

## CHAPTER 3

# Nucleotides, Nucleic Acids, and Genetic Information

## Chapter Contents

### 1 Nucleotides

### 2 Introduction to Nucleic Acid Structure

- A Nucleic Acids Are Polymers of Nucleotides
- B DNA Forms a Double Helix
- C RNA Is a Single-Stranded Nucleic Acid

### 3 Overview of Nucleic Acid Function

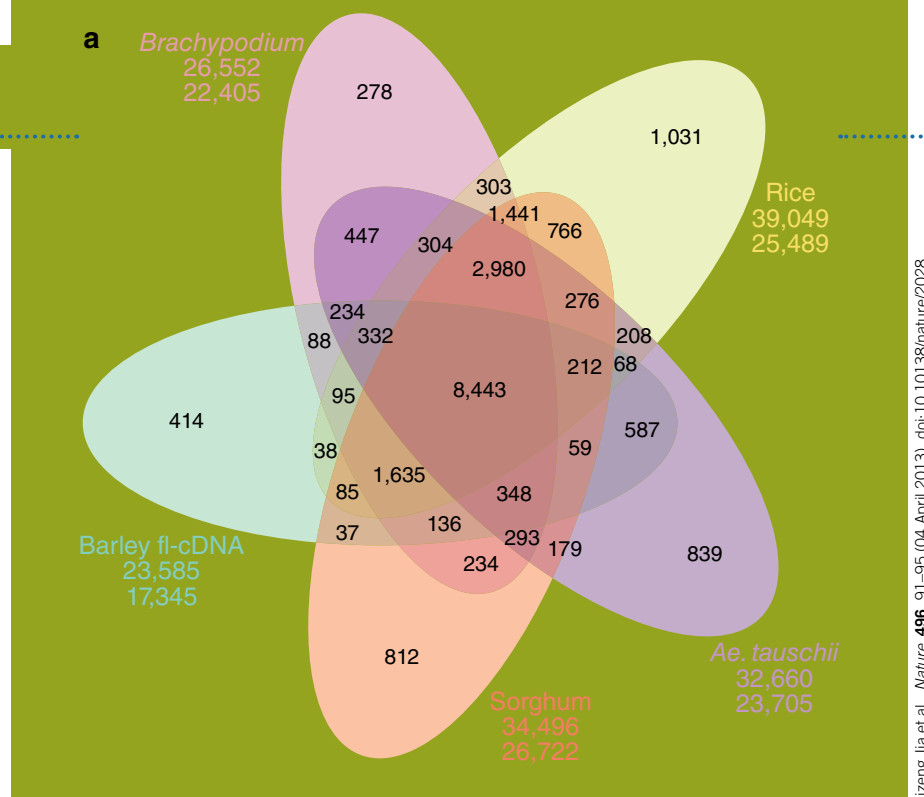
- A DNA Carries Genetic Information
- B Genes Direct Protein Synthesis

### 4 Nucleic Acid Sequencing

- A Restriction Endonucleases Cleave DNA at Specific Sequences
- B Electrophoresis Separates Nucleic Acids According to Size
- C Traditional DNA Sequencing Uses the Chain-Terminator Method
- D Next-Generation Sequencing Technologies Are Massively Parallel
- E Entire Genomes Have Been Sequenced
- F Evolution Results from Sequence Mutations

### 5 Manipulating DNA

- A Cloned DNA Is an Amplified Copy
- B DNA Libraries Are Collections of Cloned DNA
- C DNA Is Amplified by the Polymerase Chain Reaction
- D Recombinant DNA Technology Has Numerous Practical Applications



Families of genes shared by five different species of grains are depicted by overlapping shapes. By identifying the genes in samples of DNA, researchers can focus on the genetic features that have a common purpose or provide some unique function to an organism.

Despite obvious differences in lifestyle and macroscopic appearance, organisms exhibit striking similarity at the molecular level. The structures and metabolic activities of all cells rely on a common set of molecules that includes amino acids, carbohydrates, lipids, and nucleotides, as well as their polymeric forms. Each type of compound can be described in terms of its chemical makeup, its interactions with other molecules, and its physiological function. We begin our survey of biomolecules with a discussion of the **nucleotides** and their polymers, the **nucleic acids**.

Nucleotides are involved in nearly every facet of cellular life. Specifically, they participate in oxidation–reduction reactions, energy transfer, intracellular signaling, and biosynthetic reactions. Their polymers, the nucleic acids DNA and RNA, are the primary players in the storage and decoding of genetic information. Nucleotides and nucleic acids also perform structural and catalytic roles in cells. No other class of molecules participates in such varied functions or in so many functions that are essential for life.

It is becoming increasingly clear that the appearance of nucleotides permitted the evolution of organisms that could harvest and store energy from their surroundings and, most importantly, could make copies of themselves. Although the chemical and biological details of early life-forms are the subject of speculation, it is incontrovertible that life as we know it is inextricably linked to the chemistry of nucleotides and nucleic acids.

In this chapter, we briefly examine the structures of nucleotides and the nucleic acids DNA and RNA. We also consider how the chemistry of these molecules allows them to carry biological information in the form of a sequence of nucleotides. This information is expressed by the transcription of a segment of DNA to yield RNA, which is then translated to form protein. Because a cell's structure and function ultimately depend on its genetic makeup, we discuss how genomic sequences provide information about evolution, metabolism, and disease. Finally,

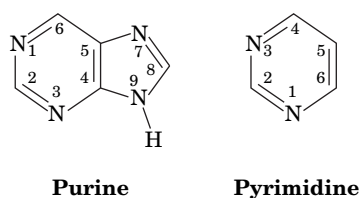
we consider some of the techniques used in manipulating DNA in the laboratory. In later chapters, we will examine in greater detail the participation of nucleotides and nucleic acids in metabolic processes. Chapter 24 includes additional information about nucleic acid structures, DNA's interactions with proteins, and DNA packaging in cells, as a prelude to several chapters discussing the roles of nucleic acids in the storage and expression of genetic information.

# 1 Nucleotides

## KEY CONCEPTS

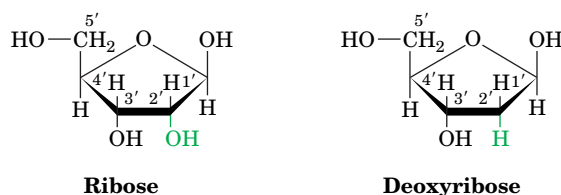
- The nitrogenous bases of nucleotides include two types of purines and three types of pyrimidines.
- A nucleotide consists of a nitrogenous base, a ribose or deoxyribose sugar, and one or more phosphate groups.
- DNA contains adenine, guanine, cytosine, and thymine deoxyribonucleotides, whereas RNA contains adenine, guanine, cytosine, and uracil ribonucleotides.

Nucleotides are ubiquitous molecules with considerable structural diversity. *There are eight common varieties of nucleotides, each composed of a nitrogenous **base** linked to a sugar to which at least one phosphate group is also attached.* The bases of nucleotides are planar, aromatic, heterocyclic molecules that are structural derivatives of either **purine** or **pyrimidine** (although they are not synthesized *in vivo* from either of these organic compounds).



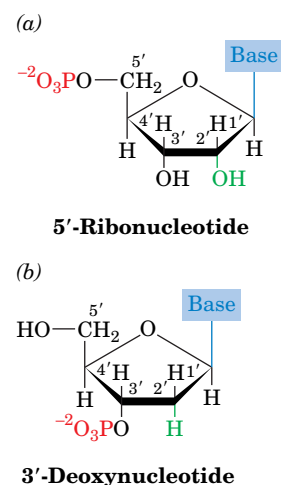
The most common purines are **adenine (A)** and **guanine (G)**, and the major pyrimidines are **cytosine (C)**, **uracil (U)**, and **thymine (T)**. The purines form bonds to a five-carbon sugar (a pentose) via their N9 atoms, whereas pyrimidines do so through their N1 atoms (**Table 3-1**).

In **ribonucleotides**, the pentose is **ribose**, while in **deoxyribonucleotides** (or just **deoxynucleotides**), the sugar is **2'-deoxyribose** (i.e., the carbon at position 2' lacks a hydroxyl group).



Note that the “primed” numbers refer to the atoms of the pentose; “unprimed” numbers refer to the atoms of the nitrogenous base.

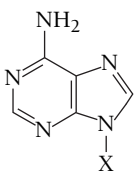
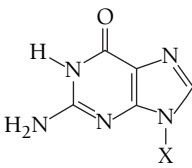
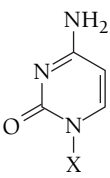
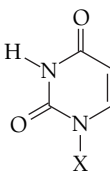
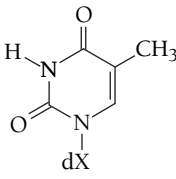
In a ribonucleotide or a deoxyribonucleotide, one or more phosphate groups are bonded to atom C3' or atom C5' of the pentose to form a 3'-nucleotide or a 5'-nucleotide, respectively (**Fig. 3-1**). When the phosphate group is absent, the compound is known as a **nucleoside**. A 5'-nucleotide can therefore be called a nucleoside-5'-phosphate. Nucleotides most commonly contain one to three



**FIG. 3-1 Chemical structures of nucleotides.** (a) A 5'-ribonucleotide and (b) a 3'-deoxynucleotide. The purine or pyrimidine base is linked to C1' of the pentose and at least one phosphate (red) is also attached. A nucleoside consists only of a base and a pentose.

