CHAPTER

10

Gene Action and Expression

10.1 Transcription—The Link Between Gene and Protein

The proteins that a particular type of cell manufactures constitute its proteome. The information in the nucleotide base sequence of a protein-encoding gene is transcribed into mRNA, which is then translated into the amino acid sequence of a protein.

10.2 Translating a Protein

The genetic code is the correspondence between RNA base triplets and particular amino acids. It is universal—the same messenger RNA codons specify the same amino acids in humans, hippos, herbs, and bacteria. Ribosomes provide structural support and enzyme activity for transfer RNA molecules to align the designated amino acids against a messenger RNA. The amino acids link, building proteins.

10.3 The Human Genome Sequence Reveals Unexpected Complexity

The human genome sequence reveals that the classic one gene-one protein paradigm is an oversimplification of the function of DNA. Only 1.5 percent of the genome encodes protein, yet those 31,000 or so genes specify more than 100,000 different proteins.

DNA replication preserves genetic information by endowing each new cell with a complete set of operating instructions. A cell uses some of the information to manufacture proteins. To do this, first the process of **transcription** copies a particular part of the DNA sequence of a chromosome into an RNA molecule that is complementary to one strand of the DNA double helix. Then the process of **translation** uses the information copied into RNA to manufacture a specific protein by aligning and joining the specified amino acids. The overall events of transcription and translation are referred to as **gene expression.**

Watson and Crick, shortly after publishing their structure of DNA in 1953, described the relationship between nucleic acids and proteins as a directional flow of information called the "central dogma" (figure 10.1). Francis Crick explained in 1957, "The specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and this sequence is a code for the amino acid sequence of a particular protein." But understanding the central dogma was only a beginning. Today, nearly half a century later, researchers all over the world are detailing the patterns of gene expression in the many types of cells that build a human body, as well as in diseased cells. The central dogma explained how a gene encodes a protein; it did not explain how a cell "knows" which genes to express. What, for example, directs a bone cell to transcribe the genes that control the synthesis of collagen, and not to transcribe those that specify muscle proteins? How does a stem cell in bone marrow "know" when to divide and send daughter cells on pathways to differentiate as white blood cells, red blood cells, or platelets—and how does the balance of blood cell production veer from normal when a person has leukemia? Genes do more than encode proteins. They also control each other's functioning, in sometimes complex hierarchies.

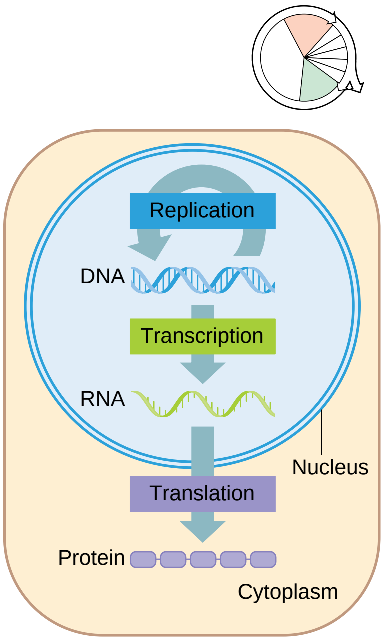


figure 10.1

DNA to RNA to protein. The central dogma of biology states that information stored in DNA is copied to RNA (transcription), which is used to assemble proteins (translation). DNA replication perpetuates genetic information. This figure repeats within the chapter, with the part under discussion highlighted.

Knowing the human genome sequence has revealed new complexities and seeming contradictions. The neat one gene-one protein picture painted by the work that followed Watson and Crick's description of DNA is correct, but a gross oversimplification. Our 31,000 or so genes actually encode between 100,000 to 200,000 proteins. Yet only a small part of our genome encodes protein. Even if the few researchers who hypothesize that the human genome contains 60,000 or more genes are correct, that is still a very small portion of the genome. This chapter presents the classical elucidation of the **genetic code**—the correspondence between gene and protein—and concludes with a consideration of the new questions posed by human genome sequence information.

10.1 Transcription—The Link Between Gene and Protein

RNA is of crucial importance in the flow of genetic information, serving as a bridge between gene and protein. Cells replicate their DNA only during S phase of the cell cycle. In contrast, transcription and translation occur continuously during the cell cycle, to supply the proteins essential for life as well as those that give a cell its specialized characteristics.

RNA and DNA share an intimate relationship, as figure 10.2 depicts. RNA is synthesized against one side of the double helix, called the **template strand,** with the assistance of an enzyme called **RNA polymerase.** The other side of the DNA double helix is the **coding strand.** RNA comes in three major types, and several less abundant types, distinguished by their functions (although we do not yet know all of them). RNA also differs in a few ways from DNA. We begin our look at transcription—the prelude to translation—by considering RNA.

RNA Structure and Types

RNA and DNA are both nucleic acids, consisting of sequences of nitrogen-containing bases joined by sugar-phosphate backbones. However, RNA is usually single-stranded, whereas DNA is double-stranded. Also, RNA has the pyrimidine base **uracil** in place of DNA's thymine. As their names imply, RNA nucleotides include the sugar ribose, rather than the deoxyribose that is part of DNA. Functionally, DNA stores genetic information, whereas RNA actively utilizes that information to enable the cell to synthesize a particular protein. Table 10.1 and

table 10.1

How RNA and DNA Differ

RNA	DNA
Usually single-stranded	Usually double-stranded
Has uracil as a base	Has thymine as a base
Ribose as the sugar	Deoxyribose as the sugar
Carries protein-encoding information	Maintains protein-encoding information
Can be catalytic	Not catalytic

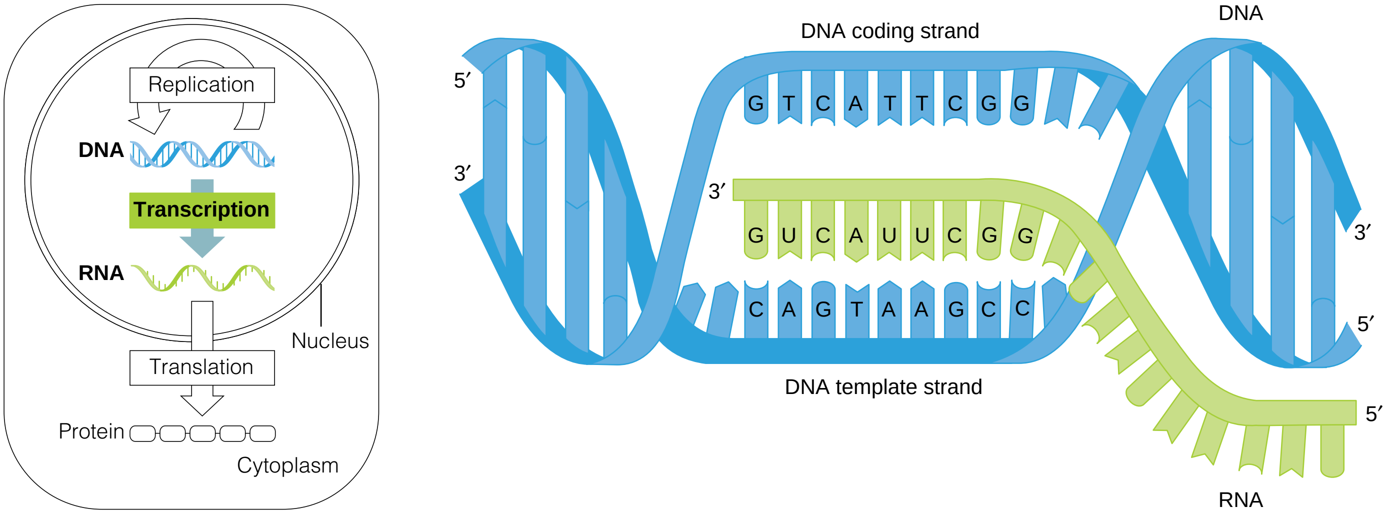


figure 10.2

The relationship among RNA, the DNA template strand, and the DNA coding strand. The RNA sequence is complementary to that of the DNA template strand and therefore is the same sequence as the DNA coding strand, with uracil (U) in place of thymine (T).

figure 10.3 summarize the differences between RNA and DNA.

As RNA is synthesized along DNA, it folds into three-dimensional shapes, or **conformations,** that are determined by complementary base pairing within the same RNA molecule. These shapes are very important for RNA's functioning. The three major types of RNA are messenger RNA, ribosomal RNA, and transfer RNA (table 10.2).

Messenger RNA (mRNA) carries the information that specifies a particular protein product. Each three mRNA bases in a row form a genetic code word, or **codon,** that specifies a certain amino acid. Because genes vary in length, so do mRNA molecules. Most mRNAs are 500 to 2,000 bases long. At any one time, different cell types have different mRNA molecules, present in differing amounts, that reflect their functions. A muscle cell has many mRNAs that specify the abundant contractile proteins actin and myosin, whereas a skin cell contains many mRNAs corresponding to the gene that encodes the scaly protein keratin. Geneticists use DNA microarrays to identify the types and amounts of mRNAs by converting the mRNAs of a cell into DNA called complementary or cDNA.

The information encoded in an mRNA sequence cannot be utilized without the participation of two other major classes of RNA. **Ribosomal RNA** (rRNA)

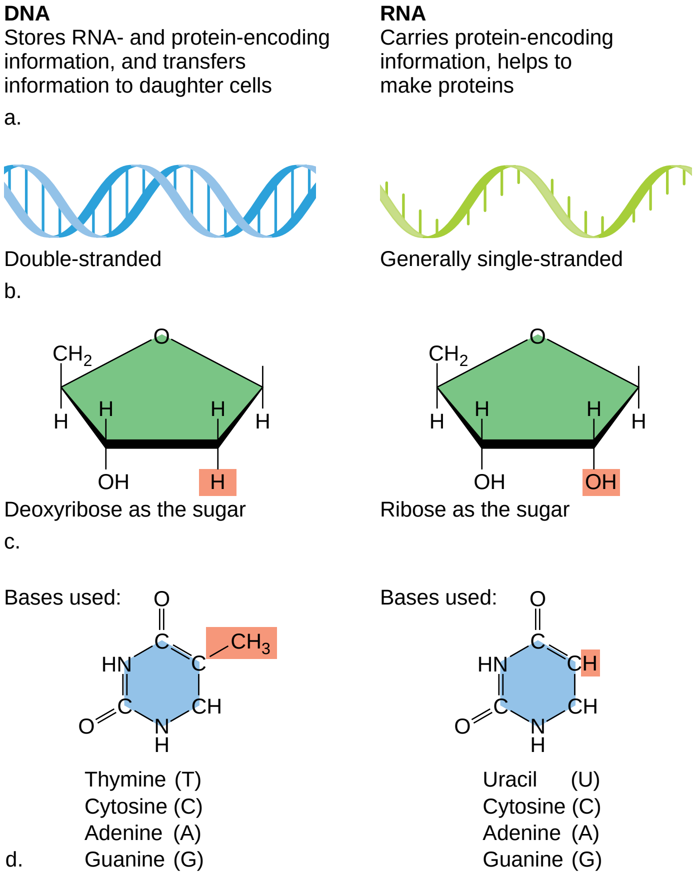


figure 10.3

DNA and RNA differences. (*a*) DNA is double-stranded; RNA is usually single-stranded (*b*). DNA nucleotides include deoxyribose, whereas RNA nucleotides have ribose (*c*). Finally, DNA nucleotides include the pyrimidine thymine, whereas RNA has uracil (*d*).

molecules range from 100 to nearly 3,000 nucleotides long. This type of RNA associates with certain proteins to form a ribosome. Recall from chapter 2 that a ribosome is a structural support for protein synthesis (figure 10.4). A ribosome has two subunits that are separate in the cytoplasm but join at the initiation of protein synthesis. The larger ribosomal subunit has three types of rRNA molecules, and the small subunit has one. Ribosomal RNA, however, is much more than a structural support. Certain rRNAs catalyze the formation of bonds between amino acids. Such an RNA with enzymatic function is called a **ribozyme.** Other rRNAs help to align the ribosome and mRNA.

The third major type of RNA molecule is **transfer RNA** (tRNA). These molecules are "connectors" that bind mRNA codons at one end and specific amino acids at the other. A tRNA molecule is only 75 to 80 nucleotides long. Some of its bases weakly bond with each other, folding the tRNA into loops that form a characteristic cloverleaf shape (figure 10.5). One loop of the tRNA has three bases in a row that form the **anticodon,** which is complementary to an mRNA codon. The end of the tRNA opposite the anticodon strongly bonds to a specific amino acid. A tRNA with a particular anticodon sequence always carries the same amino acid. (There are 20 types of amino acids in organisms.) For example, a tRNA with the anticodon sequence GAA always picks up the amino acid phenylalanine. Special enzymes attach amino acids to tRNAs that bear the appropriate anticodons.

Transcription Factors

Study of the control of gene expression began in 1961, when French biologists François Jacob and Jacques Monod described the remarkable ability of *E. coli* to produce the enzymes to metabolize the sugar lactose only when lactose is present in the cell's surroundings. What "tells" a simple bacterial cell to transcribe those products it needs—at exactly the right time?

