relies on the use of *Fischer projections*.



In a Fischer projection, the bonds to the central carbon are represented by horizontal and vertical lines from the substituent atoms to the carbon atom, which is assumed to be at the center of the cross. By convention, the horizontal bonds are assumed to project out of the page toward the viewer, whereas the vertical bonds are assumed to project into the page away from the viewer. The Glossary of Compounds found at the back of the book is a structural glossary of the key molecules in biochemistry, presented both as stereochemically accurate structures and as Fischer projections.

For depicting molecular architecture in more detail, five types of models will be used: space filling, ball and stick, skeletal, ribbon, and surface representations ([Figure 1.16](#)). The first three types show structures at the atomic level.

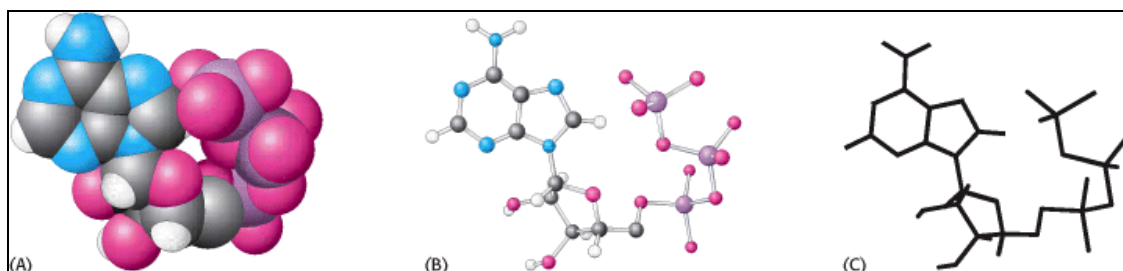


Figure 1.16. Molecular Representations. Comparison of (A) space-filling, (B) ball-and-stick, and (C) skeletal models of ATP.

1. Space-filling models. The space-filling models are the most realistic. The size and position of an atom in a space-filling model are determined by its bonding properties and van der Waals radius, or contact distance (Section 1.3.1). A van der Waals radius describes how closely two atoms can approach each other when they are not linked by a covalent bond. The colors of the model are set by convention.

Carbon, black	Hydrogen, white	Nitrogen, blue
Oxygen, red	Sulfur, yellow	Phosphorus, purple

Space-filling models of several simple molecules are shown in Figure 1.17.

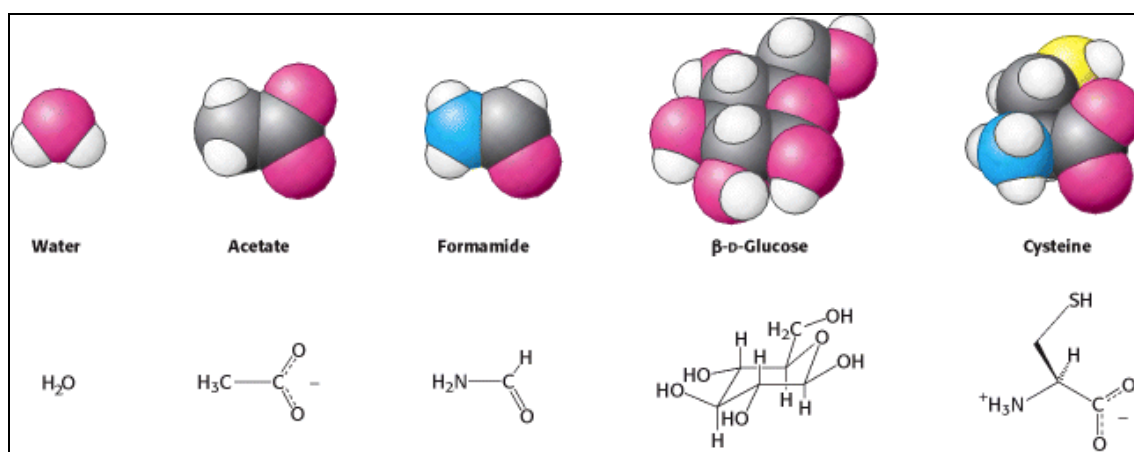


Figure 1.17. Space-Filling Models. Structural formulas and space-filling representations of selected molecules are shown.

2. Ball-and-stick models. Ball-and-stick models are not as realistic as space-filling models, because the atoms are depicted as spheres of radii smaller than their van der Waals radii. However, the bonding arrangement is easier to see because the bonds are explicitly represented as sticks. In an illustration, the taper of a stick, representing parallax, tells which of a pair of bonded atoms is closer to the reader. A ball-and-stick model reveals a complex structure more clearly than a space-filling model does.

3. Skeletal models. An even simpler image is achieved with a skeletal model, which shows only the molecular framework. In skeletal models, atoms are not shown explicitly. Rather, their positions are implied by the junctions and ends of bonds. Skeletal models are frequently used to depict larger, more complex structures.

As biochemistry has advanced, more attention has been focused on the structures of biological macromolecules and their complexes. These structures comprise thousands or even tens of thousands of atoms. Although these structures can be depicted at the atomic level, it is difficult to discern the relevant structural features because of the large number of atoms. Thus, more schematic representations - ribbon diagrams and surface representations - have been developed for the depiction of macromolecular structures in which atoms are not shown explicitly (Figure 1.18).

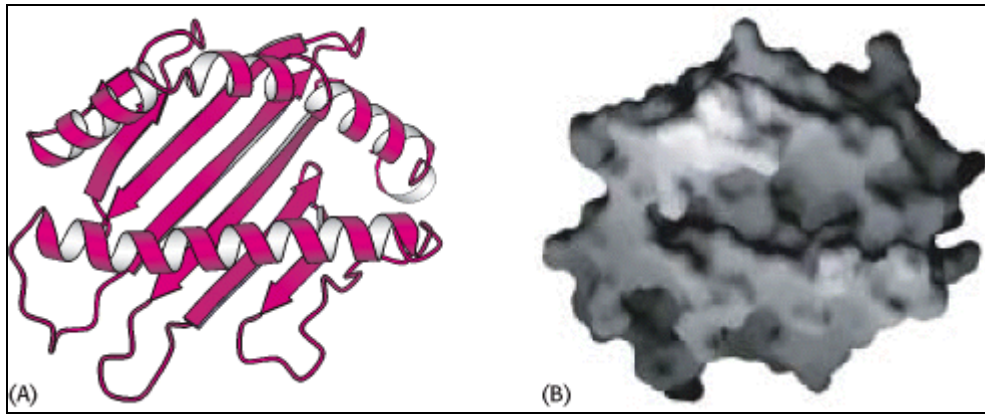


Figure 1.18. Alternative Representations of Protein Structure. A ribbon diagram (A) and a surface representation (B) of a key protein from the immune system emphasize different aspects of structure.

4. Ribbon diagrams. These diagrams are highly schematic and most commonly used to accent a few dramatic aspects of protein structure, such as the α helix (a coiled ribbon), the β strand (a broad arrow), and loops (simple lines), so as to provide simple and clear views of the folding patterns of proteins.

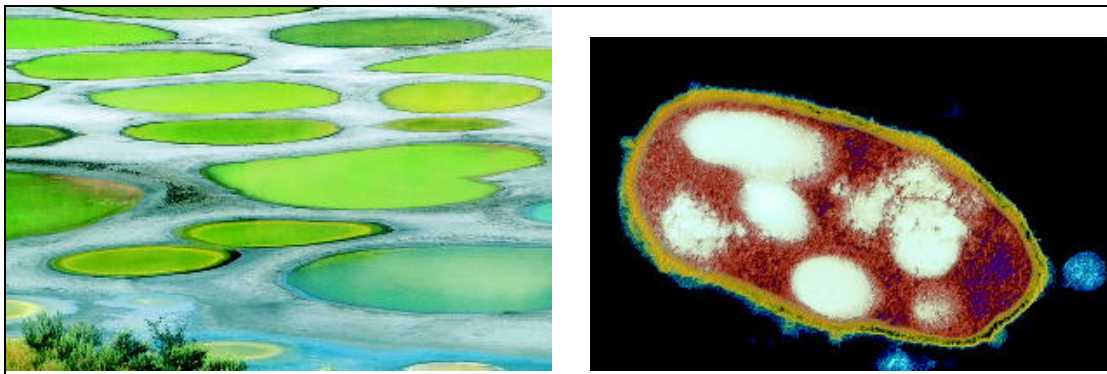
5. Surface representations. Often, the interactions between macromolecules take place exclusively at their surfaces. Surface representations have been developed to better visualize macromolecular surfaces. These representations display the overall shapes of macromolecules and can be shaded or colored to indicate particular features such as surface topography or the distribution of electric charges.

2. Biochemical Evolution

Earth is approximately 4.5 billion years old. Remarkably, there is convincing fossil evidence that organisms morphologically (and very probably biochemically) resembling certain modern bacteria were in existence 3.5 billion years ago. With the use of the results of directed studies and accidental discoveries, it is now possible to construct a hypothetical yet plausible evolutionary path from the prebiotic world to the present. A number of uncertainties remain, particularly with regard to the earliest events. Nonetheless, a consideration of the steps along this path and the biochemical problems that had to be solved provides a useful perspective from which to regard the processes found in modern organisms. *These evolutionary connections make many aspects of biochemistry easier to understand.*

We can think of the path leading to modern living species as consisting of stages, although it is important to keep in mind that these stages were almost certainly not as distinct as presented here. The first stage was the initial generation of some of the key molecules of life - nucleic acids, proteins, carbohydrates, and lipids - by nonbiological processes. The second stage was fundamental - the transition from prebiotic chemistry to replicating systems. With the passage of time, these systems became increasingly sophisticated, enabling the formation of living cells. In the third stage, mechanisms evolved for interconverting energy from chemical sources and sunlight into forms that can be utilized to drive biochemical reactions. Intertwined with these energy-conversion processes are pathways for synthesizing the components of nucleic acids, proteins, and other key substances from simpler molecules. With the development of energy-conversion processes and biosynthetic pathways, a wide variety of unicellular organisms evolved. The fourth stage was the evolution of mechanisms that allowed cells to adjust their biochemistry to different, and often changing, environments. Organisms with these capabilities could form colonies comprising groups of interacting cells, and some eventually evolved into complex multicellular organisms.

This chapter introduces key challenges posed in the evolution of life, whose solutions are elaborated in later chapters. Exploring a possible evolutionary origin for these fundamental processes makes their use, in contrast with that of potential alternatives, more understandable.



Natural selection, one of the key forces powering evolution, opens an array of improbable ecological niches to species that can adapt biochemically. (Left) Salt pools, where the salt concentration can be greater than 1.5 M, would seem to be highly inhospitable environments for life. Yet certain halophilic archaea, such as *Haloferax mediterranei* (right), possess biochemical adaptations that enable them to thrive under these harsh conditions. [(Left) Kaj R. Svensson/Science Photo Library/Photo Researchers; (right) Wanner/Eye of Science/Photo Researchers.]

2.1. Key Organic Molecules Are Used by Living Systems

Approximately 1 billion years after Earth's formation, life appeared, as already mentioned. Before life could exist, though, another major process needed to have taken place - the synthesis of the organic molecules required for living systems from simpler molecules found in the environment. The components of nucleic acids and proteins are relatively complex organic molecules, and one might expect that only sophisticated synthetic routes could produce them. However, this requirement appears not to have been the case. How did the building blocks of life come to be?

2.1.1. Many Components of Biochemical Macromolecules Can Be Produced in Simple, Prebiotic Reactions

Among several competing theories about the conditions of the *prebiotic world*, none is completely satisfactory or problem-free. One theory holds that Earth's early atmosphere was highly reduced, rich in methane (CH₄), ammonia (NH₃), water (H₂O), and hydrogen (H₂), and that this atmosphere was subjected to large amounts of solar radiation and lightning. For the sake of argument, we will assume that these conditions were indeed those of prebiotic Earth. Can complex organic molecules be synthesized under these conditions? In the 1950s, Stanley Miller and Harold Urey set out to answer this question. An electric discharge, simulating lightning, was passed through a mixture of methane, ammonia, water, and hydrogen (Figure 2.1). Remarkably, these experiments yielded a highly nonrandom mixture of organic compounds, including amino acids and other substances fundamental to biochemistry. The procedure produces the amino acids glycine and alanine in approximately 2% yield, depending on the amount of carbon supplied as methane. More complex amino acids such as glutamic acid and leucine are produced in smaller amounts (Figure 2.2). Hydrogen cyanide (HCN), another likely component of the early atmosphere, will condense on exposure to heat or light to produce adenine, one of the four nucleic acid bases (Figure 2.3). Other simple molecules combine to form the remaining bases. A wide array of sugars, including ribose, can be formed from formaldehyde under prebiotic conditions.



Figure 2.1. The Urey-Miller Experiment. An electric discharge (simulating lightning) passed through an atmosphere of CH₄, NH₃, H₂O, and H₂ leads to the generation of key organic compounds such as amino acids.

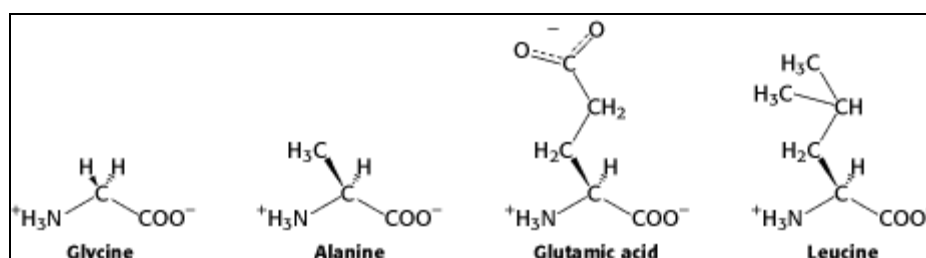


Figure 2.2. Products of Prebiotic Synthesis. Amino acids produced in the Urey-Miller experiment.

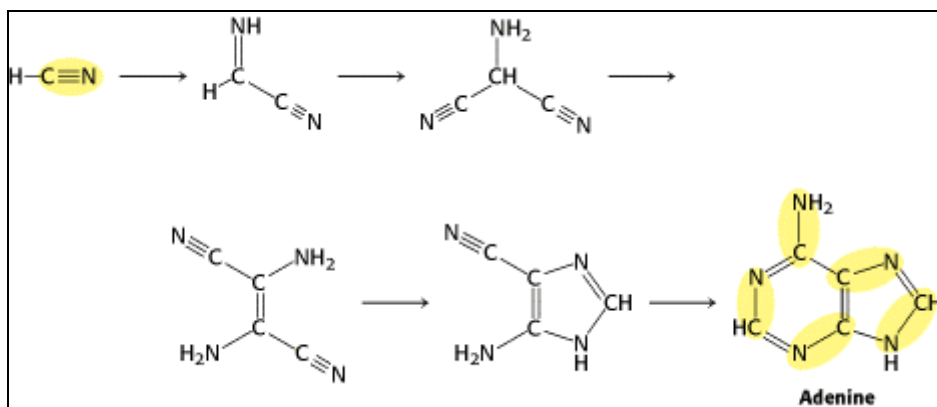


Figure 2.3. Prebiotic Synthesis of a Nucleic Acid Component. Adenine can be generated by the condensation of HCN.

2.1.2. Uncertainties Obscure the Origins of Some Key Biomolecules

The preceding observations suggest that many of the building blocks found in biology are unusually easy to synthesize and that significant amounts could have accumulated through the action of nonbiological processes. However, it is important to keep in mind that there are many uncertainties. For instance, ribose is just one of many sugars formed under prebiotic conditions. In addition, ribose is rather unstable under possible prebiotic conditions. Furthermore, ribose occurs in two mirror-image forms, only one of which occurs in modern RNA. To circumvent those problems, the first nucleic acid-like molecules have been suggested to have been bases attached to a different backbone and only later in evolutionary time was ribose incorporated to form nucleic acids as we know them today. Despite these uncertainties, an assortment of prebiotic molecules did arise in some fashion, and from this assortment *those with properties favorable for the processes that we now associate with life began to interact and to form more complicated compounds*. The processes through which modern organisms synthesize molecular building blocks will be discussed in [Chapters 24](#), [25](#), and [26](#).

2.2. Evolution Requires Reproduction, Variation, and Selective Pressure

Once the necessary building blocks were available, how did a living system arise and evolve? Before the appearance of life, simple molecular systems must have existed that subsequently evolved into the complex chemical systems that are characteristic of organisms. To address how this evolution occurred, we need to consider the *process* of evolution. There are several basic principles common to evolving systems, whether they are simple collections of molecules or competing populations of organisms. First, the most fundamental property of evolving systems is their ability to *replicate* or *reproduce*. Without this ability of *reproduction*, each "species" of molecule that might appear is doomed to extinction as soon as all its individual molecules degrade. For example, individual molecules of biological polymers such as ribonucleic acid are degraded by hydrolysis reactions and other processes. However, *molecules that can replicate will continue to be represented in the population even if the lifetime of each individual molecule remains short.*

A second principle fundamental to evolution is *variation*. The replicating systems must undergo changes. After all, if a system always replicates perfectly, the replicated molecule will always be the same as the parent molecule. Evolution cannot occur. The nature of these variations in living systems are considered in [Section 2.2.5](#).

A third basic principle of evolution is *competition*. Replicating molecules compete with one another for available resources such as chemical precursors, and the competition allows the process of *evolution by natural selection* to occur. Variation will produce differing populations of molecules. Some variant offspring may, by chance, be better suited for survival and replication under the prevailing conditions than are their parent molecules. The prevailing conditions exert a *selective pressure* that gives an advantage to one of the variants. Those molecules that are best able to survive and to replicate themselves will increase in relative concentration. Thus, new molecules arise that are better able to replicate under the conditions of their environment. The same principles hold true for modern organisms. Organisms reproduce, show variation among individual organisms, and compete for resources; those variants with a selective advantage will reproduce more successfully. The changes leading to variation still take place at the molecular level, but the selective advantage is manifest at the organismal level.

2.2.1. The Principles of Evolution Can Be Demonstrated in Vitro

Is there any evidence that evolution can take place at the molecular level? In 1967, Sol Spiegelman showed that replicating molecules could evolve new forms in an experiment that allowed him to observe molecular evolution in the test tube. He used as his evolving molecules RNA molecules derived from a bacterial virus called bacteriophage $Q\beta$. The genome of bacteriophage $Q\beta$, a single RNA strand of approximately 3300 bases, depends for its replication on the activity of a protein complex termed $Q\beta$ replicase. Spiegelman mixed the replicase with a starting population of $Q\beta$ RNA molecules. Under conditions in which there are ample amounts of precursors, no time constraints, and no other selective pressures, the composition of the population does not change from that of the parent molecules on replication. When selective pressures are applied, however, the composition of the population of molecules can change dramatically. For example, decreasing the time available for replication from 20 minutes to 5 minutes yielded, incrementally over 75 generations, a population of molecules dominated by a single species comprising only 550 bases. This species is replicated 15 times as rapidly as the parental $Q\beta$ RNA ([Figure 2.4](#)). Spiegelman applied other selective pressures by, for example, limiting the concentrations of precursors or adding compounds that inhibit the replication process. In each case, new species appeared that replicated more effectively under the conditions imposed.

The process of evolution demonstrated in these studies depended on the existence of machinery for the replication of RNA fragments in the form of the $Q\beta$ replicase. As noted in [Chapter 1](#), one of the most elegant characteristics of nucleic acids is that the mechanism for their replication follows naturally from their molecular structure. This observation suggests that nucleic acids, perhaps RNA, could have become *self-replicating*. Indeed, the results of studies have revealed that single-stranded nucleic acids can serve as templates for the synthesis of their complementary strands and that this synthesis can occur spontaneously - that is, without biologically derived replication machinery. However, investigators have not yet found

conditions in which an RNA molecule is fully capable of independent selfreplication from simple starting materials.

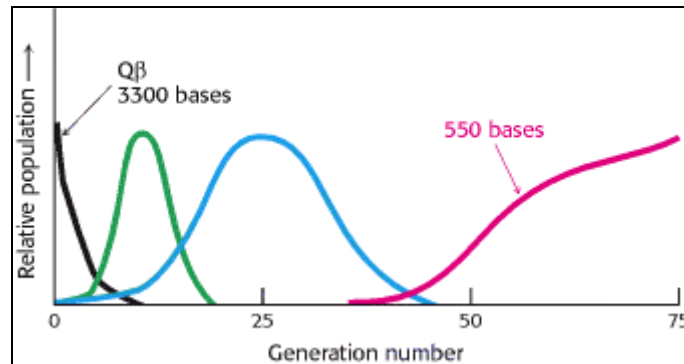


Figure 2.4. Evolution in a Test Tube. Rapidly replicating species of RNA molecules were generated from Q β RNA by exerting selective pressure. The green and blue curves correspond to species of intermediate size that accumulated and then became extinct in the course of the experiment.

2.2.2. RNA Molecules Can Act As Catalysts

The development of capabilities beyond simple replication required the generation of specific catalysts. A *catalyst* is a molecule that accelerates a particular chemical reaction without itself being chemically altered in the process. The properties of catalysts will be discussed in detail in [Chapters 8 and 9](#). Some catalysts are highly specific; they promote certain reactions without substantially affecting closely related processes. Such catalysts allow the reactions of specific pathways to take place in preference to those of potential alternative pathways. Until the 1980s, all biological catalysts, termed *enzymes*, were believed to be proteins. Then, Tom Cech and Sidney Altman independently discovered that certain RNA molecules can be effective catalysts. These RNA catalysts have come to be known as *ribozymes*. The discovery of ribozymes suggested the possibility that catalytic RNA molecules could have played fundamental roles early in the evolution of life.

The catalytic ability of RNA molecules is related to their ability to adopt specific yet complex structures. This principle is illustrated by a "hammerhead" ribozyme, an RNA structure first identified in plant viruses ([Figure 2.5](#)). This RNA molecule promotes the cleavage of specific RNA molecules at specific sites; this cleavage is necessary for certain aspects of the viral life cycle. The ribozyme, which requires Mg²⁺ ion or other ions for the cleavage step to take place, forms a complex with its substrate RNA molecule that can adopt a reactive conformation.

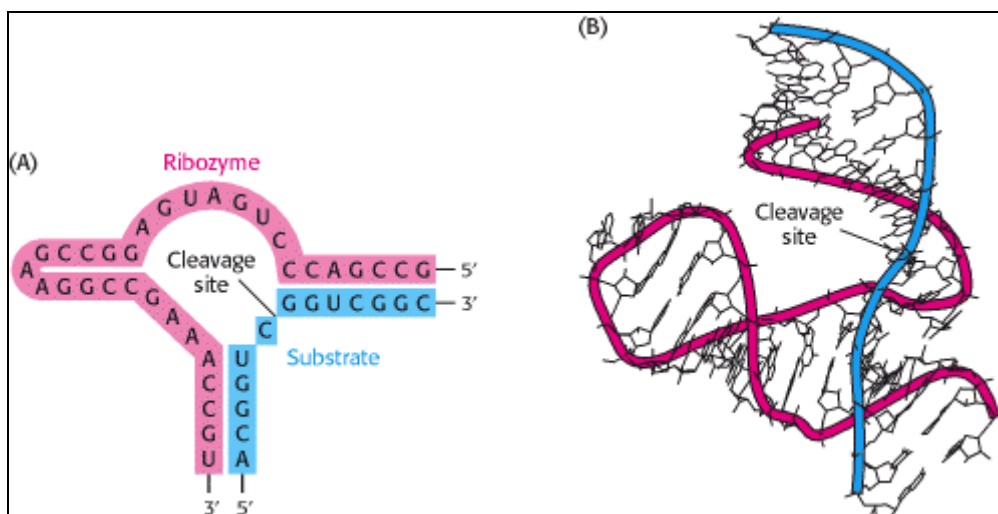


Figure 2.5. Catalytic RNA. (A) The base-pairing pattern of a "hammerhead" ribozyme and its substrate. (B) The folded conformation of the complex. The ribozyme cleaves the bond at the cleavage site. The paths of the nucleic acid backbones are highlighted in red and blue.

The existence of RNA molecules that possess specific binding and catalytic properties makes plausible the idea of an early "RNA world" inhabited by life forms dependent on RNA molecules to play all major roles, including those important in heredity, the storage of information, and the promotion of specific reactions - that is, biosynthesis and energy metabolism.

2.2.3. Amino Acids and Their Polymers Can Play Biosynthetic and Catalytic Roles

In the early RNA world, the increasing populations of replicating RNA molecules would have consumed the building blocks of RNA that had been generated over long periods of time by prebiotic reactions. A shortage of these compounds would have favored the evolution of alternative mechanisms for their synthesis. A large number of pathways are possible. Examining the biosynthetic routes utilized by modern organisms can be a source of insight into which pathways survived. A striking observation is that simple amino acids are used as building blocks for the RNA bases (Figure 2.6). For both purines (adenine and guanine) and pyrimidines (uracil and cytosine), an amino acid serves as a core onto which the remainder of the base is elaborated. In addition, nitrogen atoms are donated by the amino group of the amino acid aspartic acid and by the amide group of the glutamine side chain.

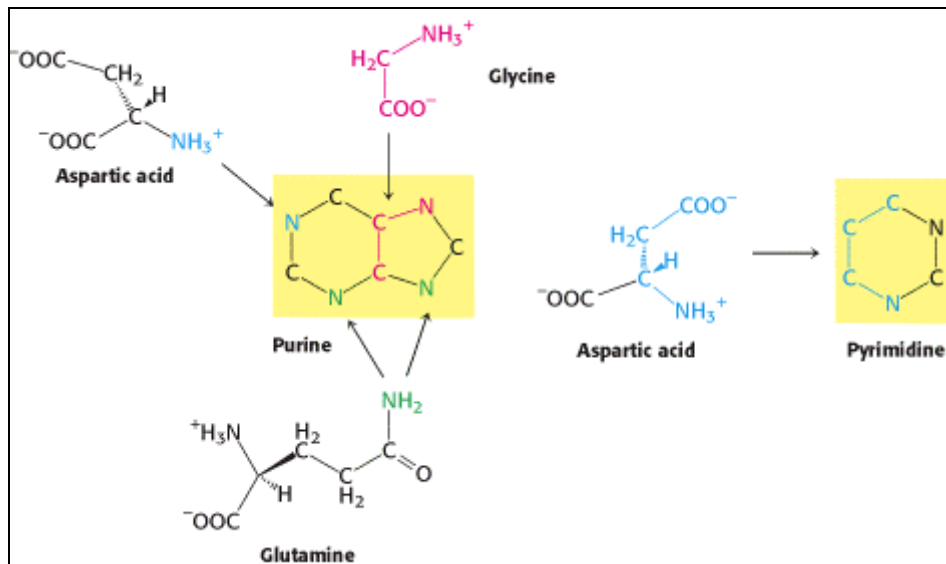


Figure 2.6. Biosynthesis of RNA Bases. Amino acids are building blocks for the biosynthesis of purines and pyrimidines.

Amino acids are chemically more versatile than nucleic acids because their side chains carry a wider range of chemical functionality. Thus, amino acids or short polymers of amino acids linked by *peptide bonds*, called *polypeptides* (Figure 2.7), may have functioned as components of ribozymes to provide a specific reactivity. Furthermore, longer polypeptides are capable of spontaneously folding to form well-defined three-dimensional structures, dictated by the sequence of amino acids along their polypeptide chains. The ability of polypeptides to fold spontaneously into elaborate structures, which permit highly specific chemical interactions with other molecules, may have favored the expansion of their roles in the course of evolution and is crucial to their dominant position in modern organisms. Today, most biological catalysts (enzymes) are not nucleic acids but are instead large polypeptides called *proteins*.

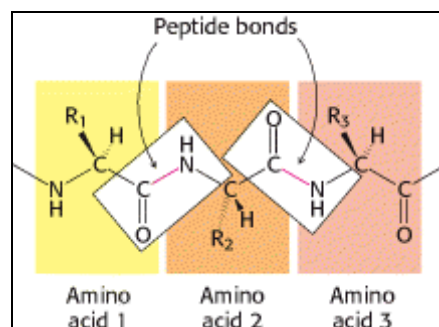


Figure 2.7. An Alternative Functional Polymer. Proteins are built of amino acids linked by peptide bonds.

2.2.4. RNA Template-Directed Polypeptide Synthesis Links the RNA and Protein Worlds

Polypeptides would have played only a limited role early in the evolution of life because their structures are not suited to self-replication in the way that nucleic acid structures are. However, polypeptides could have been included in evolutionary processes indirectly. For example, if the properties of a particular polypeptide favored the survival and replication of a class of RNA molecules, then these RNA molecules could have evolved ribozyme activities that promoted the synthesis of that polypeptide. This method of producing polypeptides with specific amino acid sequences has several limitations. First, it seems likely that only relatively short specific polypeptides could have been produced in this manner. Second, it would have been difficult to accurately link the particular amino acids in the polypeptide in a reproducible manner. Finally, a different ribozyme would have been required for each polypeptide. A critical point in evolution was reached when an apparatus for polypeptide synthesis developed that allowed *the sequence of bases in an RNA molecule to directly dictate the sequence of amino acids in a polypeptide*. A code evolved that established a relation between a specific sequence of three bases in RNA and an amino acid. We now call this set of three-base combinations, each encoding an amino acid, the *genetic code*. A decoding, or *translation*, system exists today as the *ribosome* and associated factors that are responsible for essentially all polypeptide synthesis from RNA templates in modern organisms. The essence of this mode of polypeptide synthesis is illustrated in [Figure 2.8](#).

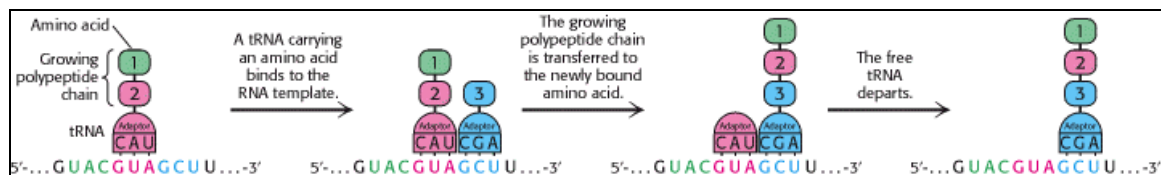


Figure 2.8. Linking the RNA and Protein Worlds. Polypeptide synthesis is directed by an RNA template. Adaptor RNA molecules, with amino acids attached, sequentially bind to the template RNA to facilitate the formation of a peptide bond between two amino acids. The growing polypeptide chain remains attached to an adaptor RNA until the completion of synthesis.

An RNA molecule (*messenger RNA*, or *mRNA*), containing in its base sequence the information that specifies a particular protein, acts as a template to direct the synthesis of the polypeptide. Each amino acid is brought to the template attached to an adapter molecule specific to that amino acid. These adapters are specialized RNA molecules (called *transfer RNAs* or *tRNAs*). After initiation of the polypeptide chain, a tRNA molecule with its associated amino acid binds to the template through specific Watson-Crick base-pairing interactions. Two such molecules bind to the ribosome and peptide-bond formation is catalyzed by an RNA component (called *ribosomal RNA* or *rRNA*) of the ribosome. The first RNA departs (with neither the polypeptide chain nor an amino acid attached) and another tRNA with its associated amino acid bonds to the ribosome. The growing polypeptide chain is transferred to this newly bound amino acid with the formation of a new peptide bond. This cycle then repeats itself. This scheme allows the sequence of the RNA template to encode the sequence of the polypeptide and thereby makes possible the production of long polypeptides with specified sequences. The mechanism of protein synthesis will be discussed in [Chapter 29](#). Importantly, the ribosome is composed largely of RNA and is a highly sophisticated ribozyme, suggesting that it might be a surviving relic of the RNA world.

2.2.5. The Genetic Code Elucidates the Mechanisms of Evolution

The sequence of bases that encodes a functional protein molecule is called a *gene*. The genetic code - that is, the relation between the base sequence of a gene and the amino acid sequence of the polypeptide whose synthesis the gene directs - applies to all modern organisms with only very minor exceptions. This universality reveals that the genetic code was fixed early in the course of evolution and has been maintained to the present day.

We can now examine the mechanisms of evolution. Earlier, we considered how variation is required for evolution. We can now see that such variations in living systems are changes that alter the meaning of the genetic message. These variations are called *mutations*. A mutation can be as simple as a change in a single nucleotide (called a *point mutation*), such that a sequence of bases that encoded a particular amino

acid may now encode another (Figure 2.9A). A mutation can also be the insertion or deletion of several nucleotides.

Other types of alteration permit the more rapid evolution of new biochemical activities. For instance, entire sections of the coding material can be duplicated, a process called *gene duplication* (Figure 2.9B). One of the duplication products may accumulate mutations and eventually evolve into a gene with a different, but related, function. Furthermore, parts of a gene may be duplicated and added to parts of another to give rise to a completely new gene, which encodes a protein with properties associated with each parent gene. Higher organisms contain many large families of enzymes and other macromolecules that are clearly related to one another in the same manner. Thus, gene duplication followed by specialization has been a crucial process in evolution. It allows the generation of macromolecules having particular functions without the need to start from scratch. The accumulation of genes with subtle and large differences allows for the generation of more complex biochemical processes and pathways and thus more complex organisms.

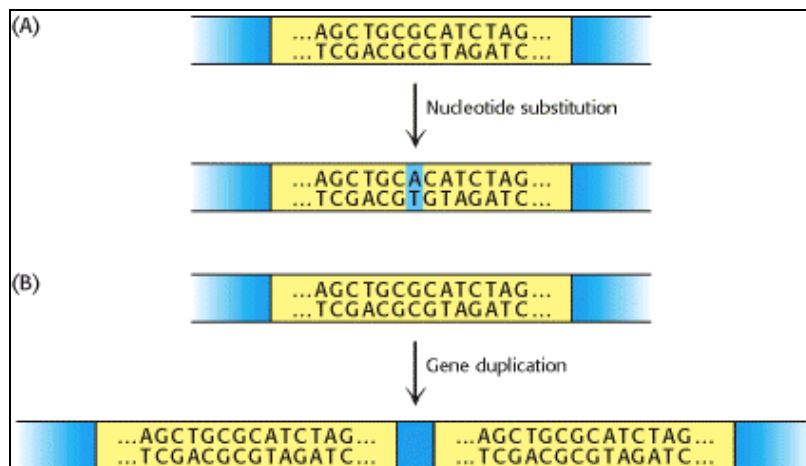


Figure 2.9. Mechanisms of Evolution. A change in a gene can be (A) as simple as a single base change or (B) as dramatic as partial or complete gene duplication.

2.2.6. Transfer RNAs Illustrate Evolution by Gene Duplication

Transfer RNA molecules are the adaptors that associate an amino acid with its correct base sequence. Transfer RNA molecules are structurally similar to one another: each adopts a three-dimensional cloverleaf pattern of base-paired groups (Figure 2.10). Subtle differences in structure enable the protein-synthesis machinery to distinguish transfer RNA molecules with different amino acid specificities.

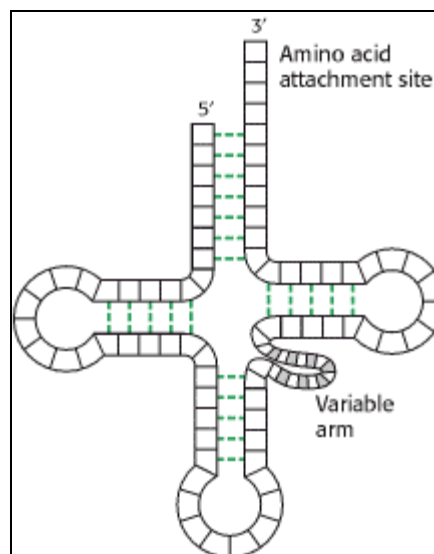


Figure 2.10. Cloverleaf Pattern of tRNA. The pattern of base-pairing interactions observed for all transfer RNA molecules reveals that these molecules had a common evolutionary origin.

This family of related RNA molecules likely was generated by gene duplication followed by specialization. A nucleic acid sequence encoding one member of the family was duplicated, and the two copies evolved independently to generate molecules with specificities for different amino acids. This process was repeated, starting from one primordial transfer RNA gene until the 20 (or more) distinct members of the transfer RNA family present in modern organisms arose.

2.2.7. DNA Is a Stable Storage Form for Genetic Information

It is plausible that RNA was utilized to store genetic information early in the history of life. However, in modern organisms (with the exception of some viruses), the RNA derivative DNA (*deoxyribonucleic acid*) performs this function (Sections 1.1.1 and 1.1.3). The 2'-hydroxyl group in the ribose unit of the RNA backbone is replaced by a hydrogen atom in DNA (Figure 2.11).

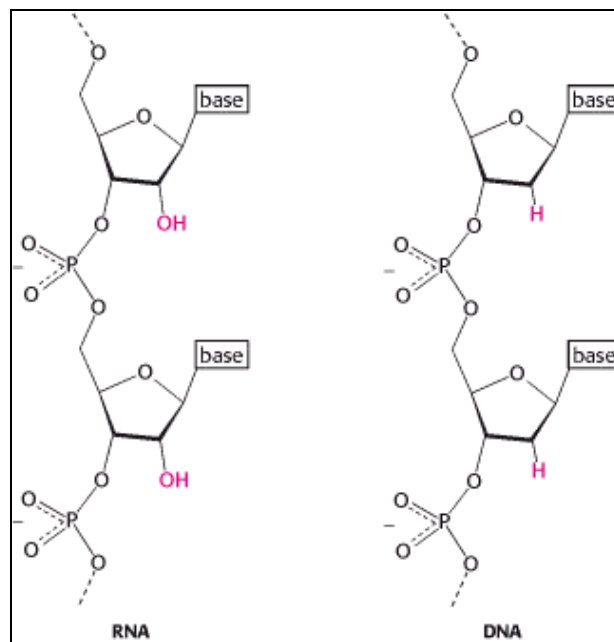
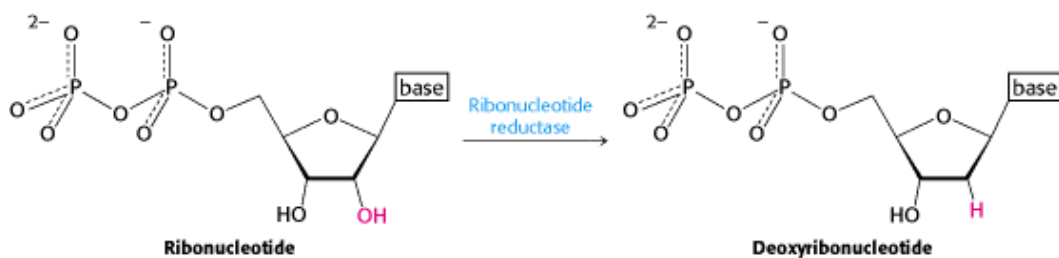


Figure 2.11. RNA and DNA Compared. Removal of the 2'-hydroxyl group from RNA to form DNA results in a backbone that is less susceptible to cleavage by hydrolysis and thus enables more-stable storage of genetic information.

What is the selective advantage of DNA over RNA as the genetic material? The genetic material must be extremely stable so that sequence information can be passed on from generation to generation without degradation. RNA itself is a remarkably stable molecule; negative charges in the sugar-phosphate backbone protect it from attack by hydroxide ions that would lead to hydrolytic cleavage. However, the 2'-hydroxyl group makes the RNA susceptible to base-catalyzed hydrolysis. The removal of the 2'-hydroxyl group from the ribose decreases the rate of hydrolysis by approximately 100-fold under neutral conditions and perhaps even more under extreme conditions. Thus, the conversion of the genetic material from RNA into DNA would have substantially increased its chemical stability.

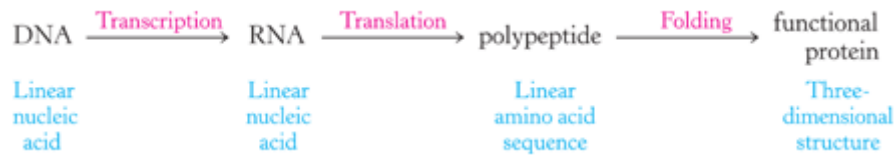
The evolutionary transition from RNA to DNA is recapitulated in the biosynthesis of DNA in modern organisms. In all cases, the building blocks used in the synthesis of DNA are synthesized from the corresponding building blocks of RNA by the action of enzymes termed *ribonucleotide reductases*. These enzymes convert ribonucleotides (a base and phosphate groups linked to a *ribose* sugar) into deoxyribonucleotides (a base and phosphates linked to *deoxyribose* sugar).



The properties of the ribonucleotide reductases vary substantially from species to species, but evidence suggests that they have a common mechanism of action and appear to have evolved from a common primordial enzyme.

The covalent structures of RNA and DNA differ in one other way. Whereas RNA contains *uracil*, DNA contains a methylated uracil derivative termed *thymine*. This modification also serves to protect the integrity of the genetic sequence, although it does so in a less direct manner. As we will see in [Chapter 27](#), the methyl group present in thymine facilitates the repair of damaged DNA, providing an additional selective advantage.

Although DNA replaced RNA in the role of storing the genetic information, RNA maintained many of its other functions. RNA still provides the template that directs polypeptide synthesis, the adaptor molecules, the catalytic activity of the ribosomes, and other functions. Thus, the genetic message is *transcribed* from DNA into RNA and then *translated* into protein.



This flow of sequence information from DNA to RNA to protein (to be considered in detail in [Chapters 5](#), [28](#), and [29](#)) applies to all modern organisms (with minor exceptions for certain viruses).

2.3. Energy Transformations Are Necessary to Sustain Living Systems

Most of the reactions that lead to the biosynthesis of nucleic acids and other biomolecules are not thermodynamically favorable under most conditions; they require an input of energy to proceed. Thus, they can proceed only if they are coupled to processes that release energy. How can energy-requiring and energy-releasing reactions be linked? How is energy from the environment transformed into a form that living systems can use? Answering these questions fundamental to biochemistry is the objective of much of this book.

2.3.1. ATP, a Common Currency for Biochemical Energy, Can Be Generated Through the Breakdown of Organic Molecules

Just as most economies simplify trade by using currency rather than bartering, biochemical systems have evolved common currencies for the exchange of energy. The most important of these currencies are molecules related to *adenosine triphosphate* (ATP) that contain an array of three linked phosphates (Figure 2.12). The bonds linking the phosphates persist in solution under a variety of conditions, but, when they are broken, an unusually large amount of energy is released that can be used to promote other processes. The roles of ATP and its use in driving other processes will be presented in detail in Chapter 14 and within many other chapters throughout this book.

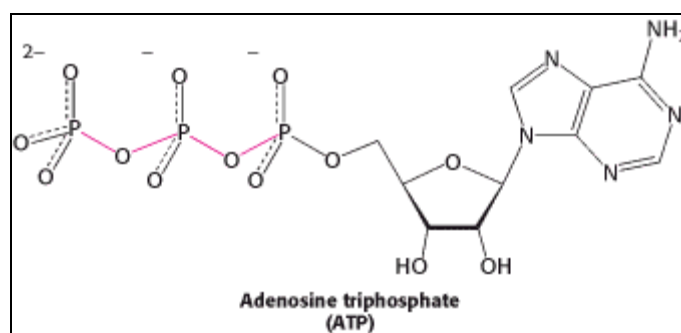
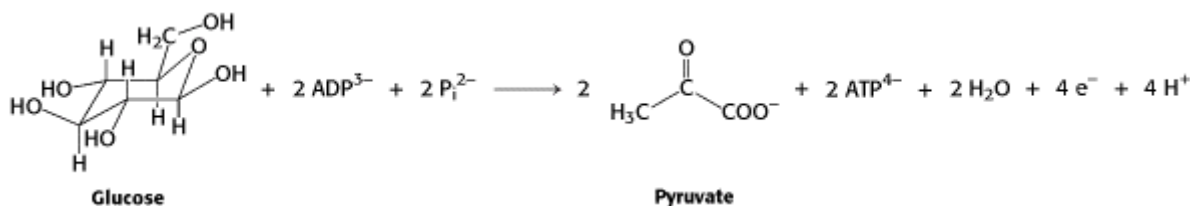


Figure 2.12. ATP, the Energy Currency of Living Systems. The phosphodiester bonds (red) release considerable energy when cleaved by hydrolysis or other processes.

ATP must be generated in appropriate quantities to be available for such reactions. The energy necessary for the synthesis of ATP can be obtained by the breakdown of other chemicals. Specific enzymes have evolved to couple these degradative processes to the phosphorylation of adenosine diphosphate (ADP) to yield ATP. Amino acids such as glycine, which were probably present in relatively large quantities in the prebiotic world and early in evolution, were likely sources of energy for ATP generation. The degradation of glycine to acetic acid may be an ATP-generation system that functioned early in evolution (Figure 2.13). In this reaction, the carbon-nitrogen bond in glycine is cleaved by reduction (the addition of electrons), and the energy released from the cleavage of this bond drives the coupling of ADP and orthophosphate (P_i) to produce ATP.

Amino acids are still broken down to produce ATP in modern organisms. However, sugars such as glucose are a more commonly utilized energy source because they are more readily metabolized and can be stored. The most important process for the direct synthesis of ATP in modern organisms is *glycolysis*, a complex process that derives energy from glucose.



Glycolysis presumably evolved as a process for ATP generation after carbohydrates such as glucose were being produced in significant quantities by other pathways. Glycolysis will be discussed in detail in [Chapter 16](#).

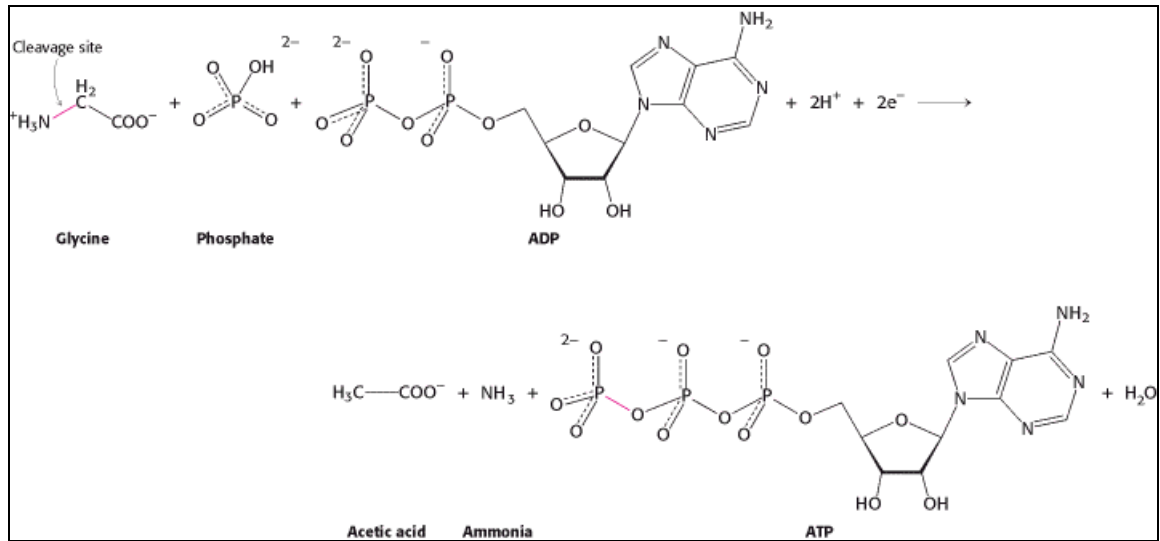
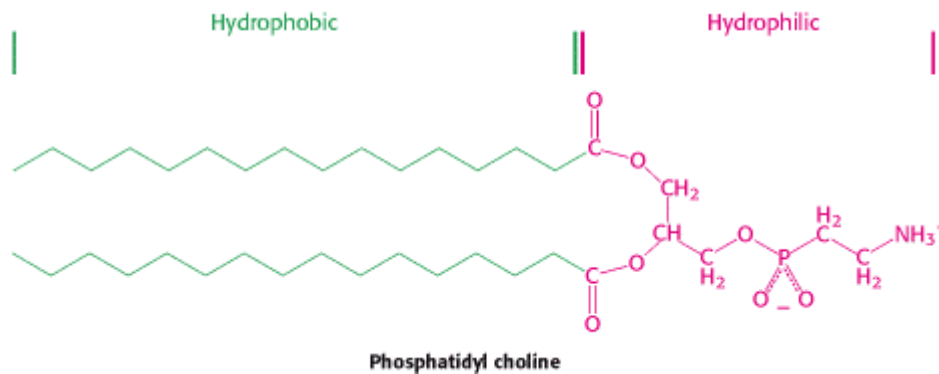


Figure 2.13. A Possible Early Method for Generating ATP. The synthesis of ATP might have been driven by the degradation of glycine.

2.3.2. Cells Were Formed by the Inclusion of Nucleic Acids Within Membranes

Modern organisms are made up of *cells*. A cell is composed of nucleic acids, proteins, and other biochemicals surrounded by a *membrane* built from lipids. These membranes completely enclose their contents, and so cells have a defined inside and outside. A typical membrane-forming lipid is phosphatidyl choline.



The most important feature of membrane-forming molecules such as phosphatidyl choline is that they are *amphipathic* - that is, they contain both *hydrophilic* (water-loving) and *hydrophobic* (water-avoiding) components. Membrane-forming molecules consist of fatty acids, whose long alkyl groups are hydrophobic, connected to shorter hydrophilic "head groups." When such lipids are in contact with water, they spontaneously aggregate to form specific structures such that the hydrophobic parts of the molecules are packed together away from water, whereas the hydrophilic parts are exposed to the aqueous solution. The structure that is important for membrane formation is the *lipid bilayer* ([Figure 2.14](#)). A bilayer is formed from two layers of lipids arranged such that the fatty acid tails of each layer interact with each other to form a hydrophobic interior while the hydrophilic head groups interact with the aqueous solution on each side. Such bilayer structures can fold onto themselves to form hollow spheres having interior compartments filled with water. The hydrophobic interior of the bilayer serves as a barrier between two aqueous phases. If such structures are formed in the presence of other molecules such as nucleic acids and proteins, these molecules can become trapped inside, thus forming cell-like structures. The structures of lipids and lipid bilayers will be considered in detail in [Chapter 12](#).

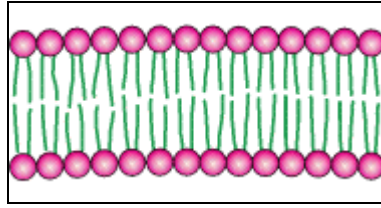


Figure 2.14. Schematic View of a Lipid Bilayer. These structures define the boundaries of cells.

At some stage in evolution, sufficient quantities of appropriate amphipathic molecules must have accumulated from biosynthetic or other processes to allow some nucleic acids to become entrapped and cell-like organisms to form. Such compartmentalization has many advantages. When the components of a cell are enclosed in a membrane, the products of enzymatic reactions do not simply diffuse away into the environment but instead are contained where they can be used by the cell that produced them. The containment is aided by the fact that nearly all biosynthetic intermediates and other biochemicals include one or more charged groups such as phosphates or carboxylates. Unlike more nonpolar or neutral molecules, charged molecules do not readily pass through lipid membranes.

2.3.3. Compartmentalization Required the Development of Ion Pumps

Despite its many advantages, the enclosure of nucleic acids and proteins within membranes introduced several complications. Perhaps the most significant were the effects of *osmosis*. Membranes are somewhat permeable to water and small nonpolar molecules, whereas they are impermeable to macromolecules such as nucleic acids. When macromolecules are concentrated inside a compartment surrounded by such a semipermeable membrane, osmotic forces drive water through the membrane into the compartment. Without counterbalancing effects, the flow of water will burst the cell ([Figure 2.15](#)).

Osmosis-

The movement of a solvent across a membrane in the direction that tends to equalize concentrations of solute on the two sides of the membrane.

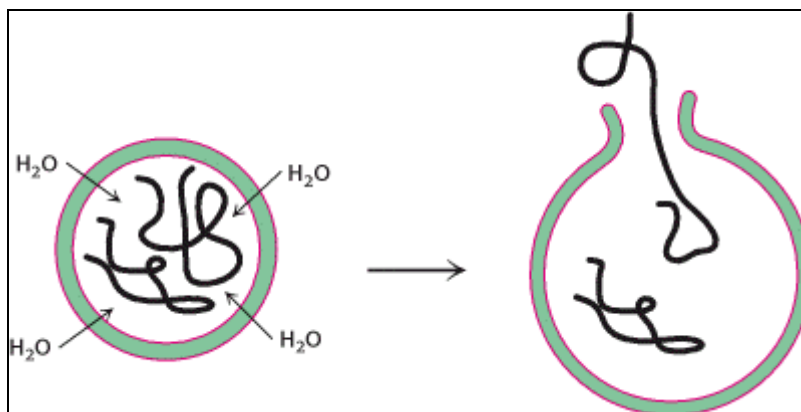


Figure 2.15. The "Osmotic Crisis." A cell consisting of macromolecules surrounded by a semipermeable membrane will take up water from outside the cell and burst.

Modern cells have two distinct mechanisms for resisting these osmotic forces. One mechanism is to toughen the cell membrane by the introduction of an additional structure such as a cell wall. However, such a chemically elaborate structure may not have evolved quickly, especially because it must completely surround a cell to be effective. The other mechanism is the use of *energy-dependent ion pumps*. These pumps can lower the concentration of ions inside a cell relative to the outside, favoring the flow of water molecules from inside to outside. The resulting unequal distribution of ions across an inherently impermeable membrane is called an *ion gradient*. Appropriate ion gradients can balance the osmotic forces and maintain a cell at a constant volume. Membrane proteins such as ion pumps will be considered in [Chapter 13](#).

Ion gradients can prevent osmotic crises, but they require energy to be produced. Most likely, an ATP-driven proton pump was the first existing component of the machinery for generating an ion gradient

(Figure 2.16). Such pumps, which are found in essentially all modern cells, hydrolyze ATP to ADP and inorganic phosphate and utilize the energy released to transport protons from the inside to the outside of a cell. The pump thus establishes a proton gradient that, in turn, can be coupled to other membrane-transport processes such as the removal of sodium ions from the cell. The proton gradient and other ion gradients generated from it act together to counteract osmotic effects and prevent the cell from swelling and bursting.

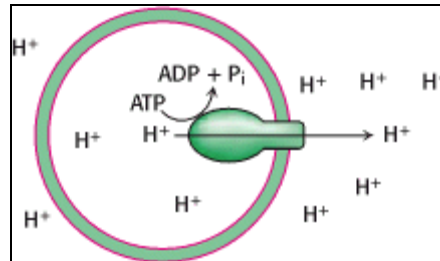


Figure 2.16. Generating an Ion Gradient. ATP hydrolysis can be used to drive the pumping of protons (or other ions) across a membrane.

2.3.4. Proton Gradients Can Be Used to Drive the Synthesis of ATP

Enzymes act to accelerate reactions, but they cannot alter the position of chemical equilibria. An enzyme that accelerates a reaction in the forward direction must also accelerate the reaction to the same extent in the reverse direction. Thus, the existence of an enzyme that utilized the hydrolysis of ATP to generate a proton gradient presented a tremendous opportunity for the evolution of alternative systems for generating ATP. Such an enzyme could synthesize ATP by reversing the process that produces the gradient. Enzymes, now called *ATP synthases*, do in fact use proton gradients to drive the bonding of ADP and P_i to form ATP (Figure 2.17). These proteins will be considered in detail in Chapter 18.

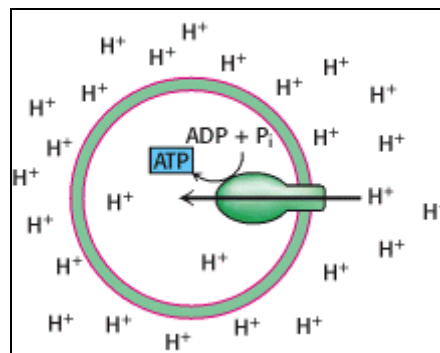


Figure 2.17. Use of Proton Gradients to Synthesize ATP. ATP can be synthesized by the action of an ATP-driven proton pump running in reverse.

Organisms have evolved a number of elaborate mechanisms for the generation of proton gradients across membranes. An example is *photosynthesis*, a process first used by bacteria and now also used by plants to harness the light energy from the sun. The essence of photosynthesis is the light-driven transfer of an electron across a membrane. The fundamental processes are illustrated in Figure 2.18.

The photosynthetic apparatus, which is embedded in a membrane, contains pigments that efficiently absorb light from the sun. The absorbed light provides the energy to promote an electron in the pigment molecule to an excited state. The high-energy electron can then jump to an appropriate acceptor molecule located in the part of the membrane facing the inside of the cell. The acceptor molecule, now reduced, binds a proton from a water molecule, generating a hydroxide ion inside the cell. The electronic "hole" left in the pigment on the outside of the membrane can then be filled by the donation of an electron from a suitable reductant on the outside of the membrane. Because the generation of a hydroxide ion inside the cell is equivalent to the generation of a proton outside the cell, a proton gradient develops across the membrane. Protons flow down this gradient through ATP synthases to generate ATP.

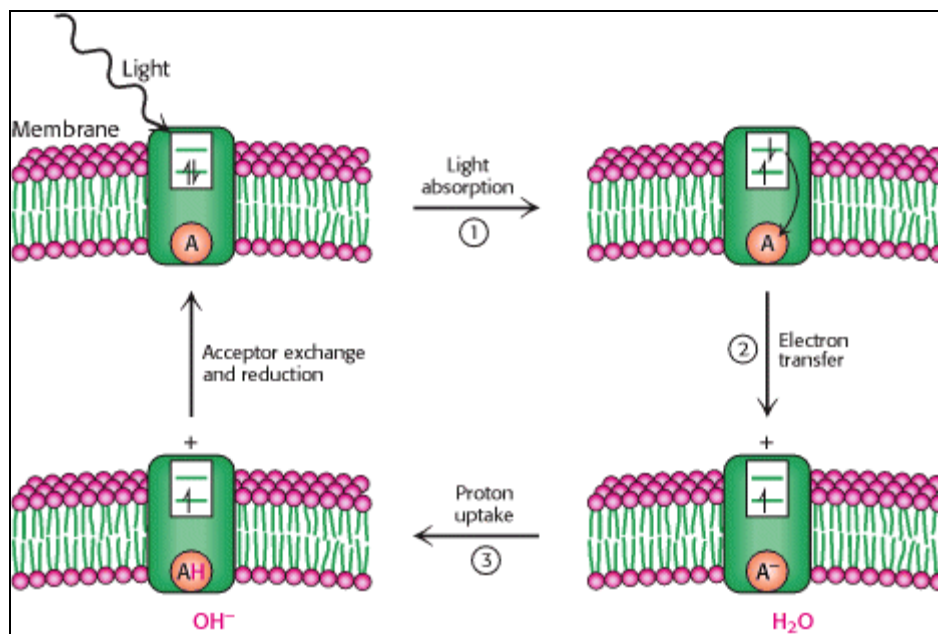
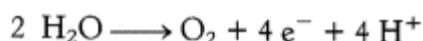


Figure 2.18. Photosynthesis. Absorption of light (1) leads to electron transfer across a membrane (2). For each electron transfer, one excess hydroxide ion is generated inside the cell (3). The process produces a proton gradient across the membrane that can drive ATP synthesis.

Photosynthesis is but one of a range of processes in different organisms that lead to ATP synthesis through the action of proteins evolutionarily related to the primordial ATP-driven pumps. In animals, the degradation of carbohydrates and other organic compounds is the source of the electron flow across membranes that can be used to develop proton gradients. The formation of ATP-generating proton gradients by fuel metabolism will be considered in [Chapter 18](#) and by light absorption in [Chapter 19](#).

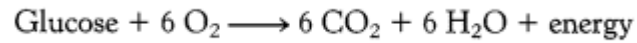
2.3.5. Molecular Oxygen, a Toxic By-Product of Some Photosynthetic Processes, Can Be Utilized for Metabolic Purposes

As stated earlier, photosynthesis generates electronic "holes" in the photosynthetic apparatus on the outside of the membrane. These holes are powerful oxidizing agents; that is, they have very high affinities for electrons and can pull electrons from many types of molecules. They can even oxidize water. Thus, for many photosynthetic organisms, the electron donor that completes the photosynthetic cycle is water. The product of water oxidation is oxygen gas - that is, molecular oxygen (O₂).

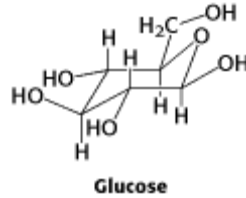


The use of water as the electron donor significantly increases the efficiency of photosynthetic ATP synthesis because the generation of one molecule of oxygen is accompanied not only by the release of four electrons (e⁻), but also by the release of four protons on one side of the membrane. Thus, an additional proton is released for each proton equivalent produced by the initial electron-transfer process, so twice as many protons are available to drive ATP synthesis. Oxygen generation will be considered in [Chapter 19](#).

Oxygen was present in only small amounts in the atmosphere before organisms evolved that could oxidize water. The "pollution" of the air with oxygen produced by photosynthetic organisms greatly affected the course of evolution. Oxygen is quite reactive and thus extremely toxic to many organisms. Many biochemical processes have evolved to protect cells from the deleterious effects of oxygen and other reactive species that can be generated from this molecule. Subsequently, organisms evolved mechanisms for taking advantage of the high reactivity of oxygen to promote favorable processes. Most important among these mechanisms are those for the oxidation of organic compounds such as glucose. Through the action of oxygen, a glucose molecule can be completely converted into carbon dioxide and water, releasing enough energy to synthesize approximately 30 molecules of ATP.



This number represents a 15-fold increase in ATP yield compared with the yield from the breakdown of glucose in the absence of oxygen in the process of glycolysis. This increased efficiency is apparent in everyday life; our muscles exhaust their fuel supply and tire quickly if they do not receive enough oxygen and are forced to use glycolysis as the sole ATP source. The role of oxygen in the extraction of energy from organic molecules will be considered in [Chapter 18](#).



2.4. Cells Can Respond to Changes in Their Environments

The environments in which cells grow often change rapidly. For example, cells may consume all of a particular food source and must utilize others. To survive in a changing world, cells evolved mechanisms for adjusting their biochemistry in response to signals indicating environmental change. The adjustments can take many forms, including changes in the activities of preexisting enzyme molecules, changes in the rates of synthesis of new enzyme molecules, and changes in membrane-transport processes.

Initially, the detection of environmental signals occurred inside cells. Chemicals that could pass into cells, either by diffusion through the cell membrane or by the action of transport proteins, and could bind directly to proteins inside the cell and modulate their activities. An example is the use of the sugar arabinose by the bacterium *Escherichia coli* (Figure 2.19). *E. coli* cells are normally unable to use arabinose efficiently as a source of energy. However, if arabinose is their only source of carbon, *E. coli* cells synthesize enzymes that catalyze the conversion of this sugar into useful forms. This response is mediated by arabinose itself. If present in sufficient quantity outside the cell, arabinose can enter the cell through transport proteins. Once inside the cell, arabinose binds to a protein called AraC. This binding alters the structure of AraC so that it can now bind to specific sites in the bacterial DNA and increase RNA transcription from genes encoding enzymes that metabolize arabinose. The mechanisms of gene regulation will be considered in Chapter 31.

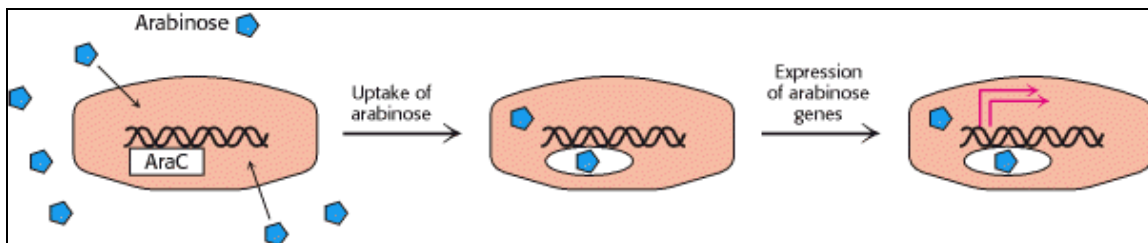
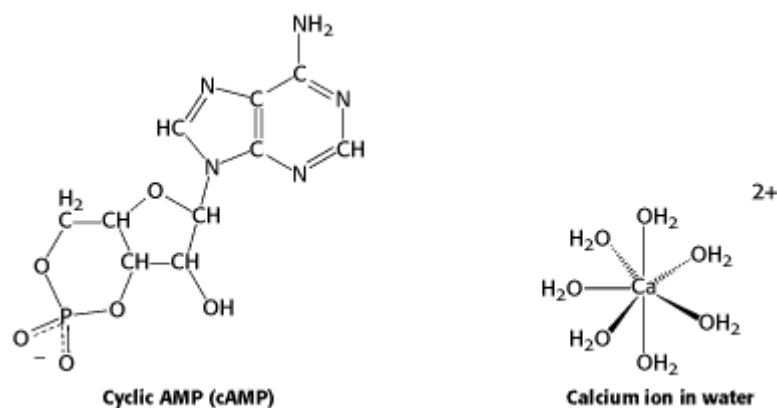


Figure 2.19. Responding to Environmental Conditions. In *E. coli* cells, the uptake of arabinose from the environment triggers the production of enzymes necessary for its utilization.

Subsequently, mechanisms appeared for detecting signals at the cell surface. Cells could thus respond to signaling molecules even if those molecules did not pass into the cell. Receptor proteins evolved that, embedded in the membrane, could bind chemicals present in the cellular environment. Binding produced changes in the protein structure that could be detected at the inside surface of the cell membrane. By this means, chemicals outside the cell could influence events inside the cell. Many of these *signal-transduction pathways* make use of substances such as cyclic adenosine monophosphate (cAMP) and calcium ion as "second messengers" that can diffuse throughout the cell, spreading the word of environmental change.



The second messengers may bind to specific sensor proteins inside the cell and trigger responses such as the activation of enzymes. Signal-transduction mechanisms will be considered in detail in Chapter 15 and in many other chapters throughout this book.

2.4.1. Filamentous Structures and Molecular Motors Enable Intracellular and Cellular Movement

The development of the ability to move was another important stage in the evolution of cells capable of adapting to a changing environment. Without this ability, nonphotosynthetic cells might have starved after consuming the nutrients available in their immediate vicinity.

Bacteria swim through the use of filamentous structures termed *flagella* that extend from their cell membranes (Figure 2.20). Each bacterial cell has several flagella, which, under appropriate conditions, form rotating bundles that efficiently propel the cell through the water. These flagella are long polymers consisting primarily of thousands of identical protein subunits. At the base of each flagellum are assemblies of proteins that act as motors to drive its rotation. The rotation of the flagellar motor is driven by the flow of protons from outside to inside the cell. Thus, energy stored in the form of a proton gradient is transduced into another form, rotatory motion.



Figure 2.20. Bacteria with Flagella. A bacterium (*Proteus mirabilis*) swims through the rotation of filamentous structures called flagella. [Fred E. Hossler/ Visuals Unlimited.]

