Chapter 3

DNA, RNA, and Protein

Early geneticists studied the transmission of characters from parents to offspring, but they did not know the underlying chemical mechanisms. In many ways, molecular biology is the merger of biochemistry with genetics. The revelation that genes are made of DNA—in other words, that biological information is carried by the nucleic acids, DNA and RNA—has transformed biology. Understanding the chemical nature of DNA has provided a mechanistic basis not just for heredity but also for a variety of other phenomena from cell growth and division to cancer. In this chapter we discuss how DNA was discovered to be the genetic material. We then review the chemical nature of DNA and RNA and explain how these molecules provide a physical mechanism for storing biological information, distributing it inside a growing cell, and finally passing it from one generation to the next. Finally, we introduce the molecules that are encoded by the genes and carry out most of the day-to-day operations of the cell—the proteins.

1. History of DNA as the Genetic Material

Until early in the nineteenth century, it was believed that living matter was quite different from inanimate matter and was not subject to the normal laws of chemistry. In other words, organisms were thought to be made from chemical components unique to living creatures. Furthermore, there was supposedly a special vital force that mysteriously energized living creatures. Then, in 1828, Friedrich Wohler demonstrated the conversion in a test tube of ammonium cyanate, a

Molecular Biology.

laboratory chemical, to urea, a "living" molecule also generated by animals. This was the first demonstration that there was nothing magical about the chemistry of living matter.

Further experiments showed that the molecules found in living organisms were often very large and complex. Consequently, their complete chemical analysis was time consuming and is indeed, still continuing today. The demystification of life chemistry reached its peak in the 1930s when the Russian biochemist Alexander Oparin wrote a book outlining his proposal for the chemical origin of life. Although the nature of the genetic material was still unknown, Oparin put forward the idea that life, with its complex molecular composition, evolved from small molecules in the primeval ocean as a result of standard physical and chemical forces (see Ch. 26).

Until the time of World War II, the chemical nature of the *inherited genetic information* remained very vague and elusive. **DNA** was actually discovered in 1869 by Frederich Miescher who extracted it from the pus from infected wounds! However, it was nearly a century before its true significance was revealed by Oswald Avery. In 1944, Avery found that the virulent nature of some strains of bacteria that caused pneumonia could be transmitted to related harmless strains by a chemical extract. Avery purified the essential molecule and demonstrated that it was DNA, although he did not use the name "DNA," since its structure was then uncharacterized. When DNA from virulent strains was added to harmless strains, some took up the DNA and were "transformed" into virulent strains. Avery concluded that the genes were made of DNA and that somehow genetic information was encoded in this molecule.

> Despite their complexity, living organisms obey the laws of chemistry.

> Avery found that purified DNA could carry genetic information from one strain of bacterium to another. This revealed that DNA was the genetic material.

2. Nucleic Acid Molecules Carry Genetic Information

Chapter 2 discussed how the fundamentals of modern genetics were laid when Mendel found that hereditary information consists of discrete fundamental units now called genes. Each gene is responsible for a single inherited property or characteristic of the organism. The realization that genes are made up of DNA molecules opened the way both to a deeper understanding of life and to its artificial alteration by genetic engineering.

Genetic information is encoded by molecules named **nucleic acids** because they were originally isolated from the nucleus of eukaryotic cells. There are two related types of nucleic acid, **deoxyribonucleic acid** (**DNA**) and **ribonucleic acid** (**RNA**). The master copy of each cell's genome is stored on long molecules of DNA, which may each contain many thousands of genes. Each gene is thus a linear segment of a long DNA molecule. In contrast, RNA molecules are much shorter, are used to transmit the genetic information to the cell machinery, and carry only one or a few genes. (Certain viruses use RNA to encode their genomes as well as transmit genetic information to the cell machinery. These RNA viruses have short genomes, rarely more than a dozen genes, as opposed to the hundreds or thousands of genes carried on the DNA genomes of cells.)

> Genetic information is carried on long linear polymers, the nucleic acids. Two classes of nucleic acid, DNA and RNA, divide up the responsibility of storing and deploying the genetic information.

3. Chemical Structure of Nucleic Acids

DNA and RNA are linear polymers made of subunits known as **nucleotides**. The information in each gene is determined by the order of the different nucleotides, just as the information in this sentence is due to the order of the 26 possible letters of the alphabet. There are four different nucleotides in each type of nucleic acid and their order determines the genetic information (Fig. 3.01).

---

deoxyribonucleic acid (DNA)   Nucleic acid polymer of which the genes are made
DNA   Deoxyribonucleic acid, nucleic acid polymer of which the genes are made
nucleic acid   Class of polymer molecule consisting of nucleotides that carries genetic information
nucleotide   Monomer or subunit of a nucleic acid, consisting of a pentose sugar plus a base plus a phosphate group
ribonucleic acid (RNA)   Nucleic acid that differs from DNA in having ribose in place of deoxyribose

FIGURE 3.01
**The Order of the
Nucleotides Encodes the
Genetic Information**

Nucleotides are ordered along
a string of DNA or RNA. It is the
ordering of the different nucleotides
that dictates the nature of the
information within the nucleic acid.

KEY

= 5-carbon sugar

= phosphate

= base
(4 alternatives)
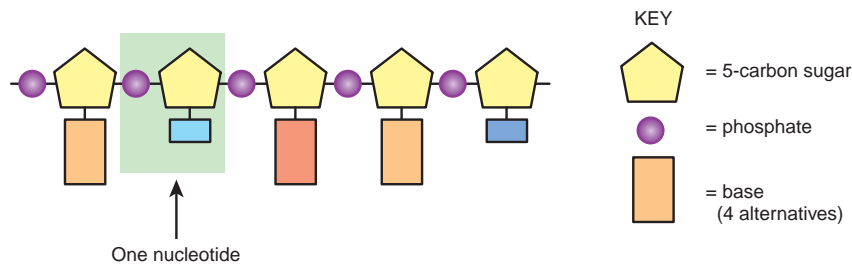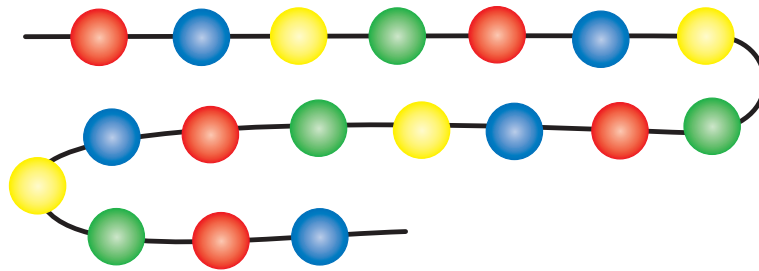
One nucleotide
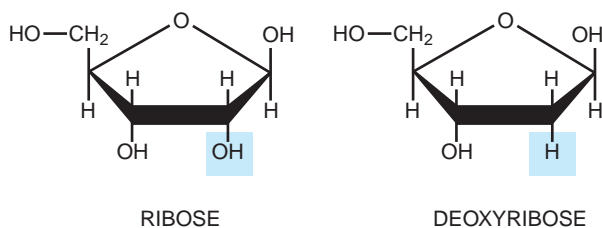
FIGURE 3.02
**Three Views of a
Nucleotide**

The three components of a
nucleotide are shown to the left.
The structures on the right show
the pentose sugar (deoxyribose)
connected to the phosphate and
the base.
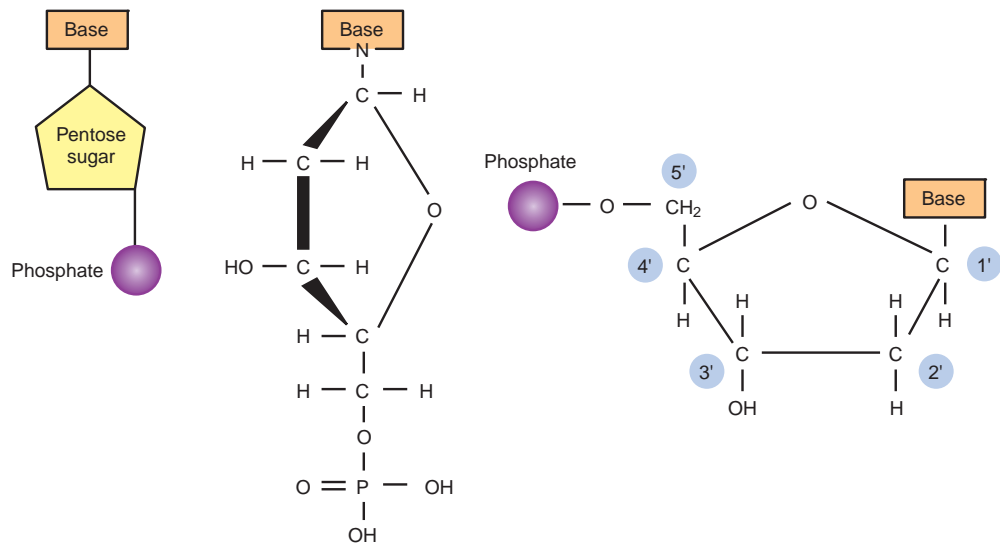
Each nucleotide has three components: a **phosphate group**, a five-carbon sugar, and a nitrogen-containing **base** (Fig. 3.02). The phosphate groups and the sugars form the backbone of each strand of DNA or RNA. The bases are joined to the sugars and stick out sideways.
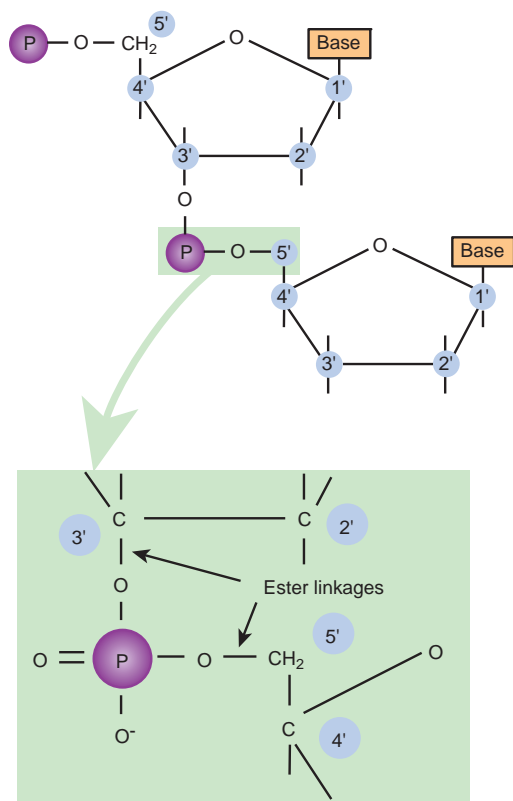
In DNA, the sugar is always **deoxyribose**; whereas, in RNA, it is **ribose**. Both sugars are **pentoses**, or five-carbon sugars. Deoxyribose has one less oxygen than ribose (Fig. 3.03). It is this chemical difference that gave rise to the names deoxyribonucleic acid and ribonucleic acid. Both sugars have five-membered rings consisting of four carbon atoms and one oxygen atom. The fifth carbon forms a side chain to the ring. The five carbon atoms of the sugar are numbered 1', 2', 3', 4', and 5' as shown in Figure 3.02. By convention, in nucleic acids, numbers with prime marks refer to the sugars and numbers without prime marks refer to the positions around the rings of the bases.

---

base   Alkaline chemical substance, in molecular biology especially refers to the cyclic nitrogen compounds found in DNA and RNA
deoxyribose   The sugar with five carbon atoms that is found in DNA
pentose   A five-carbon sugar, such as ribose or deoxyribose
phosphate group   Group of four oxygen atoms surrounding a central phosphorus atom found in the backbone of DNA and RNA
ribose   The five-carbon sugar found in RNA

Ribose is the five-carbon sugar (pentose) found in RNA. Deoxyribose is the pentose of DNA, and it has one less oxygen than ribose as it has hydrogen in place of the hydroxyl group on position 2' of the ribose ring.

The nucleotides that form the backbone of DNA and RNA are joined together by linkages involving their phosphate groups. One nucleotide is linked via its 5'-carbon to the oxygen of the phosphate group and another nucleotide is linked via its 3'-carbon to the other side of the central phosphate. These linkages are termed phosphodiester groups.

Nucleotides are joined by linking the phosphate on the 5'-carbon of the (deoxy) ribose of one nucleotide to the 3'-hydroxyl of the next as shown in Figure 3.04. The phosphate group is joined to the sugar on either side by ester linkages, and the overall structure is therefore a **phosphodiester** linkage. The phosphate group linking the sugars has a negative charge.

> Two nucleotides are linked together via a phosphodiester bond between the 5'-carbon of the sugar in one nucleotide to the 3'-hydroxyl of the next nucleotide.

3.1. DNA and RNA Each Have Four Bases

There are five different types of nitrogenous bases associated with nucleotides. DNA contains the bases **adenine**, **guanine**, **cytosine**, and **thymine**. These are often abbreviated to A, G, C, and T, respectively. RNA contains A, G, and C, but T is replaced by **uracil** (U). From the viewpoint of genetic information, T in DNA, and U in RNA are equivalent.