Jacob and Monod discovered and described how a modified form of lactose turned on the genes whose encoded proteins break down the sugar. Jacob and Monod named the set of genes that are coordinately controlled an operon. Wrote Jacob and

table 10.2

Major Types of RNA

Type of RNA	Size (number of nucleotides)	Function
mRNA	500–3,000+	Encodes amino acid sequence
rRNA	100–3,000	Associates with proteins to form ribosomes, which structurally support and catalyze protein synthesis
tRNA	75–80	Binds mRNA codon on one end, amino acid on the other, linking a gene's message to the amino acid sequence it encodes

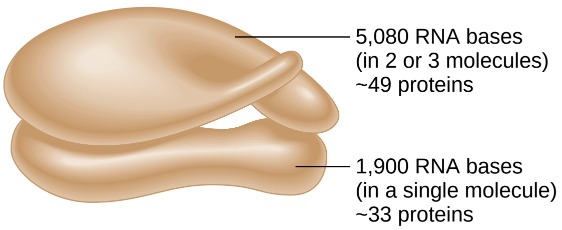


figure 10.4

The ribosome. A ribosome from a eukaryotic cell, shown here, has two subunits, containing 82 proteins and 4 rRNA molecules altogether.

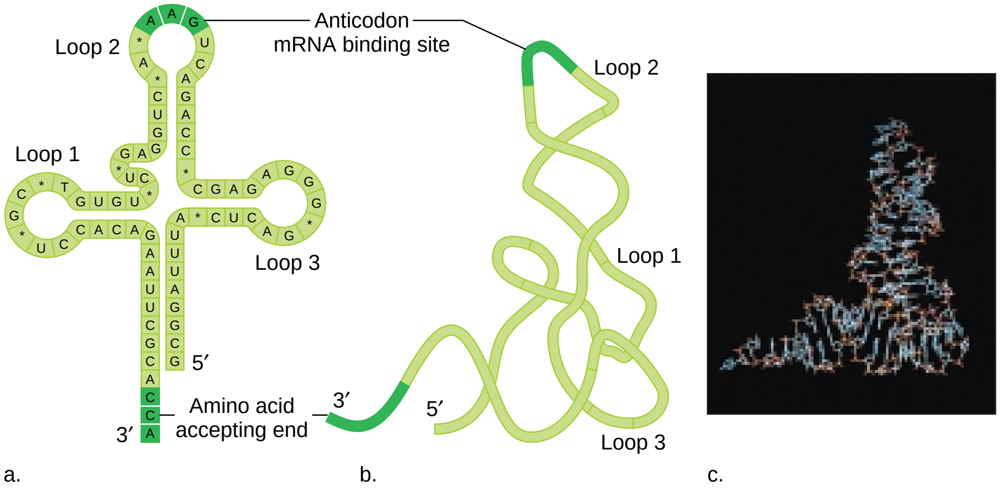


figure 10.5

Transfer RNA. (*a*) Certain nucleotide bases within a tRNA molecule hydrogen bond with each other to give the molecule a "cloverleaf" conformation that can be represented in two dimensions. The darker bases at the top form the anticodon, the sequence that binds a complementary mRNA codon. Each tRNA terminates with the sequence CCA, where a particular amino acid covalently bonds. Three-dimensional representations of a tRNA (*b*) and (*c*) depict the loops that interact with the ribosome to give tRNA its functions in translation.

Monod in 1961, "The genome contains not only a series of blueprints, but a coordinated program of protein synthesis and means of controlling its execution." Operons were originally described in several types of bacteria, but the genome sequence of the roundworm *Caenorhabditis elegans* revealed that nearly a quarter of its genes are organized into operon-like groups, too.

In bacteria, operons act like switches, turning gene transcription on or off. In multicellular eukaryotes like ourselves, genetic control is more complex because different cell types express different subsets of genes. To manage such complexity, groups of proteins called **transcription factors** come together, forming an apparatus that binds DNA at certain sequences and initiates transcription at specific sites on a chromosome. The transcription factors, activated by signals from outside the cell, set the stage for transcription to begin by forming a pocket for RNA polymerase—the enzyme that actually builds an RNA chain.

Several types of transcription factors are required to transcribe a gene. Because transcription factors are proteins, they too are gene-encoded. The DNA sequences that transcription factors bind may be located near the genes they control, or as far as 40,000 bases away. DNA may form loops so that the genes encoding proteins that act together come near each other for transcription. Proteins in the nucleus may help bring certain genes and their associated transcription factors in close proximity, much as books on a specialized topic might be grouped together in a library for easier access.

About 2,000 transcription factors are known, and defects in them cause some diseases. For example, Huntington disease results from a defect in a gene whose encoded protein, huntington, is a transcription factor. Part of its normal function is to turn on transcription of a gene that encodes brain-derived neurotrophic factor (BDNF). Neurons in a part of the brain called the striatum require BDNF to stay alive. With abnormal huntington, not enough BDNF interacts with the striatal neurons, and they die. It may take many years to produce the uncontrollable movements and other changes characteristic of the disorder.

Many transcription factors have regions in common, called motifs, that fold into similar three-dimensional shapes, or conformations. These motifs generally enable the transcription factor to bind DNA. They have very colorful names, such as "helix-turn-helix," "zinc fingers," and "leucine zippers," that reflect their distinctive shapes.

Steps of Transcription

How do transcription factors and RNA polymerase "know" where to bind to DNA to begin transcribing a specific gene? Transcription factors and RNA polymerase are attracted to a **promoter,** which is a special sequence that signals the start of the gene. Figure 10.6 shows one order in which transcription factors bind, to set up a site to receive RNA polymerase, called a preinitiation complex. The first transcription factor to bind, called a TATA binding protein, is attracted to a DNA sequence called a TATA box, which consists of the base sequence TATA surrounded by long stretches of G and C. Once the first transcription factor binds, it attracts others in groups and finally RNA polymerase joins the complex, binding just in front of the start of the gene sequence. The coming together of these components constitutes transcription initiation.

Complementary base pairing underlies transcription, just as it does DNA replication. In the next stage, transcription elongation, enzymes unwind the DNA double helix, and RNA nucleotides bond with exposed complementary bases on the DNA template strand (figure 10.2). RNA polymerase adds the RNA nucleotides in the sequence the DNA specifies, moving along the DNA strand in a 3′ to 5′ direction, synthesizing the RNA molecule in a 5′ to 3′ direction. A terminator sequence in the DNA indicates where the gene's RNA-encoding region ends. This is transcription termination.

For a particular gene, RNA is transcribed using only one strand of the DNA double helix as the template. The other

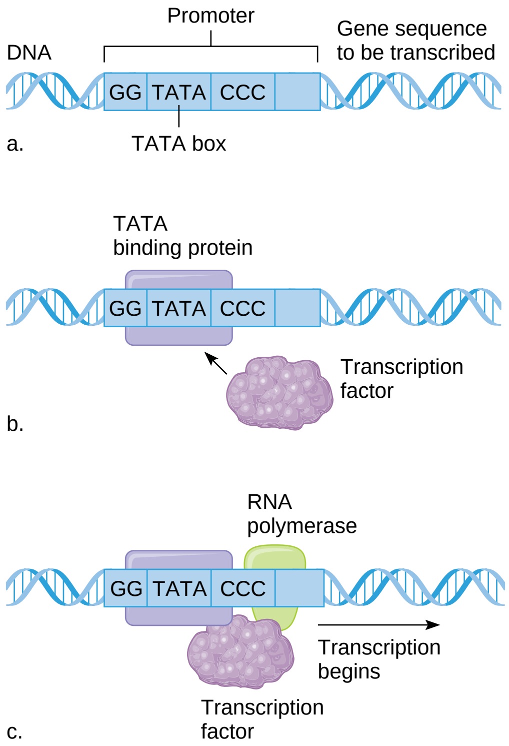


figure 10.6

Setting the stage for transcription to begin. (*a*) The promoter region of a gene has specific sequences recognized by proteins that initiate transcription. (*b*) A binding protein recognizes the TATA region and binds to the DNA. This allows other transcription factors to bind. (*c*) The presence of the necessary transcription factors allows RNA polymerase to bind and begin making RNA.

DNA strand that isn't transcribed is called the coding strand because its sequence is identical to that of the RNA, except with thymine (T) in place of uracil (U). Several RNAs may be transcribed from the same DNA template strand simultaneously (figure 10.7). Since RNA is relatively short-lived, a cell must constantly transcribe certain genes to maintain supplies of essential proteins. However, different genes on the same chromosome may be transcribed from different halves of the double helix.

To determine the sequence of RNA bases transcribed from a gene, write the RNA bases that are complementary to the template DNA strand, using uracil opposite adenine. For example, if a DNA template strand has the sequence

CCTAGCTAC

then it is transcribed into RNA with the sequence

GGAUCGAUG

The coding DNA sequence is:

GGATCGATG

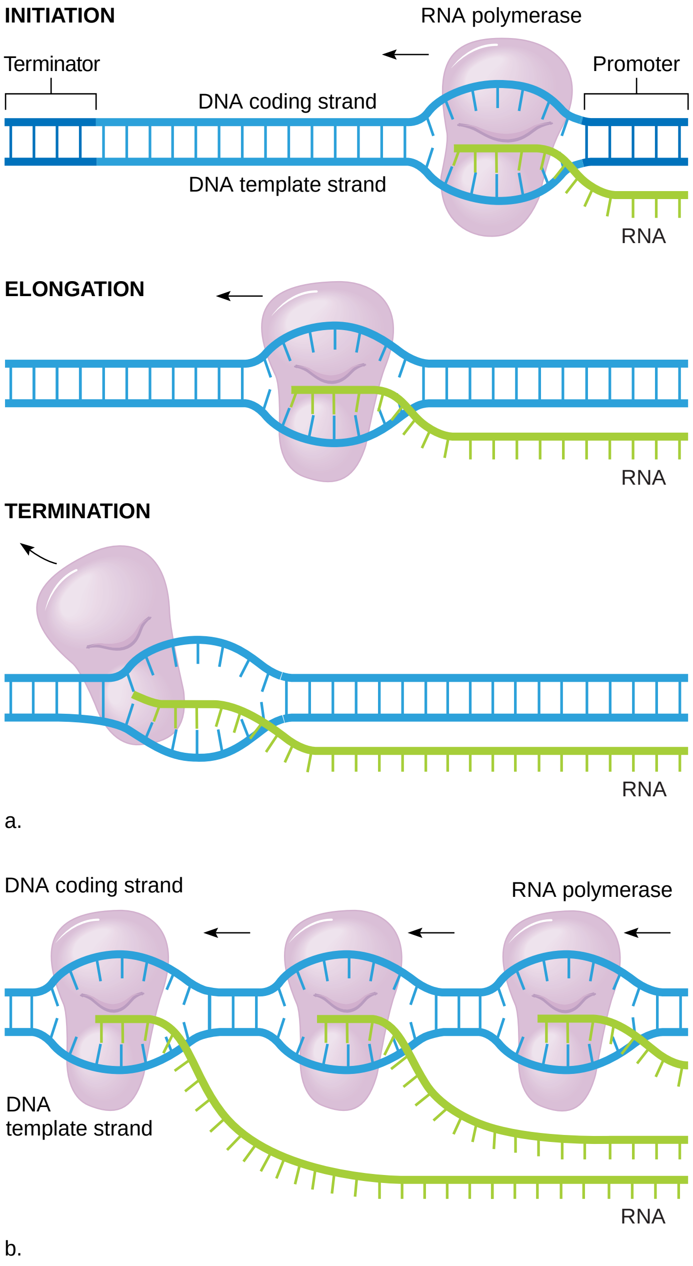


figure 10.7

Transcription of RNA from DNA. (*a*) Transcription occurs in three stages: initiation, elongation, and termination. Initiation is the control point that determines which genes are transcribed and when RNA nucleotides are added during elongation, and a terminator sequence in the gene signals the end of transcription. (*b*) Many identical copies of RNA are simultaneously transcribed, with one RNA polymerase starting after another.

RNA Processing

In bacteria and archaea, RNA is translated into protein as soon as it is transcribed from DNA because a nucleus does not physically separate the two processes. In cells of eukaryotes, mRNA must first exit the nucleus to enter the cytoplasm, where protein synthesis occurs. RNA is altered before it participates in protein synthesis in these more complex cells. (However, recent evidence indicates that some protein synthesis does occur in the nucleus.)

After mRNA is transcribed, a short sequence of modified nucleotides, called a cap, is added to the 5′ end of the molecule. At the 3′ end, a special polymerase adds 200 adenines, forming a "poly A tail." The cap and poly A tail may mark which mRNAs should exit the nucleus.

In addition to these modifications, not all of an mRNA is translated into an amino acid sequence in eukaryotic cells. Parts of mRNAs called **introns** (short for "intervening sequences") are transcribed but are later removed. The ends of the remaining molecule are spliced together before the mRNA is translated.

The mRNA prior to intron removal is called pre-mRNA. Introns are excised by small RNA molecules that are ribozymes, which associate with proteins to form small nuclear ribonucleoproteins (snRNPs), or "snurps." Four snurps form a structure called a spliceosome that cuts introns out and knits exons together to form the mature mRNA that exits the nucleus. The parts of mRNA that are translated are called **exons** (figure 10.8).

