3

The Human Genome: Structure and Function of Genes and Chromosomes

Over the past 20 years, remarkable progress has been made in our understanding of the structure and function of genes and chromosomes at the molecular level. More recently, this has been supplemented by an in-depth understanding of the organization of the human genome at the level of its DNA sequence. These advances have come about in large measure through the applications of molecular genetics and genomics to many clinical situations, thereby providing the tools for a distinctive new approach to medical genetics. In this chapter, we present an overview of the organization of the human genome and the aspects of molecular genetics that are required for an understanding of the genetic approach to medicine. This chapter is not intended to provide an extensive description of the wealth of new information about gene structure and regulation. To supplement the information discussed here, Chapter 4 describes many experimental approaches of modern molecular genetics that are becoming critical to the practice and understanding of human and medical genetics.

The increased knowledge of genes, and of their organization in the genome, has had an enormous impact on medicine and on our perception of human physiology. As Nobel laureate Paul Berg stated presciently at the dawn of this new era:

> Just as our present knowledge and practice of medicine relies on a sophisticated knowledge of human anatomy, physiology, and biochemistry, so will dealing with disease in the future demand a detailed understanding of the molecular anatomy, physiology, and biochemistry of the human genome. . . . We shall need a more detailed knowledge of how human genes are organized and how they function and are regulated. We shall also have to have physicians who are as conversant with the molecular anatomy and physiology of chromosomes and genes as the cardiac surgeon is with the structure and workings of the heart.

DNA STRUCTURE: A BRIEF REVIEW

DNA is a polymeric nucleic acid macromolecule composed of three types of units: a five-carbon sugar, deoxyribose; a nitrogen-containing base; and a phosphate group (Fig. 3–1). The bases are of two types, purines and pyrimidines. In DNA, there are two purine bases, adenine (A) and guanine (G), and two pyrimidine bases, thymine (T) and cytosine (C). Nucleotides, each composed of a base, a phosphate, and a sugar moiety, polymerize into long polynucleotide chains by 5′–3′ phosphodiester bonds formed between adjacent deoxyribose units (Fig. 3–2). In the human genome, these polynucleotide chains (in their double-helix form) are hundreds of millions of nucleotides long, ranging in size from approximately 50 million base pairs (for the smallest chromosome, chromosome 21) to 250 million base pairs (for the largest chromosome, chromosome 1).

The anatomical structure of DNA carries the chemical information that allows the exact transmission of genetic information from one cell to its daughter cells and from one generation to the next. At the same time, the primary structure of DNA specifies the amino acid sequences of the polypeptide chains of proteins, as described later in this chapter. DNA has elegant features that give it these properties. The native state of DNA, as elucidated by James Watson and Francis Crick in 1953, is a double helix (Fig. 3–3). The helical structure resembles a right-handed spiral staircase in which its two polynucleotide chains run in opposite directions, held together by hydrogen bonds between pairs of bases: A of one chain paired with T of the other, and G with C (see Fig. 3–3). Consequently,

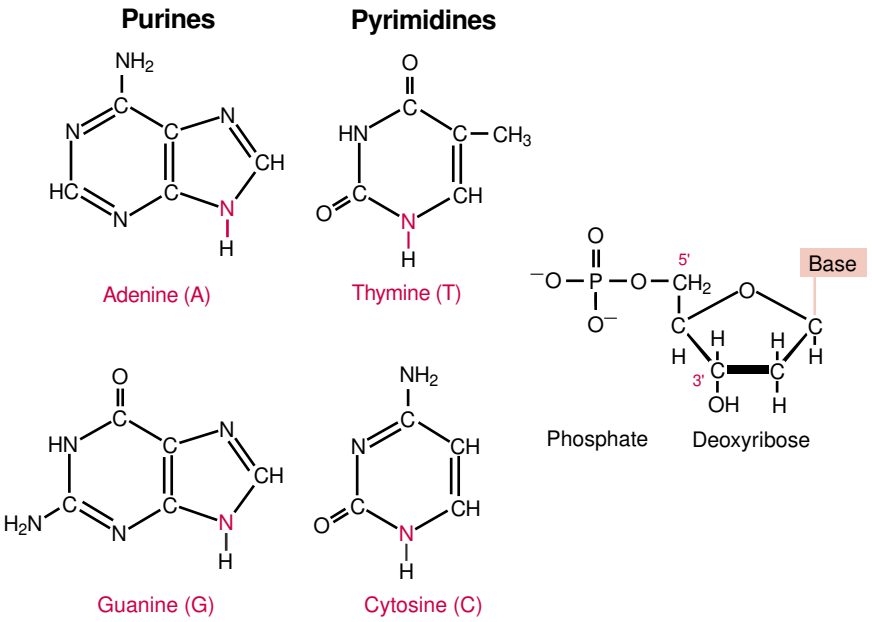


Figure 3–1. The four bases of DNA and the general structure of a nucleotide in DNA. Each of the four bases bonds with deoxyribose (through the nitrogen shown in red) and a phosphate group to form the corresponding nucleotides.

knowledge of the sequence of nucleotide bases on one strand automatically allows one to determine the sequence of bases on the other strand. The double-stranded structure of DNA molecules allows them to replicate precisely by separation of the two strands, followed by synthesis of two new complementary strands, in accordance with the sequence of the original template strands (Fig. 3–4). Similarly, when necessary, the base complementarity allows efficient and correct repair of damaged DNA molecules.