**? Compare the net charge on a nucleoside and a nucleotide.**

**TABLE 3-1** Names and Abbreviations of Nucleic Acid Bases, Nucleosides, and Nucleotides

| Base Formula                                                                       | Base<br>(X = H)      | Nucleoside<br>(X = ribose <sup>a</sup> ) | Nucleotide <sup>b</sup><br>(X = ribose phosphate <sup>a</sup> ) |
|------------------------------------------------------------------------------------|----------------------|------------------------------------------|-----------------------------------------------------------------|
|    | Adenine<br>Ade<br>A  | Adenosine<br>Ado<br>A                    | Adenylic acid<br>Adenosine monophosphate<br>AMP                 |
|    | Guanine<br>Gua<br>G  | Guanosine<br>Guo<br>G                    | Guanylic acid<br>Guanosine monophosphate<br>GMP                 |
|    | Cytosine<br>Cyt<br>C | Cytidine<br>Cyd<br>C                     | Cytidylic acid<br>Cytidine monophosphate<br>CMP                 |
|   | Uracil<br>Ura<br>U   | Uridine<br>Urd<br>U                      | Uridylic acid<br>Uridine monophosphate<br>UMP                   |
|  | Thymine<br>Thy<br>T  | Deoxythymidine<br>dThd<br>dT             | Deoxythymidylic acid<br>Deoxythymidine monophosphate<br>dTMP    |

<sup>a</sup>The presence of a 2'-deoxyribose unit in place of ribose, as occurs in DNA, is implied by the prefixes "deoxy" or "d." For example, the deoxynucleoside of adenine is deoxyadenosine, or dA. However, for thymine-containing residues, which rarely occur in RNA, the prefix is redundant and may be dropped. The presence of a ribose unit may be explicitly implied by the prefix "ribo."

<sup>b</sup>The position of the phosphate group in a nucleotide may be explicitly specified as in, for example, 3'-AMP and 5'-GMP.

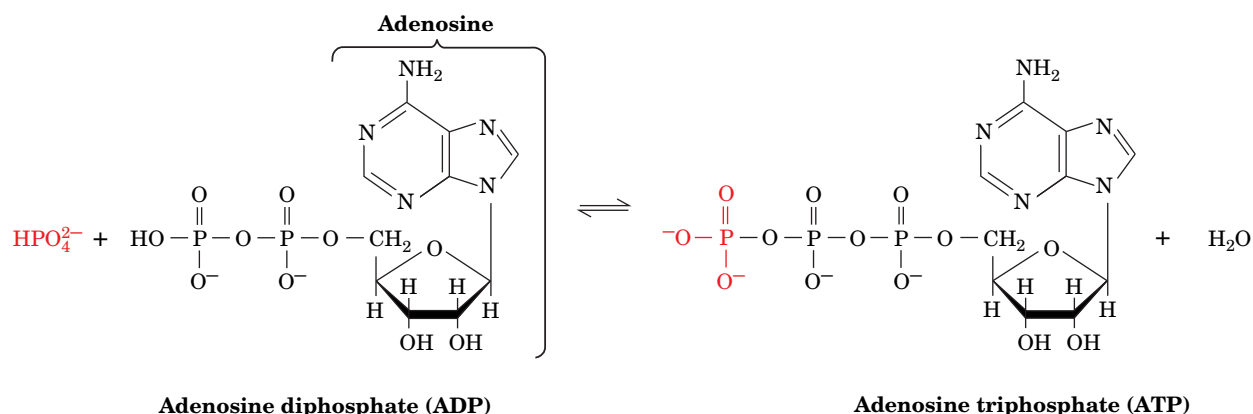
**? Without looking at the table, give the name of each base and its corresponding nucleoside.**

phosphate groups at the C5' position and are called nucleoside monophosphates, diphosphates, and triphosphates.

The structures, names, and abbreviations of the common bases, nucleosides, and nucleotides are given in Table 3-1. Ribonucleotides are components of **RNA (ribonucleic acid)**, whereas deoxynucleotides are components of **DNA (deoxyribonucleic acid)**. Adenine, guanine, and cytosine occur in both ribonucleotides and deoxynucleotides (accounting for six of the eight common nucleotides), but uracil occurs primarily in ribonucleotides and thymine occurs in deoxynucleotides. Free nucleotides, which are anionic, are almost always associated with the counterion  $Mg^{2+}$  in cells.

**Nucleotides Participate in Metabolic Reactions.** The bulk of the nucleotides in any cell are found in polymeric forms, as either DNA or RNA, whose primary functions are information storage and transfer. However, free nucleotides and nucleotide derivatives perform an enormous variety of metabolic functions not related to the management of genetic information.

Perhaps the best known nucleotide is **adenosine triphosphate (ATP)**, a nucleotide containing adenine, ribose, and a triphosphate group. ATP is often mistakenly referred to as an energy-storage molecule, but it is more accurately termed an energy carrier or energy transfer agent. The process of photosynthesis or the breakdown of metabolic fuels such as carbohydrates and fatty acids leads to the formation of ATP from **adenosine diphosphate (ADP)**:

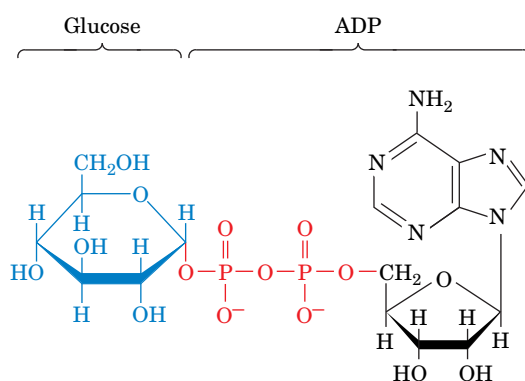


ATP diffuses throughout the cell to provide energy for other cellular work, such as biosynthetic reactions, ion transport, and cell movement. The chemical potential energy of ATP is made available when it transfers one (or two) of its phosphate groups to another molecule. This process can be represented by the reverse of the preceding reaction, namely, the hydrolysis of ATP to ADP. (As we will see in later chapters, the interconversion of ATP and ADP in the cell is not freely reversible, and free phosphate groups are seldom released directly from ATP.) The degree to which ATP participates in routine cellular activities is illustrated by calculations indicating that while the concentration of cellular ATP is relatively moderate ( $\sim 5$  mM), humans typically recycle their own weight of ATP each day.

Nucleotide derivatives participate in a wide variety of metabolic processes. For example, starch synthesis in plants proceeds by repeated additions of glucose units donated by ADP–glucose (**Fig. 3-2**). Other nucleotide derivatives, as we will see in later chapters, carry groups that undergo oxidation–reduction reactions. The attached group, which may be a small molecule such as glucose (**Fig. 3-2**) or even another nucleotide, is typically linked to the nucleotide through a mono- or diphosphate group.

### CHECKPOINT

- Identify the purines and pyrimidines commonly found in nucleic acids.
- Practice drawing the structures of adenine, adenosine, and adenyate.
- Describe the chemical differences between a ribonucleoside triphosphate and a deoxyribonucleoside monophosphate.



**FIG. 3-2 ADP–glucose.** In this nucleotide derivative, glucose (*blue*) is attached to adenosine (*black*) by a diphosphate group (*red*).

? Indicate the bonds that formed as a result of condensation reactions.

