

C H A P T E R

1

Molecular Biology and Biological Chemistry

Biology has at least 50
more interesting years.

*James D. Watson, Nobel laureate,
December 31, 1984*

The Genetic Material

Nucleotides
Orientation
Base pairing
The central dogma of molecular biology

Gene Structure and Information Content

Promoter sequences
The genetic code
Open reading frames
Introns and exons

Protein Structure and Function

Primary structure
Secondary, tertiary, and quaternary structure

The Nature of Chemical Bonds

Anatomy of an atom
Valence
Electronegativity
Hydrophilicity and hydrophobicity

Molecular Biology Tools

Restriction enzyme digests
Gel electrophoresis
Blotting and hybridization
Cloning
Polymerase chain reaction
DNA sequencing

Genomic Information Content

C-value paradox
Reassociation kinetics

The most distinguishing characteristic of living things is their ability to store, utilize, and pass on information. Bioinformatics strives to determine what information is biologically important and to decipher how it is used to precisely control the chemical environment within living organisms. Since that information is stored at a molecular level, the relatively small number of tools available to molecular biologists provides our most direct insights into that information content. This chapter provides a brief introduction or review of the format in which genetic information is maintained and used by living organisms as well as the experimental techniques that are routinely used to study it in molecular biology laboratories. Since that information is most relevant in terms of the effects that it has on the chemistry of life, that too is briefly reviewed.

The Genetic Material

DNA (deoxyribonucleic acid) is the genetic material. This is a profoundly powerful statement to molecular biologists. To a large extent, it represents the answer to questions that have been pondered by philosophers and scientists for thousands of years: “What is the basis of inheritance?” and “What allows living things to be different from nonliving things?” Quite simply, it is the information stored in DNA that allows the organization of inanimate molecules into functioning, living cells and organisms that are able to regulate their internal chemical composition, growth, and reproduction. As a direct result, it is also what allows us to inherit our mother’s curly hair, our father’s blue eyes, and even our uncle’s too-large nose. The various units that govern those characteristics at the genetic level, be it chemical composition or nose size, are called **genes**. Prior to our understanding of the chemical structure of DNA in the 1950s, what and how information was passed on from one generation to the next was largely a matter of often wild conjecture.

Nucleotides

Genes themselves contain their information as a specific **sequence** of nucleotides that are found in DNA molecules. Only four different bases are used in DNA molecules: guanine, adenine, thymine, and cytosine (**G, A, T, and C**). Each base is attached to a phosphate group and a deoxyribose sugar to form a nucleotide. The only thing that makes one nucleotide different from another is which nitrogenous base it contains (Figure 1.1). Differences between each of the four nitrogenous bases is fairly obvious even in representations of their structures such as those in Figure 1.1, and the enzymatic machinery of living cells routinely and reliably distinguishes between them. And, very much like binary uses strings of zeros and ones and the English alphabet uses combinations of 26 different letters to convey information, all of the information within each gene comes simply from the order in which those four nucleotides are found along lengthy DNA molecules. Complicated genes can be many thousands of nucleotides long, and

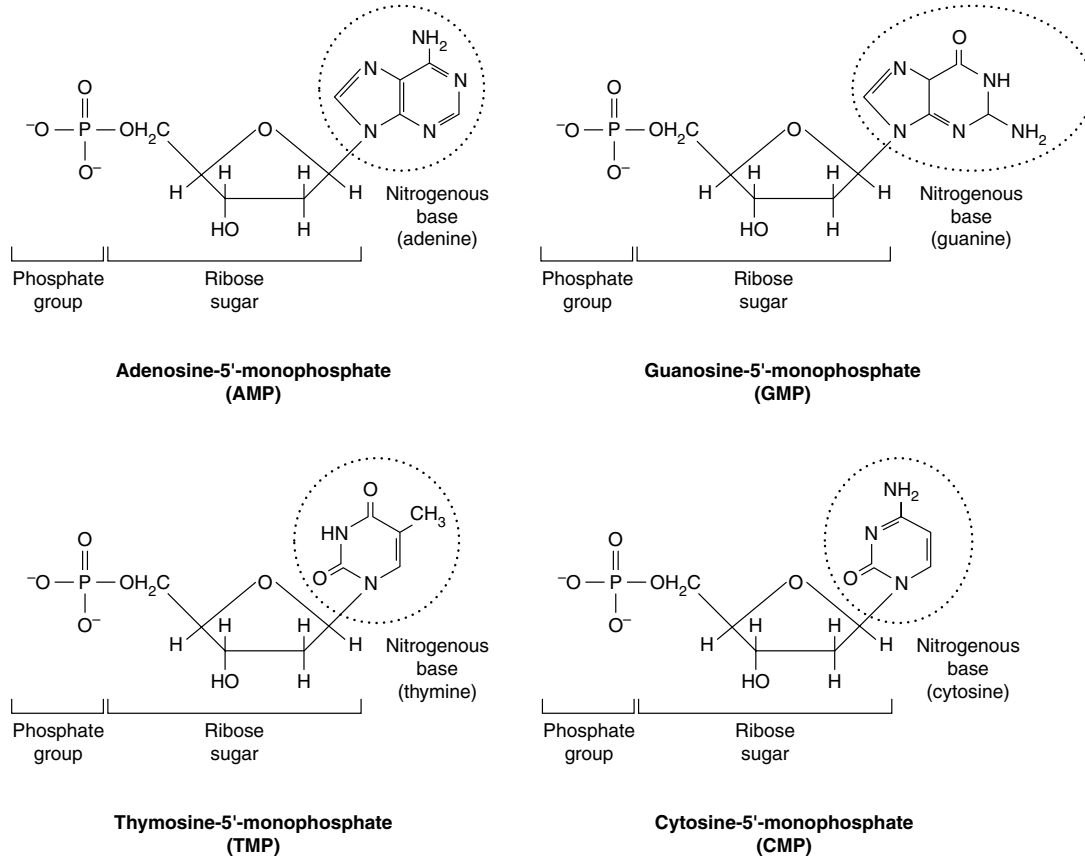


FIGURE 1.1 Chemical structure of the four nucleotides used to make DNA. Each nucleotide can be considered to be made of three component parts: (1) a phosphate group, (2) a central deoxyribose sugar, and (3) one of four different nitrogenous bases.

all of an organism's genetic instructions, its **genome**, can be maintained in millions or even billions of nucleotides.

Orientation

Strings of nucleotides can be attached to each other to make long **polynucleotide** chains or, when considered on a very large scale, **chromosomes**. The attachment between any two nucleotides is always made by way of a **phosphodiester bond** that connects the phosphate group of one nucleotide to the deoxyribose sugar of another (Figure 1.2). (Ester bonds are those that involve links made by oxygen atoms—phosphodiester bonds have a total of two ester bonds, one on each side of a phosphorous atom.)

All living things make these phosphodiester bonds in precisely the same way. Notice in Figure 1.2 that each of the five carbon atoms in a deoxyribose sugar has

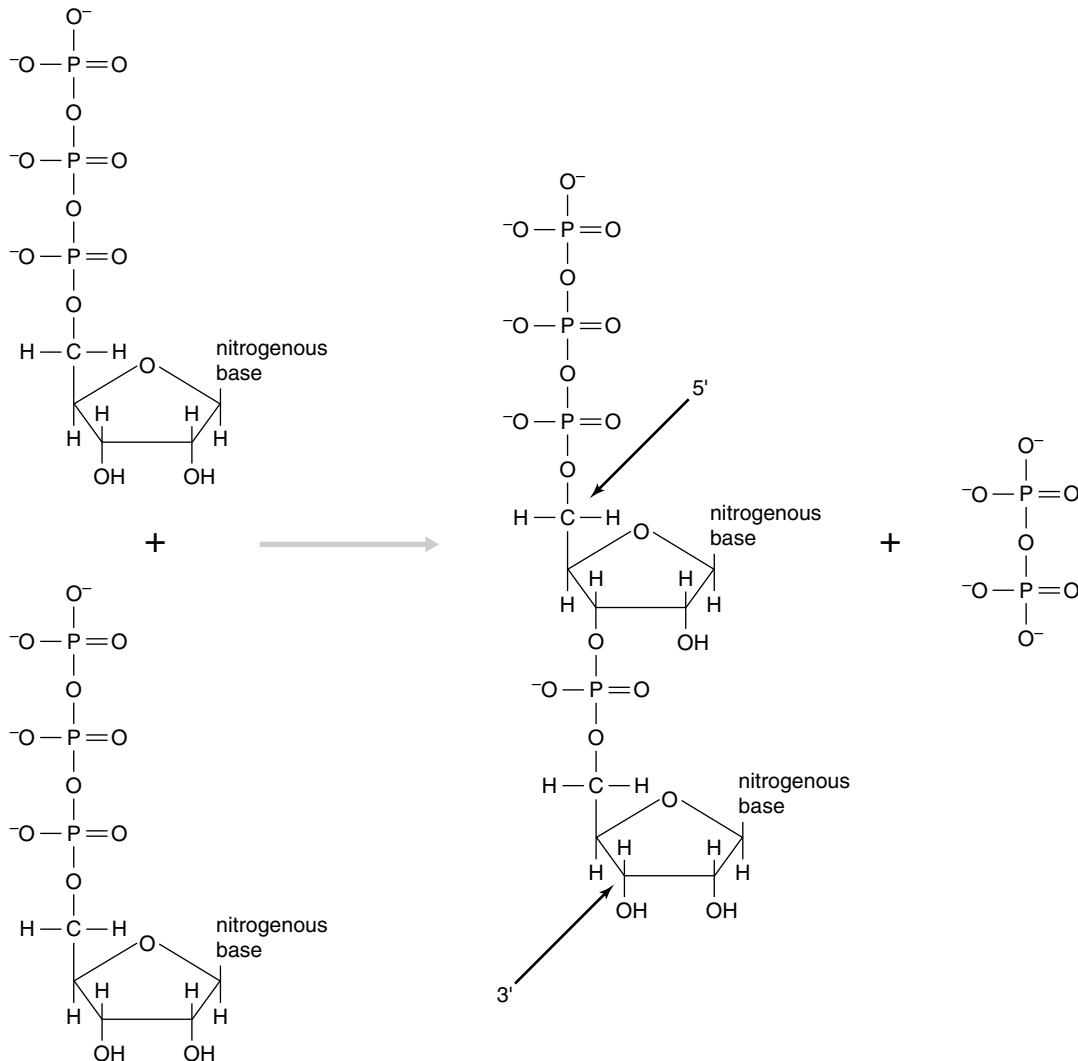


FIGURE 1.2 *The making of a phosphodiester bond. Nucleotides are added to growing DNA and RNA molecules only at their 3' ends.*

been assigned a specific numeric designation (1' through 5') by organic chemists. The phosphate group(s) of any single, unattached nucleotide are always found on its 5' carbon. Those phosphate groups are used to bridge the gap between the 5' carbon of an incoming deoxyribose sugar and the 3' carbon of a deoxyribose sugar at the end of a preexisting polynucleotide chain. As a result, one end of a string of nucleotides always has a 5' carbon that is not attached to another nucleotide, and the other end of the molecule always has an unattached 3' carbon. The difference between the 5' and 3' ends of a polynucleotide chain may seem subtle.

However, the orientation it confers to DNA molecules is every bit as important to cells as is our knowing that in written English we read from left to right and from top to bottom to understand the information content.