---

adenine (A)   A purine base that pairs with thymine, found in DNA or RNA
cytosine (C)   One of the pyrimidine bases found in DNA or RNA and which pairs with guanine
guanine (G)   A purine base found in DNA or RNA that pairs with cytosine
phosphodiester   The linkage between nucleotides in a nucleic acid that consists of a central phosphate group esterified to sugar hydroxyl groups on either side
thymine (T)   A pyrimidine base found in DNA that pairs with adenine
uracil (U)   A pyrimidine base found in RNA that may pair with adenine
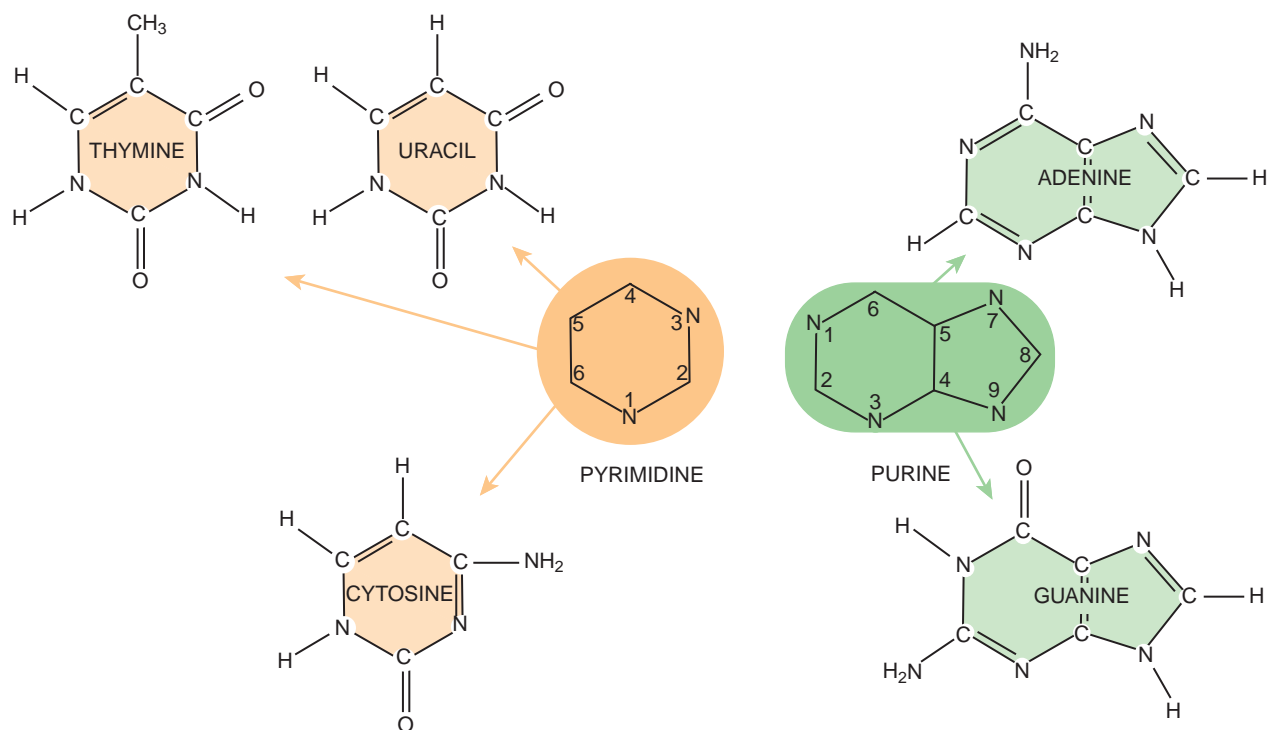
FIGURE 3.05
The Bases of the Nucleic Acids

The four bases of DNA are adenine, guanine, cytosine, and thymine. In RNA, uracil replaces thymine. Pyrimidine bases contain one-ring structures, whereas purine bases contain two-ring structures.

> Purines have two rings and pyrimidines have a single ring.

The bases found in nucleic acids are of two types, **pyrimidines** and **purines**. The smaller pyrimidine bases contain a single ring whereas the purines have a fused double ring. Adenine and guanine are purines; and thymine, uracil and cytosine are pyrimidines. The purine and pyrimidine ring systems and their derivatives are shown in Figure 3.05.

3.2. Nucleosides Are Bases Plus Sugars; Nucleotides Are Nucleosides Plus Phosphate

A base plus a sugar is known as a **nucleoside**. A base plus a sugar plus phosphate is known as a nucleotide. If necessary, one may distinguish between **deoxynucleosides** or **deoxynucleotides** where the sugar is deoxyribose, and **ribonucleosides** or **ribonucleotides** that contain ribose. The names of the nucleosides are similar to the names of the corresponding bases (see Table 3.01). The nucleotides do not have names of their own but are referred to as phosphate derivatives of the corresponding nucleoside. For example, the nucleotide of adenine is **adenosine monophosphate**, or **AMP**.

> Nucleotides have the sugar base attached to a phosphate and nitrogenous base. Nucleosides are missing the phosphate group.

Three-letter abbreviations for the bases such as ade, gua, etc., are sometimes used when writing biochemical pathways or for the names of genes involved in nucleotide metabolism. When writing the sequence of a nucleic acid, the single letter abbreviations are used (A, T, G, and C for DNA or A, U, G, and C for RNA). The letter N is often used to refer to an unspecified base.

---

adenosine monophosphate (AMP)   The nucleotide consisting of adenine, (deoxy)ribose, and one phosphate
deoxynucleoside   A nucleoside containing deoxyribose as the sugar
deoxynucleotide   A nucleotide containing deoxyribose as the sugar
nucleoside   The union of a purine or pyrimidine base with a pentose sugar
purine   Type of nitrogenous base with a double ring found in DNA and RNA
pyrimidine   Type of nitrogenous base with a single ring found in DNA and RNA
ribonucleoside   A nucleoside whose sugar is ribose (not deoxyribose)
ribonucleotide   A nucleotide whose sugar is ribose (not deoxyribose)

4. Double-Stranded DNA Forms a Double Helix

A strand of nucleic acid may be represented in various ways, either in full or abbreviated to illustrate the linkages (Fig. 3.06). As illustrated above, nucleotides are linked by joining the 5'-phosphate of one to the 3'-hydroxyl group of the next. Typically,

TABLE 3.01	Naming Bases, Nucleosides, and Nucleotides		
Base	**Abbreviations**	**Nucleoside**	**Nucleotide**
Adenine	ade A	adenosine	adenosine monophosphate (AMP)
Guanine	gua G	guanosine	guanosine monophosphate (GMP)
Cytosine	cyt C	cytidine	cytidine monophosphate (CMP)
Thymine	thy T	thymidine	thymidine monophosphate (TMP)
Uracil	ura U	uridine	uridine monophosphate (UMP)
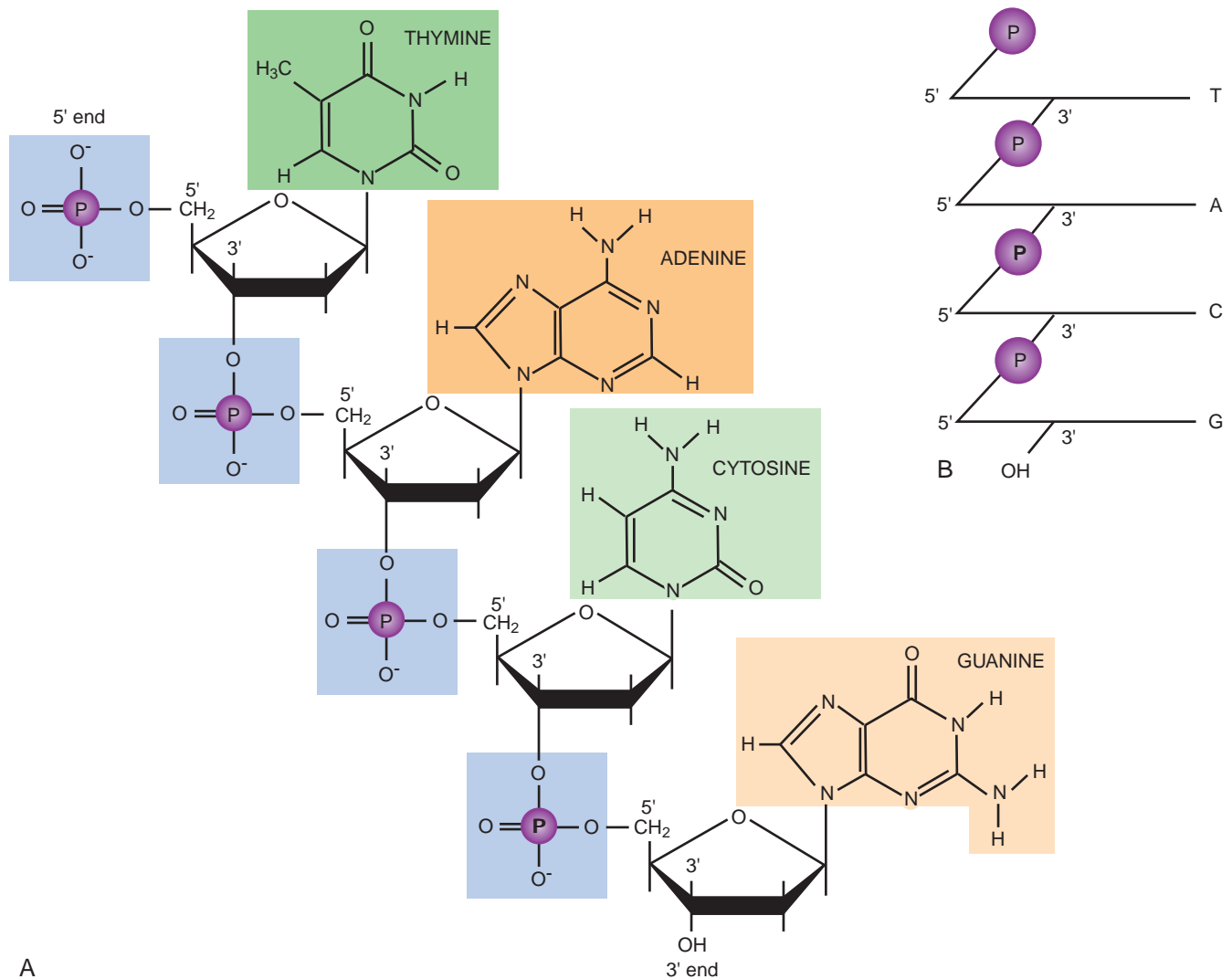
FIGURE 3.06
Some Variations in the Ways Nucleic Acids are Represented

A) More elaborate drawings show the chemical structures of the nucleic acid components, including the pentose sugar, phosphate groups, and bases. B) Simple line drawings may be used to summarize the linkage of sugars by the 5' and 3' phosphodiester bonds. Here, the protruding bases have been abbreviated to a single letter.

there is a free phosphate group at the 5'-end of the chain and a free hydroxyl group at the 3'-end of a nucleic acid strand. Consequently, a strand of nucleic acid has polarity and it matters in which direction the bases are read off. The 5'-end is regarded as the beginning of a DNA or RNA strand. This is because genetic information is read starting at the 5'-end. (In addition, when DNA is replicated, nucleic acids are synthesized starting at the 5'-end as described in Ch. 9.)

RNA is normally found as a single-stranded molecule, whereas DNA is double-stranded. Note that the two strands of a DNA molecule are **antiparallel**, as they point in opposite directions. This means that the 5'-end of one strand is opposite the 3'-end of the other strand (Fig. 3.07). Not only is DNA double-stranded, but the two separate strands are wound around each other in a helical arrangement. This is the famous **double helix** first proposed by Francis Crick and James Watson in 1953 (Fig. 3.08). The DNA double helix is stabilized both by hydrogen bonds between the bases (see below) and by stacking of the aromatic rings of the bases in the center of the helix.

In order to determine the molecular arrangement of the phosphates, sugars, and bases, Watson and Crick interpreted X-ray crystallographic data. In 1950 Maurice Wilkins and his assistant Raymond Gosling took the first images of DNA using X-ray diffraction. Gosling's work was continued by Rosalind Franklin who joined Wilkins' group the following year. Watson and Crick used an X-ray diffraction picture taken by Rosalind Franklin and Raymond Gosling in 1952 as the basis for their structural model. Rosalind Franklin died in 1958 of cancer aged 37, probably due to the effects of the X-rays. Unraveling the chemical basis for inheritance won Watson, Crick, and Wilkins the Nobel Prize in Physiology or Medicine for 1962 "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material" (Fig. 3.09).

The complementary structure of the DNA double helix is critical to replication of the genes, as described in more detail in Chapter 10.

FIGURE 3.07
Representations of Double-Stranded DNA

On the left DNA is represented as a double line consisting of two complementary strands. Actually, DNA forms a double helix, as shown to the right.

LINEAR REPRESENTATION

DOUBLE-HELIX REPRESENTATION
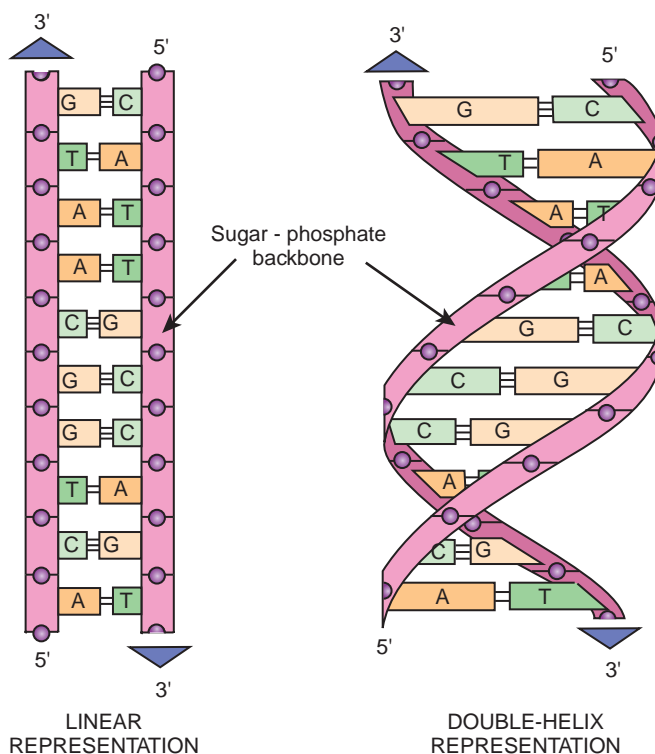
antiparallel   Parallel, but running in opposite directions
double helix   Structure formed by twisting two strands of DNA spirally around each other

NATURE

No. 4356   April 25, 1953

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3', 5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed gelices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base.

This figure is purely diagrammatic. The Two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

There is a residue on each chain every 3·4. A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrmidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid is insufficient for a rigorous test of our structure. So far as we can tell. It is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on inter-atomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the
Study of the Molecular Structure of
Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.