## 2 Introduction to Nucleic Acid Structure

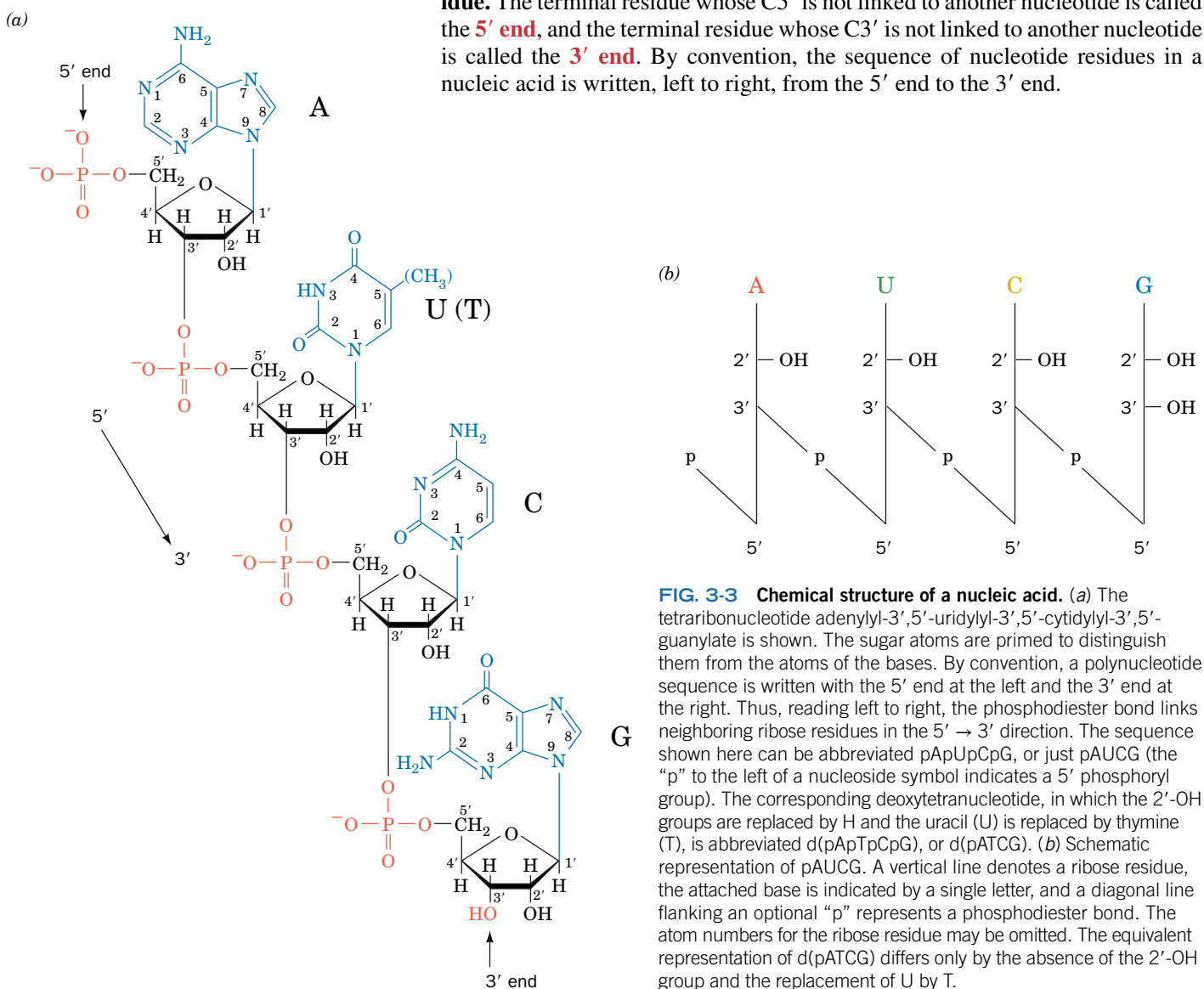
### KEY CONCEPTS

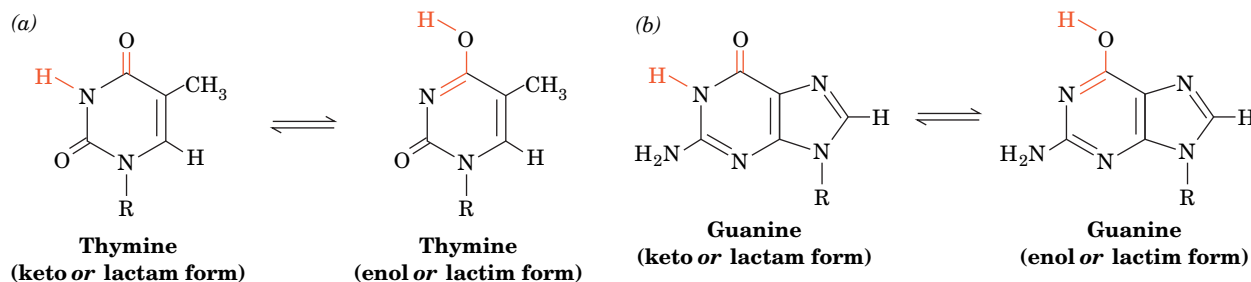
- Phosphodiester bonds link nucleotide residues in DNA and RNA.
- In the DNA double helix, two antiparallel polynucleotide strands wind around each other, interacting via hydrogen bonding between bases in opposite strands.
- RNA molecules are usually single-stranded and may form intramolecular base pairs.

Nucleotides can be joined to each other to form the polymers that are familiar to us as RNA and DNA. In this section, we describe the general features of these nucleic acids. Nucleic acid structure is considered further in Chapter 24.

### A Nucleic Acids Are Polymers of Nucleotides

The nucleic acids are chains of nucleotides whose phosphates bridge the 3' and 5' positions of neighboring ribose units (**Fig. 3-3**). The phosphates of these **poly-nucleotides** are acidic, so at physiological pH, nucleic acids are polyanions. The linkage between individual nucleotides is known as a **phosphodiester bond**, so named because the phosphate is esterified to two ribose units. Each nucleotide that has been incorporated into the polynucleotide is known as a **nucleotide residue**. The terminal residue whose C5' is not linked to another nucleotide is called the **5' end**, and the terminal residue whose C3' is not linked to another nucleotide is called the **3' end**. By convention, the sequence of nucleotide residues in a nucleic acid is written, left to right, from the 5' end to the 3' end.





**FIG. 3-4 Tautomeric forms of bases.** Some of the possible tautomeric forms of (a) thymine and (b) guanine are shown. Cytosine and adenine can undergo similar proton shifts.

? Draw the tautomers of adenine and cytosine.

The properties of a polymer such as a nucleic acid may be very different from the properties of the individual units, or **monomers**, before polymerization. As the size of the polymer increases from **dimer**, **trimer**, **tetramer**, and so on through **oligomer** (Greek: *oligo*, few), physical properties such as charge and solubility may change. In addition, a polymer of nonidentical residues has a property that its component monomers lack—namely, it contains information in the form of its sequence of residues.

**Chargaff's Rules Describe the Base Composition of DNA.** Although there appear to be no rules governing the nucleotide composition of typical RNA molecules, DNA has equal numbers of adenine and thymine residues ( $A = T$ ) and equal numbers of guanine and cytosine residues ( $G = C$ ). These relationships, known as **Chargaff's rules**, were discovered in the late 1940s by Erwin Chargaff, who devised the first reliable quantitative methods for the compositional analysis of DNA.

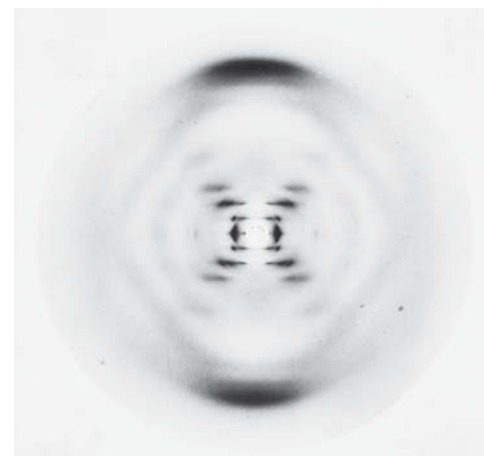
DNA's base composition varies widely among different organisms. It ranges from ~25 to 75 mol %  $G + C$  in different species of bacteria. However, it is more or less constant among related species; for example, in mammals  $G + C$  ranges from 39 to 46%. The significance of Chargaff's rules was not immediately appreciated, but we now know that the structural basis for the rules derives from DNA's double-stranded nature.

## B DNA Forms a Double Helix

The determination of the structure of DNA by James Watson and Francis Crick in 1953 is often said to mark the birth of modern molecular biology. The **Watson–Crick structure** of DNA not only provided a model of what is arguably the central molecule of life, but it also suggested the molecular mechanism of heredity. Watson and Crick's accomplishment, which is ranked as one of science's major intellectual achievements, was based in part on two pieces of evidence in addition to Chargaff's rules: the correct tautomeric forms of the bases and indications that DNA is a helical molecule.

The purine and pyrimidine bases of nucleic acids can assume different tautomeric forms (**tautomers** are readily interconverted isomers that differ only in hydrogen positions; Fig. 3-4). X-Ray, nuclear magnetic resonance (NMR), and spectroscopic investigations have firmly established that the nucleic acid bases are overwhelmingly in the keto tautomeric forms shown in Fig. 3-3. In 1953, however, this was not generally appreciated. Information about the dominant tautomeric forms was provided by Jerry Donohue, an officemate of Watson and Crick and an expert on the X-ray structures of small organic molecules.

Evidence that DNA is a helical molecule was provided by an X-ray diffraction photograph of a DNA fiber taken by Rosalind Franklin (Fig. 3-5). The appearance of the photograph enabled Crick, an X-ray crystallographer by training, to deduce (a) that DNA is a helical molecule and (b) that its planar aromatic bases form a stack that is parallel to the fiber axis.

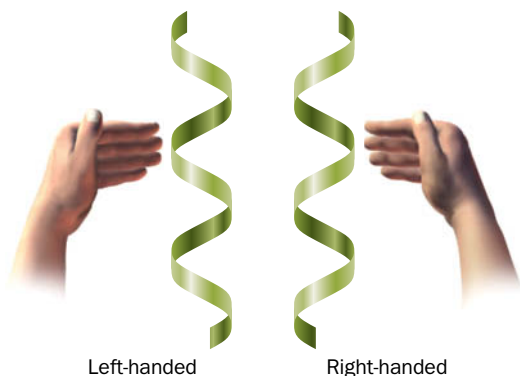
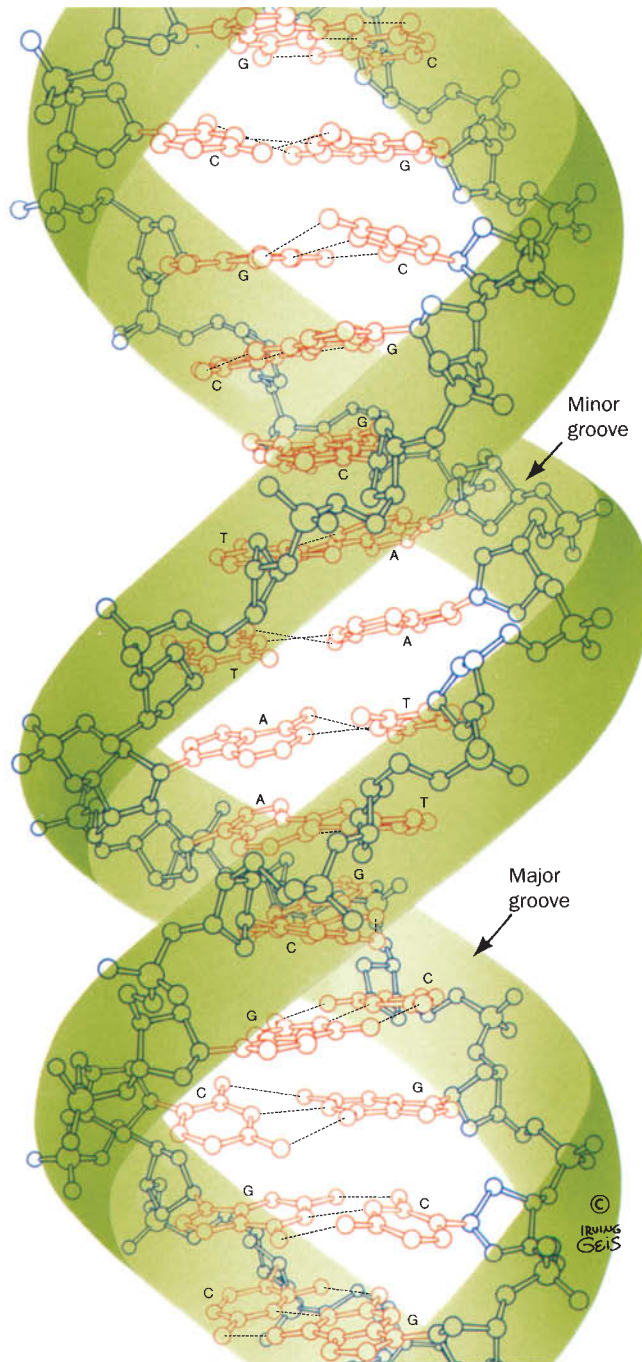


Courtesy of Maurice Wilkins, King's College, London.

**FIG. 3-5 An X-ray diffraction photograph of a vertically oriented DNA fiber.** This photograph, taken by Rosalind Franklin, provided key evidence for the elucidation of the Watson–Crick structure. The central X-shaped pattern indicates a helix, whereas the heavy black arcs at the top and bottom of the diffraction pattern reveal the spacing of the stacked bases (3.4 Å).