Base Pairing

A common theme throughout all biological systems and at all levels is the idea that structure and function are intimately related. Watson and Crick's appreciation that the DNA molecules within cells typically exist as double-stranded molecules was an invaluable clue as to how DNA might act as the genetic material. What they reported in their classic 1953 paper describing the structure of DNA is that the information content on one of those strands was essentially redundant with the information on the other. DNA could be replicated and faithfully passed on from one generation to another simply by separating the two strands and using each as a template for the synthesis of a new strand.

As we have already discussed, the information content in a DNA molecule comes from the specific sequence of its nucleotides. While the information content on each strand of a double-stranded DNA molecule is redundant it is not exactly the same—it is **complementary**. For every G on one strand, a C is found on its complementary strand and vice versa. For every A on one strand, a T is found on its complementary strand and vice versa. The interaction between G's and C's and between A's and T's is both specific and stable. The nitrogenous base guanine with its two-ringed structure is simply too large to pair with a two-ringed adenine or another guanine in the space that usually exists between two DNA strands. By the same token, the nitrogenous base thymine with its single-ringed structure is too small to interact with another single-ringed cytosine or thymine. Space is not a barrier to interaction between G's and T's or A's and C's but their chemical natures are incompatible, as will be described later in this chapter. Only the pairing between the nitrogenous bases G and C (Figure 1.3a) and the pairing between the nitrogenous bases A and T (Figure 1.3b) have both the right spacing and interaction between their chemical groups to form stable **base pairs**. In fact, the chemical interaction (specifically, three hydrogen bonds that form between G's and C's and two hydrogen bonds that form between A's and T's) between the two different kinds of base pairs is actually so stable and energetically favorable that it alone is responsible for holding the two complementary strands together.

Although the two strands of a DNA molecule are complementary they are not in the same 5'/3' orientation. Instead, the two strands are said to be **antiparallel** to each other, with the 5' end of one strand corresponding to the 3' end of its complementary strand and vice versa. Consequently, if one strand's nucleotide sequence is 5'-GTATCC-3', the other strand's sequence will be 3'-CATAGG-5'. By convention, and since most cellular processes involving DNA occur in the 5' to 3' direction, the other strand's sequence would typically be presented as: 5'-GGATAC-3'. Strictly speaking, the two strands of a double-stranded DNA molecule are *reverse* complements of each other. Sequence features that are 5' to a particular reference point are commonly described as being “upstream” while those that are 3' are described as being “downstream.”

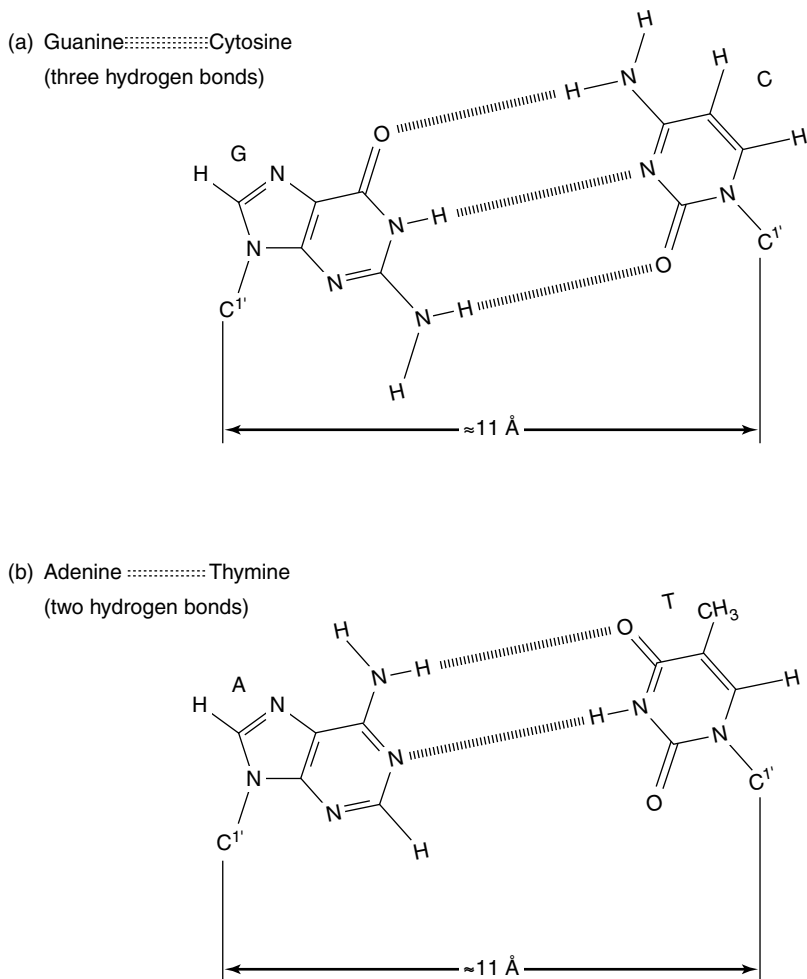


FIGURE 1.3 Base pairing between the nitrogenous bases in DNA molecules. (a) Guanine and cytosine are capable of specifically interacting by way of three hydrogen bonds, while (b) adenine and thymine interact by way of two hydrogen bonds.

The Central Dogma of Molecular Biology

While the specific sequence of nucleotides in a DNA molecule can have important information content for a cell, it is actually proteins that do the work of altering a cell's chemistry by acting as biological catalysts called **enzymes**. In chemistry catalysts are molecules that allow specific chemical reactions to proceed more quickly than they would have otherwise occurred. Catalysts are neither consumed nor altered in the course of such a chemical process and can be used to catalyze the same reaction many times. The term *gene* is used in many different ways, but one of its narrowest and simplest definitions is that genes spell out the instructions needed to make the enzyme catalysts produced by cells. The

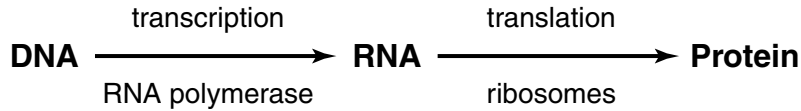


FIGURE 1.4 *The central dogma of molecular biology. Information in cells passes from DNA to RNA to proteins. RNA is made from DNA molecules during transcription by RNA polymerases. Proteins are made from the information content of RNA molecules as they are translated by ribosomes. DNA polymerases also make copies of DNA molecules during the replication process of cell division.*

process by which information is extracted from the nucleotide sequence of a gene and then used to make a protein is essentially the same for all living things on Earth and is described by the grandly named **central dogma** of molecular biology. Quite simply, information stored in DNA is used to make a more transient, single-stranded polynucleotide called RNA (ribonucleic acid) that is in turn used to make proteins (Figure 1.4). The process of making an RNA copy of a gene is called **transcription** and is accomplished through the enzymatic activity of an **RNA polymerase**. There is a one-to-one correspondence between the nucleotides used to make RNA (G, A, U, and C where “U” is an abbreviation for uracil) and the nucleotide sequences in DNA (G, A, T, and C, respectively). The process of converting that information from nucleotide sequences in RNA to the amino acid sequences that make a protein is called **translation** and is performed by a complex of proteins and RNA called **ribosomes**. Protein synthesis and structure are discussed at the end of this chapter.

Gene Structure and Information Content

Formatting and its interpretation are important considerations for any information storage system, be it a written text or a cell’s DNA molecule. All cells go about interpreting their genetic instructions in the same way and rely on specific signals to “punctuate” their genes. Much of the “language” of DNA and the rules of its interpretation were worked out very early in the history of life on earth and, because of their central importance, have changed very little over the course of billions of years. As a result, both prokaryotic (bacteria) and eukaryotic (more complicated organisms like yeast, plants, pets, and people) organisms all use not only the same “alphabet” of nucleotides but also use essentially the same format and approach for storing and utilizing their genetic information.

Promoter Sequences

Gene expression, the process of using the information stored in DNA to make an RNA molecule and then a corresponding protein, can have significant energetic and opportunity costs for a cell. Organisms that express unneeded proteins are less likely to survive and reproduce relative to competitors that regulate their

gene expression more appropriately. As a result, all cells place particular emphasis on controlling gene expression at its very start by making two crucial distinctions. First, they must reliably distinguish between those parts of an organism's genome that correspond to the beginnings of genes and those that do not. Second, they must be able to determine which genes code for proteins that are needed at any particular time.

Since RNA polymerases are responsible for the initiation of gene expression through their synthesis of RNA copies of genes, it is reasonable that the burden of making those two distinctions falls on them. Certainly not every nucleotide in a genome can correspond to the start of a gene any more than every letter on a printed page can correspond to the beginning of a sentence with useful information content. By the same token, RNA polymerases cannot simply look for any *one* particular nucleotide, like A, when looking for the start of a gene because each nucleotide occurs by chance so frequently throughout a cell's DNA. However, particular combinations of nucleotides are not as likely to occur by chance, and the greater the number of nucleotides involved, the smaller a chance occurrence becomes. The probability (P) that a string of nucleotides will occur by chance alone can be determined by the relatively simple formula $P = (1/4)^n$ if all nucleotides are present at the same frequency and where n is the string's length. Prokaryotic RNA polymerases actually scan along DNA looking for a specific set of approximately 13 nucleotides (1 nucleotide that serves as a transcriptional start site, 6 that are 10 nucleotides 5' to the start site, and 6 more that are 35 nucleotides 5' to it) that mark the beginning of genes. Those nucleotides, taken as a whole and in the proper positions relative to each other, are called **promoter sequences**. Given that most prokaryotic genomes are only a few million nucleotides long, these promoter sequences, which should occur only by chance about once in every 70 million nucleotides, allow RNA polymerases to uniquely identify the beginnings of genes with great statistical confidence. Eukaryotic genomes tend to be several orders of magnitude larger than those of prokaryotes and, as a result, eukaryotic RNA polymerases tend to recognize larger and more complex promoter sequences so that they too can reliably recognize the beginning of genes.

Two French biochemists, F. Jacob and J. Monod, were the first to obtain direct molecular insights into how cells distinguish between genes that should be transcribed and those that should not. Their work on prokaryotic gene regulation earned them a Nobel Prize in 1965 and revealed that the expression of structural genes (those that code for proteins involved in cell structure or metabolism) was controlled by specific regulatory genes. The proteins encoded by these regulatory genes are typically capable of binding to a cell's DNA near the promoter of the genes whose expression they control in some circumstances but not in others. It is the ability of these regulatory proteins to bind or not bind to specific nucleotide sequences in a fashion that is dependent on their ability to sense a cell's chemical environment that allows living things to respond appropriately to their environment. When the binding of these proteins makes it easier for an RNA polymerase to initiate transcription, **positive regulation** is said to have occurred. **Negative regulation** describes those situations where binding of the regulatory protein prevents transcription from occurring. Eventually, most prokaryotic structural

genes were found to be turned on or off by just one or two regulatory proteins. Eukaryotes, with their substantially more complicated genomes and transcriptional needs, use larger numbers (usually seven or more) and combinations of regulatory proteins to control the expression of their structural genes.