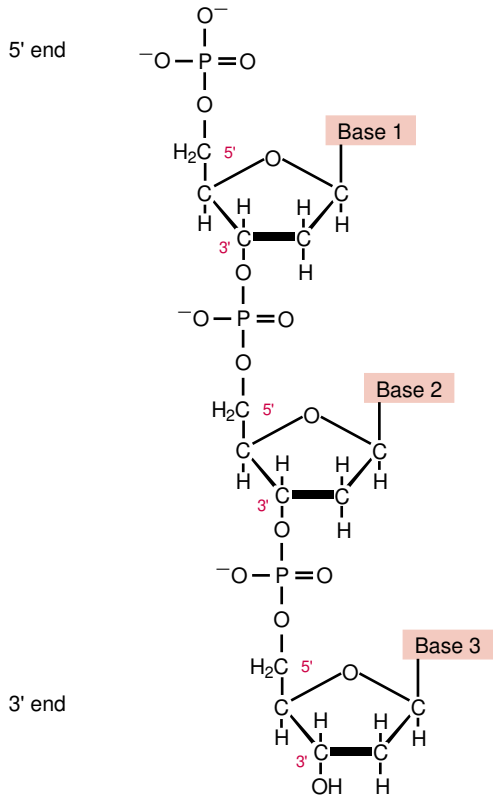


Figure 3–2. A portion of a DNA polynucleotide chain, showing the 3′–5′ phosphodiester bonds that link adjacent nucleotides.

THE CENTRAL DOGMA: DNA → RNA → PROTEIN

Genetic information is contained in DNA in the chromosomes within the cell nucleus, but protein synthesis, during which the information encoded in the DNA is used, takes place in the cytoplasm. This compartmentalization reflects the fact that the human organism is a **eukaryote**. This means that human cells have a genuine nucleus containing the DNA, which is separated by a nuclear membrane from the cytoplasm. In contrast, in prokaryotes like the intestinal bacterium *Escherichia coli*, DNA is not enclosed within a nucleus. Because of the compartmentalization of eukaryotic cells, information transfer from the nucleus to the cytoplasm is a very complex process that has been a focus of attention among molecular and cellular biologists.

The molecular link between these two related types of information (the DNA code of genes and the amino acid code of proteins) is **ribonucleic acid (RNA)**. The chemical structure of RNA is similar to that of DNA, except that each nucleotide in RNA has a ribose sugar component instead of a deoxyribose; in addition, uracil (U) replaces thymine as one of the pyrimidines of RNA (Fig. 3–5). An additional difference between RNA and DNA is that RNA in most organisms exists as a single-stranded molecule, whereas DNA exists as a double helix.

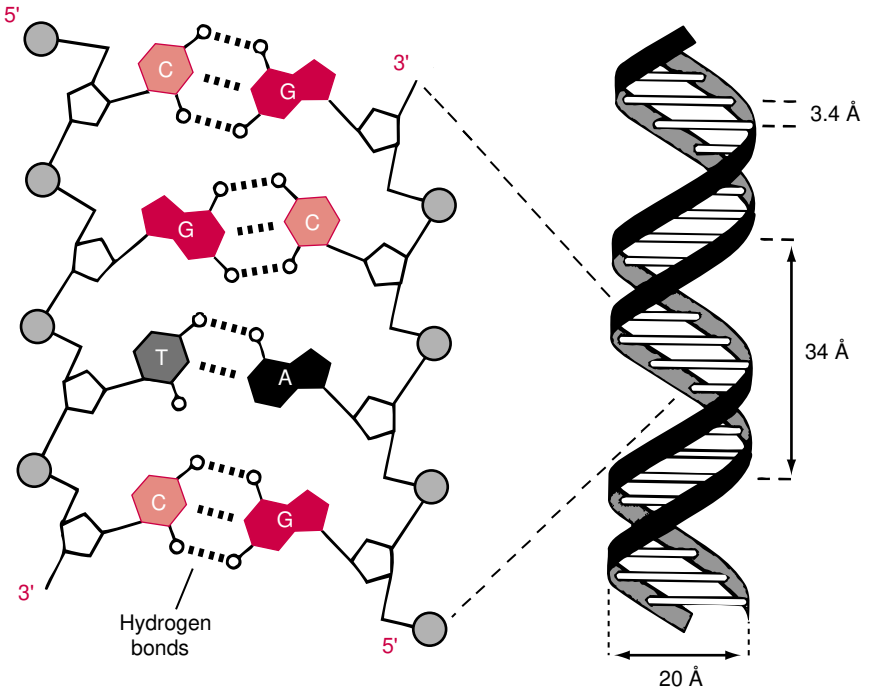


Figure 3–3. The structure of DNA. *Left*, A two-dimensional representation of the two complementary strands of DNA, showing the AT and GC base pairs. Note that the orientation of the two strands is antiparallel. *Right*, The double-helix model of DNA, as proposed by Watson and Crick. The horizontal "rungs" represent the paired bases. The helix is said to be right-handed because the strand going from lower left to upper right crosses over the opposite strand. (Based on Watson JD, Crick FHC [1953] Molecular structure of nucleic acids—A structure for deoxyribose nucleic acid. Nature 171:737–738.)

The informational relationships among DNA, RNA, and protein are intertwined: DNA directs the synthesis and sequence of RNA, RNA directs the synthesis and sequence of polypeptides, and specific proteins are involved in the synthesis and metabolism of DNA and RNA. This flow of information is referred to as the "central dogma" of molecular biology.