Other mechanisms for motion, also depending on filamentous structures, evolved in other cells. The most important of these structures are *microfilaments* and *microtubules*. Microfilaments are polymers of the protein *actin*, and microtubules are polymers of two closely related proteins termed α - and β -*tubulin*. Unlike a bacterial flagellum, these filamentous structures are highly dynamic: they can rapidly increase or decrease in length through the addition or subtraction of component protein molecules. Microfilaments and microtubules also serve as tracks on which other proteins move, driven by the hydrolysis of ATP. Cells can change shape through the motion of *molecular motor proteins* along such filamentous structures that are changing in shape as a result of dynamic polymerization (Figure 2.21). Coordinated shape changes can be a means of moving a cell across a surface and are crucial to cell division. The motor proteins are also responsible for the transport of organelles and other structures within eukaryotic cells. Molecular motors will be considered in Chapter 34.

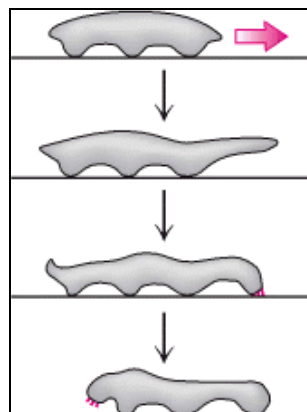


Figure 2.21. Alternative Movement. Cell mobility can be achieved by changes in cell shape.

2.4.2. Some Cells Can Interact to Form Colonies with Specialized Functions

Early organisms lived exclusively as single cells. Such organisms interacted with one another only indirectly by competing for resources in their environments. Certain of these organisms, however, developed the ability to form colonies comprising many interacting cells. In such groups, the environment of a cell is dominated by the presence of surrounding cells, which may be in direct contact with one another. These cells communicate with one another by a variety of signaling mechanisms and may respond to signals by altering enzyme activity or levels of gene expression. One result may be *cell differentiation*; differentiated cells are genetically identical but have different properties because their genes are expressed differently.

Several modern organisms are able to switch back and forth from existence as independent single cells to existence as multicellular colonies of differentiated cells. One of the most well characterized is the slime mold *Dictyostelium*. In favorable environments, this organism lives as individual cells; under conditions of starvation, however, the cells come together to form a cell aggregate. This aggregate, sometimes called a *slug*, can move as a unit to a potentially more favorable environment where it then forms a multicellular structure, termed a *fruiting body*, that rises substantially above the surface on which the cells are growing. Wind may carry cells released from the top of the fruiting body to sites where the food supply is more plentiful. On arriving in a well-stocked location, the cells grow, reproduce, and live as individual cells until the food supply is again exhausted ([Figure 2.22](#)).



Figure 2.22. Unicellular to Multicellular Transition in *Dictyostelium*. This scanning electron micrograph shows the transformation undergone by the slime mold *Dictyostelium*. Hundreds of thousands of single cells aggregate to form a migrating slug, seen in the lower left. Once the slug comes to a stop, it gradually elongates to form the fruiting body. [Courtesy of M. J. Grimsom and R. L. Blanton, Texas Tech University.]

The transition from unicellular to multicellular growth is triggered by cell-cell communication and reveals much about signaling processes between and within cells. Under starvation conditions, *Dictyostelium* cells release the signal molecule cyclic AMP. This molecule signals surrounding cells by binding to a membrane-bound protein receptor on the cell surface. The binding of cAMP molecules to these receptors triggers several responses, including movement in the direction of higher cAMP concentration, as well as the generation and release of additional cAMP molecules ([Figure 2.23](#)).

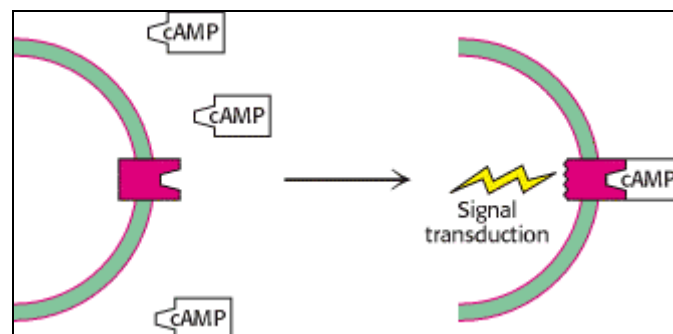


Figure 2.23. Intracellular Signaling. Cyclic AMP, detected by cell-surface receptors, initiates the formation of aggregates in *Dictyostelium*.

The cells aggregate by following cAMP gradients. Once in contact, they exchange additional signals and then differentiate into distinct *cell types*, each of which expresses the set of genes appropriate for its eventual role in forming the fruiting body (Figure 2.24). The life cycles of organisms such as *Dictyostelium* foreshadow the evolution of organisms that are multicellular throughout their lifetimes. It is also interesting to note the cAMP signals starvation in many organisms, including human beings.

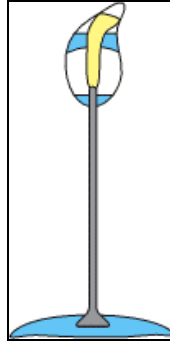


Figure 2.24. Cell Differentiation in *Dictyostelium*. The colors represent the distribution of cell types expressing similar sets of genes in the *Dictyostelium* fruiting body.

2.4.3. The Development of Multicellular Organisms Requires the Orchestrated Differentiation of Cells

The fossil record indicates that macroscopic, multicellular organisms appeared approximately 600 million years ago. Most of the organisms familiar to us consist of many cells. For example, an adult human being contains approximately 100,000,000,000,000 cells. The cells that make up different organs are distinct and, even within one organ, many different cell types are present. Nonetheless, the DNA sequence in each cell is identical. The differences between cell types are the result of differences in how these genes are expressed.

Each multicellular organism begins as a single cell. For this cell to develop into a complex organism, the embryonic cells must follow an intricate program of regulated gene expression, cell division, and cell movement. The developmental program relies substantially on the responses of cells to the environment created by neighboring cells. Cells in specific positions within the developing embryo divide to form particular tissues, such as muscle. Developmental pathways have been extensively studied in a number of organisms, including the nematode *Caenorhabditis elegans* (Figure 2.25), a 1-mm-long worm containing 959 cells. A detailed map describing the fate of each cell in *C. elegans* from the fertilized egg to the adult is shown in Figure 2.26. Interestingly, proper development requires not only cell division but also the death of specific cells at particular points in time through a process called programmed cell death or *apoptosis*.



Figure 2.25. The Nematode *Caenorhabditis elegans*. This organism serves as a useful model for development. [Sinclair Stammers Science Photo Library/Photo Researchers.]

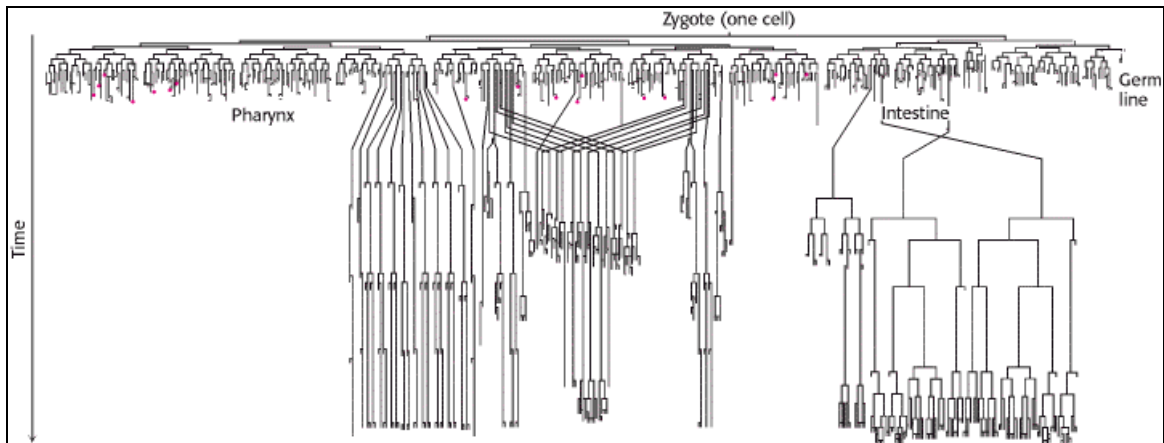


Figure 2.26. Developmental Pathways of *C. elegans*. The nematode develops from a single cell, called a zygote, into a complex organism. The fate of each individual cell in *C. elegans* is known and can be followed by referring to the cell-lineage diagram. The labels indicate cells that form specific organs. Cells that undergo programmed cell death are shown in red.

Investigations of genes and proteins that control development in a wide range of organisms have revealed a great many common features. Many of the molecules that control human development are evolutionarily related to those in relatively simple organisms such as *C. elegans*. Thus, solutions to the problem of controlling development in multicellular organisms arose early in evolution and have been adapted many times in the course of evolution, generating the great diversity of complex organisms.

2.4.4. The Unity of Biochemistry Allows Human Biology to Be Effectively Probed Through Studies of Other Organisms

All organisms on Earth have a common origin (Figure 2.27). How could complex organisms such as human beings have evolved from the simple organisms that existed at life's start? The path outlined in this chapter reveals that most of the fundamental processes of biochemistry were largely fixed early in the history of life. The complexity of organisms such as human beings is manifest, at a biochemical level, in the interactions between overlapping and competing pathways, which lead to the generation of intricately connected groups of specialized cells. The evolution of biochemical and physiological complexity is made possible by the effects of gene duplication followed by specialization. Paradoxically, the reliance on gene duplication also makes this complexity easier to comprehend. Consider, for example, the protein kinases - enzymes that transfer phosphoryl groups from ATP to specific amino acids in proteins. These enzymes play essential roles in many signal-transduction pathways and in the control of cell growth and differentiation. The human genome encodes approximately 500 proteins of this class; even a relatively simple, unicellular organism such as brewer's yeast has more than 100 protein kinases. Yet each of these enzymes is the evolutionary descendant of a common ancestral enzyme. Thus, *we can learn much about the essential behavior of this large collection of proteins through studies of a single family member*. After the essential behavior is understood, we can evaluate the specific adaptations that allow each family member to perform its particular biological functions.

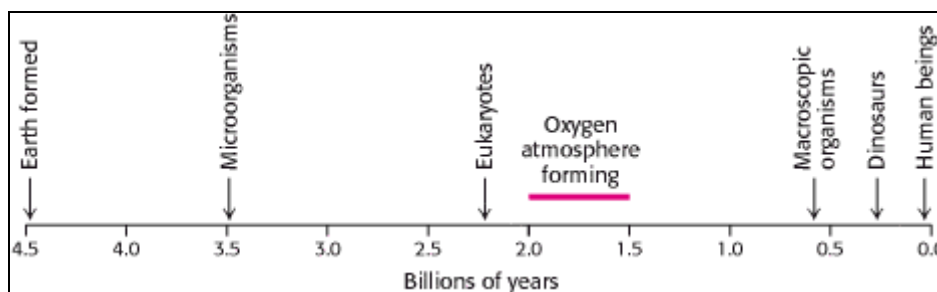


Figure 2.27. A Possible Time Line for Biochemical Evolution. Key events are indicated.

Most central processes in biology have been characterized first in relatively simple organisms, often through a combination of genetic, physiological, and biochemical studies. Many of the processes controlling early embryonic development were elucidated by the results of studies of the fruit fly. The events controlling DNA replication and the cell cycle were first deciphered in yeast. Investigators can

now test the functions of particular proteins in mammals by disrupting the genes that encode these proteins in mice and examining the effects. The investigations of organisms linked to us by common evolutionary pathways are powerful tools for exploring all of biology and for developing new understanding of normal human function and disease.

Summary

Key Organic Molecules Are Used by Living Systems

The evolution of life required a series of transitions, beginning with the generation of organic molecules that could serve as the building blocks for complex biomolecules. How these molecules arose is a matter of conjecture, but experiments have established that they could have formed under hypothesized prebiotic conditions.

Evolution Requires Reproduction, Variation, and Selective Pressure

The next major transition in the evolution of life was the formation of replicating molecules. Replication, coupled with variation and selective pressure, marked the beginning of evolution. Variation was introduced by a number of means, from simple base substitutions to the duplication of entire genes. RNA appears to have been an early replicating molecule. Furthermore, some RNA molecules possess catalytic activity. However, the range of reactions that RNA is capable of catalyzing is limited. With time, the catalytic activity was transferred to proteins - linear polymers of the chemically versatile amino acids. RNA directed the synthesis of these proteins and still does in modern organisms through the development of a genetic code, which relates base sequence to amino acid sequence. Eventually, RNA lost its role as the gene to the chemically similar but more stable nucleic acid DNA. In modern organisms, RNA still serves as the link between DNA and protein.

Energy Transformations Are Necessary to Sustain Living Systems

Another major transition in evolution was the ability to transform environmental energy into forms capable of being used by living systems. ATP serves as the cellular energy currency that links energy-yielding reactions with energy-requiring reactions. ATP itself is a product of the oxidation of fuel molecules, such as amino acids and sugars. With the evolution of membranes - hydrophobic barriers that delineate the borders of cells - ion gradients were required to prevent osmotic crises. These gradients were formed at the expense of ATP hydrolysis. Later, ion gradients generated by light or the oxidation of fuel molecules were used to synthesize ATP.

Cells Can Respond to Changes in Their Environments

The final transition was the evolution of sensing and signaling mechanisms that enabled a cell to respond to changes in its environment. These signaling mechanisms eventually led to cell-cell communication, which allowed the development of more-complex organisms. The record of much of what has occurred since the formation of primitive organisms is written in the genomes of extant organisms. Knowledge of these genomes and the mechanisms of evolution will enhance our understanding of the history of life on Earth as well as our understanding of existing organisms.

Key Terms

prebiotic world

reproduction

variation

competition

selective pressure

catalyst

enzyme

ribozyme

RNA world

proteins

genetic code

translation

gene

mutation

gene duplication

ATP (adenosine triphosphate)

membrane

ion pump

ion gradient

photosynthesis

signal transduction pathway

molecular motor protein

cell differentiation

unity of biochemistry

Problems

1. **Finding the fragments.** Identify the likely source (CH_4 , NH_3 , H_2O , or H_2) of each atom in alanine generated in the Miller-Urey experiment.

Answer:

The amino group comes from ammonia. All of the carbon atoms are derived from methane. The hydrogen atoms bonded to the carbon atoms remain with the methane during bond formation or they may come from hydrogen gas. The oxygen atoms of the carboxyl group are from water.

2. **Following the populations.** In an experiment analogous to the Spiegelman experiment, suppose that a population of RNA molecules consists of 99 identical molecules, each of which replicates once in 15 minutes, and 1 molecule that replicates once in 5 minutes. Estimate the composition of the population after 1, 10, and 25 "generations" if a generation is defined as 15 minutes of replication. Assume that all necessary components are readily available.

Answer:

We start with 99 identical RNA molecules (which we will call L) that replicate in 15 minutes and 1 variant molecule (which we will call S) that replicates in 5 minutes. After 15 minutes, we will have $2 \times 99 = 198$ molecules of L and $2^3 \times 1 = 8$ molecules of S since it replicates 3 times in 15 minutes. Thus, the population now contains $8/(8 + 198) = 3.9\%$ S. After 10 generations, each molecule of L will have replicated 10 times while each molecule of S will have replicated 30 times. The population will contain $1 \times 2^{30}/(1 \times 2^{30} + 99 \times 2^{10}) = 99.991\%$ S. After 25 generations, the population will contain essentially all S and no L.

3. **Selective advantage.** Suppose that a replicating RNA molecule has a mutation (genotypic change) and the phenotypic result is that it binds nucleotide monomers more tightly than do other RNA molecules in its population. What might the selective advantage of this mutation be? Under what conditions would you expect this selective advantage to be most important?

Answer:

The mutation permits more efficient use of substrates and thus would be most beneficial when substrate is present in low concentrations.

4. **Opposite of randomness.** Ion gradients prevent osmotic crises, but they require energy to be produced. Why does the formation of a gradient require an energy input?

Answer:

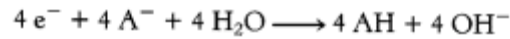
The formation of an ion gradient requires a reduction in entropy, which requires an input of free energy.

5. **Coupled gradients.** How could a proton gradient with a higher concentration of protons inside a cell be used to pump ions out of a cell?

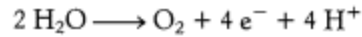
Answer:

The decrease in free energy that results when the protons run down the ion gradient could be used to pump ions out of the cell against a concentration gradient, an energy-requiring process.

6. **Proton counting.** Consider the reactions that take place across a photosynthetic membrane. On one side of the membrane, the following reaction takes place:



whereas, on the other side of the membrane, the reaction is:



How many protons are made available to drive ATP synthesis for each reaction cycle?

Answer:

Two protons per electron, or eight. The generation of four hydroxyl ions (OH^{-}) is equivalent to the generation of four protons (H^{+}) on the other side of the membrane from which the reaction is taking place. The oxidation of water produces four more protons.

7. **An alternative pathway.** To respond to the availability of sugars such as arabinose, a cell must have at least two types of proteins: a transport protein to allow the arabinose to enter the cell and a gene-control protein, which binds the arabinose and modifies gene expression. To respond to the availability of some very hydrophobic molecules, a cell requires only one protein. Which one and why?

Answer:

Only the gene-control protein is necessary. The hydrophobic molecule can pass through the membrane on its own.

8. **How many divisions?** In the development pathway of *C. elegans*, cell division is initially synchronous - that is, all cells divide at the same rate. Later in development, some cells divide more frequently than do others. How many times does each cell divide in the synchronous period? Refer to Figure 2.26.

Answer:

Approximately eight times.

Selected Readings

Where to start

N.R. Pace. 2000. The universal nature of biochemistry *Proc. Natl. Acad. Sci. U. S. A.* 98: 805-808. ([PubMed](#)) ([Full Text in PMC](#))

L.E. Orgel. 1987. Evolution of the genetic apparatus: A review *Cold Spring Harbor Symp. Quant. Biol.* 52: 9-16. ([PubMed](#))

A. Lazcano and S.L. Miller. 1996. The origin and early evolution of life: Prebiotic chemistry, the pre-RNA world, and time *Cell* 85: 793-798. ([PubMed](#))

L.E. Orgel. 1998. The origin of life: A review of facts and speculations *Trends Biochem. Sci.* 23: 491-495. ([PubMed](#))

Books

Darwin, C., 1975. *On the Origin of Species, a Facsimile of the First Edition*. Harvard University Press.

Gesteland, R. F., Cech, T., and Atkins, J. F., 1999. *The RNA World*. Cold Spring Harbor Laboratory Press.

Dawkins, R., 1996. *The Blind Watchmaker*. Norton.

Smith, J. M., and Szathmáry, E., 1995. *The Major Transitions in Evolution*. W. H. Freeman and Company.

Prebiotic chemistry

S.L. Miller. 1987. Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harbor Symp. Quant. Biol.* 52: 17-27. ([PubMed](#))

F.H. Westheimer. 1987. Why nature chose phosphates *Science* 235: 1173-1178. ([PubMed](#))

M. Levy and S.L. Miller. 1998. The stability of the RNA bases: Implications for the origin of life *Proc. Natl. Acad. Sci. U. S. A.* 95: 7933-7938. ([PubMed](#)) ([Full Text in PMC](#))

R. Sanchez, J. Ferris, and L.E. Orgel. 1966. Conditions for purine synthesis: Did prebiotic synthesis occur at low temperatures? *Science* 153: 72-73. ([PubMed](#))

In vitro evolution

D.R. Mills, R.L. Peterson, and S. Spiegelman. 1967. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule *Proc. Natl. Acad. Sci. U. S. A.* 58: 217-224. ([PubMed](#))

R. Levisohn and S. Spiegelman. 1969. Further extracellular Darwinian experiments with replicating RNA molecules: Diverse variants isolated under different selective conditions *Proc. Natl. Acad. Sci. U. S. A.* 63: 805-811. ([PubMed](#))

D.S. Wilson and J.W. Szostak. 1999. In vitro selection of functional nucleic acids *Annu. Rev. Biochem.* 68: 611-647. ([PubMed](#))

Replication and catalytic RNA

T.R. Cech. 1993. The efficiency and versatility of catalytic RNA: Implications for an RNA world *Gene* 135: 33-36. ([PubMed](#))

L.E. Orgel. 1992. Molecular replication *Nature* 358: 203-209. ([PubMed](#))

W.S. Zielinski and L.E. Orgel. 1987. Autocatalytic synthesis of a tetranucleotide analogue *Nature* 327: 346-347. ([PubMed](#))

K.E. Nelson, M. Levy, and S.L. Miller. 2000. Peptide nucleic acids rather than RNA may have been the first genetic molecule *Proc. Natl. Acad. Sci. U. S. A.* 97: 3868-3871. ([PubMed](#)) ([Full Text in PMC](#))

Transition from RNA to DNA

P. Reichard. 1997. The evolution of ribonucleotide reduction *Trends Biochem. Sci.* 22: 81-85. ([PubMed](#))

A. Jordan and P. Reichard. 1998. Ribonucleotide reductases *Annu. Rev. Biochem.* 67: 71-98. ([PubMed](#))

Membranes

T.H. Wilson and P.C. Maloney. 1976. Speculations on the evolution of ion transport mechanisms *Fed. Proc.* 35: 2174-2179. ([PubMed](#))

T.H. Wilson and E.C. Lin. 1980. Evolution of membrane bioenergetics *J. Supramol. Struct.* 13: 421-446. ([PubMed](#))

Multicellular organisms and development

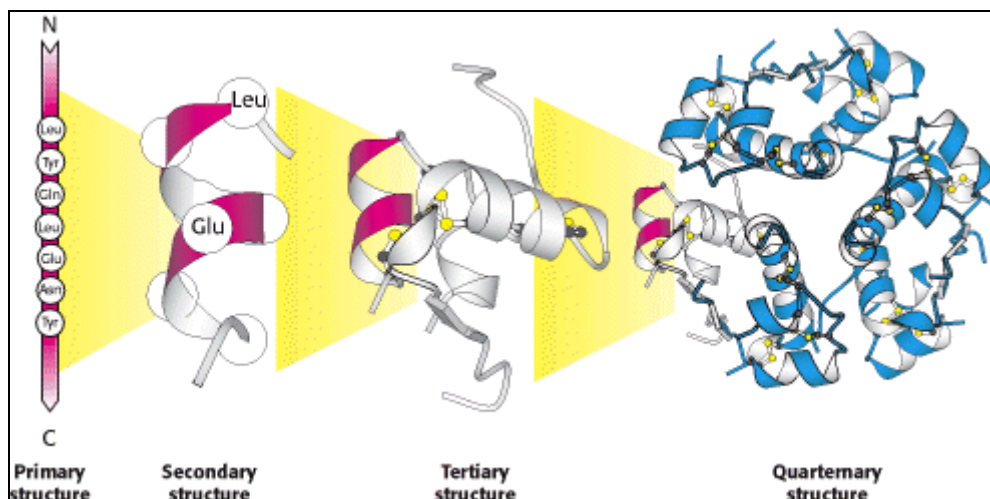
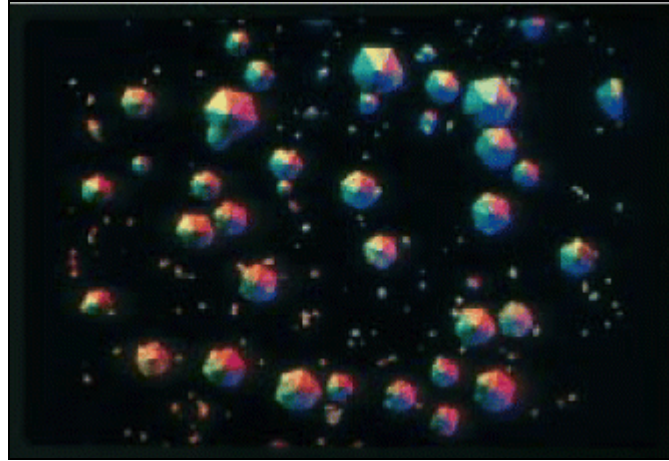
G. Mangiarotti, S. Bozzaro, S. Landfear, and H.F. Lodish. 1983. Cell-cell contact, cyclic AMP, and gene expression during development of *Dictyostelium discoideum* *Curr. Top. Dev. Biol.* 18: 117-154. ([PubMed](#))

C. Kenyon. 1988. The nematode *Caenorhabditis elegans* *Science* 240: 1448-1453. ([PubMed](#))

J. Hodgkin, R.H. Plasterk, and R.H. Waterston. 1995. The nematode *Caenorhabditis elegans* and its genome *Science* 270: 410-414. ([PubMed](#))

3. Protein Structure and Function

Proteins are the most versatile macromolecules in living systems and serve crucial functions in essentially all biological processes. They function as catalysts, they transport and store other molecules such as oxygen, they provide mechanical support and immune protection, they generate movement, they transmit nerve impulses, and they control growth and differentiation. Indeed, much of this text will focus on understanding what proteins do and how they perform these functions.



Crystals of human insulin. Insulin is a protein hormone, crucial for maintaining blood sugar at appropriate levels. (Below) Chains of amino acids in a specific sequence (the primary structure) define a protein like insulin. These chains fold into well-defined structures (the tertiary structure) - in this case a single insulin molecule. Such structures assemble with other chains to form arrays such as the complex of six insulin molecules shown at the far right (the quaternary structure). These arrays can often be induced to form well-defined crystals (photo at left), which allows determination of these structures in detail. [(Left) Alfred Pasieka/Peter Arnold.]

Several key properties enable proteins to participate in such a wide range of functions.

1. Proteins are linear polymers built of monomer units called amino acids. The construction of a vast array of macromolecules from a limited number of monomer building blocks is a recurring theme in biochemistry. Does protein function depend on the linear sequence of amino acids? The function of a protein is directly dependent on its three-dimensional structure ([Figure 3.1](#)). Remarkably, proteins spontaneously fold up into three-dimensional structures that are determined by the sequence of amino acids in the protein polymer. Thus, *proteins are the embodiment of the transition from the one-dimensional world of sequences to the three-dimensional world of molecules capable of diverse activities.*

2. Proteins contain a wide range of functional groups. These functional groups include alcohols, thiols, thioethers, carboxylic acids, carboxamides, and a variety of basic groups. When combined in various sequences, this array of functional groups accounts for the broad spectrum of protein function. For

instance, the chemical reactivity associated with these groups is essential to the function of *enzymes*, the proteins that catalyze specific chemical reactions in biological systems (see [Chapters 8-10](#)).

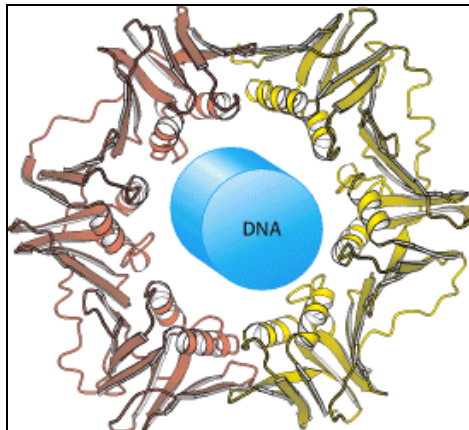


Figure 3.1. Structure Dictates Function. A protein component of the DNA replication machinery surrounds a section of DNA double helix. The structure of the protein allows large segments of DNA to be copied without the replication machinery dissociating from the DNA.

3. Proteins can interact with one another and with other biological macromolecules to form complex assemblies. The proteins within these assemblies can act synergistically to generate capabilities not afforded by the individual component proteins ([Figure 3.2](#)). These assemblies include macro-molecular machines that carry out the accurate replication of DNA, the transmission of signals within cells, and many other essential processes.

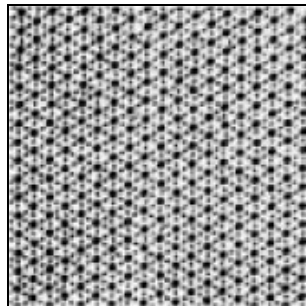


Figure 3.2. A Complex Protein Assembly. An electron micrograph of insect flight tissue in cross section shows a hexagonal array of two kinds of protein filaments. [Courtesy of Dr. Michael Reedy.]

4. Some proteins are quite rigid, whereas others display limited flexibility. Rigid units can function as structural elements in the cytoskeleton (the internal scaffolding within cells) or in connective tissue. Parts of proteins with limited flexibility may act as hinges, springs, and levers that are crucial to protein function, to the assembly of proteins with one another and with other molecules into complex units, and to the transmission of information within and between cells ([Figure 3.3](#)).

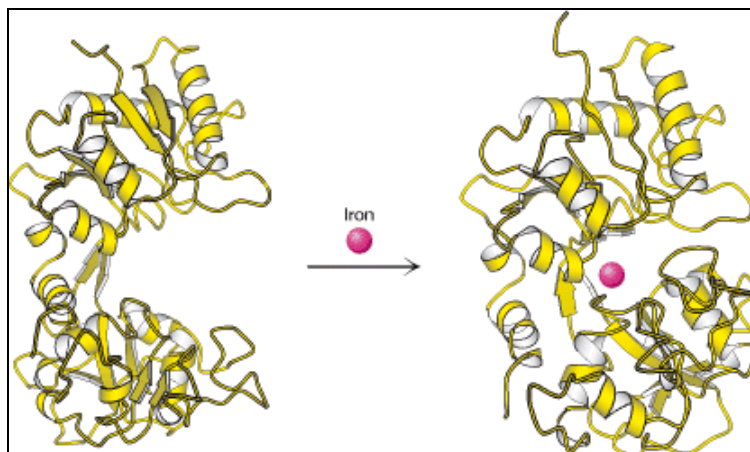


Figure 3.3. Flexibility and Function. Upon binding iron, the protein lactoferrin undergoes conformational changes that allow other molecules to distinguish between the iron-free and the iron-bound forms.

3.1. Proteins Are Built from a Repertoire of 20 Amino Acids

Amino acids are the building blocks of proteins. An α -amino acid consists of a central carbon atom, called the α carbon, linked to an amino group, a carboxylic acid group, a hydrogen atom, and a distinctive R group. The R group is often referred to as the *side chain*. With four different groups connected to the tetrahedral α -carbon atom, α -amino acids are *chiral*; the two mirror-image forms are called the L isomer and the D isomer (Figure 3.4).

Notation for distinguishing stereoisomers

The four different substituents of an asymmetric carbon atom are assigned a priority according to atomic number. The lowest-priority substituent, often hydrogen, is pointed away from the viewer. The configuration about the carbon is called *S*, from the Latin *sinis-ter* for "left," if the progression from the highest to the lowest priority is counterclockwise. The configuration is called *R*, from the Latin *rectus* for "right," if the progression is clockwise.

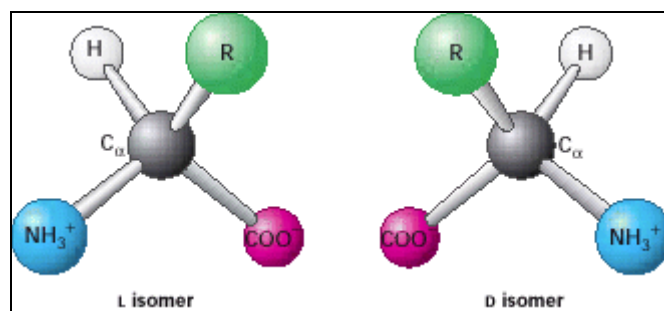


Figure 3.4. The L and D Isomers of Amino Acids. R refers to the side chain. The L and D isomers are mirror images of each other.

Only L amino acids are constituents of proteins. For almost all amino acids, the L isomer has *S* (rather than *R*) absolute configuration (Figure 3.5). Although considerable effort has gone into understanding why amino acids in proteins have this absolute configuration, no satisfactory explanation has been arrived at. It seems plausible that the selection of L over D was arbitrary but, once made, was fixed early in evolutionary history.

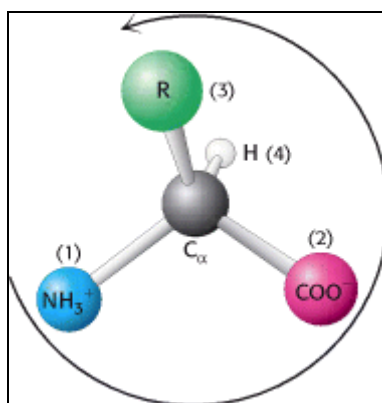


Figure 3.5. Only L Amino Acids Are Found in Proteins. Almost all L amino acids have an *S* absolute configuration (from the Latin *sinister* meaning "left"). The counterclockwise direction of the arrow from highest- to lowest-priority substituents indicates that the chiral center is of the *S* configuration.

Amino acids in solution at neutral pH exist predominantly as *dipolar ions* (also called *zwitterions*). In the dipolar form, the amino group is protonated ($-\text{NH}_3^+$) and the carboxyl group is deprotonated ($-\text{COO}^-$). The ionization state of an amino acid varies with pH (Figure 3.6). In acid solution (e.g., pH 1), the amino group is protonated ($-\text{NH}_3^+$) and the carboxyl group is not dissociated ($-\text{COOH}$). As the pH is raised, the carboxylic acid is the first group to give up a proton, inasmuch as its $\text{p}K_a$ is near 2. The dipolar form

persists until the pH approaches 9, when the protonated amino group loses a proton. For a review of acid-base concepts and pH, see the appendix to this chapter.

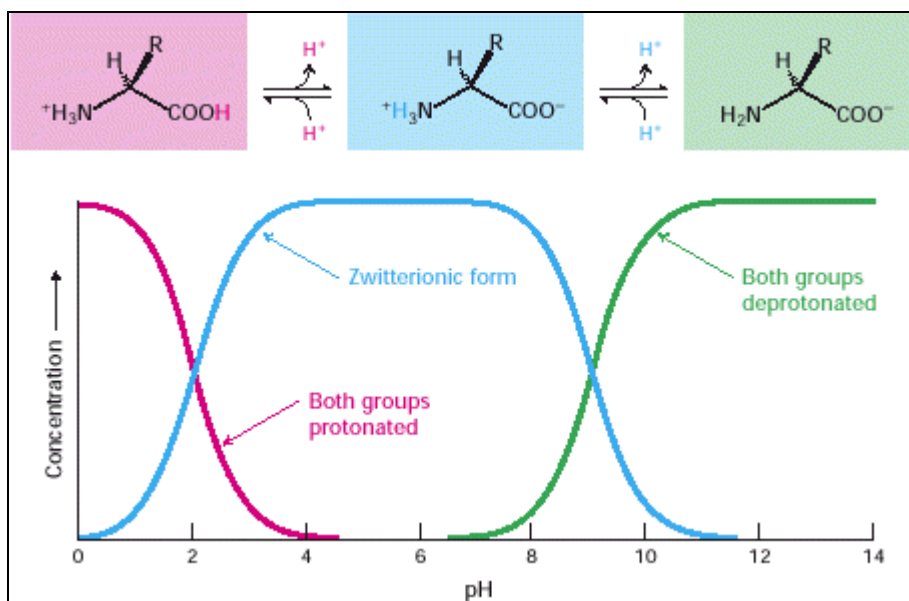


Figure 3.6. Ionization State as a Function of pH. The ionization state of amino acids is altered by a change in pH. The zwitterionic form predominates near physiological pH.

Twenty kinds of side chains varying in *size*, *shape*, *charge*, *hydrogen-bonding capacity*, *hydrophobic character*, and *chemical reactivity* are commonly found in proteins. Indeed, all proteins in all species - bacterial, archaeal, and eukaryotic - are constructed from the same set of 20 amino acids. This fundamental alphabet of proteins is several billion years old. The remarkable range of functions mediated by proteins results from the diversity and versatility of these 20 building blocks. Understanding how this alphabet is used to create the intricate three-dimensional structures that enable proteins to carry out so many biological processes is an exciting area of biochemistry and one that we will return to in [Section 3.6](#).

Let us look at this set of amino acids. The simplest one is *glycine*, which has just a hydrogen atom as its side chain. With two hydrogen atoms bonded to the α -carbon atom, glycine is unique in being *achiral*. *Alanine*, the next simplest amino acid, has a methyl group ($-CH_3$) as its side chain ([Figure 3.7](#)).

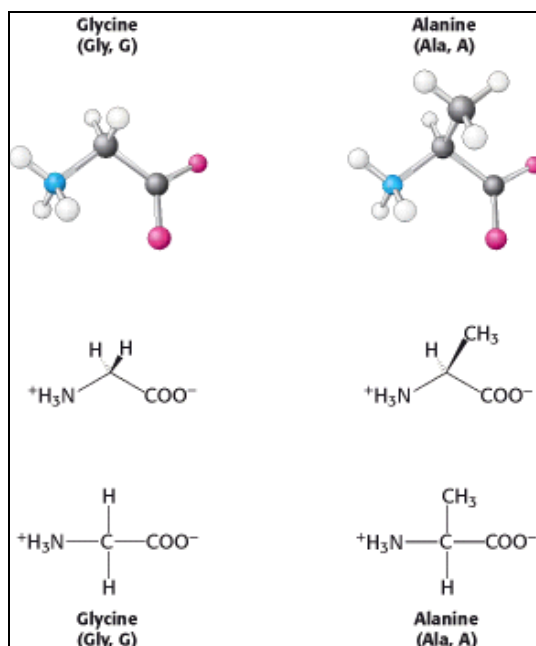


Figure 3.7. Structures of Glycine and Alanine. (Top) Ball-and-stick models show the arrangement of atoms and bonds in space. (Middle) Stereochemically realistic formulas show the geometrical arrangement of bonds around atoms (see Chapters 1 Appendix). (Bottom) Fischer projections show all bonds as being perpendicular for a simplified representation (see Chapters 1 Appendix).

Larger hydrocarbon side chains are found in *valine*, *leucine*, and *isoleucine* (Figure 3.8). *Methionine* contains a largely *aliphatic* side chain that includes a *thioether* (-S-) group. The side chain of *isoleucine* includes an additional chiral center; only the isomer shown in Figure 3.8 is found in proteins. The larger aliphatic side chains are *hydrophobic* - that is, they tend to cluster together rather than contact water. The three-dimensional structures of water-soluble proteins are stabilized by this tendency of hydrophobic groups to come together, called *the hydrophobic effect* (see Section 1.3.4). The different sizes and shapes of these hydrocarbon side chains enable them to pack together to form compact structures with few holes. *Proline* also has an aliphatic side chain, but it differs from other members of the set of 20 in that its side chain is bonded to both the nitrogen and the α -carbon atoms (Figure 3.9). *Proline* markedly influences protein architecture because its ring structure makes it more conformationally restricted than the other amino acids.

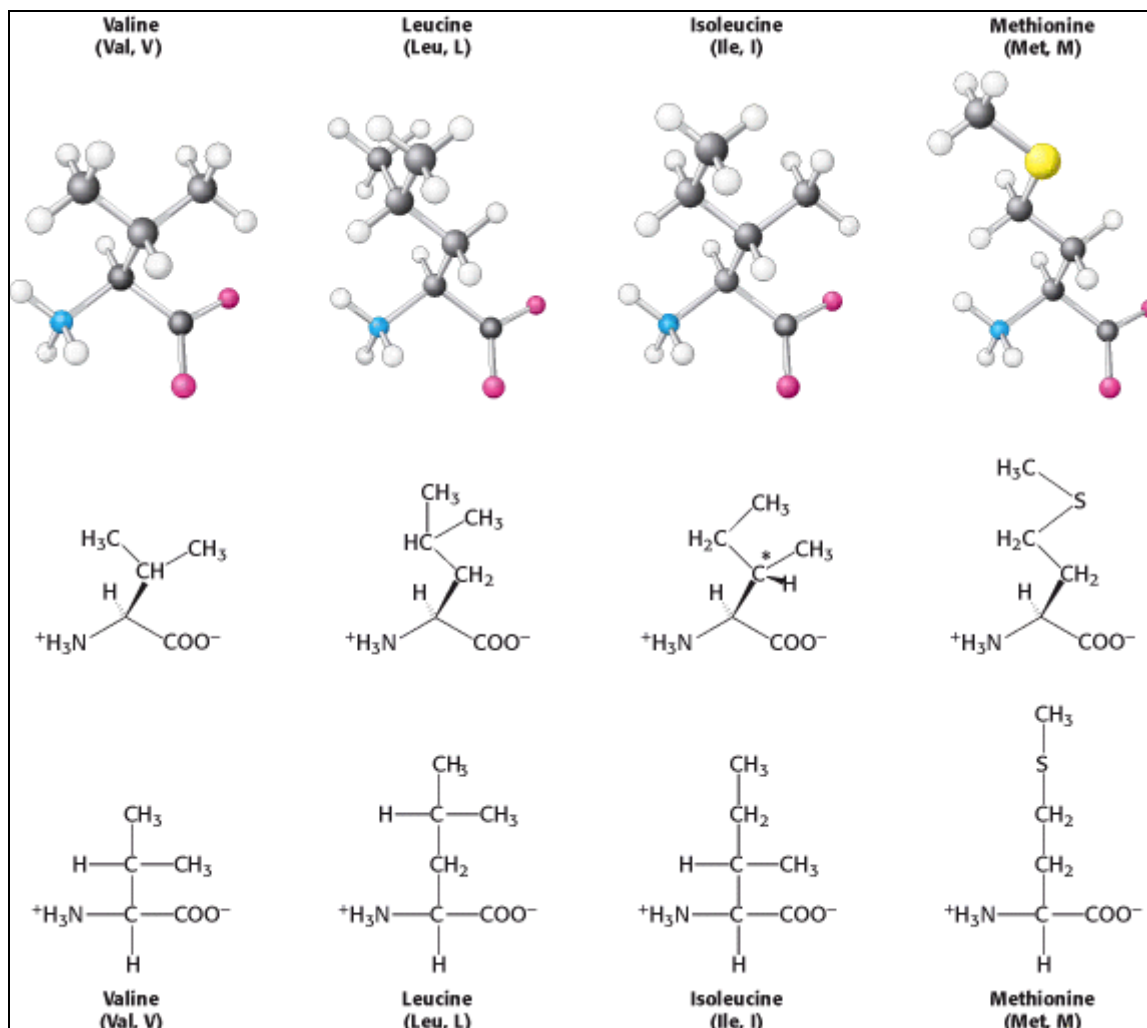


Figure 3.8. Amino Acids with Aliphatic Side Chains. The additional chiral center of isoleucine is indicated by an asterisk.

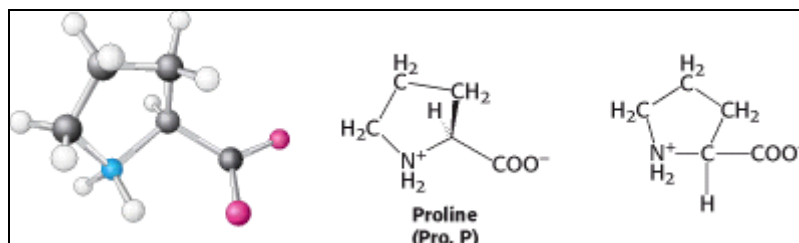


Figure 3.9. Cyclic Structure of Proline. The side chain is joined to both the α carbon and the amino group.

Three amino acids with relatively simple *aromatic side chains* are part of the fundamental repertoire (Figure 3.10). *Phenylalanine*, as its name indicates, contains a phenyl ring attached in place of one of the hydrogens of alanine. The aromatic ring of *tyrosine* contains a hydroxyl group. This hydroxyl group is reactive, in contrast with the rather inert side chains of the other amino acids discussed thus far.

Tryptophan has an indole ring joined to a methylene ($-\text{CH}_2-$) group; the indole group comprises two fused rings and an NH group. Phenylalanine is purely hydrophobic, whereas tyrosine and tryptophan are less so because of their hydroxyl and NH groups. The aromatic rings of tryptophan and tyrosine contain delocalized π electrons that strongly absorb ultraviolet light (Figure 3.11).

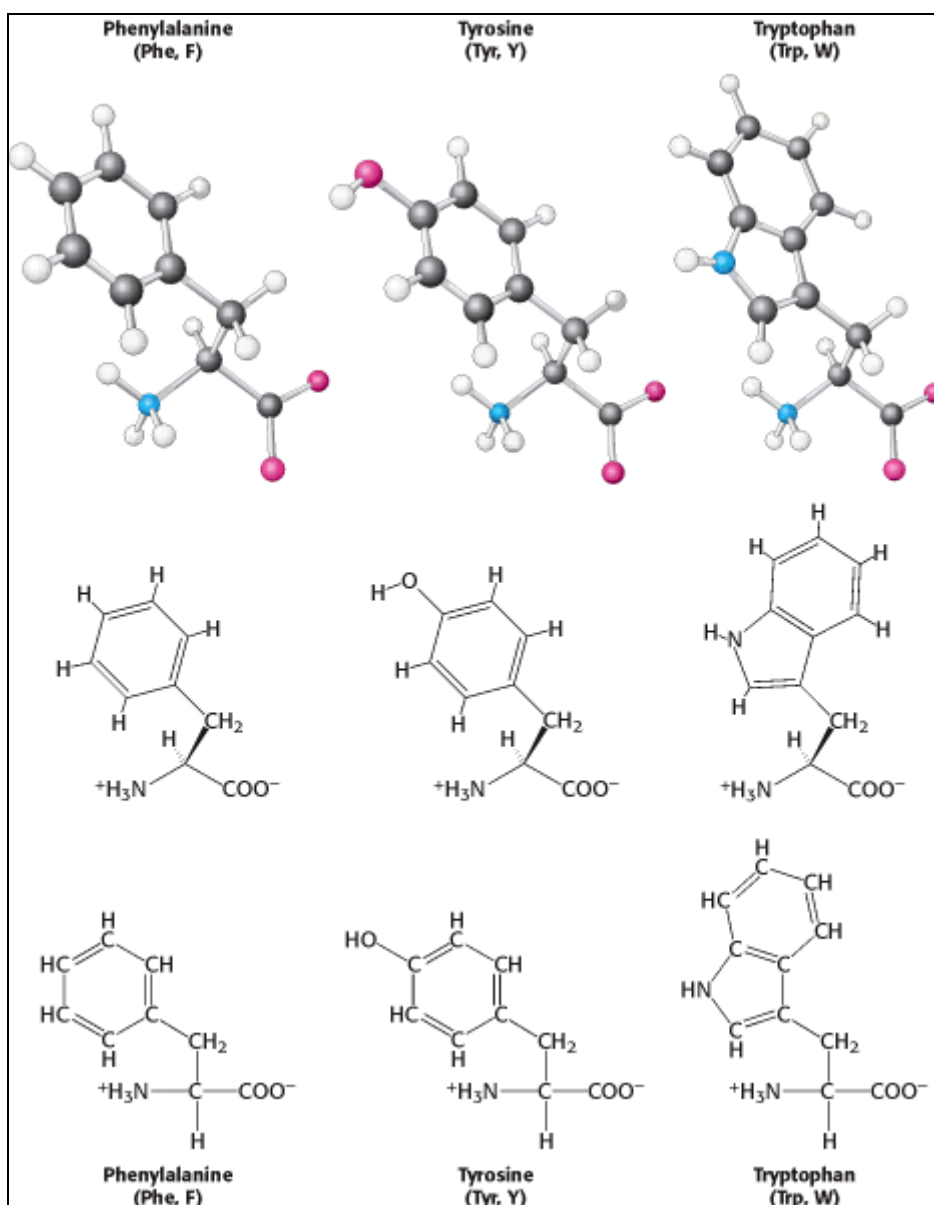


Figure 3.10. Amino Acids with Aromatic Side Chains. Phenylalanine, tyrosine, and tryptophan have hydrophobic character. Tyrosine and tryptophan also have hydrophilic properties because of their $-\text{OH}$ and $-\text{NH}-$ groups, respectively.

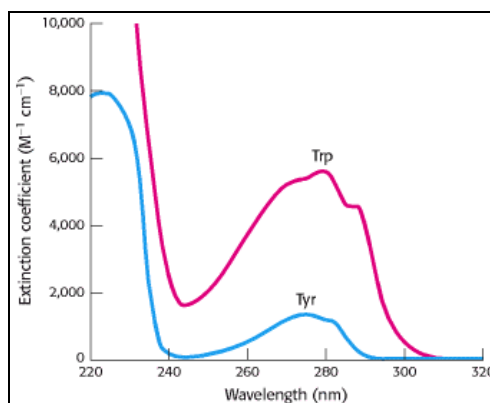


Figure 3.11. Absorption Spectra of the Aromatic Amino Acids Tryptophan (Red) and Tyrosine (Blue). Only these amino acids absorb strongly near 280 nm. [Courtesy of Greg Gatto].

A compound's *extinction coefficient* indicates its ability to absorb light. Beer's law gives the absorbance (A) of light at a given wavelength:

$$A = \epsilon cl \quad \text{Beer's law}$$

where ϵ is the extinction coefficient [in units that are the reciprocals of molarity and distance in centimeters ($M^{-1} \text{ cm}^{-1}$)], c is the concentration of the absorbing species (in units of molarity, M), and l is the length through which the light passes (in units of centimeters). For tryptophan, absorption is maximum at 280 nm and the extinction coefficient is $3400 M^{-1} \text{ cm}^{-1}$ whereas, for tyrosine, absorption is maximum at 276 nm and the extinction coefficient is a less-intense $1400 M^{-1} \text{ cm}^{-1}$. Phenylalanine absorbs light less strongly and at shorter wavelengths. The absorption of light at 280 nm can be used to estimate the concentration of a protein in solution if the number of tryptophan and tyrosine residues in the protein is known.

Two amino acids, *serine* and *threonine*, contain aliphatic *hydroxyl groups* (Figure 3.12). Serine can be thought of as a hydroxylated version of alanine, whereas threonine resembles valine with a hydroxyl group in place of one of the valine methyl groups. The hydroxyl groups on serine and threonine make them much more *hydrophilic* (water loving) and *reactive* than alanine and valine. Threonine, like isoleucine, contains an additional asymmetric center; again only one isomer is present in proteins.

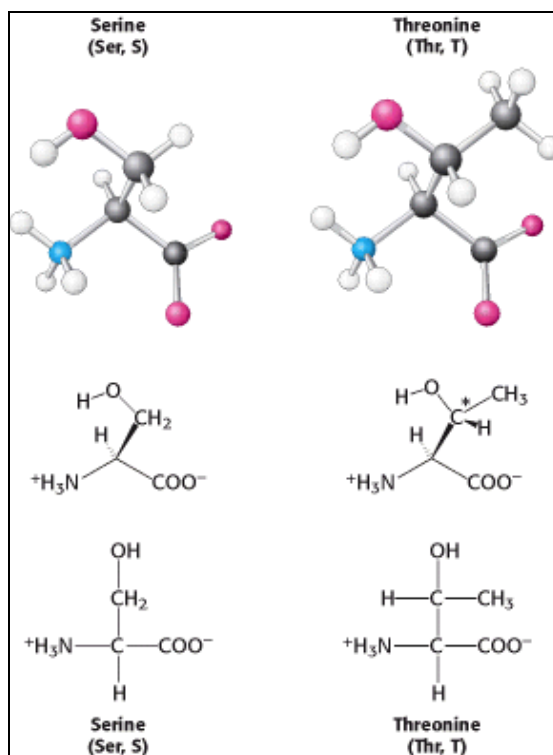


Figure 3.12. Amino Acids Containing Aliphatic Hydroxyl Groups. Serine and threonine contain hydroxyl groups that render them hydrophilic. The additional chiral center in threonine is indicated by an asterisk.

Cysteine is structurally similar to serine but contains a *sulfhydryl*, or *thiol* ($-SH$), group in place of the hydroxyl ($-OH$) group (Figure 3.13). The sulfhydryl group is much more reactive. Pairs of sulfhydryl groups may come together to form disulfide bonds, which are particularly important in stabilizing some proteins, as will be discussed shortly.

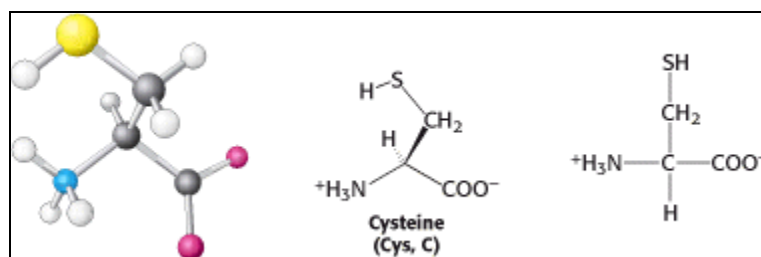


Figure 3.13. Structure of Cysteine.

We turn now to amino acids with very polar side chains that render them highly hydrophilic. *Lysine* and *arginine* have relatively long side chains that terminate with groups that are *positively charged* at neutral pH. Lysine is capped by a primary amino group and arginine by a guanidinium group. *Histidine* contains an imidazole group, an aromatic ring that also can be positively charged (Figure 3.14).

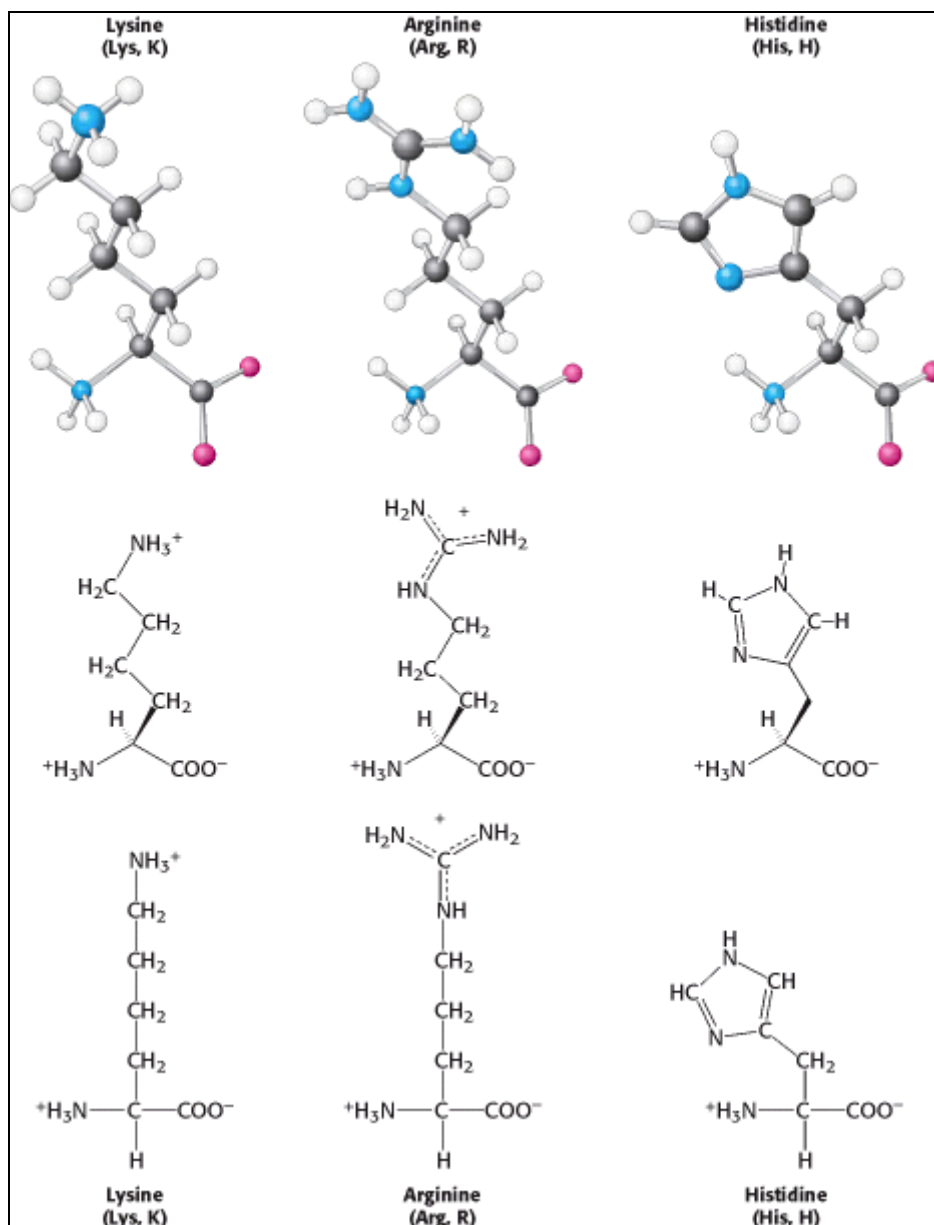


Figure 3.14. The Basic Amino Acids Lysine, Arginine, and Histidine.



With a pK_a value near 6, the imidazole group can be uncharged or positively charged near neutral pH, depending on its local environment (Figure 3.15). Indeed, histidine is often found in the active sites of enzymes, where the imidazole ring can bind and release protons in the course of enzymatic reactions.

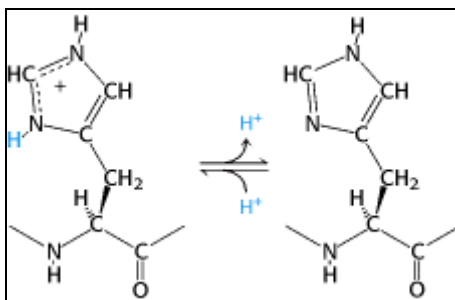


Figure 3.15. Histidine Ionization. Histidine can bind or release protons near physiological pH.

The set of amino acids also contains two with *acidic side chains*: *aspartic acid* and *glutamic acid* (Figure 3.16). These amino acids are often called *aspartate* and *glutamate* to emphasize that their side chains are usually negatively charged at physiological pH. Nonetheless, in some proteins these side chains do accept protons, and this ability is often functionally important. In addition, the set includes uncharged derivatives of aspartate and glutamate - *asparagine* and *glutamine* - each of which contains a terminal *carboxamide* in place of a carboxylic acid (Figure 3.16).

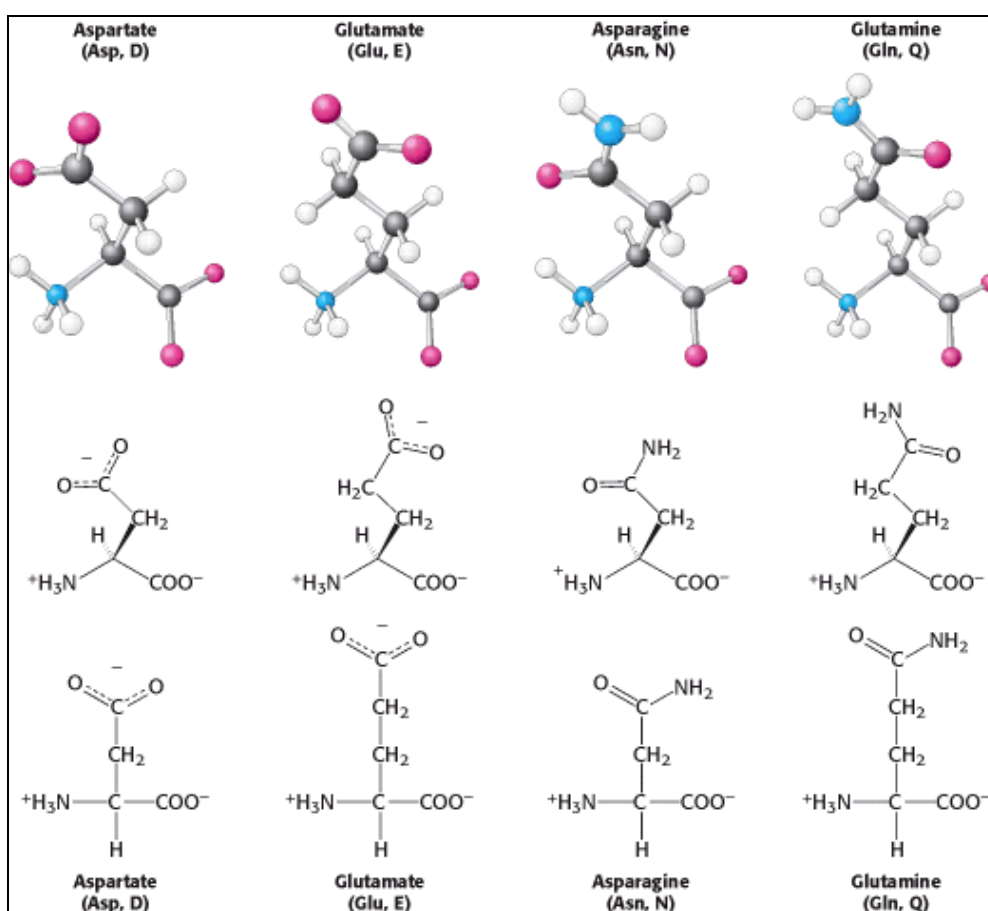


Figure 3.16. Amino Acids with Side-Chain Carboxylates and Carboxamides.

Seven of the 20 amino acids have readily ionizable side chains. These 7 amino acids are able to donate or accept protons to facilitate reactions as well as to form ionic bonds. Table 3.1 gives equilibria and typical pK_a values for ionization of the side chains of tyrosine, cysteine, arginine, lysine, histidine, and aspartic and glutamic acids in proteins. Two other groups in proteins - the terminal α -amino group and the terminal α -carboxyl group - can be ionized, and typical pK_a values are also included in Table 3.1.

Group	Acid	\rightleftharpoons	Base	Typical pK_a^*
Terminal α -carboxyl group		\rightleftharpoons		3.1
Aspartic acid Glutamic acid		\rightleftharpoons		4.1
Histidine		\rightleftharpoons		6.0
Terminal α -amino group		\rightleftharpoons		8.0
Cysteine		\rightleftharpoons		8.3
Tyrosine		\rightleftharpoons		10.9
Lysine		\rightleftharpoons		10.8
Arginine		\rightleftharpoons		12.5

* pK_a values depend on temperature, ionic strength, and the microenvironment of the ionizable group.

Table 3.1. Typical pK_a values of ionizable groups in proteins

Amino acids are often designated by either a three-letter abbreviation or a one-letter symbol (Table 3.2). The abbreviations for amino acids are the first three letters of their names, except for asparagine (Asn), glutamine (Gln), isoleucine (Ile), and tryptophan (Trp). The symbols for many amino acids are the first letters of their names (e.g., G for glycine and L for leucine); the other symbols have been agreed on by convention. These abbreviations and symbols are an integral part of the vocabulary of biochemists.

How did this particular set of amino acids become the building blocks of proteins? First, as a set, they are diverse; their structural and chemical properties span a wide range, endowing proteins with the versatility to assume many functional roles. Second, as noted in Section 2.1.1, many of these amino acids were probably available from prebiotic reactions. Finally, excessive intrinsic reactivity may have eliminated other possible amino acids. For example, amino acids such as homoserine and homocysteine tend to form five-membered cyclic forms that limit their use in proteins; the alternative amino acids that are found in proteins - serine and cysteine - do not readily cyclize, because the rings in their cyclic forms are too small (Figure 3.17).

Amino acid	Three-letter abbreviation	One-letter abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic Acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
Asparagine or aspartic acid	Asx	B
Glutamine or glutamic acid	Glx	Z

Table 3.2. Abbreviations for amino acids

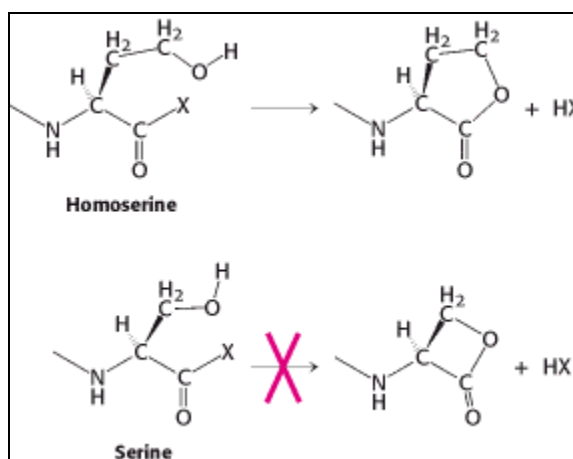


Figure 3.17. Undesirable Reactivity in Amino Acids. Some amino acids are unsuitable for proteins because of undesirable cyclization. Homoserine can cyclize to form a stable, five-membered ring, potentially resulting in peptide-bond cleavage. Cyclization of serine would form a strained, four-membered ring and thus is unfavored. X can be an amino group from a neighboring amino acid or another potential leaving group.

3.2. Primary Structure: Amino Acids Are Linked by Peptide Bonds to Form Polypeptide Chains

Proteins are *linear polymers* formed by linking the α -carboxyl group of one amino acid to the α -amino group of another amino acid with a *peptide bond* (also called an *amide bond*). The formation of a dipeptide from two amino acids is accompanied by the loss of a water molecule (Figure 3.18). The equilibrium of this reaction lies on the side of hydrolysis rather than synthesis. Hence, the biosynthesis of peptide bonds requires an input of free energy. Nonetheless, peptide bonds are quite *stable kinetically*; the lifetime of a peptide bond in aqueous solution in the absence of a catalyst approaches 1000 years.

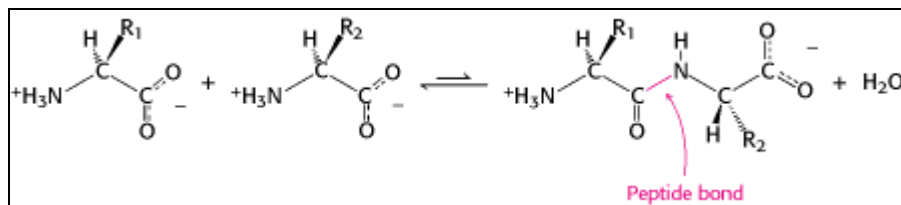


Figure 3.18. Peptide-Bond Formation. The linking of two amino acids is accompanied by the loss of a molecule of water.

A series of amino acids joined by peptide bonds form a *polypeptide chain*, and each amino acid unit in a polypeptide is called a *residue*. A *polypeptide chain has polarity* because its ends are different, with an α -amino group at one end and an α -carboxyl group at the other. By convention, *the amino end is taken to be the beginning of a polypeptide chain*, and so the sequence of amino acids in a polypeptide chain is written starting with the aminoterminal residue. Thus, in the pentapeptide Tyr-Gly-Gly-Phe-Leu (YGGFL), tyrosine is the amino-terminal (N-terminal) residue and leucine is the carboxyl-terminal (C-terminal) residue (Figure 3.19). Leu-Phe-Gly-Gly-Tyr (LFGGY) is a different pentapeptide, with different chemical properties.