The Genetic Code

While nucleotides are the building blocks that cells use to make their information storage and transfer molecules (DNA and RNA, respectively), amino acids are the units that are strung together to make the proteins that actually do most of the work of altering a cell's chemical environment. The function of a protein is intimately dependent on the order in which its amino acids are linked by ribosomes during translation and, as has already been discussed, that order is determined by the instructions transcribed into RNA molecules by RNA polymerases. However, although only four different nucleotides (nt) are used to make DNA and RNA molecules, 20 different amino acids (each with its own distinctive chemistry) are used in protein synthesis (Figure 1.5a) ($1 \text{ nt} \neq 1 \text{ aa}$; $4^1 < 20$). There cannot be a simple one-to-one correspondence between the nucleotides of genes and the amino acids of the proteins they encode. The 16 different possible pairs of nucleotides also fall short of the task ($2 \text{ nt} \neq 1 \text{ aa}$; $4^2 < 20$). However, the four nucleotides can be arranged in a total of 64 different combinations of three ($4^3 = 64$). As a result, it is necessary for ribosomes to use a **triplet code** to translate the information in DNA and RNA into the amino acid sequence of proteins. With only three exceptions, each group of three nucleotides (a **codon**) in an RNA copy of the coding portion of a gene corresponds to a specific amino acid (Table 1.1). The three codons that do not instruct ribosomes to insert a specific amino acid are called **stop codons** (functionally equivalent to a period at the end of a sentence) because they cause translation to be terminated. This same genetic code seems to have been in place since the earliest history of life on earth and, with only a few exceptions, is universally used by all living things today.

Notice in Table 1.1 that 18 of the 20 different amino acids are coded for by more than one codon. This feature of the genetic code is called **degeneracy**. It is therefore possible for mistakes to occur during DNA replication or transcription that have no effect on the amino acid sequence of a protein. This is especially true of mutations (heritable changes in the genetic material) that occur in the third (last) position of a codon. Each amino acid can be assigned to one of essentially four different categories: nonpolar, polar, positively charged, and negatively charged (Figure 1.5b). A single change within a triplet codon is usually not sufficient to cause a codon to code for an amino acid in a different group. In short, the genetic code is remarkably robust and minimizes the extent to which mistakes in the nucleotide sequences of genes can change the functions of the proteins they encode.

Open Reading Frames

Translation by ribosomes starts at translation-initiation sites on RNA copies of genes and proceeds until a stop codon is encountered. Just as three codons of the

(a) Side chain

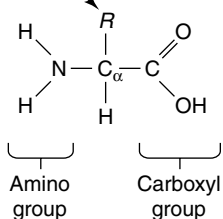


FIGURE 1.5 (a) Chemical structure of a generic amino acid. The amino group, alpha carbon, and carboxyl groups are identical for all 20 amino acids while each has its own distinctive R group. (b) Chemical structure of the 20 different amino acids complete with their distinctive R groups. Amino acids are grouped according to the properties of their side chains, shown in black. Standard three-letter and one-letter abbreviations for each of the amino acids are shown in parentheses.

(b)

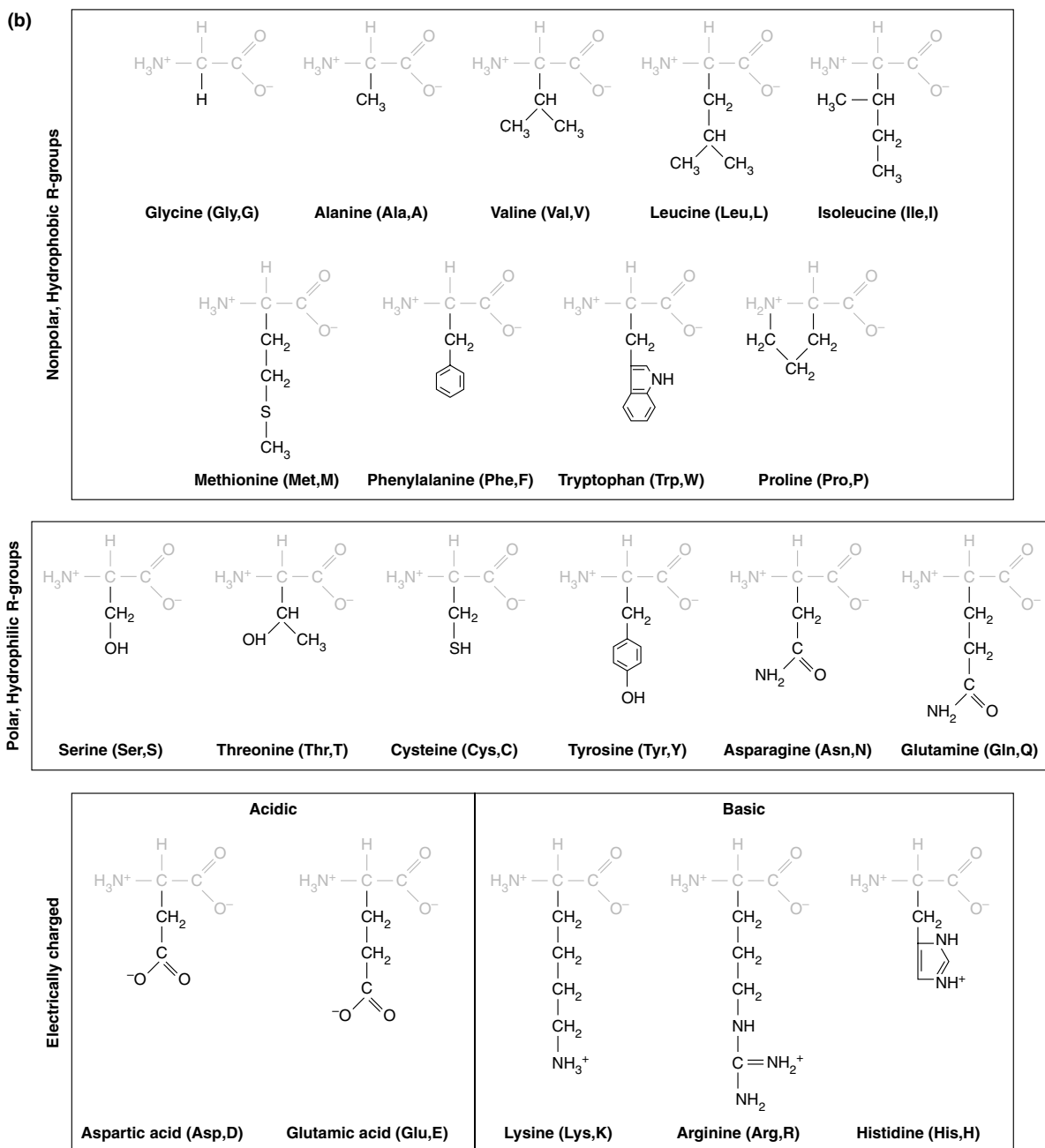


TABLE 1.1 A summary of the coding assignments of the 64 triplet codons. *Standard three-letter abbreviations for each of the most commonly used 20 amino acids are shown. The universality of the genetic code encompasses animals (including humans), plants, fungi, archaea, bacteria, and viruses. Small variations in the code exist in mitochondria and certain microbes. For instance, in a limited number of bacterial genes, a special UGA codon, normally a termination codon, is used as a codon for an unusual, 21st naturally occurring, amino acid selenocysteine. A 22nd naturally occurring amino acid, pyrrolysine, is coded for by UAG (a stop codon for most organisms) in some bacterial and eukaryotic species.*

(5') . . . pNpNpN . . . (3') in mRNA					
Base at 5' End of Codon ↓	Middle Base of Codon →				Base at 3' End of Codon ↓
	U	C	A	G	
U	phe (UUU)	ser	tyr	cys	U
	phe	ser	tyr	cys	C
	leu	ser	termination	termination	A
	leu	ser	termination	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 (and initiation)	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

genetic code are reserved as stop codons, one triplet codon is always used as a **start codon**. Specifically, the codon AUG is used both to code for the amino acid methionine as well as to mark the precise spot along an RNA molecule where translation begins in both prokaryotes and eukaryotes. Accurate translation can only occur when ribosomes examine codons in the phase or **reading frame** that is established by a gene's start codon. Unless a mistake involving some multiple of three nucleotides occurs, alterations of a gene's reading frame change every amino acid coded downstream of the alteration, and such alterations typically result in the production of a truncated version of the protein due to ribosomes encountering a premature stop codon.

Most genes code for proteins that are hundreds of amino acids long. Since stop codons occur in a randomly generated sequence at about every 20th triplet codon (3 codons out of 64), one of the reading frames of the RNA copies of most genes has unusually long runs of codons in which no stop codons occur. These strings of codons uninterrupted by stop codons are known as **open reading frames** (ORFs) and are a distinguishing feature of many prokaryotic and eukaryotic genes.

Introns and Exons

The messenger RNA (mRNA) copies of prokaryotic genes correspond perfectly to the DNA sequences present in the organism's genome with the exception that the nucleotide uracil (U) is used in place of thymine (T). In fact, translation by ribosomes almost always begins while RNA polymerases are still actively transcribing a prokaryotic gene.

Eukaryotic RNA polymerases also use uracil in place of thymine, but much more striking differences are commonly found between the mRNA molecules seen by ribosomes and the nucleotide sequences of the eukaryotic genes that code for them. In eukaryotes the two steps of gene expression are physically separated by the nuclear membrane, with transcription occurring exclusively within the nucleus and translation occurring only after mRNAs have been exported to the cytoplasm. As a result, the RNA molecules transcribed by eukaryotic RNA polymerases can be modified before ribosomes ever encounter them. The most dramatic modification that is made to the primary RNA transcripts of most eukaryotic genes is called **splicing** and involves the precise excision of internal sequences known as **introns** and the rejoining of the **exons** that flank them (Figure 1.6). Splicing is far from a trivial process and most eukaryotic genes have a large number of sometimes very large introns. An extreme example is the gene associated with the disease cystic fibrosis in humans, which has 24 introns and is over 1 million nucleotides (1 mega base pair or 1 Mb) long even though the mRNA seen by ribosomes is only about 1,000 nucleotides (1 kilo base pair or 1 kb) long. Failure to appropriately splice the introns out of a primary eukaryotic RNA transcript typically introduces frame shifts or premature stop codons that render useless any protein translated by a ribosome. Regardless of the tissue or even the organism being considered, the vast majority of eukaryotic introns conform to what is known as the "GT-AG rule," meaning that the first two nucleotides in the DNA sequence of all introns begin with the dinucleotide GT and end with the dinucleotide AG. Pairs of nucleotides occur too often just by chance to be a sufficient signal for the enzyme complexes responsible for splicing in eukaryotes, **spliceosomes**, and approximately six additional nucleotides at the 5' and 3' ends of introns are also scrutinized—sometimes differently in some cell types relative to others. This **alternative splicing** allows a huge increase in the diversity of proteins that eukaryotic organisms can use and is accomplished by often subtle modifications of spliceosomes and accessory proteins that are responsible for recognizing intron/exon boundaries.