Genetic information is stored in DNA by means of a code (the **genetic code**, discussed later) in which the sequence of adjacent bases ultimately determines the sequence of amino acids in the encoded polypeptide. First, RNA is synthesized from the DNA template through a process known as **transcription**. The RNA, carrying the coded information in a form called **messenger RNA (mRNA)**, is then transported from the nucleus to the cytoplasm, where the RNA sequence is decoded, or translated, to determine the sequence of amino acids in the protein being synthesized. The process of **translation** occurs on **ribosomes**, which are cytoplasmic organelles with binding sites for all of the interacting molecules, including the mRNA, involved in protein synthesis. Ribosomes are themselves made up of many different structural proteins in association with a specialized type of RNA known as **ribosomal RNA (rRNA)**. Translation involves yet a third type of RNA, **transfer RNA (tRNA)**, which provides the molecular link between the coded base sequence of the mRNA and the amino acid sequence of the protein.

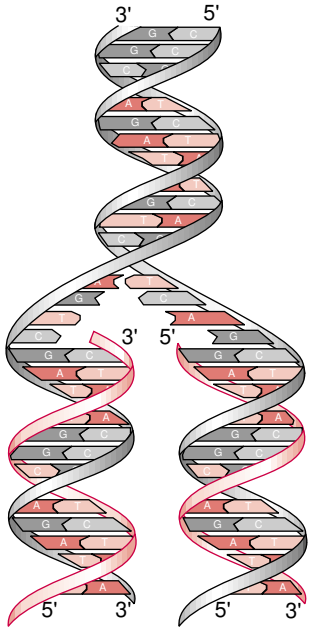


Figure 3–4. Replication of a DNA double helix, resulting in two identical daughter molecules, each composed of one parental strand (*black*) and one newly synthesized strand (*red*).

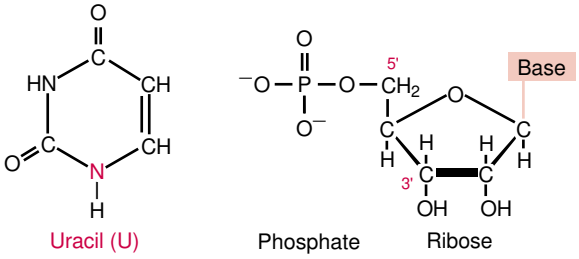


Figure 3–5. The pyrimidine uracil and the structure of a nucleotide in RNA. Note that the sugar ribose replaces the sugar deoxyribose of DNA. Compare with Figure 3–1.

Because of the interdependent flow of information represented by the central dogma, one can begin discussion of the molecular genetics of gene expression at any of its three informational levels: DNA, RNA, or protein. We begin by examining the structure of genes as a foundation for discussion of the genetic code, transcription, and translation.

Gene Structure and Organization

In its simplest form, a gene can be visualized as a segment of a DNA molecule containing the code for the amino acid sequence of a polypeptide chain and the regulatory sequences necessary for expression. This description, however, is inadequate for genes in the human genome (and indeed in most eukaryotic genomes), because few genes exist as continuous coding sequences. Rather, the vast majority of genes are interrupted by one or more noncoding regions. These intervening sequences, called **introns**, are initially transcribed into RNA in the nucleus but are not present in the mature mRNA in the cytoplasm. Thus, information from the intronic sequences is not normally represented in the final protein product. Introns alternate with coding sequences, or **exons**, that ultimately encode the amino acid sequence of the protein (Fig. 3–6). Although a few genes in the human genome