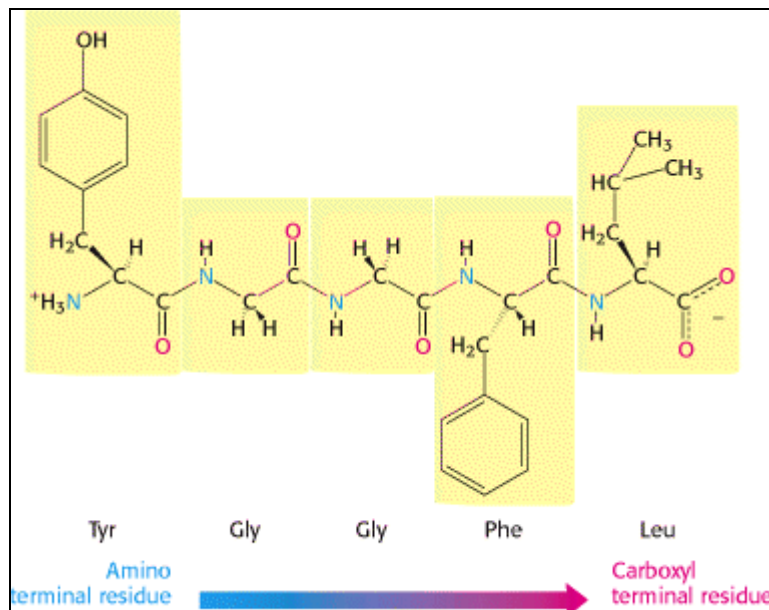


Figure 3.19. Amino Acid Sequences Have Direction. This illustration of the pentapeptide Tyr-Gly-Gly-Phe-Leu (YGGFL) shows the sequence from the amino terminus to the carboxyl terminus. This pentapeptide, Leu-enkephalin, is an opioid peptide that modulates the perception of pain. The reverse pentapeptide, Leu-Phe-Gly-Gly-Tyr (LFGGY), is a different molecule and shows no such effects.

A polypeptide chain consists of a regularly repeating part, called the *main chain* or *backbone*, and a variable part, comprising the distinctive *side chains* (Figure 3.20). The polypeptide backbone is rich in hydrogen-bonding potential. Each residue contains a carbonyl group, which is a good hydrogen-bond acceptor and, with the exception of proline, an NH group, which is a good hydrogen-bond donor. These groups interact with each other and with functional groups from side chains to stabilize particular structures, as will be discussed in detail.

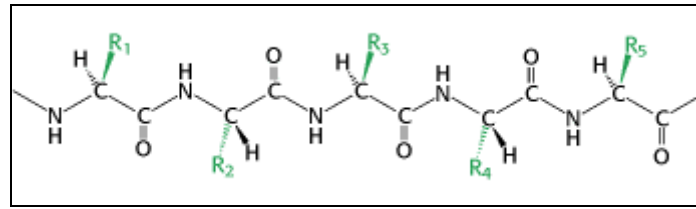


Figure 3.20. Components of a Polypeptide Chain. A polypeptide chain consists of a constant backbone (shown in black) and variable side chains (shown in green).

Most natural polypeptide chains contain between 50 and 2,000 amino acid residues and are commonly referred to as *proteins*. Peptides made of small numbers of amino acids are called *oligopeptides* or simply *peptides*. The mean molecular weight of an amino acid residue is about 110, and so the molecular weights of most proteins are between 5500 and 220,000. We can also refer to the mass of a protein, which is expressed in units of daltons; one *dalton* is equal to one atomic mass unit. A protein with a molecular weight of 50,000 has a mass of 50,000 daltons, or 50 kd (kilodaltons).

Dalton

A unit of mass very nearly equal to that of a hydrogen atom. Named after John Dalton (1766-1844), who developed the atomic theory of matter.

Kilodalton (kd)

A unit of mass equal to 1000 daltons.

In some proteins, the linear polypeptide chain is cross-linked. The most common cross-links are *disulfide bonds*, formed by the oxidation of a pair of cysteine residues (Figure 3.21). The resulting unit of linked cysteines is called *cystine*. Extracellular proteins often have several disulfide bonds, whereas intracellular proteins usually lack them. Rarely, nondisulfide cross-links derived from other side chains are present in some proteins. For example, collagen fibers in connective tissue are strengthened in this way, as are fibrin blood clots.

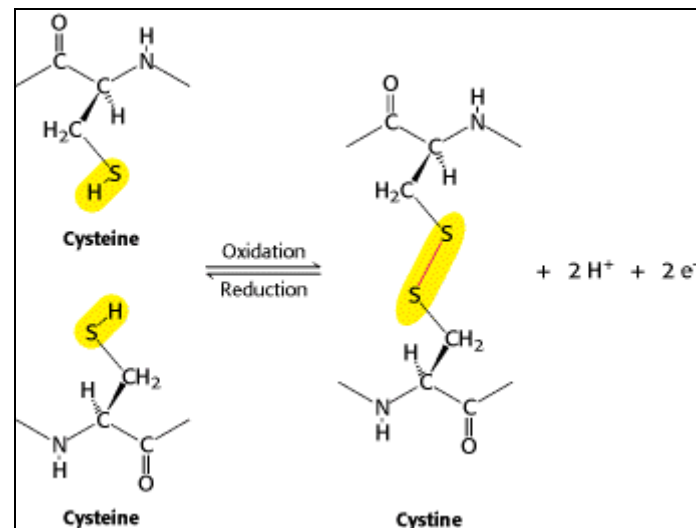


Figure 3.21. Cross-Links. The formation of a disulfide bond from two cysteine residues is an oxidation reaction.

3.2.1. Proteins Have Unique Amino Acid Sequences That Are Specified by Genes

In 1953, Frederick Sanger determined the amino acid sequence of insulin, a protein hormone (Figure 3.22). This work is a landmark in biochemistry because it showed for the first time that a protein has a precisely defined amino acid sequence. Moreover, it demonstrated that insulin consists only of L amino acids linked by peptide bonds between α -amino and α -carboxyl groups. This accomplishment stimulated other scientists to carry out sequence studies of a wide variety of proteins. Indeed, the complete amino

acid sequences of more than 100,000 proteins are now known. *The striking fact is that each protein has a unique, precisely defined amino acid sequence.* The amino acid sequence of a protein is often referred to as its *primary structure*.

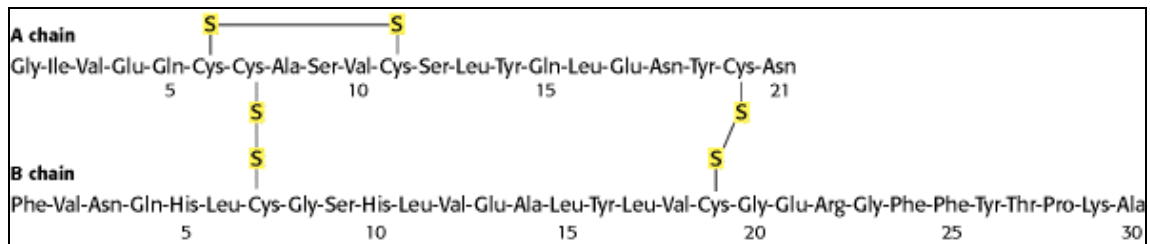


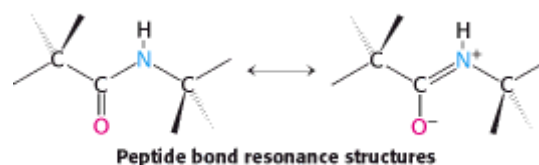
Figure 3.22. Amino Acid Sequence of Bovine Insulin.

A series of incisive studies in the late 1950s and early 1960s revealed that the amino acid sequences of proteins are genetically determined. The sequence of nucleotides in DNA, the molecule of heredity, specifies a complementary sequence of nucleotides in RNA, which in turn specifies the amino acid sequence of a protein. In particular, each of the 20 amino acids of the repertoire is encoded by one or more specific sequences of three nucleotides ([Section 5.5](#)).

Knowing amino acid sequences is important for several reasons. First, knowledge of the sequence of a protein is usually essential to elucidating its mechanism of action (e.g., the catalytic mechanism of an enzyme). Moreover, proteins with novel properties can be generated by varying the sequence of known proteins. Second, amino acid sequences determine the three-dimensional structures of proteins. Amino acid sequence is the link between the genetic message in DNA and the three-dimensional structure that performs a protein's biological function. Analyses of relations between amino acid sequences and three-dimensional structures of proteins are uncovering the rules that govern the folding of polypeptide chains. Third, sequence determination is a component of molecular pathology, a rapidly growing area of medicine. Alterations in amino acid sequence can produce abnormal function and disease. Severe and sometimes fatal diseases, such as sickle-cell anemia and cystic fibrosis, can result from a change in a single amino acid within a protein. Fourth, the sequence of a protein reveals much about its evolutionary history (see [Chapter 7](#)). Proteins resemble one another in amino acid sequence only if they have a common ancestor. Consequently, molecular events in evolution can be traced from amino acid sequences; molecular paleontology is a flourishing area of research.

3.2.2. Polypeptide Chains Are Flexible Yet Conformationally Restricted

Examination of the geometry of the protein backbone reveals several important features. First, *the peptide bond is essentially planar* ([Figure 3.23](#)). Thus, for a pair of amino acids linked by a peptide bond, six atoms lie in the same plane: the α -carbon atom and CO group from the first amino acid and the NH group and α -carbon atom from the second amino acid. The nature of the chemical bonding within a peptide explains this geometric preference. The peptide bond has considerable *double-bond character*, which prevents rotation about this bond.



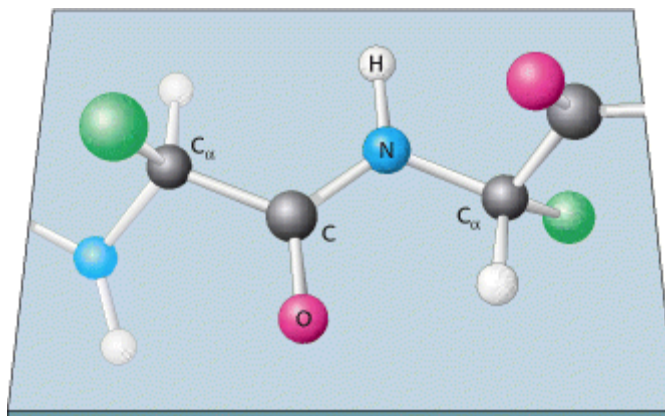


Figure 3.23. Peptide Bonds Are Planar. In a pair of linked amino acids, six atoms (C_{α} , C, O, N, H, and C_{α}) lie in a plane. Side chains are shown as green balls.

The inability of the bond to rotate constrains the conformation of the peptide backbone and accounts for the bond's planarity. This double-bond character is also expressed in the length of the bond between the CO and NH groups. The C-N distance in a peptide bond is typically 1.32 Å, which is between the values expected for a C-N single bond (1.49 Å) and a C=N double bond (1.27 Å), as shown in [Figure 3.24](#). Finally, the peptide bond is uncharged, allowing polymers of amino acids linked by peptide bonds to form tightly packed globular structures.

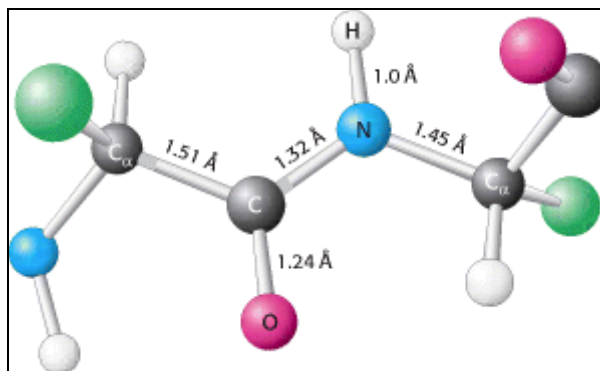


Figure 3.24. Typical Bond Lengths Within a Peptide Unit. The peptide unit is shown in the trans configuration.

Two configurations are possible for a planar peptide bond. In the trans configuration, the two α -carbon atoms are on opposite sides of the peptide bond. In the cis configuration, these groups are on the same side of the peptide bond. *Almost all peptide bonds in proteins are trans.* This preference for trans over cis can be explained by the fact that steric clashes between groups attached to the α -carbon atoms hinder formation of the cis form but do not occur in the trans configuration ([Figure 3.25](#)). By far the most common cis peptide bonds are X-Pro linkages. Such bonds show less preference for the trans configuration because the nitrogen of proline is bonded to two tetrahedral carbon atoms, limiting the steric differences between the trans and cis forms ([Figure 3.26](#)).

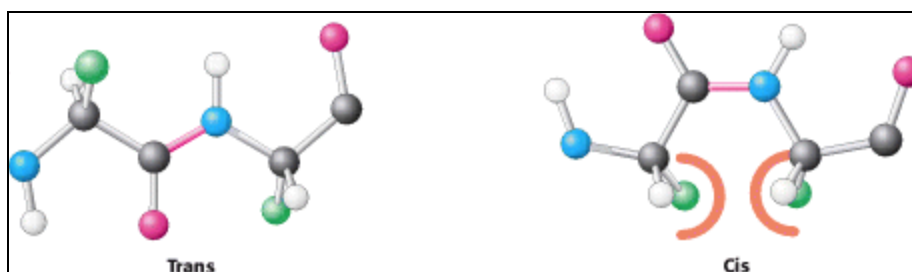


Figure 3.25. Trans and Cis Peptide Bonds. The trans form is strongly favored because of steric clashes that occur in the cis form.

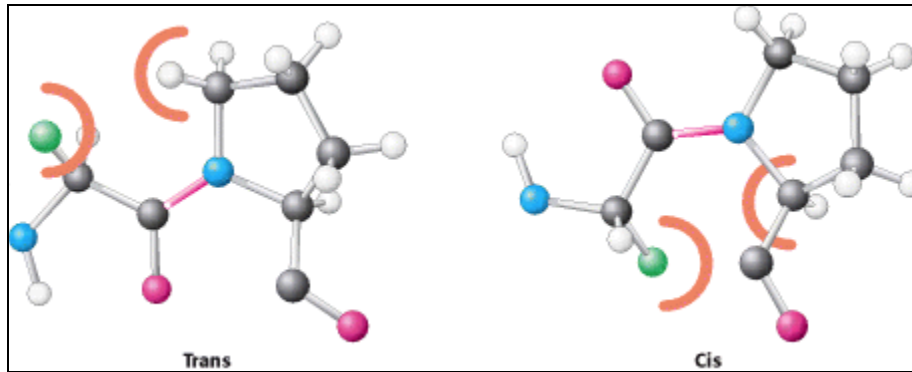


Figure 3.26. Trans and Cis X-Pro Bonds. The energies of these forms are relatively balanced because steric clashes occur in both forms.

In contrast with the peptide bond, the bonds between the amino group and the α -carbon atom and between the α -carbon atom and the carbonyl group are pure single bonds. The two adjacent rigid peptide units may rotate about these bonds, taking on various orientations. *This freedom of rotation about two bonds of each amino acid allows proteins to fold in many different ways.* The rotations about these bonds can be specified by dihedral angles (Figure 3.27). The angle of rotation about the bond between the nitrogen and the α -carbon atoms is called *phi* (ϕ). The angle of rotation about the bond between the α -carbon and the carbonyl carbon atoms is called *psi* (ψ). A clockwise rotation about either bond as viewed from the front of the back group corresponds to a positive value. The ϕ and ψ angles determine the path of the polypeptide chain.

Dihedral angle

A measure of the rotation about a bond, usually taken to lie between -180° and $+180^\circ$. Dihedral angles are sometimes called torsion angles.

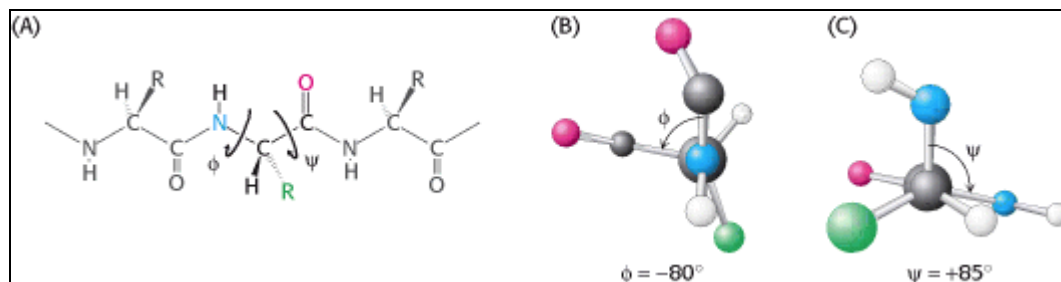


Figure 3.27. Rotation About Bonds in a Polypeptide. The structure of each amino acid in a polypeptide can be adjusted by rotation about two single bonds. (A) Phi (ϕ) is the angle of rotation about the bond between the nitrogen and the α -carbon atoms, whereas psi (ψ) is the angle of rotation about the bond between the α -carbon and the carbonyl carbon atoms. (B) A view down the bond between the nitrogen and the α -carbon atoms, showing how ϕ is measured. (C) A view down the bond between the α -carbon and the carbonyl carbon atoms, showing how ψ is measured.

Are all combinations of ϕ and ψ possible? G. N. Ramachandran recognized that many combinations are forbidden because of steric collisions between atoms. The allowed values can be visualized on a two-dimensional plot called a *Ramachandran diagram* (Figure 3.28). Three-quarters of the possible (ϕ , ψ) combinations are excluded simply by local steric clashes. *Steric exclusion, the fact that two atoms cannot be in the same place at the same time, can be a powerful organizing principle.*

The ability of biological polymers such as proteins to fold into welldefined structures is remarkable thermodynamically. Consider the equilibrium between an unfolded polymer that exists as a random coil - that is, as a mixture of many possible conformations - and the folded form that adopts a unique conformation. The favorable entropy associated with the large number of conformations in the unfolded form opposes folding and must be overcome by interactions favoring the folded form. Thus, highly flexible polymers with a large number of possible conformations do not fold into unique structures. *The rigidity of the peptide unit and the restricted set of allowed ϕ and ψ angles limits the number of structures accessible to the unfolded form sufficiently to allow protein folding to occur.*

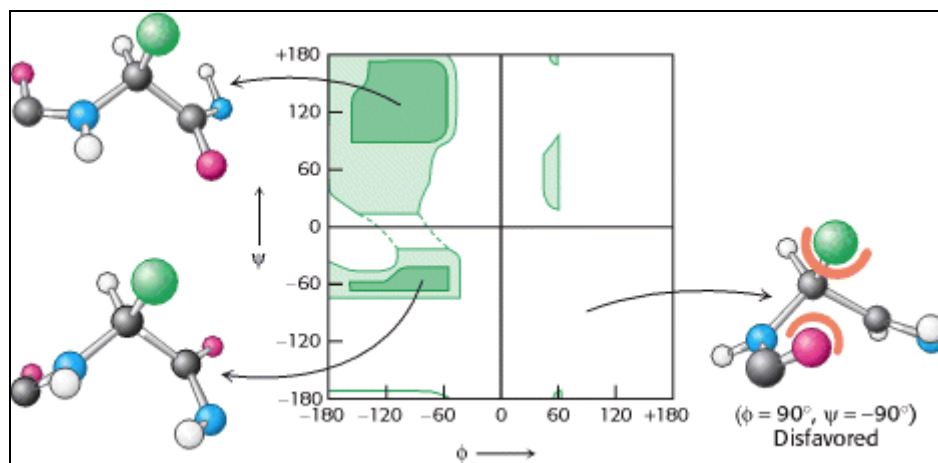


Figure 3.28. A Ramachandran Diagram Showing the Values of ϕ and ψ . Not all ϕ and ψ values are possible without collisions between atoms. The most favorable regions are shown in dark green; borderline regions are shown in light green. The structure on the right is disfavored because of steric clashes.

3.3. Secondary Structure: Polypeptide Chains Can Fold Into Regular Structures Such as the Alpha Helix, the Beta Sheet, and Turns and Loops

Can a polypeptide chain fold into a regularly repeating structure? In 1951, Linus Pauling and Robert Corey proposed two periodic structures called the α helix (alpha helix) and the β pleated sheet (beta pleated sheet). Subsequently, other structures such as the β turn and omega (Ω) loop were identified. Although not periodic, these common turn or loop structures are well defined and contribute with α helices and β sheets to form the final protein structure.

3.3.1. The Alpha Helix Is a Coiled Structure Stabilized by Intrachain Hydrogen Bonds

In evaluating potential structures, Pauling and Corey considered which conformations of peptides were sterically allowed and which most fully exploited the hydrogen-bonding capacity of the backbone NH and CO groups. The first of their proposed structures, the α helix, is a rodlike structure (Figure 3.29). A tightly coiled backbone forms the inner part of the rod and the side chains extend outward in a helical array. The α helix is stabilized by hydrogen bonds between the NH and CO groups of the main chain. In particular, the CO group of each amino acid forms a hydrogen bond with the NH group of the amino acid that is situated four residues ahead in the sequence (Figure 3.30). Thus, except for amino acids near the ends of an α helix, all the main-chain CO and NH groups are hydrogen bonded. Each residue is related to the next one by a rise of 1.5 Å along the helix axis and a rotation of 100 degrees, which gives 3.6 amino acid residues per turn of helix. Thus, amino acids spaced three and four apart in the sequence are spatially quite close to one another in an α helix. In contrast, amino acids two apart in the sequence are situated on opposite sides of the helix and so are unlikely to make contact. The pitch of the α helix, which is equal to the product of the translation (1.5 Å) and the number of residues per turn (3.6), is 5.4 Å. The screw sense of a helix can be right-handed (clockwise) or left-handed (counterclockwise). The Ramachandran diagram reveals that both the right-handed and the left-handed helices are among allowed conformations (Figure 3.31). However, right-handed helices are energetically more favorable because there is less steric clash between the side chains and the backbone. Essentially all α helices found in proteins are right-handed. In schematic diagrams of proteins, α helices are depicted as twisted ribbons or rods (Figure 3.32).

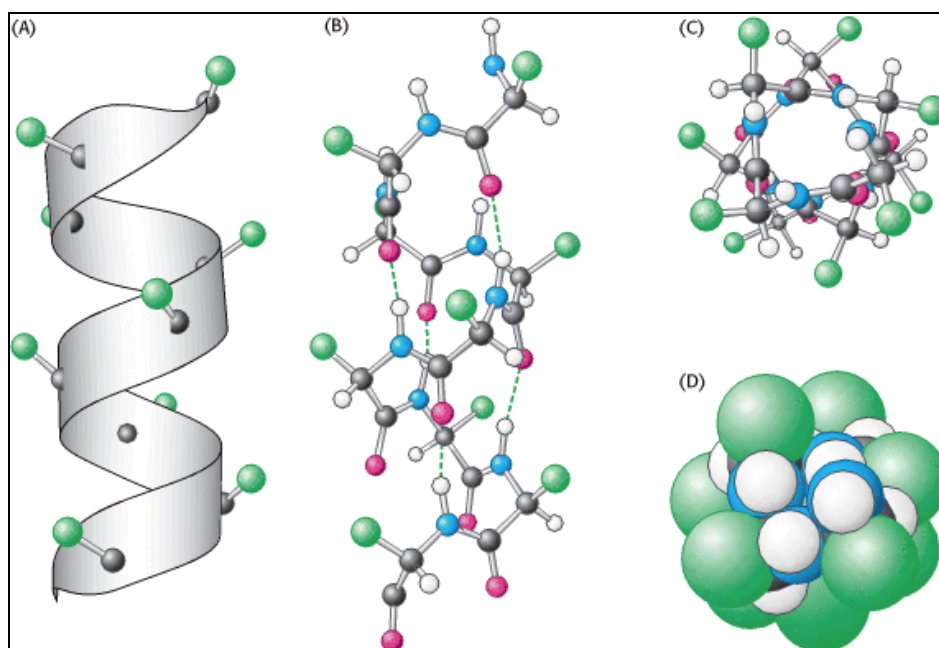


Figure 3.29. Structure of the α Helix. (A) A ribbon depiction with the α -carbon atoms and side chains (green) shown. (B) A side view of a ball-and-stick version depicts the hydrogen bonds (dashed lines) between NH and CO groups. (C) An end view shows the coiled backbone as the inside of the helix and the side chains (green) projecting outward. (D) A space-filling view of part C shows the tightly packed interior core of the helix.

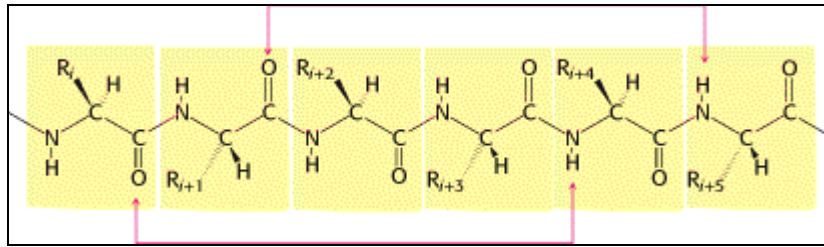


Figure 3.30. Hydrogen-Bonding Scheme For an α helix. In the α helix, the CO group of residue n forms a hydrogen bond with the NH group of residue $n+4$.

Screw sense

Describes the direction in which a helical structure rotates with respect to its axis. If, viewed down the axis of a helix, the chain turns in a clockwise direction, it has a right-handed screw sense. If the turning is counterclockwise, the screw sense is left-handed.

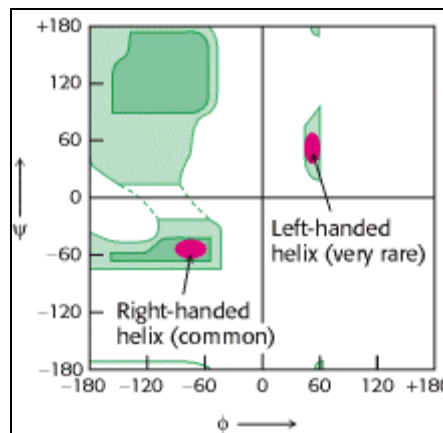


Figure 3.31. Ramachandran Diagram for Helices. Both right- and left-handed helices lie in regions of allowed conformations in the Ramachandran diagram. However, essentially all α helices in proteins are right-handed.

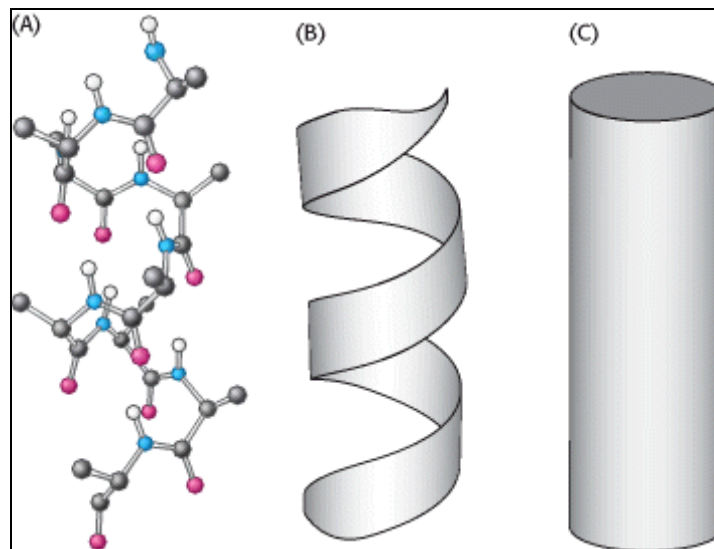


Figure 3.32. Schematic Views OF α Helices. (A) A ball-and-stick model. (B) A ribbon depiction. (C) A cylindrical depiction.

Pauling and Corey predicted the structure of the α helix 6 years before it was actually seen in the x-ray reconstruction of the structure of myoglobin. *The elucidation of the structure of the α helix is a landmark in biochemistry because it demonstrated that the conformation of a polypeptide chain can be predicted if the properties of its components are rigorously and precisely known.*

The α -helical content of proteins ranges widely, from nearly none to almost 100%. For example, about 75% of the residues in ferritin, a protein that helps store iron, are in α helices (Figure 3.33). Single α

helices are usually less than 45 Å long. However, two or more α helices can entwine to form a very stable structure, which can have a length of 1000 Å (100 nm, or 0.1 μm) or more (Figure 3.34). Such α -helical coiled coils are found in myosin and tropomyosin in muscle, in fibrin in blood clots, and in keratin in hair. The helical cables in these proteins serve a mechanical role in forming stiff bundles of fibers, as in porcupine quills. The cytoskeleton (internal scaffolding) of cells is rich in so-called intermediate filaments, which also are two-stranded α -helical coiled coils. Many proteins that span biological membranes also contain α helices.

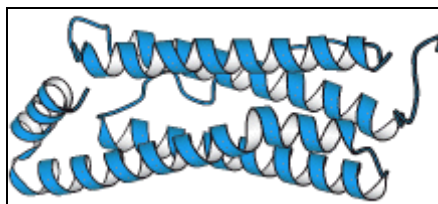


Figure 3.33. A Largely α Helical Protein. Ferritin, an iron-storage protein, is built from a bundle of α helices.

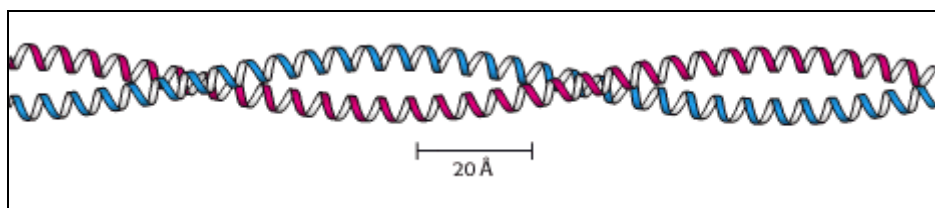


Figure 3.34. An α -Helical Coiled Coil. The two helices wind around one another to form a superhelix. Such structures are found in many proteins including keratin in hair, quills, claws, and horns.

3.3.2. Beta Sheets Are Stabilized by Hydrogen Bonding Between Polypeptide Strands

Pauling and Corey discovered another periodic structural motif, which they named the β pleated sheet (β because it was the second structure that they elucidated, the α helix having been the first). The β pleated sheet (or, more simply, the β sheet) differs markedly from the rodlike α helix. A polypeptide chain, called a β strand, in a β sheet is almost fully extended rather than being tightly coiled as in the α helix. A range of extended structures are sterically allowed (Figure 3.35).

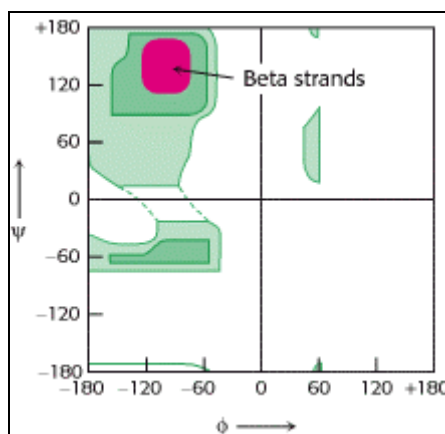


Figure 3.35. Ramachandran Diagram For β Strands. The red area shows the sterically allowed conformations of extended, β -strand-like structures.

The distance between adjacent amino acids along a β strand is approximately 3.5 Å, in contrast with a distance of 1.5 Å along an α helix. The side chains of adjacent amino acids point in opposite directions (Figure 3.36). A β sheet is formed by linking two or more β strands by hydrogen bonds. Adjacent chains in a β sheet can run in opposite directions (antiparallel β sheet) or in the same direction (parallel β sheet). In the antiparallel arrangement, the NH group and the CO group of each amino acid are respectively hydrogen bonded to the CO group and the NH group of a partner on the adjacent chain (Figure 3.37). In the parallel arrangement, the hydrogen-bonding scheme is slightly more complicated. For each amino

acid, the NH group is hydrogen bonded to the CO group of one amino acid on the adjacent strand, whereas the CO group is hydrogen bonded to the NH group on the amino acid two residues farther along the chain (Figure 3.38). Many strands, typically 4 or 5 but as many as 10 or more, can come together in β sheets. Such β sheets can be purely antiparallel, purely parallel, or mixed (Figure 3.39).

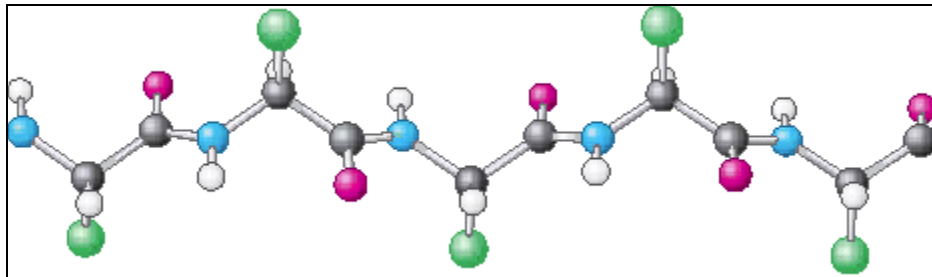


Figure 3.36. Structure of a β Strand. The side chains (green) are alternately above and below the plane of the strand.

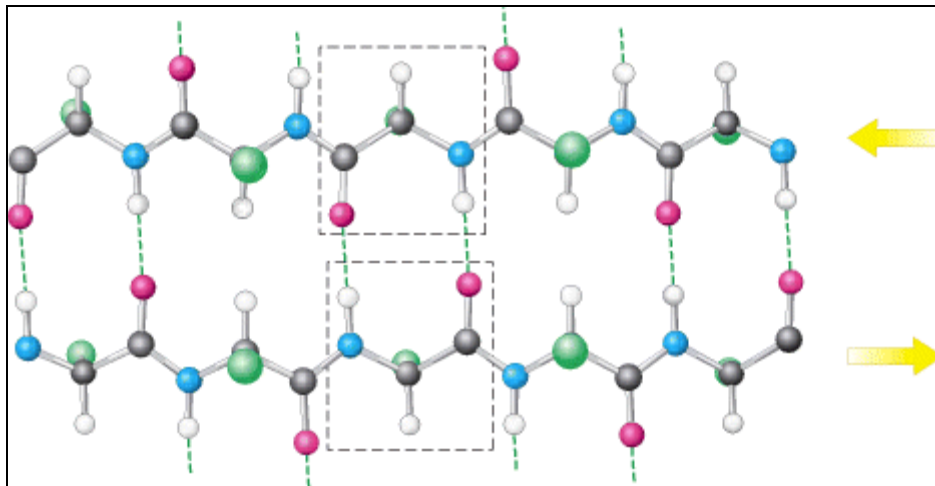


Figure 3.37. An Antiparallel β Sheet. Adjacent β strands run in opposite directions. Hydrogen bonds between NH and CO groups connect each amino acid to a single amino acid on an adjacent strand, stabilizing the structure.

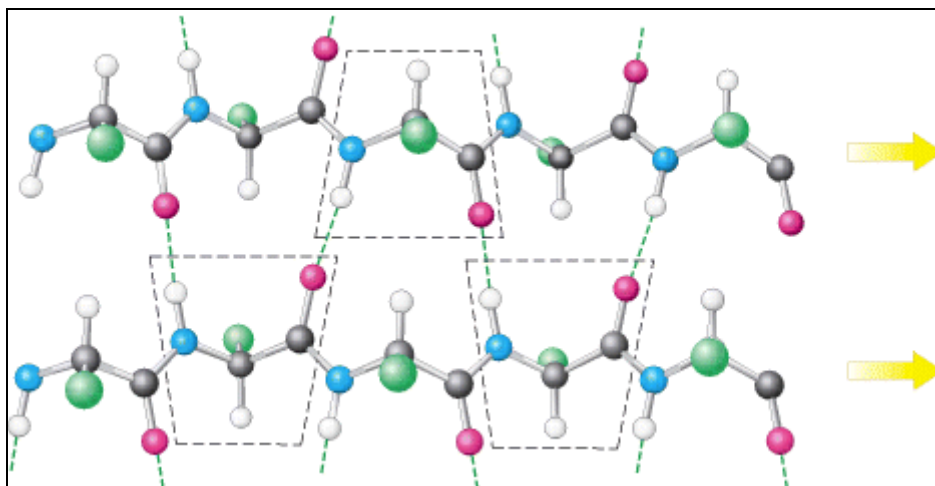


Figure 3.38. A Parallel β Sheet. Adjacent β strands run in the same direction. Hydrogen bonds connect each amino acid on one strand with two different amino acids on the adjacent strand.

In schematic diagrams, β strands are usually depicted by broad arrows pointing in the direction of the carboxyl-terminal end to indicate the type of β sheet formed - parallel or antiparallel. More structurally diverse than α helices, β sheets can be relatively flat but most adopt a somewhat twisted shape (Figure 3.40). The β sheet is an important structural element in many proteins. For example, fatty acid-binding proteins, important for lipid metabolism, are built almost entirely from β sheets (Figure 3.41).

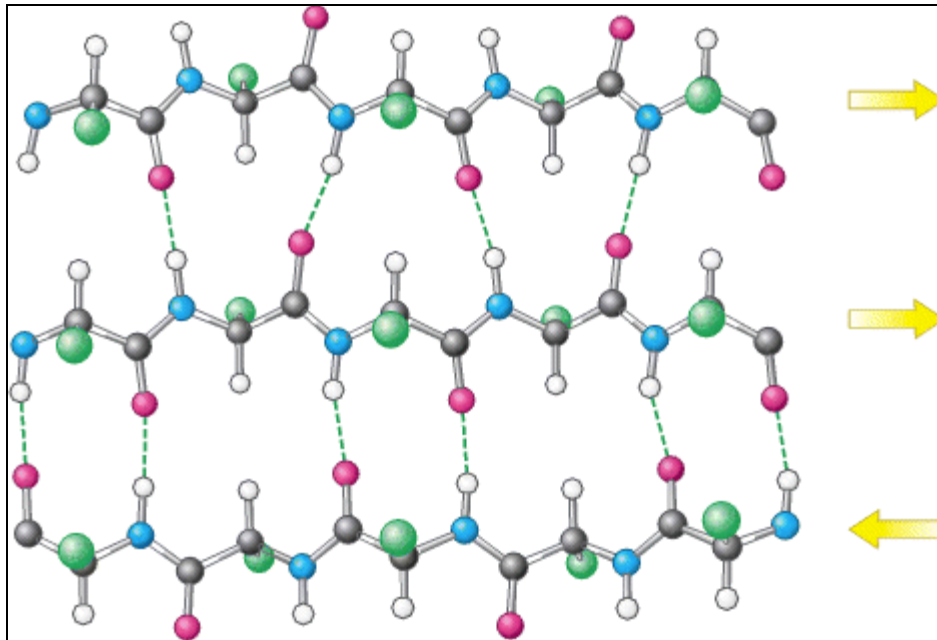


Figure 3.39. Structure of a Mixed β Sheet.

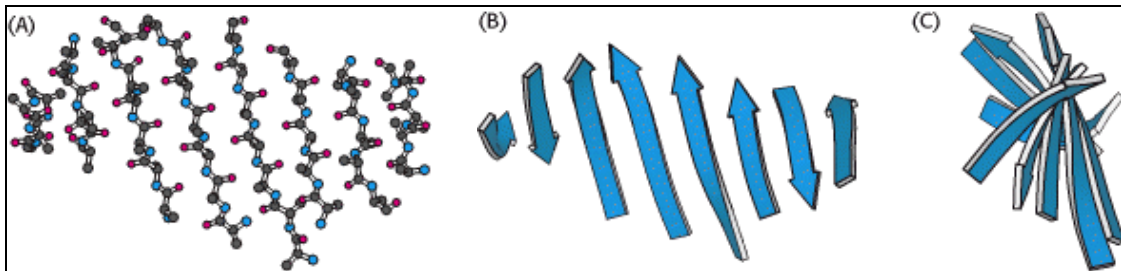


Figure 3.40. A Twisted β Sheet. (A) A ball-and-stick model. (B) A schematic model. (C) The schematic view rotated by 90 degrees to illustrate the twist more clearly.

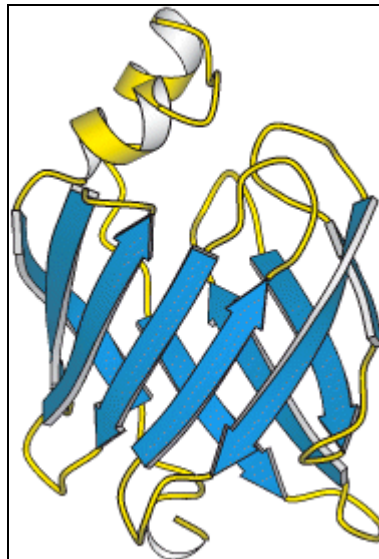


Figure 3.41. A Protein Rich in β Sheets. The structure of a fatty acid-binding protein.

3.3.3. Polypeptide Chains Can Change Direction by Making Reverse Turns and Loops

Most proteins have compact, globular shapes, requiring reversals in the direction of their polypeptide chains. Many of these reversals are accomplished by a common structural element called the *reverse turn* (also known as the β turn or *hairpin bend*), illustrated in [Figure 3.42](#). In many reverse turns, the CO group

of residue i of a polypeptide is hydrogen bonded to the NH group of residue $i + 3$. This interaction stabilizes abrupt changes in direction of the polypeptide chain. In other cases, more elaborate structures are responsible for chain reversals. These structures are called *loops* or sometimes Ω *loops* (omega loops) to suggest their overall shape. Unlike α helices and β strands, loops do not have regular, periodic structures. Nonetheless, loop structures are often rigid and well defined (Figure 3.43). Turns and loops invariably lie on the surfaces of proteins and thus often participate in interactions between proteins and other molecules. The distribution of α helices, β strands, and turns along a protein chain is often referred to as its *secondary structure*.

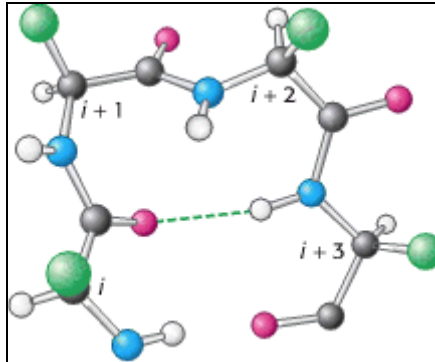


Figure 3.42. Structure of a Reverse Turn. The CO group of residue i of the polypeptide chain is hydrogen bonded to the NH group of residue $i + 3$ to stabilize the turn

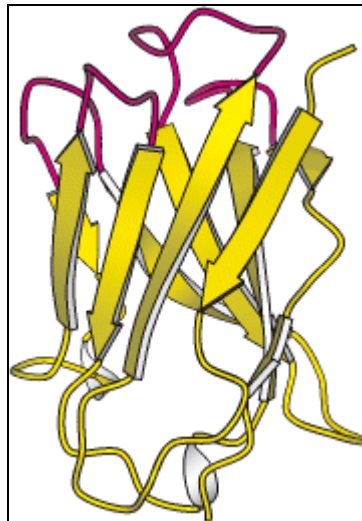


Figure 3.43. Loops on a Protein Surface. A part of an antibody molecule has surface loops (shown in red) that mediate interactions with other molecules..

3.4. Tertiary Structure: Water-Soluble Proteins Fold Into Compact Structures with Nonpolar Cores

Let us now examine how amino acids are grouped together in a complete protein. X-ray crystallographic and nuclear magnetic resonance studies (Section 4.5) have revealed the detailed three-dimensional structures of thousands of proteins. We begin here with a preview of *myoglobin*, the first protein to be seen in atomic detail.

Myoglobin, the oxygen carrier in muscle, is a single polypeptide chain of 153 amino acids (see also Chapters 7 and 10). The capacity of myoglobin to bind oxygen depends on the presence of *heme*, a nonpolypeptide *prosthetic (helper) group* consisting of protoporphyrin IX and a central iron atom. *Myoglobin is an extremely compact molecule*. Its overall dimensions are $45 \times 35 \times 25 \text{ \AA}$, an order of magnitude less than if it were fully stretched out (Figure 3.44). About 70% of the main chain is folded into eight α helices, and much of the rest of the chain forms turns and loops between helices.

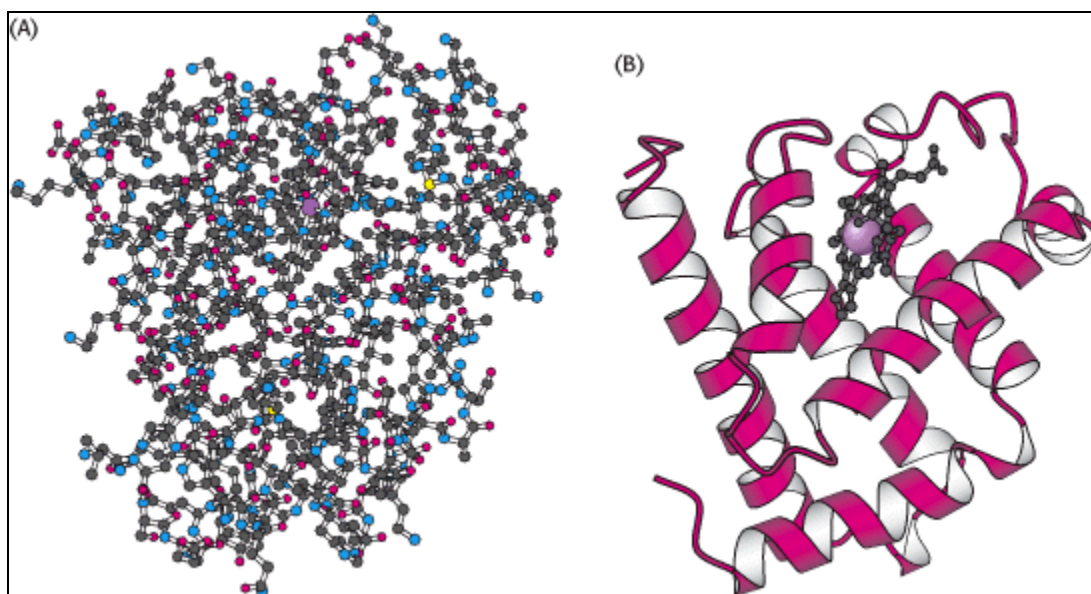


Figure 3.44. Three-Dimensional Structure of Myoglobin. (A) This ball-and-stick model shows all nonhydrogen atoms and reveals many interactions between the amino acids. (B) A schematic view shows that the protein consists largely of α helices. The heme group is shown in black and the iron atom is shown as a purple sphere.

The folding of the main chain of myoglobin, like that of most other proteins, is complex and devoid of symmetry. The overall course of the polypeptide chain of a protein is referred to as its *tertiary structure*. A unifying principle emerges from the distribution of side chains. The striking fact is that *the interior consists almost entirely of nonpolar residues* such as leucine, valine, methionine, and phenylalanine (Figure 3.45). Charged residues such as aspartate, glutamate, lysine, and arginine are absent from the inside of myoglobin. The only polar residues inside are two histidine residues, which play critical roles in binding iron and oxygen. The outside of myoglobin, on the other hand, consists of both polar and nonpolar residues. The spacefilling model shows that there is very little empty space inside.

This contrasting distribution of polar and nonpolar residues reveals a key facet of protein architecture. In an aqueous environment, protein folding is driven by the strong tendency of hydrophobic residues to be excluded from water (see Section 1.3.4). Recall that a system is more thermodynamically stable when hydrophobic groups are clustered rather than extended into the aqueous surroundings. *The polypeptide chain therefore folds so that its hydrophobic side chains are buried and its polar, charged chains are on the surface*. Many α helices and β strands are amphipathic; that is, the α helix or β strand has a hydrophobic face, which points into the protein interior, and a more polar face, which points into solution. The fate of the main chain accompanying the hydrophobic side chains is important, too. An unpaired peptide NH or CO group markedly prefers water to a nonpolar milieu. The secret of burying a segment of main chain in a hydrophobic environment is pairing all the NH and CO groups by hydrogen bonding. This pairing is neatly accomplished in an α helix or β sheet. Van der Waals interactions between tightly packed hydrocarbon side chains also contribute to the stability of proteins. We can now understand why the set of 20 amino acids contains several that differ subtly in size and shape. They provide a palette from

which to choose to fill the interior of a protein neatly and thereby maximize van der Waals interactions, which require intimate contact.

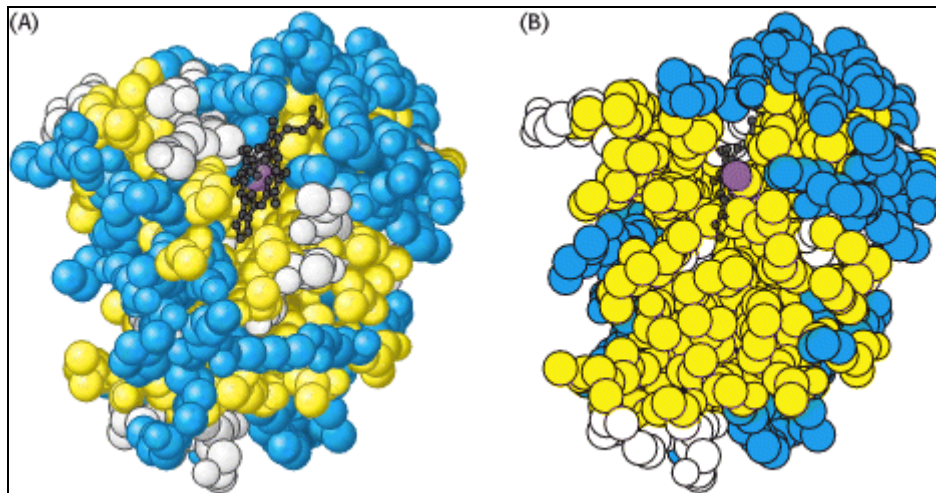


Figure 3.45. Distribution of Amino Acids in Myoglobin. (A) A space-filling model of myoglobin with hydrophobic amino acids shown in yellow, charged amino acids shown in blue, and others shown in white. The surface of the molecule has many charged amino acids, as well as some hydrophobic amino acids. (B) A cross-sectional view shows that mostly hydrophobic amino acids are found on the inside of the structure, whereas the charged amino acids are found on the protein surface.

Some proteins that span biological membranes are "the exceptions that prove the rule" regarding the distribution of hydrophobic and hydrophilic amino acids throughout three-dimensional structures. For example, consider porins, proteins found in the outer membranes of many bacteria (Figure 3.46). The permeability barriers of membranes are built largely of alkane chains that are quite hydrophobic (Section 12.4). Thus, porins are covered on the outside largely with hydrophobic residues that interact with the neighboring alkane chains. In contrast, the center of the protein contains many charged and polar amino acids that surround a water-filled channel running through the middle of the protein. Thus, because porins function in hydrophobic environments, they are "inside out" relative to proteins that function in aqueous solution.

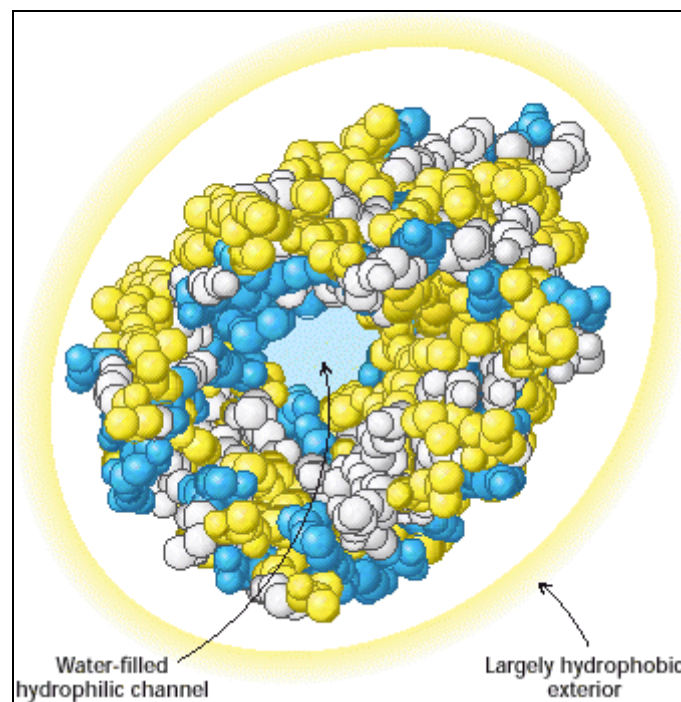


Figure 3.46. "Inside Out" Amino Acid Distribution in Porin. The outside of porin (which contacts hydrophobic groups in membranes) is covered largely with hydrophobic residues, whereas the center includes a water-filled channel lined with charged and polar amino acids.

Some polypeptide chains fold into two or more compact regions that may be connected by a flexible segment of polypeptide chain, rather like pearls on a string. These compact globular units, called

domains, range in size from about 30 to 400 amino acid residues. For example, the extracellular part of CD4, the cell-surface protein on certain cells of the immune system to which the human immunodeficiency virus (HIV) attaches itself, comprises four similar domains of approximately 100 amino acids each (Figure 3.47). Often, proteins are found to have domains in common even if their overall tertiary structures are different.

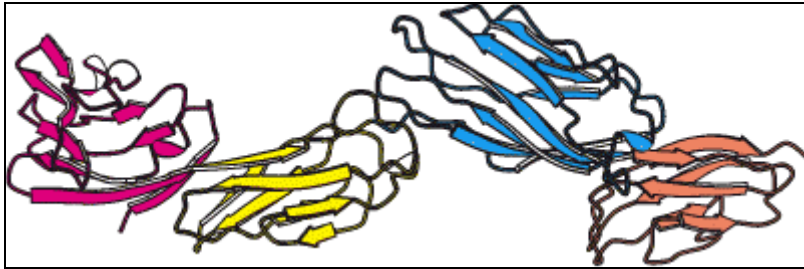


Figure 3.47. Protein Domains. The cell-surface protein CD4 consists of four similar domains

3.5. Quaternary Structure: Polypeptide Chains Can Assemble Into Multisubunit Structures

Four levels of structure are frequently cited in discussions of protein architecture. So far, we have considered three of them. *Primary structure* is the amino acid sequence. *Secondary structure* refers to the spatial arrangement of amino acid residues that are nearby in the sequence. Some of these arrangements are of a regular kind, giving rise to a periodic structure. The α helix and β strand are elements of secondary structure. *Tertiary structure* refers to the spatial arrangement of amino acid residues that are far apart in the sequence and to the pattern of disulfide bonds. We now turn to proteins containing more than one polypeptide chain. Such proteins exhibit a fourth level of structural organization. Each polypeptide chain in such a protein is called a *subunit*. *Quaternary structure* refers to the spatial arrangement of subunits and the nature of their interactions. The simplest sort of quaternary structure is a *dimer*, consisting of two identical subunits. This organization is present in the DNA-binding protein Cro found in a bacterial virus called λ (Figure 3.48). More complicated quaternary structures also are common. More than one type of subunit can be present, often in variable numbers. For example, human hemoglobin, the oxygen-carrying protein in blood, consists of two subunits of one type (designated α) and two subunits of another type (designated β), as illustrated in Figure 3.49. Thus, the hemoglobin molecule exists as an $\alpha_2\beta_2$ tetramer. Subtle changes in the arrangement of subunits within the hemoglobin molecule allow it to carry oxygen from the lungs to tissues with great efficiency (Section 10.2).

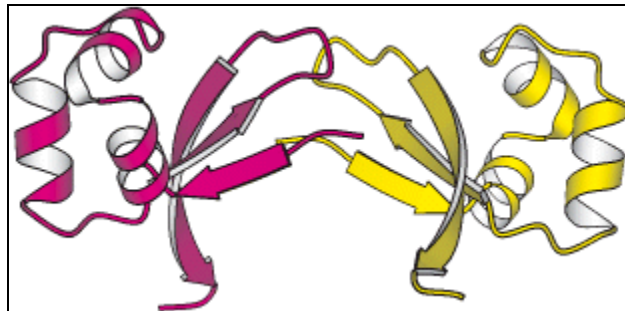


Figure 3.48. Quaternary Structure. The Cro protein of bacteriophage λ is a dimer of identical subunits.



Figure 3.49. The $\alpha_2\beta_2$ Tetramer of Human Hemoglobin. The structure of the two identical α subunits (red) is similar to but not identical with that of the two identical β subunits (yellow). The molecule contains four heme groups (black with the iron atom shown in purple).

Viruses make the most of a limited amount of genetic information by forming coats that use the same kind of subunit repetitively in a symmetric array. The coat of rhinovirus, the virus that causes the common cold, includes 60 copies each of four subunits (Figure 3.50). The subunits come together to form a nearly spherical shell that encloses the viral genome.

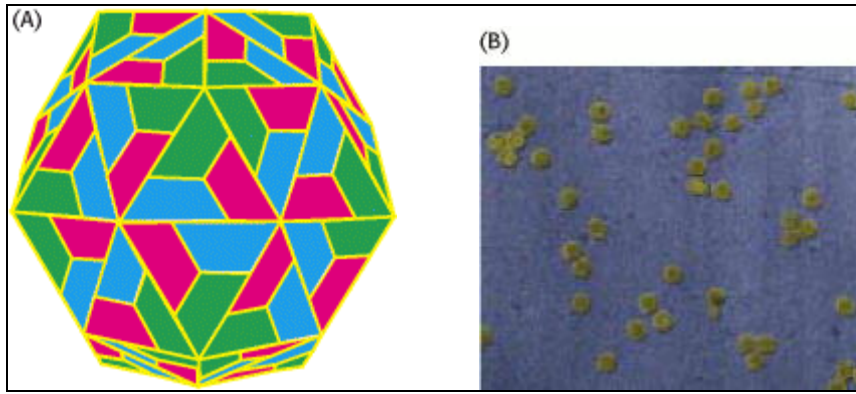


Figure 3.50. Complex Quaternary Structure. The coat of rhinovirus comprises 60 copies of each of four subunits. (A) A schematic view depicting the three types of subunits (shown in red, blue, and green) visible from outside the virus. (B) An electron micrograph showing rhinovirus particles. [Courtesy of Norm Olson, Dept. of Biological Sciences, Purdue University.]

3.6. The Amino Acid Sequence of a Protein Determines Its Three-Dimensional Structure

How is the elaborate three-dimensional structure of proteins attained, and how is the three-dimensional structure related to the one-dimensional amino acid sequence information? The classic work of Christian Anfinsen in the 1950s on the enzyme ribonuclease revealed the relation between the amino acid sequence of a protein and its conformation. Ribonuclease is a single polypeptide chain consisting of 124 amino acid residues cross-linked by four disulfide bonds (Figure 3.51). Anfinsen's plan was to destroy the three-dimensional structure of the enzyme and to then determine what conditions were required to restore the structure.

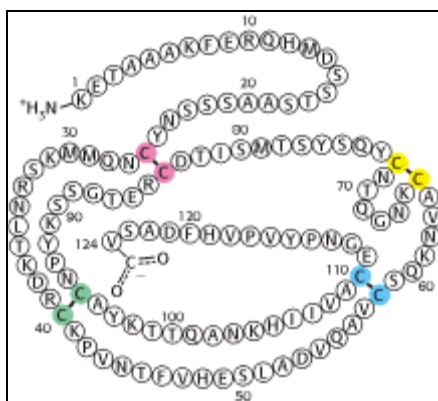


Figure 3.51. Amino Acid Sequence of Bovine Ribonuclease. The four disulfide bonds are shown in color. [After C. H. W. Hirs, S. Moore, and W. H. Stein, *J. Biol. Chem.* 235 (1960):633.]

Agents such as urea or guanidinium chloride effectively disrupt the noncovalent bonds, although the mechanism of action of these agents is not fully understood. The disulfide bonds can be cleaved reversibly by reducing them with a reagent such as β -mercaptoethanol (Figure 3.52). In the presence of a large excess of β -mercaptoethanol, a protein is produced in which the disulfides (cystines) are fully converted into sulfhydryls (cysteines).

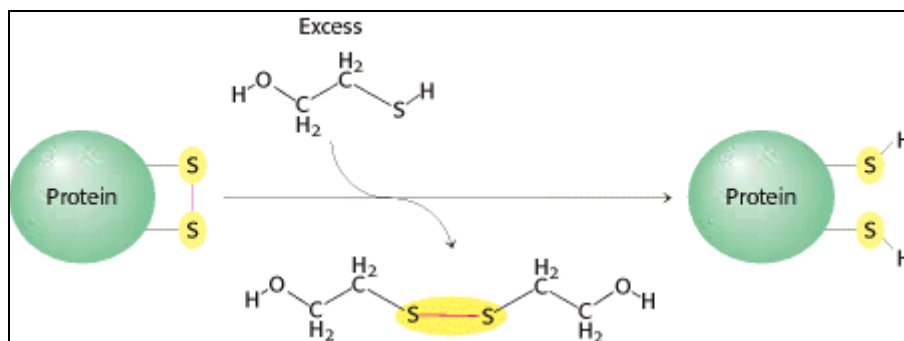
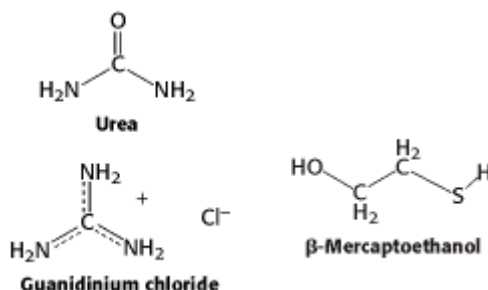


Figure 3.52. Role of β -Mercaptoethanol in Reducing Disulfide Bonds. Note that, as the disulfides are reduced, the β -mercaptoethanol is oxidized and forms dimers.

Most polypeptide chains devoid of cross-links assume a *random-coil conformation* in 8 M urea or 6 M guanidinium chloride, as evidenced by physical properties such as viscosity and optical activity. When ribonuclease was treated with β -mercaptoethanol in 8 M urea, the product was a fully reduced, randomly

coiled polypeptide chain *devoid of enzymatic activity*. In other words, ribonuclease was *denatured* by this treatment (Figure 3.53).

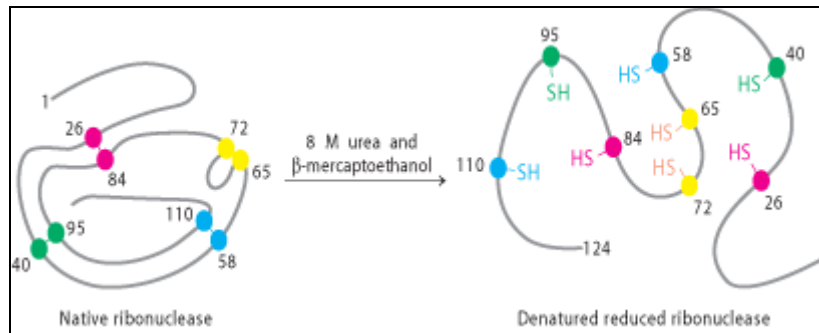


Figure 3.53. Reduction and Denaturation of Ribonuclease.

Anfinsen then made the critical observation that the denatured ribonuclease, freed of urea and β -mercaptoethanol by dialysis, slowly regained enzymatic activity. He immediately perceived the significance of this chance finding: the sulfhydryl groups of the denatured enzyme became oxidized by air, and the enzyme spontaneously refolded into a catalytically active form. Detailed studies then showed that nearly all the original enzymatic activity was regained if the sulfhydryl groups were oxidized under suitable conditions. All the measured physical and chemical properties of the refolded enzyme were virtually identical with those of the native enzyme. These experiments showed that *the information needed to specify the catalytically active structure of ribonuclease is contained in its amino acid sequence*. Subsequent studies have established the generality of this central principle of biochemistry: *sequence specifies conformation*. The dependence of conformation on sequence is especially significant because of the intimate connection between conformation and function.

A quite different result was obtained when reduced ribonuclease was reoxidized while it was still in 8 M urea and the preparation was then dialyzed to remove the urea. Ribonuclease reoxidized in this way had only 1% of the enzymatic activity of the native protein. Why were the outcomes so different when reduced ribonuclease was reoxidized in the presence and absence of urea? The reason is that the wrong disulfides formed pairs in urea. There are 105 different ways of pairing eight cysteine molecules to form four disulfides; only one of these combinations is enzymatically active. The 104 wrong pairings have been picturesquely termed "scrambled" ribonuclease. Anfinsen found that scrambled ribonuclease spontaneously converted into fully active, native ribonuclease when trace amounts of β -mercaptoethanol were added to an aqueous solution of the protein (Figure 3.54). The added β -mercaptoethanol catalyzed the rearrangement of disulfide pairings until the native structure was regained in about 10 hours. *This process was driven by the decrease in free energy as the scrambled conformations were converted into the stable, native conformation of the enzyme*. The native disulfide pairings of ribonuclease thus contribute to the stabilization of the thermodynamically preferred structure.

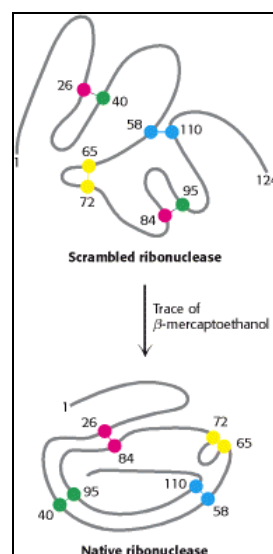


Figure 3.54. Reestablishing Correct Disulfide Pairing. Native ribonuclease can be reformed from scrambled ribonuclease in the presence of a trace of β -mercaptoethanol.

Similar refolding experiments have been performed on many other proteins. In many cases, the native structure can be generated under suitable conditions. For other proteins, however, refolding does not proceed efficiently. In these cases, the unfolding protein molecules usually become tangled up with one another to form aggregates. Inside cells, proteins called *chaperones* block such illicit interactions (Sections 11.3.6).

3.6.1. Amino Acids Have Different Propensities for Forming Alpha Helices, Beta Sheets, and Beta Turns

How does the amino acid sequence of a protein specify its three-dimensional structure? How does an unfolded polypeptide chain acquire the form of the native protein? These fundamental questions in biochemistry can be approached by first asking a simpler one: What determines whether a particular sequence in a protein forms an α helix, a β strand, or a turn? Examining the frequency of occurrence of particular amino acid residues in these secondary structures (Table 3.3) can be a source of insight into this determination. Residues such as alanine, glutamate, and leucine tend to be present in α helices, whereas valine and isoleucine tend to be present in β strands. Glycine, asparagine, and proline have a propensity for being in turns.

Amino acid	α helix	β sheet	Turn
Ala	1.29	0.90	0.78
Cys	1.11	0.74	0.80
Leu	1.30	1.02	0.59
Met	1.47	0.97	0.39
Glu	1.44	0.75	1.00
Gln	1.27	0.80	0.97
His	1.22	1.08	0.69
Lys	1.23	0.77	0.96
Val	0.91	1.49	0.47
Ile	0.97	1.45	0.51
Phe	1.07	1.32	0.58
Tyr	0.72	1.25	1.05
Trp	0.99	1.14	0.75
Thr	0.82	1.21	1.03
Gly	0.56	0.92	1.64
Ser	0.82	0.95	1.33
Asp	1.04	0.72	1.41
Asn	0.90	0.76	1.28
Pro	0.52	0.64	1.91
Arg	0.96	0.99	0.88

The amino acids are grouped according to their preference for α helices (top group), β sheets (second group), or turns (third group). Arginine shows no significant preference for any of the structures. After T. E. Creighton, *Proteins: Structures and Molecular Properties*, 2d ed. (W. H. Freeman and Company, 1992), p. 256.

Table 3.3. Relative frequencies of amino acid residues in secondary structures

The results of studies of proteins and synthetic peptides have revealed some reasons for these preferences. The α helix can be regarded as the default conformation. Branching at the β -carbon atom, as in valine, threonine, and isoleucine, tends to destabilize α helices because of steric clashes. These residues are readily accommodated in β strands, in which their side chains project out of the plane containing the main chain. Serine, aspartate, and asparagine tend to disrupt α helices because their side chains contain hydrogen-bond donors or acceptors in close proximity to the main chain, where they compete for main-chain NH and CO groups. Proline tends to disrupt both α helices and β strands because it lacks an NH group and because its ring structure restricts its ϕ value to near -60 degrees. Glycine readily fits into all structures and for that reason does not favor helix formation in particular.