Introns range in size from 65 to 10,000 or more bases; the average intron is 3,365 bases. The average exon, in contrast, is 145

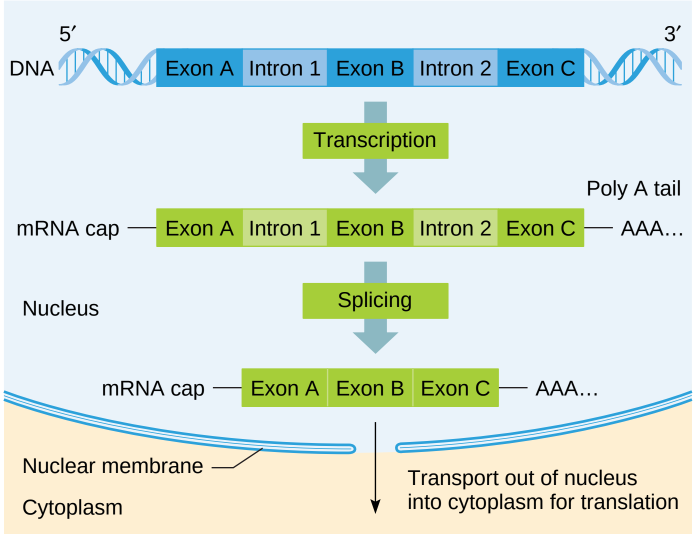


figure 10.8

Messenger RNA processing—the maturing of the message. Several steps carve the mature mRNA. First, a large region of DNA containing the gene is transcribed. Then a modified nucleotide cap and poly A tail are added, and introns are spliced out. Finally, the mature mRNA is transported out of the nucleus. Some mRNAs remain in the nucleus and are translated there—something that was only recently learned.

bases long. Many genes are riddled with introns—the human collagen gene, for example, contains 50. The gene whose absence causes Duchenne muscular dystrophy is especially interesting in its intron/exon organization. The gene is 2,500,000 bases, but its corresponding mRNA sequence is only 14,000 bases. The gene contains 80 introns. The number, size, and organization of introns vary from gene to gene. In many genes, introns take up more space than exons. We know from the human genome sequence that the coding portion of the average human gene is 1,340 bases, whereas the average total size of a gene is 27,000 bases.

The discovery of introns in 1977 arose out of then-new DNA sequencing technology. Certain gene sequences, when compared to the protein sequences that they encode, were found to have extra sections. The experiment that led to the detection of introns examined the 600-base rabbit beta globin gene, which encodes a 146 amino acid chain. Because three DNA bases encode one amino acid, a gene of only 438 bases would suffice to specify 146 amino acids ($3 \times 146 = 438$), but that is not the case. Since then, introns have been found in many genes, almost exclusively in eukaryotes.

The existence of introns surprised geneticists, who likened gene structure to a sentence in which all of the information contributes to the meaning. Why would protein-encoding genes consist of meaningful pieces scattered amidst apparent genetic gibberish? In the 1980s, geneticists confident that introns were aberrations arrogantly called them "junk DNA," Francis Crick among them. The persistent finding of introns, however, eventually convinced researchers that introns must have some function, or they would not have been retained through evolution. Said one speaker at a recent genomics conference, "Anyone who still thinks that introns have no function, please volunteer to have them removed, so we can see what they do." He had no takers.

The human genome project has revealed that the pervasiveness and size of our introns distinguishes our genome from those of our closest relatives. The human genome has many introns, with an average size 10 times that of an intron in the fruit fly or roundworm. The exon/intron organization that is a hallmark of many human genes can be compared to the signal/noise in a message. Our genome has more "noise" than do others so far sequenced, and we do not know why. Introns have complicated the "annotation" phase of the human genome project, which locates the protein-encoding parts. Finding the exons in a gene is a little like a word search puzzle, in which a square of letters harbors hidden words. Computer programs can hunt among strings of A, T, G, and C—raw DNA sequence—for the telltale signs of a protein-encoding gene, and then distinguish the exons from the introns. For example, a short sequence that indicates the start of a protein-encoding gene is called an open reading frame. Another clue is that dinucleotide repeats, such as CGCGCG, often flank introns, forming splice sites that signal the spliceosome where to cut and paste the DNA.

A quarter century after their discovery, introns remain somewhat of a mystery. We do not know why some genes have introns and some do not. Introns may be ancient genes that have lost their original function, or they may be remnants of the DNA of viruses that once infected the cell. Genes-in-pieces may be one way that our genome maximizes its informational content. For example, introns may enable exons to combine in different ways, even from different genes, much as a person can assemble many outfits from a few basic pieces of clothing. The fact that some disease-causing mutations disrupt intron/exon splice sites suggests that this cutting and pasting of gene parts is essential to health. For some genes, the mRNA is cut to different sizes in different tissues, which is called alternate splicing. The trimming of genes may explain how cells that make up different tissues use the same protein in slightly different ways. This is the case for apolipoprotein B (apo B). Recall from chapter 7 that apolipoproteins transport fats. In the small intestine, the mRNA that encodes apo B is short, and the protein binds and carries dietary fat. In the liver, however, the mRNA is not shortened, and the longer protein transports fats manufactured in the liver, which do not come from food.

Once introns have been spliced out, enzymes check the remaining RNA molecule for accuracy, much as enzymes proofread newly replicated DNA. Messenger RNAs that are too short or too long may be stopped from exiting the nucleus. A proofreading mechanism also monitors tRNAs, ensuring that the correct conformation takes shape.

Key Concepts

RNA differs from DNA in that it is single-stranded, contains uracil instead of thymine and ribose instead of deoxyribose, and has different functions. Messenger RNA transmits information to build proteins. Each three mRNA bases in a row forms a codon that specifies a particular amino acid. Ribosomal RNA and proteins form ribosomes, which physically support protein synthesis and help catalyze bonding between amino acids. Transfer RNAs connect particular mRNA codons to particular amino acids. • Bacterial operons are simple gene control systems. In more complex organisms, transcription factors control gene expression. • Transcription proceeds as RNA polymerase inserts complementary RNA bases opposite the template strand of DNA. • Messenger RNA gains a modified nucleotide cap and a poly A tail. Introns are transcribed and cut out, and exons are reattached. Introns are common, numerous, and large in human genes. Certain genes are transcribed into different-sized RNAs in different cell types.

10.2 Translating a Protein

Transcription copies the information encoded in a DNA base sequence into the complementary language of RNA. The next step is translating mRNA into the specified sequence of amino acids. Particular mRNA codons (three bases in a row) correspond to particular amino acids (figure 10.9). This correspondence between the chemical languages of mRNA and protein is the genetic code.

Francis Crick hypothesized that an "adaptor" would enable the RNA message to attract and link amino acids into proteins. He wrote in an unpublished paper in early 1955, "In its simplest form there would be 20 different kinds of adaptor molecule, one for each amino acid, and 20 different enzymes to join the amino acids to their adaptors." He was describing tRNAs, but the solution was more complex than originally thought. In the 1960s, many researchers deciphered the genetic code, determining which mRNA codons correspond to which amino acids. Marshall Nirenberg led the effort, for which he won the Nobel Prize.

The news media, on announcing the sequencing of the human genome in June 2000, widely reported that the "human genetic code" had been cracked. This was not the case. The genetic code is not unique to humans, and it was cracked decades ago. The code is the correspondence between nucleic acid triplet and amino acid, not the sequence itself.

Deciphering the Genetic Code

The researchers who deciphered the genetic code used a combination of logic and experiments. More recently, annotation of the human genome sequence has confirmed and extended the earlier work, revealing new nuances in genetic coding. Figure 10.9 summarizes the relationship of DNA, RNA, and protein. It is helpful in understanding how the genetic code works to ask the questions that researchers asked in the 1960s.

Question 1—How Many RNA Bases Specify One Amino Acid?

Because the number of different protein building blocks (20) exceeds the number of different mRNA building blocks (4), each codon must contain more than one mRNA base. In other words, if a codon consisted of only one mRNA base, then codons could specify only four different amino acids, with one corresponding to each of the four bases: A, C, G, and U (figure 10.10). If each codon

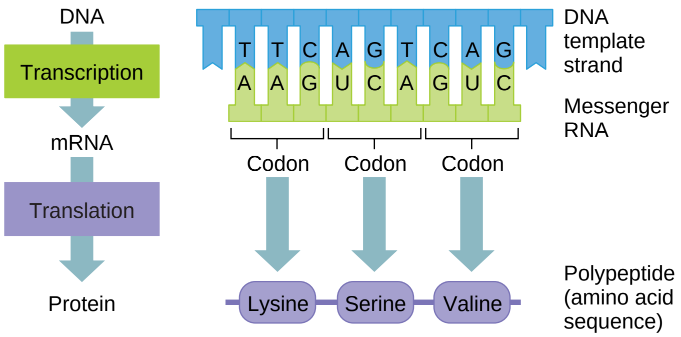


figure 10.9

From DNA to RNA to protein. Messenger RNA is transcribed from a locally unwound portion of DNA. In translation, transfer RNA matches up mRNA codons with amino acids.

Size of a genetic code word (codon) To encode a protein alphabet of 20 amino acids:	**Logic**
mRNA genetic code of 1 letter	
U C A G	4 combinations: **not sufficient**
mRNA genetic code of 2 letters	
UU CC AA GG UC CU UA AU UG GU CA AC CG GC AG GA	16 combinations: **not sufficient**
mRNA genetic code of 3 letters	
UUU UUC UUA UUG UCU UCC UCA UCG UAU UAC UAA UAG UGU UGC UGA UGG CUU CUC CUA CUG CCU CCC CCA CCG CAU CAC CAA CAG CGU CGC CGA CGG AUU AUC AUA AUG ACU ACC ACA ACG AAU AAC AAA AAG AGU AGC AGA AGG GUU GUC GUA GUG GCU GCC GCA GCG GAU GAC GAA GAG GGU GGC GGA GGG	64 combinations: **sufficient**

figure 10.10

Codon size. An exercise in logic reveals the triplet nature of the genetic code.

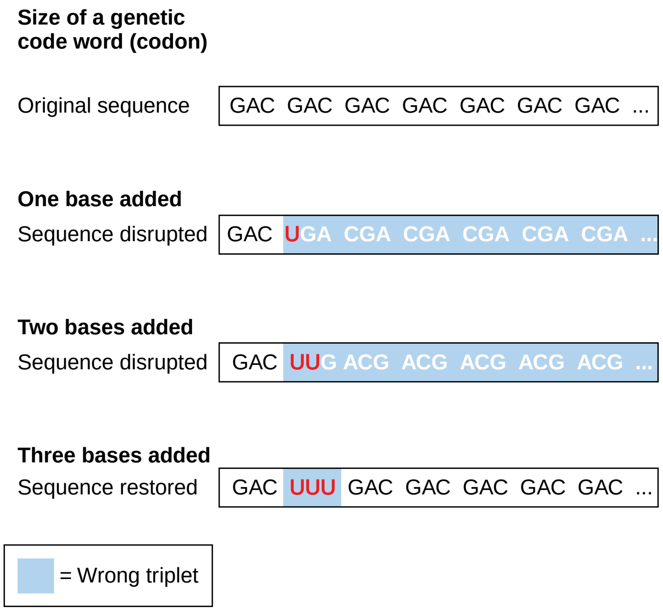
consisted of two bases, then 16 (4^2) different amino acids could be specified, one corresponding to each of the 16 possible orders of two RNA bases. This still is inadequate to encode the 20 amino acids found in organisms. If a codon consisted of three bases, then the genetic code could specify as many as 64 (4^3) different amino acids. Because 20 different amino acids require at least 20 different codons, the minimum number of bases in a codon is three.

Francis Crick and his coworkers conducted experiments on a type of virus called T4 that confirmed the triplet nature of the genetic code. They exposed the viruses to chemicals that add or remove one, two, or three bases, and examined a viral gene whose sequence and protein product were known. Altering the sequence by one or two bases produced a different amino acid sequence. The change disrupted the **reading frame,** which is the particular sequence of amino acids encoded from a certain starting point in a DNA sequence. However, adding or deleting three contiguous bases added or deleted only one amino acid in the protein product. This did not disrupt the reading frame. The rest of the amino acid sequence was retained. The code, the researchers deduced, is triplet (figure 10.11). To confirm the triplet nature of the genetic code, other experiments showed that if a base is added at one point in the gene, and a base deleted at another point, then the reading frame is disrupted only between these sites, resulting in a protein with a stretch of the wrong amino acids. In yet other experiments, if one base was inserted but two removed, the reading frame never returned to the specified normal amino acid sequence.

Question 2—Does a DNA Sequence Contain Information in an Overlapping Manner?

Consider the hypothetical mRNA sequence AUCAGUCUA. If the genetic code is triplet and a DNA sequence accessed in a nonoverlapping manner (that is, each three bases in a row form a codon, but any one base is part of only one codon), then this sequence contains only three codons: AUC, AGU, and CUA. If the DNA sequence is overlapping, the sequence contains seven codons: AUC, UCA, CAG, AGU, GUC, UCU, and CUA.

An overlapping DNA sequence seems economical in that it packs maximal information into a limited number of bases (figure 10.12). However, an overlapping sequence constrains protein structure because certain amino acids must always be followed by certain others. For example, the amino acid the first codon specifies, AUC, would always be followed by an amino acid whose codon begins with UC. Experiments that sequence proteins show that no specific type of amino acid always follows another. Any amino acid can follow any other amino acid in a protein's sequence. Therefore, the protein-encoding DNA sequence is not overlapping. There are a few exceptions,



Size of a genetic code word (codon)
Original sequence
GAC GAC GAC GAC GAC GAC GAC ...
One base added
Sequence disrupted
GAC UGA CGA CGA CGA CGA CGA ...
Two bases added
Sequence disrupted
GAC UUG ACG ACG ACG ACG ACG ...
Three bases added
Sequence restored
GAC UUU GAC GAC GAC GAC GAC ...
= Wrong triplet

figure 10.11

Three at a time. Adding or deleting one or two nucleotides to a DNA sequence disrupts the encoded amino acid sequence. However, adding or deleting three bases does not disrupt the reading frame. Therefore, the code is triplet. This is a simplified representation of the Crick experiment.