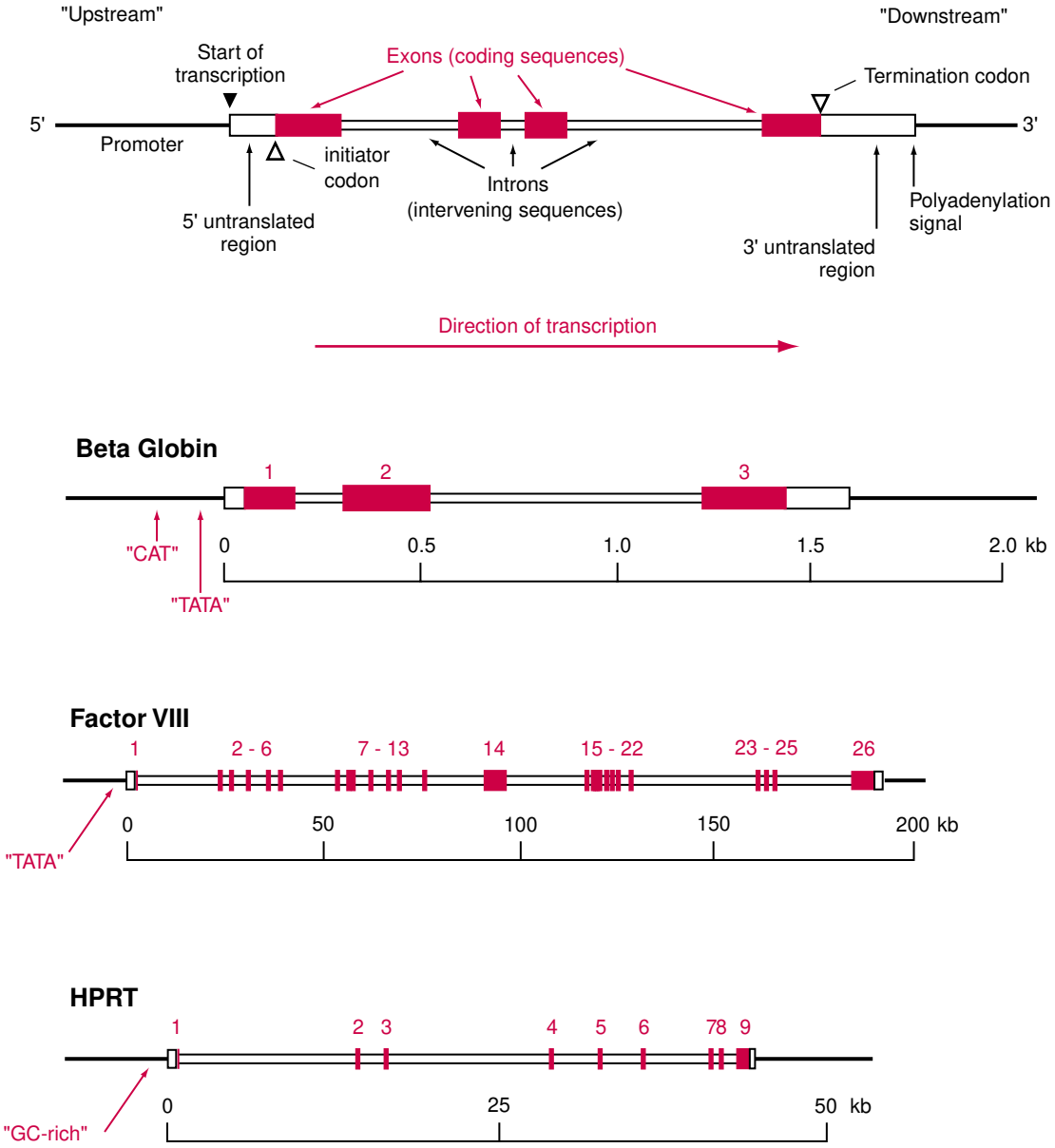


Figure 3–6. General structure of a typical human gene. Individual features are labeled in the figure and discussed in the text. Examples of three medically important human genes are presented at the bottom of the figure. Individual exons are numbered. Different mutations in the β-globin gene cause a variety of important hemoglobinopathies. Mutations in the factor VIII gene cause hemophilia A. Mutations in the hypoxanthine phosphoribosyltransferase (*HPRT*) gene lead to Lesch-Nyhan syndrome.

have no introns, most genes contain at least one and usually several introns. Surprisingly, in many genes, the cumulative length of introns makes up a far greater proportion of a gene's total length than do the exons. Whereas some genes are only a few kilobases (kb, where 1 kb = 1000 base pairs) in length, others, like the factor VIII gene shown in Figure 3–6, stretch on for hundreds of kb. There are a few exceptionally large genes, including the X-linked dystrophin gene (mutations in which lead to Duchenne muscular dystrophy) that spans more than 2 million base pairs (2000 kb), of which less than 1 percent consists of coding exons.

Structural Features of a Typical Human Gene

A schematic representation of a portion of chromosomal DNA containing a typical gene is shown in Figure 3–6, along with the structure of several medically relevant genes. Together, they illustrate the range of features that characterize human genes. In Chapters 1 and 2, we briefly defined "gene" in general terms. At this point, we can provide a molecular definition of a gene. In typical circumstances, we define a gene as *a sequence of chromosomal DNA that is required for production of a functional product*, be it a polypeptide or a functional RNA molecule. As is clear from Figure 3–6, a gene includes not only the actual coding sequences but also adjacent nucleotide sequences required for the proper expression of the gene—that is, for the production of a normal mRNA molecule, in the correct amount, in the correct place, and at the correct time during development or during the cell cycle.

The adjacent nucleotide sequences provide the molecular "start" and "stop" signals for the synthesis of mRNA transcribed from the gene. At the 5′ end of the gene lies a **promoter** region, which includes sequences responsible for the proper initiation of transcription. Within the 5′ region are several DNA elements whose sequence is conserved among many different genes. This conservation, together with functional studies of gene expression in many laboratories, indicates that these particular sequence elements play an important role in regulation. There are several different types of promoter found in the human genome, with different regulatory properties that specify the developmental patterns, as well as the levels of expression, of a particular gene in different tissues. The roles of individual conserved promoter elements, identified in Figure 3–6, are discussed in greater detail in the section on "Fundamentals of Gene Expression." Both promoters and other regulatory elements (located either 5′ or 3′ of a gene or in its introns) can be sites of mutation in genetic disease that can interfere with the normal expression of a gene. These regulatory elements, including **enhancers, silencers**, and **locus control regions** (LCRs), are discussed more fully later in this chapter.

At the 3′ end of the gene lies an untranslated region of importance that contains a signal for addition of a sequence of adenosine residues (the so-called polyA tail) to the end of the mature mRNA. Although it is generally accepted that such closely neighboring regulatory sequences are part of what is called a "gene," the precise dimensions of any particular gene will remain somewhat uncertain until the potential functions of more distant sequences are fully characterized.

Gene Families

Many genes belong to families of closely related DNA sequences, recognized as families because of similarity of the nucleotide sequence of the genes themselves or of the amino acid sequence of the encoded polypeptides.