Can one predict the secondary structure of proteins by using this knowledge of the conformational preferences of amino acid residues? Predictions of secondary structure adopted by a stretch of six or fewer residues have proved to be about 60 to 70% accurate. What stands in the way of more accurate prediction? Note that the conformational preferences of amino acid residues are not tipped all the way to one structure (see Table 3.3). For example, glutamate, one of the strongest helix formers, prefers α helix to β strand by only a factor of two. The preference ratios of most other residues are smaller. Indeed, some penta- and hexapeptide sequences have been found to adopt one structure in one protein and an entirely different structure in another (Figure 3.55). Hence, some amino acid sequences do not uniquely determine secondary structure. Tertiary interactions - interactions between residues that are far apart in the sequence - may be decisive in specifying the secondary structure of some segments. *The context is often crucial in determining the conformational outcome.* The conformation of a protein evolved to work in a particular environment or context.

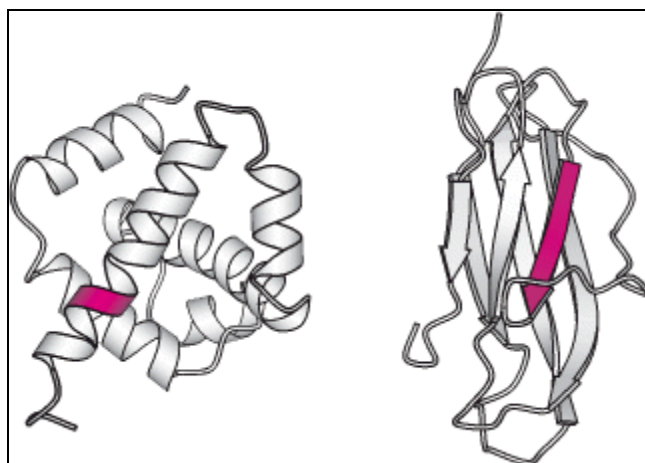


Figure 3.55. Alternative Conformations of a Peptide Sequence. Many sequences can adopt alternative conformations in different proteins. Here the sequence VDLLKN shown in red assumes an α helix in one protein context (left) and a β strand in another (right).

Pathological conditions can result if a protein assumes an inappropriate conformation for the context. Striking examples are *prion diseases*, such as Creutzfeldt-Jacob disease, kuru, and mad cow disease. These conditions result when a brain protein called a prion converts from its normal conformation (designated PrP^C) to an altered one (PrP^{Sc}). This conversion is self-propagating, leading to large aggregates of PrP^{Sc}. The role of these aggregates in the generation of the pathological conditions is not yet understood.

3.6.2. Protein Folding Is a Highly Cooperative Process

As stated earlier, proteins can be denatured by heat or by chemical denaturants such as urea or guanidium chloride. For many proteins, a comparison of the degree of unfolding as the concentration of denaturant increases has revealed a relatively sharp transition from the folded, or native, form to the unfolded, or denatured, form, suggesting that only these two conformational states are present to any significant extent (Figure 3.56). A similar sharp transition is observed if one starts with unfolded proteins and removes the denaturants, allowing the proteins to fold.

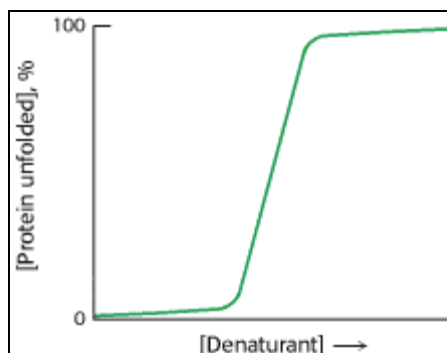


Figure 3.56. Transition from Folded to Unfolded State. Most proteins show a sharp transition from the folded to unfolded form on treatment with increasing concentrations of denaturants.

Protein folding and unfolding is thus largely an "all or none" process that results from a *cooperative transition*. For example, suppose that a protein is placed in conditions under which some part of the protein structure is thermodynamically unstable. As this part of the folded structure is disrupted, the interactions between it and the remainder of the protein will be lost. The loss of these interactions, in turn, will destabilize the remainder of the structure. Thus, conditions that lead to the disruption of any part of a protein structure are likely to unravel the protein completely. The structural properties of proteins provide a clear rationale for the cooperative transition.

The consequences of cooperative folding can be illustrated by considering the contents of a protein solution under conditions corresponding to the middle of the transition between the folded and unfolded forms. Under these conditions, the protein is "half folded." Yet the solution will contain no half-folded molecules but, instead, will be a 50/50 mixture of fully folded and fully unfolded molecules (Figure 3.57). Structures that are partly intact and partly disrupted are not thermodynamically stable and exist only transiently. Cooperative folding ensures that partly folded structures that might interfere with processes within cells do not accumulate.

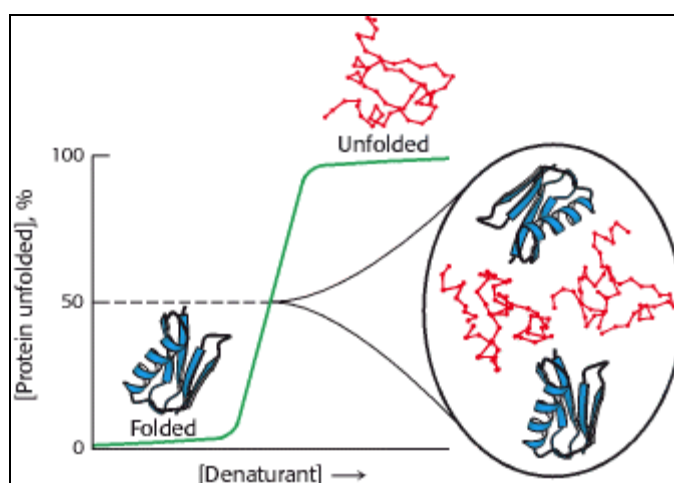


Figure 3.57. Components of a Partially Denatured Protein Solution. In a half-unfolded protein solution, half the molecules are fully folded and half are fully unfolded.

3.6.3. Proteins Fold by Progressive Stabilization of Intermediates Rather Than by Random Search

The cooperative folding of proteins is a thermodynamic property; its occurrence reveals nothing about the kinetics and mechanism of protein folding. How does a protein make the transition from a diverse ensemble of unfolded structures into a unique conformation in the native form? One possibility a priori would be that all possible conformations are tried out to find the energetically most favorable one. How long would such a random search take? Consider a small protein with 100 residues. Cyrus Levinthal calculated that, if each residue can assume three different conformations, the total number of structures would be 3^{100} , which is equal to 5×10^{47} . If it takes 10^{-13} s to convert one structure into another, the total search time would be $5 \times 10^{47} \times 10^{-13}$ s, which is equal to 5×10^{34} s, or 1.6×10^{27} years. Clearly, it would

take much too long for even a small protein to fold properly by randomly trying out all possible conformations. The enormous difference between calculated and actual folding times is called *Levinthal's paradox*.

The way out of this dilemma is to recognize the power of *cumulative selection*. Richard Dawkins, in *The Blind Watchmaker*, asked how long it would take a monkey poking randomly at a typewriter to reproduce Hamlet's remark to Polonius, "Methinks it is like a weasel" (Figure 3.58). An astronomically large number of keystrokes, of the order of 10^{40} , would be required. However, suppose that we preserved each correct character and allowed the monkey to retype only the wrong ones. In this case, only a few thousand keystrokes, on average, would be needed. The crucial difference between these cases is that the first employs a completely random search, whereas, in the second, *partly correct intermediates are retained*.

```

200 ?T(\G{+s x[A.N5~, #ATxSGpn eD@
400 oDr' Jh7s DFR:W4l'u+^v6zpJse0i
600 e2ih'8zs n527x8l8d_ih=#ldseb.
800 S#dh>}/s ]tZqC%lP%DK<!!^aseZ.
1000 V0th>nLs ut/!s]l_kwojjwMasef.
1200 jutH+rvs it is(lukh?SCW-ase5.
1400 Iiithdn4s it is0l/Ks/IxwLase~.
1600 M?thinrs it is lXK?T" woasel.
1800 MStthinws it is lwkN7Kkw(asel.
2000 Mhthin`s it is likv,aww asel.
2200 MMthinns it is lik5avwlasel.
2400 MethinXs it is likydaqw)asel.
2600 Methin4s it is lik2dasweasel.
2800 MethinHs it is likeLaTweasel.
2883 Methinks it is like a weasel.

200 }z-Ng)W4{[cu!kO{d6jS!NlEyUx)p
400 "W hi\kR.<6CFA%4-YIG!iT%6{(16
600 .L-hinkm4(uMGP^IAwoE6klw=yiS
800 AthinkaPa_vYH liR\Hb,Uo4\-"(
1000 OFthinksP)fzO li0v) /+Eln26B
1200 6ithinksMvt -V likm+gl#K-}BFk
1400 vxthinksaEt Dw like.SlGoutks.
1600 :Othinks<it MC likesN2[eaVe4.
1800 uxthinksqit Or likeQh)weaoew.
2000 Y/thinks it id like7alwea)e6.
2200 Methinks it iw like a{weaWel.
2400 Methinks it is like a;weasel.
2431 Methinks it is like a weasel.

```

Figure 3.58. Typing Monkey Analogy. A monkey randomly poking a typewriter could write a line from Shakespeare's *Hamlet*, provided that correct keystrokes were retained. In the two computer simulations shown, the cumulative number of keystrokes is given at the left of each line.

The essence of protein folding is the retention of partly correct intermediates. However, the protein-folding problem is much more difficult than the one presented to our simian Shakespeare. First, the criterion of correctness is not a residue-by-residue scrutiny of conformation by an omniscient observer but rather the total free energy of the transient species. Second, proteins are only marginally stable. The free-energy difference between the folded and the unfolded states of a typical 100-residue protein is 10 kcal mol^{-1} (42 kJ mol^{-1}), and thus each residue contributes on average only $0.1 \text{ kcal mol}^{-1}$ (0.42 kJ mol^{-1}) of energy to maintain the folded state. This amount is less than that of thermal energy, which is $0.6 \text{ kcal mol}^{-1}$ (2.5 kJ mol^{-1}) at room temperature. This meager stabilization energy means that correct intermediates, especially those formed early in folding, can be lost. The analogy is that the monkey would be somewhat free to undo its correct keystrokes. Nonetheless, the interactions that lead to cooperative folding can stabilize intermediates as structure builds up. Thus, local regions, which have significant structural preference, though not necessarily stable on their own, will tend to adopt their favored structures and, as they form, can interact with one other, leading to increasing stabilization.

3.6.4. Prediction of Three-Dimensional Structure from Sequence Remains a Great Challenge

The amino acid sequence completely determines the three-dimensional structure of a protein. However, the prediction of three-dimensional structure from sequence has proved to be extremely difficult. As we have seen, the local sequence appears to determine only between 60% and 70% of the secondary structure; long-range interactions are required to fix the full secondary structure and the tertiary structure.

Investigators are exploring two fundamentally different approaches to predicting three-dimensional structure from amino acid sequence. The first is *ab initio prediction*, which attempts to predict the folding of an amino acid sequence without any direct reference to other known protein structures. Computer-based calculations are employed that attempt to minimize the free energy of a structure with a given amino acid sequence or to simulate the folding process. The utility of these methods is limited by the vast number of possible conformations, the marginal stability of proteins, and the subtle energetics of weak interactions in aqueous solution. The second approach takes advantage of our growing knowledge of the three-dimensional structures of many proteins. In these *knowledge-based methods*, an amino acid sequence of unknown structure is examined for compatibility with any known protein structures. If a significant match is detected, the known structure can be used as an initial model. Knowledge-based methods have been a source of many insights into the three-dimensional conformation of proteins of known sequence but unknown structure.

3.6.5. Protein Modification and Cleavage Confer New Capabilities

Proteins are able to perform numerous functions relying solely on the versatility of their 20 amino acids. However, many proteins are covalently modified, through the attachment of groups other than amino acids, to augment their functions (Figure 3.59). For example, *acetyl groups* are attached to the amino termini of many proteins, a modification that makes these proteins more resistant to degradation. The addition of *hydroxyl groups* to many proline residues stabilizes fibers of newly synthesized collagen, a fibrous protein found in connective tissue and bone. The biological significance of this modification is evident in the disease scurvy: a deficiency of vitamin C results in insufficient hydroxylation of collagen and the abnormal collagen fibers that result are unable to maintain normal tissue strength. Another specialized amino acid produced by a finishing touch is γ -*carboxyglutamate*. In vitamin K deficiency, insufficient carboxylation of glutamate in prothrombin, a clotting protein, can lead to hemorrhage. Many proteins, especially those that are present on the surfaces of cells or are secreted, acquire *carbohydrate units* on specific asparagine residues. The addition of sugars makes the proteins more hydrophilic and able to participate in interactions with other proteins. Conversely, the addition of a *fatty acid* to an α -amino group or a cysteine sulfhydryl group produces a more hydrophobic protein.

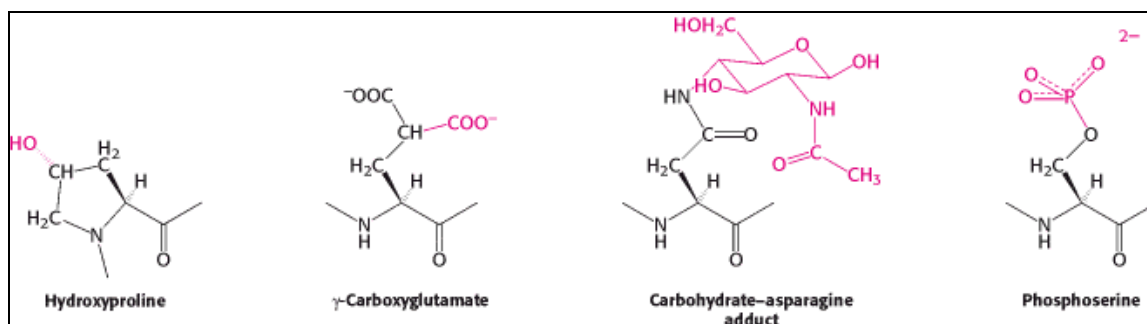


Figure 3.59. **Finishing Touches.** Some common and important covalent modifications of amino acid side chains are shown.

Many hormones, such as epinephrine (adrenaline), alter the activities of enzymes by stimulating the phosphorylation of the hydroxyl amino acids serine and threonine; *phosphoserine* and *phosphothreonine* are the most ubiquitous modified amino acids in proteins. Growth factors such as insulin act by triggering the phosphorylation of the hydroxyl group of tyrosine residues to form *phosphotyrosine*. The phosphoryl groups on these three modified amino acids are readily removed; thus they are able to act as reversible switches in regulating cellular processes. The roles of phosphorylation in signal transduction will be discussed extensively in [Chapter 15](#).

The preceding modifications consist of the addition of special groups to amino acids. Other special groups are generated by chemical rearrangements of side chains and, sometimes, the peptide backbone. For example, certain jellyfish produce a fluorescent green protein (Figure 3.60). The source of the fluorescence is a group formed by the spontaneous rearrangement and oxidation of the sequence Ser-Tyr-Gly within the center of the protein. This protein is of great utility to researchers as a marker within cells (Section 4.3.5).

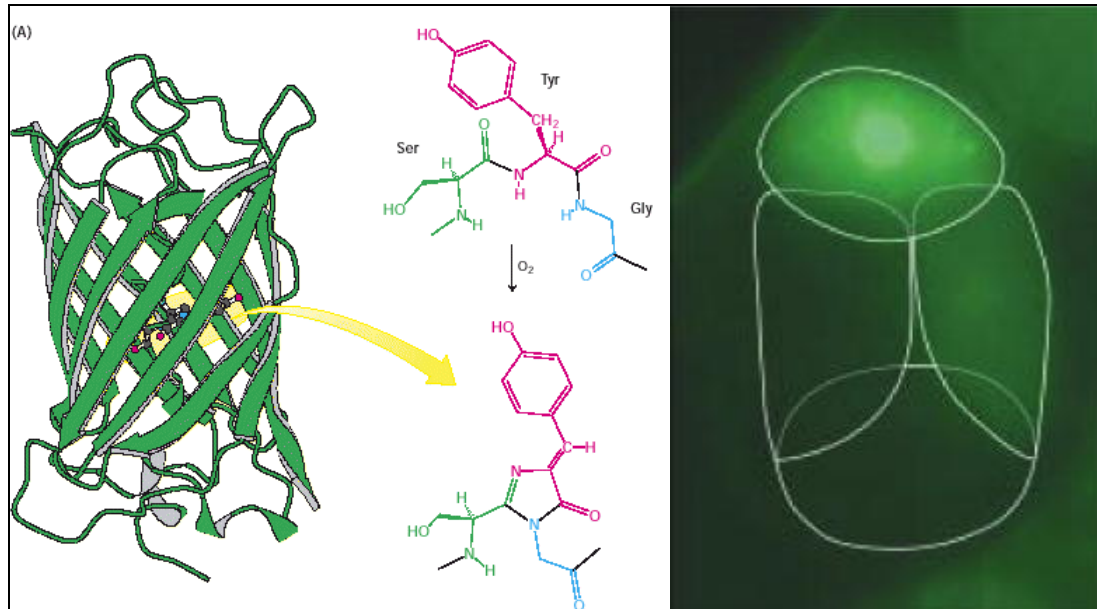


Figure 3.60. Chemical Rearrangement in GFP. (A) The structure of green fluorescent protein (GFP). The rearrangement and oxidation of the sequence Ser-Tyr-Gly is the source of fluorescence. (B) Fluorescence micrograph of a four-cell embryo (cells are outlined) from the roundworm *C. elegans* containing a protein, PIE-1, labeled with GFP. The protein is expressed only in the cell (top) that will give rise to the germline. [(B) Courtesy of Geraldine Seydoux.]

Finally, many proteins are cleaved and trimmed after synthesis. For example, digestive enzymes are synthesized as inactive precursors that can be stored safely in the pancreas. After release into the intestine, these precursors become activated by peptide-bond cleavage. In blood clotting, peptide-bond cleavage converts soluble fibrinogen into insoluble fibrin. A number of polypeptide hormones, such as adrenocorticotrophic hormone, arise from the splitting of a single large precursor protein. Likewise, many virus proteins are produced by the cleavage of large polyprotein precursors. We shall encounter many more examples of modification and cleavage as essential features of protein formation and function. Indeed, these finishing touches account for much of the versatility, precision, and elegance of protein action and regulation.

Summary

Proteins are the workhorses of biochemistry, participating in essentially all cellular processes. Protein structure can be described at four levels. The primary structure refers to the amino acid sequence. The secondary structure refers to the conformation adopted by local regions of the polypeptide chain. Tertiary structure describes the overall folding of the polypeptide chain. Finally, quaternary structure refers to the specific association of multiple polypeptide chains to form multisubunit complexes.

Proteins Are Built from a Repertoire of 20 Amino Acids

Proteins are linear polymers of amino acids. Each amino acid consists of a central tetrahedral carbon atom linked to an amino group, a carboxylic acid group, a distinctive side chain, and a hydrogen. These tetrahedral centers, with the exception of that of glycine, are chiral; only the L isomer exists in natural proteins. All natural proteins are constructed from the same set of 20 amino acids. The side chains of these 20 building blocks vary tremendously in size, shape, and the presence of functional groups. They can be grouped as follows: (1) aliphatic side chains - glycine, alanine, valine, leucine, isoleucine, methionine, and proline; (2) aromatic side chains - phenylalanine, tyrosine, and tryptophan; (3) hydroxyl-containing aliphatic side chains - serine and threonine; (4) sulfhydryl-containing cysteine; (5) basic side chains - lysine, arginine, and histidine; (6) acidic side chains - aspartic acid and glutamic acid; and (7) carboxamide-containing side chains - asparagine and glutamine. These groupings are somewhat arbitrary and many other sensible groupings are possible.

Primary Structure: Amino Acids Are Linked by Peptide Bonds to Form Polypeptide Chains

The amino acids in a polypeptide are linked by amide bonds formed between the carboxyl group of one amino acid and the amino group of the next. This linkage, called a peptide bond, has several important properties. First, it is resistant to hydrolysis so that proteins are remarkably stable kinetically. Second, the peptide group is planar because the C-N bond has considerable double-bond character. Third, each peptide bond has both a hydrogen-bond donor (the NH group) and a hydrogen-bond acceptor (the CO group). Hydrogen bonding between these backbone groups is a distinctive feature of protein structure. Finally, the peptide bond is uncharged, which allows proteins to form tightly packed globular structures having significant amounts of the backbone buried within the protein interior. Because they are linear polymers, proteins can be described as sequences of amino acids. Such sequences are written from the amino to the carboxyl terminus.

Secondary Structure: Polypeptide Chains Can Fold into Regular Structures Such as the Alpha Helix, the Beta Sheet, and Turns and Loops

Two major elements of secondary structure are the α helix and the β strand. In the β helix, the polypeptide chain twists into a tightly packed rod. Within the helix, the CO group of each amino acid is hydrogen bonded to the NH group of the amino acid four residues along the polypeptide chain. In the β strand, the polypeptide chain is nearly fully extended. Two or more β strands connected by NH-to-CO hydrogen bonds come together to form β sheets.

Tertiary Structure: Water-Soluble Proteins Fold into Compact Structures with Nonpolar Cores

The compact, asymmetric structure that individual polypeptides attain is called tertiary structure. The tertiary structures of water-soluble proteins have features in common: (1) an interior formed of amino acids with hydrophobic side chains and (2) a surface formed largely of hydrophilic amino acids that

interact with the aqueous environment. The driving force for the formation of the tertiary structure of water-soluble proteins is the hydrophobic interactions between the interior residues. Some proteins that exist in a hydrophobic environment, in membranes, display the inverse distribution of hydrophobic and hydrophilic amino acids. In these proteins, the hydrophobic amino acids are on the surface to interact with the environment, whereas the hydrophilic groups are shielded from the environment in the interior of the protein.

Quaternary Structure: Polypeptide Chains Can Assemble into Multisubunit Structures

Proteins consisting of more than one polypeptide chain display quaternary structure, and each individual polypeptide chain is called a subunit. Quaternary structure can be as simple as two identical subunits or as complex as dozens of different subunits. In most cases, the subunits are held together by noncovalent bonds.

The Amino Acid Sequence of a Protein Determines Its Three-Dimensional Structure

The amino acid sequence completely determines the three-dimensional structure and, hence, all other properties of a protein. Some proteins can be unfolded completely yet refold efficiently when placed under conditions in which the folded form of the protein is stable. The amino acid sequence of a protein is determined by the sequences of bases in a DNA molecule. This one-dimensional sequence information is extended into the three-dimensional world by the ability of proteins to fold spontaneously. Protein folding is a highly cooperative process; structural intermediates between the unfolded and folded forms do not accumulate.

The versatility of proteins is further enhanced by covalent modifications. Such modifications can incorporate functional groups not present in the 20 amino acids. Other modifications are important to the regulation of protein activity. Through their structural stability, diversity, and chemical reactivity, proteins make possible most of the key processes associated with life.

Key Terms

side chain (R group)

L amino acid

dipolar ion (zwitterion)

peptide bond (amide bond)

disulfide bond

primary structure

phi (ϕ) angle

psi (ψ) angle

Ramachandran diagram

α helix

β pleated sheet

β strand

reverse turn (β turn; hairpin turn)

secondary structure

tertiary structure

domain

subunit

quaternary structure

cooperative transition

Appendix: Acid-Base Concepts

Ionization of Water

Water dissociates into hydronium (H_3O^+) and hydroxyl (OH^-) ions. For simplicity, we refer to the hydronium ion as a hydrogen ion (H^+) and write the equilibrium as



The equilibrium constant K_{eq} of this dissociation is given by

$$K_{\text{eq}} = [\text{H}^+][\text{OH}^-]/[\text{H}_2\text{O}] \quad (1)$$

in which the terms in brackets denote molar concentrations. Because the concentration of water (55.5 M) is changed little by ionization, expression 1 can be simplified to give

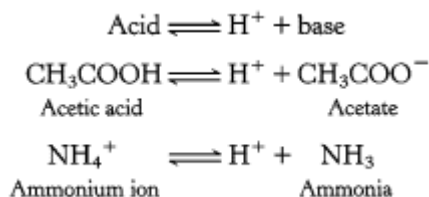
$$K_{\text{w}} = [\text{H}^+][\text{OH}^-] \quad (2)$$

in which K_{w} is the ion product of water. At 25°C, K_{w} is 1.0×10^{-14} .

Note that the concentrations of H^+ and OH^- are reciprocally related. If the concentration of H^+ is high, then the concentration of OH^- must be low, and vice versa. For example, if $[\text{H}^+] = 10^{-2}$ M, then $[\text{OH}^-] = 10^{-12}$ M.

Definition of Acid and Base

An acid is a proton donor. A base is a proton acceptor.



The species formed by the ionization of an acid is its conjugate base. Conversely, protonation of a base yields its conjugate acid. Acetic acid and acetate ion are a conjugate acid-base pair.

Definition of pH and pK

The pH of a solution is a measure of its concentration of H^+ . The pH is defined as

$$\text{pH} = \log_{10}(1/[\text{H}^+]) = -\log_{10}[\text{H}^+] \quad (3)$$

The ionization equilibrium of a weak acid is given by



The apparent equilibrium constant K_{a} for this ionization is

$$K_{\text{a}} = [\text{H}^+][\text{A}^-]/[\text{HA}] \quad (4)$$

The $\text{p}K_{\text{a}}$ of an acid is defined as

$$\text{p}K_{\text{a}} = -\log K_{\text{a}} = \log(1/K_{\text{a}}) \quad (5)$$

Inspection of equation 4 shows that the $\text{p}K_{\text{a}}$ of an acid is the pH at which it is half dissociated, when $[\text{A}^-]=[\text{HA}]$.

Henderson-Hasselbalch Equation

What is the relation between pH and the ratio of acid to base? A useful expression can be derived from equation 4. Rearrangement of that equation gives

$$1/[H^+] = 1/K_a[A^-]/[HA] \quad (6)$$

Taking the logarithm of both sides of equation 6 gives

$$\log(1/[H^+]) = \log(1/K_a) + \log([A^-]/[HA]) \quad (7)$$

Substituting pH for $\log 1/[H^+]$ and pK_a for $\log 1/K_a$ in equation 7 yields

$$pH = pK_a + \log([A^-]/[HA]) \quad (8)$$

which is commonly known as the Henderson-Hasselbalch equation.

The pH of a solution can be calculated from equation 8 if the molar proportion of A^- to HA and the pK_a of HA are known. Consider a solution of 0.1 M acetic acid and 0.2 M acetate ion. The pK_a of acetic acid is 4.8. Hence, the pH of the solution is given by

$$pH = 4.8 + \log(0.2/0.1) = 4.8 + \log 2.0 = 4.8 + 0.3 = 5.1$$

Conversely, the pK_a of an acid can be calculated if the molar proportion of A^- to HA and the pH of the solution are known.

Buffers

An acid-base conjugate pair (such as acetic acid and acetate ion) has an important property: it resists changes in the pH of a solution. In other words, it acts as a *buffer*. Consider the addition of OH^- to a solution of acetic acid (HA):



A plot of the dependence of the pH of this solution on the amount of OH^- added is called a *titration curve* (Figure 3.61). Note that there is an inflection point in the curve at pH 4.8, which is the pK_a of acetic acid. In the vicinity of this pH, a relatively large amount of OH^- produces little change in pH. In other words, the buffer maintains the value of pH near a given value, despite the addition of other either protons or hydroxide ions. In general, a weak acid is most effective in buffering against pH changes in the vicinity of its pK_a value.

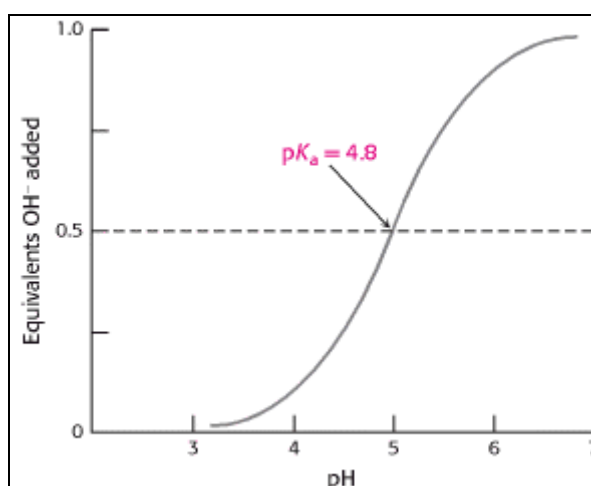


Figure 3.61. Titration Curve of Acetic Acid.

pK_a Values of Amino Acids

An amino acid such as glycine contains two ionizable groups: an α -carboxyl group and a protonated α -amino group. As base is added, these two groups are titrated (Figure 3.62). The pK_a of the α -COOH group is 2.4, whereas that of the α -NH₃⁺ group is 9.8. The pK_a values of these groups in other amino acids are similar (Table 3.4). Some amino acids, such as aspartic acid, also contain an ionizable side chain. The pK_a values of ionizable side chains in amino acids range from 3.9 (aspartic acid) to 12.5 (arginine).

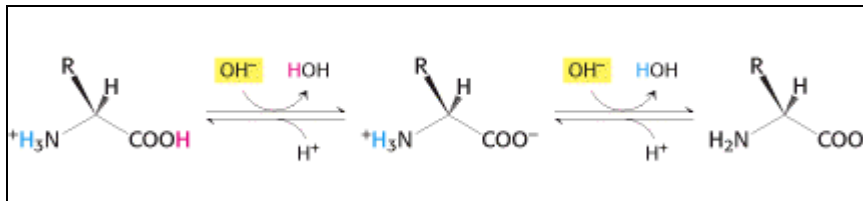


Figure 3.62. Titration of the α -Carboxyl and α -Amino Groups of an Amino Acid.

Amino acid	pK _a values (25°C)		
	α -COOH group	α -NH ₃ ⁺ group	Side chain
Alanine	2.3	9.9	
Glycine	2.4	9.8	
Phenylalanine	1.8	9.1	
Serine	2.1	9.2	
Valine	2.3	9.6	
Aspartic acid	2.0	10.0	3.9
Glutamic acid	2.2	9.7	4.3
Histidine	1.8	9.2	6.0
Cysteine	1.8	10.8	8.3
Tyrosine	2.2	9.1	10.9
Lysine	2.2	9.2	10.8
Arginine	1.8	9.0	12.5

After J. T. Edsall and J. Wyman, *Biophysical Chemistry* (Academic Press, 1958), Chapter 8.

Table 3.4. pK_a values of some amino acids

Problems

1. **Shape and dimension.** (a) Tropomyosin, a 70-kd muscle protein, is a two-stranded α -helical coiled coil. Estimate the length of the molecule. (b) Suppose that a 40-residue segment of a protein folds into a two-stranded antiparallel β structure with a 4-residue hairpin turn. What is the longest dimension of this motif?

Answer:

(a) Each strand is 35 kd and hence has about 318 residues (the mean residue mass is 110 daltons). Because the rise per residue in an α helix is 1.5 Å, the length is 477 Å. More precisely, for an α -helical coiled coil the rise per residue is 1.46 Å so that the length will be 464 Å.

(b) Eighteen residues in each strand (40 minus 4 divided by 2) are in a β -sheet conformation. Because the rise per residue is 3.5 Å, the length is 63 Å.

2. **Contrasting isomers.** Poly-L-leucine in an organic solvent such as dioxane is α helical, whereas poly-L-isoleucine is not. Why do these amino acids with the same number and kinds of atoms have different helix-forming tendencies?

Answer:

The methyl group attached to the β -carbon atom of isoleucine sterically interferes with α -helix formation. In leucine, this methyl group is attached to the γ -carbon atom, which is farther from the main chain and hence does not interfere.

3. **Active again.** A mutation that changes an alanine residue in the interior of a protein to valine is found to lead to a loss of activity. However, activity is regained when a second mutation at a different position changes an isoleucine residue to glycine. How might this second mutation lead to a restoration of activity?

Answer:

The first mutation destroys activity because valine occupies more space than alanine does, and so the protein must take a different shape, assuming that this residue lies in the closely packed interior. The second mutation restores activity because of a compensatory reduction of volume; glycine is smaller than isoleucine.

4. **Shuffle test.** An enzyme that catalyzes disulfide-sulfhydryl exchange reactions, called protein disulfide isomerase (PDI), has been isolated. PDI rapidly converts inactive scrambled ribonuclease into enzymatically active ribonuclease. In contrast, insulin is rapidly inactivated by PDI. What does this important observation imply about the relation between the amino acid sequence of insulin and its three-dimensional structure?

Answer:

The native conformation of insulin is not the thermodynamically most stable form since it contains two separate chains linked by disulfide bonds. Insulin is formed from proinsulin, a single-chain precursor, that is cleaved to form insulin with 33 residues once the disulfide bonds have formed.

5. **Stretching a target.** A protease is an enzyme that catalyzes the hydrolysis of the peptide bonds of target proteins. How might a protease bind a target protein so that its main chain becomes fully extended in the vicinity of the vulnerable peptide bond?

Answer:

A segment of the main chain of the protease could hydrogen bond to the main chain of the substrate to form an extended parallel or antiparallel pair of β strands.

6. Often irreplaceable. Glycine is a highly conserved amino acid residue in the evolution of proteins. Why?

Answer:

Glycine has the smallest side chain of any amino acid. Its size often is critical in allowing polypeptide chains to make tight turns or to approach one another closely.

7. Potential partners. Identify the groups in a protein that can form hydrogen bonds or electrostatic bonds with an arginine side chain at pH 7.

Answer:

Glutamate, aspartate, and the terminal carboxylate can form salt bridges with the guanidinium group of arginine. In addition, this group can be a hydrogen-bond donor to the side chains of glutamine, asparagine, serine, threonine, aspartate, and glutamate, and to the main-chain carbonyl group.

8. Permanent waves. The shape of hair is determined in part by the pattern of disulfide bonds in keratin, its major protein. How can curls be induced?

Answer:

Disulfide bonds in hair are broken by adding a thiol and applying gentle heat. The hair is curled, and an oxidizing agent is added to re-form disulfide bonds to stabilize the desired shape.

9. Location is everything. Proteins that span biological membranes often contain α helices. Given that the insides of membranes are highly hydrophobic (Section 12.2.1), predict what type of amino acids would be in such a helix. Why is an α helix particularly suited to exist in the hydrophobic environment of the interior of a membrane?

Answer:

The amino acids would be hydrophobic in nature. An α helix is especially suited to cross a membrane because all of the amide hydrogen atoms and carbonyl oxygen atoms of the peptide backbone take part in intrachain hydrogen bonds, thus stabilizing these polar atoms in a hydrophobic environment.

10. Issues of stability. Proteins are quite stable. The lifetime of a peptide bond in aqueous solution is nearly 1000 years. However, the ΔG° of hydrolysis of proteins is negative and quite large. How can you account for the stability of the peptide bond in light of the fact that hydrolysis releases much energy?

Answer:

The energy barrier that must be crossed to go from the polymerized state to the hydrolyzed state is large even though the reaction is thermodynamically favorable.

11. **Minor species.** For an amino acid such as alanine, the major species in solution at pH 7 is the zwitterionic form. Assume a pK_a value of 8 for the amino group and a pK_a value of 3 for the carboxylic acid and estimate the ratio of the concentration of neutral amino acid species (with the carboxylic acid protonated and the amino group neutral) to that of the zwitterionic species at pH 7.

Answer:

Using the Henderson-Hasselbach equation, we find the ratio of alanine-COOH to alanine-COO⁻ at pH 7 to be 10^{-4} . The ratio of alanine-NH₂ to alanine-NH₃⁺, determined in the same fashion, is 10^{-1} . Thus, the ratio of neutral alanine to zwitterionic species is $10^{-4} \times 10^{-1} = 10^{-5}$.

12. **A matter of convention.** All L amino acids have an S absolute configuration except L-cysteine, which has the R configuration. Explain why L-cysteine is designated as the R absolute configuration.

Answer:

The assignment of absolute configuration requires the assignment of priorities to the four groups connected to a tetrahedral carbon. For all amino acids except cysteine, the priorities are: (1) amino group; (2) carbonyl group; (3) side chain; (4) hydrogen. For cysteine, because of the sulfur atom in its side chain, the side chain has a greater priority than does the carbonyl group, leading to the assignment of an R rather than S configuration.

13. **Hidden message.** Translate the following amino acid sequence into one-letter code: Leu-Glu-Ala-Arg-Asn-Ile-Asn-Gly-Ser-Cys-Ile-Glu-Asn-Cys-Glu-Ile-Ser-Gly-Arg-Glu-Ala-Thr.

Answer:

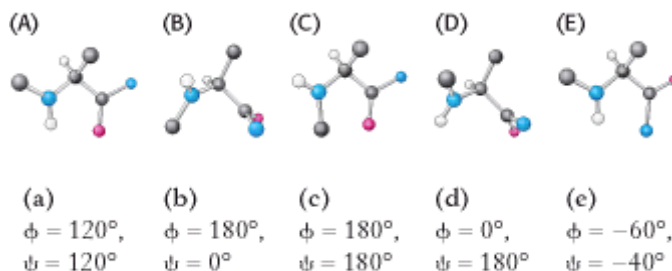
LEARNINGSOURCEISGREAT.

14. **Who goes first?** Would you expect Pro-X peptide bonds to tend to have cis conformations like those of X-Pro bonds? Why or why not?

Answer:

No, Pro-X would have the characteristics of any other peptide bond. The steric hindrance in X-Pro arises because the R group of Pro is bonded to the amino group. Hence, in X-Pro, the proline R group is near the R group of X. This would not be the case in Pro-X.

15. **Matching.** For each of the amino acid derivatives shown below (A-E), find the matching set of ϕ and ψ values (a-e).



Answer:

A, c; B, e; C, d; D, a; E, b.

- 16. Concentrate on the concentration.** A solution of a protein whose sequence includes three tryptophan residues, no tyrosine residues, and no phenylalanine residues has an absorbance of 0.1 at 280 nm in a cell with a path length of 1 cm. Estimate the concentration of the protein in units of molarity. If the protein has a molecular mass of 100 kd, estimate the concentration in units of milligrams of protein per milliliter of solution.

Answer:

With the use of Beer's law and the value of ϵ obtained from Section 3.1 ($\epsilon = 3400 \text{ M}^{-1} \text{ cm}^{-1}$), the concentration of tryptophan is found to be $\approx 30 \mu\text{M}$. Because there are three molecules of tryptophan per molecule of protein, the concentration of protein is $\approx 10 \mu\text{M}$. There is 1 mg of protein per milliliter of solution.

Selected Readings

Where to start

J.S. Richardson. 1981. The anatomy and taxonomy of protein structure *Adv. Protein Chem.* 34: 167-339. ([PubMed](#))

R.F. Doolittle. 1985. *Proteins Sci. Am.* 253: (4) 88-99. ([PubMed](#))

F.M. Richards. 1991. The protein folding problem *Sci. Am.* 264: (1) 54-57. ([PubMed](#))

A.L. Weber and S.L. Miller. 1981. Reasons for the occurrence of the twenty coded protein amino acids *J. Mol. Evol.* 17: 273-284. ([PubMed](#))

Books

Branden, C., Tooze, J., 1999. *Introduction to Protein Structure* (2d ed.). Garland.

Perutz, M. F., 1992. *Protein Structure: New Approaches to Disease and Therapy*. W. H. Freeman and Company.

Creighton, T. E., 1992. *Proteins: Structures and Molecular Principles* (2d ed.). W. H. Freeman and Company.

Schultz, G. E., and Schirmer, R. H., 1979. *Principles of Protein Structure*. Springer-Verlag.

Conformation of proteins

J.S. Richardson, D.C. Richardson, N.B. Tweedy, K.M. Gernert, T.P. Quinn, M.H. Hecht, B.W. Erickson, Y. Yan, R.D. McClain, M.E. Donlan, and M.C. Suries. 1992. Looking at proteins: Representations, folding, packing, and design *Biophys. J.* 63: 1186-1220.

C. Chothia and A.V. Finkelstein. 1990. The classification and origin of protein folding patterns *Annu. Rev. Biochem.* 59: 1007-1039. ([PubMed](#))

Alpha helices, beta sheets, and loops

K.T. O'Neil and W.F. DeGrado. 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids *Science* 250: 646-651. ([PubMed](#))

C. Zhang and S.H. Kim. 2000. The anatomy of protein beta-sheet topology *J. Mol. Biol.* 299: 1075-1089. ([PubMed](#))

L. Regan. 1994. Protein structure: Born to be beta *Curr. Biol.* 4: 656-658. ([PubMed](#))

J.F. Leszczynski and G.D. Rose. 1986. Loops in globular proteins: A novel category of secondary structure *Science* 234: 849-855. ([PubMed](#))

R. Srinivasan and G.D. Rose. 1999. A physical basis for protein secondary structure *Proc. Natl. Acad. Sci. U. S. A.* 96: 14258-14263. ([PubMed](#)) ([Full Text in PMC](#))

Domains

M.J. Bennett, S. Choe, and D. Eisenberg. 1994. Domain swapping: Entangling alliances between proteins *Proc. Natl. Acad. Sci. U. S. A.* 91: 3127-3131. ([PubMed](#)) ([Full Text in PMC](#))

M. Bergdoll, L.D. Eltis, A.D. Cameron, P. Dumas, and J.T. Bolin. 1998. All in the family: Structural and evolutionary relationships among three modular proteins with diverse functions and variable assembly *Protein Sci.* 7: 1661-1670. ([PubMed](#))

K.P. Hopfner, E. Kopetzki, G.B. Kresse, W. Bode, R. Huber, and R.A. Engh. 1998. New enzyme lineages by subdomain shuffling *Proc. Natl. Acad. Sci. U. S. A.* 95: 9813-9818. ([PubMed](#)) ([Full Text in PMC](#))

C.P. Ponting, J. Schultz, R.R. Copley, M.A. Andrade, and P. Bork. 2000. Evolution of domain families *Adv. Protein Chem.* 54: 185-244. ([PubMed](#))

Protein folding

C.B. Anfinsen. 1973. Principles that govern the folding of protein chains *Science* 181: 223-230. ([PubMed](#))

R.L. Baldwin and G.D. Rose. 1999. Is protein folding hierarchic? I. Local structure and peptide folding *Trends Biochem. Sci.* 24: 26-33. ([PubMed](#))

R.L. Baldwin and G.D. Rose. 1999. Is protein folding hierarchic? II. Folding intermediates and transition states *Trends Biochem. Sci.* 24: 77-83. ([PubMed](#))

J.P. Staley and P.S. Kim. 1990. Role of a subdomain in the folding of bovine pancreatic trypsin inhibitor *Nature* 344: 685-688. ([PubMed](#))

J.L. Neira and A.R. Fersht. 1999. Exploring the folding funnel of a polypeptide chain by biophysical studies on protein fragments *J. Mol. Biol.* 285: 1309-1333. ([PubMed](#))

Covalent modification of proteins

R.G. Krishna and F. Wold. 1993. Post-translational modification of proteins *Adv. Enzymol. Relat. Areas. Mol. Biol.* 67: 265-298. ([PubMed](#))

J.M. Aletta, T.R. Cimato, and M.J. Ettinger. 1998. Protein methylation: A signal event in post-translational modification *Trends Biochem. Sci.* 23: 89-91. ([PubMed](#))

Glazer, A. N., DeLange, R. J., and Sigman, D. S., 1975. *Chemical Modification of Proteins*. North-Holland.

R.Y. Tsien. 1998. The green fluorescent protein *Annu. Rev. Biochem.* 67: 509-544. ([PubMed](#))

Molecular graphics

P. Kraulis. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures *J. Appl. Cryst.* 24: 946-950.

T. Ferrin, C. Huang, L. Jarvis, and R. Langridge. 1988. The MIDAS display system *J. Mol. Graphics* 6: 13-27.

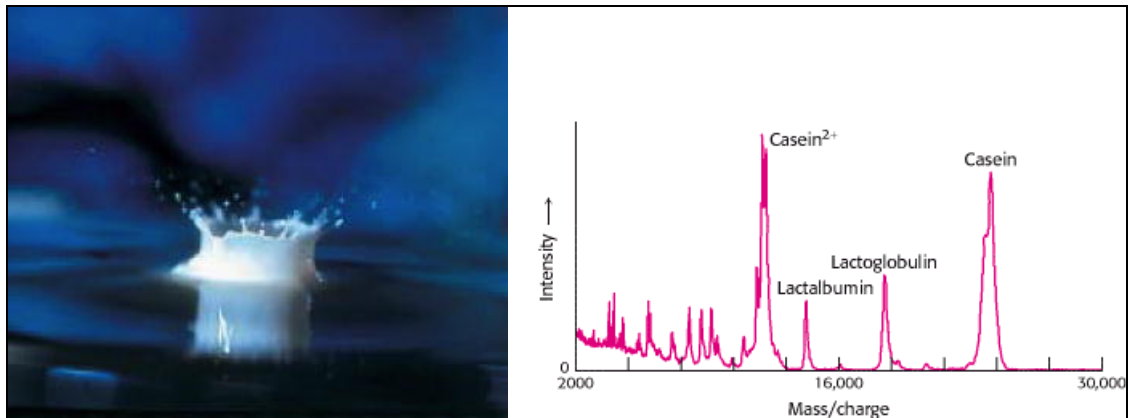
D.C. Richardson and J.S. Richardson. 1994. Kinemages: Simple macromolecular graphics for interactive teaching and publication *Trends Biochem. Sci.* 19: 135-138. ([PubMed](#))

4. Exploring Proteins

In the preceding chapter, we saw that proteins play crucial roles in nearly all biological processes - in catalysis, signal transmission, and structural support. This remarkable range of functions arises from the existence of thousands of proteins, each folded into a distinctive three-dimensional structure that enables it to interact with one or more of a highly diverse array of molecules. A major goal of biochemistry is to determine how amino acid sequences specify the conformations of proteins. Other goals are to learn how individual proteins bind specific substrates and other molecules, mediate catalysis, and transduce energy and information.

The purification of the protein of interest is the indispensable first step in a series of studies aimed at exploring protein function. Proteins can be separated from one another on the basis of solubility, size, charge, and binding ability. When a protein has been purified, the amino acid sequence can be determined. The strategy is to divide and conquer, to obtain specific fragments that can be readily sequenced. Automated peptide sequencing and the application of recombinant DNA methods are providing a wealth of amino acid sequence data that are opening new vistas. To understand the physiological context of a protein, antibodies are choice probes for locating proteins in vivo and measuring their quantities. Monoclonal antibodies able to probe for specific proteins can be obtained in large amounts. The synthesis of peptides is possible, which makes feasible the synthesis of new drugs, functional protein fragments, and antigens for inducing the formation of specific antibodies. Nuclear magnetic resonance (NMR) spectroscopy and x-ray crystallography are the principal techniques for elucidating three-dimensional structure, the key determinant of function.

The exploration of proteins by this array of physical and chemical techniques has greatly enriched our understanding of the molecular basis of life and makes it possible to tackle some of the most challenging questions of biology in molecular terms.



Milk, a source of nourishment for all mammals, is composed, in part, of a variety of proteins. The protein components of milk are revealed by the technique of MALDI-TOF mass spectrometry, which separates molecules on the basis of their mass to charge ratio. [(Left) Jean Paul Iris/FPG (Right) courtesy of Brian Chait.]

4.0.1. The Proteome Is the Functional Representation of the Genome

Many organisms are yielding their DNA base sequences to gene sequencers, including several metazoans. The roundworm *Caenorhabditis elegans* has a genome of 97 million bases and about 19,000 protein-encoding genes, whereas that of the fruit fly *Drosophila melanogaster* contains 180 million bases and about 14,000 genes. The incredible progress being made in gene sequencing has already culminated in the elucidation of the complete sequence of the human genome, all 3 billion bases with an estimated 40,000 genes. But this genomic knowledge is analogous to a list of parts for a car - it does not explain how the parts work together. A new word has been coined, the *proteome*, to signify a more complex level of information content, the level of *functional information*, which encompasses the type, functions, and interactions of proteins that yield a functional unit.

The term proteome is derived from *proteins* expressed by the *genome*. Whereas the genome tells us what is possible, the proteome tells us what is functionally present - for example, which proteins interact to form a signal-transduction pathway or an ion channel in a membrane. The proteome is not a fixed characteristic of the cell. Rather, because it represents the functional expression of information, it varies with cell type, developmental stage, and environmental conditions, such as the presence of hormones. The proteome is much larger than the genome because of such factors as alternatively spliced RNA, the posttranslational modification of proteins, the temporal regulation of protein synthesis, and varying protein-protein interactions. Unlike the genome, the proteome is not static.

An understanding of the proteome is acquired by investigating, characterizing, and cataloging proteins. An investigator often begins the process by separating a particular protein from all other biomolecules in the cell.

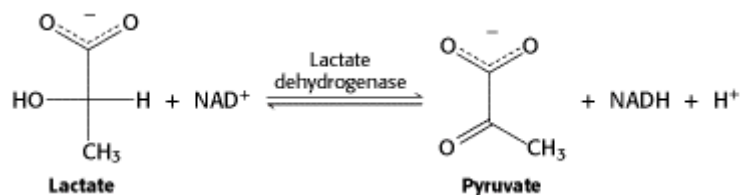
4.1. The Purification of Proteins Is an Essential First Step in Understanding Their Function

An adage of biochemistry is, Never waste pure thoughts on an impure protein. Starting from pure proteins, we can determine amino acid sequences and evolutionary relationships between proteins in diverse organisms and we can investigate a protein's biochemical function. Moreover, crystals of the protein may be grown from pure protein, and from such crystals we can obtain x-ray data that will provide us with a picture of the protein's tertiary structure - the actual *functional* unit.

4.1.1. The Assay: How Do We Recognize the Protein That We Are Looking For?

Purification should yield a sample of protein containing only one type of molecule, the protein in which the biochemist is interested. This protein sample may be only a fraction of 1% of the starting material, whether that starting material consists of cells in culture or a particular organ from a plant or animal. How is the biochemist able to isolate a particular protein from a complex mixture of proteins?

The biochemist needs a test, called an *assay*, for some unique identifying property of the protein so that he or she can tell when the protein is present. Determining an effective assay is often difficult; but the more specific the assay, the more effective the purification. For enzymes, which are protein catalysts ([Chapter 8](#)), the assay is usually based on the reaction that the enzyme catalyzes in the cell. Consider the enzyme lactate dehydrogenase, an important player in the anaerobic generation of energy from glucose as well as in the synthesis of glucose from lactate. Lactate dehydrogenase carries out the following reaction:



Nicotinamide adenine dinucleotide [reduced (NADH); [Section 14.3.1](#)] is distinguishable from the other components of the reaction by its ability to absorb light at 340 nm. Consequently, we can follow the progress of the reaction by examining how much light the reaction mixture absorbs at 340 nm in unit time - for instance, within 1 minute after the addition of the enzyme. Our assay for enzyme activity during the purification of lactate dehydrogenase is thus the increase in absorbance of light at 340 nm observed in 1 minute.

To be certain that our purification scheme is working, we need one additional piece of information - the amount of protein present in the mixture being assayed. There are various rapid and accurate means of determining protein concentration. With these two experimentally determined numbers - enzyme activity and protein concentration - we then calculate the *specific activity*, the ratio of enzyme activity to the amount of protein in the enzyme assay. The specific activity will rise as the purification proceeds and the protein mixture being assayed consists to a greater and greater extent of lactate dehydrogenase. In essence, the point of the purification is to maximize the specific activity.

4.1.2. Proteins Must Be Released from the Cell to Be Purified

Having found an assay and chosen a source of protein, we must now fractionate the cell into components and determine which component is enriched in the protein of interest. Such fractionation schemes are developed by trial and error, on the basis of previous experience. In the first step, a *homogenate* is formed by disrupting the cell membrane, and the mixture is fractionated by centrifugation, yielding a dense pellet of heavy material at the bottom of the centrifuge tube and a lighter supernatant above ([Figure 4.1](#)). The supernatant is again centrifuged at a greater force to yield yet another pellet and supernatant. The procedure, called *differential centrifugation*, yields several fractions of decreasing density, each still containing hundreds of different proteins, which are subsequently assayed for the activity being purified.

Usually, one fraction will be enriched for such activity, and it then serves as the source of material to which more discriminating purification techniques are applied.

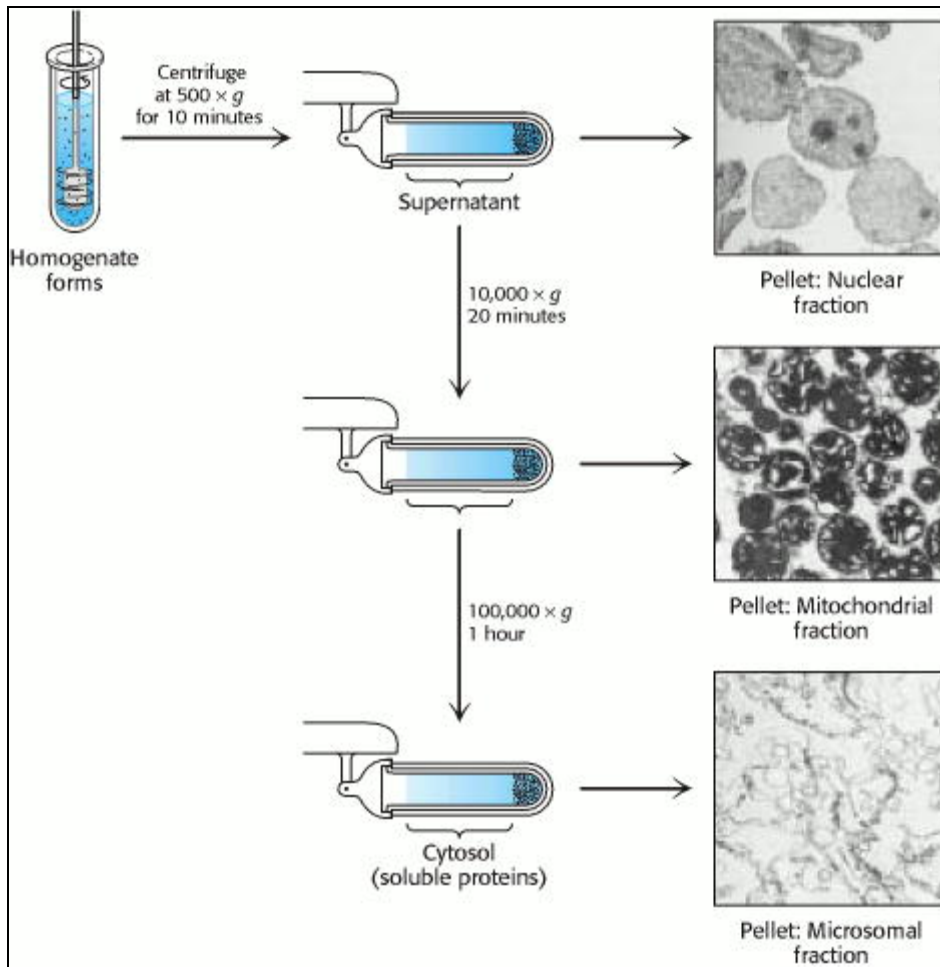


Figure 4.1. Differential Centrifugation. Cells are disrupted in a homogenizer and the resulting mixture, called the homogenate, is centrifuged in a step-by-step fashion of increasing centrifugal force. The denser material will form a pellet at lower centrifugal force than will the less-dense material. The isolated fractions can be used for further purification. [Photographs courtesy of S. Fleischer and B. Fleischer.]

4.1.3. Proteins Can Be Purified According to Solubility, Size, Charge, and Binding Affinity

Several thousand proteins have been purified in active form on the basis of such characteristics as *solubility*, *size*, *charge*, and *specific binding affinity*. Usually, protein mixtures are subjected to a series of separations, each based on a different property to yield a pure protein. At each step in the purification, the preparation is assayed and the protein concentration is determined. Substantial quantities of purified proteins, of the order of many milligrams, are needed to fully elucidate their three-dimensional structures and their mechanisms of action. Thus, the overall yield is an important feature of a purification scheme. A variety of purification techniques are available.

Salting Out.

Most proteins are less soluble at high salt concentrations, an effect called *salting out*. The salt concentration at which a protein precipitates differs from one protein to another. Hence, salting out can be used to fractionate proteins. For example, 0.8 M ammonium sulfate precipitates fibrinogen, a blood-clotting protein, whereas a concentration of 2.4 M is needed to precipitate serum albumin. Salting out is also useful for concentrating dilute solutions of proteins, including active fractions obtained from other purification steps. Dialysis can be used to remove the salt if necessary.

Dialysis.

Proteins can be separated from small molecules by *dialysis* through a semipermeable membrane, such as a cellulose membrane with pores (Figure 4.2). Molecules having dimensions significantly greater than the pore diameter are retained inside the dialysis bag, whereas smaller molecules and ions traverse the pores of such a membrane and emerge in the dialysate outside the bag. This technique is useful for removing a salt or other small molecule, but it will not distinguish between proteins effectively.

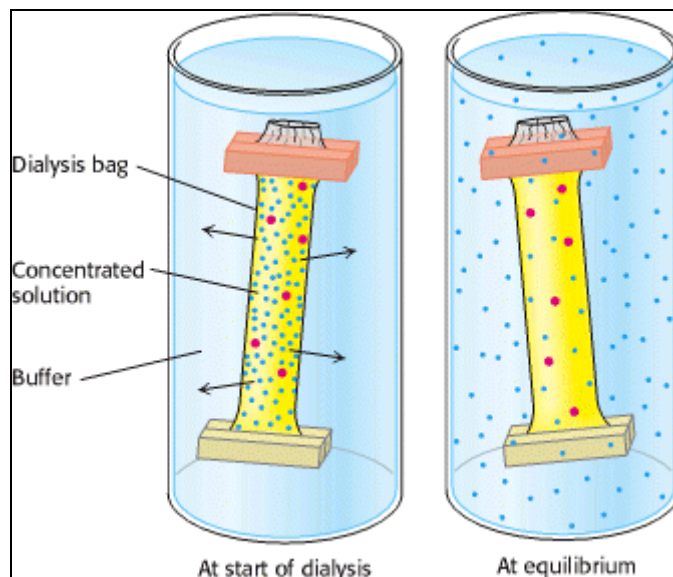


Figure 4.2. Dialysis. Protein molecules (red) are retained within the dialysis bag, whereas small molecules (blue) diffuse into the surrounding medium.

Gel-Filtration Chromatography.

More discriminating separations on the basis of size can be achieved by the technique of *gel-filtration chromatography* (Figure 4.3). The sample is applied to the top of a column consisting of porous beads made of an insoluble but highly hydrated polymer such as dextran or agarose (which are carbohydrates) or polyacrylamide. Sephadex, Sepharose, and Bio-gel are commonly used commercial preparations of these beads, which are typically 100 μm (0.1 mm) in diameter. Small molecules can enter these beads, but large ones cannot. The result is that small molecules are distributed in the aqueous solution both inside the beads and between them, whereas large molecules are located only in the solution between the beads. *Large molecules flow more rapidly through this column and emerge first because a smaller volume is accessible to them.* Molecules that are of a size to occasionally enter a bead will flow from the column at an intermediate position, and small molecules, which take a longer, tortuous path, will exit last.

Ion-Exchange Chromatography.

Proteins can be separated on the basis of their net charge by *ion-exchange chromatography*. If a protein has a net positive charge at pH 7, it will usually bind to a column of beads containing carboxylate groups, whereas a negatively charged protein will not (Figure 4.4). A positively charged protein bound to such a column can then be eluted (released) by increasing the concentration of sodium chloride or another salt in the eluting buffer because sodium ions compete with positively charged groups on the protein for binding to the column. Proteins that have a low density of net positive charge will tend to emerge first, followed by those having a higher charge density. Positively charged proteins (cationic proteins) can be separated on negatively charged carboxymethyl-cellulose (CM-cellulose) columns. Conversely, negatively charged proteins (anionic proteins) can be separated by chromatography on positively charged diethylaminoethyl-cellulose (DEAE-cellulose) columns.



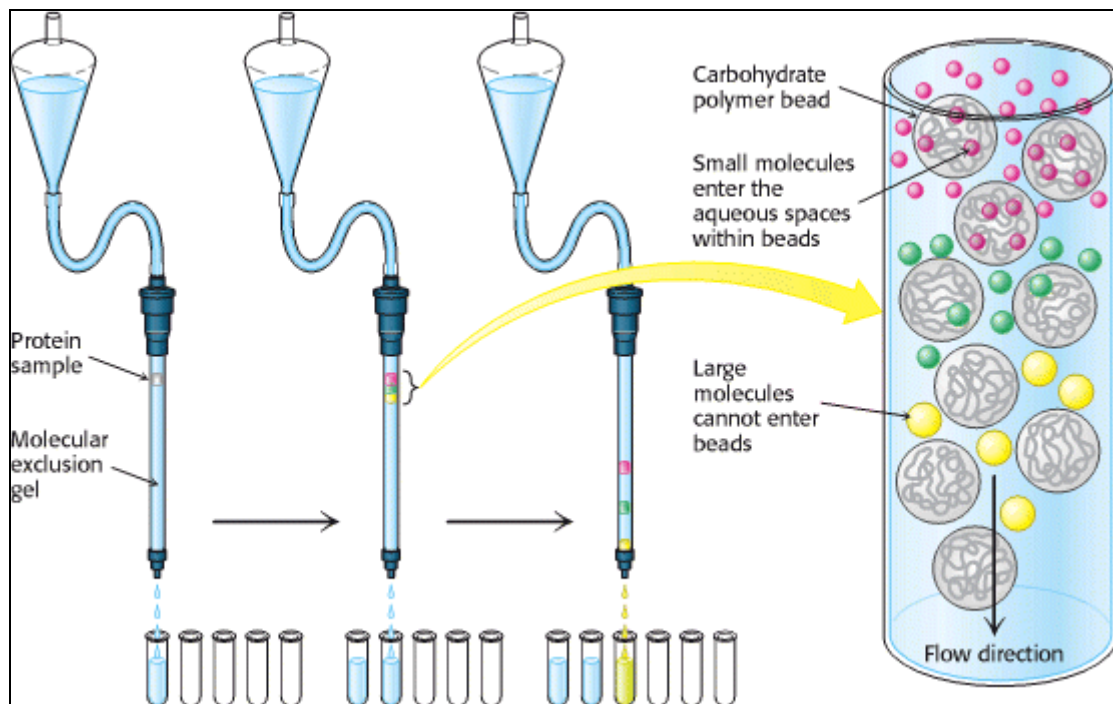


Figure 4.3. Gel Filtration Chromatography. A mixture of proteins in a small volume is applied to a column filled with porous beads. Because large proteins cannot enter the internal volume of the beads, they emerge sooner than do small ones.

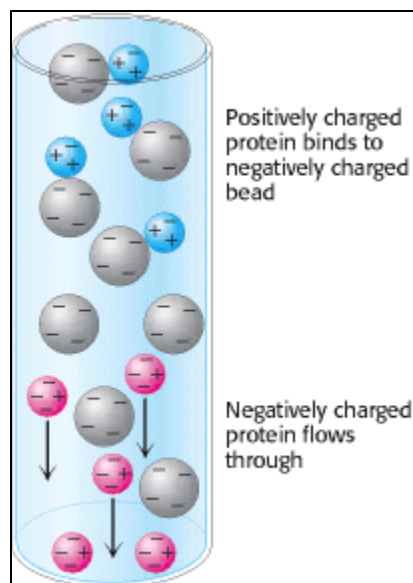


Figure 4.4. Ion-Exchange Chromatography. This technique separates proteins mainly according to their net charge.

Affinity Chromatography.

Affinity chromatography is another powerful and generally applicable means of purifying proteins. This technique takes advantage of the high affinity of many proteins for specific chemical groups. For example, the plant protein concanavalin A can be purified by passing a crude extract through a column of beads containing covalently attached glucose residues. Concanavalin A binds to such a column because it has affinity for glucose, whereas most other proteins do not. The bound concanavalin A can then be released from the column by adding a concentrated solution of glucose. The glucose in solution displaces the column-attached glucose residues from binding sites on concanavalin A (Figure 4.5). Affinity chromatography is a powerful means of isolating transcription factors, proteins that regulate gene expression by binding to specific DNA sequences. A protein mixture is percolated through a column

containing specific DNA sequences attached to a matrix; proteins with a high affinity for the sequence will bind and be retained. In this instance, the transcription factor is released by washing with a solution containing a high concentration of salt. In general, affinity chromatography can be effectively used to isolate a protein that recognizes group X by (1) covalently attaching X or a derivative of it to a column, (2) adding a mixture of proteins to this column, which is then washed with buffer to remove unbound proteins, and (3) eluting the desired protein by adding a high concentration of a soluble form of X or altering the conditions to decrease binding affinity. Affinity chromatography is most effective when the interaction of the protein and the molecule that is used as the bait is highly specific.

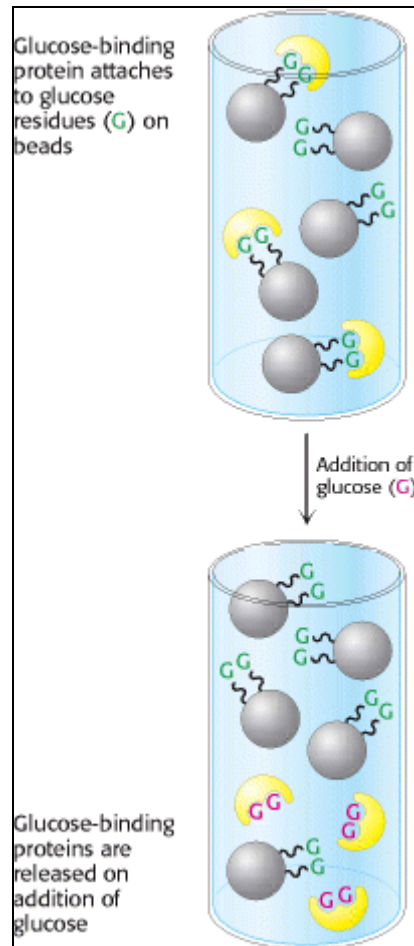


Figure 4.5. Affinity Chromatography. Affinity chromatography of concanavalin A (shown in yellow) on a solid support containing covalently attached glucose residues (G).

High-Pressure Liquid Chromatography.

The resolving power of all of the column techniques can be improved substantially through the use of a technique called *high-pressure liquid chromatography (HPLC)*, which is an enhanced version of the column techniques already discussed. The column materials themselves are much more finely divided and, as a consequence, there are more interaction sites and thus greater resolving power. Because the column is made of finer material, pressure must be applied to the column to obtain adequate flow rates. The net result is high resolution as well as rapid separation (Figure 4.6).