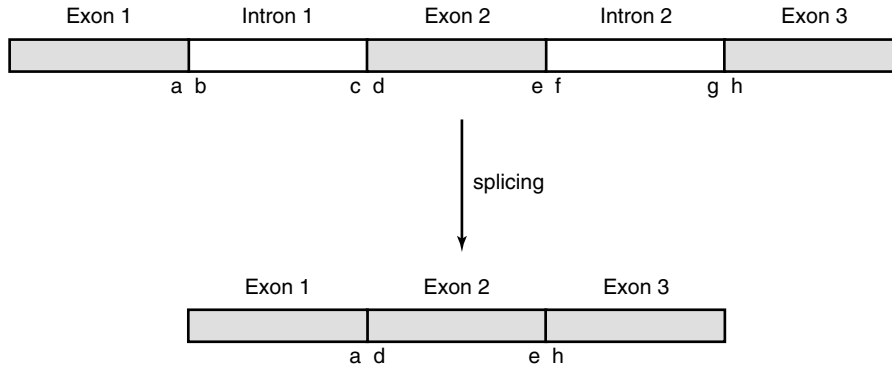


FIGURE 1.6 *Splicing of the primary transcripts of eukaryotic genes results in the removal of introns and the precise joining of exons. Intron sequences are quickly degraded once they are removed while the spliced mRNA is exported out of the nucleus for translation by ribosomes.*

Protein Structure and Function

Proteins are the molecular machinery responsible for performing most of the work of both prokaryotic and eukaryotic cells. The tasks undertaken by proteins are incredibly diverse. **Structural proteins**, such as collagen, provide rigidity and support in bones and connective tissues. Other proteins called **enzymes** act as biological catalysts, like the digestive enzyme pepsin that helps to break down and metabolize food. Proteins are also responsible for transportation of atoms and small molecules throughout an organism (e.g., hemoglobin), signaling and intercellular communication (e.g., insulin), absorbing photons to enable vision (e.g., rhodopsin), and myriad other functions.

Primary Structure

Following the genetic instructions contained in messenger RNA, proteins are translated by ribosomes as linear polymers (chains) of amino acids. The 20 amino acids have similar chemical structures (Figure 1.5a), varying only in the chemical group attached in the R position. The constant region of each amino acid is called the *backbone*, while the varying R group is called the *side chain*. The order in which the various amino acids are assembled into a protein is the sequence, or **primary structure**, of the protein. As with DNA, the protein chain has directionality. One end of the protein chain has a free amino (NH) group, while the other end of the chain terminates in a carboxylic acid (COOH) group. The individual amino acids in the protein are usually numbered starting at the **amino terminus** and proceeding toward the **carboxy terminus**.

After translation, a protein does not remain in the form of a simple linear chain. Rather, the protein collapses, folds, and is shaped into a complex globular structure. The order in which the various amino acids are assembled into a protein

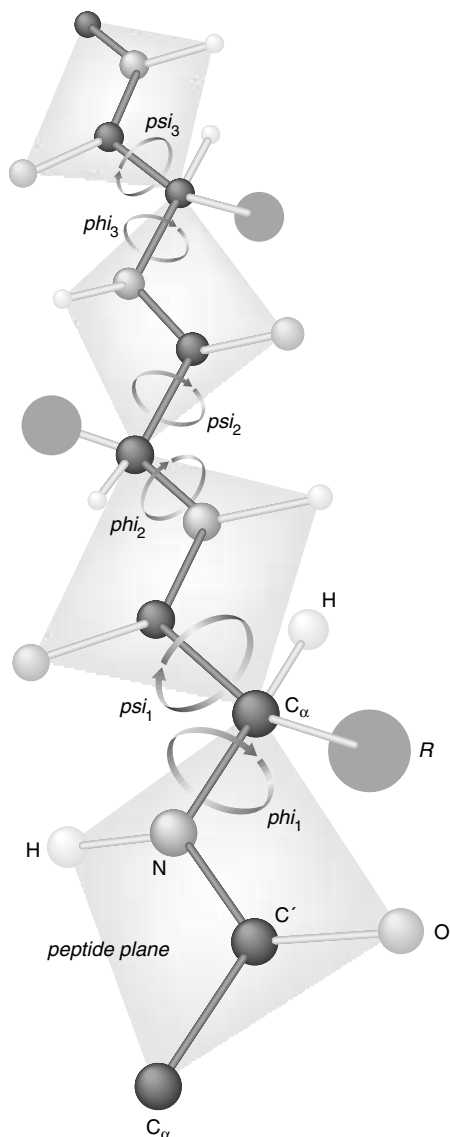


FIGURE 1.7 *Rigid and mobile regions of the protein backbone. Most of the backbone is rigid. The chemical bonds to the alpha carbons are rotatable. The angles of rotation for each alpha carbon's bonds are called phi (ϕ) and psi (ψ).*

largely determines the structure into which it will fold. The unique structure into which a particular protein will fold is called the **native structure** of the protein. The native structures of proteins give them unique properties that allow them to perform their particular roles in the context of a living organism.

The chemistry of a protein backbone forces most of the backbone to remain planar (Figure 1.7). The only “movable” segments of the protein backbone are the bonds from the nitrogen to the alpha carbon (the carbon atom to which the side chain is attached) and the bond between the alpha carbon and the carbonyl carbon (the carbon with a double bond to an oxygen atom). These two chemical bonds allow for circular (or “dihedral”) rotation, and are often called phi (ϕ) and psi (ψ), respectively. Thus, a protein consisting of 300 amino acids will have 300 phi and psi angles, often numbered ϕ_1, ψ_1 through ϕ_{300}, ψ_{300} . All of the various conformations attainable by the protein come from rotations about these 300 pairs of bonds.

Secondary, Tertiary, and Quaternary Structure

Careful examination of proteins whose structures are known reveals that a very small number of patterns in local structures are quite common. These structures, formed by regular intramolecular hydrogen bonding (described below) patterns, are found in nearly every known protein. The location and direction of these regular structures make up the **secondary structure** of the protein. The two most common structures are the α -helix and the β -sheet (Figure 1.8). Often, the secondary structures are the first portions of the protein to fold after translation. Alpha (α) helices are characterized by phi and psi angles of roughly -60° , and exhibit a spring-like helical shape with 3.6 amino acids per complete 360° turn. Beta (β) strands are characterized by regions of extended (nearly linear) backbone conformation with $\phi \approx -135^\circ$ and $\psi \approx 135^\circ$.

Beta strands assemble into one of two types of beta sheets, as illustrated in Figure 1.9 on page 16. In anti-parallel sheets, adjacent strands run in opposite directions as you move along the protein backbone from amino to carboxy terminus. In parallel sheets, the strands run in the same direc-

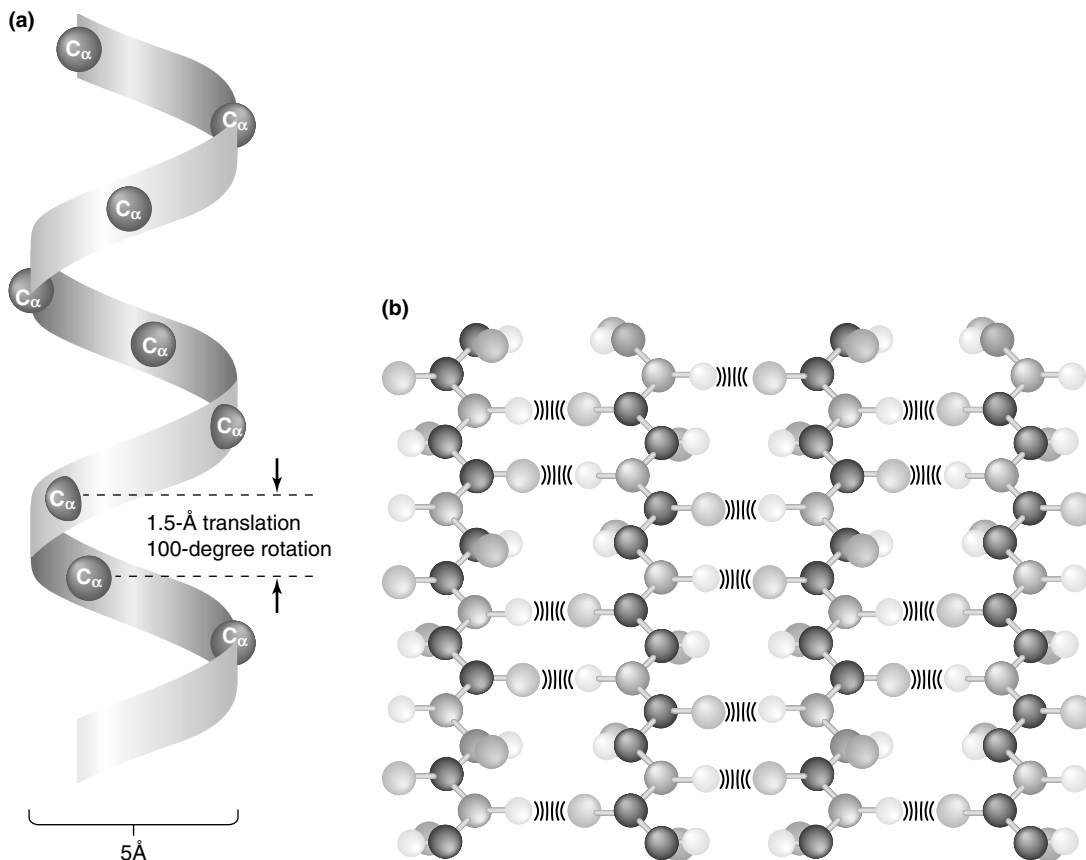


FIGURE 1.8 Elements of secondary structure: (a) the alpha (α) helix and (b) the beta (β) sheet.

tions. To make this possible, the strands of parallel beta sheets are often composed of amino acids that are nonlocal in the primary structure of the protein (Figure 1.9).