One small, but medically important gene family is composed of genes that encode the protein chains found in hemoglobins. The α-globin and β-globin gene clusters, on chromosomes 16 and 11, respectively, are shown in Figure 3–7 and are believed to have arisen by duplication of a primitive precursor gene about 500 million years ago. These two clusters contain genes coding for closely related globin chains expressed at different developmental stages, from embryo to adult. The individual genes within each cluster are more similar in sequence to one another

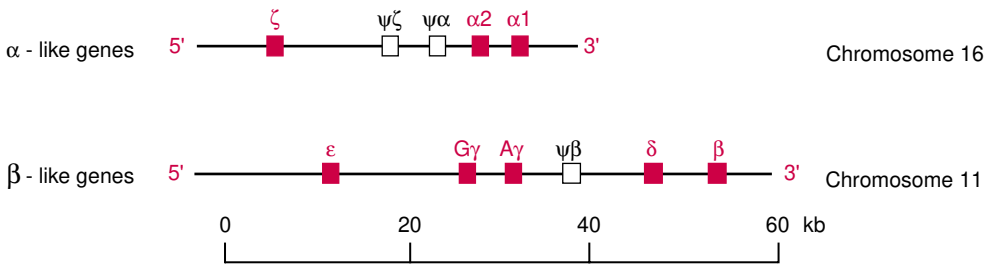


Figure 3–7. Chromosomal organization of the two clusters of human globin genes. Functional genes are indicated in red. Pseudogenes are indicated by the open boxes. (Redrawn from Nienhuis AW, Maniatis T [1987] Structure and expression of globin genes in erythroid cells. *In* Stamatoyannopoulos G, Nienhuis AW, Leder P, Majerus PW [eds] The Molecular Basis of Blood Diseases. WB Saunders, Philadelphia, pp. 28–65.)

than to genes in the other cluster; thus, each cluster is believed to have evolved by a series of sequential gene duplication events within the past 100 million years. The exon-intron patterns of the globin genes appear to have been remarkably conserved during evolution; each of the functional globin genes shown in Figure 3–7 has two introns at similar locations, although the sequences contained within the introns have accumulated far more nucleotide base changes over time than have the coding sequences of each gene. The control of expression of the various globin genes, in the normal state as well as in the many inherited hemoglobinopathies, is considered in more detail both later in this chapter and in Chapter 11.

Several of the globin genes do not produce any RNA or protein product and therefore are unlikely to have any function. DNA sequences that closely resemble known genes but are nonfunctional are called **pseudogenes**. Pseudogenes are widespread in the genome and are thought to be byproducts of evolution, representing genes that were once functional but are now vestigial, having been inactivated by mutations in coding or regulatory sequences. In some cases, as in the pseudo-α-globin and pseudo-β-globin genes, the pseudogenes presumably arose through gene duplication, followed by the introduction of numerous mutations into the extra copies of the once-functional gene. In other cases, pseudogenes have been formed by a process, called **retrotransposition**, that involves transcription, generation of a DNA copy of the mRNA, and, finally, integration of such DNA copies back into the genome. Pseudogenes created by retrotransposition lack introns and are called **processed pseudogenes**. They are not necessarily or usually on the same chromosome (or chromosomal region) as their progenitor gene.

The largest known gene family in the human genome is the so-called **immunoglobulin superfamily**, which includes many hundreds of genes involved in cell surface recognition events in the immune and nervous system, such as genes on chromosomes 2, 14, and 22 that encode the immunoglobulin heavy and light chains themselves; genes on chromosome 6 that make up the major histocompatibility complex; genes on chromosomes 7 and 14 whose products make up the T-cell receptor; and genes that are expressed primarily in neural tissues, such as genes for cell adhesion molecules or for myelin-associated glycoproteins. The structure and function of many of these genes are examined in detail in Chapter 14.

FUNDAMENTALS OF GENE EXPRESSION

The flow of information from gene to polypeptide involves several steps (Fig. 3–8). Initiation of transcription of a gene is under the influence of promoters and other regulatory elements, as well as specific proteins known as **transcription factors** that interact with specific sequences within these regions. Transcription of a gene is initiated at the transcriptional "start site" on chromosomal DNA just upstream from the coding sequences and continues along the chromosome, for anywhere from several hundred base pairs to more than a million base pairs, through both introns and exons and past the end of the coding sequences. After modification at both the 5′ and 3′ ends of the primary RNA transcript, the portions corresponding to introns are removed, and the segments corresponding to exons are spliced together. After RNA splicing, the resulting mRNA (now colinear with only the coding portions of the gene) is transported from the nucleus to the cytoplasm, where the mRNA is finally translated into the amino acid sequence of the encoded polypeptide. Each of the steps in this complex pathway is prone to error, and mutations that interfere with the individual steps have been implicated in a number of inherited genetic disorders (see Chapters 5, 11, and 12).

Transcription

Transcription of protein-coding genes by RNA polymerase II (one of several classes of RNA polymerases) is initiated upstream from the first coding sequence at the transcriptional start site, the point corresponding to the 5′ end of the final RNA product (see Figs. 3–6 and 3–8). Synthesis of the primary RNA transcript proceeds in a 5′-to-3′ direction, whereas the strand of the gene being transcribed is actually read in a 3′-to-5′ direction with respect to the direction of the deoxyribose phosphodiester backbone (see Fig. 3–2). Because the RNA synthesized corresponds both in polarity and in base sequence (substituting U for T) to the 5′-to-3′ strand of DNA, this nontranscribed DNA strand is sometimes called the "coding," or "**sense**," DNA strand. The 3′-to-5′ transcribed strand of DNA is then referred to as the "noncoding," or "**antisense**," strand. Transcription continues through both intron and exon portions of the gene, beyond the position on the chromosome that eventually corresponds to the 3′ end of the mature mRNA. Whether transcription ends at a predetermined 3′ termination point is unknown.