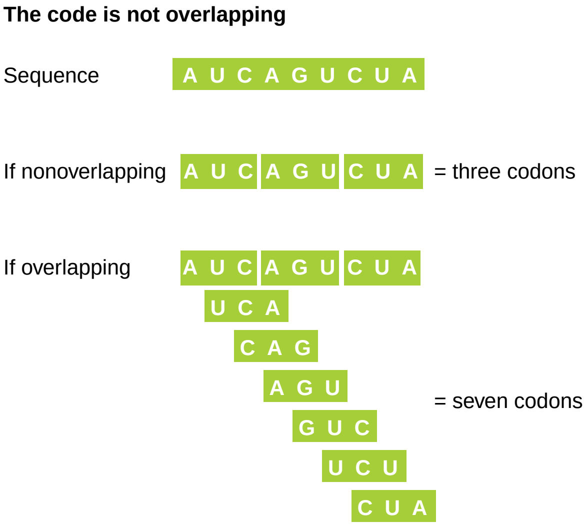


figure 10.12

The genetic code does not overlap. An overlapping genetic code may seem economical, but it is restrictive, dictating that certain amino acids must follow others in a protein's sequence. This does not happen; therefore, the genetic code is nonoverlapping.

particularly in viruses, where the same DNA sequence can be read from different starting points.

Question 3—Can mRNA Codons Signal Anything Other Than Amino Acids?

Chemical analysis eventually showed that the genetic code contains directions for starting and stopping translation. The codon AUG signals "start," and the codons UGA, UAA, and UAG each signify "stop." Another form of "punctuation" is a short sequence of bases at the start of each mRNA, called the leader sequence, that enables the mRNA to hydrogen bond with rRNA in a ribosome.

Question 4—Do All Species Use the Same Genetic Code?

All species use the same mRNA codons to specify the same amino acids, despite the popular idea of a "human" genetic code. This universality of the genetic code is part of the abundant evidence that all life on earth evolved from a common ancestor. No other mechanism as efficient at directing cellular activities has emerged and persisted. The only known exceptions to the universality of the genetic code are a few codons in the mitochondria of certain single-celled organisms. The ability of mRNA from one species to be translated in a cell of another species has made recombinant DNA technology possible, in which bacteria manufacture proteins normally made in the human body. Chapter 18 explains how such proteins are used as drugs.

Question 5—Which Codons Specify Which Amino Acids?

In 1961, Marshall Nirenberg and his coworkers at the National Institute of Health began deciphering which codons specify which amino acids, using a precise and logical series of experiments. First they synthesized mRNA molecules in the laboratory. Then they added them to test tubes that contained all the chemicals and structures needed for translation, which they had extracted from *E. coli* cells. Which amino acid would each synthetic RNA specify?

The first synthetic mRNA tested had the sequence UUUUUU. . . . In the test tube, this was translated into a peptide consisting entirely of one amino acid type: phenylalanine. Thus was revealed the first entry in the genetic code dictionary: the codon UUU specifies the amino acid phenylalanine. The number of phenylalanines always equaled one-third the number of mRNA bases, confirming that the genetic code is triplet and nonoverlapping. The next three experiments revealed that AAA codes for the amino acid lysine, GGG for glycine, and CCC for proline.

Next, the researchers synthesized chains of alternating bases. Synthetic mRNA of the sequence AUAUAU . . . introduced codons AUA and UAU. When translated, the mRNA yielded an amino acid sequence of alternating isoleucines and tyrosines. But was AUA the code for isoleucine and UAU for tyrosine, or vice versa? Another experiment answered the question.

An mRNA of sequence UUUAUAUUUAUA encoded alternating phenylalanine and isoleucine. Because the first experiment showed that UUU codes for phenylalanine, the researchers deduced that AUA must code for isoleucine. If AUA codes for isoleucine they reasoned, looking back at the previous experiment, then UAU must code for tyrosine. Table 10.3 summarizes some of these experiments.

By the end of the 1960s, researchers had deciphered the entire genetic code (table 10.4). Sixty of the possible 64 codons were found to specify particular amino acids, while the others indicate "stop" or "start." This means that some amino acids are specified by more than one codon. For example, both UUU and UUC encode phenylalanine. Different codons that specify the same amino acid are called synonymous codons, just as synonyms are words with the same meaning. The genetic code is said to be **degenerate** because each amino acid is not uniquely specified. Synonymous codons often differ from one another by the base in the third position. The corresponding base of a tRNA's anticodon is called the "wobble" position because it can bind to more than one type of base in synonymous codons. The degeneracy of the genetic code provides protection against mutation, because changes in the DNA that cause the substitution of a synonymous codon would not affect the protein's amino acid sequence.

table 10.3

Deciphering RNA Codons and the Amino Acids They Specify

Synthetic RNA	Encoded Amino Acid Chain	Puzzle Piece
UUUUUUUUUUUUUUUUUU	Phe-Phe-Phe-Phe-Phe-Phe	UUU = Phe
AAAAAAAAAAAAAAAAAA	Lys-Lys-Lys-Lys-Lys-Lys	AAA = Lys
GGGGGGGGGGGGGGGGGG	Gly-Gly-Gly-Gly-Gly-Gly	GGG = Gly
CCCCCCCCCCCCCCCCCC	Pro-Pro-Pro-Pro-Pro-Pro	CCC = Pro

AUAUAUAUAUAUAUAUAU	Ile-Tyr-Ile-Tyr-Ile-Tyr	AUA = Ile or Tyr
		UAU = Ile or Tyr

UUUAUAUUUAUAUUUAUA	Phe-Ile-Phe-Ile-Phe-Ile	AUA = Ile
		UAU = Tyr

table 10.4

The Genetic Code

	Second Letter								
First Letter	U		C		A		G		Third Letter
U	UUU	Phenylalanine (Phe)	UCU	Serine (Ser)	UAU	Tyrosine (Tyr)	UGU	Cysteine (Cys)	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine (Leu)	UCA		UAA	“stop”	UGA	“stop”	A
	UUG		UCG		UAG	“stop”	UGG	Tryptophan (Trp)	G
C	CUU	Leucine (Leu)	CCU	Proline (Pro)	CAU	Histidine (His)	CGU	Arginine (Arg)	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine (Gln)	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine (Ile)	ACU	Threonine (Thr)	AAU	Asparagine (Asn)	AGU	Serine (Ser)	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine (Lys)	AGA	Arginine (Arg)	A
	AUG	Methionine (Met) and “start”	ACG		AAG		AGG		G
G	GUU	Valine (Val)	GCU	Alanine (Ala)	GAU	Aspartic acid (Asp)	GGU	Glycine (Gly)	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic acid (Glu)	GGA		A
	GUG		GCG		GAG		GGG		G

The human genome project picked up where the genetic code experiments of the 1960s left off by identifying the DNA sequences that are transcribed into tRNAs. That is, 61 different tRNAs could theoretically exist, one for each codon that specifies an amino acid (the 64 triplets minus 3 stop codons). However, only 49 different genes were found to encode tRNAs. This is because the same type of tRNA can detect synonymous codons that differ only in whether the wobble (third) position is U or C. The same type of tRNA, for example, binds to both UUU and UUC codons, which specify the amino acid phenylalanine. Synonymous codons ending in A or G use different tRNAs.

The monumental task of deciphering the genetic code was, in an intellectual sense, even more important than today’s sequencing of genomes, for it revealed the “rules” that essentially govern life at the cellular level. As with genome projects, many research groups contributed to the effort of solving the genetic code problem. Because genetics was still a very young science, the code breakers came largely from the ranks of chemistry, physics, and math.

Some of the more exuberant personalities organized an “RNA tie club” and inducted a new member whenever someone added a new piece to the puzzle of the genetic code, anointing him (there were no prominent hers) with a tie and tie pin emblazoned with the structure of the specified amino acid (figure 10.13). By the end of the 1960s, researchers had deciphered the entire genetic code.

Building a Protein

Protein synthesis requires mRNA, tRNA molecules carrying amino acids, ribosomes, energy-storing molecules such as adenosine triphosphate (ATP), and various protein factors. These pieces come together at the

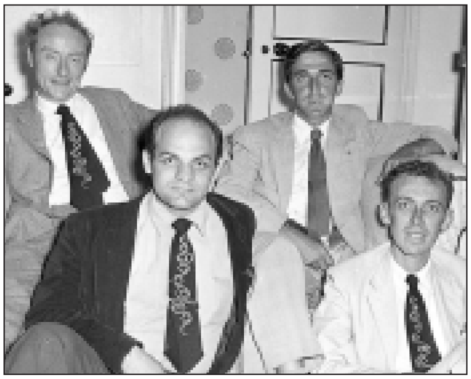


figure 10.13

The RNA tie club. In 1953, physicist-turned-biologist George Gamow started the RNA tie club, to “solve the riddle of RNA structure and to understand the way it builds proteins.” The club would have 20 members, one for each amino acid. Each honored member received a tie and tie pin labeled with the name of the particular amino acid he had worked on. Francis Crick (upper left) was tyrosine; James Watson (lower right) was proline.

beginning of translation in a stage called **translation initiation** (figure 10.14).

First, the mRNA leader sequence hydrogen bonds with a short sequence of rRNA in a small ribosomal subunit. The first mRNA codon to specify an amino acid is always AUG, which attracts an initiator tRNA that carries the amino acid methionine (abbreviated *met*). This methionine signifies the start of a polypeptide. The small ribosomal subunit, the mRNA bonded to it, and the initiator tRNA with its attached methionine form the **initiation complex.**

To start the next stage of translation, called **elongation,** a large ribosomal subunit attaches to the initiation complex. The codon adjacent to the initiation codon (AUG), which is GGA in figure 10.15*a*, then bonds to its complementary anticodon, which is part of a free tRNA that carries the amino acid glycine. The two amino acids (*met* and *gly* in the example), which are still attached to their tRNAs, align.

The part of the ribosome that holds the mRNA and tRNAs together can be described as having two sites. The positions of the sites on the ribosome remain the same with respect to each other as translation proceeds, but they cover different parts of the mRNA as the ribosome moves. The **P site** holds the growing amino acid chain, and the **A site** right next to it holds the next amino acid to be added to the chain. In figure 10.15, when the protein-to-be consists of only the first two amino acids, *met* occupies the P site and *gly* the A site.

With the help of rRNA that functions as a ribozyme, the amino acids link by an attachment called a peptide bond. Then the first tRNA is released. It will pick up another amino acid and be used again. The ribosome and its attached mRNA are now bound to a single tRNA, with two amino acids extending from it at the P site. This is the start of a polypeptide.

Next, the ribosome moves down the mRNA by one codon. The region of the mRNA that was at the A site is thus now at the P site. A third tRNA enters, carrying its amino acid (*cys* in figure 10.15*b*). This third amino acid aligns with the other two and forms a peptide bond to the second amino acid in the growing chain, now extending from the P site. The tRNA attached to the second amino acid is released and recycled. The polypeptide continues to build, one amino acid at a time. Each piece is brought in by a tRNA whose anticodon corresponds to a consecutive mRNA codon as the ribosome moves down the mRNA (figure 10.15*c*).

Elongation halts when one of the mRNA “stop” codons (UGA, UAG, or UAA) is reached, because no tRNA molecules correspond to these codons. The last tRNA leaves the ribosome, the ribosomal subunits separate from each other and are recycled, and the new polypeptide is released.

Protein synthesis is economical. A cell can produce large amounts of a particular protein from just one or two copies of a gene. A plasma cell in the immune system, for