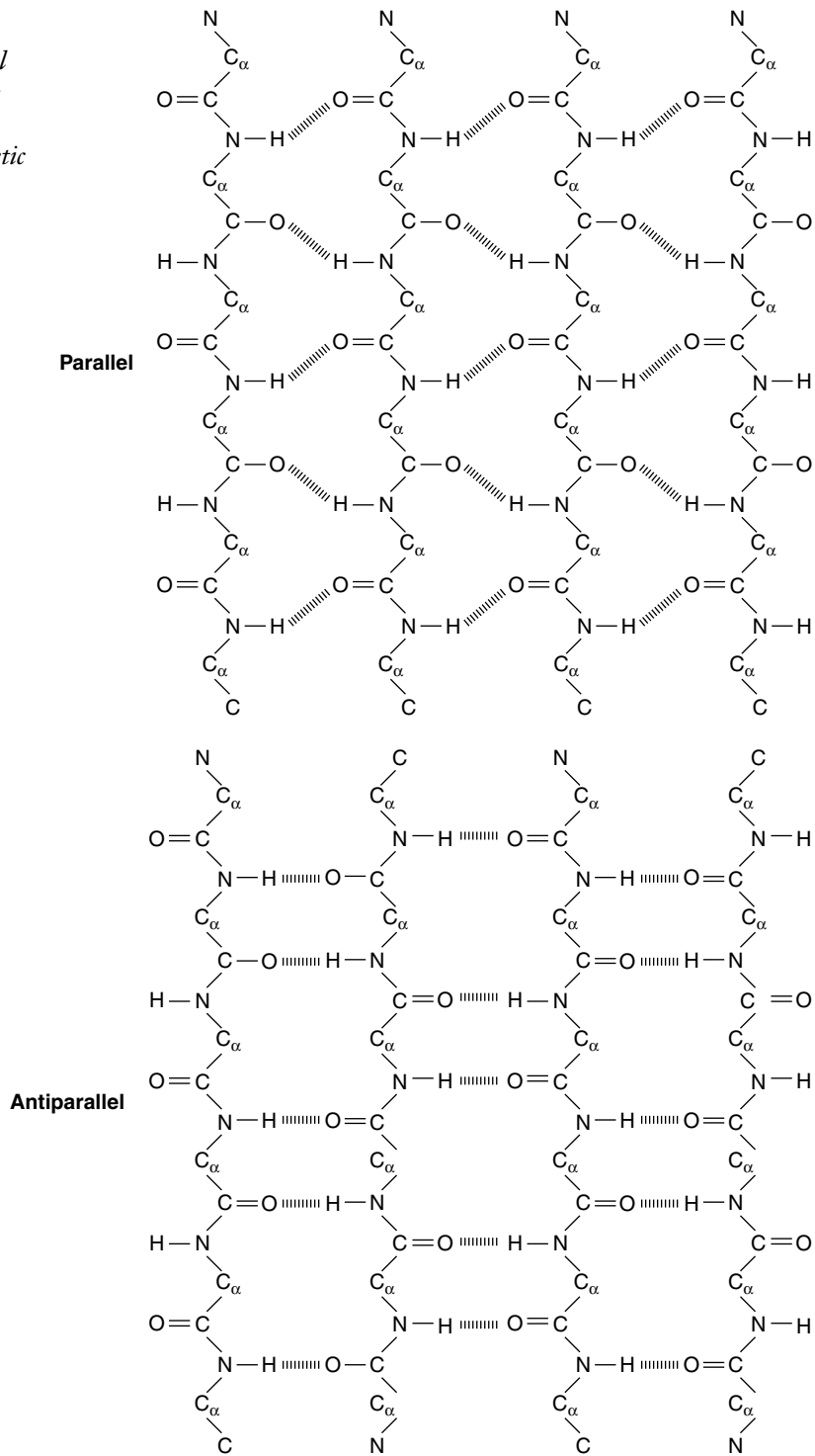
The regions of secondary structure in a protein pack together and combine with other less structured regions of the protein backbone to form an overall three-dimensional shape, which is called the **tertiary structure** of the protein. Often an active enzyme is composed of two or more protein chains that come together into a single large complex. When this occurs, the overall structure formed by the interacting proteins is commonly referred to as the **quaternary structure** of the enzyme.

The Nature of Chemical Bonds

As seen already, descriptions of nucleotides and proteins require at least a familiarity with the idea of chemical and hydrogen bonds. Much, if not all, of what we

FIGURE 1.9

Parallel and anti-parallel beta sheets. Note that the hydrogen bonds (dotted lines) give greater energetic stability in anti-parallel sheets.



consider to be essential to life can be reduced to a set of chemical reactions and the characteristics of the enzymes that control the rate at which they occur. Even a passing understanding of basic chemistry gives deep insights into the way in which enzymes function and how molecules like proteins and DNA interact. Local differences in hydrophobicity and hydrophilicity, for instance, are fundamentally important to the functioning of most enzymes. L. Pauling won a Nobel Prize in 1962 for a book he wrote that made sense of such differences at a subatomic level. The following section describes the essence of his approach.

Anatomy of an Atom

By definition, **elements** are things that cannot be further reduced by chemical reactions. Elements themselves are made of individual atoms, which, in turn, are also made of smaller, subatomic particles. These smaller component parts of elements and atoms can only be separated by physical reactions, not chemical ones. Nuclear physicists have discovered hundreds of subatomic particles. Only three, however, are stable and particularly important to the discussion of the chemistry of living things. Those three subatomic particles are neutrons (weighing 1.7×10^{-24} gram and having no charge), protons (also weighing 1.7×10^{-24} gram and possessing one positive charge), and electrons (having only 1/2000th the mass of a proton or neutron and possessing a single negative charge). The number of protons in the nucleus of an atom determines what element it is. Generally, for every proton in an atomic nucleus there is an electron in orbit around it to balance the electrical charges. Electrons move in orbits at the speed of light and a relatively long way off from the nucleus. As a result, atoms are mostly empty space.

It also takes more and more energy for an electron to be moved away from the positive charges of atomic nuclei. Similarly, it takes more energy to carry a rock to the 20th floor of a building than it does to carry it to the 19th. The further an electron is from the nucleus of its atom, the more potential energy it must have. Packets of energy can be parceled out in a number of ways at the level of atoms, and one of the more common is through light (photons). Light plus an electron often results in an electron that is residing at an orbital with a higher energy level. Electrons often release light (a packet of energy) when they go from high to low orbitals. The amount of energy required for such a transition is narrowly defined and is known as a quantum (hence the term *quantum leap*). Electrons do not have predictable orbits in the same way that planets do. In fact, the best estimates that can be made about the position of an atom's electrons are based on confidence about where the electron is most likely to be at a given time. *Orbitals* are the three-dimensional space in which an electron spends 90% of its time.

Valence

Because the negative charges of electrons are repulsive to each other, only two can share an orbital at any given time. Electrons with the lowest amounts of energy are found in an orbital closest to the nuclei of atoms known as **1s**. It is a spherical orbital and, again, it holds only two electrons. The second highest

energy level for electrons has a total of four orbitals ($2s$, and $2p$; the $2s$ orbital is also spherical and the three $2p$ orbitals are dumbbell shaped).

The chemical properties of an atom depend on its outermost shell of electrons. Since atoms are mostly empty space, nuclei never meet in normal chemical reactions—only electrons way out at the edge of the atoms ever have an opportunity to interact. Although the number of protons in an atom never changes during a chemical reaction, the relative positions (and sometimes even the number) of electrons do.

Although maintaining a balance of charges (i.e., one electron for every proton in an atom) is Nature's highest priority, there is also a strong tendency to keep an atom's outermost shell of orbitals completely full or completely empty. These potentially conflicting tendencies can be resolved by allowing the electron orbitals of atoms to overlap. The sharing of electrons that results from the overlapping of those orbitals is typically part of a long-term association between the two atoms and is the basis of **covalent bonding**. Since the atoms of some elements such as helium, ${}^2\text{He}$ (the subscript number before an atomic symbol such as He states the number of protons in an atom's nucleus), have no unpaired electrons in their outermost orbital they are not chemically reactive and are never covalently bound to other atoms. In the same way, ${}^{10}\text{Ne}$ (in which both the $1s$ orbital and all four of the level-2 orbitals are filled with a total of 10 electrons) and ${}^{18}\text{Ar}$ are also unreactive. Atoms with similar valences have similar chemical properties: Carbon, ${}^6\text{C}$ (in which each of the four level-2 orbitals has a single electron), and silicon, ${}^{14}\text{Si}$ (in which each of the four level-3 orbitals has a single electron), react very similarly and are both capable of making four covalent bonds. As a result, the number of unpaired electrons in an atom's outermost orbital, its **valence**, takes on a special significance and represents its bonding capacity: ${}^1\text{H} = 1$, ${}^8\text{O} = 2$, ${}^7\text{N} = 3$, ${}^6\text{C} = 4$. The shape and size of compounds (a complex of two or more covalently bound atoms) are largely governed by the valences of the atoms that comprise them.

Electronegativity

The chemistry of living things is complicated by the fact that different nuclei have different affinities for electrons. The higher an atom's affinity for electrons, the higher its **electronegativity**. The relative electronegativity of an atom is a function of how many electrons it needs to acquire or to donate in order to completely fill or empty its outermost shell of orbitals. For instance, ${}^1\text{H}$ and ${}^6\text{C}$ both have outermost shells of electrons that are half full. Since their electronegativities are essentially the same, atoms are shared evenly in the covalent bonds between hydrogen and carbon atoms. This is substantially different from what occurs in the covalent bonds between hydrogen and carbon with oxygen. Since ${}^8\text{O}$ must either gain just two electrons or lose six, it is much more electronegative than hydrogen or carbon. Electrons involved in the covalent bonds of water (H_2O), for instance, tend to spend more time in the vicinity of the oxygen atom than the hydrogen atom. **Polar bonds** such as these result in a slight separation of charge that makes the oxygens of water molecules slightly negative and the

hydrogens slightly positive. The slight separation of charges that result from polar covalent bonds allows for an important type of interaction between molecules called **hydrogen bonding**. Every water molecule is typically loosely associated with a network of other water molecules because the slight positive charges of their hydrogen atoms give them an affinity for the slight negative charges of the oxygens in their neighbors. Much less energy is required to break the association caused by hydrogen bonding than by covalent bonding because no electrons are shared between atoms in hydrogen bonds.

Hydrophilicity and Hydrophobicity

Chemists have found that most chemicals can be easily placed in one of just two categories: those that interact with water and those that do not. Molecules with polar bonds, like water itself, have some regions of positive and negative charge on their surfaces that are capable of forming hydrogen bonds with water. This makes them **hydrophilic** (literally, “water friendly”) and allows them to be easily dissolved in watery solutions like the interior of a living cell. Other molecules that have atoms joined by only nonpolar covalent bonds are **hydrophobic** (literally, “afraid of water”) and have much less basis of interaction with water molecules. In fact, their physical presence actually gets in the way of water molecules interacting with each other and prevents them from offsetting their partial charges. As a result, molecules such as fats that are composed primarily of carbon–carbon and carbon–hydrogen bonds are actually excluded from watery solutions and forced into associations with each other such as those observed in a cell’s lipid bilayer membrane.

Molecular Biology Tools

Recognizing the information content in the DNA sequences of prokaryotic genomes is invariably easier than the equivalent task in more complicated eukaryotic genomes. For example, while the mRNA copies of both prokaryotic and eukaryotic genes have long ORFs, the DNA sequences of eukaryotic genes themselves often do not due to the presence of introns (see Chapter 6). The problem of identifying protein coding information within eukaryotic DNA sequences is further compounded by the fact that what may be an intron in one kind of eukaryotic cell may be an exon in another (described in greater detail in Chapter 6). These problems and others associated with deciphering the information content of genomes are far from insurmountable once the rules used by cells are known. In a quickly growing number of cases, it is bioinformaticians who recognize these rules from patterns they observe in large amounts of sequence data. It is the surprisingly small number of tools commonly used by molecular biologists, however, that both generates the raw data needed for such analyses and tests the biological significance of possible underlying rules. A set of roughly six different laboratory techniques, taken together, defines the entire discipline of molecular biology. These techniques are briefly described in this section.

Restriction Enzyme Digests

The Nobel Prize–winning work of Wilkins, Watson, and Crick in 1953 told the story of how DNA could act as the genetic material. Subsequent experiments confirmed this hypothesis, but it was not until nearly 20 years later that H. Smith and others made a serendipitous discovery that allowed researchers to manipulate DNA molecules in a specific fashion and to begin to decipher DNA's actual information content. In the course of studying what causes some bacterial cells to better defend themselves against viral infections, Smith and his colleagues found that bacteria produced enzymes that introduce breaks in double-stranded DNA molecules whenever they encounter a specific string of nucleotides. These proteins, **restriction enzymes**, can be isolated from bacterial cells and used in research laboratories as precise “scissors” that let biologists cut (and later “paste” together) DNA molecules. The very first of these proteins to be characterized was given the name *EcoRI* (*Eco* because it was isolated from *Escherichia coli*; *R* because it restricted DNA; *I* because it was the first such enzyme found in *E. coli*). *EcoRI* was found to cleave DNA molecules between G and A nucleotides whenever it encountered them in the sequence 5'-GAATTC-3' (Figure 1.10). Since then over 300 types of restriction enzymes have been found in other bacterial species that recognize and cut DNA molecules at a wide variety of specific sequences. Notice that *EcoRI*, like many restriction enzymes, cleaves (or digests) double-stranded DNA molecules in a way that leaves a bit of single-stranded DNA at the end of each fragment. The nucleotide sequences of those single-stranded regions (5'-AATT-3' in the case of *EcoRI*) are naturally complementary to each other. The resulting potential for base pairing makes these **sticky ends**

capable of holding two DNA fragments together until another special enzyme called **ligase** can permanently link (or ligate) them together again by rebuilding the phosphodiester bonds that were broken by the restriction enzyme. Restriction enzymes that do not give rise to sticky ends create **blunt ends** that can be ligated to other blunt-ended DNA molecules.