The primary RNA transcript is processed by addition of a chemical "cap" structure to the 5′ end of the RNA and cleavage of the 3′ end at a specific point downstream from the end of the coding information. This cleavage is followed by addition of a polyA tail to the 3′ end of the RNA; the polyA tail appears to increase the stability of the resulting polyadenylated RNA. The location of the polyadenylation point is specified in part by the sequence AAUAAA (or a variant of this), usually found in the 3′ untranslated

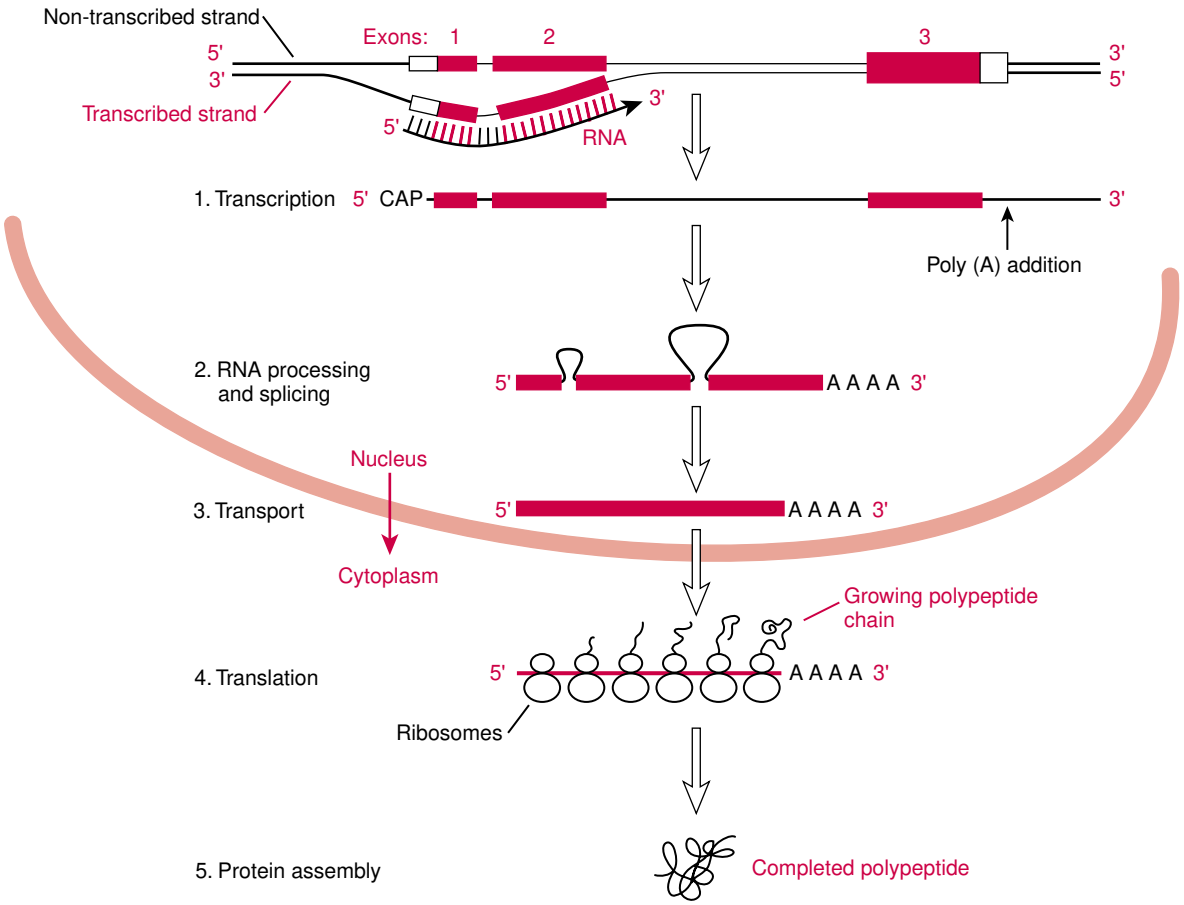


Figure 3–8. Flow of information from DNA to RNA to protein for a hypothetical gene with three exons and two introns. Steps include transcription, RNA processing and splicing, RNA transport from the nucleus to the cytoplasm, and translation.

portion of the RNA transcript. These post-transcriptional modifications take place in the nucleus, as does the process of RNA splicing. The fully processed RNA, now called mRNA, is then transported to the cytoplasm, where translation takes place (see Fig. 3–8).

Translation and the Genetic Code

In the cytoplasm, mRNA is translated into protein by the action of a variety of tRNA molecules, each specific for a particular amino acid. These remarkable molecules, each only 70 to 100 nucleotides long, have the job of transferring the correct amino acids to their positions along the mRNA template, to be added to the growing polypeptide chain. Protein synthesis occurs on ribosomes, macromolecular complexes made up of rRNA (encoded by the 18S and 28S rRNA genes) and several dozen ribosomal proteins (see Fig. 3–8).

The key to translation is a code that relates specific amino acids to combinations of three adjacent bases along the mRNA. Each set of three bases constitutes a **codon**, specific for a particular amino acid (Table 3–1). In theory, almost infinite variations are possible in the arrangement of the bases along a polynucleotide chain. At any one position, there are four possibilities (A, T, C, or G); thus, there are 4^n possible combinations in a sequence of n bases. For three bases, there are 4^3, or 64, possible triplet combinations. These 64 codons constitute the **genetic code**.