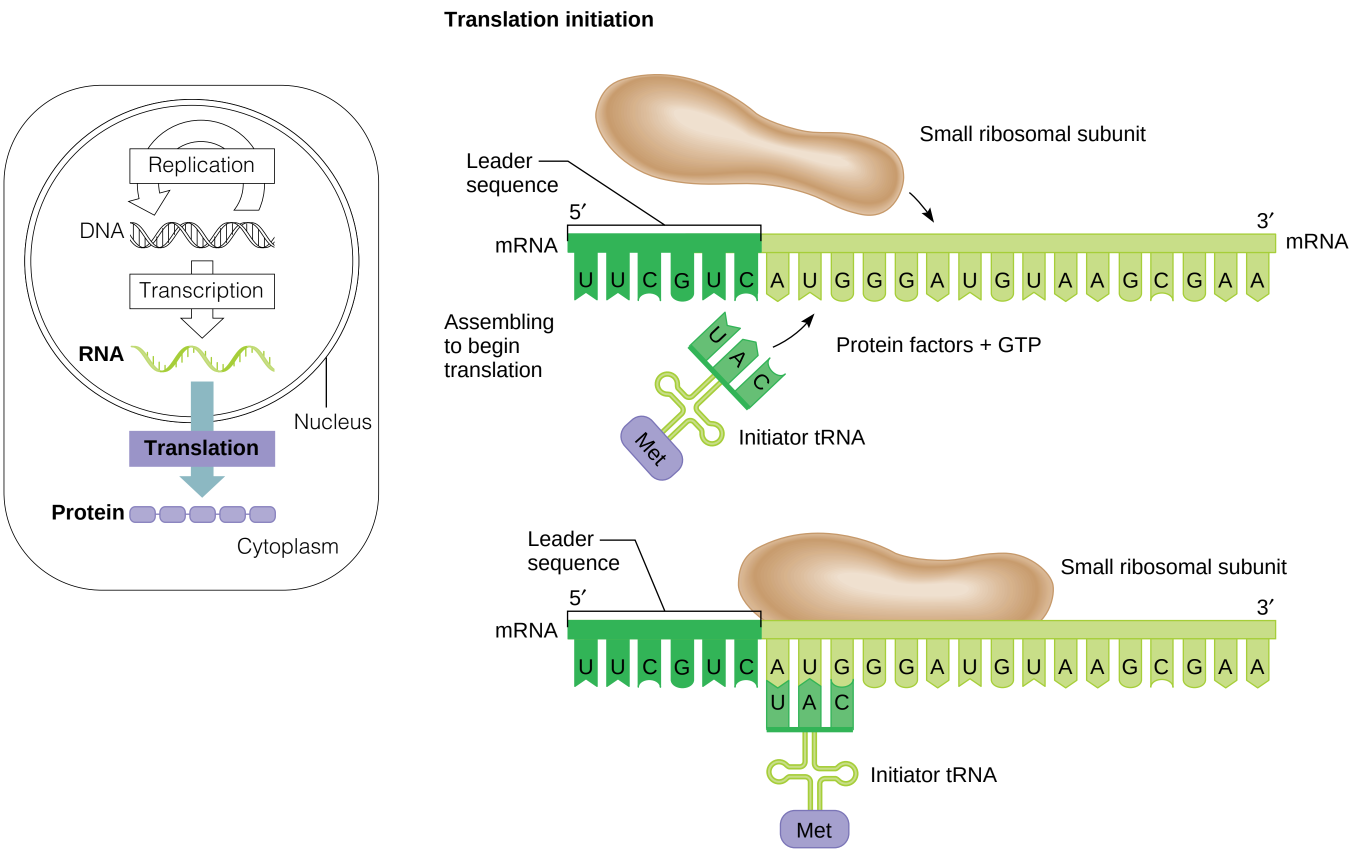
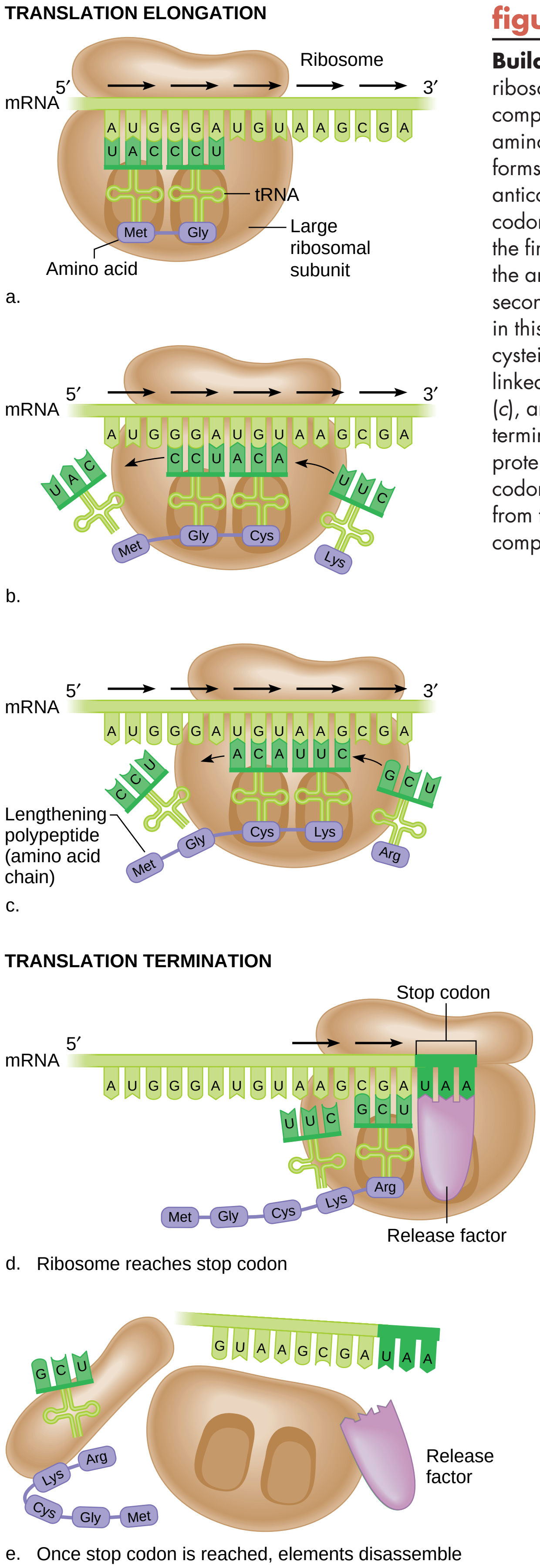


figure 10.14

Translation begins. Initiation of translation brings together a small ribosomal subunit, mRNA, and an initiator tRNA, and aligns them in the proper orientation to begin translation.



d. Ribosome reaches stop codon

e. Once stop codon is reached, elements disassemble

figure 10.15

Building a polypeptide. A large ribosomal subunit binds to the initiation complex, and a tRNA bearing a second amino acid (glycine, in this example) forms hydrogen bonds between its anticodon and the mRNA's second codon (*a*). The methionine brought in by the first tRNA forms a peptide bond with the amino acid brought in by the second tRNA, and a third tRNA arrives, in this example carrying the amino acid cysteine (*b*). A fourth amino acid is linked to the growing polypeptide chain (*c*), and the process continues until a termination codon is reached. (*d*) A protein release factor binds to the stop codon, releasing the completed protein from the tRNA and (*e*) freeing all of the components of the translation machine.

example, manufactures 2,000 identical antibody molecules per second. To mass-produce proteins on this scale, RNA, ribosomes, enzymes, and other proteins must be continually recycled. Transcription always produces multiple copies of a particular mRNA, and each mRNA may be bound to dozens of ribosomes, as figure 10.16 shows. As soon as one ribosome has moved far enough along the mRNA, another ribosome will attach. In this way, many copies of the encoded protein will be made from the same mRNA.

Protein Folding

As a protein is synthesized, it folds into a three-dimensional shape (conformation) that helps determine its function. This folding occurs because of attractions and repulsions between the protein's atoms. In addition,

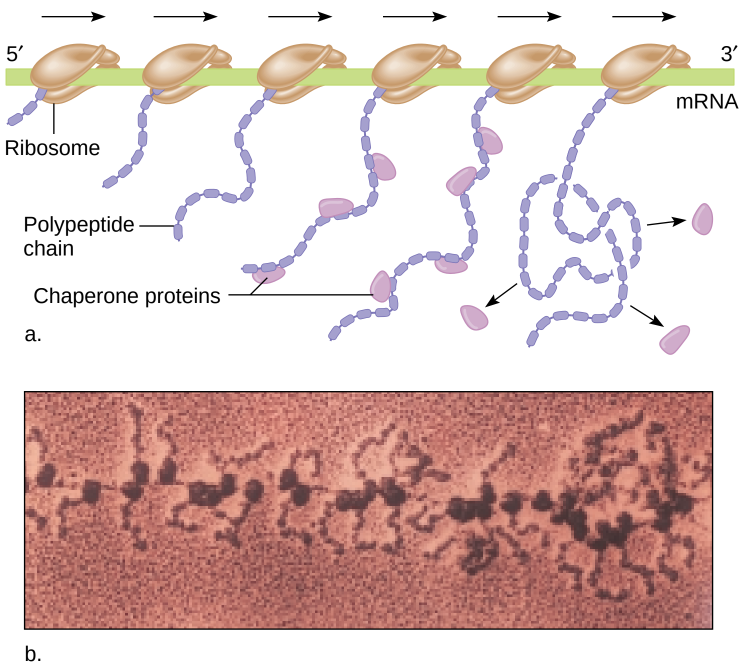


figure 10.16

Making multiple copies of a protein. Several ribosomes can translate the same protein from a single mRNA at the same time. (*a*) The ribosomes have different-sized polypeptides dangling from them—the closer a ribosome is to the end of a gene, the longer its polypeptide. Chaperone proteins help fold the polypeptide into its characteristic conformation. (*b*) In the micrograph, the ribosomes on the left have just begun translation and the polypeptides are short. Further along in translation, the polypeptides are longer. The chaperones are not visible.

thousands of water molecules surround a growing chain of amino acids, and, because some amino acids are attracted to water and some repelled by it, the water contorts the protein's shape. Sulfur atoms also affect overall conformation by bridging the two types of amino acids that contain them.

The conformation of a protein may be described at several levels. Figure 10.17 shows the four levels for hemoglobin, which carries oxygen in the blood. The amino acid sequence of a polypeptide chain determines its **primary (1°) structure.** Chemical attractions between amino acids that are close together in the 1° structure fold the polypeptide chain into its **secondary (2°) structure,** which may take the distinctive forms of loops, coils, barrels, helices, or sheets. Secondary structures wind into larger **tertiary (3°) structures** as more widely separated amino acids attract or repel in response to water molecules. Finally, proteins consisting of more than one polypeptide form a **quaternary (4°) structure.** Hemoglobin has four polypeptide chains. The liver protein ferritin has 20 identical polypeptides of 200 amino acids each. In contrast, the muscle protein myoglobin is a single polypeptide chain.

For many years, biochemists thought that protein folding was straightforward; the amino acid sequence dictated specific attractions and repulsions between parts of a protein, contorting it into its final form as it emerged from the ribosome. But these attractions and repulsions are not sufficient to fold the polypeptide into the highly specific form essential to its function. A protein apparently needs help to fold correctly.

An amino acid chain may start to fold as it emerges from the ribosome. Localized regions of shape form, and possibly break apart and form again, as translation proceeds. Experiments that isolate proteins as they are synthesized show that other proteins oversee the process of proper folding. These accessory proteins include enzymes that foster chemical bonds and chaperone proteins, which stabilize partially folded regions that are important to the molecule's final form.

Just as repair enzymes check newly replicated DNA for errors and RNAs are proofread, proteins scrutinize a folding protein to detect and dismantle incorrectly folded regions. Errors in protein folding can cause illness. Some mutations that cause cystic fibrosis, for example, prevent the encoded protein from assuming its final form and anchoring in the cell membrane, where it normally controls the flow of chloride ions. One type of Alzheimer disease is associated with a protein called amyloid that forms an abnormal, gummy mass instead of remaining as distinct molecules, because of improper folding.

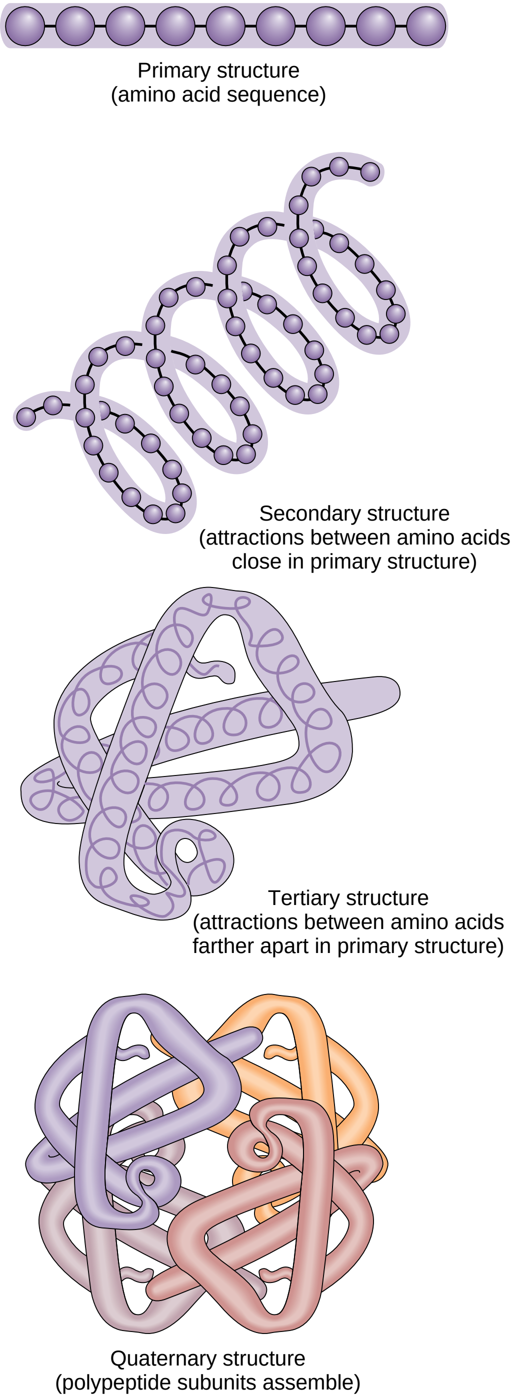

Primary structure
(amino acid sequence)
Secondary structure
(attractions between amino acids
close in primary structure)
Tertiary structure
(attractions between amino acids
farther apart in primary structure)
Quaternary structure
(polypeptide subunits assemble)

figure 10.17

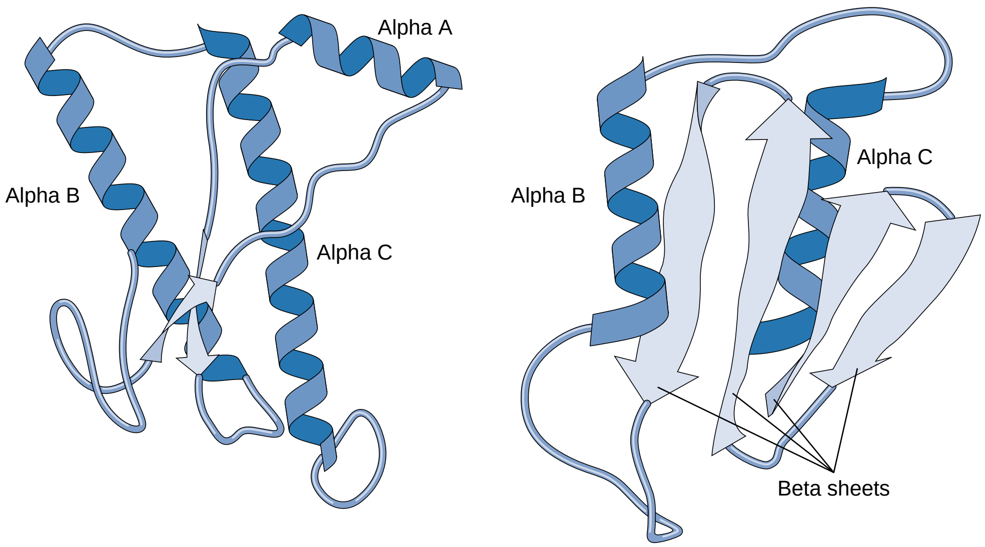
Protein conformation. The hemoglobin molecule consists of four polypeptides, called globins. Each globin surrounds a smaller organic compound that holds an iron atom. Chapter 11 revisits the hemoglobin molecule.