The string of nucleotides recognized by *EcoRI*, its **restriction site**, should occur randomly in DNA sequences only once every $(1/4)^n$ base pairs where n equals 6 or, on average, once every 4,096 base pairs. Some restriction enzymes have smaller restriction sites (such as *HinfI*, which finds and restricts at 5'-GATC-3' on average once every 256 base pairs) while others have larger sites (such as *NotI*, which finds and restricts at 5'-GCGGCCGC-3' on average once every 65,536 base pairs). Simply cutting a DNA molecule and determining how many fragments are made and the order in which the breaks occur when multiple restriction enzymes are used provide some limited insight into the specific organization and sequence of that DNA molecule. Such experiments are termed **restriction mapping**. Restriction enzymes also allowed the isolation and experimental manipulation of individual genes for the very first time.

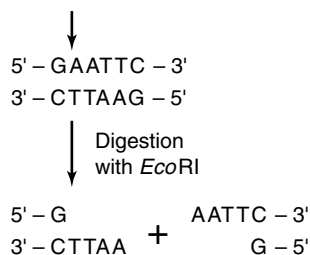


FIGURE 1.10

*Digestion with the restriction enzyme *EcoRI*. *EcoRI* introduces staggered breaks in DNA molecules whenever it encounters its recognition site (5'-GAATTC-3'). The single-stranded overhanging regions that result are capable of base pairing with each other and are referred to as sticky ends.*

Gel Electrophoresis

When dealing with a genome that is millions of base pairs long (such as *E. coli*'s) or even billions of base pairs long (such as the human genome), complete digestion with even a very specific restriction enzyme such as *NotI* can yield hundreds of thousands of DNA fragments. Separating all of those different fragments from each other is commonly accomplished by **gel electrophoresis**, another of the tools of molecular biology. In gel electrophoresis, DNA (or RNA or protein) fragments are loaded into indentations called wells at one end of a porous gel-like matrix typically made either from agarose or acrylamide. When an electric field is applied across these gels, the charged molecules naturally migrate toward one of the two electrodes generating the field. DNA (and RNA) with its negatively charged phosphate backbone is drawn toward the positively charged electrode. Very simply, small molecules have an easier time working their way through the gel's matrix than larger ones, and separation of the molecules on the basis of their size occurs (Figure 1.11). Larger molecules remain closer to the wells than smaller molecules, which migrate more quickly.

Blotting and Hybridization

Finding the single piece of DNA that contains a specific gene among hundreds or thousands is very much akin to the idea of finding a needle in a haystack even when the DNA fragments are size fractionated. Molecular biologists routinely

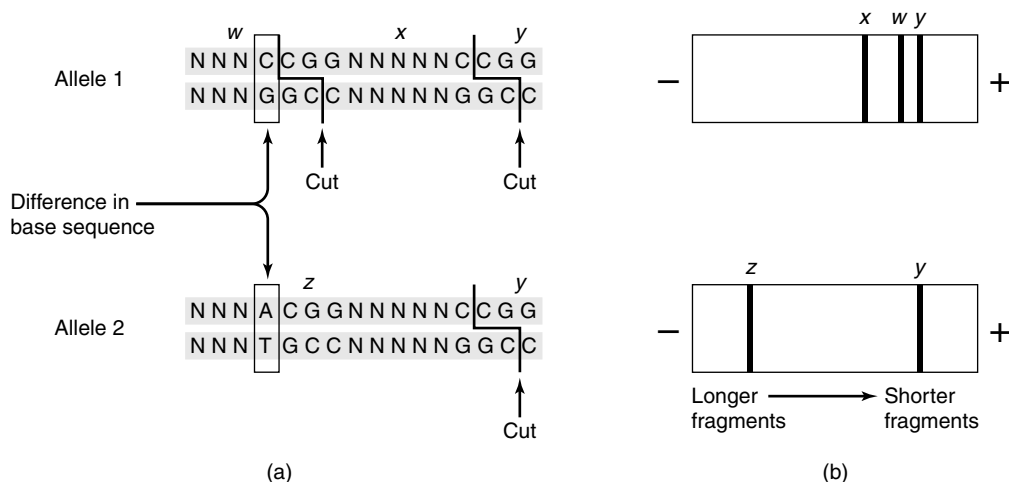


FIGURE 1.11 Gel electrophoresis allows DNA fragments to be separated on the basis of their size. (a) Differences in DNA sequences can cause differences in the places where restriction enzymes break double-stranded DNA molecules. (b) The differences in the sizes of the restriction fragments that result can be easily detected by gel electrophoresis. Allele 1 has three bands (corresponding to regions w, x, and y), while allele 2 gives rise to two bands corresponding to regions z and y.

employ another technique, **blotting and hybridization**, to draw attention to the one fragment they wish to study. In blotting, polynucleotides are transferred from the fragile gel that was used to separate them onto a more solid support such as a piece of nitrocellulose paper or a nylon membrane. This blotting process is mechanically simple and entails placing the membrane in contact with the gel and then using capillary action to pull the DNA up from the gel and onto the membrane. While water molecules can pass through the membrane (and be drawn into absorbent paper towels or a weak vacuum above it), DNA molecules are too large and remain associated with the membrane in the same relative positions that they had moved to during gel electrophoresis. Ultraviolet light or even simple baking can then be used to permanently attach the DNA fragments to the membrane.

A membrane prepared in this way is then ready for the second step in this detection process. Hybridization occurs when a labeled fragment of single-stranded DNA called a **probe** is allowed to base pair with the nucleic acids that have been transferred to a membrane. Typically 20 or more nucleotides in length, probes can be chosen on the basis of their being likely to find only one fragment of DNA on the membrane to which they can base pair. Probes can be chemically synthesized from scratch or can be fragments of DNA that have already been isolated in other experiments and even from related genes in different organisms. Any of a number of means can be used to label or tag a probe ranging from radioactivity to fluorescent dyes and even attaching enzymes that catalyze unusual reactions. These probes are allowed to wash over a membrane (often for several hours or even overnight) as part of a watery mix that also contains salt, pH buffers, and detergent. The stringency of the hybridization, particularly salt concentration and temperature, can be manipulated to allow probes to bind to sequences with less than perfect matches. At the end of the hybridization procedure, unbound probe is washed off and the membrane is examined to see where base pairing between the probe and its target sequence has occurred.

A variant of these membrane-based hybridization systems is the powerful **microarray** or DNA chip technology. Here, thousands and even tens of thousands of nucleotide sequences are each affixed to individual positions on the surface of a small silica (glass) chip. Fluorescently labeled copies of the RNA transcripts (cDNAs, described further in Chapter 6) from an organism being studied can then be washed over that surface and allowed to hybridize to complementary nucleotides. After washing, a laser is used to excite the fluorescent tags and then photodetectors quantify the amount of signal associated with each spot of known sequence. A popular application of this methodology results in the determination of relative RNA levels associated with huge numbers of known and predicted genes in a single experiment for a variety of organisms using commercially available microarrays. Quite literally, accurate measurements of *every* single gene (and even every processing variant of every gene) can be precisely assessed. The sensitivity of DNA chip technology is truly remarkable in that it can confirm the presence of as little as one transcript being present in every tenth cell of an experiment. Significant computational efforts are associated with the generation of such chips (ensuring the distinctiveness of each of the thousands of bound probes

alone is challenging) as well as the interpretation of their results. (Variation among replicate experiments, evaluation of differences between test and control conditions, and determination of expression associations are all complicated by an abundance of data.)

Cloning

While cells manipulate and extract information from single DNA molecules on a routine basis, molecular biologists typically require quantities of material that are almost visible to the naked eye (many millions of molecules) for most of their analyses. DNA sequencing reactions (described below) in particular require higher purity and larger amounts of DNA than can be practically obtained through restriction enzyme digestion of genomic DNA and gel electrophoresis. A fairly simple solution to this problem has been to invoke the assistance of cells in the generation of sufficient quantities and qualities of specific DNA molecules for such purposes. In essence, **cloning** involves the insertion of specific DNA fragments into chromosome-like carriers called **vectors** that allow their replication in (and isolation from) living cells. Since all the copies of the fragment are identical, they are known as **molecular clones** and they can be purified for immediate study or stored in collections known as libraries for future analyses.

Once a restriction fragment that contains a sequence of particular interest has been generated as described above, its sticky ends can be used to help ligate it into a vector that has been cut with a restriction enzyme that has complementary sticky ends. The first vectors to be used were derived from bacterial viruses and from small extra-chromosomal pieces of DNA in prokaryotic cells called plasmids. These vectors are easy to manipulate in the laboratory and are especially useful for cloning relatively small pieces of DNA (ranging in size from dozens to 25,000 nucleotides in length). Newer alternatives derived from bacterial and yeast chromosomes are better suited for very large fragments of DNA (ranging from 100,000 to 1,000,000 base pairs long), but are not as amenable to handling and characterization. All vectors must have several features in common to be useful to molecular biologists. Those features include sequences that allow them to be replicated inside of living cells, sequences that confer a novel ability to their host cell so their presence can be detected, and distinguishing physical traits (such as size or shape) that allow them to be separated from the host cell's DNA.

A collection of genes, each of which is cloned into a vector, is known as a **genetic library**. An ideal genomic library would contain one copy of every segment of an organism's DNA. For example, if a 4,600,000-nucleotide-long genome (such as *E. coli*'s) were completely digested with a restriction enzyme such as *EcoRI*, then a total of more than 1,000 DNA fragments with an average length of 4,096 base pairs would each need to be cloned to make a complete genomic library. The number of clones (genome size divided by average fragment length) in such a perfect genomic library defines a **genomic equivalent**. Unfortunately, making a genomic library cannot be accomplished by simply digesting the genomic DNA of a single cell and making clones of each fragment. The cloning process is not efficient and it is usually necessary to harvest DNA from hundreds

or thousands of cells to clone a single fragment. Further, the random nature of the cloning process ensures that some fragments will be cloned multiple times while others are not represented at all in one genomic equivalent. Increasing the number of clones in a genomic library increases the likelihood that it will contain at least one copy of any given segment of DNA. A genomic library with four to five genomic equivalents has, on average, four to five copies of every DNA segment and a 95% chance of containing at least one copy of any particular portion of the organism's genome. Details of this calculation are provided in Chapter 6. These realities have two practical implications: (1) Vectors that allow the cloning of larger fragments are better for making genomic libraries because fewer clones are needed to make a genomic equivalent, and (2) cloning the last 5% of a genome is often as difficult as cloning the first 95%.