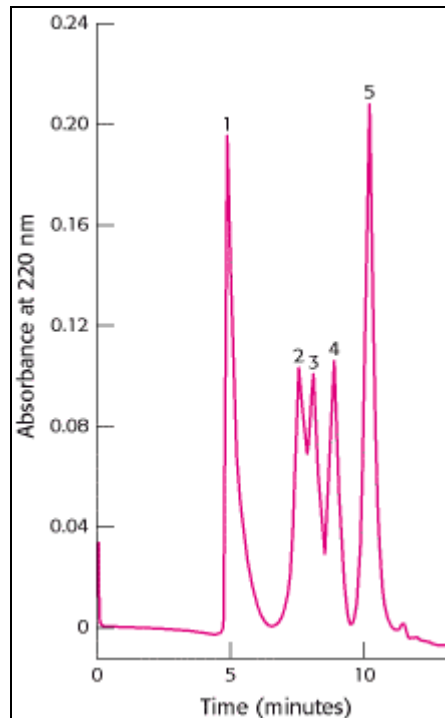


Figure 4.6. High-Pressure Liquid Chromatography (HPLC). Gel filtration by HPLC clearly defines the individual proteins because of its greater resolving power: (1) thyroglobulin (669 kd), (2) catalase (232 kd), (3) bovine serum albumin (67 kd), (4) ovalbumin (43 kd), and (5) ribonuclease (13.4 kd). [After K. J. Wilson and T. D. Schlabach. In *Current Protocols in Molecular Biology*, vol. 2, suppl. 41, F. M. Ausbel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, Eds. (Wiley, 1998), p. 10.14.1.]

4.1.4. Proteins Can Be Separated by Gel Electrophoresis and Displayed

How can we tell whether a purification scheme is effective? One way is to ascertain that the specific activity rises with each purification step. Another is to visualize the effectiveness by displaying the proteins present at each step. The technique of electrophoresis makes the latter method possible.

Gel Electrophoresis.

A molecule with a net charge will move in an electric field. This phenomenon, termed *electrophoresis*, offers a powerful means of separating proteins and other macromolecules, such as DNA and RNA. The velocity of migration (v) of a protein (or any molecule) in an electric field depends on the electric field strength (E), the net charge on the protein (z), and the frictional coefficient (f).

$$v = \frac{Ez}{f} \quad (1)$$

The electric force Ez driving the charged molecule toward the oppositely charged electrode is opposed by the viscous drag fv arising from friction between the moving molecule and the medium. The frictional coefficient f depends on both the mass and shape of the migrating molecule and the viscosity (η) of the medium. For a sphere of radius r ,

$$f = 6\pi\eta r \quad (2)$$

Electrophoretic separations are nearly always carried out in gels (or on solid supports such as paper) because the gel serves as a molecular sieve that enhances separation (Figure 4.7). Molecules that are small compared with the pores in the gel readily move through the gel, whereas molecules much larger than the pores are almost immobile. Intermediate-size molecules move through the gel with various degrees of facility. Electrophoresis is performed in a thin, vertical slab of polyacrylamide. The direction of flow is from top to bottom. Polyacrylamide gels, formed by the polymerization of acrylamide and cross-linked by methylenebisacrylamide, are choice supporting media for electrophoresis because they

are chemically inert and are readily formed (Figure 4.8). Electrophoresis is the opposite of gel filtration in that all of the molecules, regardless of size, are forced to move through the same matrix. The gel behaves as one bead of a gel-filtration column.

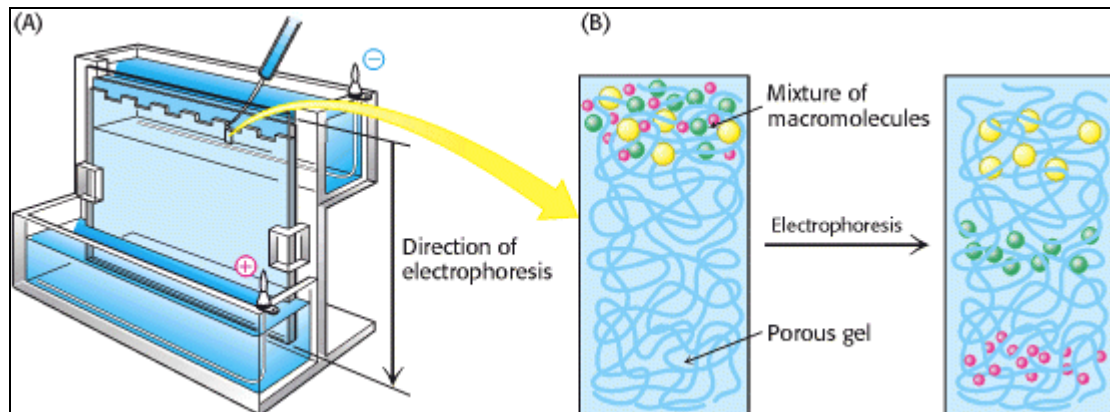


Figure 4.7. Polyacrylamide Gel Electrophoresis. (A) Gel electrophoresis apparatus. Typically, several samples undergo electrophoresis on one flat polyacrylamide gel. A microliter pipette is used to place solutions of proteins in the wells of the slab. A cover is then placed over the gel chamber and voltage is applied. The negatively charged SDS (sodium dodecyl sulfate)-protein complexes migrate in the direction of the anode, at the bottom of the gel. (B) The sieving action of a porous polyacrylamide gel separates proteins according to size, with the smallest moving most rapidly.

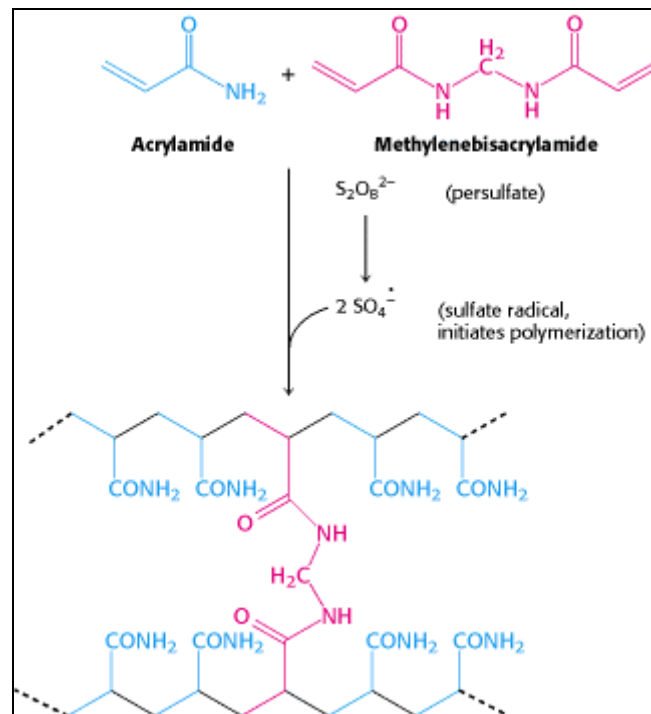


Figure 4.8. Formation of a Polyacrylamide Gel. A three-dimensional mesh is formed by co-polymerizing activated monomer (blue) and cross-linker (red).

Proteins can be separated largely on the basis of mass by electrophoresis in a polyacrylamide gel under denaturing conditions. The mixture of proteins is first dissolved in a solution of sodium dodecyl sulfate (SDS), an anionic detergent that disrupts nearly all noncovalent interactions in native proteins. Mercaptoethanol (2-thioethanol) or dithiothreitol also is added to reduce disulfide bonds. Anions of SDS bind to main chains at a ratio of about one SDS anion for every two amino acid residues. This complex of SDS with a denatured protein has a large net negative charge that is roughly proportional to the mass of the protein. The negative charge acquired on binding SDS is usually much greater than the charge on the native protein; this native charge is thus rendered insignificant. The SDS-protein complexes are then subjected to electrophoresis. When the electrophoresis is complete, the proteins in the gel can be visualized by staining them with silver or a dye such as Coomassie blue, which reveals a series of bands (Figure 4.9). Radioactive labels can be detected by placing a sheet of x-ray film over the gel, a procedure called *autoradiography*.

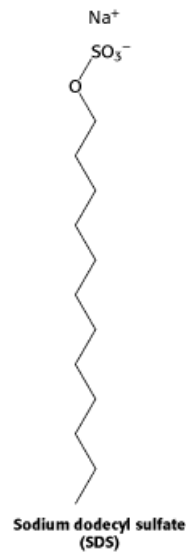


Figure 4.9. Staining of Proteins After Electrophoresis. Proteins subjected to electrophoresis on an SDS-polyacrylamide gel can be visualized by staining with Coomassie blue. [Courtesy of Kodak Scientific Imaging Systems.]

Small proteins move rapidly through the gel, whereas large proteins stay at the top, near the point of application of the mixture. The mobility of most polypeptide chains under these conditions is linearly proportional to the logarithm of their mass (Figure 4.10). Some carbohydrate-rich proteins and membrane proteins do not obey this empirical relation, however. SDS-polyacrylamide gel electrophoresis (SDS-PAGE) is rapid, sensitive, and capable of a high degree of resolution. As little as $0.1 \mu\text{g}$ ($\sim 2 \text{ pmol}$) of a protein gives a distinct band when stained with Coomassie blue, and even less ($\sim 0.02 \mu\text{g}$) can be detected with a silver stain. Proteins that differ in mass by about 2% (e.g., 40 and 41 kd, arising from a difference of about 10 residues) can usually be distinguished.

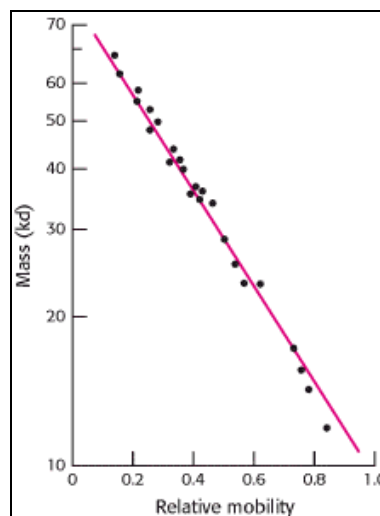


Figure 4.10. Electrophoresis Can Determine Mass. The electrophoretic mobility of many proteins in SDS-polyacrylamide gels is inversely proportional to the logarithm of their mass. [After K. Weber and M. Osborn, *The Proteins*, vol. 1, 3d ed. (Academic Press, 1975), p. 179.]

We can examine the efficacy of our purification scheme by analyzing a part of each fraction by SDS-PAGE. The initial fractions will display dozens to hundreds of proteins. As the purification progresses, the number of bands will diminish, and the prominence of one of the bands should increase. This band will correspond to the protein of interest.

Isoelectric Focusing.

Proteins can also be separated electrophoretically on the basis of their relative contents of acidic and basic residues. The *isoelectric point* (pI) of a protein is the pH at which its net charge is zero. At this pH, its electrophoretic mobility is zero because z in equation 1 is equal to zero. For example, the pI of cytochrome *c*, a highly basic electron-transport protein, is 10.6, whereas that of serum albumin, an acidic protein in blood, is 4.8. Suppose that a mixture of proteins undergoes electrophoresis in a pH gradient in a gel in the absence of SDS. Each protein will move until it reaches a position in the gel at which the pH is equal to the pI of the protein. This method of separating proteins according to their isoelectric point is called *isoelectric focusing*. The pH gradient in the gel is formed first by subjecting a mixture of *polyampholytes* (small multicharged polymers) having many pI values to electrophoresis. Isoelectric focusing can readily resolve proteins that differ in pI by as little as 0.01, which means that proteins differing by one net charge can be separated (Figure 4.11).

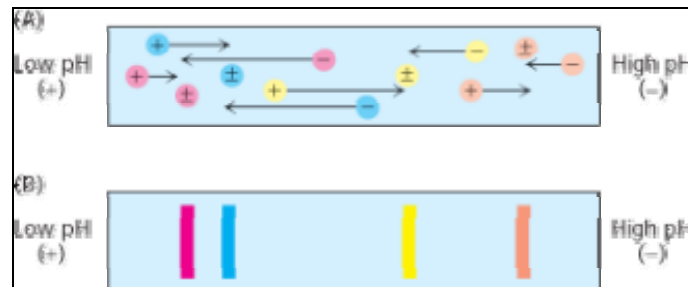


Figure 4.11. The Principle of Isoelectric Focusing. A pH gradient is established in a gel before loading the sample. (A) The sample is loaded and voltage is applied. The proteins will migrate to their isoelectric pH, the location at which they have no net charge. (B) The proteins form bands that can be excised and used for further experimentation.

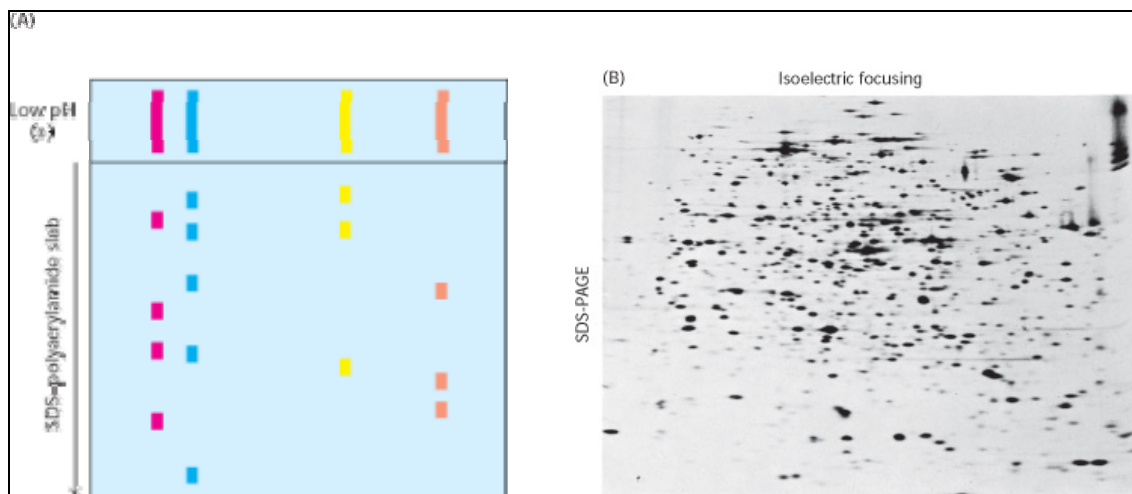


Figure 4.12. Two-Dimensional Gel Electrophoresis. (A) A protein sample is initially fractionated in one dimension by isoelectric focusing as described in Figure 4.11. The isoelectric focusing gel is then attached to an SDS-polyacrylamide gel, and electrophoresis is performed in the second dimension, perpendicular to the original separation. Proteins with the same pI are now separated on the basis of mass. (B) Proteins from *E. coli* were separated by two-dimensional gel electrophoresis, resolving more than a thousand different proteins. The proteins were first separated according to their isoelectric pH in the horizontal direction and then by their apparent mass in the vertical direction. [(B) Courtesy of Dr. Patrick H. O'Farrell.]

Two-Dimensional Electrophoresis.

Isoelectric focusing can be combined with SDS-PAGE to obtain very high resolution separations. A single sample is first subjected to isoelectric focusing. This single-lane gel is then placed horizontally on top of an SDS-polyacrylamide slab. The proteins are thus spread across the top of the polyacrylamide gel

according to how far they migrated during isoelectric focusing. They then undergo electrophoresis again in a perpendicular direction (vertically) to yield a twodimensional pattern of spots. In such a gel, proteins have been separated in the horizontal direction on the basis of isoelectric point and in the vertical direction on the basis of mass. It is remarkable that more than a thousand different proteins in the bacterium *Escherichia coli* can be resolved in a single experiment by two-dimensional electrophoresis (Figure 4.12).

Proteins isolated from cells under different physiological conditions can be subjected to two-dimensional electrophoresis, followed by an examination of the intensity of the signals. In this way, particular proteins can be seen to increase or decrease in concentration in response to the physiological state. How can we tell what protein is being regulated? A former drawback to the power of the two-dimensional gel is that, although many proteins are displayed, they are not identified. It is now possible to identify proteins by coupling two-dimensional gel electrophoresis with mass spectrometric techniques. We will consider these techniques when we examine how the mass of a protein is determined (Section 4.1.7).

4.1.5. A Protein Purification Scheme Can Be Quantitatively Evaluated

To determine the success of a protein purification scheme, we monitor the procedure at each step by determining specific activity and by performing an SDS-PAGE analysis. Consider the results for the purification of a fictitious protein, summarized in Table 4.1 and Figure 4.13. At each step, the following parameters are measured:

Total protein. The quantity of protein present in a fraction is obtained by determining the protein concentration of a part of each fraction and multiplying by the fraction's total volume.

Total activity. The enzyme activity for the fraction is obtained by measuring the enzyme activity in the volume of fraction used in the assay and multiplying by the fraction's total volume.

Specific activity. This parameter is obtained by dividing total activity by total protein.

Yield. This parameter is a measure of the activity retained after each purification step as a percentage of the activity in the crude extract. The amount of activity in the initial extract is taken to be 100%.

Purification level. This parameter is a measure of the increase in purity and is obtained by dividing the specific activity, calculated after each purification step, by the specific activity of the initial extract.

Step	Total protein (mg)	Total activity (units)	Specific activity, (units mg ⁻¹)	Yield (%)	Purification level
Homogenization	15,000	150,000	10	100	1
Salt fractionation	4,600	138,000	30	92	3
Ion-exchange chromatography	1,278	115,500	90	77	9
Molecular exclusion chromatography	68.8	75,000	1,100	50	110
Affinity chromatography	1.75	52,500	30,000	35	3,000

Table 4.1. Quantification of a purification protocol for a fictitious protein

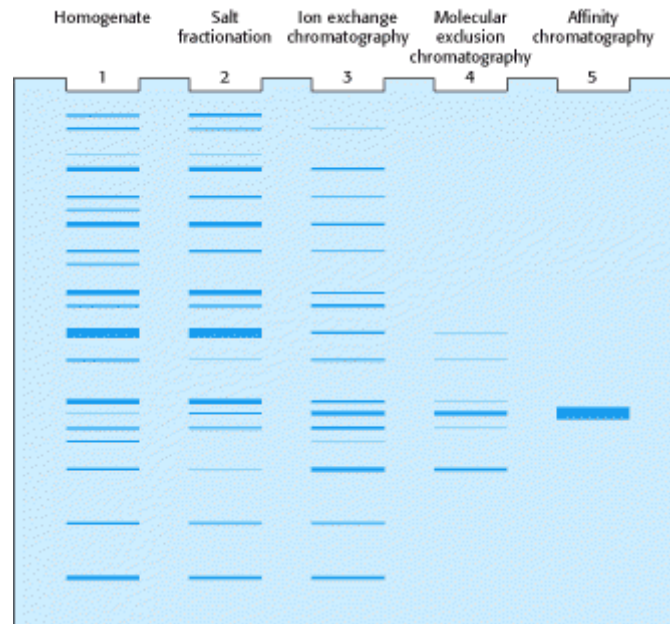


Figure 4.13. Electrophoretic Analysis of a Protein Purification. The purification scheme in Table 4.1 was analyzed by SDS-PAGE. Each lane contained 50 μg of sample. The effectiveness of the purification can be seen as the band for the protein of interest becomes more prominent relative to other bands.

As we see in [Table 4.1](#), the first purification step, salt fractionation, leads to an increase in purity of only 3-fold, but we recover nearly all the target protein in the original extract, given that the yield is 92%. After dialysis to lower the high concentration of salt remaining from the salt fractionation, the fraction is passed through an ion-exchange column. The purification now increases to 9-fold compared with the original extract, whereas the yield falls to 77%. Molecular exclusion chromatography brings the level of purification to 100-fold, but the yield is now at 50%. The final step is affinity chromatography with the use of a ligand specific for the target enzyme. This step, the most powerful of these purification procedures, results in a purification level of 3000-fold, while lowering the yield to 35%. The SDS-PAGE in [Figure 4.13](#) shows that, if we load a constant amount of protein onto each lane after each step, the number of bands decreases in proportion to the level of purification, and the amount of protein of interest increases as a proportion of the total protein present.

A good purification scheme takes into account both purification levels and yield. A high degree of purification and a poor yield leave little protein with which to experiment. A high yield with low purification leaves many contaminants (proteins other than the one of interest) in the fraction and complicates the interpretation of experiments.

4.1.6. Ultracentrifugation Is Valuable for Separating Biomolecules and Determining Their Masses

We have already seen that centrifugation is a powerful and generally applicable method for separating a crude mixture of cell components, but it is also useful for separating and analyzing biomolecules themselves. With this technique, we can determine such parameters as mass and density, learn something about the shape of a molecule, and investigate the interactions between molecules. To deduce these properties from the centrifugation data, we need a mathematical description of how a particle behaves in a centrifugal force.

A particle will move through a liquid medium when subjected to a centrifugal force. A convenient means of quantifying the rate of movement is to calculate the sedimentation coefficient, s , of a particle by using the following equation:

$$s = m(1 - \bar{v}\rho)/f$$

where m is the mass of the particle, \bar{v} is the partial specific volume (the reciprocal of the particle density), ρ is the density of the medium and f is the frictional coefficient (a measure of the shape of the particle). The $(1 - \rho)$ term is the buoyant force exerted by liquid medium.

Sedimentation coefficients are usually expressed in *Svedberg units* (*S*), equal to 10^{-13} s. The smaller the *S* value, the slower a molecule moves in a centrifugal field. The *S* values for a number of biomolecules and cellular components are listed in [Table 4.2](#) and [Figure 4.14](#).

Protein	S value (Svedberg units)	Molecular weight
Pancreatic trypsin inhibitor	1	6,520
Cytochrome <i>c</i>	1.83	12,310
Ribonuclease A	1.78	13,690
Myoglobin	1.97	17,800
Trypsin	2.5	23,200
Carbonic anhydrase	3.23	28,800
Concanavlin A	3.8	51,260
Malate dehydrogenase	5.76	74,900
Lactate dehydrogenase	7.54	146,200

From T. Creighton, *Proteins*, 2nd Edition (W. H. Freeman and Company, 1993), Table 7.1.

Table 4.2. S values and molecular weights of sample proteins

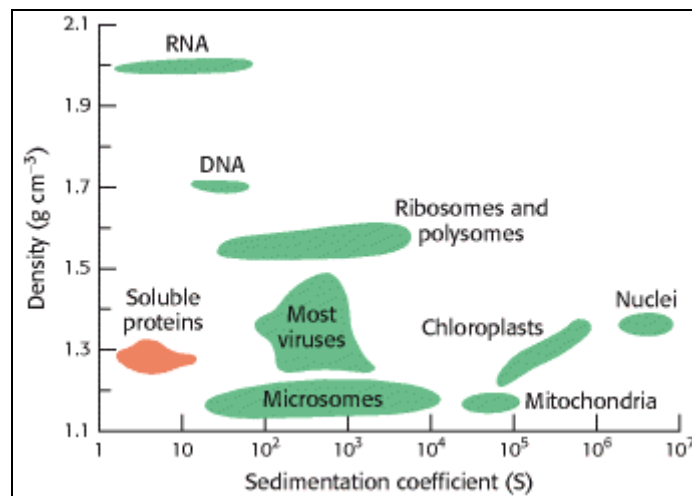


Figure 4.14. Density and Sedimentation Coefficients of Cellular Components. [After L. J. Kleinsmith and V. M. Kish, *Principles of Cell and Molecular Biology*, 2d ed. (Harper Collins, 1995), p. 138.]

Several important conclusions can be drawn from the preceding equation:

1. The sedimentation velocity of a particle depends in part on its mass. A more massive particle sediments more rapidly than does a less massive particle of the same shape and density.
2. Shape, too, influences the sedimentation velocity because it affects the viscous drag. The frictional coefficient f of a compact particle is smaller than that of an extended particle of the same mass. Hence, elongated particles sediment more slowly than do spherical ones of the same mass.
3. A dense particle moves more rapidly than does a less dense one because the opposing buoyant force ($1-\rho$) is smaller for the denser particle.
4. The sedimentation velocity also depends on the density of the solution. (ρ). Particles sink when $\rho < 1$, float when $\rho > 1$, and do not move when $\rho = 1$.

A technique called *zonal*, *band*, or most commonly *gradient* centrifugation can be used to separate proteins with different sedimentation coefficients. The first step is to form a density gradient in a centrifuge tube. Differing proportions of a low-density solution (such as 5% sucrose) and a high-density solution (such as 20% sucrose) are mixed to create a linear gradient of sucrose concentration ranging from 20% at the bottom of the tube to 5% at the top (Figure 4.15). The role of the gradient is to prevent convective flow. A small volume of a solution containing the mixture of proteins to be separated is placed on top of the density gradient. When the rotor is spun, proteins move through the gradient and separate according to their sedimentation coefficients. The time and speed of the centrifugation is determined empirically. The separated bands, or zones, of protein can be harvested by making a hole in the bottom of the tube and collecting drops. The drops can be measured for protein content and catalytic activity or another functional property. This sedimentation-velocity technique readily separates proteins differing in sedimentation coefficient by a factor of two or more.

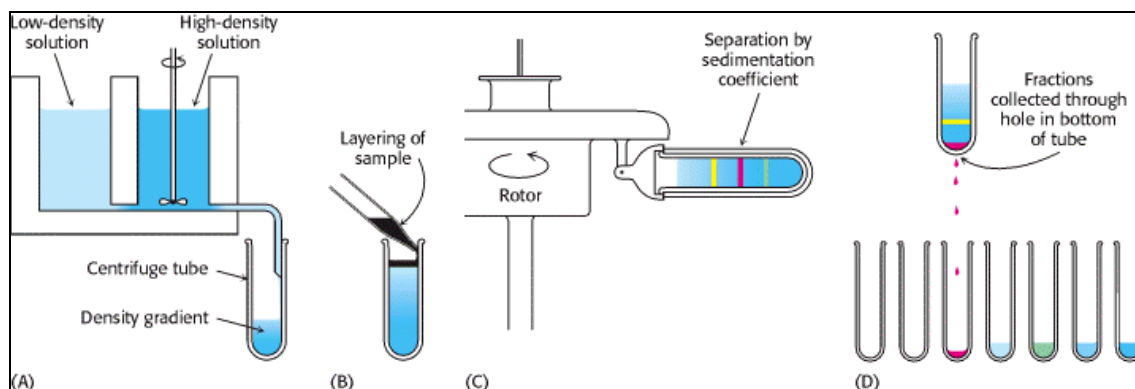


Figure 4.15. Zonal Centrifugation. The steps are as follows: (A) form a density gradient, (B) layer the sample on top of the gradient, (C) place the tube in a swinging-bucket rotor and centrifuge it, and (D) collect the samples. [After D. Freifelder, *Physical Biochemistry*, 2d ed. (W. H. Freeman and Company, 1982), p. 397.]

The mass of a protein can be directly determined by *sedimentation equilibrium*, in which a sample is centrifuged at relatively low speed so that sedimentation is counterbalanced by diffusion. *The sedimentation-equilibrium technique for determining mass is very accurate and can be applied under nondenaturing conditions in which the native quaternary structure of multimeric proteins is preserved.* In contrast, SDS-polyacrylamide gel electrophoresis (Section 4.1.4) provides an *estimate* of the mass of dissociated polypeptide chains under *denaturing* conditions. Note that, if we know the mass of the dissociated components of a multimeric protein as determined by SDS-polyacrylamide analysis and the mass of the intact multimeric protein as determined by sedimentation equilibrium analysis, we can determine how many copies of each polypeptide chain is present in the multimeric protein.

4.1.7. The Mass of a Protein Can Be Precisely Determined by Mass Spectrometry

Mass spectrometry has been an established analytical technique in organic chemistry for many years. Until recently, however, the very low volatility of proteins made mass spectrometry useless for the investigation of these molecules. This difficulty has been circumvented by the introduction of techniques for effectively dispersing proteins and other macromolecules into the gas phase. These methods are called *matrix-assisted laser desorption-ionization (MALDI)* and *electrospray spectrometry*. We will focus on MALDI spectrometry. In this technique, protein ions are generated and then accelerated through an electrical field (Figure 4.16). They travel through the flight tube, with the smallest traveling fastest and arriving at the detector first. Thus, the *time of flight (TOF)* in the electrical field is a measure of the mass (or, more precisely, the mass/charge ratio). Tiny amounts of biomolecules, as small as a few picomoles (pmol) to femtomoles (fmol), can be analyzed in this manner. A MALDI-TOF mass spectrum for a mixture of the proteins insulin and β -lactoglobulin is shown in Figure 4.17. The masses determined by MALDI-TOF are 5733.9 and 18,364, respectively, compared with calculated values of 5733.5 and 18,388. MALDI-TOF is indeed an accurate means of determining protein mass.

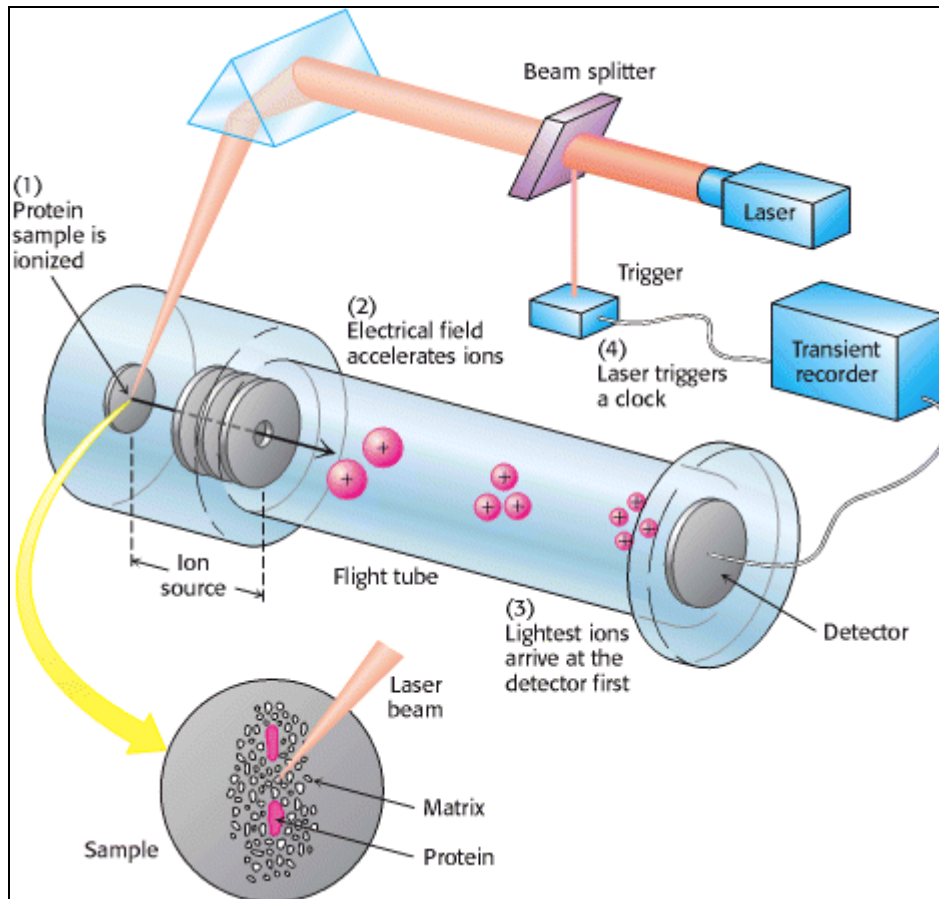


Figure 4.16. MALDI-TOF Mass Spectrometry. (1) The protein sample, embedded in an appropriate matrix, is ionized by the application of a laser beam. (2) An electrical field accelerates the ions formed through the flight tube toward the detector. (3) The lightest ions arrive first. (4) The ionizing laser pulse also triggers a clock that measures the time of flight (TOF) for the ions. [After J. T. Watson, *Introduction to Mass Spectrometry*, 3d ed. (Lippincott-Raven, 1997), p. 279.]

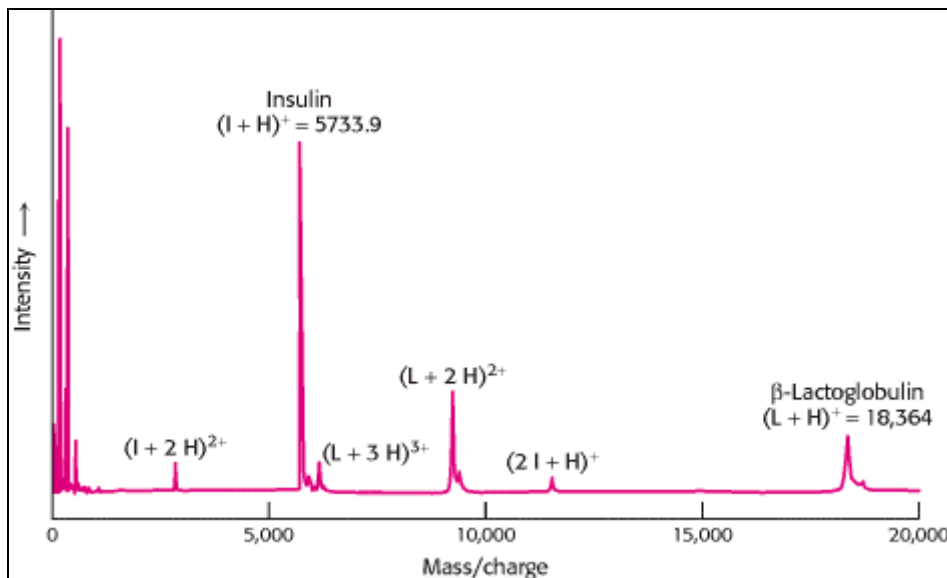


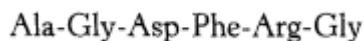
Figure 4.17. MALDI-TOF Mass Spectrum of Insulin and β -lactoglobulin. A mixture of 5 pmol each of insulin (I) and β -lactoglobulin (L) was ionized by MALDI, which produces predominately singly charged molecular ions from peptides and proteins ($I + H^+$ for insulin and $L + H^+$ for lactoglobulin). However, molecules with multiple charges as well as small quantities of a singly charged dimer of insulin, $(2I + H)^+$, also are produced. [After J. T. Watson, *Introduction to Mass Spectrometry*, 3d ed. (Lippincott-Raven, 1997), p. 282.]

Mass spectrometry has permitted the development of *peptide mass fingerprinting*. This technique for identifying peptides has greatly enhanced the utility of two-dimensional gels. Two-dimensional electrophoresis is performed as described in Section 4.1.4. The sample of interest is extracted and cleaved

specifically by chemical or enzymatic means. The masses of the protein fragments are then determined with the use of mass spectrometry. Finally, the peptide masses, or *fingerprint*, are matched against the fingerprint found in databases of proteins that have been "electronically cleaved" by a computer simulating the same fragmentation technique used for the experimental sample. This technique has provided some outstanding results. For example, of 150 yeast proteins analyzed with the use of two-dimensional gels, peptide mass fingerprinting unambiguously identified 80%. Mass spectrometry has provided name tags for many of the proteins in twodimensional gels.

4.2. Amino Acid Sequences Can Be Determined by Automated Edman Degradation

The protein of interest having been purified and its mass determined, the next analysis usually performed is to determine the protein's amino acid sequence, or primary structure. As stated previously (Section 3.2.1), a wealth of information about a protein's function and evolutionary history can often be obtained from the primary structure. Let us examine first how we can sequence a simple peptide, such as



The first step is to determine the *amino acid composition* of the peptide. The peptide is hydrolyzed into its constituent amino acids by heating it in 6 N HCl at 110°C for 24 hours. Amino acids in hydrolysates can be separated by ion-exchange chromatography on columns of sulfonated polystyrene. The identity of the amino acid is revealed by its elution volume, which is the volume of buffer used to remove the amino acid from the column (Figure 4.18), and quantified by reaction with *ninhydrin*. Amino acids treated with ninhydrin give an intense blue color, except for proline, which gives a yellow color because it contains a secondary amino group. The concentration of an amino acid in a solution, after heating with ninhydrin, is proportional to the optical absorbance of the solution. This technique can detect a microgram (10 nmol) of an amino acid, which is about the amount present in a thumbprint. As little as a nanogram (10 pmol) of an amino acid can be detected by replacing ninhydrin with *fluorescamine*, which reacts with the α -amino group to form a highly fluorescent product (Figure 4.19). A comparison of the chromatographic patterns of our sample hydrolysate with that of a standard mixture of amino acids would show that the amino acid composition of the peptide is

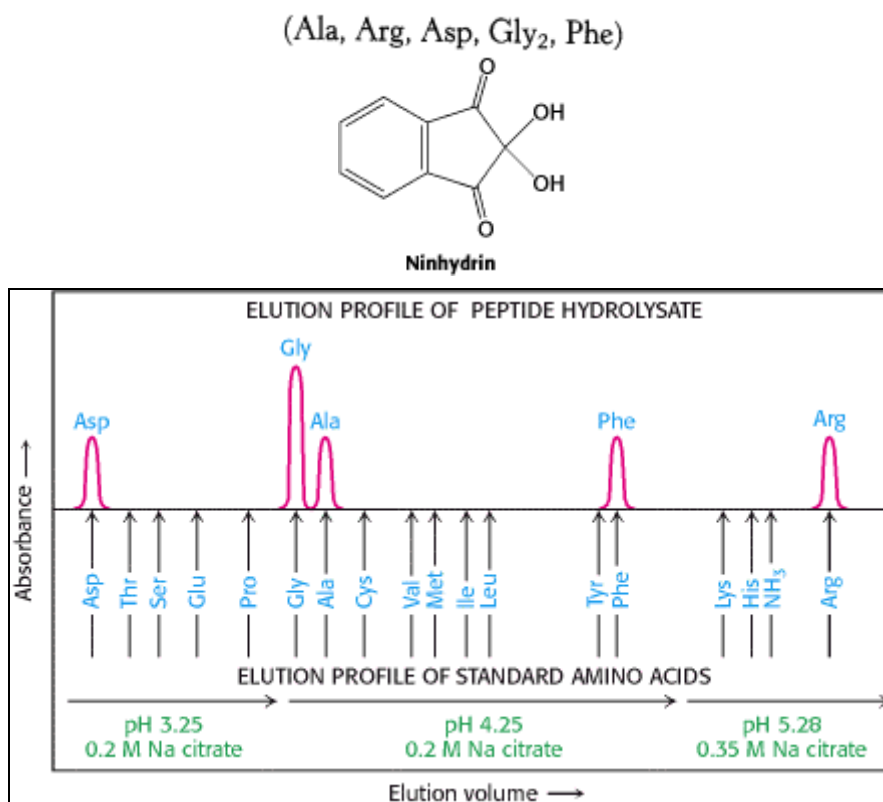


Figure 4.18. Determination of Amino Acid Composition. Different amino acids in a peptide hydrolysate can be separated by ion-exchange chromatography on a sulfonated polystyrene resin (such as Dowex-50). Buffers (in this case, sodium citrate) of increasing pH are used to elute the amino acids from the column. The amount of each amino acid present is determined from the absorbance. Aspartate, which has an acidic side chain, is first to emerge, whereas arginine, which has a basic side chain, is the last. The original peptide is revealed to be composed of one aspartate, one alanine, one phenylalanine, one arginine, and two glycine residues.

The parentheses denote that this is the amino acid composition of the peptide, not its sequence.

The next step is often to identify the N-terminal amino acid by labeling it with a compound that forms a stable covalent bond. *Fluorodinitrobenzene* (FDNB) was first used for this purpose by Frederick Sanger. *Dabsyl chloride* is now commonly used because it forms fluorescent derivatives that can be detected with high sensitivity. It reacts with an uncharged α -NH₂ group to form a sulfonamide derivative that is stable

under conditions that hydrolyze peptide bonds (Figure 4.20). Hydrolysis of our sample dabsyl-peptide in 6 N HCl would yield a dabsyl-amino acid, which could be identified as dabsyl-alanine by its chromatographic properties. *Dansyl chloride*, too, is a valuable labeling reagent because it forms fluorescent sulfonamides.

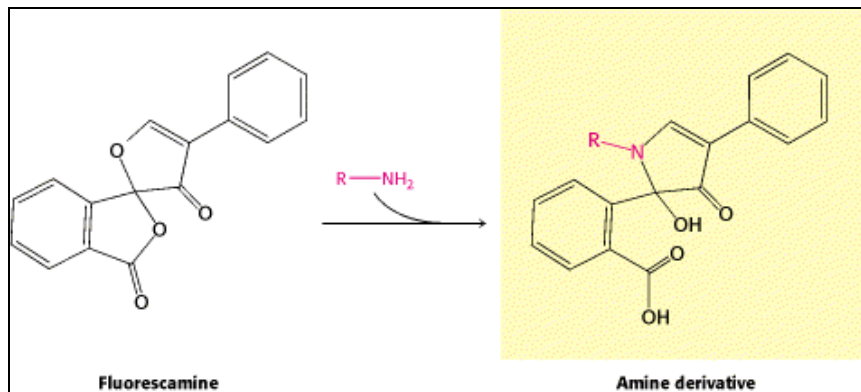


Figure 4.19. Fluorescent Derivatives of Amino Acids. Fluorescamine reacts with the α -amino group of an amino acid to form a fluorescent derivative

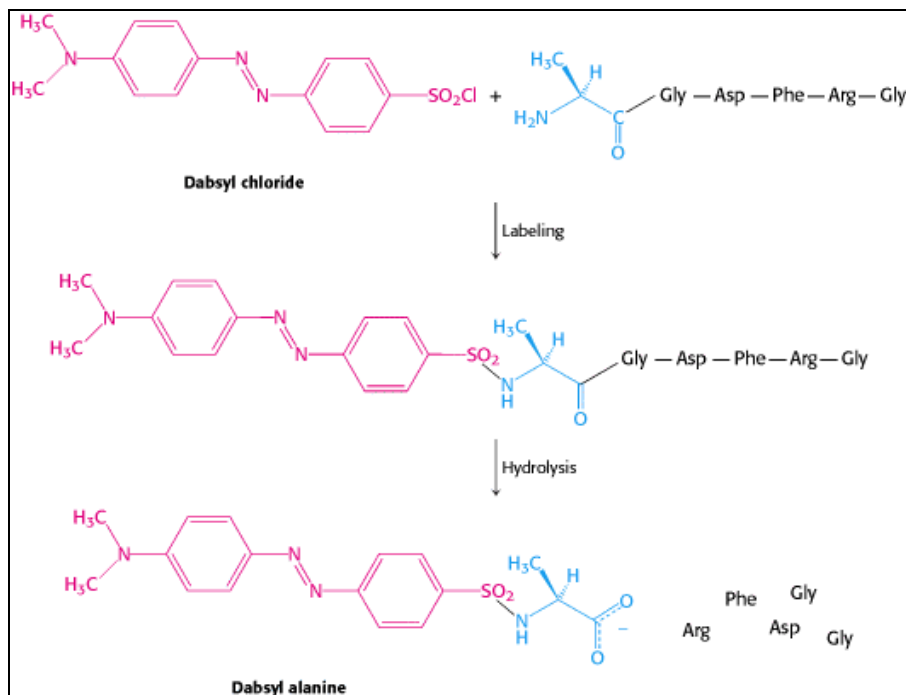
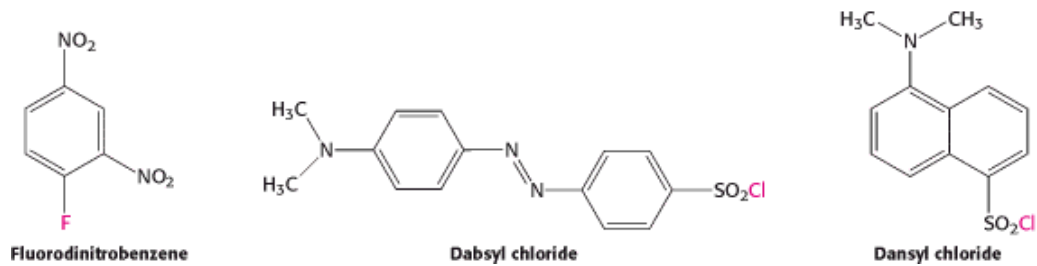


Figure 4.20. Determination of the Amino-Terminal Residue of a Peptide. Dabsyl chloride labels the peptide, which is then hydrolyzed with the use of hydrochloric acid. The dabsyl-amino acid (dabsyl-alanine in this example) is identified by its chromatographic characteristics.

Although the dabsyl method for determining the amino-terminal residue is sensitive and powerful, it cannot be used repeatedly on the same peptide, because the peptide is totally degraded in the acid-hydrolysis step and thus all sequence information is lost. Pehr Edman devised a method for labeling the amino-terminal residue and cleaving it from the peptide without disrupting the peptide bonds between the other amino acid residues. The *Edman degradation* sequentially removes one residue at a time from the

amino end of a peptide (Figure 4.21). *Phenyl isothiocyanate* reacts with the uncharged terminal amino group of the peptide to form a phenylthiocarbamoyl derivative. Then, under mildly acidic conditions, a cyclic derivative of the terminal amino acid is liberated, which leaves an intact peptide shortened by one amino acid. The cyclic compound is a phenylthiohydantoin (PTH)-amino acid, which can be identified by chromatographic procedures. The Edman procedure can then be repeated on the shortened peptide, yielding another PTH-amino acid, which can again be identified by chromatography. Three more rounds of the Edman degradation will reveal the complete sequence of the original peptide pentapeptide.

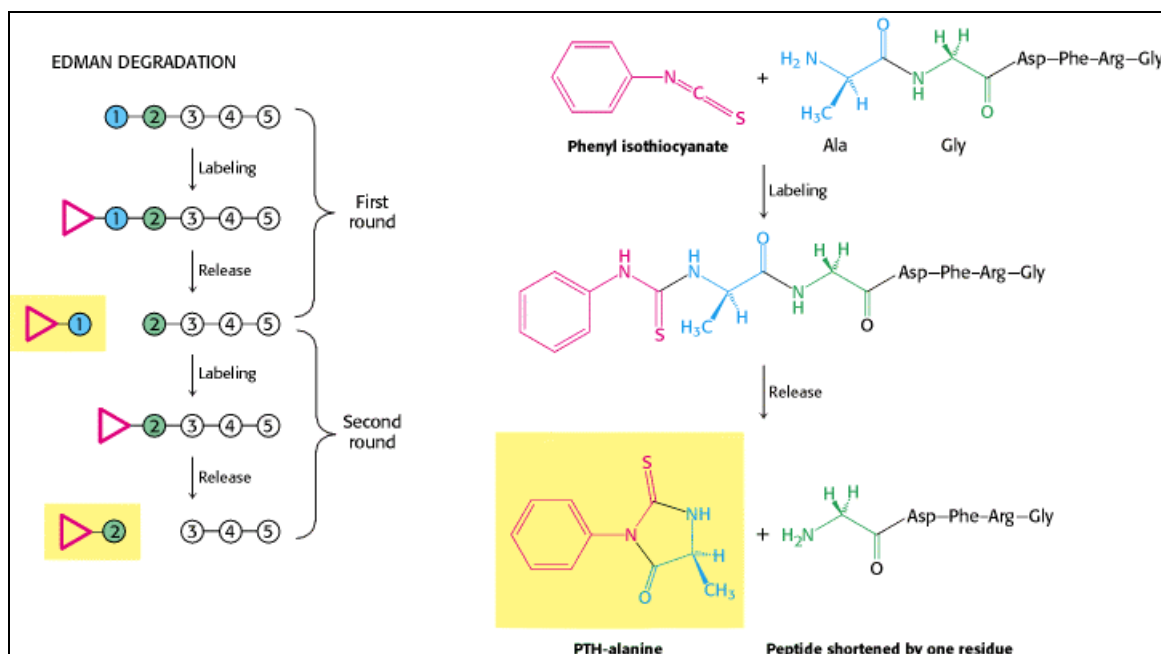


Figure 4.21. The Edman Degradation. The labeled amino-terminal residue (PTH-alanine in the first round) can be released without hydrolyzing the rest of the peptide. Hence, the amino-terminal residue of the shortened peptide (Gly-Asp-Phe-Arg-Gly) can be determined in the second round. Three more rounds of the Edman degradation reveal the complete sequence of the original peptide.

The development of automated sequencers has markedly decreased the time required to determine protein sequences. One cycle of the Edman degradation - the cleavage of an amino acid from a peptide and its identification - is carried out in less than 1 hour. By repeated degradations, the amino acid sequence of some 50 residues in a protein can be determined. High-pressure liquid chromatography provides a sensitive means of distinguishing the various amino acids (Figure 4.22). Gas-phase sequencers can analyze picomole quantities of peptides and proteins. This high sensitivity makes it feasible to analyze the sequence of a protein sample eluted from a single band of an SDS-polyacrylamide gel.

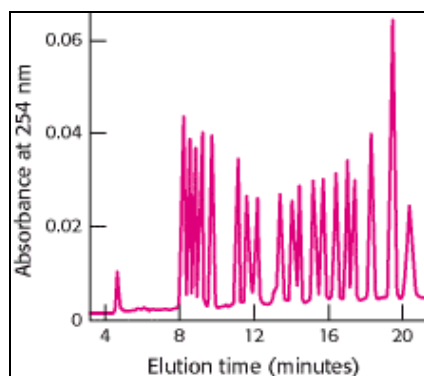


Figure 4.22. Separation of PTH-Amino Acids. PTH-amino acids can be rapidly separated by high-pressure liquid chromatography (HPLC). In this HPLC profile, a mixture of PTH-amino acids is clearly resolved into its components. An unknown amino acid can be identified by its elution position relative to the known ones.

4.2.1. Proteins Can Be Specifically Cleaved into Small Peptides to Facilitate Analysis

In principle, it should be possible to sequence an entire protein by using the Edman method. In practice, the peptides cannot be much longer than about 50 residues. This is so because the reactions of the Edman method, especially the release step, are not 100% efficient, and so not all peptides in the reaction mixture release the amino acid derivative at each step. For instance, if the efficiency of release for each round were 98%, the proportion of "correct" amino acid released after 60 rounds would be (0.98^{60}) , or 0.3 - a hopelessly impure mix. This obstacle can be circumvented by cleaving the original protein at specific amino acids into smaller peptides that can be sequenced. In essence, the strategy is to *divide and conquer*.

Specific cleavage can be achieved by chemical or enzymatic methods. For example, *cyanogen bromide* (CNBr) splits polypeptide chains only on the carboxyl side of methionine residues (Figure 4.23).

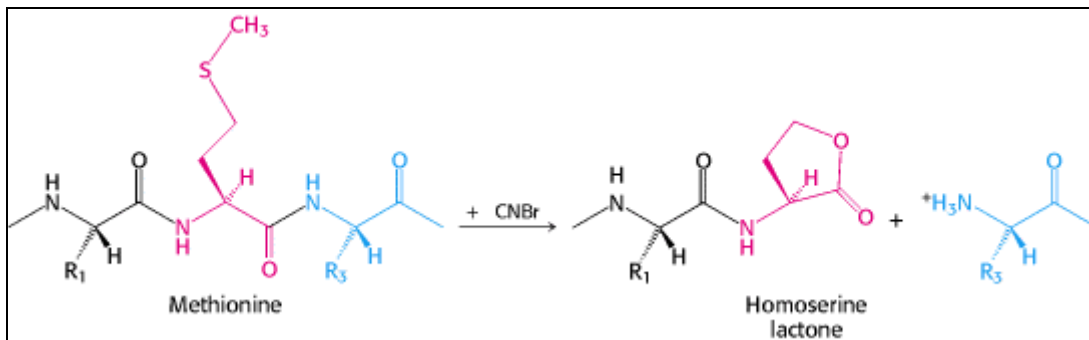


Figure 4.23. Cleavage by Cyanogen Bromide. Cyanogen bromide cleaves polypeptides on the carboxyl side of methionine residues.

A protein that has 10 methionine residues will usually yield 11 peptides on cleavage with CNBr. Highly specific cleavage is also obtained with *trypsin*, a proteolytic enzyme from pancreatic juice. Trypsin cleaves polypeptide chains on the carboxyl side of arginine and lysine residues (Figure 4.24 and Section 9.1.4). A protein that contains 9 lysine and 7 arginine residues will usually yield 17 peptides on digestion with trypsin. Each of these tryptic peptides, except for the carboxyl-terminal peptide of the protein, will end with either arginine or lysine. Table 4.3 gives several other ways of specifically cleaving polypeptide chains.

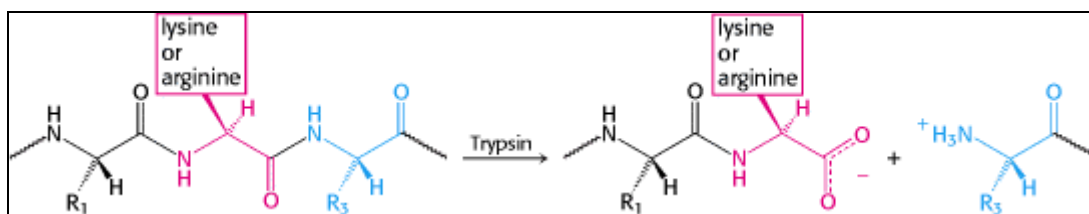


Figure 4.24. Cleavage by Trypsin. Trypsin hydrolyzes polypeptides on the carboxyl side of arginine and lysine residues.

The peptides obtained by specific chemical or enzymatic cleavage are separated by some type of chromatography. The sequence of each purified peptide is then determined by the Edman method. At this point, the amino acid sequences of segments of the protein are known, but the order of these segments is not yet defined. How can we order the peptides to obtain the primary structure of the original protein? The necessary additional information is obtained from *overlap peptides* (Figure 4.25). A second enzyme is used to split the polypeptide chain at different linkages. For example, chymotrypsin cleaves preferentially on the carboxyl side of aromatic and some other bulky nonpolar residues (Section 9.1.3). Because these chymotryptic peptides overlap two or more tryptic peptides, they can be used to establish the order of the peptides. The entire amino acid sequence of the polypeptide chain is then known.

Reagent	Cleavage site
Chemical cleavage	
Cyanogen bromide	Carboxyl side of methionine residues
<i>O</i> -Iodosobenzoate	Carboxyl side of tryptophan residues
Hydroxylamine	Asparagine-glycine bonds
2-Nitro-5-thiocyanobenzoate	Amino side of cysteine residues
Enzymatic cleavage	
Trypsin	Carboxyl side of lysine and arginine residues
Clostripain	Carboxyl side of arginine residues
Staphylococcal protease	Carboxyl side of aspartate and glutamate residues (glutamate only under certain conditions)
Thrombin	Carboxyl side of arginine
Chymotrypsin	Carboxyl side of tyrosine, tryptophan, phenylalanine, leucine, and methionine
Carboxypeptidase A	Amino side of C-terminal amino acid (not arginine, lysine, or proline)

Table 4.3. Specific cleavage of polypeptides

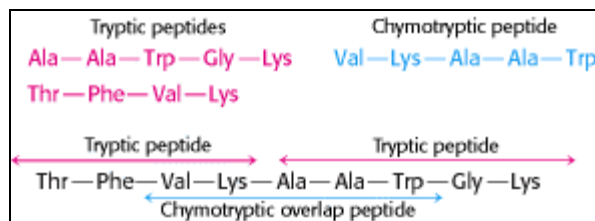


Figure 4.25. Overlap Peptides. The peptide obtained by chymotryptic digestion overlaps two tryptic peptides, establishing their order.

Additional steps are necessary if the initial protein sample is actually several polypeptide chains. SDS-gel electrophoresis under reducing conditions should display the number of chains. Alternatively, the number of distinct N-terminal amino acids could be determined. For a protein made up of two or more polypeptide chains held together by noncovalent bonds, denaturing agents, such as urea or guanidine hydrochloride, are used to dissociate the chains from one another. The dissociated chains must be separated from one another before sequence determination of the individual chains can begin. Polypeptide chains linked by disulfide bonds are separated by reduction with thiols such as β -mercaptoethanol or dithiothreitol. To prevent the cysteine residues from recombining, they are then alkylated with iodoacetate to form stable *S*-carboxymethyl derivatives (Figure 4.26). Sequencing can then be performed as heretofore described.

To complete our understanding of the protein's structure, we need to determine the positions of the original disulfide bonds. This information can be obtained by using a *diagonal electrophoresis* technique to isolate the peptide sequences containing such bonds (Figure 4.27). First, the protein is specifically cleaved into peptides under conditions in which the disulfides remain intact. The mixture of peptides is applied to a corner of a sheet of paper and subjected to electrophoresis in a single lane along one side. The resulting sheet is exposed to vapors of performic acid, which cleaves disulfides and converts them into cysteic acid residues. Peptides originally linked by disulfides are now independent and more acidic because of the formation of an SO_3^- group.

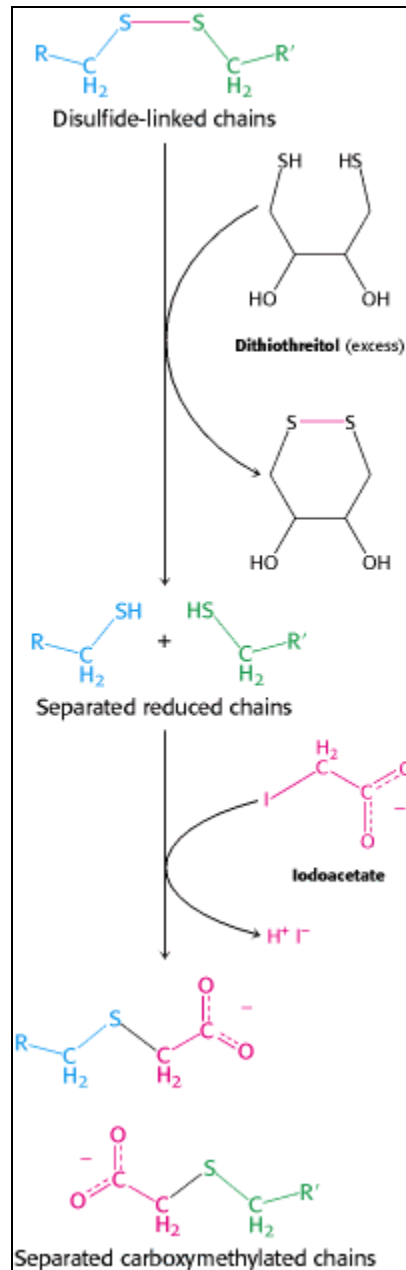


Figure 4.26. Disulfide-Bond Reduction. Polypeptides linked by disulfide bonds can be separated by reduction with dithiothreitol followed by alkylation to prevent reformation.

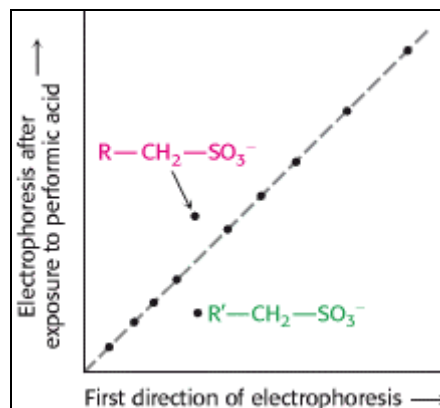
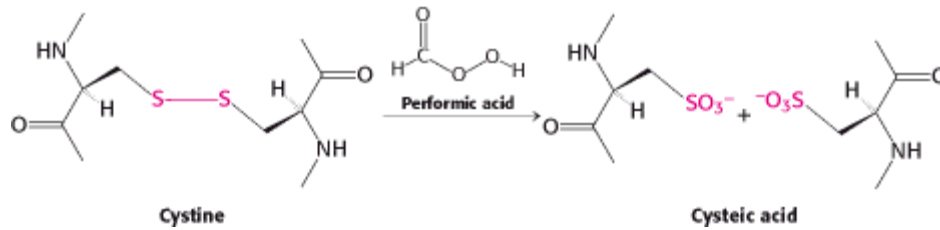


Figure 4.27. Diagonal Electrophoresis. Peptides joined together by disulfide bonds can be detected by diagonal electrophoresis. The mixture of peptides is subjected to electrophoresis in a single lane in one direction (horizontal) and then treated with performic acid, which cleaves and oxidizes the disulfide bonds. The sample is then subjected to electrophoresis in the perpendicular direction (vertical).



This mixture is subjected to electrophoresis in the perpendicular direction under the same conditions as those of the first electrophoresis. Peptides that were devoid of disulfides will have the same mobility as before, and consequently all will be located on a single diagonal line. In contrast, the newly formed peptides containing cysteic acid will usually migrate differently from their parent disulfide-linked peptides and hence will lie off the diagonal. These peptides can then be isolated and sequenced, and the location of the disulfide bond can be established.

4.2.2. Amino Acid Sequences Are Sources of Many Kinds of Insight

A protein's amino acid sequence, once determined, is a valuable source of insight into the protein's function, structure, and history.

1. The sequence of a protein of interest can be compared with all other known sequences to ascertain whether significant similarities exist. Does this protein belong to one of the established families? A search for kinship between a newly sequenced protein and the thousands of previously sequenced ones takes only a few seconds on a personal computer (Section 7.2). If the newly isolated protein is a member of one of the established classes of protein, we can begin to infer information about the protein's function. For instance, chymotrypsin and trypsin are members of the serine protease family, a clan of proteolytic enzymes that have a common catalytic mechanism based on a reactive serine residue (Section 9.1.4). If the sequence of the newly isolated protein shows sequence similarity with trypsin or chymotrypsin, the result suggests that it may be a serine protease.

2. Comparison of sequences of the same protein in different species yields a wealth of information about evolutionary pathways. Genealogical relations between species can be inferred from sequence differences between their proteins. We can even estimate the time at which two evolutionary lines diverged, thanks to the clocklike nature of random mutations. For example, a comparison of serum albumins found in primates indicates that human beings and African apes diverged 5 million years ago, not 30 million years ago as was once thought. Sequence analyses have opened a new perspective on the fossil record and the pathway of human evolution.

3. Amino acid sequences can be searched for the presence of internal repeats. Such internal repeats can reveal information about the history of an individual protein itself. Many proteins apparently have arisen by duplication of a primordial gene followed by its diversification. For example, calmodulin, a ubiquitous calcium sensor in eukaryotes, contains four similar calcium-binding modules that arose by gene duplication (Figure 4.28).

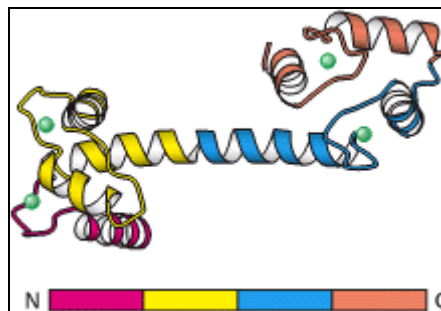


Figure 4.28. Repeating Motifs in a Protein Chain. Calmodulin, a calcium sensor, contains four similar units in a single polypeptide chain shown in red, yellow, blue, and orange. Each unit binds a calcium ion (shown in green).

4. Many proteins contain amino acid sequences that serve as signals designating their destinations or controlling their processing. A protein destined for export from a cell or for location in a membrane, for example, contains a signal sequence, a stretch of about 20 hydrophobic residues near the amino terminus that directs the protein to the appropriate membrane. Another protein may contain a stretch of amino acids that functions as a nuclear localization signal, directing the protein to the nucleus.

5. Sequence data provide a basis for preparing antibodies specific for a protein of interest. Careful examination of the amino acid sequence of a protein can reveal which sequences will be most likely to elicit an antibody when injected into a mouse or rabbit. Peptides with these sequences can be synthesized and used to generate antibodies to the protein. These specific antibodies can be very useful in determining the amount of a protein present in solution or in the blood, ascertaining its distribution within a cell, or cloning its gene (Section 4.3.3).

6. Amino acid sequences are valuable for making DNA probes that are specific for the genes encoding the corresponding proteins (Section 6.1.4). Knowledge of a protein's primary structure permits the use of reverse genetics. DNA probes that correspond to a part of the amino acid sequence can be constructed on the basis of the genetic code. These probes can be used to isolate the gene of the protein so that the entire sequence of the protein can be determined. The gene in turn can provide valuable information about the physiological regulation of the protein. Protein sequencing is an integral part of molecular genetics, just as DNA cloning is central to the analysis of protein structure and function.

4.2.3. Recombinant DNA Technology Has Revolutionized Protein Sequencing

Hundreds of proteins have been sequenced by Edman degradation of peptides derived from specific cleavages. Nevertheless, heroic effort is required to elucidate the sequence of large proteins, those with more than 1,000 residues. For sequencing such proteins, a complementary experimental approach based on recombinant DNA technology is often more efficient. As will be discussed in Chapter 6, long stretches of DNA can be cloned and sequenced, and the nucleotide sequence directly reveals the amino acid sequence of the protein encoded by the gene (Figure 4.29). Recombinant DNA technology is producing a wealth of amino acid sequence information at a remarkable rate.

DNA sequence	GGG	TTC	TTG	GGA	GCA	GCA	GGA	AGC	ACT	ATG	GGC	GCA
Amino acid sequence	Gly	Phe	Leu	Gly	Ala	Ala	Gly	Ser	Thr	Met	Gly	Ala

Figure 4.29. DNA Sequence Yields the Amino Acid Sequence. The complete nucleotide sequence of HIV-1 (human immunodeficiency virus), the cause of AIDS (acquired immune deficiency syndrome), was determined within a year after the isolation of the virus. A part of the DNA sequence specified by the RNA genome of the virus is shown here with the corresponding amino acid sequence (deduced from a knowledge of the genetic code).

Even with the use of the DNA base sequence to determine primary structure, there is still a need to work with isolated proteins. The amino acid sequence deduced by reading the DNA sequence is that of the *nascent* protein, the direct product of the translational machinery. Many proteins are modified after synthesis. Some have their ends trimmed, and others arise by cleavage of a larger initial polypeptide chain. Cysteine residues in some proteins are oxidized to form disulfide links, connecting either parts within a chain or separate polypeptide chains. Specific side chains of some proteins are altered. Amino acid sequences derived from DNA sequences are rich in information, but they do not disclose such posttranslational modifications. Chemical analyses of proteins in their final form are needed to delineate the nature of these changes, which are critical for the biological activities of most proteins. *Thus, genomic and proteomic analyses are complementary approaches to elucidating the structural basis of protein function.*

4.3. Immunology Provides Important Techniques with Which to Investigate Proteins

Immunological methods have become important tools used to purify a protein, locate it in the cell, or quantify how much of the protein is present. These methods are predicated on the exquisite specificity of antibodies for their target proteins. Labeled antibodies provide a means to tag a specific protein so that it can be isolated, quantified, or visualized.

4.3.1. Antibodies to Specific Proteins Can Be Generated

Immunological techniques begin with the generation of antibodies to a particular protein. An *antibody* (also called an *immunoglobulin*, Ig) is a protein synthesized by an animal in response to the presence of a foreign substance, called an *antigen*, and normally functions to protect the animal from infection (Chapter 33). Antibodies have specific and high affinity for the antigens that elicited their synthesis. Proteins, polysaccharides, and nucleic acids can be effective antigens. An antibody recognizes a specific group or cluster of amino acids on a large molecule called an *antigenic determinant*, or *epitope* (Figures 4.30 and 4.31). Small foreign molecules, such as synthetic peptides, also can elicit antibodies, provided that the small molecule contains a recognized epitope and is attached to a macromolecular carrier. The small foreign molecule itself is called a *hapten*. Animals have a very large repertoire of antibody-producing cells, each producing an antibody of a single specificity. An antigen acts by stimulating the proliferation of the small number of cells that were already forming an antibody capable of recognizing the antigen (Chapter 33).