Some members of a class of inherited disorders called triplet repeats tack extra glutamines onto particular proteins. The extra amino acids alter the ability of the protein to fold into its characteristic conformation. Most of the triplet repeat disorders—so-called because extra DNA triplets encode the extra amino acids—affect the brain. They are discussed further in chapter 11.

Yet another type of disorder that arises from an abnormal protein conformation is the spongiform encephalopathies, such as “mad cow disease” and similar conditions in sheep, humans, and several other types of mammals. Recall from chapter 2 that these disorders are caused by abnormal aggregation of proteins called prions. Unlike other proteins that misfold to cause disease, the normal and abnormal forms of prion protein have the same primary structure, but they are capable of folding into at least eight three-dimensional shapes (figure 10.18). Just as researchers assumed that genes were continuous, so too did they assume incorrectly that a protein can fold into just one conformation.

In addition to folding, certain proteins must be altered further before they become functional. Sometimes enzymes must shorten a polypeptide chain for it to become active. Insulin, which is 51 amino acids long, for example, is initially translated as the polypeptide proinsulin, which is 80 amino acids long. Some proteins must have sugars attached for them to become functional.

The linguistic nature of the flow of genetic information makes it ideal for computer analysis. The view of DNA sequences as a language emerged in the 1960s, as experiments revealed the linear relationship between nucleic acid sequences in genes and amino acid sequences in proteins. Yet the “rules” by which DNA sequences specify protein shapes are still not well understood, even as we routinely decipher the sequences of entire genomes.



a. Cellular prion protein (noninfectious)

b. Scrapie prion protein (infectious)

figure 10.18

One protein, multiple conformations. Biochemists once thought that the primary structure of a protein dictated one conformation. This is not true. The cellular form of prion protein, for example, does not cause disease (*a*). The scrapie form is infectious—it converts the cellular form to more of itself (*b*). Infectious prions cause scrapie in sheep, bovine spongiform encephalopathy in cows, and variant Creutzfeldt-Jakob disease in humans. Recent work showing that myoglobin also assumes different forms suggests that there is much that we do not know about protein conformation.

KEY CONCEPTS

The genetic code is triplet, nonoverlapping, continuous, universal, and degenerate. • As translation begins, mRNA, tRNA with bound amino acids, ribosomes, energy molecules, and protein factors assemble. The mRNA binds to rRNA in the small subunit of a ribosome, and the first codon attracts a tRNA bearing methionine. Next, as the chain elongates, the large ribosomal subunit attaches and the appropriate anticodon parts of tRNAs bind to successive codons in the mRNA. As the amino acids attached to the aligned tRNA molecules form peptide bonds, a polypeptide grows. The ribosome moves down the mRNA to the region that holds the amino acid chain, called the P site, and the area where a new tRNA binds, called the A site. When the ribosome reaches a “stop” codon, protein synthesis ceases. RNA, ribosomes, enzymes, and key proteins are recycled. • Protein folding begins as translation proceeds, with enzymes and chaperone proteins assisting. A protein can fold in more than one way.

10.3 The Human Genome Sequence Reveals Unexpected Complexity

For nearly half a century after Watson and Crick deduced the structure of DNA, a view of the genome as a set number of genes that specify an equal number of proteins ruled. Even the finding that many genes are split into coding (exon) and noncoding (intron) regions did little, at first, to shake the one gene–one protein way of thinking. All that has changed with the sequencing of the human genome.

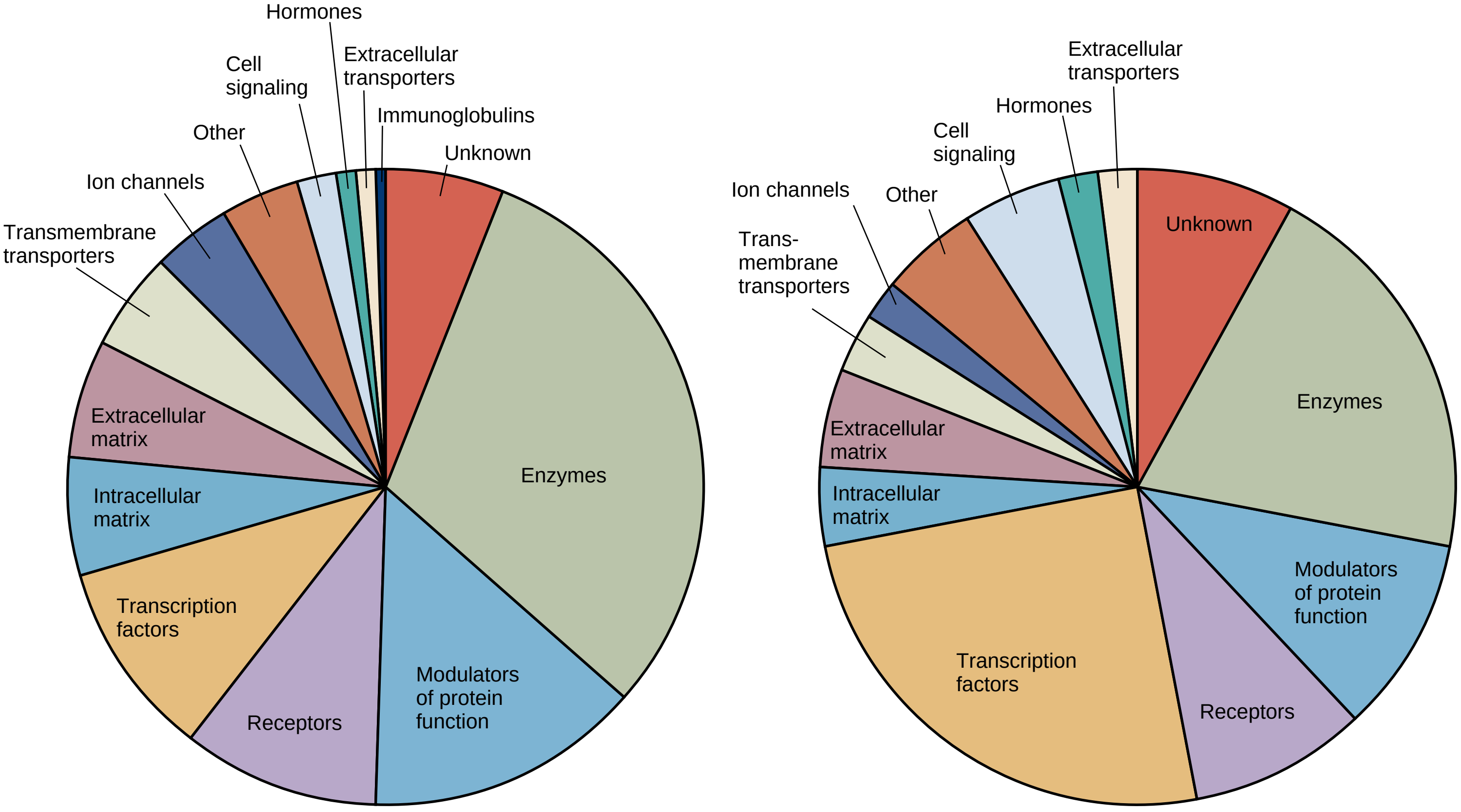
Until intense genome sequencing efforts began in the 1990s, most researchers focused on mapping, identifying, and discovering the functions of individual genes. With the genome sequence in hand, researchers can now estimate the number of protein-encoding genes and categorize proteins by function. **Proteomics** is the study of the entire collection of proteins produced in a particular cell. Charts such as figure 10.19 are being compiled for all cell types, at all stages of development, and under different conditions, to catalog the functioning of the human body at the molecular and cellular levels, in health and disease. But proteomics is just the start of genome analysis. Learning the protein-producing capabilities of cells seems simple compared to other information revealed in the genome project. Consider these facts:

- The human genome consists of approximately 3.2 billion DNA base pairs.
- Only about 1.5 percent of the human genome sequence encodes protein.
- The human genome includes approximately 31,000 protein-encoding genes.
- Human cells, considered together, can produce from 100,000 to 200,000 different proteins.

These statistics just do not add up in a one gene–one protein paradigm. Two central questions emerge, and they are the focus of this final section of the chapter: (1) How can 31,000 genes encode 100,000 to 200,000 proteins? Even if the number of protein-encoding genes exceeds 31,000, they are still fewer than the number of proteins. (2) What does the other 98.5 percent of the human genome—the part that doesn't encode protein—do?

Genome Economy: Reconciling Gene and Protein Number

The discovery of introns in 1977 first planted the idea that a number of genes could specify a larger number of proteins by mixing and matching gene parts. Research since then has revealed that about a third of the protein-encoding portion of the



a. Distribution of health-related proteins from conception through old age.

b. Distribution of health-related proteins from conception to birth.

figure 10.19

Proteomics meets medicine. One way to analyze the effects of genes is to categorize them by the functions of their protein products, and then to chart the relative abundance of each class at different stages of development or life, and in sickness and in health. The pie chart in (*a*) considers 14 categories of proteins that when abnormal or missing cause disease, and their relative abundance from conception through advanced age. The pie chart in (*b*) displays the same protein categories for the prenatal period, from conception to birth. Note, for example, that transcription factor genes are more highly expressed in the embryo and fetus, presumably because of the extensive cell differentiation that is a hallmark of this period. The relative expression of genes that encode enzymes is slightly less in the prenatal period than at other times because before birth, some metabolic needs are met by the pregnant woman. These depictions represent just one of the many new ways of looking at gene action.

genome—at least 10,000 genes—mix and match exons, each of which encodes a segment of a protein called a domain. For example, the DNA that encodes a blood-clotting protein called tissue plasminogen activator (t-PA) includes sequences from genes that encode three other proteins (plasminogen, epidermal growth factor, and fibronectin)—that's four proteins from three genes. This process of combining exons is called **exon shuffling.** Figure 10.20 illustrates schematically how two genes can give rise to seven proteins. Sequencing of extensive regions of chromosomes confirms the disparity between gene and protein number that researchers first discovered in considering the exon/intron structures of genes one at a time. For example, in a large part of chromosome 22, the first autosome to be sequenced, 245 genes are associated with 642 mRNA transcripts.

Introns may seem wasteful, little more than vast stretches of DNA bases that outnumber and outsize exons. But researchers are discovering that a DNA sequence that is an intron in one context may encode protein in another. Consider prostate specific antigen (PSA), a protein found on certain cell surfaces that is overproduced in some cases of prostate cancer (figure 10.21*a*). The gene for PSA has five exons and four introns, but it also encodes a second, different protein, called PSA-linked molecule (PSA-LM). Both genes have the same beginning DNA sequence, but the remainder of the PSA-LM gene is part of the fourth intron of the PSA sequence! The proteins seem to have antagonistic functions. That is, when the level of one is high, the other is low. Future blood tests to detect elevated risk of prostate cancer will likely consider levels of both proteins.

In another situation where introns may account for the overabundance of proteins compared to genes, a DNA sequence that is an intron in one gene's template strand may encode protein on the coding strand. That is, what is the template strand for one gene may be the coding strand for the other. This is the case for the gene that specifies neurofibromin, which when mutant causes neurofibromatosis. (This is an autosomal dominant condition that causes benign tumors beneath the skin and "café au lait" spots on the skin.) Encoded within an intron of the neurofibromin gene, but on the coding strand, are instructions for three other genes (figure 10.21*b*). It's even possible for a single RNA to be patched together from instructions on both strands, an apparently rare occurrence called trans-splicing. A gene in the fruit fly that controls chromosome structure in early development is transcribed from four exons on one strand and from two exons in the opposite orientation on the other strand. A small region of overlap allows the two mRNAs to complementary base pair. Then, a kind of cut-and-paste operation links all of the transcribed exons into a single mRNA molecule. Trans-splicing has not been identified in the human genome yet, but probably exists.

Still another way that a gene can maximize its informational content is for its encoded protein to be cut to yield two products. An inherited disorder called dentinogenesis imperfecta revealed this mechanism (figure 10.21*c*). The condition causes discolored, misshapen teeth with peeling enamel, due to abnormal dentin, which is the bone-like substance beneath the enamel that forms the bulk of the tooth. Dentin is a complex mixture of extracellular matrix proteins. Ninety percent of dentin protein is collagen, and most of the rest are proteins also found in bone. However, two proteins are unique to dentin: dentin phosphoprotein (DPP) and dentin sialoprotein (DSP). The single gene that encodes these two proteins is part of an area of chromosome 4 that seems to be devoted to teeth.