**FIG. 3-6 Three-dimensional structure of DNA.**

The repeating helix is based on the structure of the self-complementary dodecamer d(CGCGAATTCGCG) determined by Richard Dickerson and Horace Drew. The view in this ball-and-stick model is perpendicular to the helix axis. The sugar-phosphate backbones (*blue, with green ribbon outlines*) wind around the periphery of the molecule. The bases (*red*) form hydrogen-bonded pairs that occupy the core. H atoms have been omitted for clarity. The two strands run in opposite directions. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

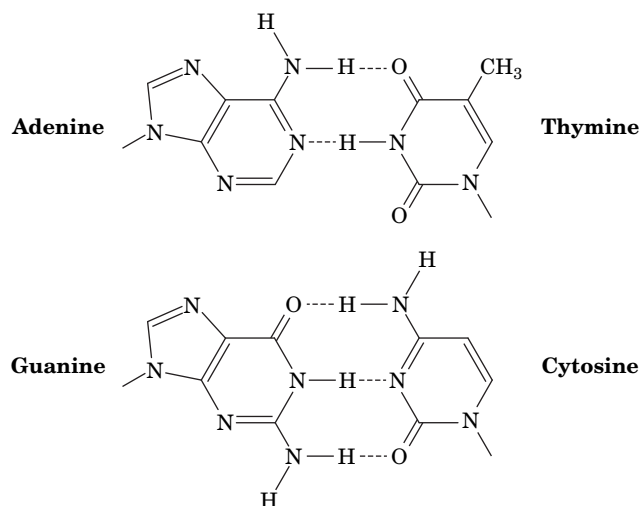
**FIG. 3-7 Diagrams of left- and right-handed helices.**

In each case, the fingers curl in the direction the helix turns when the thumb points in the direction the helix rises. Note that the handedness is retained when the helices are turned upside down.

The limited structural information, along with Chargaff's rules, provided few clues to the structure of DNA; Watson and Crick's model sprang mostly from their imaginations and model-building studies. Once the Watson–Crick model had been published, however, its basic simplicity combined with its obvious biological relevance led to its rapid acceptance. Later investigations have confirmed the general validity of the Watson–Crick model, although its details have been modified.

The Watson–Crick model of DNA has the following major features (**Fig. 3-6**):

1. Two polynucleotide chains wind around a common axis to form a **double helix**.
2. The two strands of DNA are **antiparallel** (run in opposite directions), but each forms a right-handed helix. (The difference between a right-handed and a left-handed helix is shown in **Fig. 3-7**.)



**FIG. 3-8 Complementary base pairing in DNA.** Adenine in one strand pairs with thymine in the other strand by forming specific hydrogen bonds (*dashed lines*). Similarly, guanine pairs with cytosine. This base pairing between polynucleotide chains is responsible for the double-stranded nature of DNA.

? Indicate the 5' → 3' sequence of the strand that would be complementary to the nucleic acid shown in Fig. 3-3a.

3. The bases occupy the core of the helix and sugar–phosphate chains run along the periphery, thereby minimizing the repulsions between charged phosphate groups. The surface of the double helix contains two grooves of unequal width: the **major** and **minor grooves**.
4. Each base is hydrogen bonded to a base in the opposite strand to form a planar **base pair**. The Watson–Crick structure can accommodate only two types of base pairs. Each adenine residue must pair with a thymine residue and vice versa, and each guanine residue must pair with a cytosine residue and vice versa (**Fig. 3-8**). These hydrogen-bonding interactions, a phenomenon known as **complementary base pairing**, result in the specific association of the two chains of the double helix.

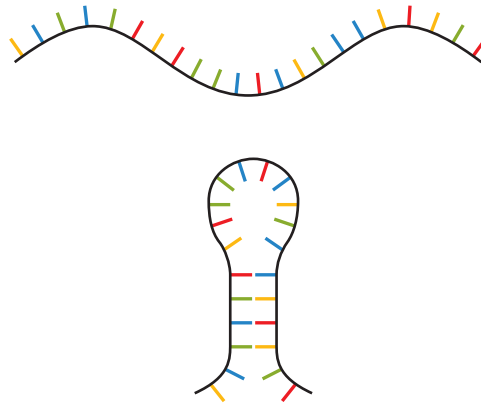
The Watson–Crick structure can accommodate any sequence of bases on one polynucleotide strand if the opposite strand has the complementary base sequence. This immediately accounts for Chargaff's rules. More importantly, it suggests that *each DNA strand can act as a **template** for the synthesis of its complementary strand and hence that hereditary information is encoded in the sequence of bases on either strand.*

**Most DNA Molecules Are Large.** The extremely large size of DNA molecules is in keeping with their role as the repository of a cell's genetic information. Of course, an organism's **genome**, its unique DNA content, may be allocated among several **chromosomes** (Greek: *chromos*, color + *soma*, body), each of which contains a separate DNA molecule. Note that many organisms are **diploid**; that is, they contain two equivalent sets of chromosomes, one from each parent. Their content of unique (**haploid**) DNA is half their total DNA. For example, humans are diploid organisms that carry 46 chromosomes per cell; their haploid number is therefore 23.

Because of their great lengths, DNA molecules are described in terms of the number of base pairs (**bp**) or thousands of base pairs (**kilobase pairs**, or **kb**). Naturally occurring DNAs vary in length from ~5 kb in small DNA-containing viruses to well over 250,000 kb in the largest mammalian chromosomes. Although DNA molecules are long and relatively stiff, they are not completely rigid. We will see in Chapter 24 that the DNA double helix forms coils and loops when it is packaged inside the cell. Furthermore, depending on the nucleotide

#### GATEWAY CONCEPT Noncovalent Interactions

The sequence of nucleotides in a polynucleotide chain is determined by covalent bonds, but hydrogen bonds, which are much weaker, allow one chain to interact with another via base pairing.



**FIG. 3-9 Formation of a stem-loop structure.** Base pairing between complementary sequences within an RNA strand allows the polynucleotide to fold back on itself.

sequence, DNA may adopt slightly different helical conformations. Finally, in the presence of other cellular components, the DNA may bend sharply or the two strands may partially unwind.

### CHECKPOINT

- Using Fig. 3-3a as a guide, draw the complete structure of a nucleoside triphosphate before and after it becomes incorporated into a polynucleotide chain. Draw the structure that would result if the newly formed phosphodiester bond were hydrolyzed.
- Explain the structural basis for Chargaff's rules.
- Using a three-dimensional computer model of the DNA molecule (such as shown in Interactive Exercise 1), identify each of the following structural features: the 3' and 5' end of each strand, the atoms that make up the sugar-phosphate backbone, the major and minor grooves, the bases in several base pairs, and the atoms that participate in hydrogen bonding in A · T and G · C base pairs.
- List the structural differences between DNA and RNA.

## C RNA Is a Single-Stranded Nucleic Acid

Single-stranded DNA is rare, occurring mainly as the hereditary material of certain viruses. In contrast, RNA occurs primarily as single strands, which usually form compact structures rather than loose extended chains (double-stranded RNA is the hereditary material of certain viruses). An RNA strand—which is identical to a DNA strand except for the presence of 2'-OH groups and the substitution of uracil for thymine—can base-pair with a complementary strand of RNA or DNA. As expected, A pairs with U (or T in DNA), and G with C. Base pairing often occurs intramolecularly, giving rise to **stem-loop** structures (Fig. 3-9) or, when loops interact with each other, to more complex structures.

The intricate structures that can potentially be adopted by single-stranded RNA molecules provide additional evidence that RNA can do more than just store and transmit genetic information. Numerous investigations have found that certain RNA molecules can specifically bind small organic molecules and can catalyze reactions involving those molecules. These findings provide substantial support for theories that *many of the processes essential for life began through the chemical versatility of small polynucleotides* (a situation known as the **RNA world**). We will further explore RNA structure and function in Section 24-2C.

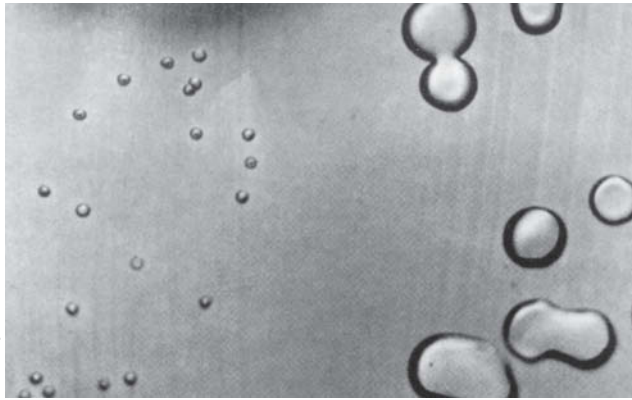
## 3 Overview of Nucleic Acid Function

### KEY CONCEPTS

- DNA carries genetic information in the form of its sequence of nucleotides.
- The nucleotide sequence of DNA is transcribed into the nucleotide sequence of messenger RNA, which is then translated into a protein, a sequence of amino acids.

DNA is the carrier of genetic information in all cells and in many viruses. Yet a period of over 75 years passed from the time the laws of inheritance were discovered by Gregor Mendel until the biological role of DNA was elucidated. Even now, many details of how genetic information is expressed and transmitted to future generations remain unclear.

From Avery, O. T., MacLeod, C. M., and McCarty, M. J. *Exp. Med.* **79**, 153 (1944). Copyright © 1944 by Rockefeller University Press.



**FIG. 3-10 Transformed pneumococci.** The large colonies are virulent pneumococci that resulted from the transformation of nonpathogenic pneumococci (smaller colonies) by DNA extracted from the virulent strain. We now know that this DNA contained a gene that was defective in the nonpathogenic strain.

Mendel's work with garden peas led him to postulate that each trait in an individual plant is determined by a pair of factors (which we now call **genes**), one inherited from each parent. But Mendel's theory of inheritance, reported in 1866, was almost universally ignored by his contemporaries, whose knowledge of anatomy and physiology provided no basis for its understanding. Eventually, genes were hypothesized to be part of chromosomes, and the pace of genetic research accelerated greatly.