[1] Pauling, L., and Corey, R. B., Nature, 171, 346 (1953); Proc. U.S. Nat. Acad. Sci., 39, 84 (1953).
[2] Furberg, S., Acta Chem. Scand., 6, 634 (1952).
[3] Chargaff, E., for references see Zamenhof, S., Brawerman, G., and Chargaff, E., Biochim. et Biophys. Acta, 9, 402 (1952).
[4] Wyatt, G. R., J. Gen. Physiol., 36, 201 (1952).
[5] Astbury, W. T., Symp. Soc. Exp. Biol. 1, Nucleic Acid, 66 (Camb. Univ. Press, 1947).
[6] Wilkins, M. H. F., and Randall, J. T., Biochim. et Biophys. Acta, 10, 192 (1953).

FIGURE 3.08
DNA is a Double Helix

This one-page paper published in *Nature* described the now-famous double helix. J.D. Watson & F.H.C. Crick (1953) Molecular Structure of Nucleic Acids, A Structure for Deoxyribose Nucleic Acid. *Nature* 171: 737.

Watson and Crick in the 1950s

James Watson (b. 1928) at left and Francis Crick (b. 1916) are shown with their model of part of a DNA molecule in 1953. (Credit: A. Barrington Brown, Science Photo Library.)
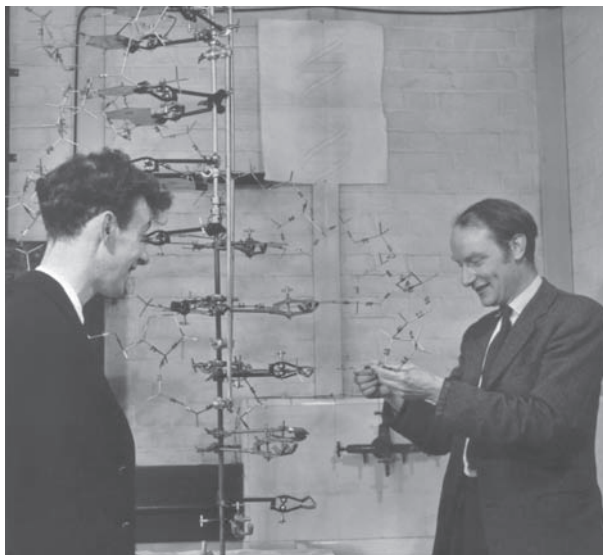
X-ray diffraction showed that two strands of DNA are twisted together forming a double helix.

DNA forms a **right-handed double helix**. To tell a right-handed helix from a left-handed helix, the observer must look down the helix axis (in either direction). In a right-handed helix, each strand turns clockwise as it moves away from the observer (in a left-handed helix it would turn counterclockwise).

Box 3.01 *The Double Helix* **by James D. Watson Published in 1968 by Atheneum, New York**

This book gives a personal account of the greatest biological advance of the twentieth century—the unraveling of the structure of the DNA double helix by James Watson and Francis Crick. Like the bases of DNA, Watson and Crick formed a complementary pair. Crick, a physicist with an annoying laugh, was supposed to be working towards a Ph.D. on protein X-ray crystallography. Watson was a homeless American biologist, wandering around Europe with a postdoctoral fellowship, looking for something to do.

Despite spending much time carousing, the intrepid heroes, Crick and Watson, beat their elders to the finish line. Watson describes with relish how the great American chemist, Linus Pauling, placed the phosphate backbone of DNA down the middle, so failing to solve the structure. The data proving the phosphate backbone was on the outside of the double helix came from Rosalind Franklin, an X-ray crystallographer at London University.

The Director of the Cavendish Laboratory at Cambridge was Sir William Bragg, the august inventor of X-ray crystallography. Despite being depicted as a stuffy has-been who nearly threw Crick out for loud-mouthed insubordination, Bragg wrote the foreword to the book. After all, when younger scientists under your direction make the greatest discovery of the century, it is no time to bear a grudge!

The biographies of great scientists are usually exceedingly dull. Who cares, after all, what Darwin liked for breakfast or what size shoes Mendel wore? It is their discoveries and how they changed the world that is fascinating. "The Double Helix" is different. Biographers are generally minor figures, understandably hesitant to criticize major achievers. Watson, himself a big name, happily lacks such respect, and cheerfully castigates other top scientists. It is this honest portrayal of the flaws and fantasies of those involved in unraveling the DNA double helix that keeps the reader's attention.

If your stomach can't stand any more sagas about caring investigators who work on into the early hours hoping that their discoveries will help sick children, this book is for you. Like most candid scientists, Watson and Crick did not work for the betterment of mankind; they did it for fun.

right-handed double helix   In a right-handed helix, as the observer looks down the helix axis (in either direction), each strand turns <u>clockwise</u> as it moves away from the observer

Box 3.02 50 Years After Determining the Structure of DNA

In 2003 the double helix celebrated its 50th anniversary. In Great Britain, the Royal Mail issued a set of five commemorative stamps illustrating the double helix together with some of the technological advances that followed, such as comparative genomics and genetic engineering. In addition, the Royal Mint issued a £2 coin depicting the DNA double helix itself (Fig. 3.10).

FIGURE 3.10
Double Helix—50th Anniversary Coin

A £2 coin commemorating the discovery of the double helix was issued in 2003 by Great Britain.

4.1. Base Pairs are Held Together by Hydrogen Bonds

In double-stranded DNA, the bases on each strand protrude into the center of the double helix where they are paired with the bases in the other strand by means of **hydrogen bonds**. Adenine (A) in one strand is always paired with thymine (T) in the other, and guanine (G) is always paired with cytosine (C) (Fig. 3.11). Consequently, the number of adenines in DNA is equal to the number of thymines, and similarly the numbers of guanine and cytosine are equal. This is called **Chargaff's rule** in honor of Edwin Chargaff who determined that there were equimolar amounts of C and G and equimolar amounts of A and T in DNA. Notice that the nucleic acid bases have amino or oxygen side-groups attached to the ring. It is these chemical groups, along with the nitrogen atoms, that are part of the rings themselves, which allow the formation of hydrogen bonds. The hydrogen bonding in DNA **base pairs** involves either oxygen or nitrogen as the atoms that carry the hydrogen, giving three alternative arrangements: O–H–O, N–H–N, and O–H–N.

Each base pair consists of one larger purine base paired with a smaller pyrimidine base. So, although the bases themselves differ in size, all of the allowed base pairs are the same width, providing for a uniform width of the helix. The A-T base pair has two hydrogen bonds and the G-C base pair is held together by three, as shown in Figure 3.11. Before the hydrogen bonds form and the bases pair off, the shared hydrogen atom is found attached to one or the other of the two bases (shown by the complete lines in Fig. 3.11). During base pairing, this hydrogen also bonds to an atom of the second base (shown by the dashed lines).

> Guanine always pairs with cytosine; adenine pairs with thymine in DNA and uracil in RNA. There is always a 1:1 ratio of G to C and A to T in DNA.

base pair   A pair of two complementary bases (A with T or G with C) held together by hydrogen bonds
Chargaff's rule   For each strand of DNA the ratio of purines to pyrimidines is always 1:1 because A always pairs with T and G always pairs with C.
hydrogen bond   Bond resulting from the attraction of a positive hydrogen atom to both of two other atoms with negative charges

FIGURE 3.11
Base Pairing by Hydrogen Bond Formation

Purines (adenine and guanine) pair with pyrimidines (thymine and cytosine) by hydrogen bonding (colored regions). When the purines and pyrimidines first come together, they form the bonds indicated by the dotted lines.

ADENINE   THYMINE

GUANINE   CYTOSINE
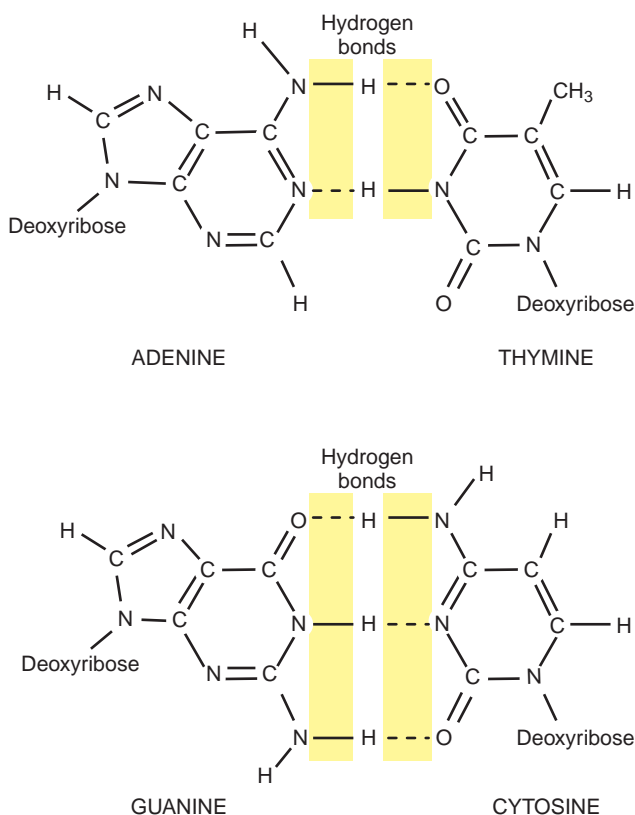
Although RNA is normally single-stranded, many RNA molecules fold up, giving double-stranded regions. In addition, a strand of RNA may be found paired with one of DNA under some circumstances. Furthermore, the genome of certain viruses consists of double-stranded RNA (see Ch. 21). In all of these cases, the uracil in RNA will base pair with adenine. Thus, the base-pairing properties of the uracil found in RNA are identical to those of the thymine of DNA.

## 4.2. Complementary Strands Reveal the Secret of Heredity

If one of the bases in a base pair of double-stranded DNA is known, then the other can be deduced. If one strand has an A, then the other will have a T, and vice versa. Similarly, G is always paired with C. This is termed complementary base pairing. The significance is that if the base sequence of either one of the strands of a DNA molecule is known, the sequence of the other strand can be deduced. Such mutually deducible sequences are known as **complementary sequences**. It is this complementary nature of a DNA double helix that allows genetic information to be inherited. Upon cell division, each daughter cell must receive a copy of the parental genome. This requires accurate duplication or replication of the DNA (Fig. 3.12). This is achieved by separating the two strands of DNA and using complementary base pairing to make a new partner for each original strand (see Ch. 10 for details).

Due to the rules for base pairing, the sequence of a DNA strand can be deduced if the sequence of its partner is known.

4.3. Melting Separates DNA Strands; Cooling Anneals Them

Hydrogen bonds are rather weak, but since a molecule of DNA usually contains millions of base pairs, the added effect of millions of weak bonds is strong enough to

complementary sequences   Two nucleic acid sequences whose bases pair with each other because A, T, G, C in one sequence correspond to T, A, C, G, respectively, in the other
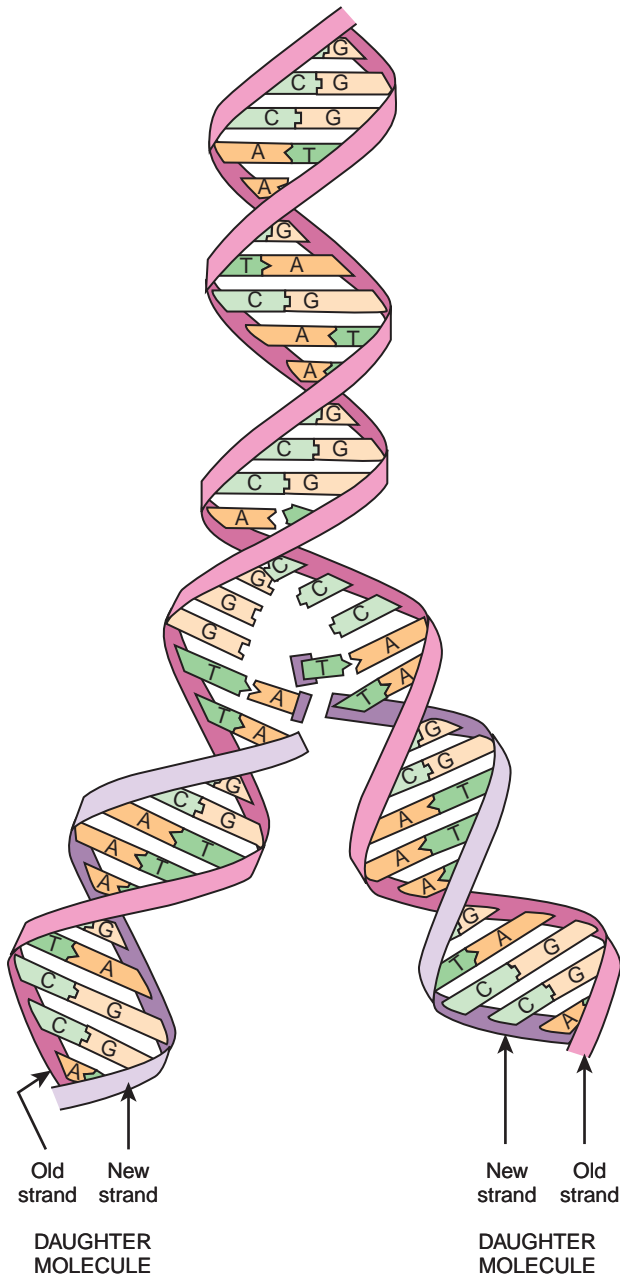
Old   New   New   Old
strand strand   strand strand

DAUGHTER   DAUGHTER
MOLECULE   MOLECULE

keep the two strands together (Fig. 3.13). When DNA is heated, the hydrogen bonds begin to break and the two strands will eventually separate if the temperature rises high enough. This is referred to as "**melting**" or **denaturation**, and each DNA molecule has a **melting temperature** (T_m) that depends on its base composition. The melting temperature of a DNA molecule is defined strictly as the temperature at the halfway point on the melting curve, as this is more accurate than trying to guess where exactly melting is complete.

Heating breaks hydrogen bonds and eventually causes the two strands of a DNA double helix to separate—the DNA "melts."