In many cases a useful alternative to a genomic library is a **cDNA library**. The portions of a genome that are typically of greatest interest are those that correspond to the regions that code for proteins. One thing that all protein coding regions have in common is the fact that they are all converted into mRNAs before they are translated by ribosomes. Those mRNAs can be separated from all the other polynucleotides within a cell by means of a special enzyme called **reverse transcriptase**, which converts them back into complementary DNA (cDNA) sequences and then clones those cDNAs as part of a library. Simply showing up in a cDNA library is often enough to attach significance to a portion of a genome since cells usually only make mRNA copies of genes that are functionally important. Further, the relative abundance of cDNAs within a library from any given organism or cell type gives an indication as to how much a particular gene is expressed. A disadvantage to cDNA sequences, though, is that they typically contain only the information that is used by ribosomes in making proteins and not the important regulatory sequences and introns usually associated with genes. As a result, complete understanding of a gene's structure and function usually comes only after characterization of both its genomic and cDNA clone. The creation of screening libraries to determine which clones contain sequences of interest is accomplished by similar kinds of blotting and hybridization strategies used to distinguish between one DNA fragment and another.

Polymerase Chain Reaction

Molecular cloning provides a means of organizing and indefinitely maintaining specific portions of a genome in a way that also allows large quantities of that region to be isolated and used in more detailed analyses. When little information about the sequence of the region is known and large quantities of a region are needed, a powerful alternative to cloning is the use of the **polymerase chain reaction (PCR)** method. Developed by K. Mullis in 1985, PCR relies on an understanding of two idiosyncrasies associated with DNA polymerases (the enzymes responsible for replicating DNA during cell division). First, like RNA polymerases, all DNA polymerases add new nucleotides onto just the 3' (and never the 5') end of a DNA strand during synthesis; hence, there is a definite directionality to DNA synthesis. Second, while it is the job of a DNA polymerase

to make double-stranded DNA molecules by using the information inherent to a single-stranded DNA molecule, DNA polymerases can only begin DNA synthesis by adding nucleotides onto the end of an existing DNA strand (Figure 1.12). PCR takes advantage of those two quirks of DNA polymerases to drive the replication of very specific regions of a genome that are of interest to a molecular biologist. One double-stranded copy of such a region can be replicated into two double-stranded copies after one round of amplification. Those two copies can each be duplicated to give rise to four copies during a second round of amplification. In just a couple of hours and after 20 to 30 such rounds of exponential amplification, a specific region of DNA is usually present in enormously higher quantities (theoretically, 2^{20} to 2^{30} or 1,048,576 to 1,073,741,824 copies, assuming that only one copy was present at the start of the process) than other DNA sequences present at the start of the process (Figure 1.12). These amplified DNA molecules are produced much more quickly and efficiently than those obtained from clones yet they can be used in many of the same ways. The amplifying nature of PCR gives it the additional advantage of being able to start with much smaller quantities of material (such as those typically associated with museum or even fossil and forensic specimens) than are usually amenable to cloning experiments.

DNA synthesis occurs only at specific segments of a genome during PCR amplification because of the specific primers that are added to the reaction mixture at the very start of the amplification process. Like the probes used in hybridization experiments, PCR primers are typically 20 or more nucleotides in length to ensure that each can bind specifically to only one target sequence within an organism's genome. The specific sequences used to make primers in the first place typically come from DNA sequence analyses of similar regions in closely related organisms and at some point usually require the more laborious process of cloning and screening described earlier.

DNA Sequencing

The ultimate molecular characterization of any piece of DNA comes from determining the order or sequence of its component nucleotides. All DNA sequencing strategies involve the same three steps: (1) the generation of a complete set of subfragments for the region being studied whose lengths differ from each other by a single nucleotide, (2) labeling of each fragment with one of four different tags that are dependent on the fragment's terminal nucleotide, and (3) separating those fragments by size in a way (usually some form of acrylamide gel electrophoresis) that allows the sequence to be read by detecting the order in which the different tags are seen.

A. M. Maxam and W. Gilbert developed the first successful DNA sequencing strategy in the late 1970s. However, the **Maxam-Gilbert method** relied on chemical degradation to generate the DNA subfragments needed for sequencing, so it quickly fell out of favor when a safer and more efficient DNA polymerase-based method was developed by F. Sanger a few years later. The Sanger approach is sometimes referred to as a **chain-termination method** because the subset of

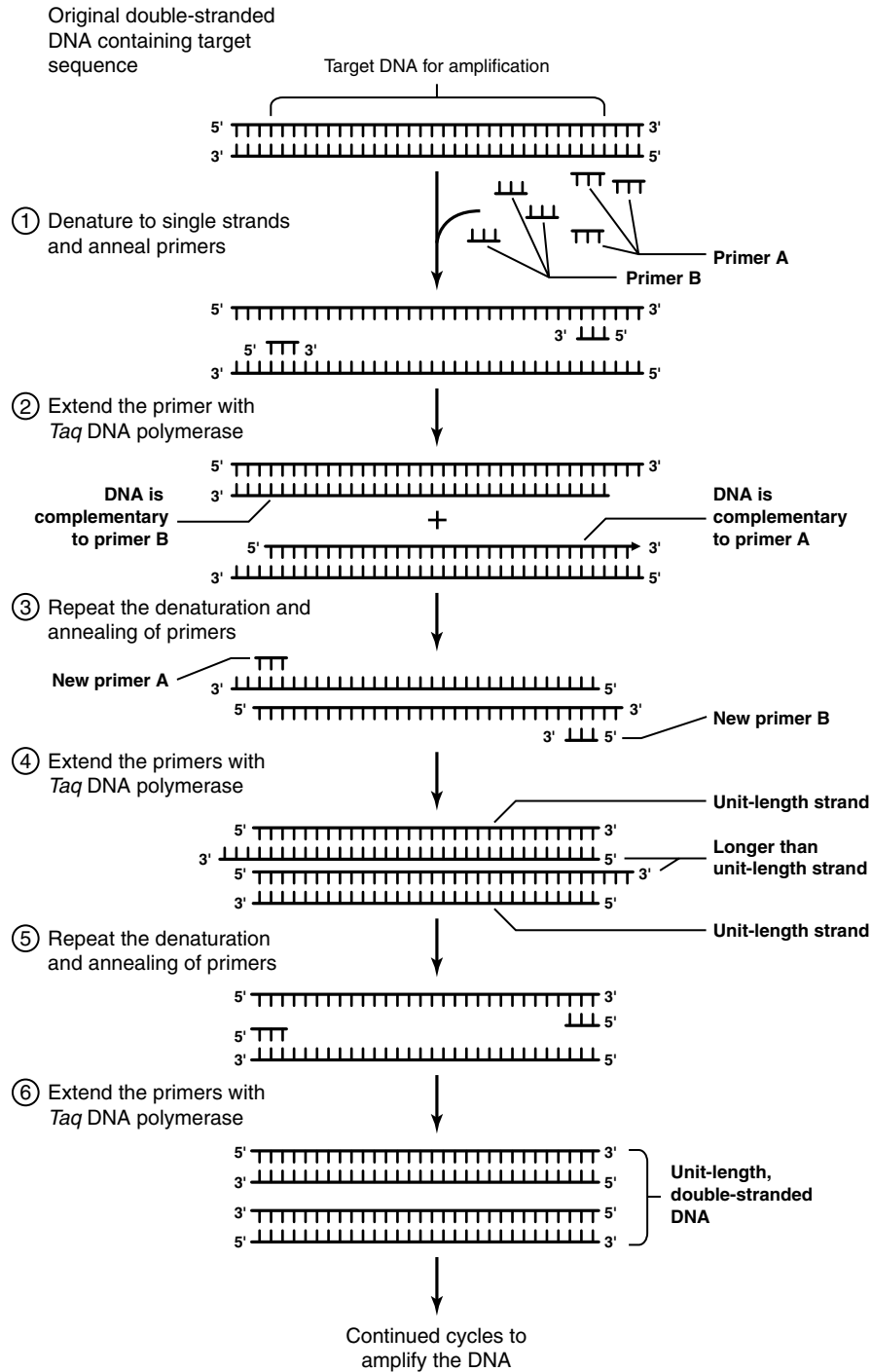


FIGURE 1.12 *The polymerase chain reaction.*

fragments needed for sequencing is generated by the incorporation of modified nucleotides that prevent a DNA polymerase from adding any additional bases to the chain (Figure 1.13). Those chain-terminating nucleotides differ from their normal counterparts in that they are missing their 3' hydroxyl group (see Figure 1.2) onto which the next nucleotide in a growing DNA strand is usually attached. They usually also have a tag such as a fluorescent dye that allows for their detection when they are size fractionated.

In an ideal sequencing reaction the modified and normal nucleotides are mixed in a ratio that allows DNA polymerases to randomly incorporate a chain terminator only once in every 500 or so nucleotides so that a complete set of sub-fragments from a region up to 1,000 base pairs long can be sequenced at one time. Improvements in the methodology and particularly in the automation of DNA sequencing now make it possible for a single analyst to generate tens and even hundreds of thousands of base pairs of DNA sequence data in a single day—quite a contrast to the 50 base pairs of sequence data for which Gilbert (along with Sanger) shared a Nobel Prize in 1980. Still, the short size (roughly 1,000 base pairs) of each piece of sequence information relative to the overall size of a genome (billions of base pairs in many eukaryotes) and even to genes (sometimes hundreds of thousands of nucleotides long) can make assembling sequences of useful size a computationally challenging task.

Genomic Information Content

As mentioned earlier, an organism's genome can be millions or billions of base pairs long. It was possible to obtain interesting insights into how complex a genome was and how much useful information it contained long before it was possible to determine the order in which its nucleotides were arranged. Even now that automated sequencing has made the sequencing of complete genomes feasible, those earliest approaches used to characterize genomes as a whole still provide useful insights into the quantity and complexity of their genetic information.

C-Value Paradox

In 1948 the discovery was made that the amount of DNA in every cell of a given organism is the same. These measures of a cell's total DNA content are referred to as **C values** (Figure 1.14). Interestingly, while genome size within a species is constant, large variations across species lines have been observed but not in a way that correlated well with organismal complexity. The absence of a perfect correlation between complexity and genome size is often called the **C-value paradox** (Figure 1.14). Total DNA amounts often differ by 100-fold or more even between very similar species. The clear (but difficult to prove) implication is that a large portion of the DNA in some organisms is expendable and does not contribute significantly to an organism's complexity.