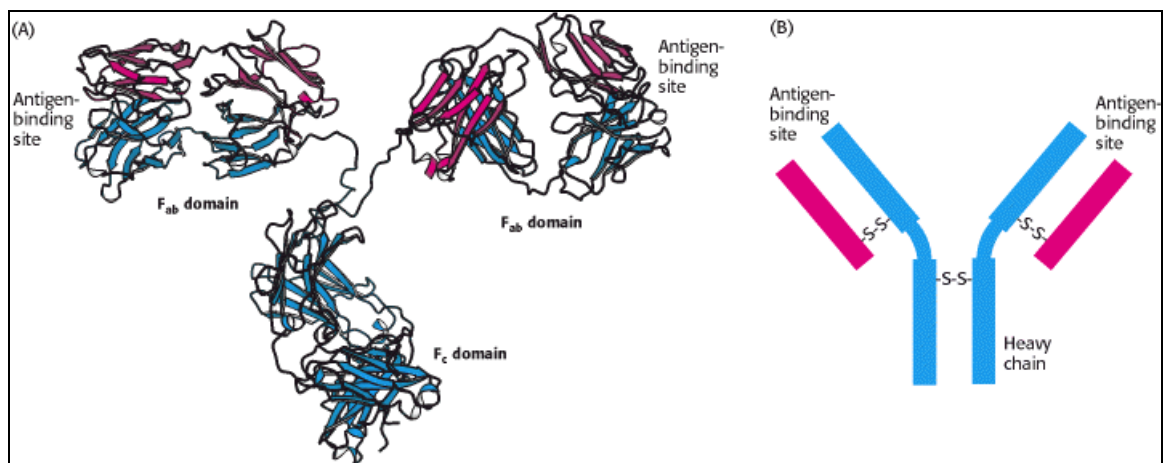


Figure 4.30. Antibody Structure. (A) IgG antibodies consist of four chains, two heavy chains (blue) and two light chains (red), linked by disulfide bonds. The heavy and light chains come together to form Fab domains, which have the antigen-binding sites at the ends. The two heavy chains form the Fc domain. The Fab domains are linked to the Fc domain by flexible linkers. (B) A more schematic representation of an IgG molecule.

Immunological techniques depend on our being able to generate antibodies to a specific antigen. To obtain antibodies that recognize a particular protein, a biochemist injects the protein into a rabbit twice, 3 weeks apart. The injected protein stimulates the reproduction of cells producing antibodies that recognize the foreign substance. Blood is drawn from the immunized rabbit several weeks later and centrifuged to separate blood cells from the supernatant, or serum. The serum, called an *antiserum*, contains antibodies to all antigens to which the rabbit has been exposed. Only some of them will be antibodies to the injected protein. Moreover, antibodies of a given specificity are not a single molecular species. For instance, 2,4-dinitrophenol (DNP) has been used as a hapten to generate antibodies to DNP. Analyses of anti-DNP antibodies revealed a wide range of binding affinities - the dissociation constants ranged from about 0.1 nM to 1 μ M. Correspondingly, a large number of bands were evident when anti-DNP antibody was subjected to isoelectric focusing. These results indicate that cells are producing many different antibodies, each recognizing a different surface feature of the same antigen. The antibodies are heterogeneous, or *polyclonal* (Figure 4.32). This heterogeneity is a barrier, which can complicate the use of these antibodies.

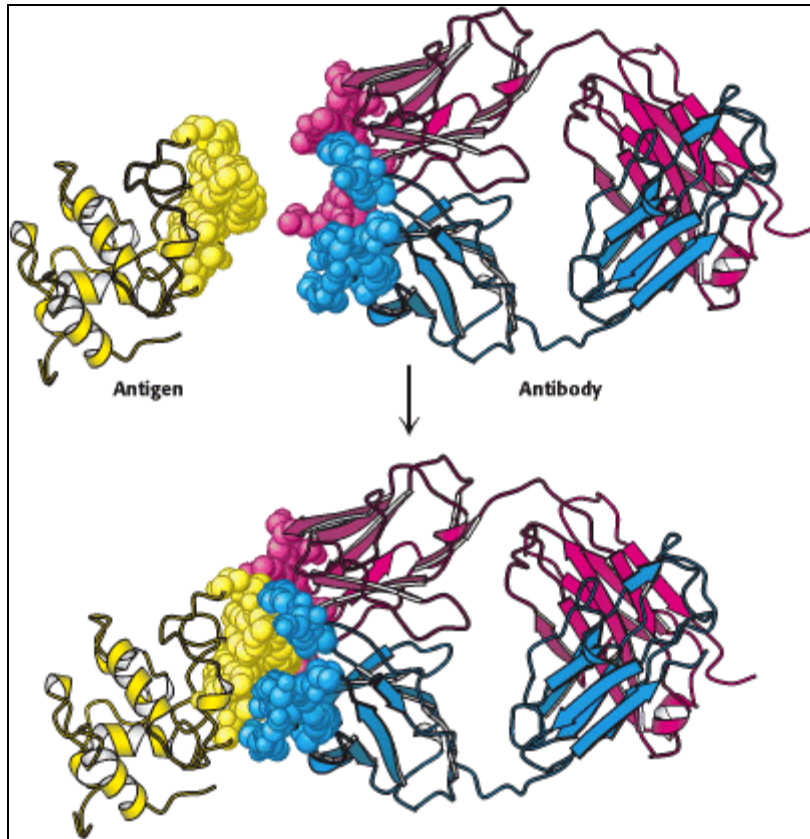


Figure 4.31. Antigen-Antibody Interactions. A protein antigen, in this case lysozyme, binds to the end of an Fab domain from an antibody. The end of the antibody and the antigen have complementary shapes, allowing a large amount of surface to be buried on binding.

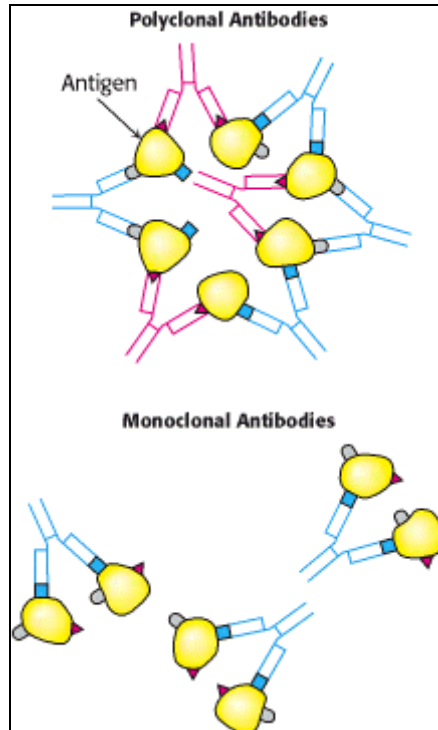


Figure 4.32. Polyclonal and Monoclonal Antibodies. Most antigens have several epitopes. Polyclonal antibodies are heterogeneous mixtures of antibodies, each specific for one of the various epitopes on an antigen. Monoclonal antibodies are all identical, produced by clones of a single antibody-producing cell. They recognize one specific epitope. [After R. A. Goldsby, T. J. Kindt, B. A. Osborne, *Kuby Immunology*, 4th ed. (W. H. Freeman and Company, 2000), p. 154.]

4.3.2. Monoclonal Antibodies with Virtually Any Desired Specificity Can Be Readily Prepared

The discovery of a means of producing *monoclonal antibodies* of virtually any desired specificity was a major breakthrough that intensified the power of immunological approaches. Just as working with impure proteins makes it difficult to interpret data and understand function, so too does working with an impure mixture of antibodies. The ideal would be to isolate a clone of cells that produce only a single antibody. The problem is that antibody-producing cells isolated from an organism die in a short time.

Immortal cell lines that produce monoclonal antibodies do exist. These cell lines are derived from a type of cancer, *multiple myeloma*, a malignant disorder of antibody-producing cells. In this cancer, a single transformed plasma cell divides uncontrollably, generating a very large number of *cells of a single kind*. They are a *clone* because they are descended from the same cell and have identical properties. The identical cells of the myeloma secrete large amounts of normal *immunoglobulin of a single kind* generation after generation. A myeloma can be transplanted from one mouse to another, where it continues to proliferate. These antibodies were useful for elucidating antibody structure, but nothing is known about their specificity and so they are useless for the immunological methods described in the next pages.

Cesar Milstein and Georges Köhler discovered that *large amounts of homogeneous antibody of nearly any desired specificity could be obtained by fusing a short-lived antibody-producing cell with an immortal myeloma cell*. An antigen is injected into a mouse, and its spleen is removed several weeks later (Figure 4.33). A mixture of plasma cells from this spleen is fused *in vitro* with myeloma cells. Each of the resulting hybrid cells, called *hybridoma cells*, indefinitely produces homogeneous antibody specified by the parent cell from the spleen. Hybridoma cells can then be screened, by using some sort of assay for the antigen-antibody interaction, to determine which ones produce antibody having the desired specificity. Collections of cells shown to produce the desired antibody are subdivided and reassayed. This process is repeated until a pure cell line, a clone producing a single antibody, is isolated. These positive cells can be grown in culture medium or injected into mice to induce myelomas. Alternatively, the cells can be frozen and stored for long periods.

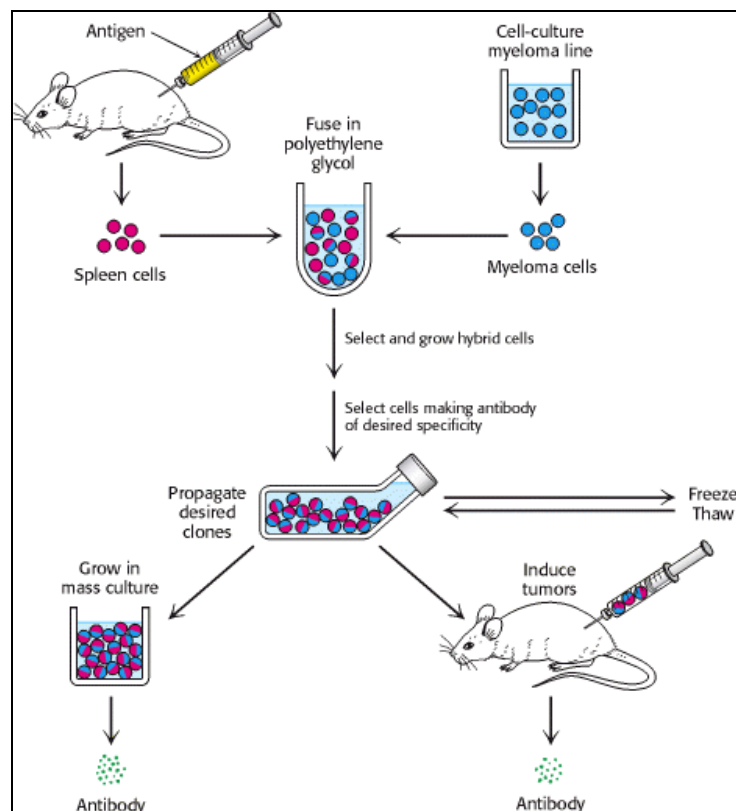


Figure 4.33. Preparation of Monoclonal Antibodies. Hybridoma cells are formed by fusion of antibody-producing cells and myeloma cells. The hybrid cells are allowed to proliferate by growing them in selective medium. They are then screened to determine which ones produce antibody of the desired specificity. [After C. Milstein. Monoclonal antibodies. Copyright © 1980 by Scientific American, Inc. All rights reserved.]

The hybridoma method of producing monoclonal antibodies has opened new vistas in biology and medicine. *Large amounts of homogeneous antibodies with tailor-made specificities can be readily prepared. They are sources of insight into relations between antibody structure and specificity. Moreover, monoclonal antibodies can serve as precise analytical and preparative reagents.* For example, a pure antibody can be obtained against an antigen that has not yet been isolated (Section 4.4). Proteins that guide development have been identified with the use of monoclonal antibodies as tags (Figure 4.34). Monoclonal antibodies attached to solid supports can be used as affinity columns to purify scarce proteins. This method has been used to purify interferon (an antiviral protein) 5,000-fold from a crude mixture. *Clinical laboratories are using monoclonal antibodies in many assays.* For example, the detection in blood of isozymes that are normally localized in the heart points to a myocardial infarction (heart attack). Blood transfusions have been made safer by antibody screening of donor blood for viruses that cause AIDS (acquired immune deficiency syndrome), hepatitis, and other infectious diseases. Monoclonal antibodies are also being evaluated for use as therapeutic agents, as in the treatment of cancer. Furthermore, the vast repertoire of antibody specificity can be tapped to generate catalytic antibodies having novel features not found in naturally occurring enzymes.



Figure 4.34. Fluorescence Micrograph of a Developing *Drosophila* Embryo. The embryo was stained with a fluorescent-labeled monoclonal antibody for the DNA-binding protein encoded by *engrailed*, an essential gene in specifying the body plan. [Courtesy of Dr. Nipam Patel and Dr. Corey Goodman.]

4.3.3. Proteins Can Be Detected and Quantitated by Using an Enzyme-Linked Immunosorbent Assay

Antibodies can be used as exquisitely specific analytic reagents to quantify the amount of a protein or other antigen. The technique is the *enzyme-linked immunosorbent assay (ELISA)*. In this method, an enzyme, which reacts with a colorless substrate to produce a colored product, is covalently linked to a specific antibody that recognizes a target antigen. If the antigen is present, the antibody-enzyme complex will bind to it, and the enzyme component of the antibody-enzyme complex will catalyze the reaction generating the colored product. Thus, the presence of the colored product indicates the presence of the antigen. Such an enzyme-linked immunosorbent assay, which is rapid and convenient, can detect less than a nanogram (10^{-9} g) of a protein. ELISA can be performed with either polyclonal or monoclonal antibodies, but the use of monoclonal antibodies yields more reliable results.

We will consider two among the several types of ELISA. *The indirect ELISA is used to detect the presence of antibody* and is the basis of the test for HIV infection. In that test, viral core proteins (the antigen) are absorbed to the bottom of a well. Antibodies from a patient are then added to the coated well and allowed to bind to the antigen. Finally, enzyme-linked antibodies to human antibodies (for instance, goat antibodies that recognize human antibodies) are allowed to react in the well and unbound antibodies are removed by washing. Substrate is then applied. An enzyme reaction suggests that the enzyme-linked antibodies were bound to human antibodies, which in turn implies that the patient had antibodies to the viral antigen (Figure 4.35).

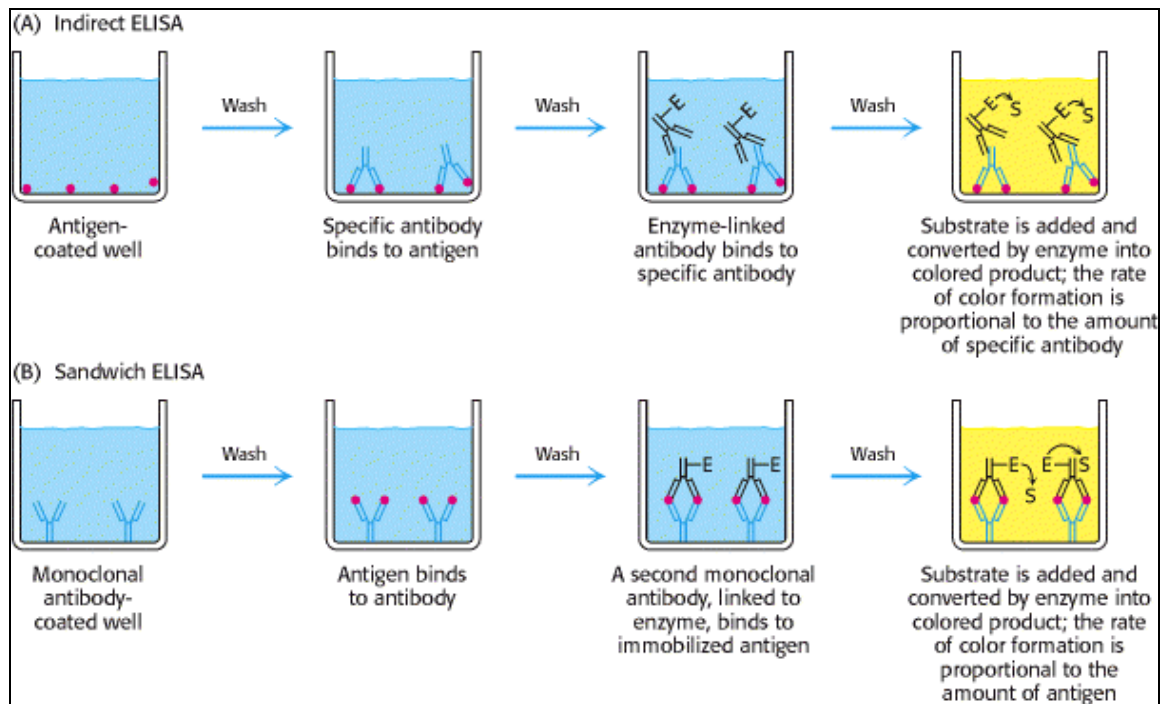


Figure 4.35. Indirect ELISA and Sandwich ELISA (A) In indirect ELISA, the production of color indicates the amount of an antibody to a specific antigen. (B) In sandwich ELISA, the production of color indicates the quantity of antigen. [After R. A. Goldsby, T. J. Kindt, B. A. Osborne, *Kuby Immunology*, 4th ed. (W. H. Freeman and Company, 2000), p. 162.]

The sandwich ELISA allows both the detection and the quantitation of antigen. Antibody to a particular antigen is first absorbed to the bottom of a well. Next, the antigen (or blood or urine containing the antigen) is added to the well and binds to the antibody. Finally, a second, different antibody to the antigen is added. This antibody is enzyme linked and is processed as described for indirect ELISA. In this case, the extent of reaction is directly proportional to the amount of antigen present. Consequently, it permits the measurement of small quantities of antigen (see [Figure 4.35](#)).

4.3.4. Western Blotting Permits the Detection of Proteins Separated by Gel Electrophoresis

Often it is necessary to detect small quantities of a particular protein in the presence of many other proteins, such as a viral protein in the blood. Very small quantities of a protein of interest in a cell or in body fluid can be detected by an immunoassay technique called *Western blotting* ([Figure 4.36](#)). A sample is subjected to electrophoresis on an SDS-polyacrylamide gel. Blotting (or more typically electroblotting) transfers the resolved proteins on the gel to the surface of a polymer sheet to make them more accessible for reaction. An antibody that is specific for the protein of interest is added to the sheet and reacts with the antigen. The antibody-antigen complex on the sheet then can be detected by rinsing the sheet with a second antibody specific for the first (e.g., goat antibody that recognizes mouse antibody). A radioactive label on the second antibody produces a dark band on x-ray film (an autoradiogram). Alternatively, an enzyme on the second antibody generates a colored product, as in the ELISA method. Western blotting makes it possible to find a protein in a complex mixture, the proverbial needle in a haystack. It is the basis for the test for infection by hepatitis C, where it is used to detect a core protein of the virus. This technique is also very useful in the cloning of genes.

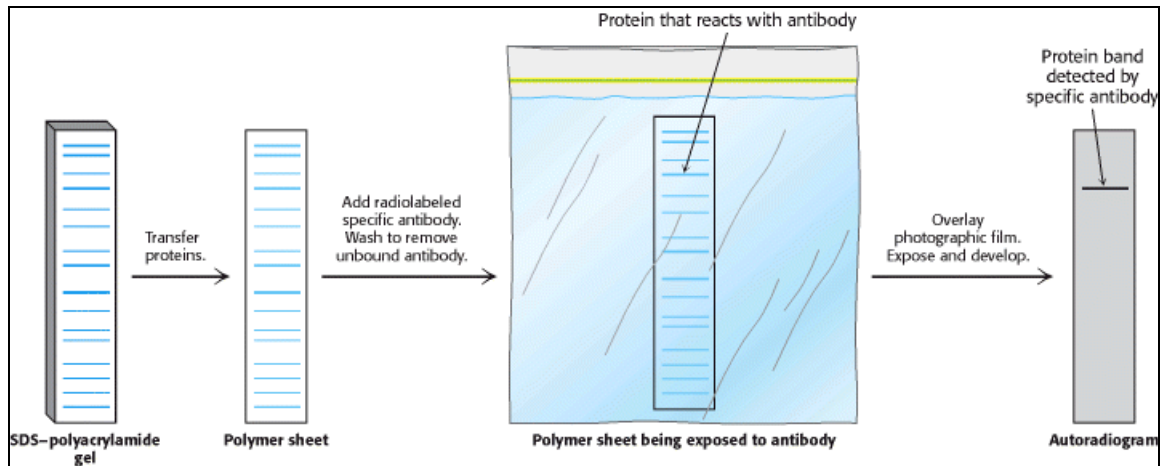


Figure 4.36. Western Blotting. Proteins on an SDS-polyacrylamide gel are transferred to a polymer sheet and stained with radioactive antibody. A band corresponding to the protein to which the antibody binds appears in the autoradiogram.

4.3.5. Fluorescent Markers Make Possible the Visualization of Proteins in the Cell

Biochemistry is often performed in test tubes or polyacrylamide gels. However, most proteins function in the context of a cell. Fluorescent markers provide a powerful means of examining proteins in their biological context. For instance, cells can be stained with fluorescence-labeled antibodies or other fluorescent proteins and examined by *fluorescence microscopy* to reveal the location of a protein of interest. Arrays of parallel bundles are evident in cells stained with antibody specific for actin, a protein that polymerizes into filaments (Figure 4.37). Actin filaments are constituents of the cytoskeleton, the internal scaffolding of cells that controls their shape and movement. By tracking protein location, fluorescent markers also provide clues to protein function. For instance, the glucocorticoid receptor protein is a transcription factor that controls gene expression in response to the steroid hormone cortisone. The receptor was linked to *green fluorescent protein (GFP)*, a naturally fluorescent protein isolated from the jellyfish *Aequorea victoria* (Section 3.6.5). Fluorescence microscopy revealed that, in the absence of the hormone, the receptor is located in the cytoplasm (Figure 4.38A). On addition of the steroid, the receptor is translocated to the nucleus, where it binds to DNA (Figure 4.38B).

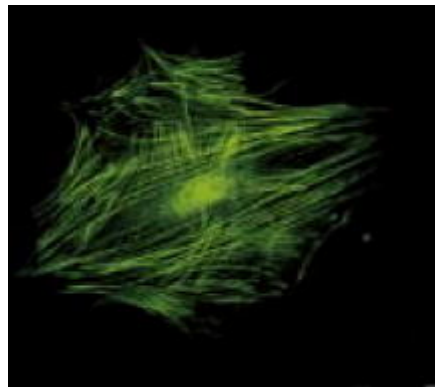


Figure 4.37. Actin Filaments. Fluorescence micrograph of actin filaments in a cell stained with an antibody specific to actin. [Courtesy of Dr. Elias Lazarides.]

The highest resolution of fluorescence microscopy is about $0.2 \mu\text{m}$ (200 nm, or 2000 \AA), the wavelength of visible light. Finer spatial resolution can be achieved by electron microscopy by using antibodies tagged with electron-dense markers. For example, ferritin conjugated to an antibody can be readily visualized by electron microscopy because it contains an electron-dense core rich in iron. Clusters of gold also can be conjugated to antibodies to make them highly visible under the electron microscope. *Immunoelectron microscopy* can define the position of antigens to a resolution of 10 nm (100 \AA) or finer (Figure 4.39).

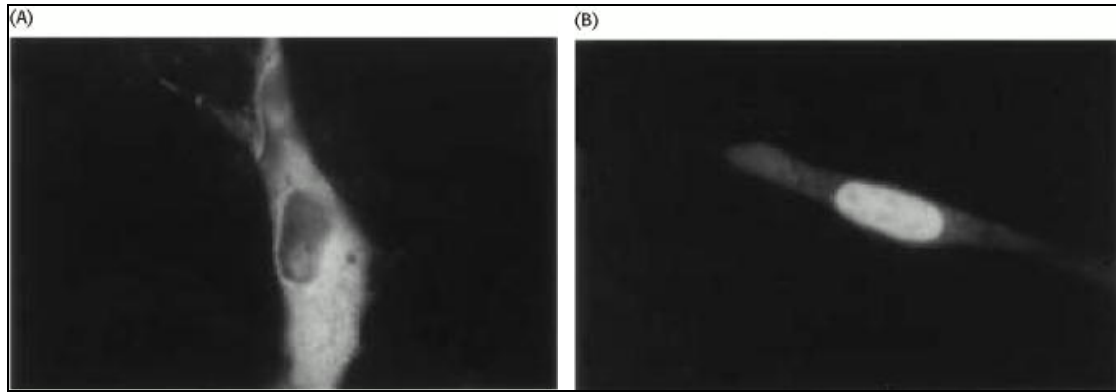


Figure 4.38. Nuclear Localization of a Steroid Receptor. (A) The receptor, made visible by attachment of the green fluorescent protein, is located predominantly in the cytoplasm of the cultured cell. (B) Subsequent to the addition of corticosterone (a glucocorticoid steroid), the receptor moves into the nucleus. [Courtesy of Professor William B. Pratt/Department of Pharmacology, University of Michigan.]



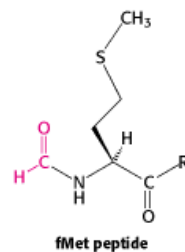
Figure 4.39. Immunoelectron Microscopy. The opaque particles (150-Å, or 15-nm, diameter) in this electron micrograph are clusters of gold atoms bound to antibody molecules. These membrane vesicles from the synapses of neurons contain a channel protein that is recognized by the specific antibody. [Courtesy of Dr. Peter Sargent.]

4.4. Peptides Can Be Synthesized by Automated Solid-Phase Methods

The ability to synthesize peptides of defined sequence is a powerful technique for extending biochemical analysis for several reasons.

1. *Synthetic peptides can serve as antigens to stimulate the formation of specific antibodies.* For instance, as discussed earlier, it is often more efficient to obtain a protein sequence from a nucleic acid sequence than by sequencing the protein itself (see also [Chapter 6](#)). Peptides can be synthesized on the basis of the nucleic acid sequence, and antibodies can be raised that target these peptides. These antibodies can then be used to isolate the intact protein from the cell.

2. *Synthetic peptides can be used to isolate receptors for many hormones and other signal molecules.* For example, white blood cells are attracted to bacteria by formylmethionyl (fMet) peptides released in the breakdown of bacterial proteins. Synthetic formylmethionyl peptides have been useful in identifying the cell-surface receptor for this class of peptide. Moreover, synthetic peptides can be attached to agarose beads to prepare affinity chromatography columns for the purification of receptor proteins that specifically recognize the peptides.



3. *Synthetic peptides can serve as drugs.* Vasopressin is a peptide hormone that stimulates the reabsorption of water in the distal tubules of the kidney, leading to the formation of more concentrated urine. Patients with diabetes insipidus are deficient in *vasopressin* (also called *antidiuretic hormone*), and so they excrete large volumes of urine (more than 5 liters per day) and are continually thirsty. This defect can be treated by administering 1-desamino-8-D-arginine vasopressin, a synthetic analog of the missing hormone ([Figure 4.40](#)). This synthetic peptide is degraded *in vivo* much more slowly than vasopressin and, additionally, does not increase the blood pressure.

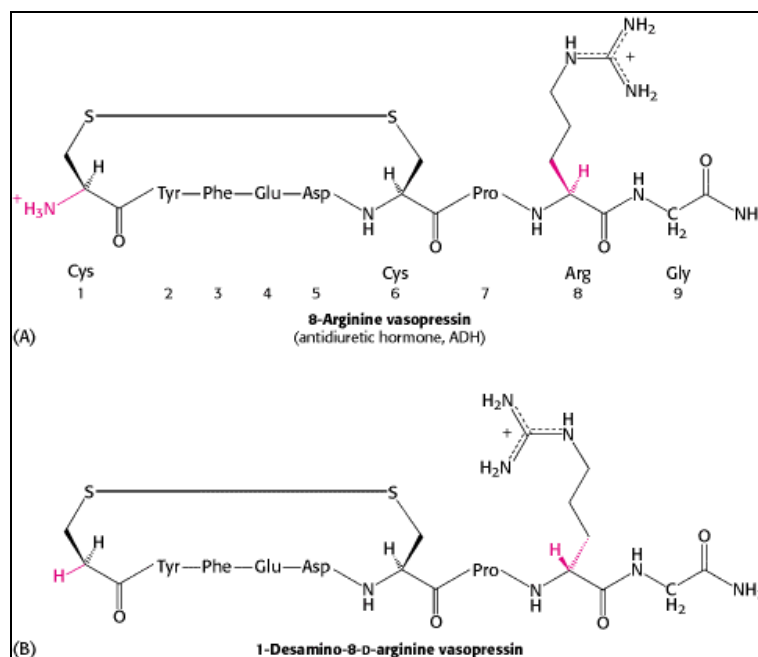


Figure 4.40. Vasopressin and Synthetic Vasopressin. Structural formulas of (A) vasopressin, a peptide hormone that stimulates water resorption, and (B) 1-desamino-8-d-arginine vasopressin, a more stable synthetic analog of this antidiuretic hormone.

4. Finally, *studying synthetic peptides can help define the rules governing the three-dimensional structure of proteins*. We can ask whether a particular sequence by itself folds into an α helix, β strand, or hairpin turn or behaves as a random coil.

How are these peptides constructed? The amino group of one amino acid is linked to the carboxyl group of another. However, a unique product is formed only if a single amino group and a single carboxyl group are available for reaction. Therefore, it is necessary to block some groups and to activate others to prevent unwanted reactions. The α -amino group of the first amino acid of the desired peptide is blocked with a *tert*-butyloxycarbonyl (*t*-Boc) group, yielding a *t*-Boc amino acid. The carboxyl group of this same amino acid is activated by reacting it with a reagent such as *dicyclohexylcarbodiimide* (DCC), as illustrated in [Figure 4.41](#). The free amino group of the next amino acid to be linked attacks the activated carboxyl, leading to the formation of a peptide bond and the release of dicyclohexylurea. The carboxyl group of the resulting dipeptide is activated with DCC and reacted with the free amino group of the amino acid that will be the third residue in the peptide. This process is repeated until the desired peptide is synthesized. Exposing the peptide to dilute acid removes the *t*-Boc protecting group from the first amino acid while leaving peptide bonds intact.

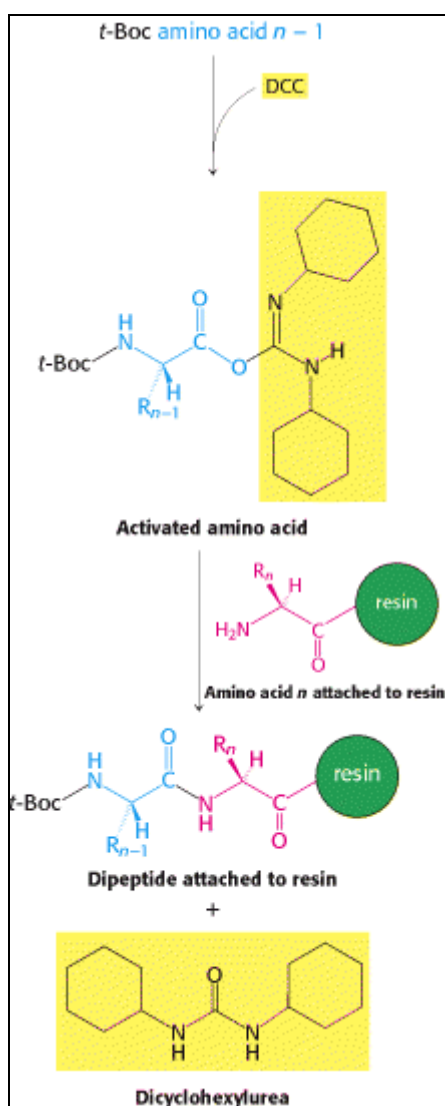


Figure 4.41. Amino Acid Activation. Dicyclohexylcarbodiimide is used to activate carboxyl groups for the formation of peptide bonds.

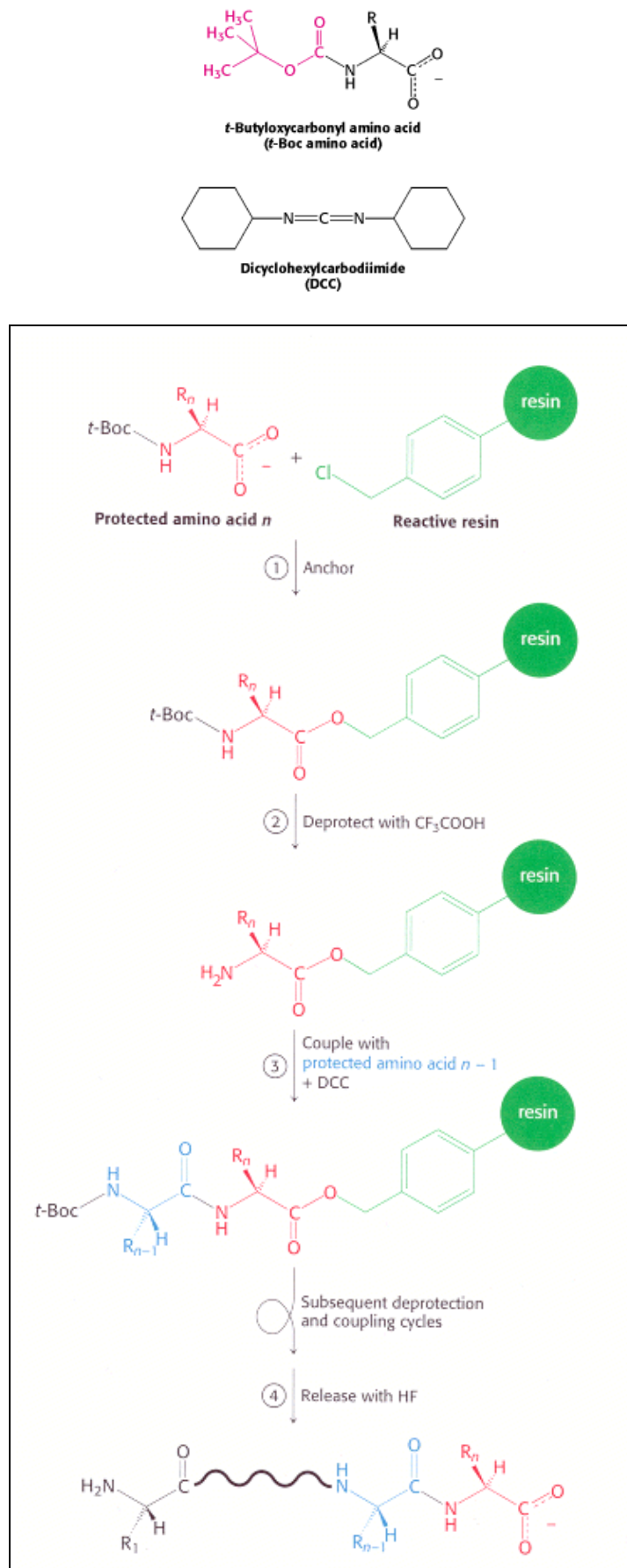


Figure 4.42. Solid-Phase Peptide Synthesis. The sequence of steps in solid-phase synthesis is: (1) anchoring of the C-terminal amino acid, (2) deprotection of the amino terminus, and (3) coupling of the next residue. Steps 2 and 3 are repeated for each added amino acid. Finally, in step 4, the completed peptide is released from the resin.

Peptides containing more than 100 amino acids can be synthesized by sequential repetition of the preceding reactions. Linking the growing peptide chain to an insoluble matrix, such as polystyrene beads, further enhances efficiency. A major advantage of this *solid-phase method* is that the desired product at each stage is bound to beads that can be rapidly filtered and washed, and so there is no need to purify intermediates. All reactions are carried out in a single vessel, eliminating losses caused by repeated transfers of products. The carboxyl-terminal amino acid of the desired peptide sequence is first anchored to the polystyrene beads ([Figure 4.42](#)). The *t*-Boc protecting group of this amino acid is then removed. The next amino acid (in the protected *t*-Boc form) and dicyclohexylcarbodiimide, the coupling agent, are added together. After the peptide bond forms, excess reagents and dicyclohexylurea are washed away, leaving the desired dipeptide product attached to the beads. Additional amino acids are linked by the same sequence of reactions. At the end of the synthesis, the peptide is released from the beads by adding hydrofluoric acid (HF), which cleaves the carboxyl ester anchor without disrupting peptide bonds. Protecting groups on potentially reactive side chains, such as that of lysine, also are removed at this time. This cycle of reactions can be readily automated, which makes it feasible to routinely synthesize peptides containing about 50 residues in good yield and purity. In fact, the solid-phase method has been used to synthesize interferons (155 residues) that have antiviral activity and ribonuclease (124 residues) that is catalytically active.

4.5. Three-Dimensional Protein Structure Can Be Determined by NMR Spectroscopy and X-Ray Crystallography

A crucial question is, What does the three-dimensional structure of a specific protein look like? Protein structure determines function, given that the specificity of active sites and binding sites depends on the precise threedimensional conformation. Nuclear magnetic resonance spectroscopy and x-ray crystallography are two of the most important techniques for elucidating the conformation of proteins.

4.5.1. Nuclear Magnetic Resonance Spectroscopy Can Reveal the Structures of Proteins in Solution

Nuclear magnetic resonance (NMR) spectroscopy is unique in being able to reveal the *atomic structure* of macromolecules *in solution*, provided that highly concentrated solutions (~ 1 mM, or 15 mg ml^{-1} for a 15-kd protein) can be obtained. This technique depends on the fact that certain atomic nuclei are intrinsically magnetic. Only a limited number of isotopes display this property, called *spin*, and the ones most important to biochemistry are listed in [Table 4.4](#). The simplest example is the hydrogen nucleus (^1H), which is a proton. The spinning of a proton generates a magnetic moment. This moment can take either of two orientations, or spin states (called α and β), when an external magnetic field is applied ([Figure 4.43](#)). The energy difference between these states is proportional to the strength of the imposed magnetic field. The α state has a slightly lower energy and hence is slightly more populated (by a factor of the order of 1.00001 in a typical experiment) because it is aligned with the field. A spinning proton in an α state can be raised to an excited state (β state) by applying a pulse of electromagnetic radiation (a radio-frequency, or RF, pulse), provided the frequency corresponds to the energy difference between the α and the β states. In these circumstances, the spin will change from α to β ; in other words, *resonance* will be obtained. A resonance spectrum for a molecule can be obtained by varying the magnetic field at a constant frequency of electromagnetic radiation or by keeping the magnetic field constant and varying electromagnetic radiation.

Nucleus	Natural abundance (% by weight Nucleus of the element)
^1H	99.984
^2H	0.016
^{13}C	1.108
^{14}N	99.635
^{15}N	0.365
^{17}O	0.037
^{23}Na	100.0
^{25}Mg	10.05
^{31}P	100.0
^{35}Cl	75.4
^{39}K	93.1

Table 4.4. Biologically important nuclei giving NMR signals

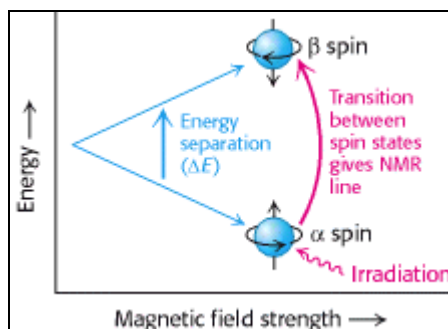


Figure 4.43. Basis of NMR Spectroscopy. The energies of the two orientations of a nucleus of spin $1/2$ (such as ^{31}P and ^1H) depend on the strength of the applied magnetic field. Absorption of electromagnetic radiation of appropriate frequency induces a transition from the lower to the upper level.

These properties can be used to examine the chemical surroundings of the hydrogen nucleus. The flow of electrons around a magnetic nucleus generates a small local magnetic field that opposes the applied field. The degree of such shielding depends on the surrounding electron density. Consequently, nuclei in different environments will change states, or resonate, at slightly different field strengths or radiation frequencies. The nuclei of the perturbed sample absorb electromagnetic radiation at a frequency that can be measured. The different frequencies, termed *chemical shifts*, are expressed in fractional units δ (parts per million, or ppm) relative to the shifts of a standard compound, such as a water-soluble derivative of tetramethylsilane, that is added with the sample. For example, a $-\text{CH}_3$ proton typically exhibits a chemical shift (δ) of 1 ppm, compared with a chemical shift of 7 ppm for an aromatic proton. The chemical shifts of most protons in protein molecules fall between 0 and 9 ppm (Figure 4.44). It is possible to resolve most protons in many proteins by using this technique of *onedimensional NMR*. With this information, we can then deduce changes to a particular chemical group under different conditions, such as the conformational change of a protein from a disordered structure to an α helix in response to a change in pH.

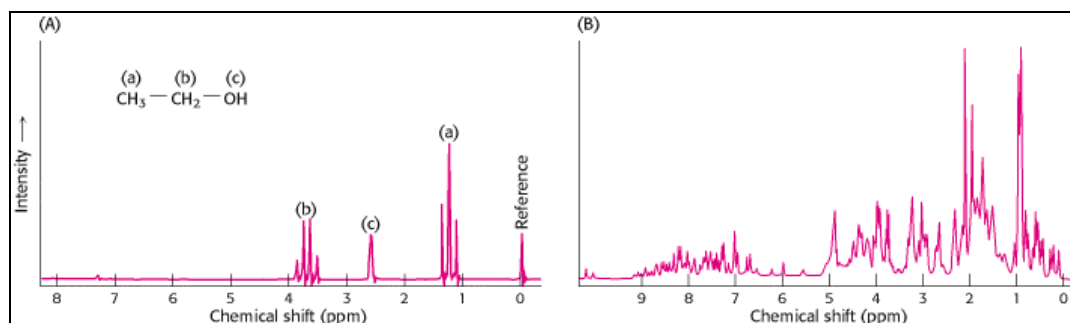


Figure 4.44. One-Dimensional NMR Spectra. (A) ^1H -NMR spectrum of ethanol ($\text{CH}_3\text{CH}_2\text{OH}$) shows that the chemical shifts for the hydrogen are clearly resolved. (B) ^1H -NMR spectrum from a 55 amino acid fragment of a protein with a role in RNA splicing shows a greater degree of complexity. A large number of peaks are present and many overlap. [(A) After C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, 1991), p. 280; (B) courtesy of Barbara Amann and Wesley McDermott.]

We can garner even more information by examining how the spins on different protons affect their neighbors. By inducing a transient magnetization in a sample through the application a radio-frequency pulse, it is possible to alter the spin on one nucleus and examine the effect on the spin of a neighboring nucleus. Especially revealing is a *two-dimensional spectrum obtained by nuclear Overhauser enhancement spectroscopy (NOESY)*, which graphically displays pairs of protons that are in close proximity, even if they are not close together in the primary structure. The basis for this technique is the *nuclear Overhauser effect (NOE)*, an interaction between nuclei that is proportional to the inverse sixth power of the distance between them. Magnetization is transferred from an excited nucleus to an unexcited one if they are less than about 5 Å apart (Figure 4.45A). In other words, the effect provides a means of detecting the location of atoms relative to one another in the three-dimensional structure of the protein. The diagonal of a NOESY spectrum corresponds to a one-dimensional spectrum. The offdiagonal peaks provide crucial new information: they identify pairs of protons that are less than 5 Å apart (Figure 4.45B). A two-dimensional NOESY spectrum for a protein comprising 55 amino acids is shown in Figure 4.46. The large number of off-diagonal peaks reveals short proton-proton distances. The three-dimensional structure of a protein can be reconstructed with the use of such proximity relations. Structures are calculated such that protons that must be separated by less than 5 Å on the basis of NOESY spectra are close to one another in the three-dimensional structure (Figure 4.47). If a sufficient number of

distance constraints are applied, the three-dimensional structure can be determined nearly uniquely. A family of related structures is generated for three reasons (Figure 4.48). First, not enough constraints may be experimentally accessible to fully specify the structure. Second, the distances obtained from analysis of the NOESY spectrum are only approximate. Finally, the experimental observations are made not on single molecules but on a large number of molecules in solution that may have slightly different structures at any given moment. Thus, the family of structures generated from NMR structure analysis indicates the range of conformations for the protein in solution. At present, NMR spectroscopy can determine the structures of only relatively small proteins (<40 kd), but its resolving power is certain to increase. The power of NMR has been greatly enhanced by the ability to produce proteins labeled uniformly or at specific sites with ^{13}C , ^{15}N , and ^2H with the use of recombinant DNA technology (Chapter 6).

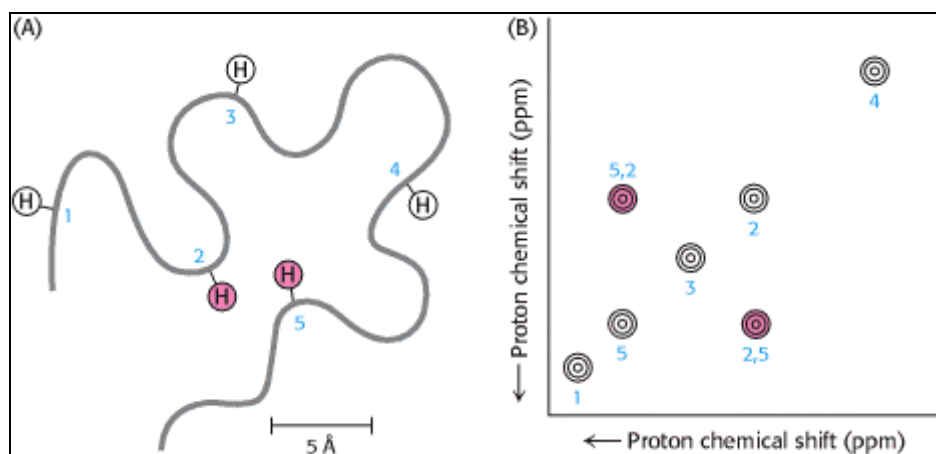


Figure 4.45. The Nuclear Overhauser Effect. The nuclear Overhauser effect (NOE) identifies pairs of protons that are in close proximity. (A) Schematic representation of a polypeptide chain highlighting five particular protons. Protons 2 and 5 are in close proximity ($\sim 4 \text{ \AA}$ apart), whereas other pairs are farther apart. (B) A highly simplified NOESY spectrum. The diagonal shows five peaks corresponding to the five protons in part A. The peaks above the diagonal and the symmetrically related one below reveal that proton 2 is close to proton 5.

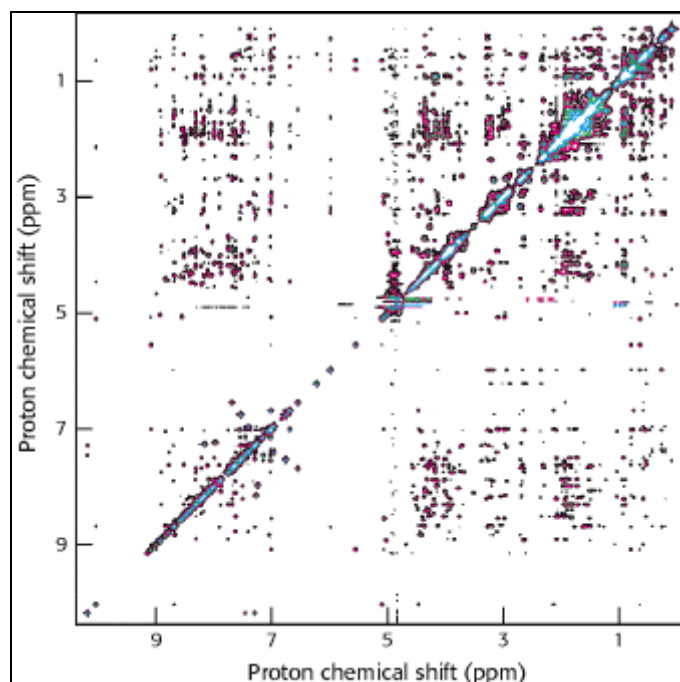


Figure 4.46. Detecting Short Proton-Proton Distances. A NOESY spectrum for a 55 amino acid domain from a protein having a role in RNA splicing. Each off-diagonal peak corresponds to a short proton-proton separation. This spectrum reveals hundreds of such short proton-proton distances, which can be used to determine the three-dimensional structure of this domain. [Courtesy of Barbara Amann and Wesley McDermott.]

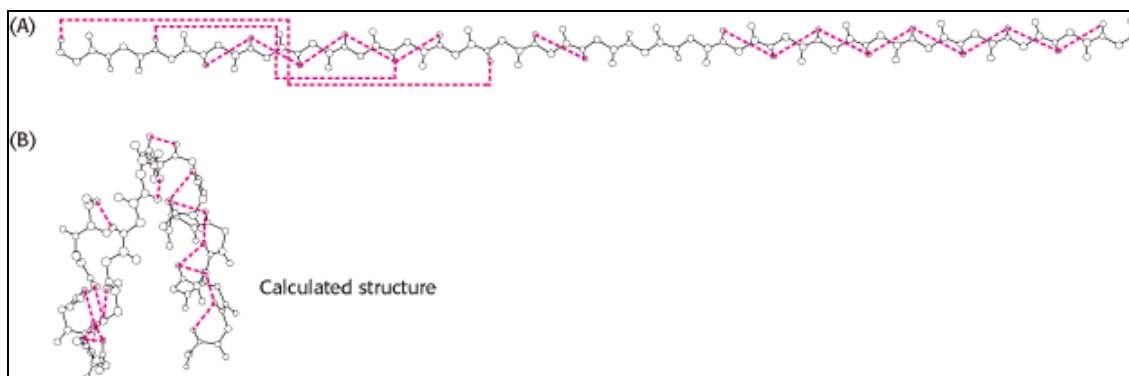


Figure 4.47. Structures Calculated on the Basis of NMR Constraints. (A) NOESY observations show that protons (connected by dotted red lines) are close to one another in space. (B) A three-dimensional structure calculated with these proton pairs constrained to be close together.

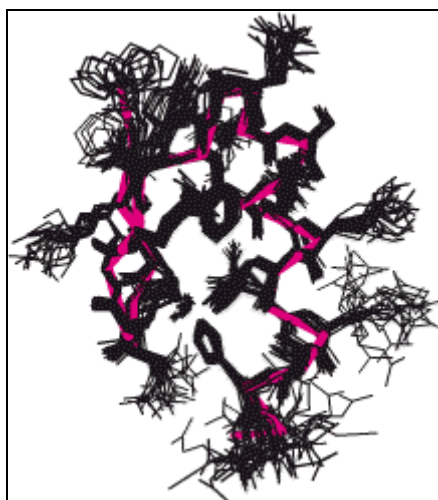


Figure 4.48. A Family of Structures. A set of 25 structures for a 28 amino acid domain from a zinc-finger-DNA-binding protein. The red line traces the average course of the protein backbone. Each of these structures is consistent with hundreds of constraints derived from NMR experiments. The differences between the individual structures are due to a combination of imperfections in the experimental data and the dynamic nature of proteins in solution. [Courtesy of Barbara Amann.]

4.5.2. X-Ray Crystallography Reveals Three-Dimensional Structure in Atomic Detail

X-ray crystallography provides the finest visualization of protein structure currently available. This technique can reveal the precise three-dimensional positions of most atoms in a protein molecule. The use of x-rays provides the best resolution because the wavelength of x-rays is about the same length as that of a covalent bond. The three components in an x-ray crystallographic analysis are a *protein crystal*, a *source of x-rays*, and a *detector* (Figure 4.49).

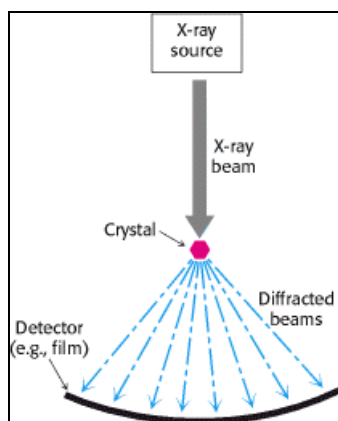


Figure 4.49. Essence of an X-Ray Crystallographic Experiment: an X-Ray Beam, a Crystal, and a Detector.

The technique requires that all molecules be precisely oriented, so the first step is to obtain crystals of the protein of interest. Slowly adding ammonium sulfate or another salt to a concentrated solution of protein to reduce its solubility favors the formation of highly ordered crystals. This is the process of salting out discussed in [Section 4.1.3](#). For example, myoglobin crystallizes in 3 M ammonium sulfate ([Figure 4.50](#)). Some proteins crystallize readily, whereas others do so only after much effort has been expended in identifying the right conditions. Crystallization is an art; the best practitioners have great perseverance and patience. Increasingly large and complex proteins are being crystallized. For example, poliovirus, an 8500-kd assembly of 240 protein subunits surrounding an RNA core, has been crystallized and its structure solved by x-ray methods. Crucially, protein crystals frequently display their biological activity, indicating that the proteins have crystallized in their biologically active configuration. For instance, enzyme crystals may display catalytic activity if the crystals are suffused with substrate.

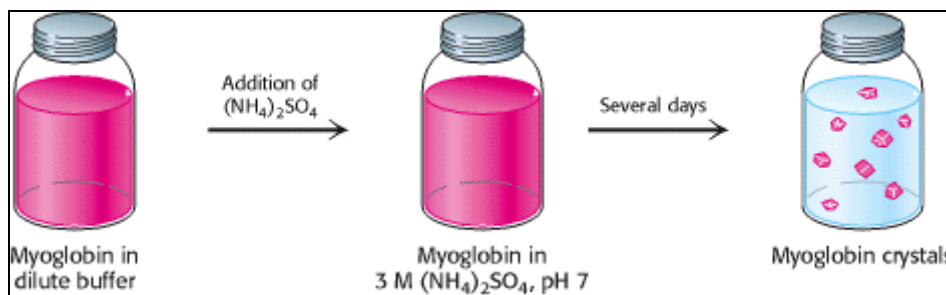


Figure 4.50. Crystallization of Myoglobin.

Next, a source of x-rays is required. A beam of x-rays of wavelength 1.54 Å is produced by accelerating electrons against a copper target. A narrow beam of x-rays strikes the protein crystal. Part of the beam goes straight through the crystal; the rest is *scattered* in various directions. Finally, these scattered, or *diffracted*, x-rays are detected by x-ray film, the blackening of the emulsion being proportional to the intensity of the scattered x-ray beam, or by a solid-state electronic detector. The scattering pattern provides abundant information about protein structure. The basic physical principles underlying the technique are:

1. *Electrons scatter x-rays.* The amplitude of the wave scattered by an atom is proportional to its number of electrons. Thus, a carbon atom scatters six times as strongly as a hydrogen atom does.
2. *The scattered waves recombine.* Each atom contributes to each scattered beam. The scattered waves reinforce one another at the film or detector if they are in phase (in step) there, and they cancel one another if they are out of phase.
3. *The way in which the scattered waves recombine depends only on the atomic arrangement.*

The protein crystal is mounted and positioned in a precise orientation with respect to the x-ray beam and the film. The crystal is rotated so that the beam can strike the crystal from many directions. This rotational motion results in an x-ray photograph consisting of a regular array of spots called *reflections*. The x-ray photograph shown in [Figure 4.51](#) is a twodimensional section through a three-dimensional array of 25,000 spots. The intensity of each spot is measured. These *intensities and their positions* are the basic experimental data of an x-ray crystallographic analysis. The next step is to reconstruct an image of the protein from the observed intensities. In light microscopy or electron microscopy, the diffracted beams are focused by lenses to directly form an image. However, appropriate lenses for focusing x-rays do not exist. Instead, the image is formed by applying a mathematical relation called a Fourier transform. For each spot, this operation yields a wave of electron density whose amplitude is proportional to the square root of the observed intensity of the spot. Each wave also has a *phase* - that is, the timing of its crests and troughs relative to those of other waves. The phase of each wave determines whether the wave reinforces or cancels the waves contributed by the other spots. These phases can be deduced from the well-understood diffraction patterns produced by electron-dense heavy-atom reference markers such as uranium or mercury at specific sites in the protein.

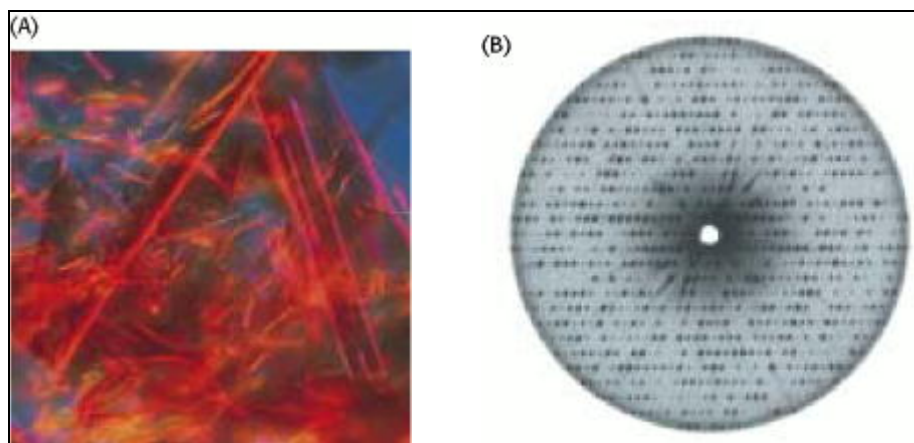


Figure 4.51. Myoglobin Crystal and X-Ray. (A) Crystal of myoglobin. (B) X-ray precession photograph of a myoglobin crystal. [(A) Mel Pollinger/Fran Heyl Associates.]

The stage is then set for the calculation of an electron-density map, which gives the density of electrons at a large number of regularly spaced points in the crystal. This three-dimensional electron-density distribution is represented by a series of parallel sections stacked on top of one another. Each section is a transparent plastic sheet (or, more recently, a layer in a computer image) on which the electron-density distribution is represented by contour lines (Figure 4.52), like the contour lines used in geological survey maps to depict altitude (Figure 4.53). The next step is to interpret the electron-density map. A critical factor is the *resolution* of the x-ray analysis, which is determined by the number of scattered intensities used in the Fourier synthesis. The fidelity of the image depends on the resolution of the Fourier synthesis, as shown by the optical analogy in Figure 4.54. A resolution of 6 Å reveals the course of the polypeptide chain but few other structural details. The reason is that polypeptide chains pack together so that their centers are between 5 Å and 10 Å apart. Maps at higher resolution are needed to delineate groups of atoms, which lie between 2.8 Å and 4.0 Å apart, and individual atoms, which are between 1.0 Å and 1.5 Å apart. The ultimate resolution of an x-ray analysis is determined by the degree of perfection of the crystal. For proteins, this limiting resolution is usually about 2 Å.

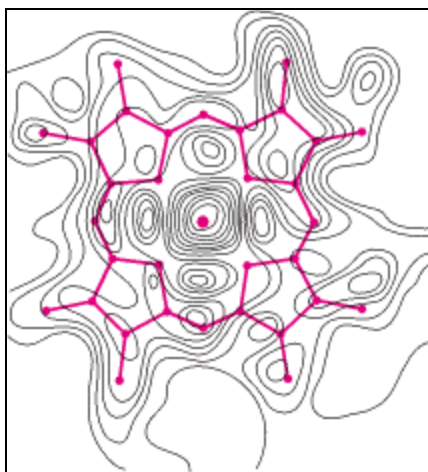


Figure 4.52. Section of the Electron-Density Map of Myoglobin. This section of the electron-density map shows the heme group. The peak of the center of this section corresponds to the position of the iron atom. [From J. C. Kendrew. The three-dimensional structure of a protein molecule. Copyright © 1961 by Scientific American, Inc. All rights reserved.]

The structures of more than 10,000 proteins had been elucidated by NMR and x-ray crystallography by mid-2000, and several new structures are now determined each day. The coordinates are collected at the Protein Data Bank (<http://www.rcsb.org/pdb>) and the structures can be accessed for visualization and analysis. Knowledge of the detailed molecular architecture of proteins has been a source of insight into how proteins recognize and bind other molecules, how they function as enzymes, how they fold, and how they evolved. This extraordinarily rich harvest is continuing at a rapid pace and is greatly influencing the entire field of biochemistry.

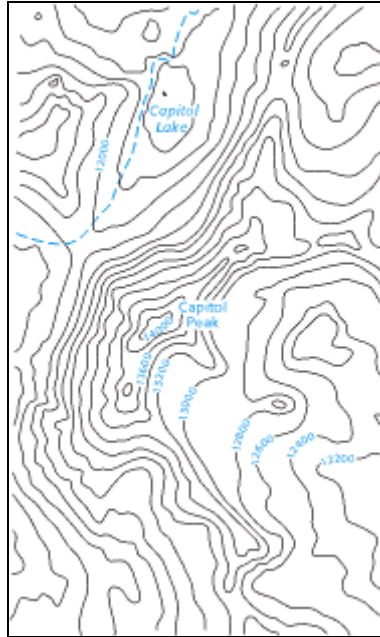


Figure 4.53. Section of a U.S. Geological Survey Map. Capitol Peak Quadrangle, Colorado.

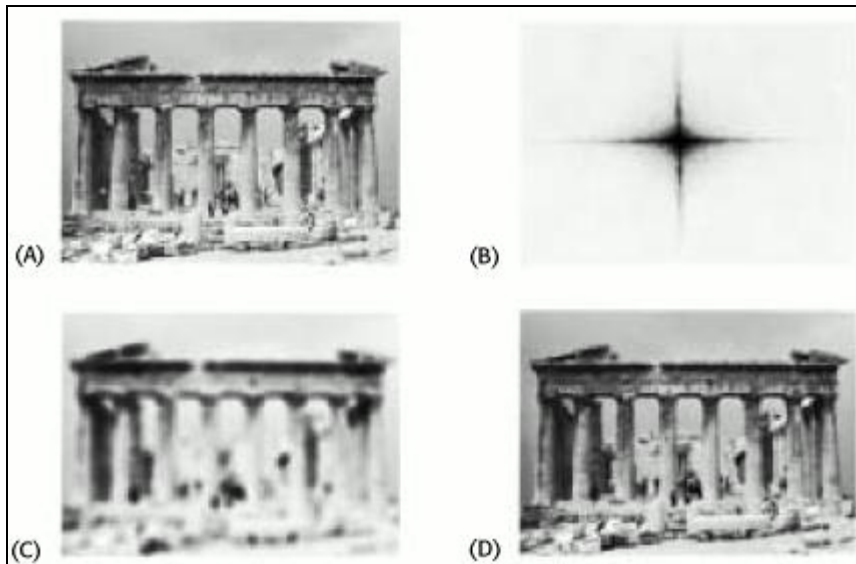


Figure 4.54. Resolution Affects the Quality of an Image. The effect of resolution on the quality of a reconstructed image is shown by an optical analog of x-ray diffraction: (A) a photograph of the Parthenon; (B) an optical diffraction pattern of the Parthenon; (C and D) images reconstructed from the pattern in part B. More data were used to obtain image D than image C, which accounts for the higher quality of image D. [(A) Courtesy of Dr. Thomas Steitz. (B) Courtesy of Dr. David DeRosier.]

Summary

The rapid progress in gene sequencing has advanced another goal of biochemistry - elucidation of the proteome. The proteome is the complete set of proteins expressed and includes information about how they are modified, how they function, and how they interact with other molecules.

The Purification of Proteins Is an Essential Step in Understanding Their Function

Proteins can be separated from one another and from other molecules on the basis of such characteristics as solubility, size, charge, and binding affinity. SDS-polyacrylamide gel electrophoresis separates the polypeptide chains of proteins under denaturing conditions largely according to mass. Proteins can also be separated electrophoretically on the basis of net charge by isoelectric focusing in a pH gradient. Ultracentrifugation and gel-filtration chromatography resolve proteins according to size, whereas ion-exchange chromatography separates them mainly on the basis of net charge. The high affinity of many proteins for specific chemical groups is exploited in affinity chromatography, in which proteins bind to columns containing beads bearing covalently linked substrates, inhibitors, or other specifically recognized groups. The mass of a protein can be precisely determined by sedimentation equilibrium measurements or by mass spectrometry.

Amino Acid Sequences Can Be Determined by Automated Edman Degradation

The amino acid composition of a protein can be ascertained by hydrolyzing it into its constituent amino acids in 6 N HCl at 110°C. The amino acids can be separated by ion-exchange chromatography and quantitated by reacting them with ninhydrin or fluorescamine. Amino acid sequences can be determined by Edman degradation, which removes one amino acid at a time from the amino end of a peptide. Phenyl isothiocyanate reacts with the terminal amino group to form a phenylthiocarbamoyl derivative, which cyclizes under mildly acidic conditions to give a phenylthiohydantoin-amino acid and a peptide shortened by one residue. Automated repeated Edman degradations by a sequenator can analyze sequences of about 50 residues. Longer polypeptide chains are broken into shorter ones for analysis by specifically cleaving them with a reagent such as cyanogen bromide, which splits peptide bonds on the carboxyl side of methionine residues. Enzymes such as trypsin, which cleaves on the carboxyl side of lysine and arginine residues, also are very useful in splitting proteins. Amino acid sequences are rich in information concerning the kinship of proteins, their evolutionary relations, and diseases produced by mutations. Knowledge of a sequence provides valuable clues to conformation and function.

Immunology Provides Important Techniques with Which to Investigate Proteins

Proteins can be detected and quantitated by highly specific antibodies; monoclonal antibodies are especially useful because they are homogeneous. Enzyme-linked immunosorbent assays and Western blots of SDS-polyacrylamide gels are used extensively. Proteins can also be localized within cells by immunofluorescence microscopy and immunoelectron microscopy.

Peptides Can Be Synthesized by Automated Solid-Phase Methods

Polypeptide chains can be synthesized by automated solid-phase methods in which the carboxyl end of the growing chain is linked to an insoluble support. The α -carboxyl group of the incoming amino acid is activated by dicyclohexylcarbodiimide and joined to the α -amino group of the growing chain. Synthetic peptides can serve as drugs and as antigens to stimulate the formation of specific antibodies. They can also be sources of insight into relations between amino acid sequence and conformation.

Three-Dimensional Protein Structure Can Be Determined by NMR Spectroscopy and X-Ray Crystallography

Nuclear magnetic resonance spectroscopy and x-ray crystallography have greatly enriched our understanding of how proteins fold, recognize other molecules, and catalyze chemical reactions. Nuclear magnetic resonance spectroscopy reveals the structure and dynamics of proteins in solution. The chemical shift of nuclei depends on their local environment. Furthermore, the spins of neighboring nuclei interact with each other in ways that provide definitive structural information.

X-ray crystallography is possible because electrons scatter x-rays; the way in which the scattered waves recombine depends only on the atomic arrangement. The three-dimensional structures of thousands of proteins are now known in atomic detail.