The melting temperature is affected by the pH and salt concentration of the solution, so these must be standardized if comparisons are to be made. Extremes of pH disrupt hydrogen bonds. A highly alkaline pH deprotonates the bases, which abolishes
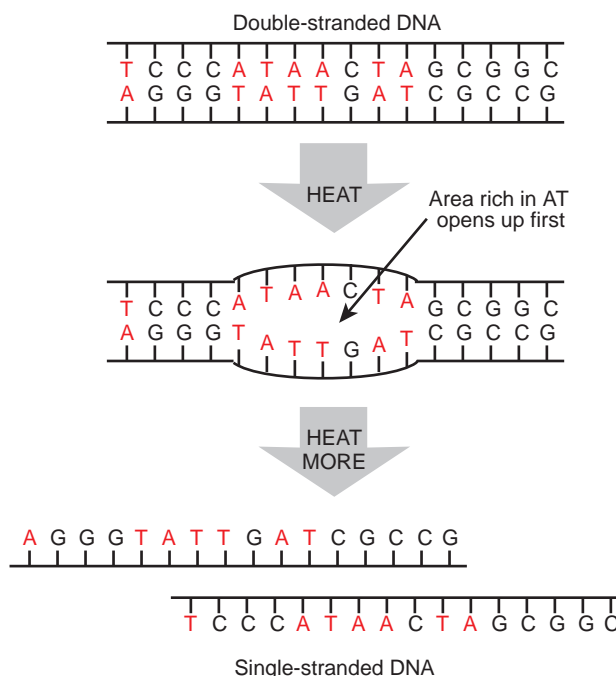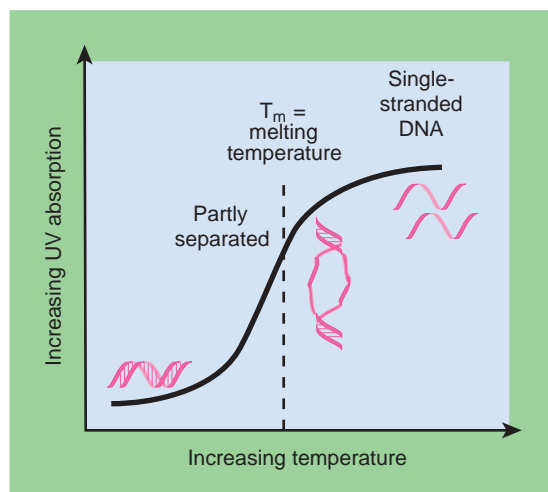
denaturation   In reference to DNA, the breaking apart of a double strand of DNA into two single strands; when used of proteins, refers to the loss of correct 3D structure
melting   When used of DNA, refers to its separation into two strands as a result of heating
melting temperature (T_m)   The temperature at which the two strands of a DNA molecule are half unpaired

Double-stranded DNA

T C C C A T A A C T A G C G G C
A G G G T A T T G A T C G C C G

HEAT → Area rich in AT opens up first

T C C C A T A A C T A G C G G C
A G G G T A T T G A T C G C C G

HEAT MORE

A G G G T A T T G A T C G C C G

T C C C A T A A C T A G C G G C

Single-stranded DNA

FIGURE 3.13
Melting of DNA

DNA strands separate, or "melt," with increasing temperature. This curve shows the measurement of DNA separation by ultraviolet absorption. As the temperature increases more UV is absorbed by the individual strands. The T_m or melting temperature is the point at which half of the double-stranded DNA has separated. During the melting process regions rich in A/T base pairs melt first since these base pairs have only two hydrogen bonds.

> Melted DNA absorbs more UV light than double-helical DNA.

> The more GC base pairs (with three hydrogen bonds) the higher the melting temperature for DNA.

their ability to form hydrogen bonds, and at a pH of 11.3 DNA becomes fully denatured. Conversely, a very low pH causes excessive protonation, which also prevents hydrogen bonding. When DNA is deliberately denatured by pH, alkaline treatment is used because unlike acid, this does not affect the glycosidic bonds between bases and deoxyribose. DNA is relatively more stable at higher ionic concentrations. This is because ions suppress the electrostatic repulsion between the negatively-charged phosphate groups on the backbone and hence exert a stabilizing effect. In pure water, DNA will melt even at room temperature.

A spectrophotometer detects the amount absorbed when light is passed through a solution containing DNA. This is compared with the light absorbed by a solution containing no DNA to determine the amount absorbed by the DNA itself. Melting is followed by measuring the absorption of ultraviolet (UV) light at a wavelength of 260 nm (the wavelength of maximum absorption), since disordered DNA absorbs more UV light than a double helix.

Overall, the higher the proportion of GC base pairs, the higher the melting temperature of a DNA molecule. This is because A/T base pairs are weaker, as they have only two hydrogen bonds, as opposed to G/C pairs, which have three. In addition, the stacking of G/C base pairs with their neighbors is also more favorable than for A/T base pairs. In the early days of molecular biology, melting temperatures were used to estimate the percentage of G/C versus A/T in samples of DNA. DNA base compositions are often cited as the **G/C ratio**. The G/C content (%G+C) is calculated from the fractional composition of bases as follows:

$$\%GC = \frac{(G + C)}{(A + T + G + C)} \times 100$$

GC ratio   The amount of G plus C divided by the total of all four bases in a sample of DNA. The GC ratio is usually expressed as a percentage

G/C contents for the DNA from different bacterial species vary from 20% to 80%, with *E. coli* having a ratio of 50%. Despite this, there is no correlation between G/C content and optimum growth temperature. Apparently, this is because the genomes of bacteria are circular DNA molecules with no free ends, and this greatly hinders unraveling at elevated temperatures. In fact, small circular DNA molecules, like plasmids, may remain base paired up to 110–120°C. In contrast, the range of G/C contents for animals (which have linear chromosomes) is much narrower, from approximately 35–45%, with humans having 40.3% G/C.

As a DNA molecule melts, regions with a high local concentration of A/T pairs will melt earlier and G/C-rich regions will stay double-stranded longer. When DNA is replicated, the two strands must first be pulled apart at a region known as the origin of replication (see Ch. 9). The DNA double helix must also be opened up when genes are transcribed to make mRNA molecules. In both cases, AT-rich tracts are found where the DNA double helix will be opened up more readily.

If the single strands of a melted DNA molecule are cooled, the single DNA strands will recognize their partners by base pairing and the double-stranded DNA will re-form. This is referred to as **annealing** or **renaturation**. For proper annealing, the DNA must be cooled slowly to allow the single strands time to find the correct partners. Furthermore, the temperature should remain moderately high to disrupt random H-bond formation over regions of just one or a few bases. A temperature 20–25°C below the T_m is suitable. If DNA from two different, but related, sources is melted and reannealed, **hybrid DNA** molecules may be obtained (Fig. 3.14).

> Upon cooling, the bases in the separated strands of DNA can pair up again and the double helix can re-form.

In the days before direct sequencing of DNA became routine, **hybridization** of DNA and/or RNA was originally used to estimate the relatedness of different organisms, especially bacteria where the amount of DNA is relatively small. Other uses for hybridization include detection of specific gene sequences and gene cloning. Several extremely useful techniques used in molecular biology rely upon hybridization, including *in situ* hybridization, which determines the location of different molecules within an organism, bacteria, or cell.

> Hybrid DNA molecules may be formed by heating and cooling a mixture of two different, but related, DNA molecules.

5. Constituents of Chromosomes

Genes are segments of large DNA molecules known as **chromosomes** (Fig. 3.15). Each chromosome is thus an exceedingly long single molecule of DNA. In addition to the DNA, which comprises the genes themselves, the chromosome has some accessory protein molecules, which help maintain its structure. The term **chromatin** refers to this mixture of DNA and protein, especially as observed with the microscope in the nuclei of eukaryotic cells. The genes are arranged in linear order. In front of each gene is a **regulatory region** of DNA involved in switching the gene on or off. In prokaryotes, groups of genes may be clustered close together with no **intergenic regions**. Such clusters are called **operons** and each is under the control of a single regulatory region. Operons are transcribed to give single mRNA molecules, each consisting of several genes.

> Genetic information includes both the genes themselves and regions of DNA involved in controlling gene expression.
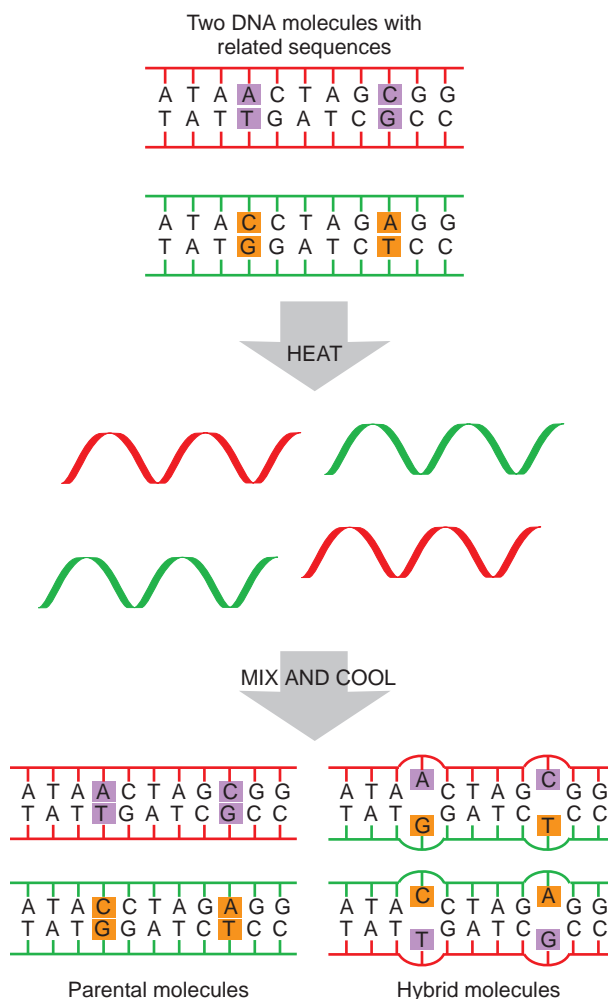
Chromosomes from bacteria are circular molecules of double-stranded DNA. Since bacteria generally have only around 3,000–4,000 genes, and the intergenic regions are very short, one chromosome is sufficient to accommodate all of their

---

annealing   The re-pairing of separated single strands of DNA to form a double helix
chromatin   Complex of DNA plus protein which constitutes eukaryotic chromosomes
chromosome   Structure containing the genes of a cell and made of a single molecule of DNA
hybrid DNA   Artificial double-stranded DNA molecule made by pairing two single strands from two different sources
hybridization   Pairing of single strands of DNA or RNA from two different (but related) sources to give a hybrid double helix
intergenic region   DNA sequence between genes
operon   A cluster of prokaryotic genes that are transcribed together to give a single mRNA (i.e., polycistronic mRNA)
regulatory region   DNA sequence in front of a gene, used for regulation rather than to encode a protein
renaturation   Re-annealing of single-stranded DNA or refolding of a denatured protein to give the original natural 3D structure

FIGURE 3.14
Annealing of DNA

When pure DNA is heated and cooled the two strands pair up again (i.e., they re-anneal). When DNA from two related sources is heated and re-annealed, hybrid strands may form. The likelihood of hybridization depends on percentage of identical sequence between the two related DNAs.

Two DNA molecules with related sequences

A T A A C T A G C G G
T A T T G A T C G C C

A T A C C T A G A G G
T A T G G A T C T C C

HEAT

MIX AND COOL

A T A A C T A G C G G
T A T T G A T C G C C

A T A C C T A G A G G
T A T G G A T C T C C

Parental molecules

A T A A C T A G C G G
T A T G G A T C T C C

A T A C C T A G A G G
T A T T G A T C G C C

Hybrid molecules

FIGURE 3.15
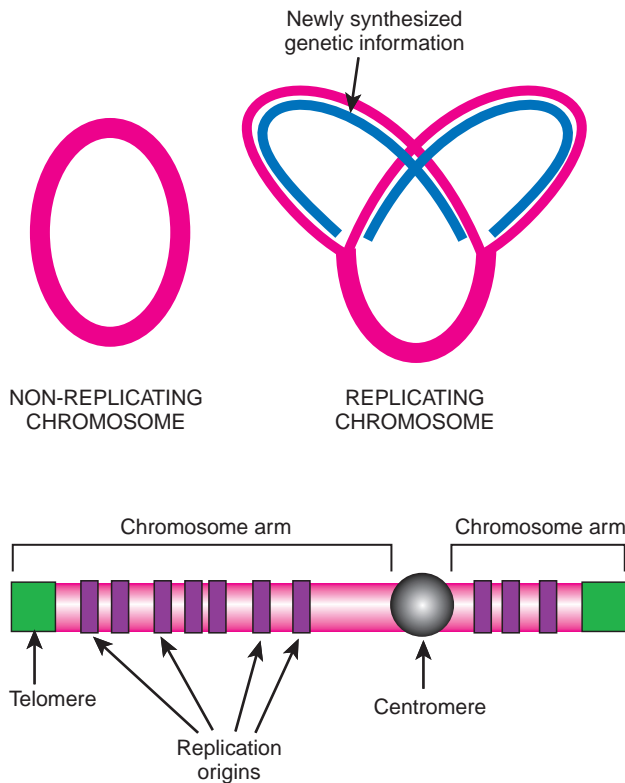*The General Pattern of Information on a Chromosome*

Genes are normally preceded by regions of DNA involved in regulation. Between the genes are regions of DNA that apparently do not carry useful genetic information. These are called intergenic regions and vary greatly in size.