## A DNA Carries Genetic Information

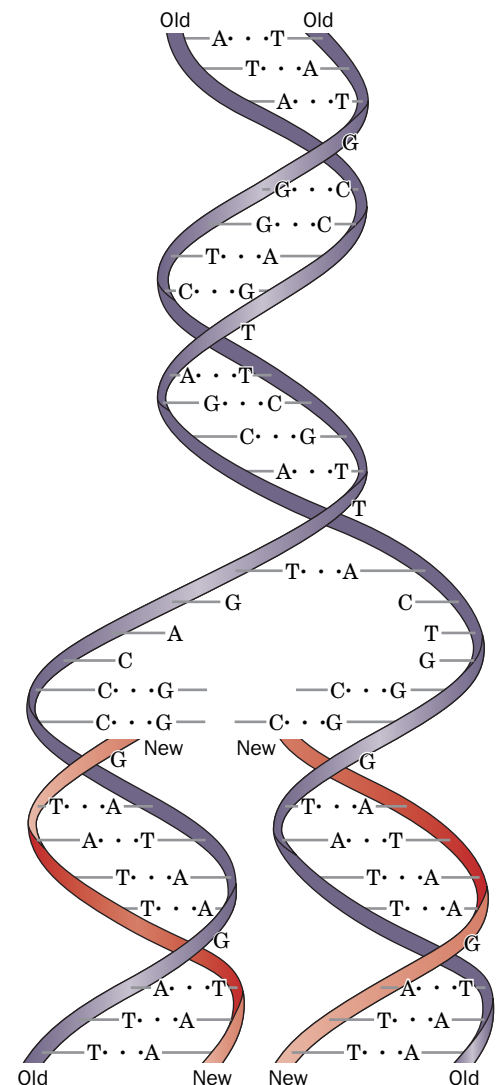
Until the 1940s, it was generally assumed that genes were made of protein, since proteins were the only biochemical entities that, at the time, seemed complex enough to serve as agents of inheritance. Nucleic acids, which had first been isolated in 1869 by Friedrich Miescher, were believed to have monotonously repeating nucleotide sequences and were therefore unlikely candidates for transmitting genetic information.

It took the efforts of Oswald Avery, Colin MacLeod, and Maclyn McCarty to demonstrate that DNA carries genetic information. Their experiments, completed in 1944, showed that DNA—not protein—extracted from a virulent (pathogenic) strain of the bacterium *Diplococcus pneumoniae* was the substance that **transformed** (permanently changed) a nonpathogenic strain of the organism to the virulent strain (**Fig. 3-10**). Avery's discovery was initially greeted with skepticism, but it influenced Erwin Chargaff, whose rules (Section 3-2A) led to subsequent models of the structure and function of DNA.

The double-stranded, or duplex, nature of DNA facilitates its **replication**. When a cell divides, each DNA strand acts as a template for the assembly of its complementary strand (**Fig. 3-11**). Consequently, every progeny cell contains a complete DNA molecule (or a complete set of DNA molecules in organisms whose genomes contain more than one chromosome). Each DNA molecule consists of one parental strand and one daughter strand. Daughter strands are synthesized by the stepwise polymerization of nucleotides that specifically pair with bases on the parental strands. The mechanism of replication, while straightforward in principle, is exceedingly complex in the cell, requiring a multitude of cellular factors to proceed with fidelity and efficiency, as we will see in Chapter 25.

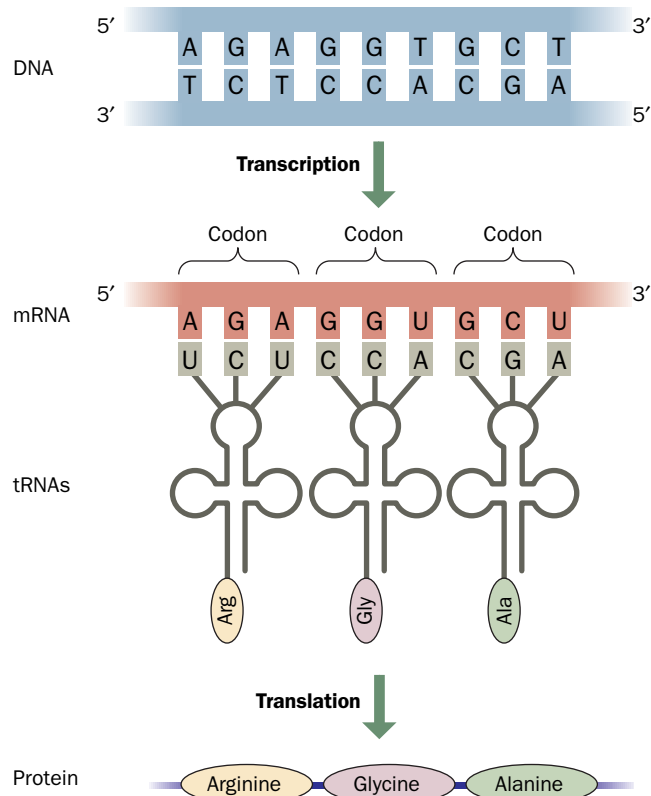
## B Genes Direct Protein Synthesis

The question of how sequences of nucleotides control the characteristics of organisms took some time to be answered. In experiments with the mold *Neurospora crassa* in the 1940s, George Beadle and Edward Tatum found that *there is a specific connection between genes and enzymes: the one gene—one enzyme theory*. Beadle and Tatum showed that certain mutant varieties of *Neurospora*



**FIG. 3-11 DNA replication.** Each strand of parental DNA (**blue**) acts as a template for the synthesis of a complementary daughter strand (**red**). Thus, the resulting double-stranded molecules are identical.

**? What types of bonds or interactions form during DNA replication?**



**FIG. 3-12 Transcription and translation.** One strand of DNA directs the synthesis of messenger RNA (mRNA). The base sequence of the transcribed RNA is complementary to that of the DNA strand. The message is translated when transfer RNA (tRNA) molecules align with the mRNA by complementary base pairing between three-nucleotide segments known as **codons**. Each tRNA carries a specific amino acid. These amino acids are covalently joined to form a protein. Thus, the sequence of bases in DNA specifies the sequence of amino acids in a protein.

? What might happen if a mutation changed one of the nucleotides in the DNA?

that were generated by irradiation with X-rays required additional nutrients in order to grow. Presumably, the offspring of the radiation-damaged cells lacked the specific enzymes necessary to synthesize those nutrients.

The link between DNA and enzymes (nearly all of which are proteins) is RNA. *The DNA of a gene is **transcribed** to produce an RNA molecule that is complementary to the DNA. The RNA sequence is then **translated** into the corresponding sequence of amino acids to form a protein (Fig. 3-12).* These transfers of biological information are summarized in the so-called **central dogma of molecular biology** formulated by Crick in 1958.

### WileyPLUS Learning Space

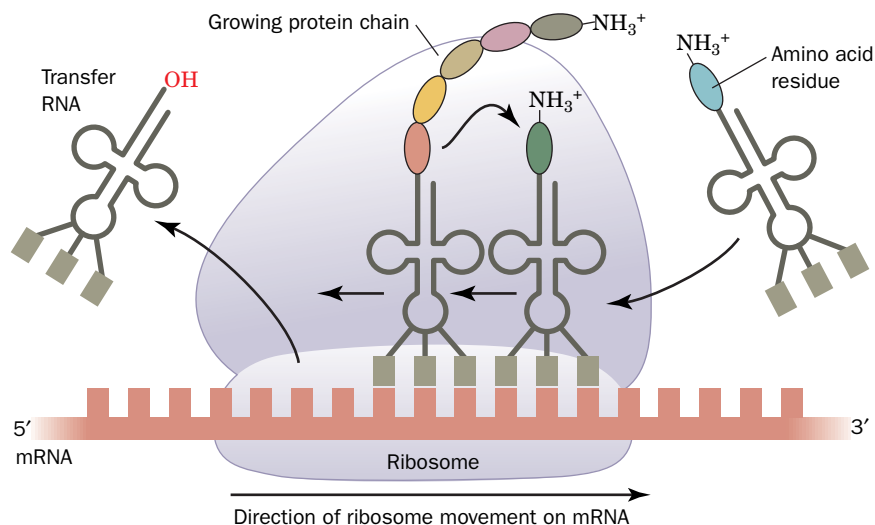
See Guided Exploration

Overview of transcription and translation



In this diagram, the arrows represent the flow of genetic information: DNA directs its own replication to produce new DNA molecules; the DNA is transcribed into RNA; and the RNA is translated into proteins.

Just as the daughter strands of DNA are synthesized from free deoxynucleoside triphosphates that pair with bases in the parent DNA strand, RNA strands are synthesized from free ribonucleoside triphosphates that pair with the complementary bases in one DNA strand of a gene (transcription is described in greater detail in Chapter 26). The RNA that corresponds to a protein-coding gene (called **messenger RNA**, or **mRNA**) makes its way to a **ribosome**, an organelle that is itself composed largely of RNA (**ribosomal RNA**, or **rRNA**). At the ribosome, each set of three nucleotides in the mRNA pairs with three complementary nucleotides in a small RNA molecule called



**FIG. 3-13 Translation.** tRNA molecules with their attached amino acids bind to complementary three-nucleotide sequences (codons) on mRNA. The ribosome facilitates the alignment of the tRNA and the mRNA, and it catalyzes the joining of amino acids to produce a protein chain. When a new amino acid is added, the preceding tRNA is ejected, and the ribosome proceeds along the mRNA.

a **transfer RNA**, or **tRNA** (Fig. 3-13). Attached to each tRNA molecule is its corresponding amino acid. The ribosome catalyzes the joining of amino acids, which are the monomeric units of proteins (protein synthesis is described in detail in Chapter 27). Amino acids are added to the growing protein chain according to the order in which the tRNA molecules bind to the mRNA. Since the nucleotide sequence of the mRNA in turn reflects the sequences of nucleotides in the gene, DNA directs the synthesis of proteins. It follows that alterations to the genetic material of an organism (**mutations**) may manifest themselves as proteins with altered structures and functions.

Using techniques that are described in the following sections and in other parts of this book, researchers can compile a catalog of all the information encoded in an organism's DNA. The study of the genome's size, organization, and gene content is known as **genomics**. By analogy, **transcriptomics** refers to the study of **gene expression**, which focuses on the set of mRNA molecules, or **transcriptome**, that is transcribed from DNA under any particular set of circumstances. Finally, **proteomics** is the study of the proteins (the **proteome**) produced as a result of transcription and translation. Although an organism's genome remains essentially unchanged throughout its lifetime, its transcriptome and proteome may vary greatly among different types of tissues, developmental stages, and environmental conditions.