Key Terms

proteome

assay

homogenate

salting out

dialysis

gel-filtration chromatography

ion-exchange chromatography

affinity chromatography

high-pressure liquid chromatography (HPLC)

gel electrophoresis

isoelectric point

isoelectric focusing

two-dimensional electrophoresis

sedimentation coefficient (Svedberg units, S)

matrix-assisted laser desorption- ionization-time of flight spectrometry (MALDI-TOF)

dabsyl chloride

dansyl chloride

Edman degradation

phenyl isothiocyanate

cyanogen bromide (CNBr)

overlap peptides

diagonal electrophoresis

antibody

antigen

antigenic determinant (epitope)

monoclonal antibodies

enzyme-linked immunosorbent assay (ELISA)

Western blotting

fluorescence microscopy

green fluorescent protein (GFP)

solid-phase method

nuclear magnetic resonance (NMR) spectroscopy

x-ray crystallography

Problems

1. **Valuable reagents.** The following reagents are often used in protein chemistry:

CNBr	Performic acid	Phenyl isothiocyanate
Urea	Dabsyl chloride	Chymotrypsin
Mercaptoethanol	6 N HCl	
Trypsin	Ninhydrin	

Which one is the best suited for accomplishing each of the following tasks?

- Determination of the amino acid sequence of a small peptide.
- Identification of the amino-terminal residue of a peptide (of which you have less than 0.1 μg).
- Reversible denaturation of a protein devoid of disulfide bonds. Which additional reagent would you need if disulfide bonds were present?
- Hydrolysis of peptide bonds on the carboxyl side of aromatic residues.
- Cleavage of peptide bonds on the carboxyl side of methionines.
- Hydrolysis of peptide bonds on the carboxyl side of lysine and arginine residues.

Answer:

(a) Phenyl isothiocyanate; (b) dansyl chloride or dabsyl chloride; (c) urea; β -mercaptoethanol to reduce disulfides; (d) chymotrypsin; (e) CNBr; (f) trypsin

2. **Finding an end.** Anhydrous hydrazine ($\text{H}_2\text{N-NH}_2$) has been used to cleave peptide bonds in proteins. What are the reaction products? How might this technique be used to identify the carboxyl-terminal amino acid?

Answer:

Each amino acid residue, except the carboxyl-terminal residue, gives rise to a hydrazide on reacting with hydrazine. The carboxyl-terminal residue can be identified because it yields a free amino acid.

3. **Crafting a new breakpoint.** Ethyleneimine reacts with cysteine side chains in proteins to form *S*-aminoethyl derivatives. The peptide bonds on the carboxyl side of these modified cysteine residues are susceptible to hydrolysis by trypsin. Why?

Answer:

The *S*-aminoethylcysteine side chain resembles that of lysine. The only difference is a sulfur atom in place of a methylene group.

4. **Spectrometry.** The absorbance A of a solution is defined as

$$A = \log_{10} (I_0/I)$$

in which I_0 is the incident light intensity and I is the transmitted light intensity. The absorbance is related to the molar absorption coefficient (extinction coefficient) ϵ (in $\text{M}^{-1} \text{cm}^{-1}$), concentration c (in M), and path length l (in cm) by

$$A = \epsilon lc$$

The absorption coefficient of myoglobin at 580 nm is $15,000 \text{ M}^{-1} \text{ cm}^{-1}$. What is the absorbance of a 1 mg ml^{-1} solution across a 1-cm path? What percentage of the incident light is transmitted by this solution?

Answer:

A 1 mg/ml solution of myoglobin (17.8 kd) corresponds to $5.62 \times 10^{-5} \text{ M}$. The absorbance of a 1-cm path length is 0.84, which corresponds to an I_0/I ratio of 6.96. Hence 14.4% of the incident light is transmitted.

- 5. A slow mover. Tropomyosin, a 93-kd muscle protein, sediments more slowly than does hemoglobin (65 kd). Their sedimentation coefficients are 2.6S and 4.31S, respectively. Which structural feature of tropomyosin accounts for its slow sedimentation?**

Answer:

Tropomyosin is rod shaped, whereas hemoglobin is approximately spherical.

- 6. Sedimenting spheres. What is the dependence of the sedimentation coefficient S of a spherical protein on its mass? How much more rapidly does an 80-kd protein sediment than does a 40-kd protein?**

Answer:

The frictional coefficient f and the mass m determine S . Specifically, f is proportional to r (see equation 2 on p. 83). Hence, f is proportional to $m^{1/3}$, and so S is proportional to $m^{2/3}$ (see the equation on p. 88). An 80-kd spherical protein sediments 1.59 times as rapidly as a 40-kd spherical protein.

- 7. Size estimate. The relative electrophoretic mobilities of a 30-kd protein and a 92-kd protein used as standards on an SDS-polyacrylamide gel are 0.80 and 0.41, respectively. What is the apparent mass of a protein having a mobility of 0.62 on this gel?**

Answer:

50 kd.

- 8. A new partnership? The gene encoding a protein with a single disulfide bond undergoes a mutation that changes a serine residue into a cysteine residue. You want to find out whether the disulfide pairing in this mutant is the same as in the original protein. Propose an experiment to directly answer this question.**

Answer:

The positions of disulfide bonds can be determined by diagonal electrophoresis (p. 96). The disulfide pairing is unaltered by the mutation if the off-diagonal peptides formed from the native and mutant proteins are the same.

- 9. Sorting cells. Fluorescence-activated cell sorting (FACS) is a powerful technique for separating cells according to their content of particular molecules. For example, a fluorescence-labeled antibody specific for a cell-surface protein can be used to detect cells containing such a molecule. Suppose that you want to isolate cells that possess a receptor enabling them to detect bacterial degradation products. However, you do not yet have an antibody directed against this receptor. Which fluorescencelabeled molecule would you prepare to identify such cells?**

Answer:

A fluorescent-labeled derivative of a bacterial degradation product (e.g., a formylmethionyl peptide) would bind to cells containing the receptor of interest.

10. Column choice. (a) The octapeptide AVGWRVKS was digested with the enzyme trypsin. Would ion exchange or molecular exclusion be most appropriate for separating the products? Explain. (b) Suppose that the peptide was digested with chymotrypsin. What would be the optimal separation technique? Explain.

Answer:

(a) Trypsin cleaves after arginine (R) and lysine (K), generating AVGWR, VK, and S. Because they differ in size, these products could be separated by molecular exclusion chromatography.

(b) Chymotrypsin, which cleaves after large aliphatic or aromatic R groups, generates two peptides of equal size (AVGW) and (RVKS). Separation based on size would not be effective. The peptide RVKS has two positive charges (R and K), whereas the other peptide is neutral. Therefore, the two products could be separated by ion-exchange chromatography.

11. Making more enzyme? In the course of purifying an enzyme, a researcher performs a purification step that results in an *increase* in the total activity to a value greater than that present in the original crude extract. Explain how the amount of total activity might increase.

Answer:

An inhibitor of the enzyme being purified might have been present and subsequently removed by a purification step. This would lead to an apparent increase in the total amount of enzyme present.

12. Protein purification problem. Complete the table below.

Purification procedure	Total protein (mg)	Total activity (units)	Specific activity (units mg ⁻¹)	Purification level	Yield (%)
Crude extract	20,000	4,000,000		1	100
(NH ₄) ₂ SO ₄ precipitation	5,000	3,000,000			
DEAE-cellulose chromatography	1,500	1,000,000			
Size-exclusion chromatography	500	750,000			
Affinity chromatography	45	675,000			

Answer:

See table below.

Purification procedure	Total protein (mg)	Total activity (units)	Specific activity (units/mg)	Purification level	Yield (%)
Crude extract	20,000	4,000,000	200	1	100
(NH ₄) ₂ SO ₄ precipitation	5,000	3,000,000	600	3	75
DEAE-cellulose chromatography	1,500	1,000,000	667	3.3	25
Size-exclusion chromatography	500	750,000	1,500	7.5	19
Affinity chromatography	45	675,000	15,000	75	17

Chapter Integration Problems

13. Quaternary structure. A protein was purified to homogeneity. Determination of the molecular weight by molecular exclusion chromatography yields 60 kd. Chromatography in the presence of 6 M urea yields a 30-kd species. When the chromatography is repeated in the presence of 6 M urea and 10 mM β -mercaptoethanol, a single molecular species of 15 kd results. Describe the structure of the molecule.

Answer:

Treatment with urea will disrupt noncovalent bonds. Thus the original 60-kd protein must be made of two 30-kd subunits. When these subunits are treated with urea and mercaptoethanol, a single 15-kd species results, suggesting that disulfide bonds link the 30-kd subunits.

14. Helix-coil transitions.

(a) NMR measurements have shown that poly-L-lysine is a random coil at pH 7 but becomes α helical as the pH is raised above 10. Account for this pH-dependent conformational transition.

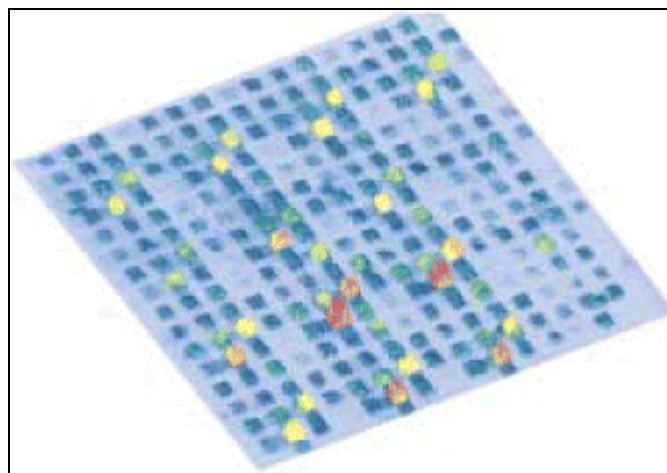
(b) Predict the pH dependence of the helix-coil transition of poly-L-glutamate.

Answer:

(a) Electrostatic repulsion between positively charged ϵ -amino groups hinders α -helix formation at pH 7. At pH 10, the side chains become deprotonated, allowing α -helix formation.

(b) Poly-L-glutamate is a random coil at pH 7 and becomes α helical below pH 4.5 because the γ -carboxylate groups become protonated.

15. Peptides on a chip. Large numbers of different peptides can be synthesized in a small area on a solid support. This high-density array can then be probed with a fluorescence-labeled protein to find out which peptides are recognized. The binding of an antibody to an array of 1024 different peptides occupying a total area the size of a thumbnail is shown in the figure below. How would you synthesize such a peptide array? [Hint: Use light instead of acid to deprotect the terminal amino group in each round of synthesis.]



Fluorescence Scan of an Array of 1024 Peptides in A 1.6-cm² Area. Each synthesis site is a 400- μ m square. A fluorescently labeled monoclonal antibody was added to the array to identify peptides that are recognized. The height and color of each square denote the fluorescence intensity. [After S. P. A. Fodor, J. O. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. *Science* 251(1991):767.]

Answer:

Light was used to direct the synthesis of these peptides. Each amino acid added to the solid support contained a photolabile protecting group instead of a *t*-Boc protecting group at its α -amino group. Illumination of selected regions of the solid support led to the release of the protecting group, which exposed the amino groups in these sites to make them reactive. The pattern of masks used in these illuminations and the sequence of reactants define the ultimate products and their locations.

Data Interpretation Problems

16. Protein sequencing I. Determine the sequence of hexapeptide based on the following data. Note: When the sequence is not known, a comma separates the amino acids. (See Table 4.3)

Amino acid composition: (2R,A,S,V,Y)

N-terminal analysis of the hexapeptide: A

Trypsin digestion: (R,A,V) and (R,S,Y)

Carboxypeptidase digestion: No digestion.

Chymotrypsin digestion: (A,R,V,Y) and (R,S)

Answer:

AVRYSR

17. Protein sequencing II. Determine the sequence of a peptide consisting of 14 amino acids on the basis of the following data.

Amino acid composition: (4S,2L,F,G,I,M,T,W,Y)

N-terminal analysis: S

Carboxypeptidase digestion: L

Trypsin digestion: (3S,2L,F,I,M,T,W) (G,K,S,Y)

Chymotrypsin digestion: (F,I,S) (G,K,L) (L,S) (M,T) (S,W) (S,Y)

N-terminal analysis of (F,I,S) peptide: S

Cyanogen bromide treatment: (2S,F,G,I,K,L,M*,T,Y) (2S,L,W)

M*, methionine detected as homoserine

Answer:

First amino acid: S

Last amino acid: L

Cyanogen bromide cleavage: M is 10th position, C-terminal residues are: (2S,L,W)

Amino-terminal residues: (G,K,S,Y), tryptic peptide, ends in K

Amino-terminal sequence: SYGK

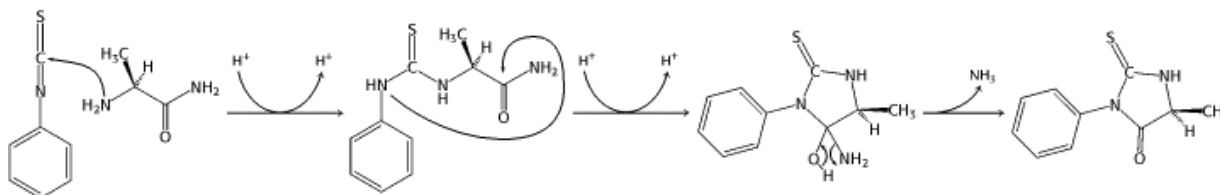
Chymotryptic peptide order: (S,Y), (G,K,L), (F,I,S), (M,T), (S,W), (S,L)

Sequence: SYGKLSIFTMSWSL

18. *Edman degradation.* Alanine amide was treated with phenyl isothiocyanate to form PTH-alanine. Write a mechanism for this reaction.

Answer:

See equation below.



Selected Readings

Where to start

M.W. Hunkapiller and L.E. Hood. 1983. Protein sequence analysis: Automated microsequencing *Science* 219: 650-659. ([PubMed](#))

B. Merrifield. 1986. Solid phase synthesis *Science* 232: 341-347. ([PubMed](#))

F. Sanger. 1988. Sequences, sequences, sequences *Annu. Rev. Biochem.* 57: 1-28. ([PubMed](#))

C. Milstein. 1980. Monoclonal antibodies *Sci. Am.* 243: (4) 66-74. ([PubMed](#))

S. Moore and W.H. Stein. 1973. Chemical structures of pancreatic ribonuclease and deoxyribonuclease *Science* 180: 458-464. ([PubMed](#))

Books

Creighton, T. E., 1993. *Proteins: Structure and Molecular Properties* (2d ed.). W. H. Freeman and Company.

Kyte, J., 1994. *Structure in Protein Chemistry*. Garland.

Van Holde, K. E., Johnson, W. C., and Ho, P.-S., 1998. *Principles of Physical Biochemistry*. Prentice Hall.

Methods in Enzymology. Academic Press. [The more than 200 volumes of this series are a treasure house of experimental procedures.]

Cantor, C. R., and Schimmel, P. R., 1980. *Biophysical Chemistry*. W. H. Freeman and Company.

Freifelder, D., 1982. *Physical Biochemistry: Applications to Biochemistry and Molecular Biology*. W. H. Freeman and Company.

Johnstone, R. A. W., 1996. *Mass Spectroscopy for Chemists and Biochemists* (2d ed.). Cambridge University Press.

Wilkins, M. R., Williams, K. L., Appel, R. D., and Hochstrasser, D. F., 1997. *Proteome Research: New Frontiers in Functional Genomics (Principles and Practice)*. Springer Verlag

Protein purification and analysis

Deutscher, M. (Ed.), 1997. *Guide to Protein Purification*. Academic Press.

Scopes, R. K., and Cantor, C., 1994. *Protein Purification: Principles and Practice* (3d ed.). Springer Verlag.

M.J. Dunn. 1997. Quantitative two-dimensional gel electrophoresis: From proteins to proteomes *Biochem. Soc. Trans.* 25: 248-254. ([PubMed](#))

R. Aebersold, G.D. Pipes, R.E. Wettenhall, H. Nika, and L.E. Hood. 1990. Covalent attachment of peptides for high sensitivity solid-phase sequence analysis *Anal. Biochem.* 187: 56-65. ([PubMed](#))

W.P. Blackstock and M.P. Weir. 1999. Proteomics: Quantitative and physical mapping of cellular proteins *Trends Biotechnol.* 17: 121-127. ([PubMed](#))

M.J. Dutt and K.H. Lee. 2000. Proteomic analysis *Curr. Opin. Biotechnol.* 11: 176-179. ([PubMed](#))

A. Pandey and M. Mann. 2000. Proteomics to study genes and genomes *Nature* 405: 837-846. ([PubMed](#))

Ultracentrifugation and mass spectrometry

- Schuster, T. M., and Laue, T. M., 1994. *Modern Analytical Ultracentrifugation*. Springer Verlag.
- D. Arnott, J. Shabanowitz, and D.F. Hunt. 1993. Mass spectrometry of proteins and peptides: Sensitive and accurate mass measurement and sequence analysis *Clin. Chem.* 39: 2005-2010. ([PubMed](#))
- B.T. Chait and S.B.H. Kent. 1992. Weighing naked proteins: Practical, high-accuracy mass measurement of peptides and proteins *Science* 257: 1885-1894. ([PubMed](#))
- I. Jardine. 1990. Molecular weight analysis of proteins *Methods Enzymol.* 193: 441-455. ([PubMed](#))
- C.G. Edmonds, J.A. Loo, R.R. Loo, H.R. Udseth, C.J. Barinaga, and R.D. Smith. 1991. Application of electrospray ionization mass spectrometry and tandem mass spectrometry in combination with capillary electrophoresis for biochemical investigations *Biochem. Soc. Trans.* 19: 943-947. ([PubMed](#))
- L. Li, R.W. Garden, and J.V. Sweedler. 2000. Single-cell MALDI: A new tool for direct peptide profiling *Trends Biotechnol.* 18: 51-160.
- D.J. Pappin. 1997. Peptide mass fingerprinting using MALDI-TOF mass spectrometry *Methods Mol. Biol.* 64: 165-173. ([PubMed](#))
- J.R. Yates and 3rd. 1998. Mass spectrometry and the age of the proteome *J. Mass Spectrom.* 33: 1-19. ([PubMed](#))

X-ray crystallography and spectroscopy

- J.P. Glusker. 1994. X-ray crystallography of proteins *Methods Biochem. Anal.* 37: 1-72. ([PubMed](#))
- J.P. Wery and R.W. Schevitz. 1997. New trends in macromolecular x-ray crystallography *Curr. Opin. Chem. Biol.* 1: 365-369. ([PubMed](#))
- A.T. Brunger. 1997. X-ray crystallography and NMR reveal complementary views of structure and dynamics *Nat. Struct. Biol.* 4 (suppl.): 862-865. ([PubMed](#))
- K. Wüthrich. 1989. Protein structure determination in solution by nuclear magnetic resonance spectroscopy *Science* 243: 45-50. ([PubMed](#))
- G.M. Clore and A.M. Gronenborn. 1991. Structures of larger proteins in solution: Three- and four-dimensional heteronuclear NMR spectroscopy *Science* 252: 1390-1399. ([PubMed](#))
- Wüthrich, K., 1986. *NMR of Proteins and Nucleic Acids*. WileyInterscience.

Monoclonal antibodies and fluorescent molecules

- G. Köhler and C. Milstein. 1975. Continuous cultures of fused cells secreting antibody of predefined specificity *Nature* 256: 495-497. ([PubMed](#))
- Goding, J. W., 1996. *Monoclonal Antibodies: Principles and Practice*. Academic Press.
- Immunology Today*, 2000. Volume 21, issue 8.
- R.Y. Tsien. 1998. The green fluorescent protein *Annu. Rev. Biochem.* 67: 509-544. ([PubMed](#))
- J.M. Kendall and M.N. Badminton. 1998. *Aequorea victoria* bioluminescence moves into an exciting era *Trends Biotechnol.* 16: 216-234. ([PubMed](#))

Chemical synthesis of proteins

K.H. Mayo. 2000. Recent advances in the design and construction of synthetic peptides: For the love of basics or just for the technology of it *Trends Biotechnol.* 18: 212-217. ([PubMed](#))

J.A. Borgia and G.B. Fields. 2000. Chemical synthesis of proteins *Trends Biotechnol.* 18: 243-251. ([PubMed](#))

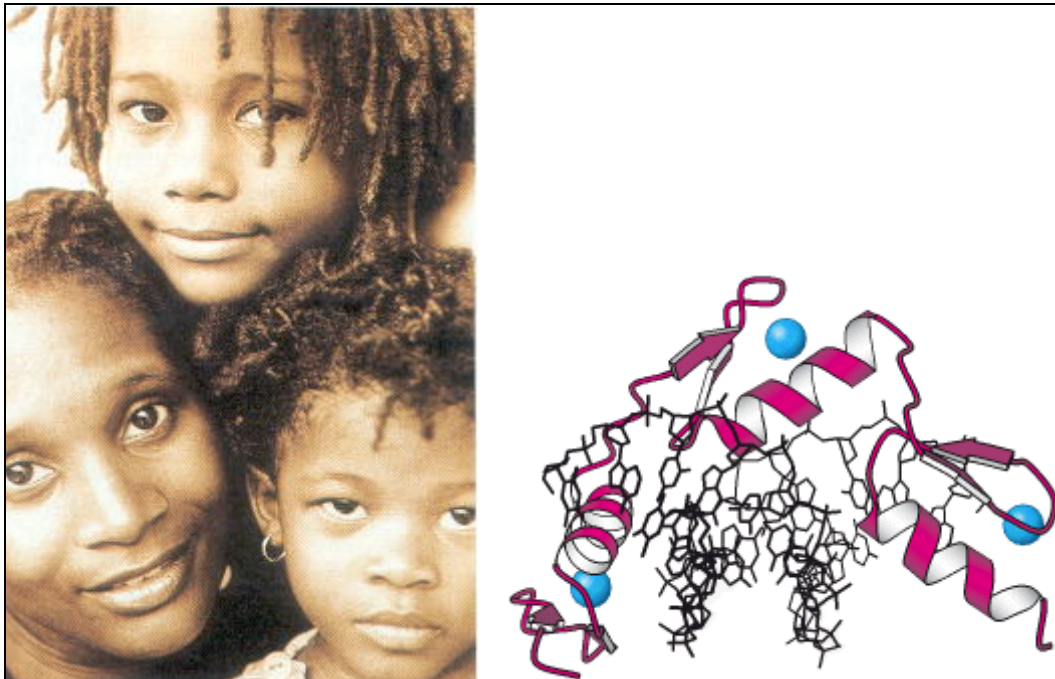
5. DNA, RNA, and the Flow of Genetic Information

DNA and RNA are long linear polymers, called nucleic acids, that carry information in a form that can be passed from one generation to the next. These macromolecules consist of a large number of linked nucleotides, each composed of a sugar, a phosphate, and a base. Sugars linked by phosphates form a common backbone, whereas the bases vary among four kinds. *Genetic information is stored in the sequence of bases along a nucleic acid chain.* The bases have an additional special property: they form specific pairs with one another that are stabilized by hydrogen bonds. The base pairing results in the formation of a double helix, a helical structure consisting of two strands. *These base pairs provide a mechanism for copying the genetic information in an existing nucleic acid chain to form a new chain.* Although RNA probably functioned as the genetic material very early in evolutionary history, the genes of all modern cells and many viruses are made of DNA. DNA is replicated by the action of DNA polymerase enzymes. These exquisitely specific enzymes copy sequences from nucleic acid templates with an error rate of less than 1 in 100 million nucleotides.

Genes specify the kinds of proteins that are made by cells, but DNA is not the direct template for protein synthesis. Rather, the templates for protein synthesis are RNA (ribonucleic acid) molecules. In particular, a class of RNA molecules called *messenger RNA* (mRNA) are the information-carrying intermediates in protein synthesis. Other RNA molecules, such as *transfer RNA* (tRNA) and *ribosomal RNA* (rRNA), are part of the protein-synthesizing machinery. All forms of cellular RNA are synthesized by RNA polymerases that take instructions from DNA templates. This process of *transcription* is followed by *translation*, the synthesis of proteins according to instructions given by mRNA templates. Thus, the flow of genetic information, or *gene expression*, in normal cells is:



This flow of information is dependent on the genetic code, which defines the relation between the sequence of bases in DNA (or its mRNA transcript) and the sequence of amino acids in a protein. The code is nearly the same in all organisms: a sequence of three bases, called a *codon*, specifies an amino acid. Codons in mRNA are read sequentially by tRNA molecules, which serve as adaptors in protein synthesis. Protein synthesis takes place on ribosomes, which are complex assemblies of rRNAs and more than 50 kinds of proteins.



Having genes in common accounts for the resemblance of a mother and her daughters. Genes must be expressed to exert an effect, and proteins regulate such expression. One such regulatory protein, a zinc-finger protein (zinc ion is blue, protein is red), is shown bound to a control or promoter region of DNA (black). [Barnaby Hall/Photonica.]

The last theme to be considered is the interrupted character of most eukaryotic genes, which are mosaics of nucleic acid sequences called *introns* and *exons*. Both are transcribed, but introns are cut out of newly synthesized RNA molecules, leaving mature RNA molecules with continuous exons. The existence of introns and exons has crucial implications for the evolution of proteins.

5.1. A Nucleic Acid Consists of Four Kinds of Bases Linked to a Sugar-Phosphate Backbone

The nucleic acids DNA and RNA are well suited to function as the carriers of genetic information by virtue of their covalent structures. These macromolecules are *linear polymers* built up from similar units connected end to end (Figure 5.1). Each monomer unit within the polymer consists of three components: a sugar, a phosphate, and a base. The sequence of bases uniquely characterizes a nucleic acid and represents a form of linear information.

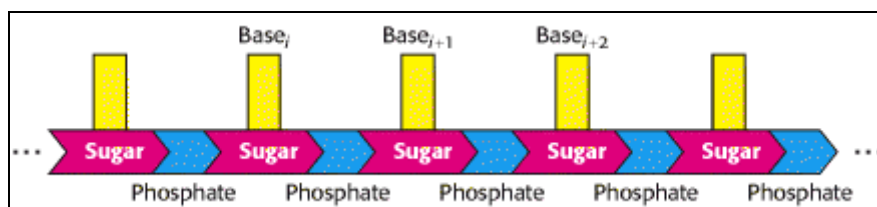


Figure 5.1. Polymeric Structure of Nucleic Acids.

5.1.1. RNA and DNA Differ in the Sugar Component and One of the Bases

The sugar in *deoxyribonucleic acid (DNA)* is *deoxyribose*. The deoxy prefix indicates that the 2' carbon atom of the sugar lacks the oxygen atom that is linked to the 2' carbon atom of *ribose* (the sugar in *ribonucleic acid*, or *RNA*), as shown in Figure 5.2. The sugars in nucleic acids are linked to one another by phosphodiester bridges. Specifically, the 3'-hydroxyl (3'-OH) group of one nucleotide is esterified to a phosphate group, which is, in turn, joined to the 5'-hydroxyl group of the adjacent sugar. The chain of sugars linked by phosphodiester bridges is referred to as the *backbone* of the nucleic acid (Figure 5.3). Whereas the backbone is constant in DNA and RNA, the bases vary from one monomer to the next. Two of the bases are derivatives of *purine* - adenine (A) and guanine (G) - and two of *pyrimidine* - cytosine (C) and thymine (T, DNA only) or uracil (U, RNA only), as shown in Figure 5.4.

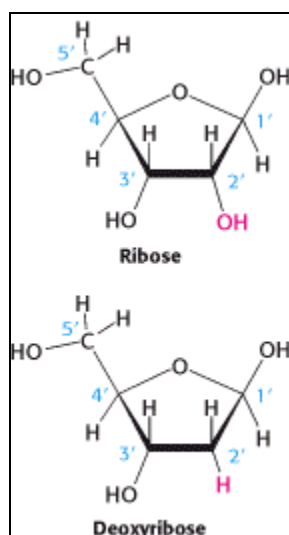


Figure 5.2. Ribose and Deoxyribose. Atoms are numbered with primes to distinguish them from atoms in bases (see Figure 5.4).

RNA, like DNA, is a long unbranched polymer consisting of nucleotides joined by 3'→5' phosphodiester bonds (see Figure 5.3). The covalent structure of RNA differs from that of DNA in two respects. As stated earlier and as indicated by its name, the sugar units in RNA are riboses rather than deoxyriboses. Ribose contains a 2'-hydroxyl group not present in deoxyribose. As a consequence, in addition to the standard 3'→5' linkage, a 2'→5' linkage is possible for RNA. This latter linkage is important in the removal of introns and the joining of exons for the formation of mature RNA (Section 28.3.4). The other difference, as already mentioned, is that one of the four major bases in RNA is uracil (U) instead of thymine (T).

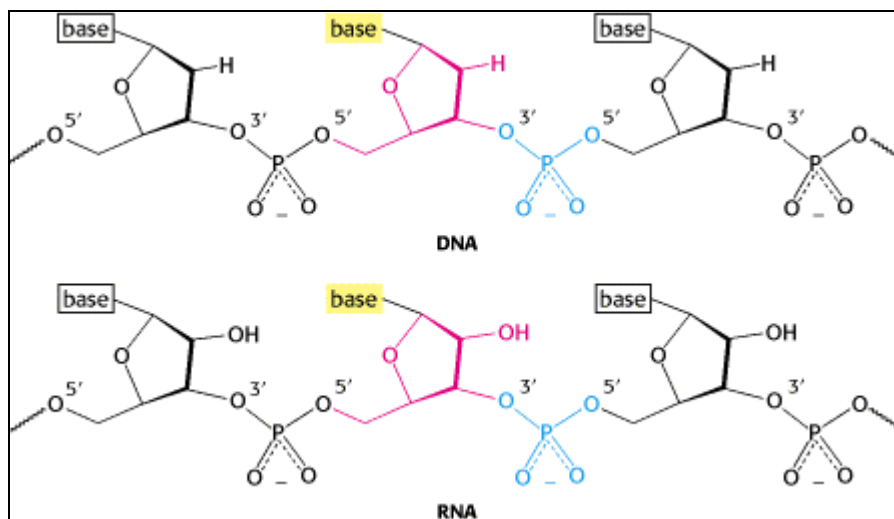


Figure 5.3. Backbones of DNA and RNA. The backbones of these nucleic acids are formed by 3'-to-5' phosphodiester linkages. A sugar unit is highlighted in red and a phosphate group in blue.

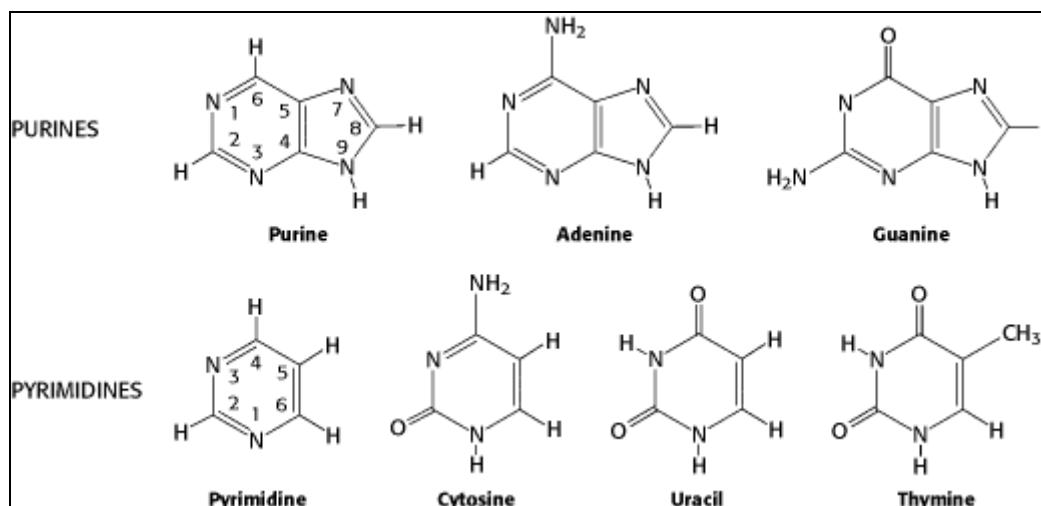


Figure 5.4. Purines and Pyrimidines. Atoms within bases are numbered without primes. Uracil instead of thymine is used in RNA.

Note that each phosphodiester bridge has a negative charge. This negative charge repels nucleophilic species such as hydroxide ion; consequently, phosphodiester linkages are much less susceptible to hydrolytic attack than are other esters such as carboxylic acid esters. This resistance is crucial for maintaining the integrity of information stored in nucleic acids. The absence of the 2'-hydroxyl group in DNA further increases its resistance to hydrolysis. The greater stability of DNA probably accounts for its use rather than RNA as the hereditary material in all modern cells and in many viruses.

5.1.2. Nucleotides Are the Monomeric Units of Nucleic Acids

A unit consisting of a base bonded to a sugar is referred to as a *nucleoside*. The four nucleoside units in RNA are called *adenosine*, *guanosine*, *cytidine*, and *uridine*, whereas those in DNA are called *deoxyadenosine*, *deoxyguanosine*, *deoxycytidine*, and *thymidine*. In each case, N-9 of a purine or N-1 of a pyrimidine is attached to C-1' of the sugar (Figure 5.5). The base lies above the plane of sugar when the structure is written in the standard orientation; that is, the configuration of the *N*-glycosidic linkage is β . A *nucleotide* is a nucleoside joined to one or more phosphate groups by an ester linkage. The most common site of esterification in naturally occurring nucleotides is the hydroxyl group attached to C-5' of the sugar. A compound formed by the attachment of a phosphate group to the C-5' of a nucleoside sugar is called a *nucleoside 5'-phosphate* or a *5'-nucleotide*. For example, ATP is *adenosine 5'-triphosphate*. Another nucleotide is deoxyguanosine 3'-monophosphate (3'-dGMP; Figure 5.6). This nucleotide differs from ATP in that it contains guanine rather than adenine, contains deoxyribose rather than ribose

(indicated by the prefix "d"), contains one rather than three phosphates, and has the phosphate esterified to the hydroxyl group in the 3' rather than the 5' position. Nucleotides are the monomers that are linked to form RNA and DNA. The four nucleotide units in DNA are called *deoxyadenylate*, *deoxyguanylate*, *deoxycytidylate*, and *deoxythymidylate*, and *thymidylate*. Note that thymidylate contains deoxyribose; by convention, the prefix deoxy is not added because thymine-containing nucleotides are only rarely found in RNA.

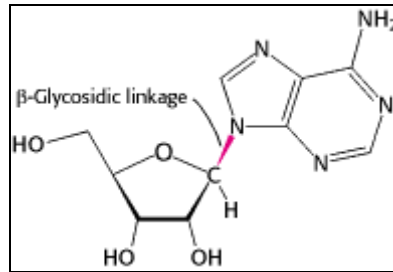


Figure 5.5. β -Glycosidic linkage in a nucleoside.

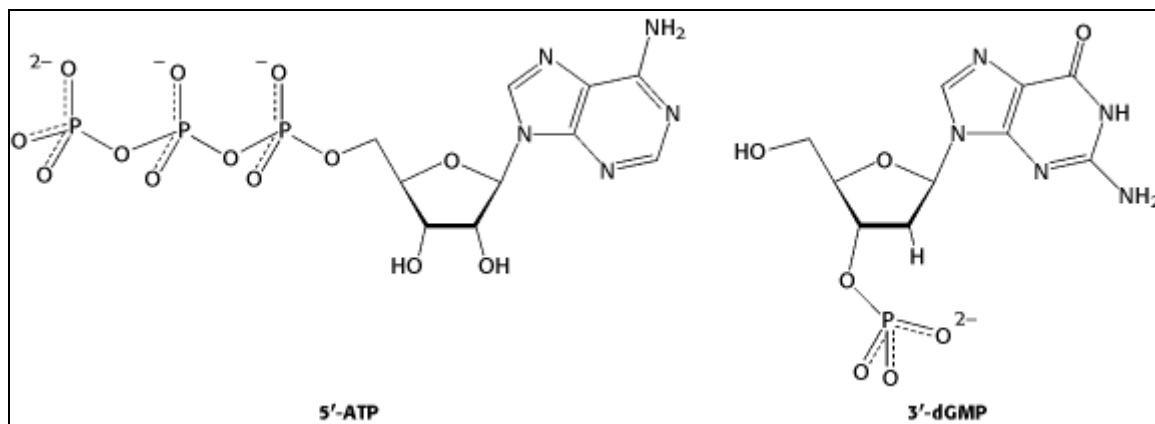


Figure 5.6. Nucleotides Adenosine 5'-triphosphate (5'-ATP) and deoxyguanosine 3'-monophosphate (3'-dGMP).

The abbreviated notations pApCpG or pACG denote a trinucleotide of DNA consisting of the building blocks deoxyadenylate monophosphate, deoxycytidylate monophosphate, and deoxyguanylate monophosphate linked by a phosphodiester bridge, where "p" denotes a phosphate group (Figure 5.7). The 5' end will often have a phosphate attached to the 5'-OH group. Note that, like a polypeptide (see Section 3.2), a DNA chain has polarity. One end of the chain has a free 5'-OH group (or a 5'-OH group attached to a phosphate), whereas the other end has a 3'-OH group, neither of which is linked to another nucleotide. By convention, the base sequence is written in the 5'-to-3' direction. Thus, the symbol ACG indicates that the unlinked 5'-OH group is on deoxyadenylate, whereas the unlinked 3'-OH group is on deoxyguanylate. Because of this polarity, ACG and GCA correspond to different compounds.

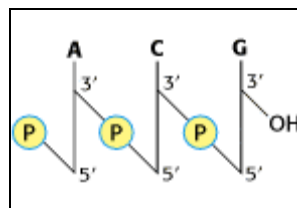


Figure 5.7. Structure of a DNA Chain. The chain has a 5' end, which is usually attached to a phosphate, and a 3' end, which is usually a free hydroxyl group.

A striking characteristic of naturally occurring DNA molecules is their length. A DNA molecule must comprise many nucleotides to carry the genetic information necessary for even the simplest organisms. For example, the DNA of a virus such as polyoma, which can cause cancer in certain organisms, is as long as 5100 nucleotides in length. We can quantify the information carrying capacity of nucleic acids in the following way. Each position can be one of four bases, corresponding to two bits of information ($2^2 = 4$). Thus, a chain of 5100 nucleotides corresponds to $2 \times 5100 = 10,200$ bits, or 1275 bytes (1 byte = 8 bits). The *E. coli* genome is a single DNA molecule consisting of two chains of 4.6 million nucleotides, corresponding to 9.2 million bits, or 1.15 megabytes, of information (Figure 5.8).



Figure 5.8. Electron Micrograph of Part of the *E. coli* genome. [Dr. Gopal Murti/Science Photo Library/Photo Researchers.]

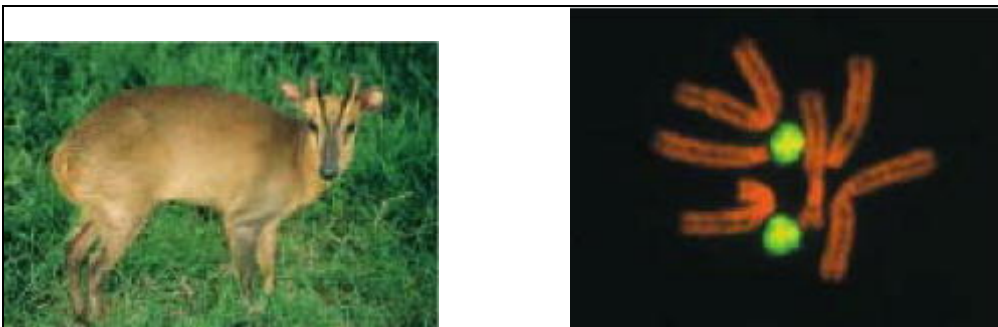


Figure 5.9. The Indian Muntjak and Its Chromosomes. Cells from a female Indian muntjak (right) contain three pairs of very large chromosomes (stained orange). The cell shown is a hybrid containing a pair of human chromosomes (stained green) for comparison. [(Left) M. Birkhead, OSF/Animals Animals. (Right) J-Y Lee, M Koi, E.J. Stanbridge, M. Oshimura, A.T Kumamoto, and A.P. Feinbert. *Nature Genetics* 7 (1994):30.]

DNA molecules from higher organisms can be much larger. The human genome comprises approximately 3 billion nucleotides, divided among 24 distinct DNA molecules (22 autosomes, x and y sex chromosomes) of different sizes. One of the largest known DNA molecules is found in the Indian muntjak, an Asiatic deer; its genome is nearly as large as the human genome but is distributed on only 3 chromosomes (Figure 5.9). The largest of these chromosomes has chains of more than 1 billion nucleotides. If such a DNA molecule could be fully extended, it would stretch more than 1 foot in length. Some plants contain even larger DNA molecules.

5.2. A Pair of Nucleic Acid Chains with Complementary Sequences Can Form a Double-Helical Structure

The covalent structure of nucleic acids accounts for their ability to carry information in the form of a sequence of bases along a nucleic acid chain. Other features of nucleic acid structure facilitate the process of *replication* - that is, the generation of two copies of a nucleic acid from one. These features depend on the ability of the bases found in nucleic acids to form *specific base pairs* in such a way that a helical structure consisting of two strands is formed. The double-helical structure of DNA facilitates the replication of the genetic material (Section 5.2.2).

5.2.1. The Double Helix Is Stabilized by Hydrogen Bonds and Hydrophobic Interactions

The existence of specific base-pairing interactions was discovered in the course of studies directed at determining the three-dimensional structure of DNA. Maurice Wilkins and Rosalind Franklin obtained x-ray diffraction photographs of fibers of DNA (Figure 5.10). The characteristics of these diffraction patterns indicated that DNA was formed of two chains that wound in a regular helical structure. From these and other data, James Watson and Francis Crick inferred a structural model for DNA that accounted for the diffraction pattern and was also the source of some remarkable insights into the functional properties of nucleic acids (Figure 5.11).

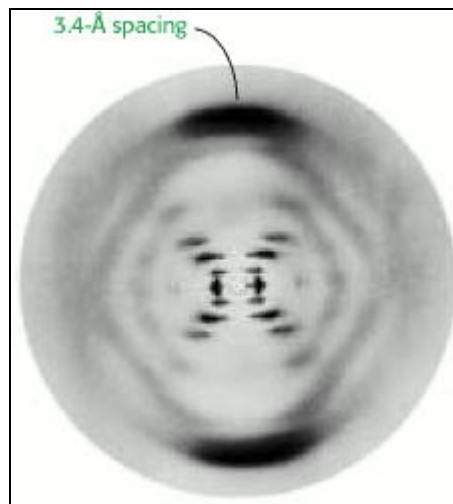


Figure 5.10. X-Ray Diffraction Photograph of a Hydrated DNA Fiber. The central cross is diagnostic of a helical structure. The strong arcs on the meridian arise from the stack of nucleotide bases, which are 3.4 Å apart. [Courtesy of Dr. Maurice Wilkins.]

The features of the Watson-Crick model of DNA deduced from the diffraction patterns are:

1. Two helical polynucleotide chains are coiled around a common axis. The chains run in opposite directions.
2. The sugar-phosphate backbones are on the outside and, therefore, the purine and pyrimidine bases lie on the inside of the helix.
3. The bases are nearly perpendicular to the helix axis, and adjacent bases are separated by 3.4 Å. The helical structure repeats every 34 Å, so there are 10 bases ($= 34 \text{ Å per repeat} / 3.4 \text{ Å per base}$) per turn of helix. There is a rotation of 36 degrees per base (360 degrees per full turn / 10 bases per turn).
4. The diameter of the helix is 20 Å.

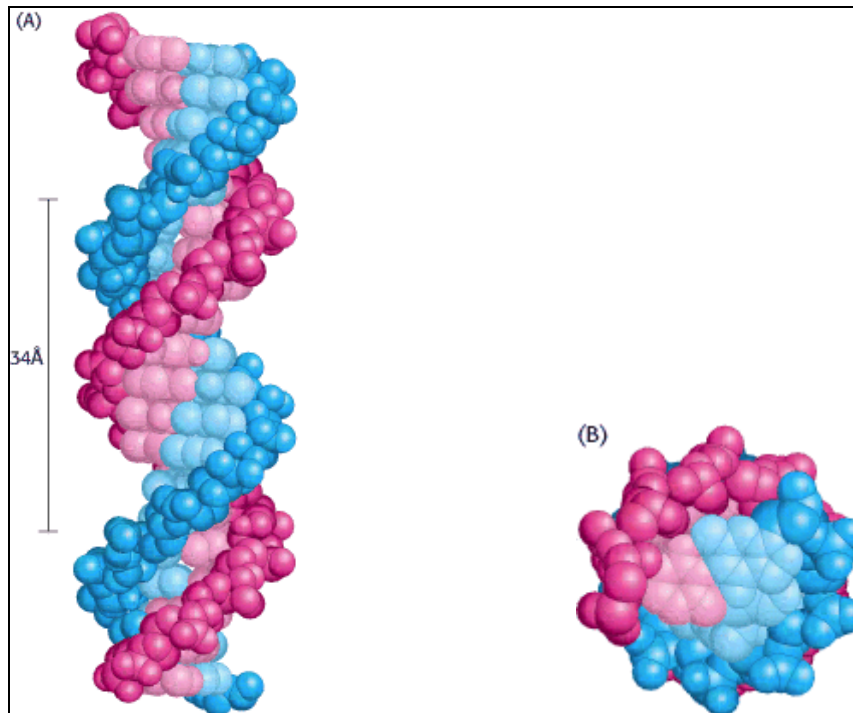


Figure 5.11. Watson-Crick Model of Double-Helical DNA. One polynucleotide chain is shown in blue and the other in red. The purine and pyrimidine bases are shown in lighter colors than the sugar-phosphate backbone. (A) Axial view. The structure repeats along the helical axis (vertical) at intervals of 34 Å, which corresponds to 10 nucleotides on each chain. (B) Radial view, looking down the helix axis.

How is such a regular structure able to accommodate an arbitrary sequence of bases, given the different sizes and shapes of the purines and pyrimidines? In attempting to answer this question, Watson and Crick discovered that guanine can be paired with cytosine and adenine with thymine to form base pairs that have essentially the same shape (Figure 5.12). These base pairs are held together by specific hydrogen bonds. This base-pairing scheme was supported by earlier studies of the base composition of DNA from different species. In 1950, Erwin Chargaff reported that the ratios of adenine to thymine and of guanine to cytosine were nearly the same in all species studied. Note in Table 5.1 that all the adenine/thymine and guanine/cytosine ratios are close to 1, whereas the adenine-to-guanine ratio varies considerably. The meaning of these equivalences was not evident until the Watson-Crick model was proposed, when it became clear that they represent an essential facet of DNA structure.

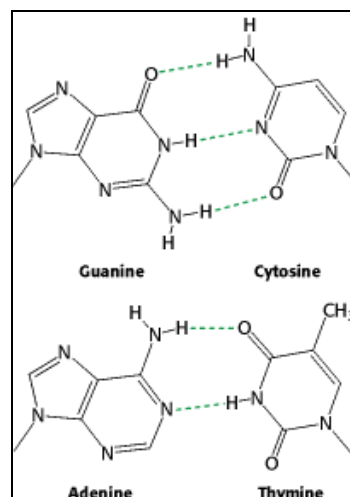


Figure 5.12. Structures of the Base Pairs Proposed by Watson and Crick.

Species	A:T	G:C	A:G
Human being	1.00	1.00	1.56
Salmon	1.02	1.02	1.43
Wheat	1.00	0.97	1.22
Yeast	1.03	1.02	1.67
<i>Escherichia coli</i>	1.09	0.99	1.05
<i>Serratia marcescens</i>	0.95	0.86	0.70

Table 5.1. Base compositions experimentally determined for a variety of organisms

The spacing of approximately 3.4 Å between nearly parallel base pairs is readily apparent in the DNA diffraction pattern (see Figure 5.10). The stacking of bases one on top of another contributes to the stability of the double helix in two ways (Figure 5.13). First, adjacent base pairs attract one another through van der Waals forces (Section 1.3.1). Energies associated with van der Waals interactions are quite small, such that typical interactions contribute from 0.5 to 1.0 kcal mol⁻¹ per atom pair. In the double helix, however, a large number of atoms are in van der Waals contact, and the net effect, summed over these atom pairs, is substantial. In addition, the double helix is stabilized by the hydrophobic effect (Section 1.3.4): base stacking, or hydrophobic interactions between the bases, results in the exposure of the more polar surfaces to the surrounding water. This arrangement is reminiscent of protein folding, where hydrophobic amino acids are interior in the protein and hydrophilic are exterior (Section 3.4). Base stacking in DNA is also favored by the conformations of the relatively rigid five-membered rings of the backbone sugars. The sugar rigidity affects both the single-stranded and the double-helical forms.

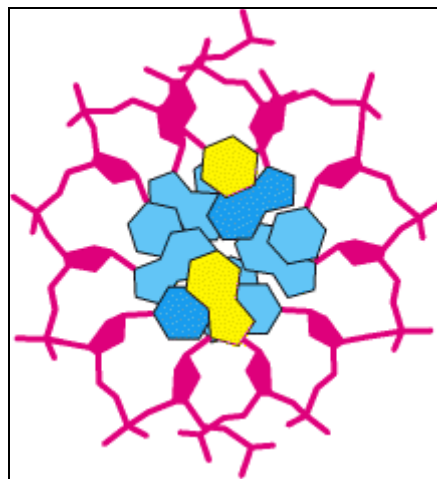


Figure 5.13. Axial View of DNA. Base pairs are stacked nearly one on top of another in the double helix.

5.2.2. The Double Helix Facilitates the Accurate Transmission of Hereditary Information

The double-helical model of DNA and the presence of specific base pairs immediately suggested how the genetic material might replicate. The sequence of bases of one strand of the double helix precisely determines the sequence of the other strand; a guanine base on one strand is always paired with a cytosine base on the other strand, and so on. Thus, separation of a double helix into its two component chains would yield two single-stranded templates onto which new double helices could be constructed, each of which would have the same sequence of bases as the parent double helix. Consequently, as DNA is replicated, one of the chains of each daughter DNA molecule would be newly synthesized, whereas the

other would be passed unchanged from the parent DNA molecule. This distribution of parental atoms is achieved by *semiconservative replication*.

Matthew Meselson and Franklin Stahl carried out a critical test of this hypothesis in 1958. They labeled the parent DNA with ^{15}N , a heavy isotope of nitrogen, to make it denser than ordinary DNA. The labeled DNA was generated by growing *E. coli* for many generations in a medium that contained $^{15}\text{NH}_4\text{Cl}$ as the sole nitrogen source. After the incorporation of heavy nitrogen was complete, the bacteria were abruptly transferred to a medium that contained ^{14}N , the ordinary isotope of nitrogen. The question asked was: What is the distribution of ^{14}N and ^{15}N in the DNA molecules after successive rounds of replication?

The distribution of ^{14}N and ^{15}N was revealed by the technique of *density-gradient equilibrium sedimentation*. A small amount of DNA was dissolved in a concentrated solution of cesium chloride having a density close to that of the DNA (1.7 g cm^{-3}). This solution was centrifuged until it was nearly at equilibrium. The opposing processes of sedimentation and diffusion created a gradient in the concentration of cesium chloride across the centrifuge cell. The result was a stable density gradient, ranging from 1.66 to 1.76 g cm^{-3} . The DNA molecules in this density gradient were driven by centrifugal force into the region where the solution's density was equal to their own. The genomic DNA yielded a narrow band that was detected by its absorption of ultraviolet light. A mixture of ^{14}N DNA and ^{15}N DNA molecules gave clearly separate bands because they differ in density by about 1% (Figure 5.14).

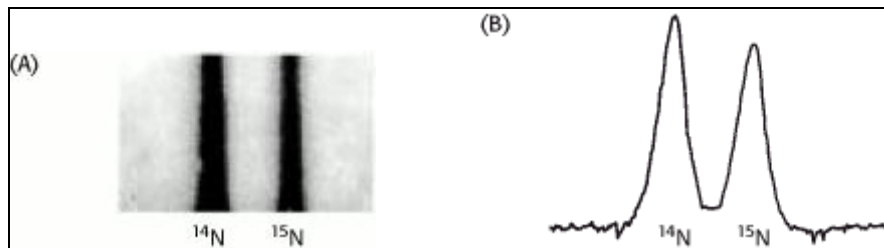


Figure 5.14. Resolution of ^{14}N DNA and ^{15}N DNA by density-gradient centrifugation. (A) Ultraviolet absorption photograph of a centrifuge cell showing the two distinct bands of DNA. (B) Densitometric tracing of the absorption photograph. [From M. Meselson and F. W. Stahl. *Proc. Natl. Acad. Sci. U.S.A.* 44(1958):671.]

DNA was extracted from the bacteria at various times after they were transferred from a ^{15}N to a ^{14}N medium and centrifuged. Analysis of these samples showed that there was a single band of DNA after one generation. The density of this band was precisely halfway between the densities of the ^{14}N DNA and ^{15}N DNA bands (Figure 5.15). The absence of ^{15}N DNA indicated that parental DNA was not preserved as an intact unit after replication. The absence of ^{14}N DNA indicated that all the daughter DNA derived some of their atoms from the parent DNA. This proportion had to be half because the density of the hybrid DNA band was halfway between the densities of the ^{14}N DNA and ^{15}N DNA bands.

After two generations, there were equal amounts of two bands of DNA. One was hybrid DNA, and the other was ^{14}N DNA. Meselson and Stahl concluded from these incisive experiments "that the nitrogen in a DNA molecule is divided equally between two physically continuous subunits; that following duplication, each daughter molecule receives one of these; and that the subunits are conserved through many duplications." Their results agreed perfectly with the Watson-Crick model for DNA replication (Figure 5.16).

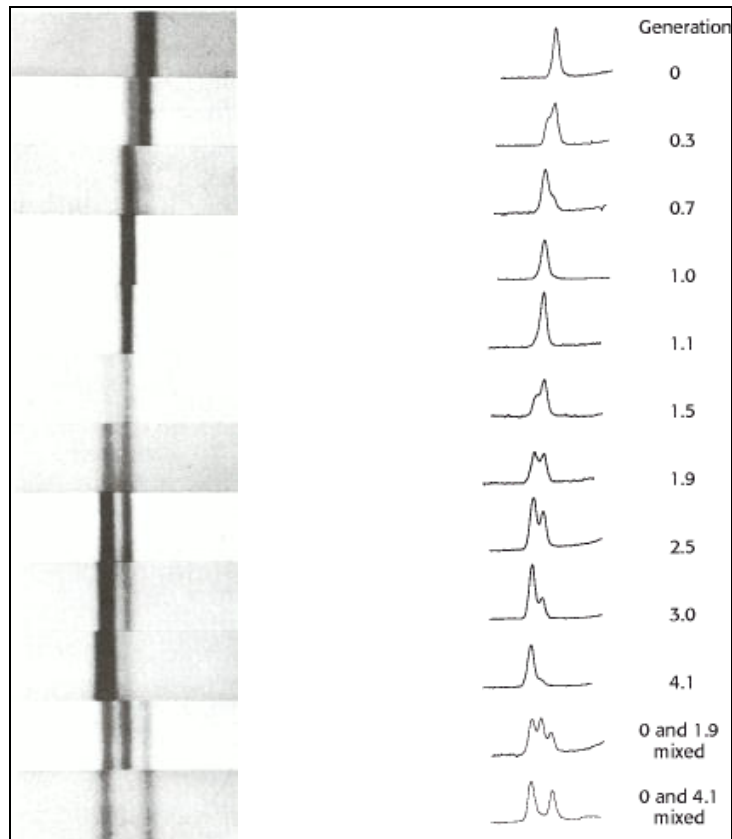


Figure 5.15. Detection of Semiconservative Replication of *E. coli* DNA by density-gradient centrifugation The position of a band of DNA depends on its content of ^{14}N and ^{15}N . After 1.0 generation, all of the DNA molecules were hybrids containing equal amounts of ^{14}N and ^{15}N . [From M. Meselson and F. W. Stahl. *Proc. Natl. Acad. Sci. U.S.A.* 44(1958):671.]

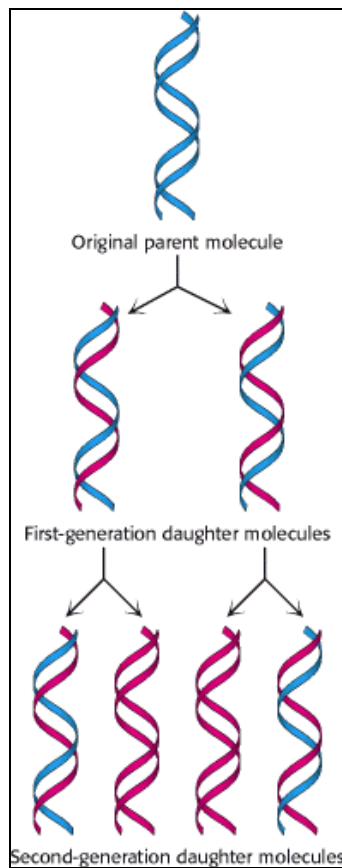


Figure 5.16. Diagram of Semiconservative Replication. Parental DNA is shown in blue and newly synthesized DNA in red. [After M. Meselson and F. W. Stahl. *Proc. Natl. Acad. Sci. U.S.A.* 44(1958):671.]

5.2.3. The Double Helix Can Be Reversibly Melted

During DNA replication and other processes, the two strands of the double helix must be separated from one another, at least in a local region. In the laboratory, the double helix can be disrupted by heating a solution of DNA. The heating disrupts the hydrogen bonds between base pairs and thereby causes the strands to separate. The dissociation of the double helix is often called *melting* because it occurs relatively abruptly at a certain temperature. The *melting temperature* (T_m) is defined as the temperature at which half the helical structure is lost. Strands may also be separated by adding acid or alkali to ionize the nucleotide bases and disrupt base pairing.

Stacked bases in nucleic acids absorb less ultraviolet light than do unstacked bases, an effect called *hypochromism*. Thus, the melting of nucleic acids is easily followed by monitoring their absorption of light, which peaks at a wavelength of 260 nm (Figure 5.17).

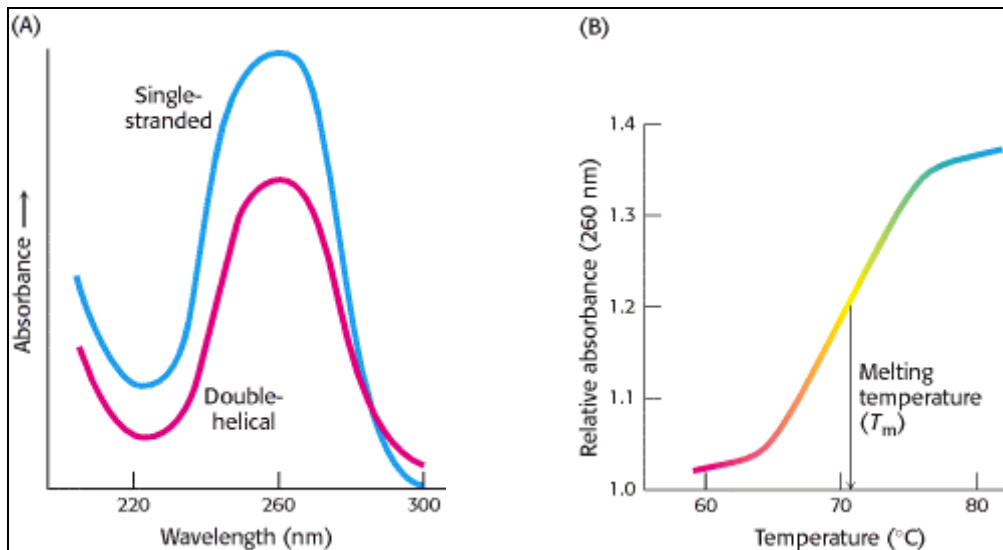


Figure 5.17. Hypochromism. (A) Single-stranded DNA absorbs light more effectively than does double-helical DNA. (B) The absorbance of a DNA solution at a wavelength of 260 nm increases when the double helix is melted into single strands.

Separated complementary strands of nucleic acids spontaneously reassociate to form a double helix when the temperature is lowered below T_m . This renaturation process is sometimes called *annealing*. The facility with which double helices can be melted and then reassociated is crucial for the biological functions of nucleic acids. Of course, inside cells, the double helix is not melted by the addition of heat. Instead, proteins called *helicases* use chemical energy (from ATP) to disrupt the structure of double-stranded nucleic acid molecules.

The ability to reversibly melt and reanneal DNA in the laboratory provides a powerful tool for investigating sequence similarity as well as gene structure and expression. For instance, DNA molecules from two different organisms can be melted and allowed to reanneal or *hybridize* in the presence of each other. If the sequences are similar, hybrid DNA duplexes, with DNA from each organism contributing a strand of the double helix, can form. Indeed, the degree of hybridization is an indication of the relatedness of the genomes and hence the organisms. Similar hybridization experiments with RNA and DNA can locate genes in a cell's DNA that correspond to a particular RNA. We will return to this important technique in Chapter 6.

5.2.4. Some DNA Molecules Are Circular and Supercoiled

The DNA molecules in human chromosomes are linear. However, electron microscopic and other studies have shown that intact DNA molecules from some other organisms are circular (Figure 5.18A). The term *circular* refers to the continuity of the DNA chains, not to their geometrical form. DNA molecules inside cells necessarily have a very compact shape. Note that the *E. coli* chromosome, fully extended, would be about 1000 times as long as the greatest diameter of the bacterium.

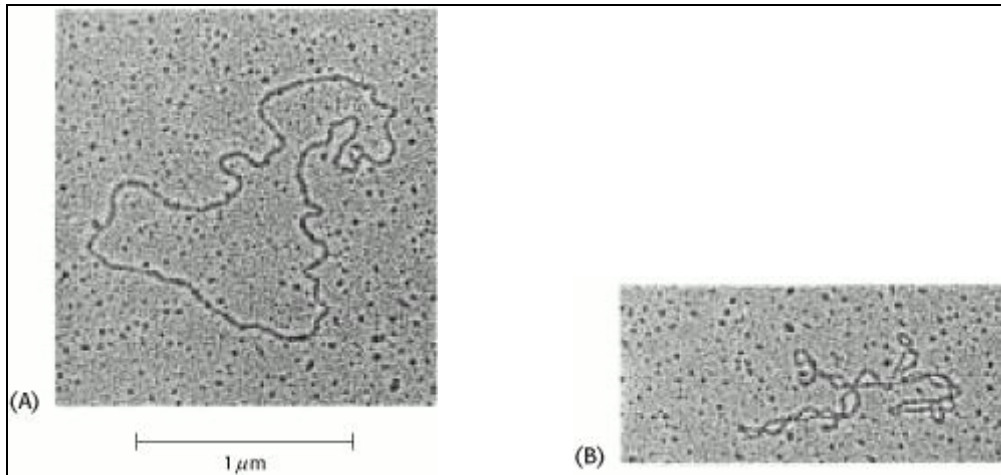


Figure 5.18. Electron Micrographs of Circular DNA from Mitochondria. (A) Relaxed form. (B) Supercoiled form. [Courtesy of Dr. David Clayton.]

A new property appears in the conversion of a linear DNA molecule into a closed circular molecule. The axis of the double helix can itself be twisted into a *superhelix* (Figure 5.18B). A circular DNA molecule without any superhelical turns is known as a *relaxed molecule*. Supercoiling is biologically important for two reasons. First, a *supercoiled DNA molecule has a more compact shape than does its relaxed counterpart*. Second, *supercoiling may hinder or favor the capacity of the double helix to unwind and thereby affects the interactions between DNA and other molecules*. These topological features of DNA will be considered further in Section 27.3.

5.2.5. Single-Stranded Nucleic Acids Can Adopt Elaborate Structures

Single-stranded nucleic acids often fold back on themselves to form well-defined structures. Early in evolutionary history, nucleic acids, particularly RNA, may have adopted complex and diverse structures both to store genetic information and to catalyze its transmission (Section 2.2.2). Such structures are also important in all modern organisms in entities such as the ribosome, a large complex of RNAs and proteins on which proteins are synthesized.

The simplest and most common structural motif formed is a *stem-loop*, created when two complementary sequences within a single strand come together to form double-helical structures (Figure 5.19). In many cases, these double helices are made up entirely of Watson-Crick base pairs. In other cases, however, the structures include mismatched or unmatched (bulged) bases. Such mismatches destabilize the local structure but introduce deviations from the standard double-helical structure that can be important for higher-order folding and for function (Figure 5.20).

Single-stranded nucleic acids can adopt structures more complex than simple stem-loops through the interaction of more widely separated bases. Often, three or more bases may interact to stabilize these structures. In such cases, hydrogen-bond donors and acceptors that ordinarily participate in Watson-Crick base pairs may participate in hydrogen bonds of nonstandard pairings. Metal ions such as magnesium ion (Mg^{2+}) often assist in the stabilization of these more elaborate structures.

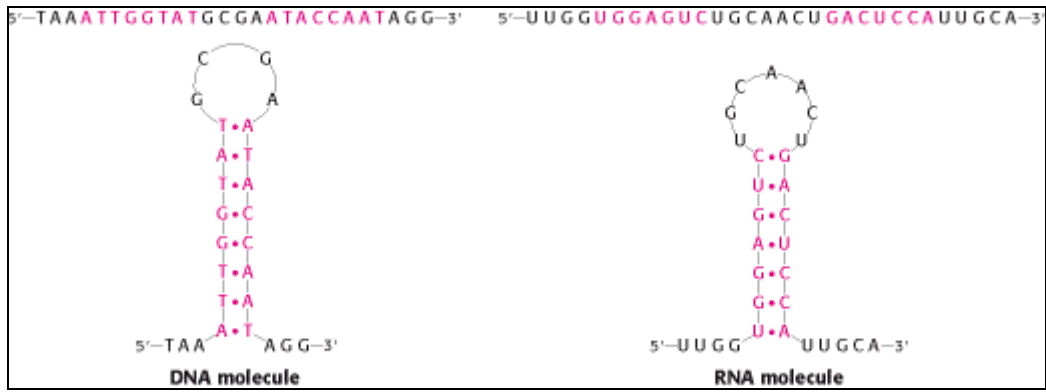


Figure 5.19. Stem-Loop Structures. Stem-loop structures may be formed from single-stranded DNA and RNA molecules.