It was abnormal DPP that had been associated with dentinogenesis imperfecta. However, DPP, because it is abundant, accounting for 50 percent of the noncollagen protein in dentin compared to 6 percent for DSP, may have overshadowed a contributory role for DSP as well. Both proteins are translated from a single mRNA molecule as the precursor protein dentin sialophosphoprotein (DSPP), which is then cut to yield DPP and DSP. The DPP may be much more abundant because it is longer-lived than DSP. That is, DSP is degraded faster. These two proteins remain somewhat of a mystery. Often when genes with similar functions are right next to each other on a chromosome, as they are, it is because they arise from gene duplication. The two genes are then very similar in sequence. This is not the case with the DNA sequences that encode DPP and DSP.

Exon shuffling is probably fairly common—researchers had predicted its existence since shortly after the discovery of introns. Less well known and studied are genes in introns, trans-splicing, and two proteins encoded on one gene. Still, software can search for evidence of these ways to maximize genome information.

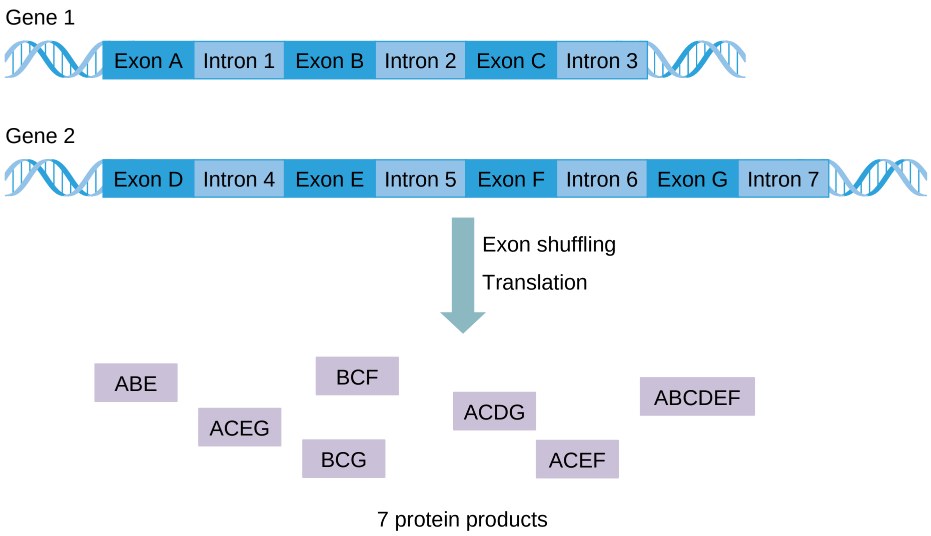


figure 10.20

Exon shuffling expands gene number. This schematic illustration shows how two genes can encode seven different proteins. Considering that many genes have many more introns than this one, it's clear that exon shuffling can generate many different proteins.

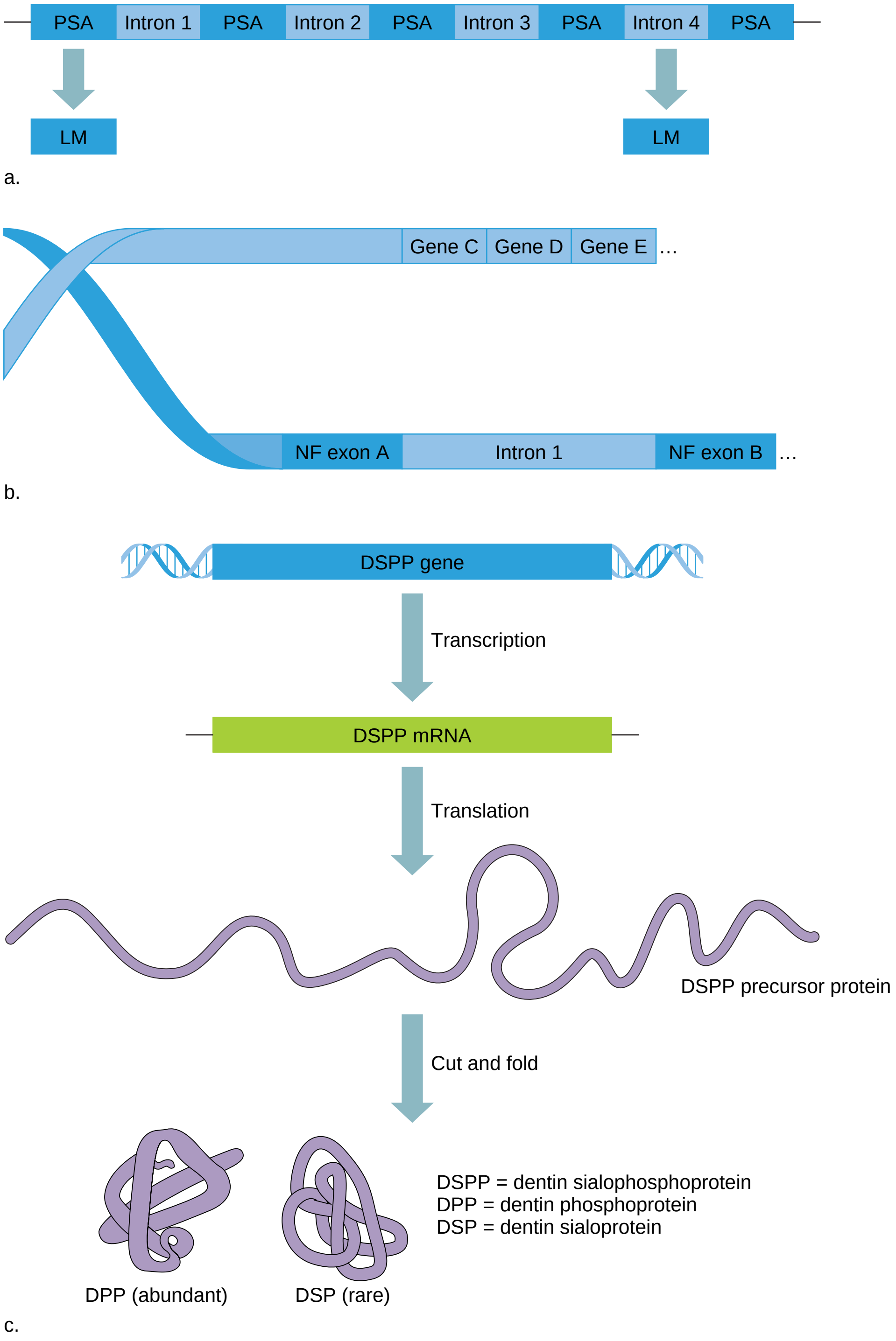


figure 10.21

Genome economy occurs in several ways. (*a*) Embedded in the PSA gene are really two protein-encoded sequences—the PSA portion consists of five exons. The PSA-LM part consists of two exons, one of which lies within an intron of PSA. (*b*) An intron of the neurofibromin gene harbors three genes on the opposite strand. (*c*) The dentin sialophosphoprotein (DSPP) gene encodes a long protein that is cleaved to yield dentin phosphoprotein (DPP) and dentin sialoprotein (DSP). The observation that DPP is more abundant than DSP indicates that their rate of degradation differs, because their rate of synthesis is presumably the same, since they come from the same transcript.

What Does the Other 98.5 Percent of the Human Genome Do?

The second quandary revealed in the human genome sequence is the role of the "other" 98.5 percent of the DNA bases. In general, this noncoding DNA falls into four categories: (1) RNAs other than mRNA (called noncoding or ncRNAs), (2) introns, (3) promoters and other control sequences, and (4) repeated sequences. Table 10.5 summarizes some of the functions of DNA other than encoding protein.

Noncoding RNAs

About a third of the human genome is transcribed into RNA types other than mRNA, the noncoding RNAs. The two best-studied ncRNAs are already familiar—tRNA and rRNA. The first draft sequence of the human genome identified 497 types of tRNA genes. The rate of transcription of a cell's tRNA genes seems to be attuned to the specific proteins that the cell manufactures—its proteome. This thrifty expression of genes is a little like the operons that enable bacteria to "sense" when enzymes are needed to dismantle certain nutrients. Human tRNA genes are dispersed among the chromosomes in clusters—25 percent of them are in a 4-million-base (4 megabase) region on chromosome 6. Altogether they account for 0.1 percent of the genome. Our 497 types of tRNA genes may seem like a lot, but frogs have thousands! This may reflect the fact that frog eggs are huge and contain may types of proteins. The 243 types of rRNA genes are clustered on six chromosomes, each cluster harboring 150 to 200 copies of a 44,000 base repeat sequence. Once transcribed from these clustered genes, the rRNAs go to the nucleolus, where they are cut to their final forms by yet another type of ncRNA called small nucleolar RNAs (snoRNAs).

Hundreds of thousands of ncRNAs are neither tRNA nor rRNA, nor snoRNAs, nor the other less-abundant types described in table 10.5. Instead, they are transcribed from DNA sequences called **pseudogenes.** A pseudogene is very similar in sequence to a particular protein-encoding gene, and it may be transcribed into RNA but it is not translated into protein. Pseudogenes may be

table 10.5

The Nonprotein Encoding Parts of the Human Genome

	Function or Characteristic
Noncoding RNA genes	
tRNA genes	Connect mRNA codon to amino acid
rRNA genes	Parts of ribosomes
Pseudogenes	DNA sequences that are very similar to known gene sequences and may be transcribed but are not translated
Small nucleolar RNAs	Process rRNA in nucleolus
Small nuclear RNAs	Parts of spliceosomes
Telomerase RNA	Part of ribonucleoprotein that adds bases to chromosome tips
Xist RNA	Inactivates one X chromosome in cells of females
Vault RNA	Part of "vault," a large ribonucleoprotein complex of unknown function
Introns	Parts of protein-encoding genes that are transcribed but cut out before the encoded protein is translated
Promoters and other control sequences	Guide enzymes that carry out DNA replication, transcription, or translation
Repeats	
Transposons	Repeats that move around the genome
Telomeres	Chromosome tips whose lengths control the cell cycle
Centromeres	Provide backdrop for proteins that form attachments for spindle fibers
Duplications of 10 to 300 kilobases	Unknown
Simple short repeats	Unknown

remnants of genes past, once-functional variants that diverged from the normal sequence too greatly to encode a working protein. Pseudogenes are incredibly common in the human genome. For example, at least 324 pseudogenes shadow our 497 tRNA genes.

Repeats

The human genome is riddled with highly repetitive sequences that appear to be gibberish, at least if we restrict the definition of genetic meaning to encoding protein. It is entirely possible that repeats represent a different type of genetic information, perhaps using a language in which meaning lies in a repeat size or number. Some types of repeats may serve a structural function of helping to hold a chromosome together.

The most abundant type of repeat is a sequence of DNA that can jump about the genome, called a transposable element, or **transposon** for short. Originally identified in corn by Barbara McClintock in the 1940s, and then in bacteria as the age of molecular biology dawned in the 1960s, transposons are now known to comprise about 45 percent of the human genome sequence. They are considered to be repeats because they are typically present in many copies. Some transposons include parts that encode enzymes that enable them to leave one chromosomal site and integrate into another. In a way, such a transposon is like a stripped-down virus or retrovirus.

Transposons are classified by size, whether they are transcribed into RNA, which enzymes they use to move, and whether they resemble bacterial transposons. For example, a class of transposons called long interspersed elements (LINEs) are 6,000 bases long and are transcribed and then trimmed to 900 bases before they reinsert into a chromosome. In contrast, short interspersed elements (SINEs) are 100 to 500 bases long and use enzymes that are encoded in LINEs. A major class of SINEs are called Alu repeats. Each Alu repeat is about 300 bases long, and a genome may contain 300,000 to 500,000 of them. Researchers still do not know what Alu repeats do, but they comprise 2 to 3 percent of the genome, and they have been increasing in number over time because they can copy themselves. Other rarer classes of repeats include those that comprise telomeres, centromeres, and rRNA gene clusters; duplications of 10,000 to 300,000 bases (10 to 300 kilobases); copies of pseudogenes; and simple repeats of one, two, or three bases. Many repeats arise from RNAs that are reverse transcribed into DNA and are then inserted into chromosomes.

Our understanding of the functions of repeats lags far behind our knowledge of the roles of the various noncoding RNA genes. Repeats may make sense in light of evolution, past and future. Pseudogenes are likely vestiges of genes that functioned in our nonhuman ancestors. Perhaps the repeats that seem to have no obvious function today can serve as raw material from which future genes may arise.

Key Concepts

Only a tiny proportion of the genome encodes protein, yet the number of proteins greatly outnumbers known protein-encoding genes. Ways that the genome is economical include exon shuffling, introns that encode protein, and single genes that encode a precursor protein that is cut in two. • The 98.5 percent of the human genome that does not specify protein encodes many types of RNA as well as introns, promoters, and other control sequences and repeats. We do not know the functions of some types of repeats.