Intergenic region

Gene

Intergenic region

Regulatory region

Gene

Intergenic region
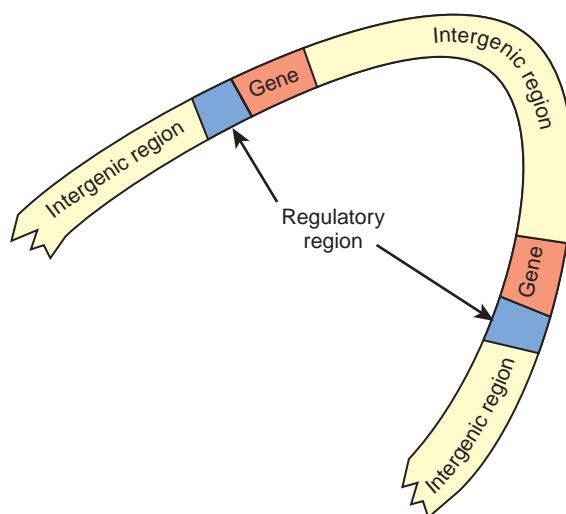
genes. When bacteria divide, the chromosome opens up at the origin of replication and replication proceeds around the circle in both directions (Fig. 3.16).

Chromosomes from higher organisms such as animals and plants are linear molecules of double-stranded DNA. They have a **centromere**, usually located more

centromere   Region of eukaryotic chromosome, usually more or less central, where the microtubules attach during mitosis and meiosis

Newly synthesized
genetic information

NON-REPLICATING
CHROMOSOME

REPLICATING
CHROMOSOME

Chromosome arm   Chromosome arm

Telomere   Centromere

Replication
origins

or less in the middle, and structures known as **telomeres** at the two ends (Fig. 3.17). Both centromeres and telomeres contain special repetitive DNA sequences allowing their recognition by particular proteins. (One exception to this rule is that the yeast, *Saccharomyces*, lacks repetitive sequences at its centromere. However, this is not a general property of fungi, as other fungi do have repetitive centromere sequences.) The centromere is used at cell division when the chromosomes replicate. The newly-divided daughter chromosomes are pulled apart by spindle fibers (or microtubules) attached to the centromeres via protein structures known as **kinetochores**.

Telomeres are critically important to maintain chromosome stability. Due to the mechanism of initiating DNA replication by an RNA primer (see Ch. 10), the far ends of linear DNA molecules are shortened by a few bases each round of replication. In those cells that are permitted to continue growing and dividing, the end sequences are restored by the enzyme **telomerase**. If telomeres become too short, cells commit suicide in order to prevent problems in cell differentiation, cancer, and aging.

Eukaryotic chromosomes are only visible under the light microscope during cell division and it is only then that a complete set of chromosomes can be visualized (Fig. 3.18). The complete set of chromosomes found in the cells of a particular individual is known as the **karyotype**. Chromosomes and specific regions of chromosomes may be identified by their staining patterns after using specific stains that emphasize regions lacking genes, known as **non-coding DNA**. This **chromosome banding technique** has been used to identify major chromosome abnormalities (Fig. 3.19).

Details of replication
mechanism and structure
vary between the linear
chromosomes of eukaryotes
and the circular chromosomes
of most bacteria.

Humans have vast amounts
of DNA making up 46 linear
chromosomes. However, most
of this sequence is non-coding
DNA as discussed further in
Chapter 4.

chromosome banding technique   Visualization of chromosome bands by using specific stains that emphasize regions lacking genes
karyotype   The complete set of chromosomes found in the cells of a particular individual
kinetochore   Protein structure that attaches to the DNA of the centromere during cell division and also binds the microtubules
non-coding DNA   DNA sequences that do not code for proteins or functional RNA molecules
telomerase   Enzyme that adds DNA to the end, or telomere, of a chromosome
telomere   Specific sequence of DNA found at the end of linear eukaryotic chromosomes

A human karyotype is a complete set of chromosomes containing 22 pairs plus one "X" and one "Y" chromosome (lower right) if the individual is male (as shown here). Females possess two "X" chromosomes. *(Credit: Alfred Pasieka, Science Photo Library.)*

Under normal circumstances, genetic information flows from DNA to RNA to protein. As a result, proteins are often referred to as "gene products." Some RNA molecules are also "gene products" as they act without being translated into protein.

6. The Central Dogma Outlines the Flow of Genetic Information

Genetic information flows from DNA to RNA to protein during cell growth. In addition, all living cells must replicate their DNA when they divide. The **central dogma** of molecular biology is a scheme showing the flow of genetic information during both the growth and division of a living cell (Fig. 3.20). During cell division each daughter cell receives a copy of the genome of the parent cell. As the genome is present in the form of DNA, cell division involves the duplication of this DNA. **Replication** is the process by which two identical copies of DNA are made from an original molecule of DNA. Replication occurs prior to cell division. An important point is that information does not flow from protein to RNA or DNA. However, information flow from RNA "backwards" to DNA is possible in certain special circumstances due to the operation of reverse transcriptase. In addition, replication of RNA occurs in viruses with an RNA genome (neither complication is shown in Fig. 3.20).

The genetic information stored as DNA is not used directly to make protein. During cell growth and metabolism, temporary, working copies of the genes known as **messenger RNA (mRNA)** are used. These are RNA copies of genetic information stored by the DNA and are made by a process called **transcription**. The mRNA molecules carry information from the genome to the cytoplasm, where the information is used by the **ribosomes** to synthesize **proteins**. In eukaryotes, mRNA is not made directly. Instead, transcription yields precursor RNA molecules (pre-mRNA) that must be processed to produce the actual mRNA as detailed in Chapter 12.

The DNA that carries the primary copy of the genes is present as gigantic molecules, each carrying hundreds or thousands of genes. In contrast, any individual mRNA molecule carries only one or a few genes' worth of information. Thus, in practice, multiple short segments of DNA are transcribed simultaneously to give many different mRNA molecules. In eukaryotes, each mRNA normally carries only a single gene, whereas in prokaryotes, anywhere from one to a dozen genes may be transcribed as a block to give an mRNA molecule carrying several genes, usually with related functions (Fig. 3.21).

central dogma   Basic plan of genetic information flow in living cells which relates genes (DNA), message (RNA), and proteins
messenger RNA (mRNA)   The molecule that carries genetic information from the genes to the rest of the cell
protein   Polymer made from amino acids; may consist of several polypeptide chains
replication   Duplication of DNA prior to cell division
ribosome   The cell's machinery for making proteins
transcription   Conversion of information from DNA into its RNA equivalent
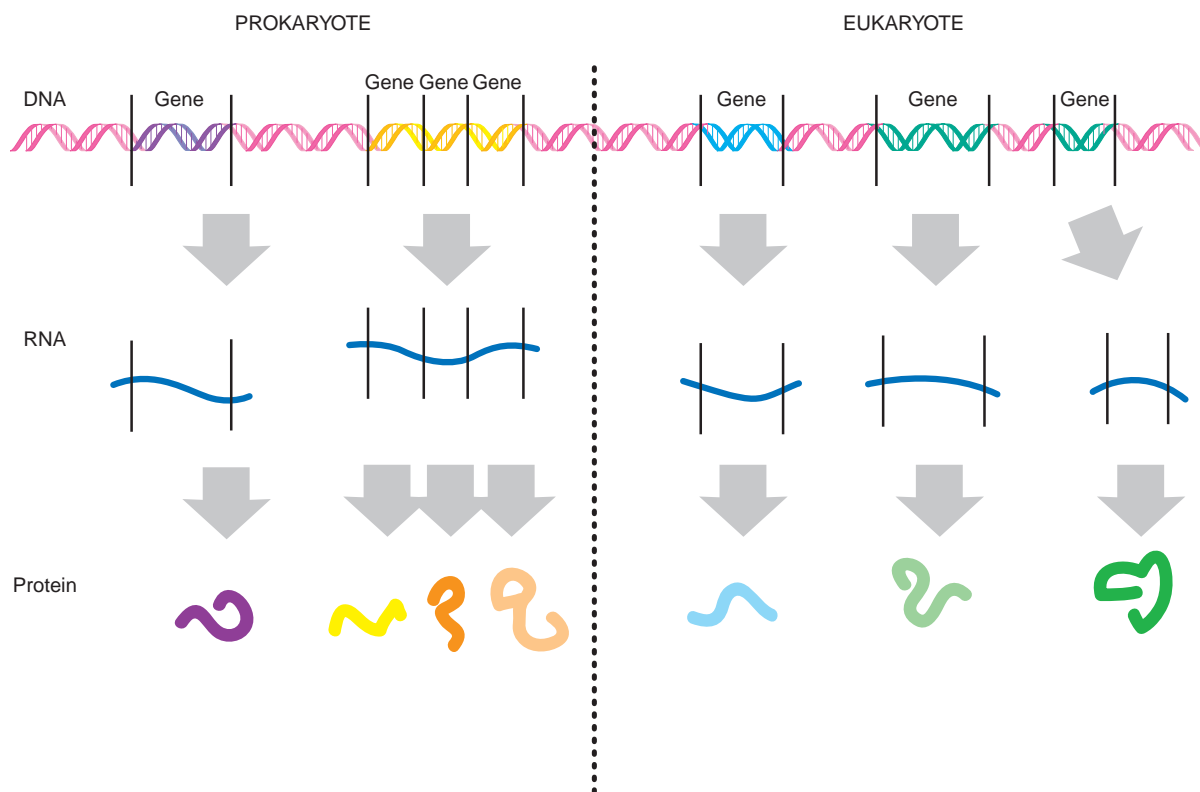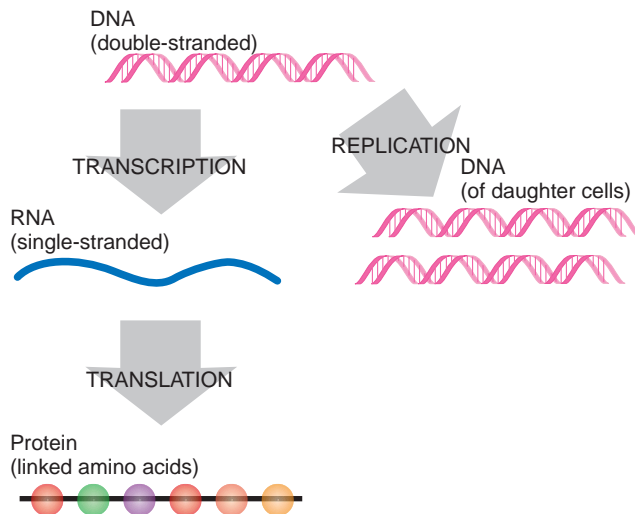
FIGURE 3.19
Banding Patterns of Human Chromosomes

Representation of the banding patterns seen in metaphase chromosomes during meiosis is shown. The bands are originally visualized by dyes. The relative distances between these bands are the same for an individual chromosome, so this is a useful way of identifying a particular chromosome. *(Credit: Dept. of Clinical Cytogenetics, Addenbrookes Hospital, Cambridge, UK, Science Photo Library.)*

FIGURE 3.20
The Central Dogma
(Simple Version)

The information flow in cells begins with DNA, which may either be replicated, giving a duplicate molecule of DNA, or be transcribed to give RNA. The RNA is read (translated) as a protein is built.

FIGURE 3.21
Differing Patterns of Transcription

In eukaryotes, each gene is transcribed to give a separate mRNA that encodes only a single protein. In prokaryotes, an mRNA molecule may carry information from a single gene or from several genes that are next to each other on the chromosome.

Translation is the synthesis of proteins using genetic information carried by mRNA. Proteins consist of one or more polymer chains known as **polypeptides**. These are made from subunits called **amino acids**. Translation thus involves transfer of information from nucleic acids to an entirely different type of macromolecule. This decoding process is carried out by ribosomes. These submicroscopic machines read the mRNA and use the information to make a polypeptide chain. Proteins, which make up about two-thirds of the organic matter in a typical cell, are directly responsible for

amino acid   Monomer from which the polypeptide chains of proteins are built
polypedtide chain   A polymer that consists of amino acids
translation   Making a protein using the information provided by mRNA

most of the processes of metabolism. Proteins perform most of the enzyme reactions and transport functions of the cell. They also provide many structural components and some act as regulatory molecules, as described below.

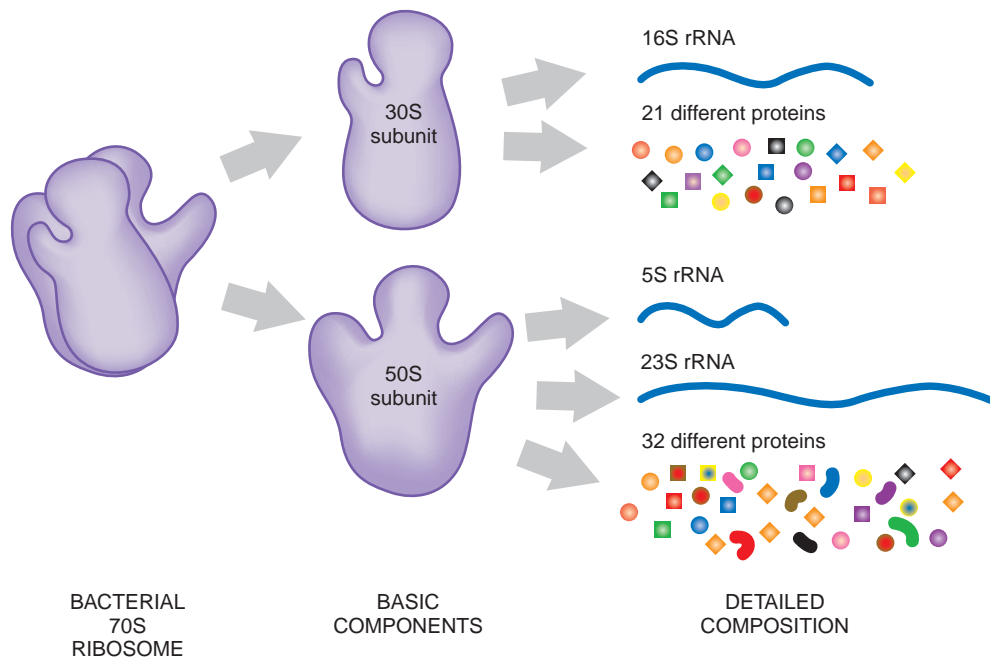## 7. Ribosomes Read the Genetic Code

This introductory section will summarize protein synthesis as it occurs in bacteria. Although the overall process is similar, the details of protein synthesis differ between bacteria and higher organisms (see Ch. 13). The bacterial ribosome consists of two subunits, small (30S) and large (50S). **S-values** indicate how fast particles sediment through a particular solution in an ultracentrifuge. They give a rough indication of size but are not linearly related to molecular weight. A complete ribosome with a 30S and a 50S subunit has an S-value of 70S (not 80S).