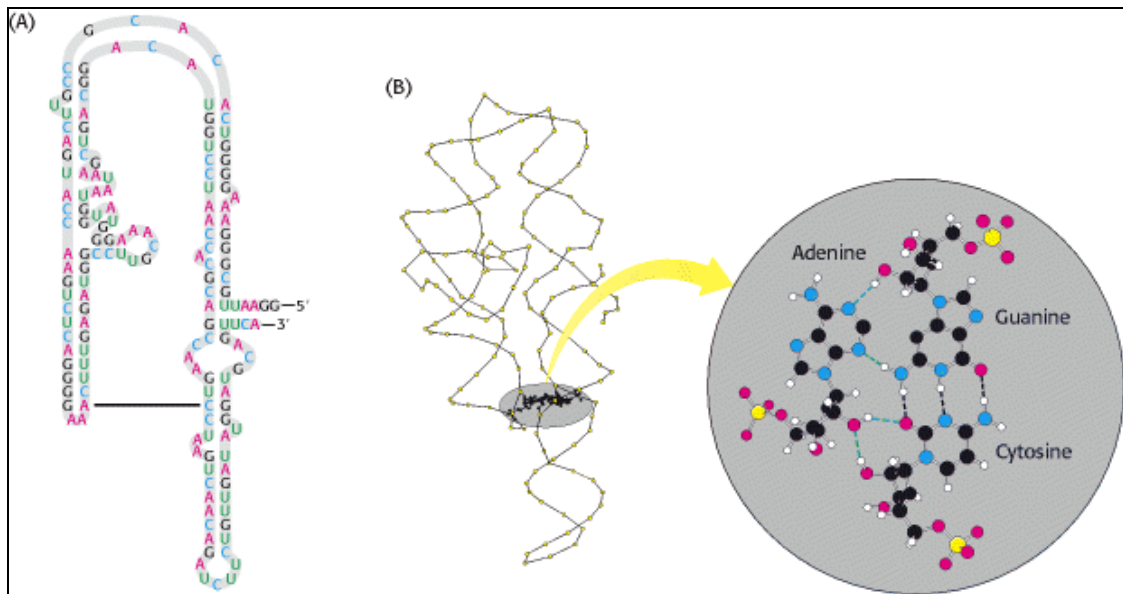


Figure 5.20. Complex Structure of an RNA Molecule. A single-stranded RNA molecule may fold back on itself to form a complex structure. (A) The nucleotide sequence showing Watson-Crick base pairs and other nonstandard base pairings in stem-loop structures. (B) The three-dimensional structure and one important long-range interaction between three bases. Hydrogen bonds within the Watson-Crick base pair are shown as dashed black lines; additional hydrogen bonds are shown as dashed green lines

5.3. DNA Is Replicated by Polymerases that Take Instructions from Templates

We now turn to the molecular mechanism of DNA replication. The full replication machinery in cells comprises more than 20 proteins engaged in intricate and coordinated interplay. In 1958, Arthur Kornberg and his colleagues isolated the first known of the enzymes, called DNA polymerases, that promote the formation of the bonds joining units of the DNA backbone.

5.3.1. DNA Polymerase Catalyzes Phosphodiester-Bond Formation

DNA polymerases catalyze the step-by-step addition of deoxyribonucleotide units to a DNA chain (Figure 5.21). Importantly, *the new DNA chain is assembled directly on a preexisting DNA template*. The reaction catalyzed, in its simplest form, is:



where dNTP stands for any deoxyribonucleotide and PP_i is a pyrophosphate molecule. The template can be a single strand of DNA or a double strand with one of the chains broken at one or more sites. If single stranded, the template DNA must be bound to a *primer* strand having a free 3'-hydroxyl group. The reaction also requires all four activated precursors - that is, the deoxynucleoside 5'-triphosphates dATP, dGTP, dTTP, and dCTP - as well as Mg^{2+} ion.

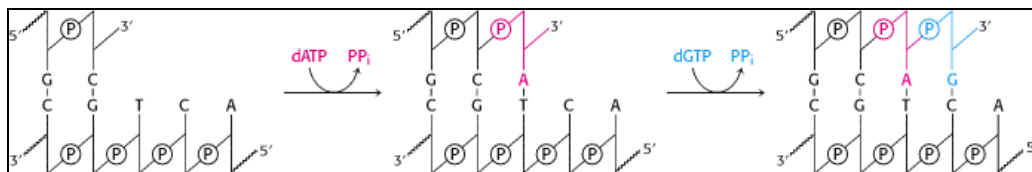


Figure 5.21. Polymerization Reaction Catalyzed by DNA Polymerases.

The chain-elongation reaction catalyzed by DNA polymerases is a nucleophilic attack by the 3'-hydroxyl group of the primer on the innermost phosphorus atom of the deoxynucleoside triphosphate (Figure 5.22). A phosphodiester bridge forms with the concomitant release of pyrophosphate. The subsequent hydrolysis of pyrophosphate by pyrophosphatase, a ubiquitous enzyme, helps drive the polymerization forward. Elongation of the DNA chain proceeds in the 5'-to-3' direction.

DNA polymerases catalyze the formation of a phosphodiester bond efficiently only if the base on the incoming nucleoside triphosphate is complementary to the base on the template strand. Thus, DNA polymerase is a *template-directed enzyme* that synthesizes a product with a base sequence complementary to that of the template. Many DNA polymerases also have a separate nuclease activity that allows them to correct mistakes in DNA by using a different reaction to remove mismatched nucleotides. These properties of DNA polymerases contribute to the remarkably high fidelity of DNA replication, which has an error rate of less than 10^{-8} per base pair.

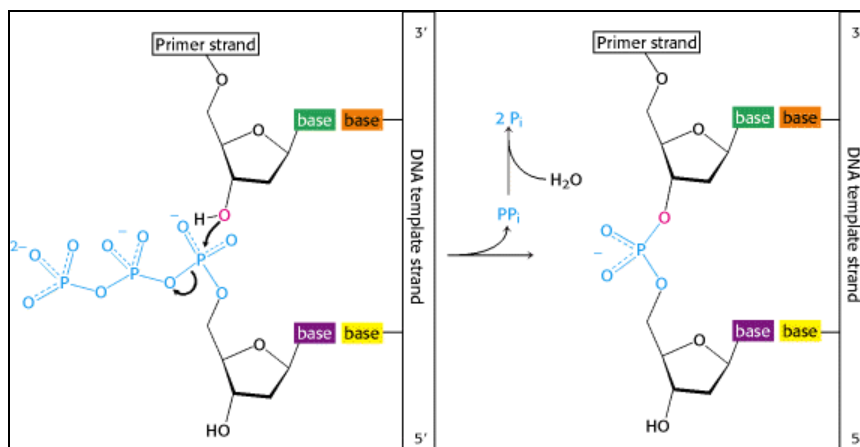


Figure 5.22. DNA Replication. The formation of a phosphodiester bridge is catalyzed by DNA polymerases.

5.3.2. The Genes of Some Viruses Are Made of RNA

Genes in all cellular organisms are made of DNA. The same is true for some viruses, but for others the genetic material is RNA. Viruses are genetic elements enclosed in protein coats that can move from one cell to another but are not capable of independent growth. One well-studied example of an RNA virus is the tobacco mosaic virus, which infects the leaves of tobacco plants. This virus consists of a single strand of RNA (6930 nucleotides) surrounded by a protein coat of 2130 identical subunits. An RNA-directed RNA polymerase catalyzes the replication of this viral RNA.

Another important class of RNA virus comprises the *retroviruses*, so called because the genetic information flows from RNA to DNA rather than from DNA to RNA. This class includes human immunodeficiency virus 1 (HIV-1), the cause of AIDS, as well as a number of RNA viruses that produce tumors in susceptible animals. Retrovirus particles contain two copies of a single-stranded RNA molecule. On entering the cell, the RNA is copied into DNA through the action of a viral enzyme called *reverse transcriptase* (Figure 5.23). The resulting double-helical DNA version of the viral genome can become incorporated into the chromosomal DNA of the host and is replicated along with the normal cellular DNA. At a later time, the integrated viral genome is expressed to form viral RNA and viral proteins, which assemble into new virus particles.

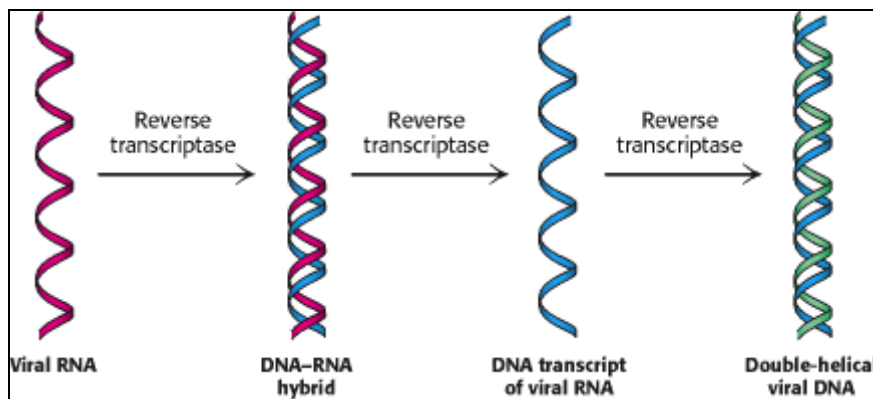


Figure 5.23. Flow of Information from RNA to DNA in Retroviruses. The RNA genome of a retrovirus is converted into DNA by reverse transcriptase, an enzyme brought into the cell by the infecting virus particle. Reverse transcriptase catalyzes the synthesis of a complementary DNA strand, the digestion of the RNA, and the subsequent synthesis of the DNA strand.

Note that RNA viruses are not vestiges of the RNA world. Instead, fragments of RNA in these viruses have evolved to encode their protein coats and other structures needed for transferring from cell to cell and replicating.

5.4. Gene Expression Is the Transformation of DNA Information Into Functional Molecules

The information stored as DNA becomes useful when it is expressed in the production of RNA and proteins. This rich and complex topic is the subject of several chapters later in this book, but here we introduce the basics of gene expression. DNA can be thought of as archival information, stored and manipulated judiciously to minimize damage (mutations). It is expressed in two steps. First, an RNA copy is made. An RNA molecule that encodes proteins can be thought of as a photocopy of the original information - it can be made in multiple copies, used, and then disposed of. Second, an RNA molecule can be further thought of as encoding directions for protein synthesis that must be translated to be of use. The information in messenger RNA is translated into a functional protein. Other types of RNA molecules exist to facilitate this translation. We now examine the transcription of DNA information into RNA, the translation of RNA information into protein, and the genetic code that links nucleotide sequence with amino acid sequence.

5.4.1. Several Kinds of RNA Play Key Roles in Gene Expression

Cells contain several kinds of RNA (Table 5.2).

Type	Relative amount (%)	Sedimentation coefficient (S)	Mass (kd)	Number of nucleotides
Ribosomal RNA (rRNA)	80	23	1.2×10^3	3700
		16	0.55×10^3	1700
		5	3.6×10^1	120
Transfer RNA (tRNA)	15	4	2.5×10^1	75
Messenger RNA (mRNA)	5	Heterogeneous		

Table 5.2. RNA molecules in *E. coli*

1. Messenger RNA is the template for protein synthesis or *translation*. An mRNA molecule may be produced for each gene or group of genes that is to be expressed in *E. coli*, whereas a distinct mRNA is produced for each gene in eukaryotes. Consequently, mRNA is a heterogeneous class of molecules. In *E. coli*, the average length of an mRNA molecule is about 1.2 kilobases (kb).

Kilobase (kb)

A unit of length equal to 1000 base pairs of a double-stranded nucleic acid molecule (or 1000 bases of a single-stranded molecule).

One kilobase of double-stranded DNA has a contour length of $0.34 \mu\text{m}$ and a mass of about 660 kd.

2. Transfer RNA carries amino acids in an activated form to the ribosome for peptide-bond formation, in a sequence dictated by the mRNA template. There is at least one kind of tRNA for each of the 20 amino acids. Transfer RNA consists of about 75 nucleotides (having a mass of about 25 kd), which makes it the smallest of the RNA molecules.

3. Ribosomal RNA (rRNA), the major component of ribosomes, plays both a catalytic and a structural role in protein synthesis (Section 29.3.1). In *E. coli*, there are three kinds of rRNA, called 23S, 16S, and 5S

RNA because of their sedimentation behavior. One molecule of each of these species of rRNA is present in each ribosome.

Ribosomal RNA is the most abundant of the three types of RNA. Transfer RNA comes next, followed by messenger RNA, which constitutes only 5% of the total RNA. Eukaryotic cells contain additional small RNA molecules. *Small nuclear RNA* (snRNA) molecules, for example, participate in the splicing of RNA exons. A small RNA molecule in the cytosol plays a role in the targeting of newly synthesized proteins to intracellular compartments and extracellular destinations.

5.4.2. All Cellular RNA Is Synthesized by RNA Polymerases

The synthesis of RNA from a DNA template is called *transcription* and is catalyzed by the enzyme *RNA polymerase* (Figure 5.24). RNA polymerase requires the following components:

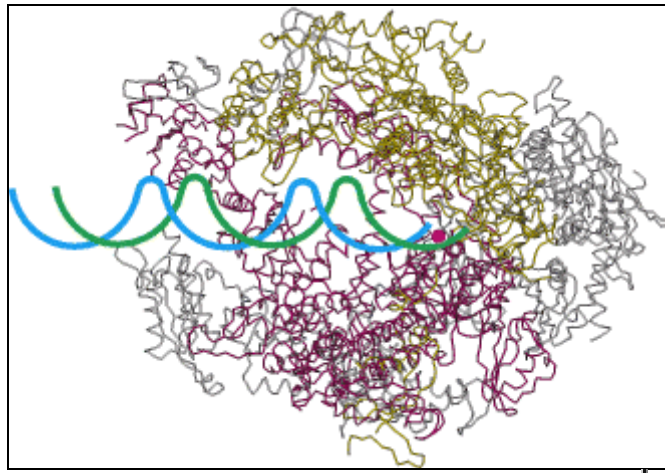
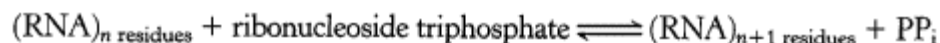


Figure 5.24. RNA Polymerase. A large enzyme comprising many subunits including β (red) and β' (blue), which form a "claw" that holds the DNA to be transcribed. The active site includes a Mg^{2+} ion at the center of the structure.

1. *A template.* The preferred template is *double-stranded DNA*. Single-stranded DNA also can serve as a template. RNA, whether single or double stranded, is not an effective template; nor are RNA-DNA hybrids.
2. *Activated precursors.* All four *ribonucleoside triphosphates* - ATP, GTP, UTP, and CTP - are required.
3. *A divalent metal ion.* Mg^{2+} or Mn^{2+} are effective.

RNA polymerase catalyzes the initiation and elongation of RNA chains. The reaction catalyzed by this enzyme is:



The synthesis of RNA is like that of DNA in several respects (Figure 5.25). First, the direction of synthesis is $5' \rightarrow 3'$. Second, the mechanism of elongation is similar: the $3'$ -OH group at the terminus of the growing chain makes a nucleophilic attack on the innermost phosphate of the incoming nucleoside triphosphate. Third, the synthesis is driven forward by the hydrolysis of pyrophosphate. In contrast with DNA polymerase, however, RNA polymerase does not require a primer. In addition, RNA polymerase lacks the nuclease capability used by DNA polymerase to excise mismatched nucleotides.

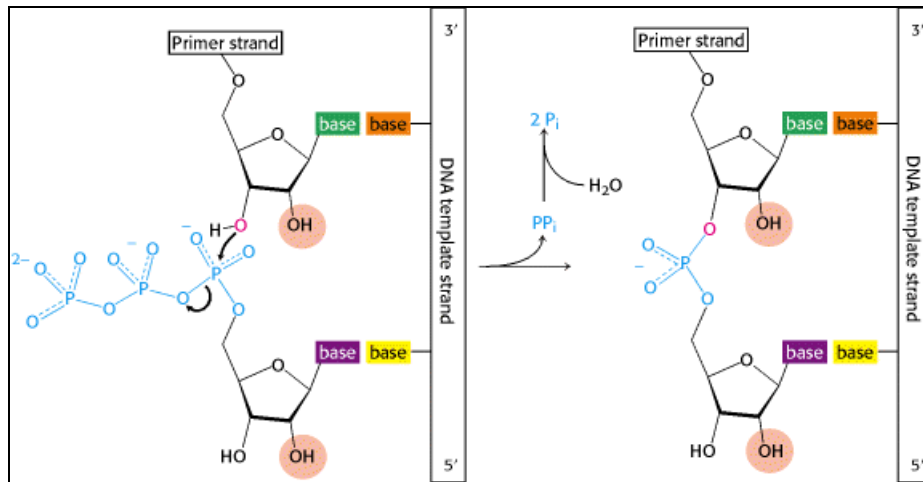


Figure 5.25. Transcription Mechanism of the Chain-Elongation Reaction Catalyzed by RNA Polymerase.

All three types of cellular RNA - mRNA, tRNA, and rRNA - are synthesized in *E. coli* by the same RNA polymerase according to instructions given by a DNA template. In mammalian cells, there is a division of labor among several different kinds of RNA polymerases. We shall return to these RNA polymerases in Chapter 28.

5.4.3. RNA Polymerases Take Instructions from DNA Templates

RNA polymerase, like the DNA polymerases described earlier, takes instructions from a DNA template. The earliest evidence was the finding that the *base composition* of newly synthesized RNA is the complement of that of the DNA template strand, as exemplified by the RNA synthesized from a template of single-stranded ϕ X174 DNA (Table 5.3). *Hybridization experiments* also revealed that RNA synthesized by RNA polymerase is complementary to its DNA template. In these experiments, DNA is melted and allowed to reassociate in the presence of mRNA. RNA-DNA hybrids will form if the RNA and DNA have complementary sequences. The strongest evidence for the fidelity of transcription came from base-sequence studies showing that the RNA sequence is the precise complement of the DNA template sequence (Figure 5.26).



Figure 5.26. Complementarity between mRNA and DNA. The base sequence of mRNA (red) is the complement of that of the DNA template strand (blue). The sequence shown here is from the tryptophan operon, a segment of DNA containing the genes for five enzymes that catalyze the synthesis of tryptophan. The other strand of DNA (black) is called the coding strand because it has the same sequence as the RNA transcript except for thymine (T) in place of uracil (U).

DNA template (plus strand of ϕ X174)	RNA	product
A 25	25	U
T 33	32	A
G 24	23	C
C 18	20	G

Table 5.3. Base composition (percentage) of RNA synthesized from a viral DNA template

5.4.4. Transcription Begins near Promoter Sites and Ends at Terminator Sites

RNA polymerase must detect and transcribe discrete genes from within large stretches of DNA. What marks the beginning of a transcriptional unit? DNA templates contain regions called *promoter sites* that specifically bind RNA polymerase and determine where transcription begins. In bacteria, two sequences on the 5' (upstream) side of the first nucleotide to be transcribed function as promoter sites (Figure 5.27A). One of them, called the *Pribnow box*, has the consensus sequence TATAAT and is centered at -10 (10 nucleotides on the 5' side of the first nucleotide transcribed, which is denoted by +1). The other, called the *-35 region*, has the consensus sequence TTGACA. The first nucleotide transcribed is usually a purine.

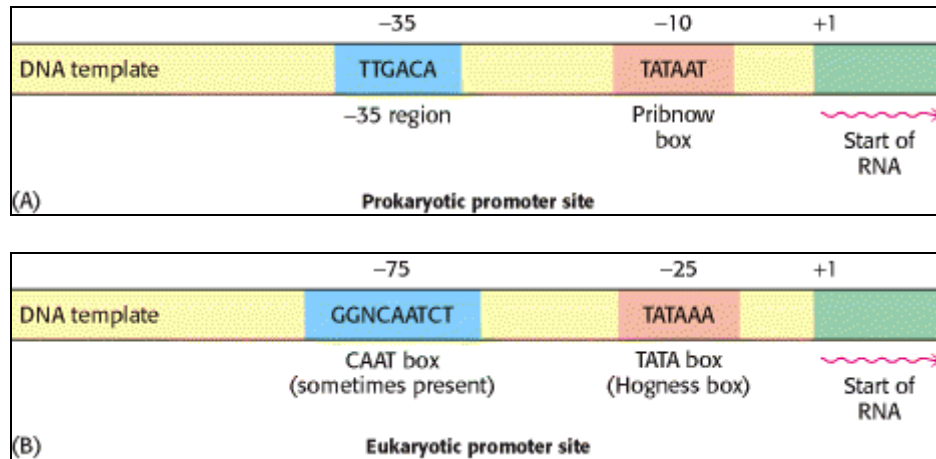


Figure 5.27. Promoter Sites for Transcription. Promoter sites are required for the initiation of transcription in both (A) prokaryotes and (B) eukaryotes. Consensus sequences are shown. The first nucleotide to be transcribed is numbered +1. The adjacent nucleotide on the 5' side is numbered -1. The sequences shown are those of the coding strand of DNA.

Consensus sequence

The base sequences of promoter sites are not all identical. However, they do possess common features, which can be represented by an idealized consensus sequence. Each base in the consensus sequence TATAAT is found in a majority of prokaryotic promoters. Nearly all promoter sequences differ from this consensus sequence at only one or two bases.

Eukaryotic genes encoding proteins have promoter sites with a TATAAA consensus sequence, called a *TATA box* or a *Hogness box*, centered at about -25 (Figure 5.27B). Many eukaryotic promoters also have a *CAAT box* with a GGNCAATCT consensus sequence centered at about -75. Transcription of eukaryotic genes is further stimulated by *enhancer sequences*, which can be quite distant (as many as several kilobases) from the start site, on either its 5' or its 3' side.

RNA polymerase proceeds along the DNA template, transcribing one of its strands until it reaches a terminator sequence. This sequence encodes a termination signal, which in *E. coli* is a *base-paired hairpin* on the newly synthesized RNA molecule (Figure 5.28). This hairpin is formed by base pairing of self-complementary sequences that are rich in G and C. Nascent RNA spontaneously dissociates from RNA polymerase when this hairpin is followed by a string of U residues. Alternatively, RNA synthesis can be terminated by the action of *rho*, a protein. Less is known about the termination of transcription in eukaryotes. A more detailed discussion of the initiation and termination of transcription will be given in Chapter 28. The important point now is that *discrete start and stop signals for transcription are encoded in the DNA template.*

In eukaryotes, the mRNA is modified after transcription (Figure 5.29). A "cap" structure is attached to the 5' end, and a sequence of adenylates the poly(A) tail is added to the 3' end. These modifications will be presented in detail in Section 28.3.1.

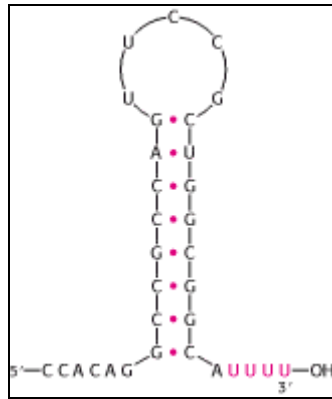


Figure 5.28. Base Sequence of the 3' end of an mRNA transcript in *E. coli*. A stable hairpin structure is followed by a sequence of uridine (U) residues.

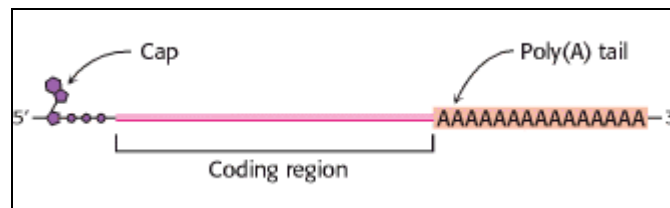


Figure 5.29. Modification of mRNA. Messenger RNA in eukaryotes is modified after transcription. A nucleotide "cap" structure is added to the 5' end, and a poly(A) tail is added at the 3' end.

5.4.5. Transfer RNA Is the Adaptor Molecule in Protein Synthesis

We have seen that mRNA is the template for protein synthesis. How then does it direct amino acids to become joined in the correct sequence to form a protein? In 1958, Francis Crick wrote:

RNA presents mainly a sequence of sites where hydrogen bonding could occur. One would expect, therefore, that whatever went onto the template in a *specific* way did so by forming hydrogen bonds. It is therefore a natural hypothesis that the amino acid is carried to the template by an adaptor molecule, and that the adaptor is the part that actually fits onto the RNA. In its simplest form, one would require twenty adaptors, one for each amino acid.

This highly innovative hypothesis soon became established as fact. *The adaptor in protein synthesis is transfer RNA*. The structure and reactions of these remarkable molecules will be considered in detail in Chapter 29. For the moment, it suffices to note that tRNA contains an *aminoacid attachment site* and a *template-recognition site*. A tRNA molecule carries a specific amino acid in an activated form to the site of protein synthesis. The carboxyl group of this amino acid is esterified to the 3'- or 2'-hydroxyl group of the ribose unit at the 3' end of the tRNA chain (Figure 5.30). The joining of an amino acid to a tRNA molecule to form an *aminoacyl-tRNA* is catalyzed by a specific enzyme called an *aminoacyl-tRNA synthetase* (or *activating enzyme*). This esterification reaction is driven by ATP. There is at least one specific synthetase for each of the 20 amino acids. The template-recognition site on tRNA is a sequence of three bases called an *anticodon* (Figure 5.31). The anticodon on tRNA recognizes a complementary sequence of three bases, called a *codon*, on mRNA.

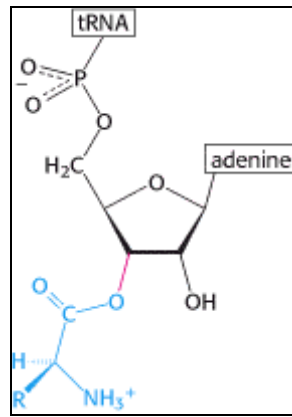


Figure 5.30. Attachment of an Amino Acid to a tRNA Molecule. The amino acid (shown in blue) is esterified to the 3'-hydroxyl group of the terminal adenosine of tRNA.

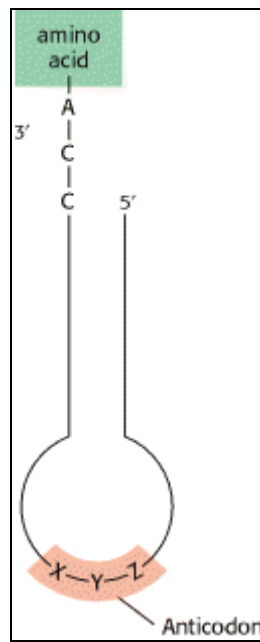


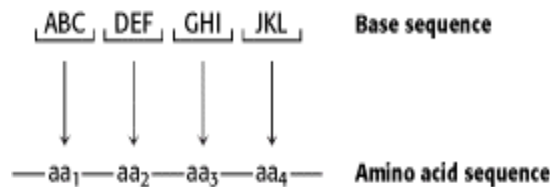
Figure 5.31. Symbolic Diagram of an Aminoacyl-tRNA. The amino acid is attached at the 3' end of the RNA. The anticodon is the template-recognition site.

5.5. Amino Acids Are Encoded by Groups of Three Bases Starting from a Fixed Point

The *genetic code* is the relation between the sequence of bases in DNA (or its RNA transcripts) and the sequence of amino acids in proteins. Experiments by Francis Crick, Sydney Brenner, and others established the following features of the genetic code by 1961:

1. *Three nucleotides encode an amino acid.* Proteins are built from a basic set of 20 amino acids, but there are only four bases. Simple calculations show that a minimum of three bases is required to encode at least 20 amino acids. Genetic experiments showed that *an amino acid is in fact encoded by a group of three bases, or codon.*

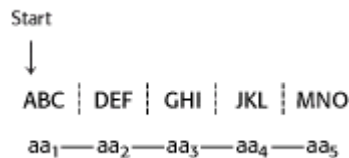
2. *The code is nonoverlapping.* Consider a base sequence ABCDEF. In an overlapping code, ABC specifies the first amino acid, BCD the next, CDE the next, and so on. In a nonoverlapping code, ABC designates the first amino acid, DEF the second, and so forth. Genetics experiments again established the code to be nonoverlapping.



3. *The code has no punctuation.* In principle, one base (denoted as Q) might serve as a "comma" between groups of three bases.

... QABCQDEFQGHQIJKLQ ...

This is not the case. Rather, *the sequence of bases is read sequentially from a fixed starting point*, without punctuation.



4. *The genetic code is degenerate.* Some amino acids are encoded by more than one codon, inasmuch as there are 64 possible base triplets and only 20 amino acids. In fact, 61 of the 64 possible triplets specify particular amino acids and 3 triplets (called stop codons) designate the termination of translation. Thus, *for most amino acids, there is more than one code word.*

5.5.1. Major Features of the Genetic Code

All 64 codons have been deciphered (Table 5.4). Because the code is highly degenerate, only tryptophan and methionine are encoded by just one triplet each. The other 18 amino acids are each encoded by two or more. Indeed, leucine, arginine, and serine are specified by six codons each. The number of codons for a particular amino acid correlates with its frequency of occurrence in proteins.

Codons that specify the same amino acid are called *synonyms*. For example, CAU and CAC are synonyms for histidine. Note that synonyms are not distributed haphazardly throughout the genetic code (depicted in Table 5.4). An amino acid specified by two or more synonyms occupies a single box (unless it is specified by more than four synonyms). The amino acids in a box are specified by codons that have the same first two bases but differ in the third base, as exemplified by GUU, GUC, GUA, and GUG. Thus, *most synonyms differ only in the last base of the triplet.* Inspection of the code shows that XYZ and

XYU always encode the same amino acid, whereas XYG and XYA usually encode the same amino acid. The structural basis for these equivalences of codons will become evident when we consider the nature of the anticodons of tRNA molecules (Section 29.3.9).

First position (5 [#] end)	Second position				Third position (3 [#] end)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note: This table identifies the amino acid encoded by each triplet. For example, the codon 5' AUG 3' on mRNA specifies methionine, whereas CAU specifies histidine, UAA, UAG, and UGA are termination signals. AUG is part of the initiation signal, in addition to coding for internal methionine residues.

Table 5.4. The genetic code

What is the biological significance of the extensive degeneracy of the genetic code? If the code were not degenerate, 20 codons would designate amino acids and 44 would lead to chain termination. The probability of mutating to chain termination would therefore be much higher with a nondegenerate code. Chain-termination mutations usually lead to inactive proteins, whereas substitutions of one amino acid for another are usually rather harmless. Thus, *degeneracy minimizes the deleterious effects of mutations*. Degeneracy of the code may also be significant in permitting DNA base composition to vary over a wide range without altering the amino acid sequence of the proteins encoded by the DNA. The G + C content of bacterial DNA ranges from less than 30% to more than 70%. DNA molecules with quite different G + C contents could encode the same proteins if different synonyms of the genetic code were consistently used.

5.5.2. Messenger RNA Contains Start and Stop Signals for Protein Synthesis

Messenger RNA is translated into proteins on *ribosomes*, large molecular complexes assembled from proteins and ribosomal RNA. How is mRNA interpreted by the translation apparatus? As already mentioned, *UAA*, *UAG*, and *UGA* designate *chain termination*. These codons are read not by tRNA molecules but rather by specific proteins called *release factors* (Section 29.4.4). Binding of the release factors to the ribosomes releases the newly synthesized protein. The start signal for protein synthesis is more complex. Polypeptide chains in bacteria start with a modified amino acid - namely, formylmethionine (fMet). A specific tRNA, the initiator tRNA, carries fMet. This fMet-tRNA recognizes the codon AUG or, less frequently, GUG. However, AUG is also the codon for an internal methionine residue, and GUG is the codon for an internal valine residue. Hence, the signal for the first amino acid in a prokaryotic polypeptide chain must be more complex than that for all subsequent ones. *AUG (or GUG) is only part of the initiation signal* (Figure 5.32). In bacteria, the initiating AUG (or GUG) codon is preceded several nucleotides away by a purine-rich sequence that base-pairs with a complementary sequence in a ribosomal RNA molecule (Section 29.3.4). In eukaryotes, the AUG closest to the 5' end of an mRNA molecule is usually the start signal for protein synthesis. This particular AUG is read by an initiator tRNA conjugated to methionine. Once the initiator AUG is located, the *reading frame* is established - groups of three nonoverlapping nucleotides are defined, beginning with the initiator AUG codon.

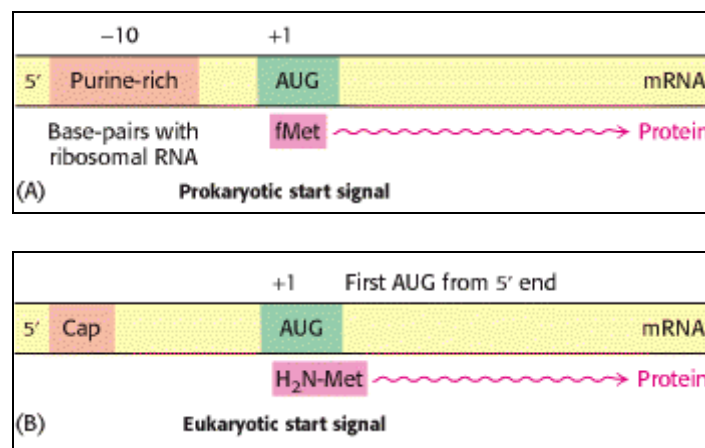
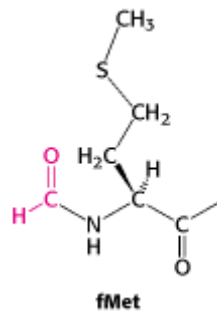


Figure 5.32. Initiation of Protein Synthesis. Start signals are required for the initiation of protein synthesis in (A) prokaryotes and (B) eukaryotes.

5.5.3. The Genetic Code Is Nearly Universal

Is the genetic code the same in all organisms? The base sequences of many wild-type and mutant genes are known, as are the amino acid sequences of their encoded proteins. In each case, the nucleotide change in the gene and the amino acid change in the protein are as predicted by the genetic code. Furthermore, mRNAs can be correctly translated by the proteinsynthesizing machinery of very different species. For example, human hemoglobin mRNA is correctly translated by a wheat germ extract, and bacteria efficiently express recombinant DNA molecules encoding human proteins such as insulin. These experimental findings strongly suggested that the genetic code is universal.

A surprise was encountered when the sequence of human mitochondrial DNA became known. Human mitochondria read UGA as a codon for tryptophan rather than as a stop signal (Table 5.5). Furthermore, AGA and AGG are read as stop signals rather than as codons for arginine, and AUA is read as a codon for methionine instead of isoleucine. Mitochondria of other species, such as those of yeast, also have genetic codes that differ slightly from the standard one. The genetic code of mitochondria can differ from that of the rest of the cell because mitochondrial DNA encodes a distinct set of tRNAs. Do any cellular protein-synthesizing systems deviate from the standard genetic code? Ciliated protozoa differ from most organisms in reading UAA and UAG as codons for amino acids rather than as stop signals; UGA is their sole termination signal. Thus, *the genetic code is nearly but not absolutely universal*. Variations clearly exist in mitochondria and in species, such as ciliates, that branched off very early in eukaryotic evolution. It is interesting to note that two of the codon reassignments in human mitochondria diminish the information content of the third base of the triplet (e.g., both AUA and AUG specify methionine). Most variations from the standard genetic code are in the direction of a simpler code.

Codon	Standard code	Mitochondrial code
UGA	Stop	Trp
UGG	Trp	Trp
AUA	Ile	Met
AUG	Met	Met
AGA	Arg	Stop
AGG	Arg	Stop

Table 5.5. Distinctive codons of human mitochondria

Why has the code remained nearly invariant through billions of years of evolution, from bacteria to human beings? A mutation that altered the reading of mRNA would change the amino acid sequence of most, if not all, proteins synthesized by that particular organism. Many of these changes would undoubtedly be deleterious, and so there would be strong selection against a mutation with such pervasive consequences.

5.6. Most Eukaryotic Genes Are Mosaics of Introns and Exons

In bacteria, polypeptide chains are encoded by a continuous array of triplet codons in DNA. For many years, genes in higher organisms also were assumed to be continuous. This view was unexpectedly shattered in 1977, when investigators in several laboratories discovered that several genes are *discontinuous*. The mosaic nature of eukaryotic genes was revealed by electron microscopic studies of hybrids formed between mRNA and a segment of DNA containing the corresponding gene (Figure 5.33). For example, the gene for the β chain of hemoglobin is interrupted within its amino acid-coding sequence by a long *intervening sequence* of 550 base pairs and a short one of 120 base pairs. Thus, the *β -globin gene is split into three coding sequences.*

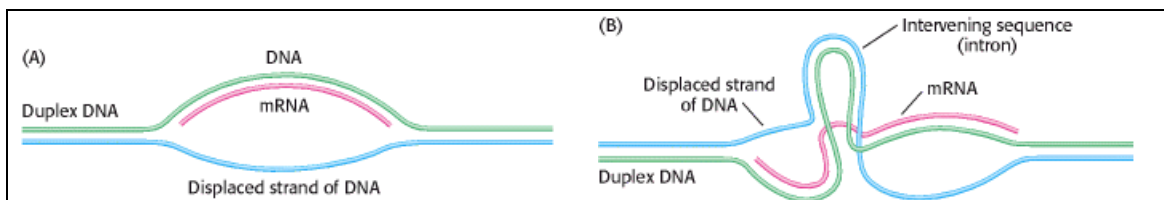
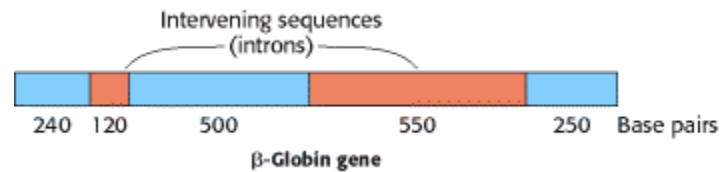


Figure 5.33. Detection of Intervening Sequences by Electron Microscopy. An mRNA molecule (shown in red) is hybridized to genomic DNA containing the corresponding gene. (A) A single loop of single-stranded DNA (shown in blue) is seen if the gene is continuous. (B) Two loops of single-stranded DNA (blue) and a loop of double-stranded DNA (blue and green) are seen if the gene contains an intervening sequence. Additional loops are evident if more than one intervening sequence is present.

5.6.1. RNA Processing Generates Mature RNA

At what stage in gene expression are intervening sequences removed? Newly synthesized RNA chains (pre-mRNA) isolated from nuclei are much larger than the mRNA molecules derived from them: in the case of β -globin RNA, the former sediment at 15S in zonal centrifugation experiments (Section 4.1.6) and the latter at 9S. In fact, the primary transcript of the β -globin gene contains two regions that are not present in the mRNA. *These intervening sequences in the 15S primary transcript are excised, and the coding sequences are simultaneously linked by a precise splicing enzyme to form the mature 9S mRNA* (Figure 5.34). Regions that are removed from the primary transcript are called *introns* (for *intervening sequences*), whereas those that are retained in the mature RNA are called *exons* (for *expressed regions*). A common feature in the expression of split genes is that their exons are ordered in the same sequence in mRNA as in DNA. Thus, split genes, like continuous genes, are colinear with their polypeptide products.

Splicing is a facile complex operation that is carried out by *spliceosomes*, which are assemblies of proteins and small RNA molecules (Section 28.3.4). This enzymatic machinery recognizes signals in the nascent RNA that specify the splice sites. *Introns nearly always begin with GU and end with an AG that is preceded by a pyrimidine-rich tract* (Figure 5.35). *This consensus sequence is part of the signal for splicing.*

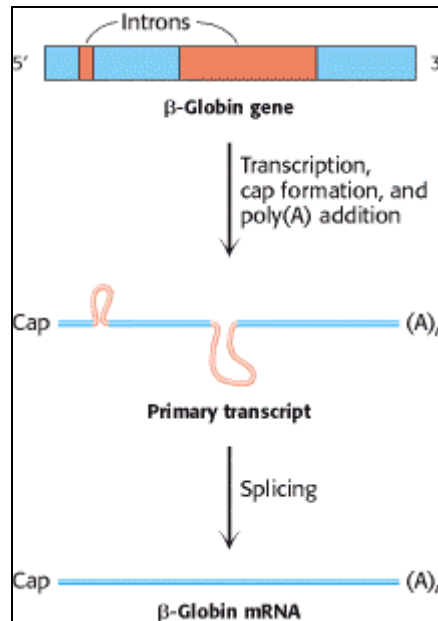


Figure 5.34. Transcription and Processing of the β -globin gene. The gene is transcribed to yield the primary transcript, which is modified by cap and poly(A) addition. The intervening sequences in the primary RNA transcript are removed to form the mRNA.

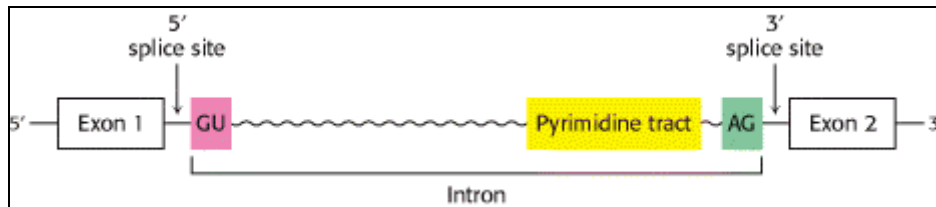


Figure 5.35. Consensus Sequence for the Splicing of mRNA Precursors.

5.6.2. Many Exons Encode Protein Domains

Most genes of higher eukaryotes, such as birds and mammals, are split. Lower eukaryotes, such as yeast, have a much higher proportion of continuous genes. In prokaryotes, split genes are extremely rare. Have introns been inserted into genes in the evolution of higher organisms? Or have introns been removed from genes to form the streamlined genomes of prokaryotes and simple eukaryotes? Comparisons of the DNA sequences of genes encoding proteins that are highly conserved in evolution suggest that *introns were present in ancestral genes and were lost in the evolution of organisms that have become optimized for very rapid growth, such as prokaryotes*. The positions of introns in some genes are at least 1 billion years old. Furthermore, a common mechanism of splicing developed before the divergence of fungi, plants, and vertebrates, as shown by the finding that mammalian cell extracts can splice yeast RNA. *Many exons encode discrete structural and functional units of proteins*. An attractive hypothesis is that *new proteins arose in evolution by the rearrangement of exons encoding discrete structural elements, binding sites, and catalytic sites*, a process called *exon shuffling*. Because it preserves functional units but allows them to interact in new ways, exon shuffling is a rapid and efficient means of generating novel genes (Figure 5.36). Introns are extensive regions in which DNA can break and recombine with no deleterious effect on encoded proteins. In contrast, the exchange of sequences between different exons usually leads to loss of function.

Another advantage conferred by split genes is the potentiality for generating a series of related proteins by splicing a nascent RNA transcript in different ways. For example, a precursor of an antibody-producing cell forms an antibody that is anchored in the cell's plasma membrane (Figure 5.37). Stimulation of such a cell by a specific foreign antigen that is recognized by the attached antibody leads to cell differentiation and proliferation. The activated antibody-producing cells then splice their nascent RNA transcript in an alternative manner to form soluble antibody molecules that are secreted rather than retained on the cell surface. We see here a clear-cut example of a benefit conferred by the complex arrangement of introns and exons in higher organisms. *Alternative splicing is a facile means of forming a*

set of proteins that are variations of a basic motif according to a developmental program without requiring a gene for each protein.

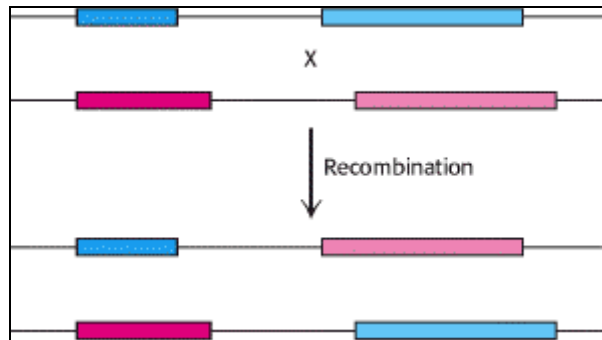


Figure 5.36. Exon Shuffling. Exons can be readily shuffled by recombination of DNA to expand the genetic repertoire.

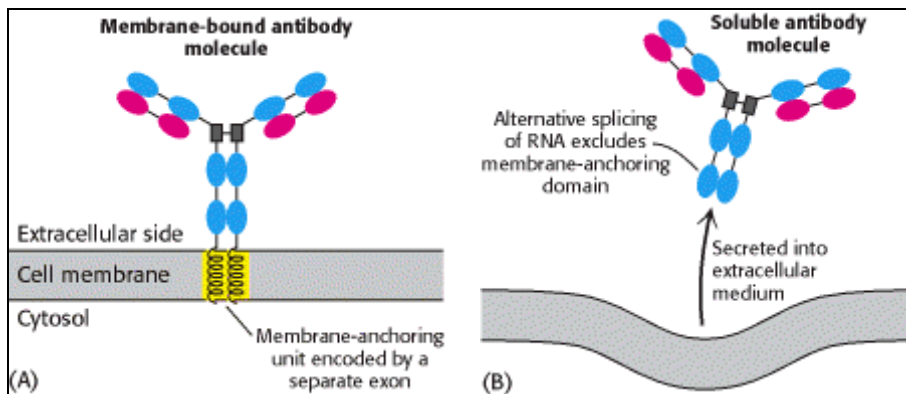


Figure 5.37. Alternative Splicing. Alternative splicing generates mRNAs that are templates for different forms of a protein: (A) a membrane-bound antibody on the surface of a lymphocyte, and (B) its soluble counterpart, exported from the cell. The membrane-bound antibody is anchored to the plasma membrane by a helical segment (highlighted in yellow) that is encoded by its o

Summary

A Nucleic Acid Consists of Four Kinds of Bases Linked to a Sugar-Phosphate Backbone

DNA and RNA are linear polymers of a limited number of monomers. In DNA, the repeating units are nucleotides, with the sugar being a deoxyribose and the bases being adenine (A), thymine (T), guanine (G), and cytosine (C). In RNA, the sugar is a ribose and the base uracil (U) is used in place of thymine. DNA is the molecule of heredity in all prokaryotic and eukaryotic organisms. In viruses, the genetic material is either DNA or RNA.

A Pair of Nucleic Acid Chains with Complementary Sequences Can Form a Double-Helical Structure

All cellular DNA consists of two very long, helical polynucleotide chains coiled around a common axis. The sugar-phosphate backbone of each strand is on the outside of the double helix, whereas the purine and pyrimidine bases are on the inside. The two chains are held together by hydrogen bonds between pairs of bases: adenine is always paired with thymine, and guanine is always paired with cytosine. Hence, one strand of a double helix is the complement of the other. The two strands of the double helix run in opposite directions. Genetic information is encoded in the precise sequence of bases along a strand. Most RNA molecules are single stranded, but many contain extensive double-helical regions that arise from the folding of the chain into hairpins.

DNA Is Replicated by Polymerases That Take Instructions from Templates

In the replication of DNA, the two strands of a double helix unwind and separate as new chains are synthesized. Each parent strand acts as a template for the formation of a new complementary strand. Thus, the replication of DNA is semiconservative - each daughter molecule receives one strand from the parent DNA molecule. The replication of DNA is a complex process carried out by many proteins, including several DNA polymerases. The activated precursors in the synthesis of DNA are the four deoxyribonucleoside 5'-triphosphates. The new strand is synthesized in the 5' → 3' direction by a nucleophilic attack by the 3'-hydroxyl terminus of the primer strand on the innermost phosphorus atom of the incoming deoxyribonucleoside triphosphate. Most important, DNA polymerases catalyze the formation of a phosphodiester bond only if the base on the incoming nucleotide is complementary to the base on the template strand. In other words, DNA polymerases are template-directed enzymes. The genes of some viruses, such as tobacco mosaic virus, are made of single-stranded RNA. An RNA-directed RNA polymerase mediates the replication of this viral RNA. Retroviruses, exemplified by HIV-1, have a single-stranded RNA genome that is transcribed into double-stranded DNA by reverse transcriptase, an RNA-directed DNA polymerase.

Gene Expression Is the Transformation of DNA Information into Functional Molecules

The flow of genetic information in normal cells is from DNA to RNA to protein. The synthesis of RNA from a DNA template is called transcription, whereas the synthesis of a protein from an RNA template is termed translation. Cells contain several kinds of RNA: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), which vary in size from 75 to more than 5000 nucleotides. All cellular RNA is synthesized by RNA polymerase according to instructions given by DNA templates. The activated intermediates are ribonucleoside triphosphates and the direction of synthesis, like that of DNA, is 5' → 3'. RNA polymerase differs from DNA polymerase in not requiring a primer.

Amino Acids Are Encoded by Groups of Three Bases Starting from a Fixed Point

The genetic code is the relation between the sequence of bases in DNA (or its RNA transcript) and the sequence of amino acids in proteins. Amino acids are encoded by groups of three bases (called codons) starting from a fixed point. Sixty-one of the 64 codons specify particular amino acids, whereas the other 3 codons (UAA, UAG, and UGA) are signals for chain termination. Thus, for most amino acids, there is more than one code word. In other words, the code is degenerate. The genetic code is nearly the same in all organisms. Natural mRNAs contain start and stop signals for translation, just as genes do for directing where transcription begins and ends.

Most Eukaryotic Genes Are Mosaics of Introns and Exons

Most genes in higher eukaryotes are discontinuous. Coding sequences (exons) in these split genes are separated by intervening sequences (introns), which are removed in the conversion of the primary transcript into mRNA and other functional mature RNA molecules. Split genes, like continuous genes, are colinear with their polypeptide products. A striking feature of many exons is that they encode functional domains in proteins. New proteins probably arose in the course of evolution by the shuffling of exons. Introns may have been present in primordial genes but were lost in the evolution of such fast-growing organisms as bacteria and yeast.

Key Terms

deoxyribonucleic acid (DNA)

deoxyribose

ribose

ribonucleic acid

purine

pyrimidine

nucleoside

nucleotide

replication

double helix

semiconservative replication

DNA polymerase

template

primer

reverse transcriptase

messenger RNA (mRNA)

translation

transfer RNA (tRNA)

ribosomal RNA (rRNA)

small nuclear RNA (snRNA)

transcription

RNA polymerase

promoter site

codon

genetic code

ribosome

intron

exon

splicing

spliceosomes

exon shuffling

alternative splicing

Problems

1. **Complements.** Write the complementary sequence (in the standard 5'→3' notation) for (a) GATCAA, (b) TCGAAC, (c) ACGCGT, and (d) TACCAT.

Answer:

(a) TTGATC; (b) GTTCGA; (c) ACGCGT; (d) ATGGTA.

2. **Compositional constraint.** The composition (in mole-fraction units) of one of the strands of a double-helical DNA molecule is [A] = 0.30 and [G] = 0.24. (a) What can you say about [T] and [C] for the same strand? (b) What can you say about [A], [G], [T], and [C] of the complementary strand?

Answer:

(a) [T] + [C] = 0.46. (b) [T] = 0.30, [C] = 0.24, and [A] + [G] = 0.46.

3. **Lost DNA.** The DNA of a deletion mutant of λ bacteriophage has a length of 15 μm instead of 17 μm . How many base pairs are missing from this mutant?

Answer:

5.7×10^3 base pairs.

4. **An unseen pattern.** What result would Meselson and Stahl have obtained if the replication of DNA were conservative (i.e., the parental double helix stayed together)? Give the expected distribution of DNA molecules after 1.0 and 2.0 generations for conservative replication.

Answer:

In conservative replication, after 1.0 generation, half of the molecules would be ^{15}N - ^{15}N , the other half ^{14}N - ^{14}N . After 2.0 generations, one-quarter of the molecules would be ^{15}N - ^{15}N , the other three-quarters ^{14}N - ^{14}N . Hybrid ^{14}N - ^{15}N molecules would not be observed in conservative replication.

5. **Tagging DNA.** (a) Suppose that you want to radioactively label DNA but not RNA in dividing and growing bacterial cells. Which radioactive molecule would you add to the culture medium? (b) Suppose that you want to prepare DNA in which the backbone phosphorus atoms are uniformly labeled with ^{32}P . Which precursors should be added to a solution containing DNA polymerase I and primed template DNA? Specify the position of radioactive atoms in these precursors.

Answer:

(a) Tritiated thymine or tritiated thymidine. (b) dATP, dGTP, dCTP, and dTTP labeled with ^{32}P in the innermost (α) phosphorus atom.

6. **Finding a template.** A solution contains DNA polymerase I and the Mg^{2+} salts of dATP, dGTP, dCTP, and TTP. The following DNA molecules are added to aliquots of this solution. Which of them would lead to DNA synthesis? (a) A single-stranded closed circle containing 1000 nucleotide units. (b) A double-stranded closed circle containing 1000 nucleotide pairs. (c) A

single-stranded closed circle of 1000 nucleotides base-paired to a linear strand of 500 nucleotides with a free 3'-OH terminus. (d) A double-stranded linear molecule of 1000 nucleotide pairs with a free 3'-OH group at each end.

Answer:

Molecules in parts *a* and *b* would not lead to DNA synthesis because they lack a 3'-OH group (a primer). The molecule in part *d* has a free 3'-OH at one end of each strand but no template strand beyond. Only the molecule in part *c* would lead to DNA synthesis.

- 7. *The right start.* Suppose that you want to assay reverse transcriptase activity. If polyriboadenylate is the template in the assay, what should you use as the primer? Which radioactive nucleotide should you use to follow chain elongation?**

Answer:

A thymidylate oligonucleotide should be used as the primer. The poly(rA) template specifies the incorporation of T; hence, radioactive TTP (labeled in the α -phosphate) should be used in the assay.

- 8. *Essential degradation.* Reverse transcriptase has ribonuclease activity as well as polymerase activity. What is the role of its ribonuclease activity?**

Answer:

The ribonuclease serves to degrade the RNA strand, a necessary step in forming duplex DNA from the RNA-DNA hybrid.

- 9. *Virus hunting.* You have purified a virus that infects turnip leaves. Treatment of a sample with phenol removes viral proteins. Application of the residual material to scraped leaves results in the formation of progeny virus particles. You infer that the infectious substance is a nucleic acid. Propose a simple and highly sensitive means of determining whether the infectious nucleic acid is DNA or RNA.**

Answer:

Treat one aliquot of the sample with ribonuclease and another with deoxyribonuclease. Test these nuclease-treated samples for infectivity.

- 10. *Mutagenic consequences.* Spontaneous deamination of cytosine bases in DNA occurs at low but measurable frequency. Cytosine is converted into uracil by loss of its amino group. After this conversion, which base pair occupies this position in each of the daughter strands resulting from one round of replication? Two rounds of replication?**

Answer:

Deamination changes the original G · C base pair into a G · U pair. After one round of replication, one daughter duplex will contain a G · C pair, and the other duplex an A · U pair. After two rounds of replication, there would be two G · C pairs, one A · U pair, and one A · T pair.

- 11. *Information content.* (a) How many different 8-mer sequences of DNA are there? (Hint: There are 16 possible dinucleotides and 64 possible trinucleotides.) (b) How many bits of information are stored in an 8-mer DNA sequence? In the *E. coli* genome? In the human genome? (c)**

Compare each of these values with the amount of information that can be stored on a personal computer diskette. A byte is equal to 8 bits.

Answer:

(a) $4^8 = 65,536$. In computer terminology, there are 64K 8-mers of DNA.

(b) A bit specifies two bases (say, A and C) and a second bit specifies the other two (G and T). Hence, two bits are needed to specify a single nucleotide (base pair) in DNA. For example, 00, 01, 10, and 11 could encode A, C, G, and T. An 8-mer stores 16 bits ($2^{16} = 65,536$), the *E. coli* genome (4×10^6 bp) stores 8×10^6 bits, and the human genome (2.9×10^9 bases) stores 5.8×10^9 bits of genetic information.

(c) A floppy diskette stores about 1.5 megabytes, which is equal to 1.2×10^7 bits. A large number of 8-mer sequences could be stored on such a diskette. The DNA sequence of *E. coli*, could be written on a single diskette. Nearly 500 diskettes would be needed to record the human DNA sequence.

12. Key polymerases. Compare DNA polymerase I and RNA polymerase from *E. coli* in regard to each of the following features: (a) activated precursors, (b) direction of chain elongation, (c) conservation of the template, and (d) need for a primer.

Answer:

(a) Deoxyribonucleoside triphosphates versus ribonucleoside triphosphates.

(b) $5' \rightarrow 3'$ for both.

(c) Semiconserved for DNA polymerase I; conserved for RNA polymerase.

(d) DNA polymerase I needs a primer, whereas RNA polymerase does not.

13. Encoded sequences. (a) Write the sequence of the mRNA molecule synthesized from a DNA template strand having the sequence



(b) What amino acid sequence is encoded by the following base sequence of an mRNA molecule? Assume that the reading frame starts at the 5' end.



(c) What is the sequence of the polypeptide formed on addition of poly(UUAC) to a cell-free protein-synthesizing system?

Answer:

(a) $5\text{'-UAACGGUACGAU-3'}$

(b) Leu-Pro-Ser-Asp-Trp-Met.

(c) Poly(Leu-Leu-Thr-Tyr).

14. A tougher chain. RNA is readily hydrolyzed by alkali, whereas DNA is not. Why?

Answer:

The 2'-OH group in RNA acts as an intramolecular nucleophile. In the alkaline hydrolysis of RNA, it forms a 2'-3' cyclic intermediate.

15. A potent blocker. How does cordycepin (3'-deoxyadenosine) block the synthesis of RNA?

Answer:

Cordycepin terminates RNA synthesis. An RNA chain containing cordycepin lacks a 3'-OH group.

16. Silent RNA. The code word GGG cannot be deciphered in the same way as can UUU, CCC, and AAA, because poly(G) does not act as a template. Poly(G) forms a triple-stranded helical structure. Why is it an ineffective template?

Answer:

Only single-stranded RNA can serve as a template for protein synthesis.

17. Two from one. Synthetic RNA molecules of defined sequence were instrumental in deciphering the genetic code. Their synthesis first required the synthesis of DNA molecules to serve as a template. H. Gobind Khorana synthesized, by organic-chemical methods, two complementary deoxyribonucleotides, each with nine residues: d(TAC)₃ and d(GTA)₃. Partly overlapping duplexes that formed on mixing these oligonucleotides then served as templates for the synthesis by DNA polymerase of long, repeating double-helical DNA chains. The next step was to obtain long polyribonucleotide chains with a sequence complementary to only one of the two DNA strands. How did he obtain only poly(UAC)? Only poly(GUA)?

Answer:

Incubation with RNA polymerase and only UTP, ATP, and CTP led to the synthesis of only poly(UAC). Only poly(GUA) was formed when GTP was used in place of CTP.

18. Overlapping or not. In a nonoverlapping triplet code, each group of three bases in a sequence ABCDEF . . . specifies only one amino acid - ABC specifies the first, DEF the second, and so forth - whereas, in a completely overlapping triplet code, ABC specifies the first amino acid, BCD the second, CDE the third, and so forth. Assume that you can mutate an individual nucleotide of a codon and detect the mutation in the amino acid sequence. Design an experiment that would establish whether the genetic code is overlapping or nonoverlapping.

Answer:

These alternatives were distinguished by the results of studies of the sequence of amino acids in mutants. Suppose that the base C is mutated to C'. In a nonoverlapping code, only amino acid 1 will be changed. In a completely overlapping code, amino acids 1, 2, and 3 will all be altered by a mutation of C to C'. The results of amino acid sequence studies of tobacco mosaic virus mutants and abnormal hemoglobins showed that alterations usually affected only a single amino acid. Hence, it was concluded that the *genetic code is nonoverlapping*.

19. Triple entendre. The RNA transcript of a region of T4 phage DNA contains the sequence 5'-AAAUGAGGA-3'. This sequence encodes three different polypeptides. What are they?

Answer:

A peptide terminating with Lys (UGA is a stop codon), -Asn-Glu-, and -Met-Arg-.

20. Valuable synonyms. Proteins generally have low contents of Met and Trp, intermediate ones of His and Cys, and high ones of Leu and Ser. What is the relation between the number of codons

of an amino acid and its frequency of occurrence in proteins? What might be the selective advantage of this relation?

Answer:

Highly abundant amino acid residues have the most codons (e.g., Leu and Ser each have six), whereas the least-abundant amino acids have the fewest (Met and Trp each have only one). Degeneracy (a) allows variation in base composition and (b) decreases the likelihood that a substitution for a base will change the encoded amino acid. If the degeneracy were equally distributed, each of the 20 amino acids would have three codons. Both benefits (a and b) are maximized by the assignment of more codons to prevalent amino acids than to less frequently used ones.

- 21. *A new translation.* A transfer RNA with a UGU anticodon is enzymatically conjugated to ¹⁴C-labeled cysteine. The cysteine unit is then chemically modified to alanine (with the use of Raney nickel, which removes the sulfur atom of cysteine). The altered aminoacyl-tRNA is added to a protein-synthesizing system containing normal components except for this tRNA. The mRNA added to this mixture contains the following sequence:**

5'-UUUUGCCAUGUUUGUGCU-3'

What is the sequence of the corresponding radiolabeled peptide?

Answer:

Phe-Cys-His-Val-Ala-Ala.

Chapter Integration Problems

- 22. *Eons ago.* The atmosphere of the primitive Earth before the emergence of life contained N₂, NH₃, H₂, HCN, CO, and H₂O. Which of these compounds is the most likely precursor of most of the atoms in adenine? Why?**

Answer:

Hydrogen cyanide. Adenine can be viewed as a pentamer of HCN.

- 23. *Back to the bench.* A protein chemist told a molecular geneticist that he had found a new mutant hemoglobin in which aspartate replaced lysine. The molecular geneticist expressed surprise and sent his friend scurrying back to the laboratory. (a) Why did the molecular geneticist doubt the reported amino acid substitution? (b) Which amino acid substitutions would have been more palatable to the molecular geneticist?**

Answer:

(a) A codon for lysine cannot be changed to one for aspartate by the mutation of a single nucleotide.

(b) Arg, Asn, Gln, Glu, Ile, Met, or Thr.

- 24. *Eons apart.* The amino acid sequences of a yeast protein and a human protein carrying out the same function are found to be 60% identical. However, the corresponding DNA sequences are only 45% identical. Account for this differing degree of identity.**

Answer:

The genetic code is degenerate. Of the 20 amino acids, 18 are specified by more than one codon. Hence, many nucleotide changes (especially in the third base of a codon) do not alter the nature of the encoded amino acid. Mutations leading to an altered amino acid are usually more deleterious than those that do not and hence are subject to more stringent selection.

Selected Readings

Where to start

G. Felsenfeld. 1985. DNA *Sci. Am.* 253: (4) 58-67. ([PubMed](#))

J.E. Darnell Jr. 1985. RNA *Sci. Am.* 253: (4) 68-78. ([PubMed](#))

R.E. Dickerson. 1983. The DNA helix and how it is read *Sci. Am.* 249: (6) 94-111.

F.H.C. Crick,. 1954.. The structure of the hereditary material *Sci. Am.* 191: (4): 54-61..

P. Chambon. 1981. Split genes *Sci. Am.* 244: (5) 60-71. ([PubMed](#))

J.D. Watson and F.H.C. Crick. 1953. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* 171: 737-738.

J.D. Watson and F.H.C. Crick. 1953. Genetic implications of the structure of deoxyribonucleic acid *Nature* 171: 964-967.

M. Meselson and F.W. Stahl. 1958. The replication of DNA in *Escherichia coli* *Proc. Natl. Acad. Sci. U.S.A.* 44: 671-682.

Books

Bloomfield, V. A., Crothers, D. M., Tinoco, I. and Hearst, J., 2000. *Nucleic Acids: Structures, Properties, and Functions*. University Science Books.

Singer, M., Berg, P., 1991. *Genes and Genomes: A Changing Perspective* . University Science Books.

Lodish, H., Berk, A., Zipursky, L., and Matsudaira, P., 1999. *Molecular Cell Biology* (4th ed.). W. H. Freeman and Company.

Lewin, B., 2000. *Genes VII*. Oxford University Press.

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M., 2000. *Molecular Biology of the Gene* (5th ed.). Benjamin Cummings.

DNA structure

Saenger, W., 1984. *Principles of Nucleic Acid Structure*. Springer Verlag.

R.E. Dickerson, H.R. Drew, B.N. Conner, R.M. Wing, A.V. Fratini, and M.L. Kopka. 1982. The anatomy of A-, B-, and Z-DNA *Science* 216: 475-485. ([PubMed](#))

Sinden, R. R., 1994. *DNA structure and function*. Academic Press.

DNA replication

Kornberg, A., and Baker, T. A., 1992. *DNA Replication* (2d ed.). W. H. Freeman and Company.

U. Hübscher, H.-P. Nasheuer, and J.E. Syväoja. 2000. Eukaryotic DNA polymerases: A growing family *Trends Biochem. Sci.* 25: 143-147. ([PubMed](#))

C.A. Brautigam and T.A. Steitz. 1998. Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes *Curr. Opin. Struct. Biol.* 8: 54-63. ([PubMed](#))

Discovery of messenger RNA

F. Jacob and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins *J. Mol. Biol.* 3: 318-356.

S. Brenner, F. Jacob, and M. Meselson. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis *Nature* 190: 576-581.

B.D. Hall and S. Spiegelman. 1961. Sequence complementarity of T2-DNA and T2-specific RNA *Proc. Natl. Acad. Sci. U.S.A.* 47: 137-146.

Genetic code

F.H.C. Crick, L. Barnett, S. Brenner, and R.J. Watts-Tobin. 1961. General nature of the genetic code for proteins *Nature* 192: 1227-1232.

Nirenberg, M., 1968. The genetic code. In *Nobel Lectures: Physiology or Medicine (1963-1970)*, pp. 372-395. American Elsevier (1973).

F.H.C. Crick. 1958. On protein synthesis *Symp. Soc. Exp. Biol.* 12: 138-163.

Woese, C. R., 1967. *The Genetic Code*. Harper & Row.

R.D. Knight, S.J. Freeland, and L.F. Landweber. 1999. Selection, history and chemistry: The three faces of the genetic code *Trends Biochem. Sci.* 24: (6) 241-247. ([PubMed](#))

Introns, exons, and split genes

P.A. Sharp. 1988. RNA splicing and genes *J. Am. Med. Assoc.* 260: 3035-3041.

R.L. Dorit, L. Schoenbach, and W. Gilbert. 1990. How big is the universe of exons? *Science* 250: 1377-1382. ([PubMed](#))

M. Cochet, F. Gannon, R. Hen, L. Maroteaux, F. Perrin, and P. Chambon. 1979. Organization and sequence studies of the 17-piece chicken conalbumin gene *Nature* 282: 567-574. ([PubMed](#))

S.M. Tilghman, D.C. Tiemeier, J.G. Seidman, B.M. Peterlin, M. Sullivan, J.V. Maizel, and P. Leder. 1978. Intervening sequence of DNA identified in the structural portion of a mouse β -globin gene *Proc. Natl. Acad. Sci. U.S.A.* 75: 725-729. ([PubMed](#))

Reminiscences and historical accounts

Watson, J. D., 1968. *The Double Helix*. Atheneum.

McCarty, M., 1985. *The Transforming Principle: Discovering That Genes Are Made of DNA*. Norton.

Cairns, J., Stent, G. S., and Watson, J. D., 2000. *Phage and the Origins of Molecular Biology*. Cold Spring Harbor Laboratory.

Olby, R., 1974. *The Path to the Double Helix*. University of Washington Press.

Portugal, F. H., and Cohen, J. S., 1977. *A Century of DNA: A History of the Discovery of the Structure and Function of the Genetic Substance*. MIT Press.

Judson, H., 1996. *The Eighth Day of Creation*. Cold Spring Harbor Laboratory.

Sayre, A. 2000. *Rosalind Franklin and DNA*. Norton.