Summary

10.1 Transcription—The Link Between Gene and Protein

1. The DNA sequence of a gene that encodes a protein is **transcribed** into RNA and **translated** into protein. The overall process is called **gene expression.** But much of the genome does not encode protein. The proteins produced in a particular cell type constitute its **proteome.**
2. RNA is transcribed from the **template strand** of DNA. The other strand is called the **coding strand.**
3. RNA is a single-stranded nucleic acid similar to DNA but containing uracil and ribose rather than thymine and deoxyribose.
4. Several types of RNA participate in protein synthesis (translation). **Messenger RNA** (mRNA) carries a protein-encoding gene's information. **Ribosomal RNA** (rRNA) associates with certain proteins to form ribosomes, which physically support protein synthesis. **Transfer RNA** (tRNA) is cloverleaf-shaped, with a three-base sequence called the **anticodon** that is complementary to mRNA on one end. It bonds to a particular amino acid at the other end.
5. **Transcription factors** regulate which genes or subsets of genes are transcribed in a particular cell type. Operons control gene expression in bacteria.
6. Transcription begins when transcription factors help **RNA polymerase** bind to a gene's starting region, or promoter. RNAP then adds RNA nucleotides to a growing chain, in a sequence complementary to the DNA template strand.
7. After a gene is transcribed, the mRNA receives a "cap" of modified nucleotides at one end and a poly-A tail at the other.
8. Many genes are in pieces. After transcription, segments called **exons** are translated into protein, but segments called **introns** are removed. Introns may outnumber and outsize exons. Researchers aren't certain what introns do, but they are probably not "junk."

10.2 Translating a Protein

9. Each three consecutive mRNA bases form a **codon** that specifies a particular amino acid. The **genetic code** is the correspondence between each codon and the amino acid it specifies. Of the 64 different possible codons, 61 specify amino acids and three signal stop. Because 61 codons specify the 20 amino acids, more than one type of codon may encode a single amino acid. The genetic code is nonoverlapping, triplet, universal, and degenerate.
10. In the 1960s, researchers used logic and clever experiments that used synthetic RNAs to decipher the genetic code. The sequencing of the human genome has refined that information.
11. Translation requires tRNA, ribosomes, energy-storage molecules, enzymes, and protein factors. An **initiation complex** forms when mRNA, a small ribosomal subunit, and a tRNA carrying methionine join. The amino acid chain begins to elongate when a large ribosomal subunit joins the small one. Next, a second tRNA binds by its **anticodon** to the next mRNA codon, and its amino acid bonds with the first amino acid. Transfer RNAs add amino acids, forming a polypeptide. The ribosome moves down the mRNA as the chain grows. The P site bears the amino acid chain, and the A site holds the newest tRNA. When the ribosome reaches a "stop" codon, it falls apart into its two subunits and is released. The new polypeptide breaks free.
12. After translation, some polypeptides are cleaved, and some aggregate to form larger proteins. The cell uses or secretes the protein, which must have a particular **conformation** to be active and functional.
13. A protein's **primary structure** is its amino acid sequence. Its **secondary structure** forms as amino acids close in the primary structure attract one another. **Tertiary structure** appears as more widely separated amino acids approach or repel in response to water molecules. **Quaternary structure** forms when a protein consists of more than one polypeptide. Chaperone proteins help mold conformation. Some proteins can fold into several conformations, some of which can cause disease.

10.3 The Human Genome Sequence Reveals Unexpected Complexity

14. Only 1.5 percent of the 3.2 billion base pairs of the human genome encode protein, yet those 31,000 or so genes specify 100,000 to 200,000 distinct proteins.
15. Several mechanisms explain how a set number of genes can encode a larger number of proteins. These include exon shuffling, use of introns, and cutting proteins translated from a single gene.
16. The 98.5 percent of the human genome that does not encode protein apparently encodes several types of RNA, control sequences, and repeats.

Review Questions

1. Explain how complementary base pairing is responsible for
   a. the structure of the DNA double helix.
   b. DNA replication.
   c. transcription of RNA from DNA.
   d. the attachment of mRNA to a ribosome.
   e. codon/anticodon pairing.
   f. tRNA conformation.
2. A retrovirus has RNA as its genetic material. When it infects a cell, it uses enzymes to copy its RNA into DNA, which then integrates into the host cell's chromosome. Is this flow of genetic information consistent with the central dogma? Why or why not?
3. Genomics is highly dependent upon computer algorithms that search DNA sequences for indications of specialized functions. Explain the significance of detecting the following sequences:
   a. a promoter
   b. a sequence of 75 to 80 bases that folds into a cloverleaf shape
   c. a gene with a sequence very similar to that of a known protein-encoding gene, but that isn't translated into protein
   d. 200 copies of a 44,000 base long sequence

e. RNAs with poly A tails

f. a sequence that is very similar to part of a known virus that is found at several sites in a genome

4. Many antibiotic drugs work by interfering with protein synthesis in the bacteria that cause infections. Explain how each of the following antibiotic mechanisms disrupts genetic function in bacteria.

a. Transfer RNAs misread mRNA codons, binding with the incorrect codon and bringing in the wrong amino acid.

b. The first amino acid is released from the initiation complex before translation can begin.

c. Transfer RNA cannot bind to the ribosome.

d. Ribosomes cannot move.

e. A tRNA picks up the wrong amino acid.

5. Define and distinguish between transcription and translation.

6. List the differences between RNA and DNA.

7. Where in a cell do DNA replication, transcription, and translation occur?

8. How does transcription control cell specialization?

9. How can the same mRNA codon be at an A site on a ribosome at one time, but at a P site at another?

10. Describe the events of transcription initiation.

11. List the three major types of RNA and their functions.

12. State three ways that RNA is altered after it is transcribed.

13. What are the components of a ribosome?

14. Why was the discovery of introns a surprise? of ribozymes?

15. Why would an overlapping genetic code be restrictive?

16. How are the processes of transcription and translation economical?

17. How does the shortening of proinsulin to insulin differ from the shortening of apolipoprotein B?

18. What factors determine how a protein folds into its characteristic conformation?

19. Why would two-nucleotide codons be insufficient to encode the number of amino acids in biological proteins?

20. Cite two ways that RNA helps in its own synthesis, and two ways that proteins help in their own synthesis.

21. In the 1960s, a gene was defined as a continuous sequence of DNA, located permanently at one place on a chromosome, that specifies a sequence of amino acids. State three ways that this statement is incomplete.

22. Until recently, a protein was thought to have only one conformation. Which protein provides evidence that this is incorrect?

23. How can one of the two dental proteins implicated in dentinogenesis imperfecta be much more abundant than the other, if they are both transcribed and translated from the same gene?

24. The four mRNA codons that specify the amino acid leucine are CUU, CUC, CUA, and CUG. Only three types of tRNAs recognize these four codons. How is this possible? Which two codons does a single tRNA recognize?

Applied Questions

1. The *BRCA1* gene that, when missing several bases, causes a form of breast cancer has 24 exons and 23 introns.

a. How many splice sites does the gene contain? (A splice site is the junction of an exon and an intron.)

b. In a woman with *BRCA1* breast cancer, an entire exon is missing, or "skipped." How many splice sites does her affected copy of the gene have?

2. When researchers compared the number of mRNA transcripts that correspond to a part of chromosome 19 to the number of protein-encoding genes in the region, they found 1,859 transcripts and 544 genes. State three mechanisms that could account for the discrepancy.

3. List the sequences of RNA that would be transcribed from the following DNA template sequences.

a. TTACACTTGCTTGAGAGTC

b. ACTTGGGCTATGCTCATTA

c. GGCTGCAATAGCCGTAGAT

d. GGAATACGTCTAGCTAGCA

4. Given the following partial mRNA sequences, reconstruct the corresponding DNA template sequences.

a. GCUAUCUGUCAUAAAAGAGGA

b. GUGGCGUAUUCUUUUCCGGGUAGG

c. GAGGGAAUUCUUUCUCAACGAAGU

d. AGGAAAACCCCUCUUAUUAUAGAU

5. List three different mRNA sequences that could encode the following amino acid sequence:

histidine-alanine-arginine-serine-leucine-valine-cysteine

6. Write a DNA sequence that would encode the following amino acid sequence:

valine-tryptophan-lysine-proline-phenylalanine-threonine

7. In the film *Jurassic Park,* which is about genetically engineered dinosaurs, a cartoon character named Mr. DNA talks about the billions of genetic codes in the DNA. Why is this statement incorrect?

8. In investigating the genetic code, when the researchers examined synthetic RNA of sequence ACACACACACACACA, they found that it encoded the amino acid sequence *thr-his-thr-his-thr-his.* How did the researchers determine the codon assignments for ACA and CAC?

9. Cystic fibrosis is caused by an abnormal chloride channel protein that makes up part of the cell membrane. How might a defect in protein folding cause the cystic fibrosis phenotype at the cellular level?

10. Figure 10.19 shows the distribution of types of proteins that, when abnormal or absent from a certain cell type, cause disease. Such charts have been constructed for different stages of development—prenatal, under a year, childhood, puberty to age 50, and over age 50. Explain the observation that transcription factors account for:

- 9 percent of proteins overall (throughout development and life)
- 25 percent of proteins before birth
- 7 percent of proteins from birth to one year
- 6 percent of proteins from childhood to age 50 years
- 5 percent of proteins for those over 50 years

11. Titin is a muscle protein named for its gargantuan size—its gene has the largest known coding sequence—80,781 DNA bases. How many amino acids long is it?

12. On the television program *The X Files,* Agent Scully discovers an extraterrestrial life form that has a triplet genetic code, but with five different bases, instead of the four of earthly inhabitants. How many different amino acids can this code specify?

13. In malignant hyperthermia, a person develops a life-threateningly high fever after taking certain types of anesthetic drugs. In one family, the mutation deletes three contiguous bases in exon 44. How many amino acids are missing from the protein?

14. A mutation in a gene called RPGR-interacting protein causes visual loss. The encoded protein is 1,259 amino acids long. What is the minimal size of this gene?

15. Parkinson disease caused rigidity, tremors, and other motor symptoms. Only 2 percent of cases are inherited, and these tend to have an early onset of symptoms. Some inherited cases result from mutations in a gene that encodes the protein parkin, which has 12 exons. Indicate whether each of the following mutations in the parkin gene would result in a smaller protein, a larger protein, or not change the size of the protein.

a. deletion of exon 3

b. deletion of six contiguous nucleotides in exon 1

c. duplication of exon 5

d. disruption of the splice site between exon 8 and intron 8

e. deletion of intron 2

16. How can repeated sequences impart information?

Suggested Readings

Caron, Huib, et al. February 16, 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* 291:1289–92. The transcriptome map includes gene expression profiles by chromosome region for various normal and diseased cells.

Dahlberg, Albert. May 4, 2001. The ribosome in action. *Science* 292:868–69. A review of recent research reveals the structures of interacting ribosomes, tRNA, and mRNA, in bacteria.

Gilbert, Walter. February 9, 1978. Why genes in pieces? *Nature* 271:501. A classic and insightful look at the enigma of introns.

Hoagland, Mahlon. 1990. *Toward the habit of truth.* New York: W. W. Norton. The story of the RNA tie club, by a member.

Jiminez-Sanchez, Gerardo, et al. February 15, 2001. Human disease genes. *Nature* 409:853–58. Human inherited disease results from defects in proteins that fall into a few categories.

Kay, Lily E. 2001. *Who Wrote the Book of Life?* Stanford, CA: Stanford University Press. The story of how a group of mostly physicists-turned-biologists deciphered the genetic code, in the 1960s.

Lewis, Ricki. February 1996. On cracked codes, cell walls, and human fungi. *The American Biology Teacher* 58:16. A funny look at errors in genetic code usage.

Lewis, Ricki, and Barry Palevitz. June 11, 2001. Genome economy. *The Scientist* 15:19. The article that forms the basis for section 10.3 of this book shows how the genome specifies many proteins with few genes.

Pollack, Andrew. July 24, 2001. Scientists are starting to add letters to life's alphabet. *The New York Times,* p. F1. Investigators at the Scripps Research Institute in La Jolla, California, are attempting to create life forms that use a more extensive genetic code.

Prusiner, Stanley. May 17, 2001. Shattuck lecture—neurodegenerative diseases and prions. *The New England Journal of Medicine* 344:1516–20. The normal and pathogenic forms of prion protein have the same amino acids sequence, but different conformations.

Solovitch, Sara. July 2001. The citizen scientists. *Wired.* Frustrated by nonexistent or slow research, parents of sick children have been instrumental in discovering the genes behind some rare disorders.

Tupler, Rosella. February 15, 2001. Expressing the human genome. *Nature* 409:832–33. Sequencing the genome will seem easy compared to the task of tracking gene expression in the more than 260 different cell types.

Vogel, Gretchen. February 16, 2001. Why sequence the junk? *Science* 291:1184. DNA as "junk" is a value judgment often based on lack of data.

On the Web

Check out the resources on our website at

www.mhhe.com/lewisgenetics5

On the web for this chapter, you will find additional study questions, vocabulary review, useful links to case studies, tutorials, popular press coverage, and much more. To investigate specific topics mentioned in this chapter, also try the links below:

Annotated Human Genomic Database
www.DoubleTwist.com/genome

Cold Spring Harbor Laboratory Learning Center
vector.cshl.org/resources/resources.html

Functional Genomics Website
www.sciencegenomics.org

Genome Jokes and Cartoons
cagle.slate.msn.com/news/gene

Genome News Network
www.celera.com/genomics/genomic.cfm

The Human Transcriptome Map
http://bioinfo.amc.uva.nl/HTM/l

National Center for Biotechnology Information Splash Page
www.ncbi.nlm.nih.gov/genome/guide/human

Online Mendelian Inheritance in Man
www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

dentinogenesis imperfecta	125490
Duchenne muscular dystrophy	310200
epidermal growth factor	131530
fibronectin	135600
Huntington disease	143100
neurofibromatosis I	162200
tissue plasminogen activator	173370