By weight, the ribosome itself consists of about two-thirds **ribosomal RNA** (**rRNA**) and one-third protein. In bacteria, the large subunit has two rRNA molecules, 5S rRNA and 23S rRNA, and the small subunit has just the one 16S rRNA. In addition to the rRNA, there are 52 different proteins, 31 in the large subunit and the other 21 in the small subunit (Fig. 3.22). The rRNA molecules are NOT themselves translated into protein; instead, they form part of the machinery of the ribosome that translates the mRNA.

> Proteins are made by a subcellular machine, the ribosome that uses the genetic code to read information encoded by nucleic acids.

7.1. The Genetic Code Dictates the Amino Acid Sequence of Proteins

There are 20 amino acids in proteins but only four different bases in the mRNA. So nature cannot simply use one base of a nucleic acid to code for a single amino acid when making a protein. During translation, the bases of mRNA are read off in groups of three, which are known as **codons**. Each codon represents a particular amino acid. Since there are four different bases, there are 64 possible groups of three bases; that is,

16S rRNA

21 different proteins

30S subunit

5S rRNA

23S rRNA

32 different proteins

50S subunit

BACTERIAL 70S RIBOSOME

BASIC COMPONENTS

DETAILED COMPOSITION

■ FIGURE 3.22
Structural Components of a Ribosome

The bacterial ribosome can be broken down into two smaller subunits and finally into RNA molecules and proteins.

codon Group of three RNA or DNA bases that encodes a single amino acid
ribosomal RNA (rRNA) Class of RNA molecule that makes up part of the structure of a ribosome
S-value The sedimentation coefficient is the velocity of sedimentation divided by the centrifugal field. It is dependent on mass and is measured in Svedberg units

64 different codons in the **genetic code**. However, there are only 20 different amino acids making up proteins, so some amino acids are encoded by more than one codon. Three of the codons are used for punctuation to stop the growing chain of amino acids (Fig. 3.23). In addition, the codon, AUG, encoding methionine, acts as a start codon. Thus, newly made polypeptide chains start with the amino acid methionine. (Much less often, GUG-encoding valine may also act as the start codon. However, even if the start codon is GUG the first amino acid of the newly made protein is still methionine (not valine).)

> The bases of DNA or RNA are grouped in threes for decoding.

To read the codons a set of adapter molecules is needed. These molecules, known as **transfer RNA** (**tRNA**), recognize the codon on the mRNA at one end and carry the corresponding amino acid attached to their other end (Fig. 3.24). These adapters represent a third class of RNA and were named tRNA since they transport amino acids to the ribosome in addition to recognizing the codons of mRNA. Since there are numerous codons, there are many different tRNAs. (Actually, there are fewer different tRNA molecules than codons as some tRNA molecules can read multiple codons—see Ch. 13 for details.) At one end, the tRNA has an **anticodon** consisting of three bases that are complementary to the three bases of the codon on the mRNA. The codon and anticodon recognize each other by base pairing and are held together by hydrogen bonds. At its other end, each tRNA carries the amino acid corresponding to the codon it recognizes.

The small (30S) subunit binds the mRNA, and the large (50S) subunit is responsible for making the new polypeptide chain. Figure 3.25 shows the relationship between the mRNA and the tRNAs in a stylized way. In practice, only two tRNA molecules are base paired to the mRNA at any given time. After binding to the mRNA, the ribosome moves along it, adding a new amino acid to the growing polypeptide chain each time it reads a codon from the message (Fig. 3.26).

8. Various Classes of RNA Have Different Functions

Originally, genes were regarded as units of heredity and alleles were defined as alternative versions of a gene. However, these concepts have been broadened as knowledge of
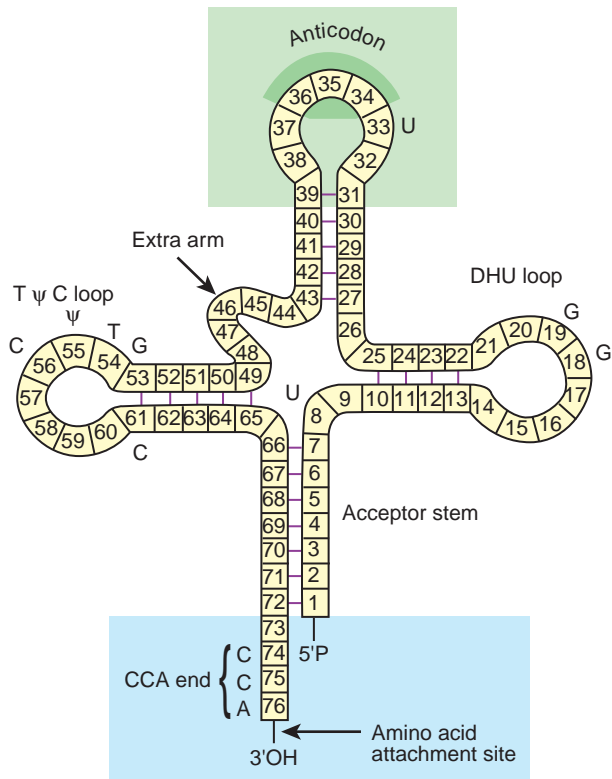
FIGURE 3.23
The Genetic Code

A codon consisting of three base pairs determines each amino acid to be added to a growing polypeptide chain. The codon table shows the 64 different codons, in RNA language, alongside the amino acids they encode. Three of the codons act as stop signals. AUG (methionine) and GUG (valine) act as start codons.

1st base	2nd (middle) base				3rd base
	U	C	A	G	
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA stop UAG stop	UGU Cys UGC Cys UGA stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

anticodon   Group of three complementary bases on tRNA that recognizes and binds to a codon on the mRNA
**genetic code**   The code for converting the base sequence in nucleic acids, read in groups of three, into the sequence of a polypeptide chain
transfer RNA (tRNA)   RNA molecules that carry amino acids to a ribosome

Transfer RNA Contains the Anticodon

Each tRNA molecule has an anticodon that is complementary to the codon carried on the mRNA. The codon and anticodon bind together by base pairing. At the far end of the tRNA is the acceptor stem ending in the bases CCA (cytosine, cytosine, and adenine). Here, the amino acid that corresponds to the codon is attached.
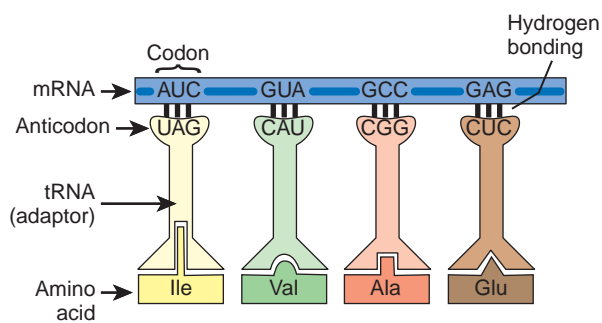
Stylized Relationship of Charged tRNA to mRNA

This figure shows a stylized relationship between the mRNA and tRNA molecules. In protein translation, there are at most two tRNAs attached to an mRNA. These are held together with the 30S subunit of the ribosome (not shown).
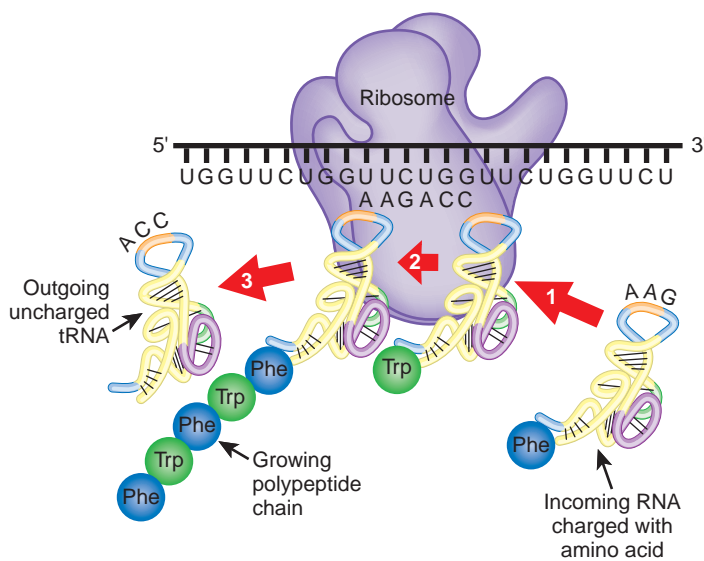
Ribosome Elongating a Polypeptide Chain

A new amino acid is added to the polypeptide chain each time a new tRNA arrives at the ribosome, bringing its attached amino acid. The anticodon of the tRNA binds to the mRNA. The large subunit crosslinks the incoming amino acid to the growing chain, such that the incoming tRNA ends up carrying the growing polypeptide chain. The 30S subunit of the ribosome then moves one step along the mRNA. This results in ejection of the left-most tRNA and readies the mRNA to accept the next incoming tRNA. The polypeptide chain continues to grow until a "stop codon" is reached.

TABLE 3.02	Major Classes of Non-Translated RNA
Ribosomal RNA (rRNA)	comprises major portion of ribosome and is involved in synthesis of polypeptide chains
Transfer RNA (tRNA)	carries amino acids to ribosome and recognizes codons on mRNA
Small nuclear RNA (snRNA)	involved in the processing of mRNA molecules in the nucleus of eukaryotic cells
microRNA (miRNA)	small RNAs encoded by the genome and used to regulate gene expression
Short interfering RNA (siRNA)	short RNA created by enzymatic cleavage of a larger double-stranded RNA and used in defense against viruses
Guide RNA	involved in processing of RNA or DNA in some organisms
Regulatory RNA	functions in the regulation of gene expression by binding to proteins or DNA or to other RNA molecules
Antisense RNA	functions in regulating gene expression by base pairing to mRNA
Recognition RNA	part of a few enzymes (e.g., telomerase); enables them to recognize certain short DNA sequences
Ribozymes	enzymatically-active RNA molecules

RNA is not so simple after all. Several classes of RNA exist that carry out a variety of roles in addition to carrying information for protein synthesis. (See especially Ch. 18 for novel insights into the role of RNA in regulation.)

genome structure has increased. Molecular insights led first to the view of genes as segments of DNA encoding proteins—the one gene—one enzyme model of Beadle and Tatum. In this case, mRNA acts as an intermediary between the DNA, which is used for storage of genetic information, and the protein, which functions in running the cell. The concept of a gene was then further extended to include segments of DNA that encode RNA molecules that are not translated into protein but function as RNA. The most common examples are the ribosomal RNA and tRNA involved in protein synthesis. The term "gene products" therefore includes such non-translated RNA molecules as well as proteins. For convenience, the major classes of non-translated RNA are summarized in Table 3.02.

In addition to the chemical differences discussed above (ribose instead of deoxyribose and uracil instead of thymine), RNA differs from DNA in several respects. RNA is usually single-stranded, although most RNA molecules do fold up, thus producing double-stranded regions. RNA molecules are usually much shorter than DNA and only carry the information for one or a few genes. Moreover, RNA is usually much shorter-lived than DNA, which is used for long-term storage of the genome. Some classes of RNA molecules, especially tRNA, contain unusual, chemically-modified bases that are never found in DNA (see Ch. 12).

The above differences in function between RNA and DNA apply to living cells. However, certain viruses carry their genomes as either single- or double-stranded RNA. In such cases, multiple genes will obviously be present on these RNA genomes. Furthermore, double-stranded viral RNA can form a double helix, similar though not identical in structure to that of DNA. The properties of viruses and the novel aspects of their genomes are discussed more fully in Chapter 21.

## 9. Proteins Carry Out Many Cell Functions

Proteins are made from a linear chain of monomers, known as amino acids, and are folded into a variety of complex 3D shapes. A chain of amino acids is called a **polypeptide chain**. There are 20 different amino acids used in making proteins. All have a central carbon atom, the **alpha carbon**, surrounded by a hydrogen atom, an amino group ($-NH_2$), a carboxyl group ($-COOH$), and a variable side chain, the

Typically, about 60% of the organic matter in a cell is protein. Most of the cell's activities and many of its structures depend on its proteins.

alpha carbon   The central carbon atom of an amino acid to which the amino, carboxyl, and R-groups are attached
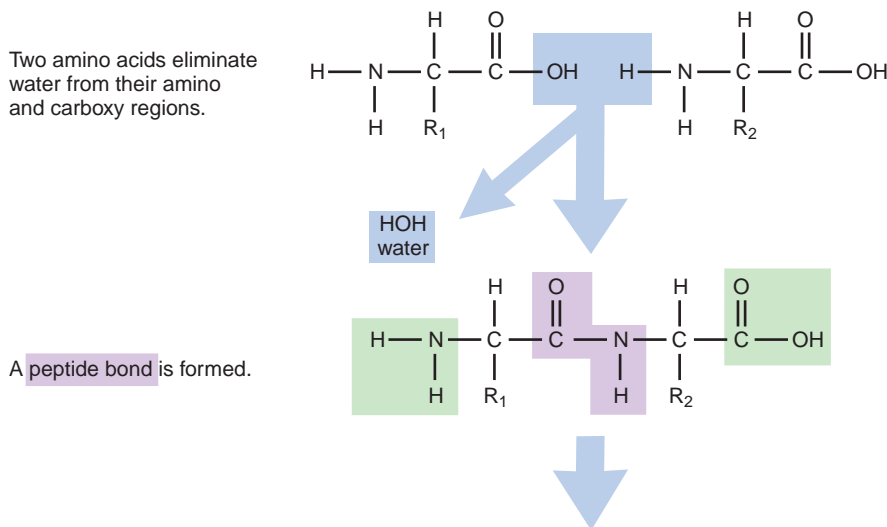**polypeptide chain**   A polymer that consists of amino acids

R-group (Fig. 3.27). Amino acids are joined together by **peptide bonds** (Fig. 3.28). The first amino acid in the chain retains its free amino group, and this end is often called the **amino-** or **N-terminus** of the polypeptide chain. The last amino acid to be added is left with a free carboxyl group, and this end is often called the **carboxy-** or **C-terminus**.