#### GATEWAY CONCEPT The Central Dogma

Although we sometimes think of a cell's nucleus as a sort of command center, it does not broadcast information to the rest of the cell. Instead, as the central dogma reminds us, genetic information in the form of DNA is relatively inert and is simply copied—via replication or transcription.

#### CHECKPOINT

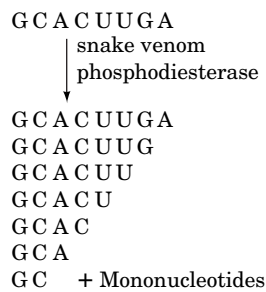
- Explain why the double-stranded nature of DNA is relevant for copying and transmitting genetic information when a cell divides.
- Summarize the steps of the central dogma. What role does RNA play in each?

## 4 Nucleic Acid Sequencing

### KEY CONCEPTS

- In the laboratory, nucleic acids can be cut at specific sequences by restriction enzymes.
- Nucleic acid fragments are separated by size using electrophoresis.
- In the chain-termination method, DNA polymerase generates DNA fragments that are randomly terminated. The identities of the terminator nucleotides of successive fragments reveal the original DNA sequence.
- The human genome contains ~21,000 genes, corresponding to about 1.2% of its 3 billion nucleotides.
- Sequence differences reveal evolutionary changes.

Much of our current understanding of protein structure and function rests squarely on information gleaned not from the proteins themselves, but indirectly from their genes. *The ability to determine the sequence of nucleotides in nucleic acids*



**FIG. 3-14 Determining the sequence of an oligonucleotide using nonspecific enzymes.** The oligonucleotide is partially digested with snake venom phosphodiesterase, which breaks the phosphodiester bonds between nucleotide residues, starting at the 3' end of the oligonucleotide. The result is a mixture of fragments of all lengths, which are then separated. Comparing the base composition of a pair of fragments that differ in length by one nucleotide establishes the identity of the 3'-terminal nucleotide in the larger fragment. Analysis of each pair of fragments reveals the sequence of the original oligonucleotide.

has made it possible to deduce the amino acid sequences of their encoded proteins and, to some extent, the structures and functions of those proteins. Nucleic acid sequencing has also revealed information about the regulation of genes. Certain portions of genes that are not actually transcribed into RNA nevertheless influence how often a gene is transcribed and translated—that is, **expressed**. Moreover, efforts to elucidate the sequences in hitherto unmapped regions of DNA have led to the discovery of new genes and new regulatory elements. *Once in hand, a nucleic acid sequence can be duplicated, modified, and expressed, making it possible to study proteins that could not otherwise be obtained in useful quantities.* In this section, we describe how nucleic acids are sequenced and what information the sequences may reveal. In the following section, we discuss the manipulation of purified nucleic acid sequences for various purposes.

The overall strategy for sequencing any polymer of nonidentical units is

1. Cleave the polymer into fragments that are small enough to be fully sequenced.
2. Determine the sequence of residues in each fragment.
3. Determine the order of the fragments in the original polymer by aligning fragments that contain overlapping sequences.

The first efforts to sequence RNA used nonspecific enzymes to generate relatively small fragments whose nucleotide composition was then determined by partial digestion with an enzyme that selectively removed nucleotides from one end or the other (**Fig. 3-14**). Sequencing RNA in this manner was tedious and time-consuming. Using such methods, it took Robert Holley 7 years to determine the sequence of a 76-residue tRNA molecule.

After 1975, dramatic progress was made in nucleic acid sequencing technology. The advances were made possible by the discovery of enzymes that could cleave DNA at specific sites and by the development of rapid sequencing techniques for DNA. Because most specific DNA sequences are normally present in a genome in only a single copy, most sequencing projects take advantage of methods to amplify segments of DNA by cloning or copying them (Section 3-5).

## A Restriction Endonucleases Cleave DNA at Specific Sequences

Many bacteria are able to resist infection by **bacteriophages** (viruses that are specific for bacteria) by virtue of a **restriction–modification system**. The bacterium modifies certain nucleotides in specific sequences of its own DNA by adding a methyl ( $-\text{CH}_3$ ) group in a reaction catalyzed by a **modification methylase**. A **restriction endonuclease**, which recognizes the same nucleotide sequence as does the methylase, cleaves any DNA that has not been modified on at least one of its two strands. (An **endonuclease** cleaves a nucleic acid within the polynucleotide strand; an **exonuclease** cleaves a nucleic acid by removing one of its terminal residues.) This system destroys foreign (phage) DNA containing a recognition site that has not been modified by methylation. The host DNA is always at least half methylated, because although the daughter strand is not methylated until shortly after it is synthesized, the parental strand to which it is paired is already modified (and thus protects both strands of the DNA from cleavage by the restriction enzyme).

Type II restriction endonucleases are particularly useful in the laboratory. These enzymes cleave DNA within the four- to eight-base sequence that is recognized by their corresponding modification methylase. (Type I and Type III restriction endonucleases cleave DNA at sites other than their recognition sequences.) More than 11,000 Type II restriction enzymes with more than 270 different recognition sequences have been characterized. Some of the more widely used restriction enzymes are listed in **Table 3-2**. A restriction enzyme is named by the first letter of the genus and the first two letters of the species of the

**TABLE 3-2** Recognition and Cleavage Sites of Some Restriction Enzymes

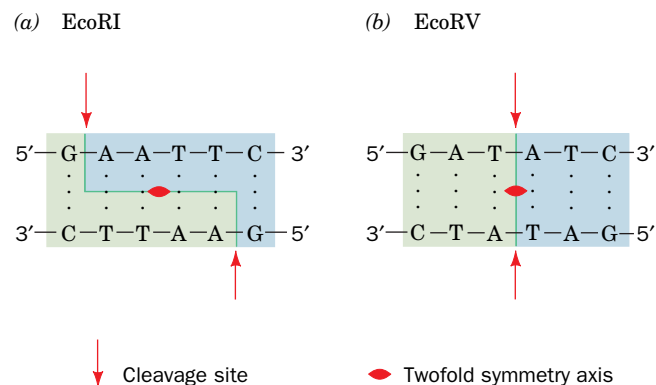
| Enzyme  | Recognition Sequence <sup>a</sup> | Microorganism                                |
|---------|-----------------------------------|----------------------------------------------|
| AluI    | AG↓CT                             | <i>Arthrobacter luteus</i>                   |
| BamHI   | G↓GATCC                           | <i>Bacillus amyloliquefaciens</i> H          |
| BglII   | A↓GATCT                           | <i>Bacillus globigii</i>                     |
| EcoRI   | G↓AATTC                           | <i>Escherichia coli</i> RY13                 |
| EcoRII  | ↓CC(↓)GG                          | <i>Escherichia coli</i> R245                 |
| EcoRV   | GAT↓ATC                           | <i>Escherichia coli</i> J62 pLG74            |
| HaeII   | RGCGC↓Y                           | <i>Haemophilus aegyptius</i>                 |
| HaeIII  | GG↓CC                             | <i>Haemophilus aegyptius</i>                 |
| HindIII | A↓AGCTT                           | <i>Haemophilus influenzae</i> R <sub>d</sub> |
| HpaII   | C↓CGG                             | <i>Haemophilus parainfluenzae</i>            |
| MspI    | C↓CGG                             | <i>Moraxella</i> species                     |
| PstI    | CTGCA↓G                           | <i>Providencia stuartii</i> 164              |
| PvuII   | CAG↓CTG                           | <i>Proteus vulgaris</i>                      |
| SalI    | G↓TCGAC                           | <i>Streptomyces albus</i> G                  |
| TaqI    | T↓CGA                             | <i>Thermus aquaticus</i>                     |
| XhoI    | C↓TCGAG                           | <i>Xanthomonas holcicola</i>                 |

<sup>a</sup>The recognition sequence is abbreviated so that only one strand, reading 5' to 3', is given. The cleavage site is represented by an arrow (↓). R and Y represent a purine nucleotide and a pyrimidine nucleotide, respectively.

Source: Roberts, R.J. and Macelis, D., REBASE—the restriction enzyme database, <http://rebase.neb.com>.

bacterium that produced it, followed by its serotype or strain designation, if any, and a roman numeral if the bacterium contains more than one type of restriction enzyme. For example, EcoRI is produced by *E. coli* strain RY13.

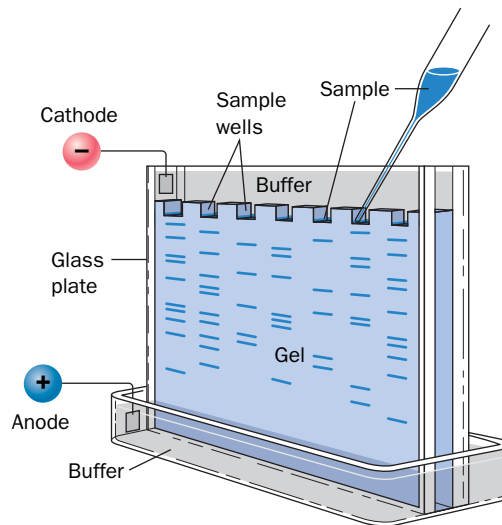
Interestingly, most Type II restriction endonucleases recognize and cleave palindromic DNA sequences. A **palindrome** is a word or phrase that reads the same forward or backward. Two examples are “refer” and “Madam, I’m Adam.” In a palindromic DNA segment, the sequence of nucleotides is the same in each strand, and the segment is said to have twofold symmetry (Fig. 3-15). Most restriction enzymes cleave the two strands of DNA at positions that are staggered, producing DNA fragments with complementary single-strand extensions. Restriction fragments with such **sticky ends** can



**FIG. 3-15** Restriction sites. The recognition sequences for Type II restriction endonucleases are palindromes, sequences with a twofold axis of symmetry. (a) Recognition site for EcoRI, which generates DNA fragments with sticky ends. (b) Recognition site for EcoRV, which generates blunt-ended fragments.