Because there are only 20 amino acids and 64 possible codons, most amino acids are specified by more than one codon; hence the code is said to be **degenerate**. For instance, the base in the third position of the triplet can often be either purine (A or G) or either pyrimidine (T or C) or, in some cases, any one of the four bases, without altering the coded message (see Table 3–1). Leucine and arginine are each specified by six codons. Only methionine and tryptophan are each specified by a single, unique codon. Three of the codons are called **stop** (or **nonsense**) **codons** because they designate termination of translation of the mRNA at that point.

Translation of a processed mRNA is always initiated at a codon specifying methionine. Methionine is therefore the first encoded (amino-terminal) amino acid of each polypeptide chain, although it is usually removed before protein synthesis is completed. The

TABLE 3–1

The Genetic Code

First Base	Second Base: U		C		A		G		Third Base
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC	phe	UCC	ser	UAC	tyr	UGC	cys	C
	UUA	leu	UCA	ser	UAA	stop	UGA	stop	A
	UUG	leu	UCG	ser	UAG	stop	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC	leu	CCC	pro	CAC	his	CGC	arg	C
	CUA	leu	CCA	pro	CAA	gln	CGA	arg	A
	CUG	leu	CCG	pro	CAG	gln	CGG	arg	G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC	ile	ACC	thr	AAC	asn	AGC	ser	C
	AUA	ile	ACA	thr	AAA	lys	AGA	arg	A
	AUG	met	ACG	thr	AAG	lys	AGG	arg	G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC	val	GCC	ala	GAC	asp	GGC	gly	C
	GUA	val	GCA	ala	GAA	glu	GGA	gly	A
	GUG	val	GCG	ala	GAG	glu	GGG	gly	G

Abbreviations for amino acids:

ala (A)	alanine	leu (L)	leucine
arg (R)	arginine	lys (K)	lysine
asn (N)	asparagine	met (M)	methionine
asp (D)	aspartic acid	phe (F)	phenylalanine
cys (C)	cysteine	pro (P)	proline
gln (Q)	glutamine	ser (S)	serine
glu (E)	glutamic acid	thr (T)	threonine
gly (G)	glycine	trp (W)	tryptophan
his (H)	histidine	tyr (Y)	tyrosine
ile (I)	isoleucine	val (V)	valine

Other abbreviation:

stop	termination codon

Codons are shown in terms of messenger RNA, which are complementary to the corresponding DNA codons.

codon for methionine (the initiator codon, AUG) establishes the **reading frame** of the mRNA; each subsequent codon is read in turn to predict the amino acid sequence of the protein.

The molecular links between codons and amino acids are the specific tRNA molecules. A particular site on each tRNA forms a three-base **anticodon** that is complementary to a specific codon on the mRNA. Bonding between the codon and anticodon brings the appropriate amino acid into the next position on the ribosome for attachment by formation of a peptide bond to the carboxyl end of the growing polypeptide chain. The ribosome then slides along the mRNA exactly three bases, bringing the next codon into line for recognition by another tRNA with the next amino acid. Thus, proteins are synthesized from the amino terminus to the carboxyl terminus, which corresponds to translation of the mRNA in a 5′-to-3′ direction.

As mentioned earlier, translation ends when a stop codon (UGA, UAA, or UAG) is encountered in the same reading frame as the initiator codon. (Stop codons in either of the other two, unused reading frames are not read and therefore have no effect on translation.) The completed polypeptide is then released from the ribosome, which becomes available to begin synthesis of another protein.

Post-Translational Processing

Many proteins undergo extensive post-translational modifications. The polypeptide chain that is the primary translation product is folded and bonded into a specific three-dimensional structure that is determined by the amino acid sequence itself. Two or more polypeptide chains, products of the same gene or of different genes, may combine to form a single mature protein complex. For example, two α-globin chains and two β-globin chains associate noncovalently to form the tetrameric $\alpha_2\beta_2$ hemoglobin molecule. The protein products may also be modified chemically by, for example, addition of phosphate or carbohydrates at specific sites. Other modifications may involve cleavage of the protein, either to remove specific amino-terminal sequences after they have

functioned to direct a protein to its correct location within the cell (e.g., proteins that function within the nucleus or mitochondria) or to split the molecule into smaller polypeptide chains. For example, the two chains that make up mature insulin, one 21 and the other 30 amino acids long, are originally part of an 82 amino-acid primary translation product, called proinsulin.

Gene Expression in Action: The Beta-Globin Gene

The flow of information outlined in the preceding sections can best be appreciated by reference to a particular well-studied gene, the β-globin gene. The β-globin chain is a 146 amino-acid polypeptide, encoded by a gene that occupies approximately 1.6 kb on the short arm of chromosome 11. The gene has three exons and two introns (see Fig. 3–6). The β-globin gene, as well as the other genes in the β-globin cluster (see Fig. 3–7), is transcribed in a centromere-to-telomere direction. This orientation, however, is different for other genes in the genome and depends on which strand of the chromosomal double helix is the coding strand for a particular gene.

DNA sequences required for accurate initiation of transcription of the β-globin gene are located in the promoter within approximately 200 base pairs upstream from the transcription start site. The double-stranded DNA sequence of this region of the β-globin gene, the corresponding RNA sequence, and the translated sequence of the first 10 amino acids are depicted in Figure 3–9 to illustrate the relationships among these three information levels. As mentioned previously, it is the 3′-to-5′ strand of the DNA that serves as template and is actually transcribed, but it is the 5′-to-3′ strand of DNA that most directly corresponds to the 5′-to-3′ sequence of the mRNA (and, in fact, is identical to it except that U is substituted for T). Because of this correspondence, the 5′-to-3′ DNA strand of a gene (i.e., the strand that is *not* transcribed) is the strand generally reported in the scientific literature or in databases.