Some proteins consist of a single polypeptide chain; others contain more than one. To function properly, many proteins need extra components, called **cofactors** or **prosthetic groups**, which are not made of amino acids. Many proteins use single metal atoms as cofactors; others need more complex organic molecules.

Carbon atom

NH_2 (amino) group

Hydrogen atom

COOH (carboxyl) group

Variable group
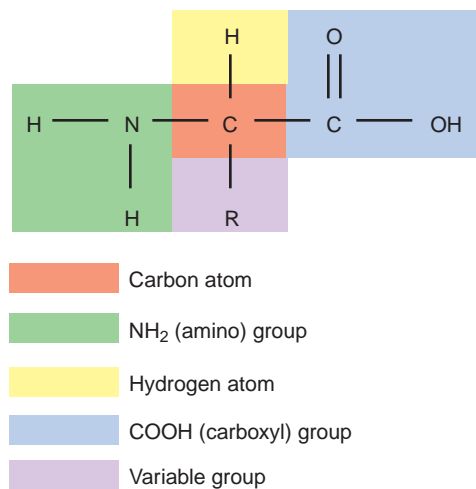
■ **FIGURE 3.27**
*General Features of Amino Acids*

Almost all amino acids found in proteins have an amino group, hydrogen atom, and carboxyl group surrounding a central carbon. In addition, the central carbon has an R-group that varies from one amino acid to the next. In glycine, the simplest amino acid, the R group is a single hydrogen atom. In proline, the R group consists of a ring structure that bonds to the nitrogen atom shown. This therefore only has a single attached hydrogen and becomes an imino group (−NH−).

Two amino acids eliminate water from their amino and carboxy regions.

HOH water

A peptide bond is formed.

■ **FIGURE 3.28**
Formation of a Polypeptide Chain

A polypeptide chain is formed as amino and carboxyl groups on two neighboring amino acids combine and eliminate water. The linkage formed is known as a peptide bond. No matter how many amino acids are added, the growing chain always has an N- or amino-terminus and a C- or carboxy-terminus.

A polypeptide is formed when multiple amino acids (AA) join, giving a linear structure with an N-terminus (NH_2) and a carboxy terminus (COOH).

H_2N - - AA_1- - AA_2- - AA_3- - AA_4- - AA_{n-2}- - AA_{n-1}- - AA_n- - COOH
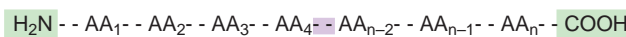
amino- or N-terminus   The end of a polypeptide chain that is made first and that has a free amino group
carboxy- or C-terminus   The end of a polypeptide chain that is made last and has a free carboxy group
cofactor   Extra chemical group non-covalently attached to a protein that is not part of the polypeptide chain
peptide bond   Type of chemical linkage holding amino acids together in a protein molecule
prosthetic group   Extra chemical group covalently attached to a protein that is not part of the polypeptide chain
R-group   Chemical group forming side chain of amino acid

9.1. The Structure of Proteins Has Four Levels of Organization

For a protein to be functional, the polypeptide chains must be folded into their correct 3D structures. The structures of biological polymers, both protein and nucleic acid, are often divided into levels of organization (Fig. 3.29). The first level, or **primary structure**, is the linear order of the monomers—that is, the sequence of the amino acids for a protein. **Secondary structure** is the folding or coiling of the original polymer chains by means of hydrogen bonding. In proteins, hydrogen bonding between peptide groups results in several possible helical or wrinkled sheet-like structures (see Ch. 14 for details).

The next level is the **tertiary structure**. The polypeptide chain, with its pre-formed regions of secondary structure, is then folded to give the final 3D structure. This level of folding depends on the side chains of the individual amino acids. In certain cases, proteins known as chaperonins help other proteins to fold correctly. As there are 20 different amino acids, a great variety of final 3D conformations is possible. Nonetheless, many proteins are roughly spherical. Lastly, **quaternary structure** is the assembly of several individual polypeptide chains to give the final structure. Not all proteins have more than one polypeptide chain; some just have one, so they have no quaternary structure.

> Proteins have four different levels of structure, from the primary sequence of amino acids to the assembly of multiple polypeptide chains.

9.2. Proteins Vary in Their Biological Roles

Functionally, proteins may be divided into four main categories: **structural proteins**, **enzymes**, **regulatory proteins,** and **transport proteins**:

1. Structural proteins make up many subcellular structures. The flagella with which bacteria swim around, the microtubules used to control traffic flow inside cells of higher organisms, the fibers involved in contractions of a muscle cell, and the outer coats of viruses are examples of structures constructed using proteins.

2. Enzymes are proteins that catalyze chemical reactions. An enzyme first binds another molecule, known as its **substrate**, and then performs chemical operations with it. Some enzymes bind only a single substrate molecule; others may bind two or more and combine them to make the final product. In any case, the enzyme needs an **active site**, a pocket or cleft in the protein, where the substrate binds and the reaction occurs. The active site of the protein is produced by the folding up of its polypeptide chain correctly so that amino acid residues that were spread out at great distances in the linear chain now come together and can cooperate in the enzyme reaction (Fig. 3.30).

3. Although regulatory proteins are not enzymes, they do bind other molecules and so they also need active sites to accommodate these. Regulatory proteins vary enormously. Many of them can bind both small signal molecules and DNA. The presence or absence of the signal molecule determines whether or not a gene is switched on (Fig. 3.31).

---

active site   Special site or pocket on a protein where the substrate binds and the enzyme reaction occurs
enzyme   A protein or RNA molecule that catalyzes a chemical reaction
primary structure   The linear order in which the subunits of a polymer are arranged
quaternary structure   Aggregation of more than one polymer chain in the final structure
regulatory protein   A protein that regulates the expression of a gene or the activity of another protein
secondary structure   Initial folding up of a polymer due to hydrogen bonding
structural protein   A protein that forms part of a cellular structure
substrate   The molecule altered by the action of an enzyme
transport protein   A protein that carries other molecules across membranes or around the body
**tertiary structure**   Final 3D folding of a polymer chain

(a) Primary structure

(b) Secondary structures

α helix   β sheet

(c) Tertiary structure

(d) Quaternary structure
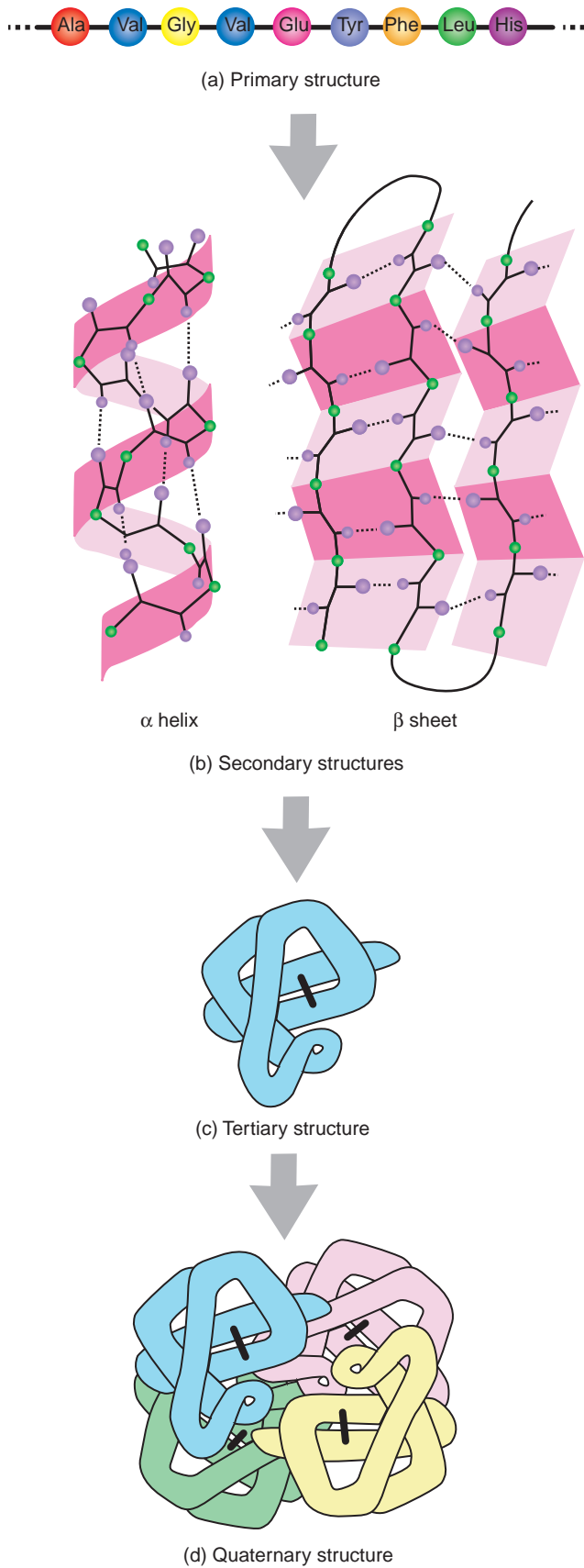
Four Levels of Protein Structure

The final protein structure is best understood by following the folding process from simple to complex. The primary structure is the specific order of the amino acids (a). The secondary structure is due to regular folding of the polypeptide chain due to hydrogen bonding (b). The tertiary structure results from further folding of the polypeptide due to interactions between the amino acid side chains (c). Finally, the quaternary structure is the assembly of multiple polypeptide chains (d).
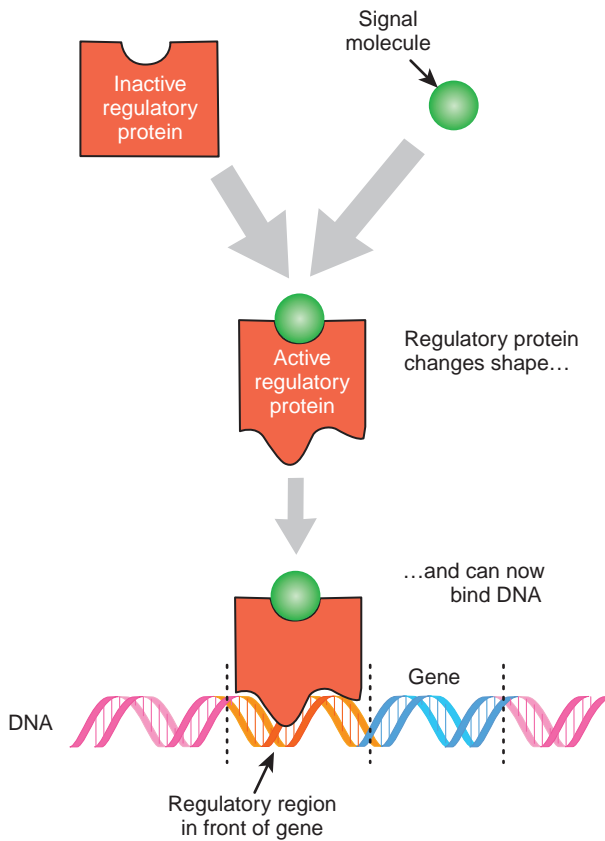
FIGURE 3.30
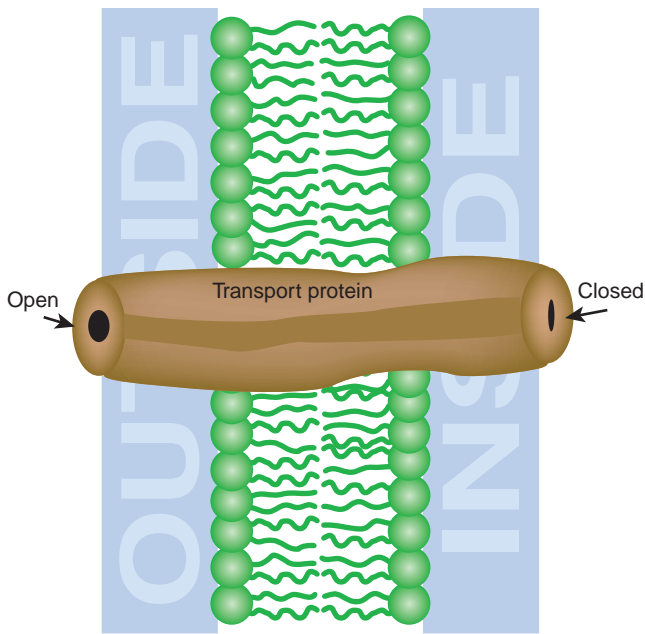*Polypeptide Forms an Active Site After Folding*

Folding of the protein brings together several regions of the polypeptide chain that are needed to perform its biological role. The active site forms a pocket for binding the substrate. Some of the amino acid residues at the active site are also involved in chemical reactions with the substrate.

Functional active site components

Polypeptide chain

FOLDING

Substrate

Pocket formed by active site residues
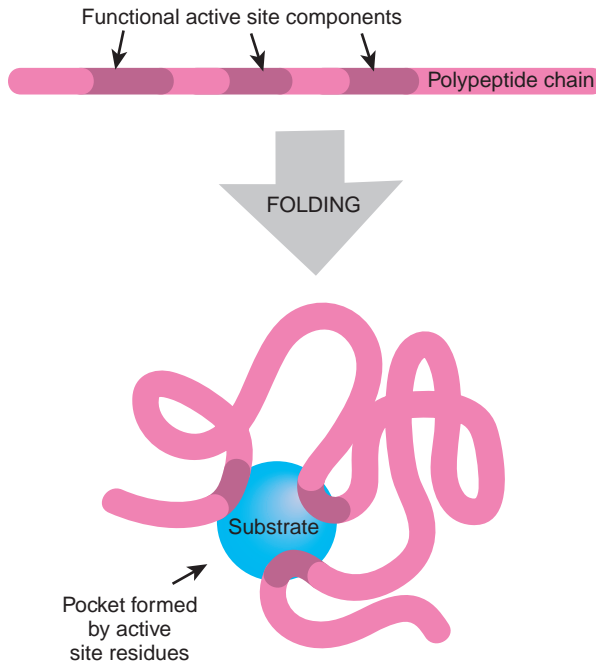
FIGURE 3.31
Regulatory Protein

Regulatory proteins usually exist in two conformations. Receiving a signal promotes a change in shape. The regulatory protein may then bind to DNA and alter the expression of a gene.