**FIG. 3-16 Apparatus for gel electrophoresis.** Samples are applied in slots at the top of the gel and electrophoresed in parallel lanes. Negatively charged molecules such as DNA migrate through the gel matrix toward the anode in response to an applied electric field. Because smaller molecules move faster, the molecules in each lane are separated according to size. Following electrophoresis, the separated molecules may be visualized by staining or fluorescence.

? What would happen if the sample contained positively charged molecules?

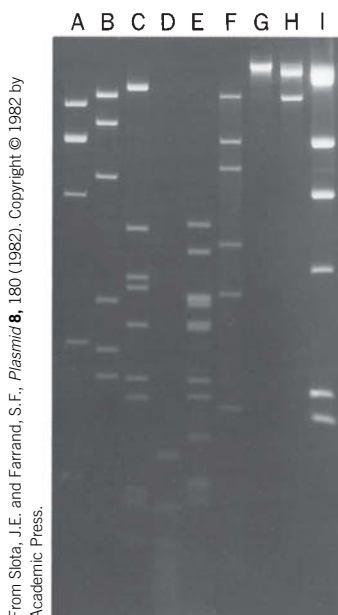


associate by base pairing with other restriction fragments generated by the same restriction enzyme. Some restriction endonucleases cleave the two strands of DNA at the symmetry axis to yield restriction fragments with fully base-paired **blunt ends**.

## B Electrophoresis Separates Nucleic Acids According to Size

Treating a DNA molecule with a restriction endonuclease produces a series of precisely defined fragments that can be separated according to size. **Gel electrophoresis** is commonly used for the separation. In principle, a charged molecule moves in an electric field with a velocity proportional to its overall charge density, size, and shape. For molecules with a relatively homogeneous composition (such as nucleic acids), shape and charge density are constant, so the velocity depends primarily on size. Electrophoresis is carried out in a gel-like matrix, usually made from **agarose** (carbohydrate polymers that form a loose mesh) or **polyacrylamide** (a more rigid cross-linked synthetic polymer). The gel is typically held between two glass plates (Fig. 3-16) or inside a narrow capillary tube (Section 5-2D). The molecules to be separated are applied to one end of the gel, and the molecules move through the pores in the matrix under the influence of an electric field. Smaller molecules move more rapidly through the gel and therefore migrate farther in a given time.

Following electrophoresis, the separated molecules may be visualized in the gel by an appropriate technique, such as addition of a stain that binds tightly to the DNA, by radioactive labeling, or by their fluorescence. Depending on the dimensions of the gel and the visualization technique used, samples containing less than a nanogram of material can be separated and detected by gel electrophoresis. Several samples can be electrophoresed simultaneously. For example, the fragments obtained by digesting a DNA sample with different restriction endonucleases can be visualized side by side (Fig. 3-17). The



From Sota, J.E. and Farrand, S.F., *Plasmid* 8, 180 (1982). Copyright © 1982 by Academic Press.

**FIG. 3-17 Electrophoretogram of restriction digests.** The plasmid pAgK84 has been digested with (A) BamHI, (B) PstI, (C) BglII, (D) HaeIII, (E) HincII, (F) SacI, (G) XbaI, and (H) HpaI. Lane I contains bacteriophage  $\lambda$  digested with HindIII as a standard since these fragments have known sizes. The restriction fragments in each lane are made visible by fluorescence against a black background.

sizes of the various fragments can be determined by comparing their electrophoretic mobilities to the mobilities of fragments of known size.

## C Traditional DNA Sequencing Uses the Chain-Terminator Method

Until recent years, the most widely used technique for sequencing DNA was the **chain-terminator method**, which was devised by Frederick Sanger. The first step in this procedure is to obtain single polynucleotide strands. Complementary DNA strands can be separated by heating, which breaks the hydrogen bonds between bases. Next, polynucleotide fragments that terminate at positions corresponding to each of the four nucleotides are generated. Finally, the fragments are separated and detected.

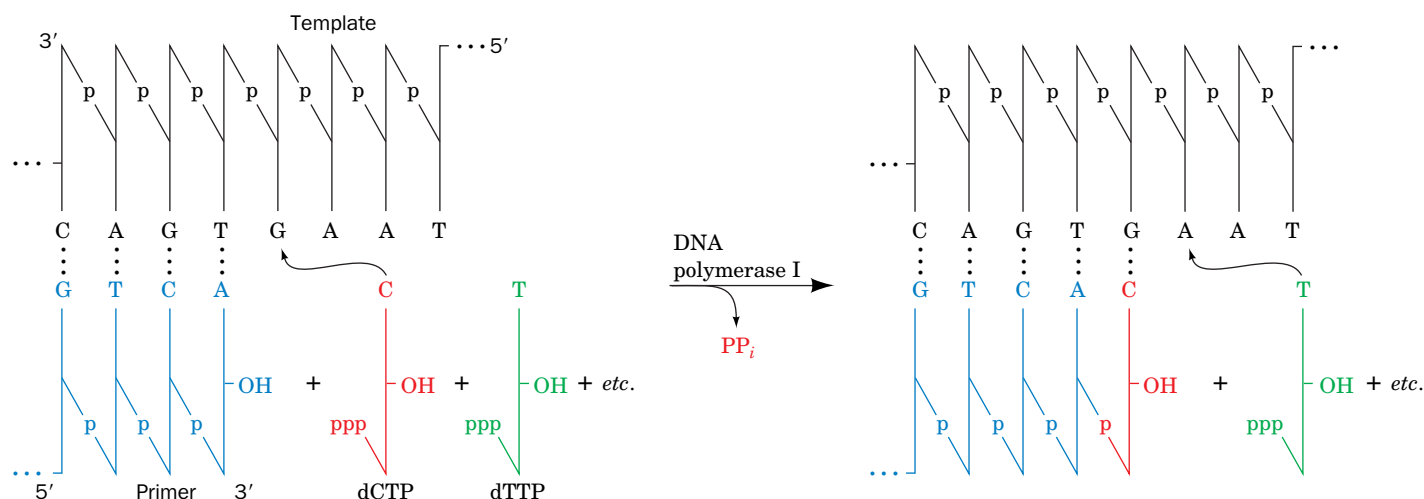
**DNA Polymerase Copies a Template Strand.** The chain-terminator method (also called the **dideoxy method**) uses an *E. coli* enzyme to make complementary copies of the single-stranded DNA being sequenced. The enzyme is a fragment of **DNA polymerase I**, one of the enzymes that participates in replication of bacterial DNA (Section 25-2A). Using the single DNA strand as a template, DNA polymerase I assembles the four deoxynucleoside triphosphates (**dNTPs**), dATP, dCTP, dGTP, and dTTP, into a complementary polynucleotide chain that it elongates in the 5' → 3' direction (Fig. 3-18).

DNA polymerase I can sequentially add deoxynucleotides only to the 3' end of a polynucleotide that is base-paired to the template strand. Replication is initiated in the presence of a short polynucleotide (a **primer**) that is complementary to the 3' end of the template DNA, and thus becomes the 5' end of the new strand. If the DNA being sequenced is a restriction fragment, as it usually is, it begins and ends with a restriction site. The primer can therefore be a short DNA segment with the sequence of this restriction site.

### WileyPLUS Learning Space

#### See Guided Exploration

DNA sequence determination by the chain-terminator method

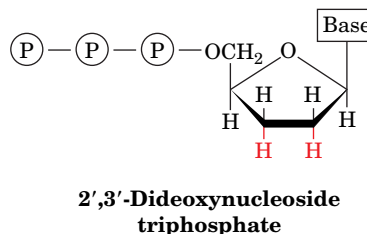


**FIG. 3-18 Action of DNA polymerase I.** Using a single DNA strand as a template, the enzyme elongates the primer by stepwise addition of complementary nucleotides. Incoming nucleotides pair with bases on the template strand and are joined to the growing polynucleotide strand

in the 5' → 3' direction. The polymerase-catalyzed reaction requires a free 3'-OH group on the growing strand. **Pyrophosphate** ( $\text{O}_3\text{P}-\text{O}-\text{PO}_3^{4-}$ ;  $\text{PP}_i$ ) is released with each nucleotide addition.

? List the substrates and products of the DNA polymerase reaction.

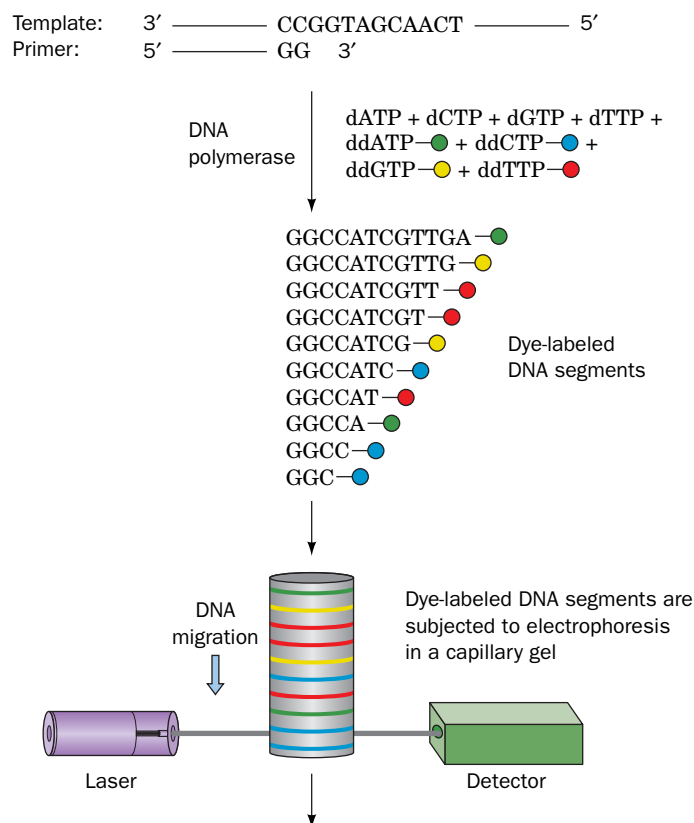
**DNA Synthesis Terminates after Specific Bases.** In the chain-terminator technique (**Fig. 3-19**), the DNA to be sequenced is incubated with DNA polymerase I, a suitable primer, and the four dNTP **substrates** (reactants in enzymatic reactions) for the polymerization reaction. The key component of the reaction mixture is a small amount of a **2',3'-dideoxynucleoside triphosphate (ddNTP)**,