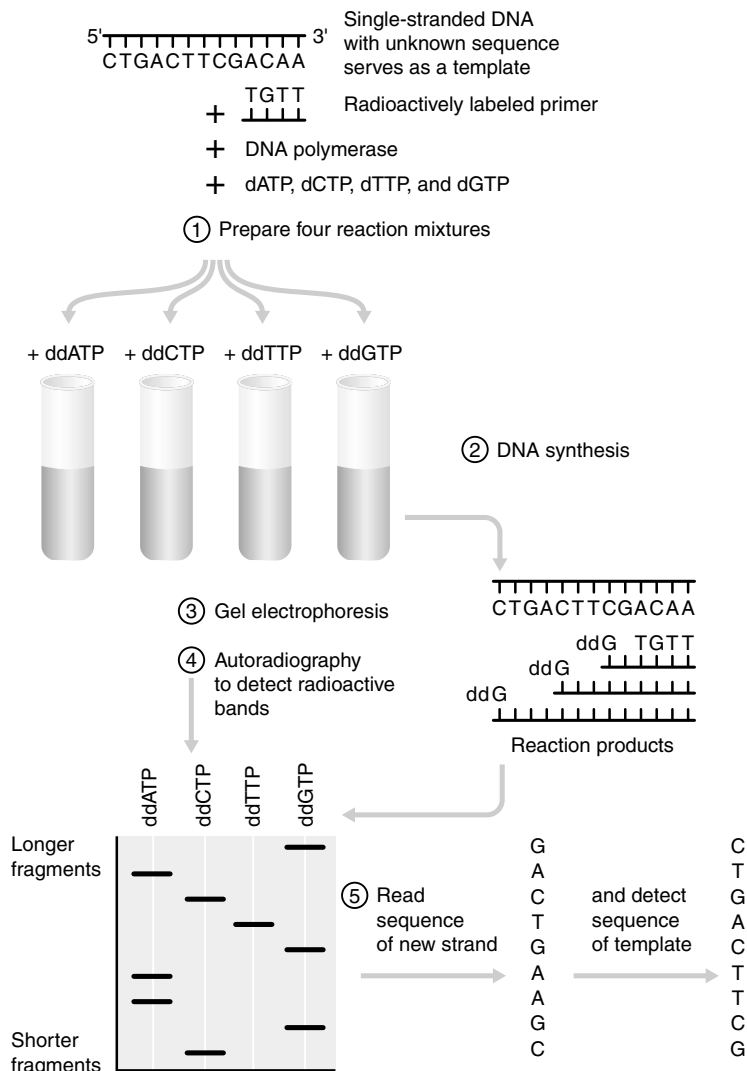


FIGURE 1.13 *The Sanger dideoxy method of DNA sequencing.*

Reassociation Kinetics

When the complementary strands of double-stranded DNA are separated (denatured) by heat or alkali treatment, they can readily reform (renature) a conventional double-stranded structure when conditions are returned to those typically encountered inside a cell. Quite a bit can be learned about the structure of genomes simply by examining the way in which their denatured DNA renatures. In the simplest terms, the more unique a sequence in a genome, the more time it will take for each strand to find and hybridize to its complement. Studies

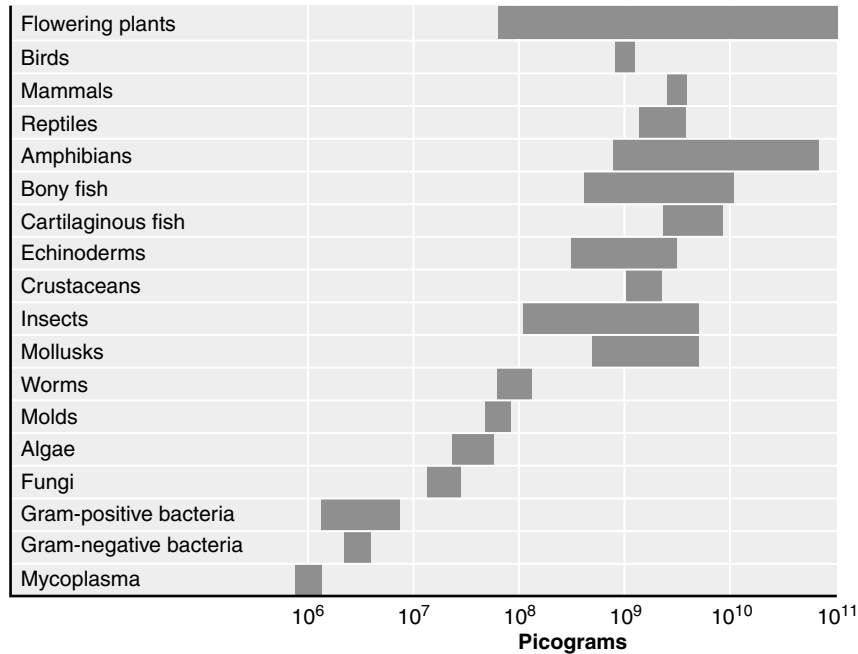


FIGURE 1.14 *The DNA contents of the haploid genomes of a variety of different organisms. C values are generally correlated to morphological complexity in simpler eukaryotes but vary significantly among more complex eukaryotes. The range of DNA content within a phylum is indicated by the shaded areas.*

by R. Britten and others in the 1960s revealed that the time course of DNA re-
 naturation could be conveniently described by a **cot equation**. The cot equation
 relates the fraction of single-stranded DNA remaining (c/c_0) after t seconds of re-
 naturing multiplied by the amount of denatured genomic DNA at the start of the
 experiment (c_0). A specific value, $c_0 t_{1/2}$, derived from such experiments can be ob-
 tained for any organism and is directly proportional to the number of nucleotides
 of nonrepeated sequence within the organism's DNA. Measuring the time ($t_{1/2}$)
 required for one-half of the single-stranded DNA to renature ($c/c_0 = 0.5$) allows
 an experimental determination of the total amount of unique genetic information
 encoded in a genome. Since the $c_0 t_{1/2}$ is the product of the concentration and time
 required to proceed halfway, a greater $c_0 t_{1/2}$ implies a slower reaction and reflects
 a situation in which there are fewer copies of any particular sequence within a
 given mass of DNA.

For example, if the c_0 of DNA is 10 picograms (pg), it will contain 2,000
 copies of each sequence in a bacterial genome whose size is 0.005 pg, but will con-
 tain only 2 copies of each sequence present in a eukaryotic genome of size 5 pg.
 So, the same absolute amount of DNA (c_0) will provide a concentration of each
 eukaryotic sequence that is $2,000/2 = 1,000$ times lower than that of each bacte-
 rial sequence. Since the rate of renaturation depends on the concentration of

complementary sequences, for the eukaryotic sequences to renature at the same rate as the bacterial sequences it would be necessary to have 1,000 times as much eukaryotic DNA. If the starting concentrations of genomic DNA are the same, however, the $c_0t_{1/2}$ of the eukaryotic reaction will be 1,000 times longer than the $c_0t_{1/2}$ of the bacterial reaction if no particular sequence occurs more than one time in each genome. If an organism's genome contains multiple copies of the same sequence, then its $c_0t_{1/2}$ values should be less than those for another with the same genome size but no repeated sequences. In short, $c_0t_{1/2}$ is a measure of the total length of different sequences within a genome and can be used to describe genomic complexity. While the total amount of DNA present within any given genome (its C value) may not be indicative of the overall complexity of an organism, the amount of single-copy DNA it contains (its $c_0t_{1/2}$) usually is (Figure 1.14). Disparities between C values and $c_0t_{1/2}$ usually indicate that an organism contains multiple copies of disposable DNA sequences often referred to as **junk DNA**. Repeated sequences within this junk DNA differ widely in terms of their complexity (ranging from single and dinucleotide repeat units to repeat units that are hundreds or even thousands of nucleotides long) and distribution (arranged in local clusters or scattered relatively randomly) within a genome, as will be discussed in greater detail in Chapter 6.

Chapter Summary

DNA is an information storage molecule for cells. The specific order of its four different nucleotides is transcribed by RNA polymerases into mRNAs that are then translated by ribosomes into proteins. Twenty different amino acids are used to make proteins, and the specific order and composition of those building blocks play an important role in establishing and maintaining the structure and function of enzymes. Molecular biologists have a fairly limited set of tools to study DNA and its information content. Restriction enzymes cut DNA molecules when they encounter specific strings of nucleotides. Electrophoresis allows such DNA fragments to be separated on the basis of their size and charge. Blotting and hybridization techniques allow specific DNA fragments to be found within a mixture of other DNA fragments, while cloning allows specific molecules to be propagated and used over and over again. PCR is a popular and versatile alternative to cloning that allows specific DNA fragments to be amplified and characterized. Ultimate characterization of a DNA molecule comes from determining the order of its nucleotides and can be accomplished by DNA sequencing techniques. Reassociation kinetics have revealed that a cell's DNA content (its C value) does not always correspond directly to an organism's information content due to the large amounts of "junk DNA" found in complex organisms.

Readings for Greater Depth

Numerous textbooks give excellent overviews of molecular biology. For a concise description of genes and our understanding of them, try P. Portin, 1993, The concept of the gene: Short history and present status, *Q. Rev. Biol.* **68**: 173–223.

Discovery of the structure of DNA won Watson and Crick a Nobel Prize. Their classic single-page paper describing their insight is J. D. Watson and F. H. C. Crick, 1953, Genetical implications of the structure of deoxyribonucleic acid, *Nature* **171**: 964–967.

Watson himself relates the dramatic and often unscientific race to discover the structure of DNA in a very informative and often humorous book that has also been made into a major motion picture: J. D. Watson, 1968, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. Atheneum, New York.

Reassociation experiments and “junk DNA” are both described in R. J. Britten, D. E. Graham, and B. R. Neufeld, 1974, Analysis of repeating DNA sequences by reassociation. *Methods Enzymol.* **29**: 363–418.

A general text that describes the structure of genes and the experimental techniques used to study them is B. Lewin, 2000, *Genes VII*, Oxford University Press, New York.

Questions and Problems

- * **1.1** Deoxyribonucleic acid (DNA) differs from ribonucleic acid (RNA) in two ways: (1) RNA uses the nitrogenous base uracil in place of DNA's thymine, and (2) the hydroxyl (OH) group attached to the 2' carbon of the deoxyribose sugar of RNA is replaced with just a hydrogen (H) in DNA. Sketch the chemical structures of the deoxyribose sugar used by DNA and the ribose sugar used by RNA.
- 1.2** What is the complementary sequence to the following string of nucleotides? Be sure to label the 5' and 3' ends of the sequence that you write. 5'-GGATCGTAGCCTA-3'.
- * **1.3** Diagram the “central dogma” of molecular biology complete with labels that indicate the portions that correspond to transcription and translation and indicate what enzymes are responsible for those important steps.
- 1.4** Organic molecules that contain hydroxyl groups (—OH) are called alcohols. Would you expect such molecules to be hydrophobic or hydrophilic? Why?
- * **1.5** Examine the chemical structures of the amino acid R groups shown in Figure 1.5b. What atom(s) is found in the R groups that are in the hydrophilic amino acids that generally is absent in the nonpolar group?

1.6 How frequently would you expect to find the sequence of nucleotides provided in Question 1.2 in a DNA molecule simply as a result of random chance? Assume that each of the four nucleotides occurs with the same frequency.

* **1.7** How many nucleotides long would a DNA sequence need to be in order for it to not be found by chance more than once in a genome whose size is 3 billion base pairs long?

1.8 Distinguish between positive and negative regulation of gene expression.

* **1.9** What sequence of amino acids would the following RNA sequence code for if it were to be translated by a ribosome?: 5'-AUG GGA UGU CGC CGA AAC-3'. What sequence of amino acids would it code for if the first nucleotide were deleted and another "A" were added to the 3' end of the RNA sequence?

1.10 A circular piece of DNA known to be 4,000 bp long is cut into two pieces when treated with the restriction enzyme *EcoRI*: One piece is 3,000 bp long and the other is 1,000 bp long. Another restriction enzyme, *BamHI*, cuts the same DNA molecule into three pieces of the following lengths: 2,500, 1,200, and 300 bp. When both *EcoRI* and *BamHI* are used to cut the DNA molecule together, fragments of the following sizes are generated: 1,600, 1,200, 900, 200, and 100 bp. Use this information to make a restriction enzyme map of this circular DNA molecule.

* **1.11** How does a cDNA library differ from a genomic library?