Inactive regulatory protein

Signal molecule

Active regulatory protein

Regulatory protein changes shape…

…and can now bind DNA

Gene

DNA

Regulatory region in front of gene

4. Transport proteins are found mostly in biological membranes, as shown in Figure 3.32, where they carry material from one side to the other. Nutrients, such as sugars, must be transported into cells of all organisms, whereas waste products are deported. Multicellular organisms also have transport proteins to carry materials around the body. An example is hemoglobin, which carries oxygen in blood.

Since proteins are so abundant within the cell, they assume many roles, including providing structure, regulating other proteins or genes, catalyzing reactions, and importing and exporting various solutes or molecules.

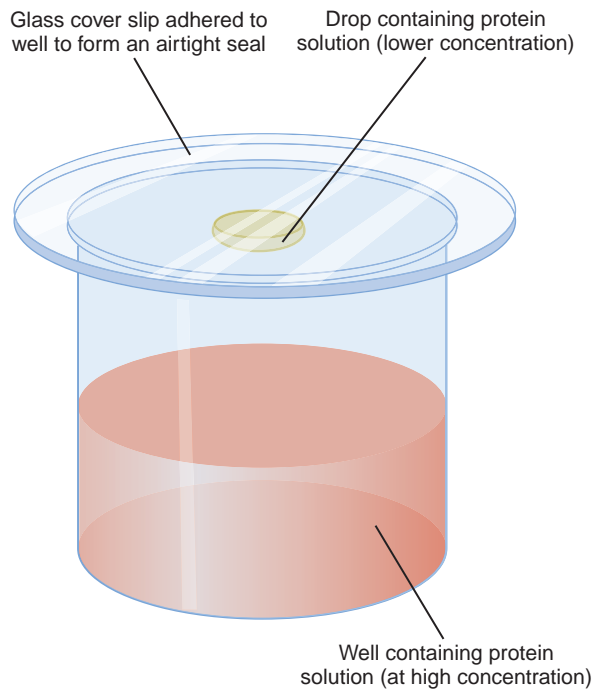## 9.3. Protein Structure is Elucidated by X-Ray Crystallography

Figuring out the exact 3D structure of a protein provides a lot of insight into its function and in the case of enzymes, the transition states of its reactions. The structure of many proteins has been elucidated using X-ray crystallography, the same technique used to elucidate the structure of DNA. The first step, and by far the most challenging, is to consolidate the pure protein into a regularly ordered crystal. The key to creating a good protein crystal is to slowly condense the protein by removing water from the protein solution. In the hanging drop method, a drop of low concentration protein solution is placed on a glass or acrylic cover slip (Fig. 3.33). This is inverted over a well containing the same protein sample at a higher concentration. The drop will become more concentrated as the water vapor moves toward equilibrium with the solution in the well. As the drop becomes more concentrated, the proteins begin to order themselves in crystals. Other methods can also be used to create crystals, such as the sitting drop method or the microdialysis methods, but each of these follow the principle of the hanging drop method.

To determine the structure of a protein, the soluble protein must be first concentrated into orderly crystals, and then X-rays are passed through the crystals. The pattern of X-ray diffraction is then interpreted and used to create a model for the actual protein structure.

The second step of X-ray crystallography is to pass X-rays through the protein crystal. The X-rays are scattered into regular repeating patterns as they pass through the atoms of the protein. The patterns of scattered X-rays are evaluated by computer programs and interpreted into a set of data that describes the location and orientation of the amino acids within the protein (Fig. 3.34).
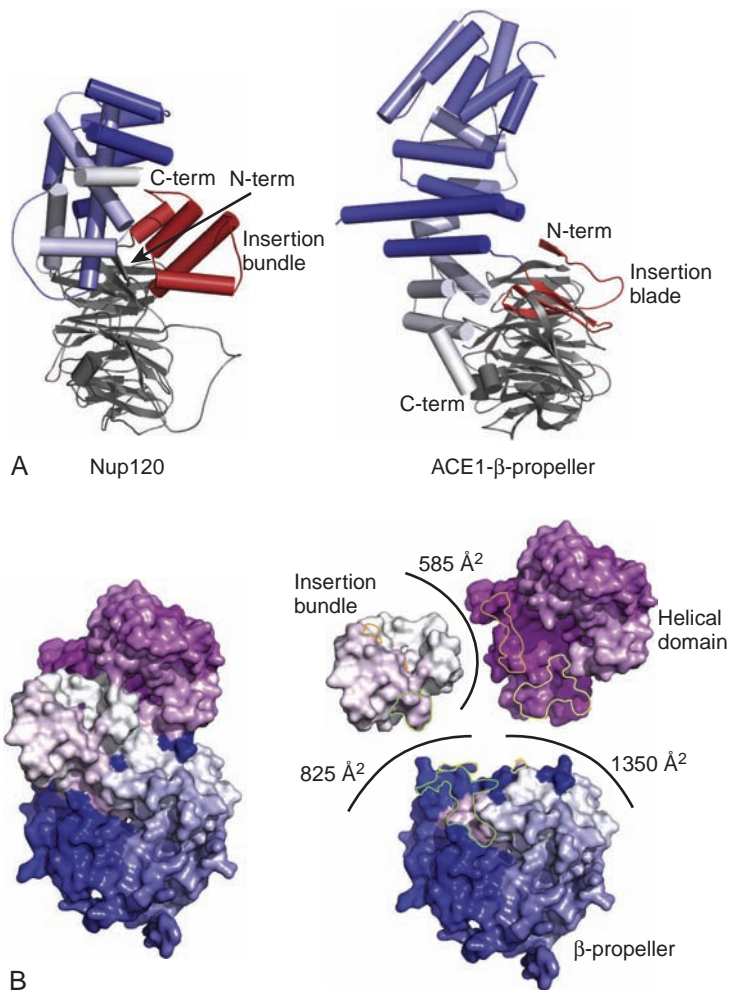
FIGURE 3.33
Hanging Drop Method

The more concentrated protein solution is at the bottom of the chamber, and the drop contains a lower concentration. Water molecules move to an equilibrium, meaning that the proteins in the drop become more and more concentrated. As they concentrate, the proteins order themselves into a regular repeating array—a crystal.

Glass cover slip adhered to well to form an airtight seal

Drop containing protein solution (lower concentration)

Well containing protein solution (at high concentration)

FIGURE 3.34
X-ray Crystal Structure of NUP120 and ACE1

The figure shows the structure of two different proteins that form the nuclear pore complex in yeast. These proteins provide the structure of the pore and lie adjacent to the lipid membrane. A. Overall structure. B. Surface model shown with subunits together and pulled apart. (Credit: Leksa, et al. (2009) Structure 17(8): 1082–1091.

C-term   N-term
Insertion bundle

N-term
Insertion blade

C-term

A   Nup120   ACE1-β-propeller

585 Å²
Insertion bundle

Helical domain

825 Å²

1350 Å²

β-propeller

B

As defined above, the central dogma of molecular biology describes the key cellular processes that convert the information stored in DNA into proteins that perform most of the functions in the cell. One of the key molecules in the central dogma is the ribosome. The actual structure of the ribosome was elusive for many years simply because it is a large complex of protein and RNA. Such a large complex was simply too difficult to crystallize into an ordered manner for X-ray diffraction. Although a formidable project, the structure of the 50S or the 30S subunits was finally crystallized by three independent groups within the same year. A discovery so grand that it earned the three primary investigators, Thomas A. Steitz, Ada E. Yonath, and Venkatraman Ramakrishnan, a Nobel prize in Chemistry. This associated article describes how the discovery of the ribosome structure has advanced our understanding of how the catalytic activity and the structure of the ribosome are intimately linked for proper protein production.

FOCUS ON RELEVANT RESEARCH

Key Concepts

- Genetic material can be classified as DNA or RNA based upon the chemical structure of the molecule. Both DNA and RNA are long polymers made of nucleotides composed of a phosphate group, a five-carbon sugar, and a nitrogen-containing base.
- DNA nucleotides have deoxyribose attached to a phosphate group and one nitrogenous base (guanine, cytosine, adenine, or thymine). RNA nucleotides have ribose which has an extra hydroxyl group attached to the 2'-carbon instead of deoxyribose, and uracil instead of thymine. Nucleotides are connected via phosphodiester bonds. The bases of a nucleotide polymer can connect via hydrogen bonds such that adenine always pairs to thymine (or uracil in RNA), and guanine always attaches to cytosine to form antiparallel double-stranded DNA or RNA.
- Two DNA strands with complementary bases are normally found connected in an antiparallel direction, where the 5'-phosphate of one strand is opposite the 3'-hydroxyl group of the other strand. The double strands spontaneously twist into a helix.
- The complementary nature of DNA is the basis for heredity.
- Two DNA strands can be melted or denatured by heat and then re-annealed by cooling. The temperature that is halfway up the melting curve is called the T_m or melting temperature. Annealing two related DNA strands is called hybridization.
- Long DNA strands form chromosomes that have different structures depending upon the organism of origin. Most bacteria have circular chromosomes, whereas humans have linear chromosomes with centromeres and telomeres.
- The central dogma of molecular biology shows the flow of genetic material during the growth and division of living cells. Replication is when one DNA molecule is copied and then transferred to the daughter cell after cell division. Transcription creates temporary copies of a gene in the form called mRNA. Translation creates the proteins used to carry out vital cellular functions by the decoding of the mRNA sequence into amino acid chains via the ribosome and tRNA.
- Ribosomes are large molecular machines consisting of both protein and RNA (ribosomal RNA—rRNA). Interestingly, the RNA components are responsible for the catalytic activity. Other types of RNA are also important to cellular

function, including transfer RNA (tRNA), microRNA (miRNA), short inter-fering RNA (siRNA), **antisense RNA**, and **ribozymes**.

- The genetic code in DNA and mRNA is based on three nucleotides read together as a codon.

- Proteins carry out many of the cellular functions. Structural proteins provide structure, enzymes catalyze chemical reactions, regulatory proteins regulate other molecules, and transport proteins move other molecules in and out of the cell.

- Protein structure is controlled by the interaction of the amino acid side chains with each other and the environment. Proteins have primary structure, the linear order of amino acids, secondary structure, the folding or coiling of the amino acids due to hydrogen bonding, tertiary structure, where the coils or sheets fold into a 3D shape, and in some proteins, quaternary structure where multiple polypeptide chains combine into one complex. Protein structure can be elucidated via X-ray diffraction patterns.

Review Questions

1. What are nucleic acids? Name the two types of nucleic acids.
2. What are the three components of a nucleotide?
3. Name the bond that links the nucleotides.
4. Name the four bases in DNA. Name the four bases in RNA.
5. What is the difference between: a) a purine and a pyrimidine; b) a nucleotide and a nucleoside?
6. What is at the beginning and the end of a DNA or RNA molecule?
7. Name the bond that holds together the double strands of DNA.
8. Describe the Watson-Crick model of a DNA molecule.
9. What is the difference between A/T pairing and G/C pairing?
10. What are chromosomes? What is the difference between the prokaryotic and eukaryotic chromosomes?
11. How are genes arranged on a chromosome?
12. How are DNA strands duplicated? What is this process called?
13. What is a centromere? What is its function?
14. What are telomeres? What is the function of telomerase? What is the importance of telomeres?
15. What is a karyotype? Give an example. When can it be visualized?
16. Describe a technique used to identify chromosome abnormalities.
17. What is the central dogma of molecular biology?
18. Define replication, transcription, and translation.
19. What are the differences between prokaryotic and eukaryotic mRNA?
20. What are the structural components of a bacterial ribosome?
21. What is the function of ribosomal RNA?
22. What is a codon? How many codons are possible in the genetic code?
23. What is the function of a start codon? What is the first amino acid in a newly made protein?
24. What are tRNAs? What are their functions?
25. Name the three RNAs involved in the formation of protein.
26. List the major differences between DNA and RNA.
27. What is the general structure of an amino acid? How many different amino acids are used to make proteins?
28. What are proteins? How are they formed?
29. What are the N-terminus and C-terminus of a protein?
30. What are the four levels of folding that makes the protein active?
31. List the four main functional roles of proteins.

antisense RNA RNA complementary in sequence to mRNA and which, therefore, base pairs with it and prevents translation
ribozyme RNA molecule that acts as an enzyme

Conceptual Questions

1. a. Draw the complementary sequence for the following single strand of DNA:

    ```
    5' CTATCGATTCAACGAAATTCGCAAGGCATT 3'
    ```

    b. Transcribe the double-stranded DNA from question 1a into a single-stranded mRNA using the top strand as the template.
    c. Translate the mRNA from question 2b into protein using the codon chart.
2. Explain the central dogma of molecular biology.
3. A scientist was given a solution containing ribosomes and was told to isolate the different subunits from the mixture. Present an experimental technique to isolate ribosomal subunits.
4. A researcher studying the gene for curly hair in mice has found that one nucleotide is different between the gene for hair shape enzyme in the curly-haired mouse and straight-haired mouse. This mutation alters the gene so that a different amino acid is added at the location. Explain the results of this research based on your knowledge of protein functions.
5. Calculate the percent content of each of the four bases for the following organisms. Deduce the information based on the provided data.

Organism	Guanine (G)	Cytosine (C)	Thymine (T)	Adenine (A)
Streptomyces coelicolor		38%		
Saccharomyces cerevisiae			19%	
Arabidopsis thaliana	18%			
Plasmodium falciparum				30%