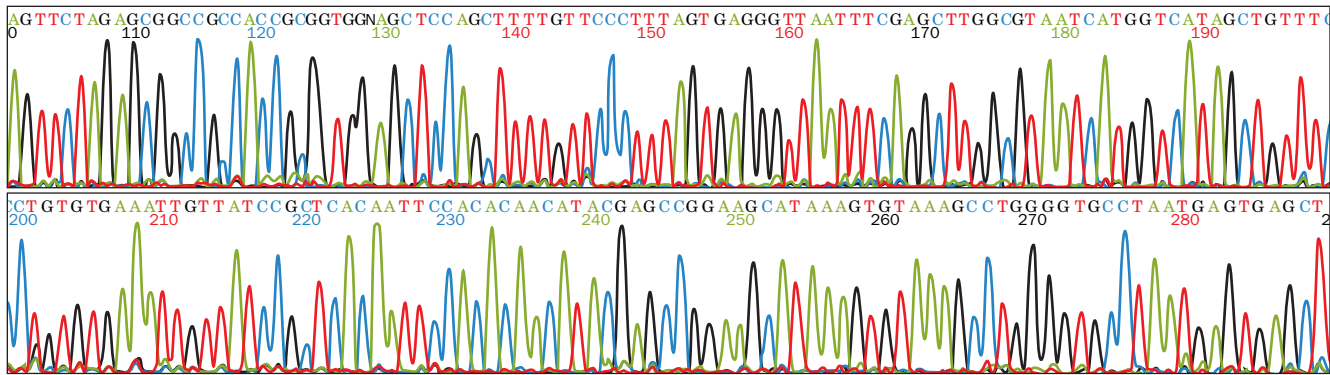
which lacks the 3'-OH group of deoxynucleotides. When the dideoxy analog is incorporated into the growing polynucleotide in place of the corresponding normal nucleotide, chain growth is terminated because addition of the next nucleotide requires a free 3'-OH group.

By using only a small amount of the ddNTP, a series of truncated chains is generated, each of which ends with the dideoxy analog at one of the positions occupied by the corresponding base. Each ddNTP bears a different fluorescent “tag” so that the products of the polymerase reaction can be readily detected. Gel electrophoresis separates the newly synthesized DNA segments, which differ in size by one nucleotide. Thus, the sequence of the replicated strand can be directly read from the gel. Note that the sequence obtained by the chain-terminator method is complementary to the DNA strand being sequenced.

The most advanced sequencing devices that employ the chain-terminator method identify each DNA fragment as it exits the bottom of a capillary



**FIG. 3-19 Automated DNA sequencing by the chain-terminator method.** The reaction mixture includes the single-stranded DNA to be sequenced (the template), a primer, the four deoxynucleoside triphosphates (represented as dATP, etc.), and small amounts of the four fluorescently labeled dideoxynucleoside triphosphates (ddATP, etc.). DNA polymerase extends the primer until a chain-terminating dideoxy nucleotide is added, generating a set of DNA fragments that differ by one nucleotide. The mixture is subjected to gel electrophoresis in a capillary tube, which separates the fragments according to size. As each polynucleotide passes the detector, its 3'-terminal nucleotide is identified according to its laser-stimulated fluorescence.



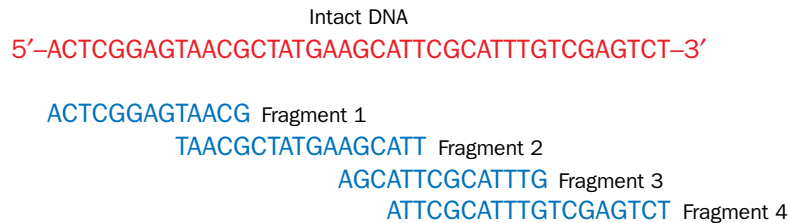
**FIG. 3-20 DNA sequence data.** Each of the four colored curves represents the electrophoretic pattern of fragments containing one of the dideoxynucleotides: Green, red, black, and blue peaks correspond to fragments ending in ddATP, ddTTP, ddGTP, and ddCTP, respectively. The 3'-terminal base of each oligonucleotide, identified by the fluorescence of its gel band, is indicated by a single letter (A, T, G, or C). This portion of the readout corresponds to nucleotides 100–290 of the DNA segment being sequenced. [Courtesy of Mark Adams, The Institute for Genomic Research, Rockville, Maryland.]

electrophoresis tube, so the sequence data take the form of a series of peaks (Fig. 3-20). Sample preparation and data analysis are fully automated, and sequences of DNA (called **reads**) of up to 1000 nucleotides can be obtained from a single reaction mixture before cumulative errors reduce the confidence of identification to unacceptable levels. Moreover, these systems contain arrays of up to 384 capillary tubes and hence can simultaneously sequence as many as 384 DNA segments.

The many reads that have been sequenced must be correctly assembled into the far longer strand of DNA from which they originated. This is done by comparing reads with overlapping sequences as is schematically indicated in Fig. 3-21. The overlapping sets of DNA fragments are generated by separately cleaving the DNA with at least two restriction endonucleases that have different sequence specificities. Alternatively, fragments may be generated by subjecting a solution of the stiff double-stranded DNA to high-frequency sound waves, a process called **sonication**, thereby mechanically breaking (shearing) the DNA at random sites. For even relatively small chromosomes (the *E. coli* genome has 4600 kb; the average human chromosome has ~125,000 kb), assembly is a highly computationally intensive process. In addition, the chain-terminator method has an error rate of ~0.1% (which would lead to ~125,000 mistakes in a 125,000-kb chromosome). To minimize these errors, the DNA must be independently sequenced multiple times—normally with at least a 10-fold redundancy.

## D Next-Generation Sequencing Technologies Are Massively Parallel

The value of DNA sequence information, particularly the enormous datasets obtained by sequencing entire genomes, has fostered the development of so-called

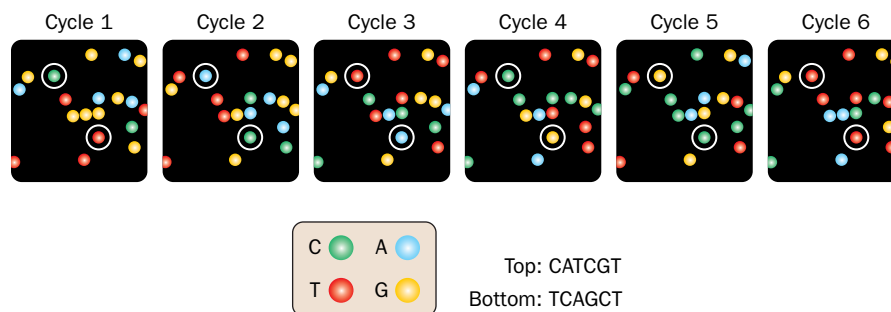


**FIG. 3-21 The order of the sequenced fragments in their DNA of origin is determined by matching the sequences of overlapping sets of fragments.** Thus, the 3' end of fragment 1 is identical to the 5' end of fragment 2, whose 3' end is identical to the 5' end of fragment 3, etc.

next-generation sequencing technologies that offer various trade-offs among cost, speed, and accuracy. Like the chain-terminator method, many of the newer methods take advantage of the ability of a DNA template strand to direct the synthesis of its complementary copy (a procedure known as **sequencing by synthesis**). But in addition, these methods can simultaneously determine hundreds of thousands to billions of different reads; that is, they carry out massively parallel sequencing reactions. Two of the several DNA sequencing technologies that are in use are described below.

In **pyrosequencing** (which employs instrumentation made by 454 Life Sciences, Inc. and hence is also known as **454 sequencing**), segments of the DNA to be sequenced are immobilized on the surfaces of microscopic plastic beads under dilution conditions such that no more than one DNA molecule is attached to a bead. The DNA on each bead is then replicated (amplified) many-fold by a variant of the process described in Section 3-5C, and the beads are deposited in wells in a fiber-optic slide that can hold only one bead per well. A primer and DNA polymerase are added, and then a dNTP is introduced. If DNA polymerase adds that nucleotide to the growing DNA strand, pyrophosphate ( $PP_i$ ) is released and undergoes a chemical reaction sequence involving the firefly enzyme **luciferase**, which generates a flash of light. Solutions of each of the four dNTPs are successively washed across the immobilized DNA template, and a detector records whether light is produced in the presence of a particular dNTP. By multiply repeating this process, the sequence of nucleotides complementary to the template strand can be deduced. Pyrosequencing can accurately determine reads of up to 700 nucleotides, somewhat shorter than the reads determined by chain-terminator sequencing. Each fiber-optic slide contains numerous wells so that as many as  $\sim 1$  million templates can be sequenced simultaneously. Consequently, the pyrosequencing system is  $\sim 1000$ -fold faster than the most advanced chain-terminator sequencing systems.

In **Illumina sequencing** (using instrumentation produced by Illumina, Inc.), which is also a sequencing-by-synthesis method, numerous DNA segments are attached to a glass plate and amplified in place to form clusters of millions of identical DNA molecules. To determine their sequences, a solution containing the four dNTPs that are each linked to a different fluorescent group and chemically blocked at their 3' positions is flowed over the plate so that only a single nucleotide is added, by DNA polymerase, to each primer strand. The unreacted dNTPs are then washed away and the fluorescent groups that are linked to the primer strands are excited by a laser and identified according to their color of fluorescence by a digital camera (**Fig. 3-22**). The fluorescent and 3'-blocking groups are then chemically removed and the synthesis-and-identification cycle is repeated. The Illumina system can accurately determine reads of 30 to 300 nucleotides, significantly shorter than those of the chain-terminator and pyrosequencing methods, but it can determine up to 3 billion reads per run.



**FIG. 3-22** Schematic diagram of six successive cycles from an Illumina DNA sequencing device. The different nucleotide derivatives (C, A, T, and G) each fluoresce with a different color. [Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Genetics*, Vol. 11, No. 1, 31-46, Copyright (2009).]

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/036104054243010100>