In accordance with this convention, the complete sequence of approximately 2.0 kb of chromosome 11 that includes the β-globin gene is shown in Figure 3–10. (It is sobering to reflect that this page of nucleotides represents only 0.000067 percent of the sequence of the entire human genome!) Within these 2.0 kb are contained most, but not all, of the sequence elements required to encode and regulate the expression of this gene. Indicated in Figure 3–10 are many of the important structural features of the β-globin gene, including conserved promoter sequence elements, intron and exon boundaries, RNA splice sites, the initiator and termination codons, and the polyadenylation signal, all of which are known to be mutated in various inherited defects of the β-globin gene (see Chapter 11).

Initiation of Transcription

The β-globin promoter, like many other gene promoters, consists of a series of relatively short functional elements that are thought to interact with specific proteins (generically called **transcription factors**) that regulate transcription, including, in the case of the globin genes, those proteins that restrict expression of these genes to erythroid cells, the cells in which hemoglobin is produced. One important promoter sequence is the "TATA box," a conserved region rich in adenines and thymines that is approximately 25 to 30 base pairs upstream of the start site of transcription (see Figs. 3–6 and 3–10). The TATA box appears to be important for determining the position of the start of transcription, which in the β-globin gene is approximately 50 base pairs upstream from the translation initiation site (see Fig. 3–9). Thus, in this gene there are about 50 base pairs of sequence that are transcribed but are not translated. In other genes, this 5′ transcribed but untranslated region (called the 5′ UTR) can be much longer and can, in fact, be interrupted by one or more introns. A second conserved region, the so-called CAT box (actually CCAAT), is a few dozen base pairs farther upstream (see Fig. 3–10). Both experimentally

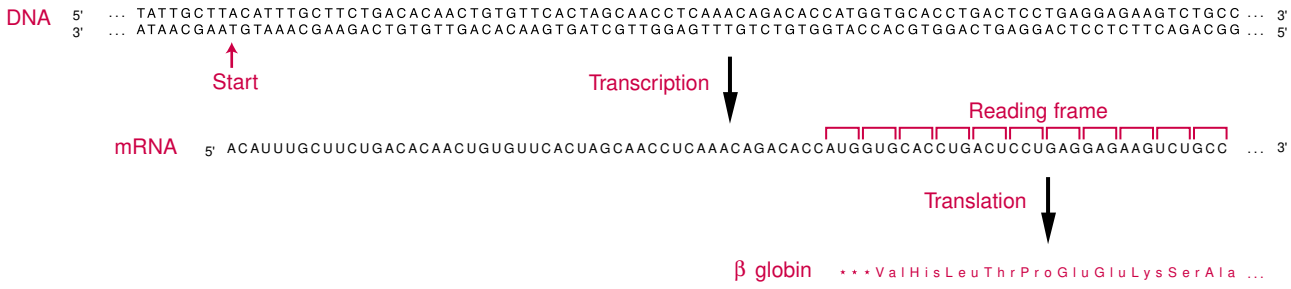


Figure 3–9. Structure and nucleotide sequence of the 5′ end of the human β-globin gene on the short arm of chromosome 11. Transcription of the 3′- to-5′ (*lower*) strand begins at the indicated start size to produce β-globin mRNA. The translational reading frame is determined by the AUG initiator codon (***); subsequent codons specifying amino acids are indicated in red. The other two potential frames are not used.

Figure 3–10. Nucleotide sequence of the complete human β-globin gene. The sequence of the 5′-to-3′ strand of the gene is shown. Capital letters represent sequences corresponding to mature mRNA. Lowercase letters indicate introns and flanking sequences. The CAT and TATA box sequences in the 5′ flanking region are boxed. The ATG initiator codon (AUG in mRNA) and the TAA stop codon (UAA in mRNA) are shown in red. The amino-acid sequence of β-globin is shown above the coding sequence; the three-letter abbreviations in Table 3–1 are used here. The GT and AG dinucleotides important for RNA splicing at the intron/exon junctions are boxed. (From Lawn RM, Efstratiadis A, O'Connell C, et al [1980] The nucleotide sequence of the human β-globin gene. Cell 21: 647–651.)

induced and naturally occurring mutations in either of these sequence elements, as well as in other regulatory sequences even farther upstream, lead to a sharp reduction in the level of transcription, thereby demonstrating the importance of these elements for normal gene expression. Many mutations in these regulatory elements have been identified in patients with the disorder β-thalassemia (see Chapter 11).

Not all gene promoters contain the two specific elements described. In particular, genes that are

constitutively expressed in most or all tissues (called "housekeeping" genes) often lack the CAT and TATA boxes that are more typical of tissue-specific genes. Promoters of many housekeeping genes often contain a high proportion of cytosines and guanines in relation to the surrounding DNA (see the promoter of the hypoxanthine phosphoribosyltransferase gene in Fig. 3–6). Such CG-rich promoters are often located in regions of the genome called **CG** (or **CpG**) **islands**, so named because of the unusually high concentration of the dinucleotide 5′-CG-3′ that stands out from the more general AT-rich chromosomal landscape. Some of the CG-rich sequence elements found in these promoters are thought to serve as binding sites for specific transcription factors.

In addition to the sequences that constitute a promoter itself, there are other sequence elements that can markedly alter the efficiency of transcription. The best characterized of these "activating" sequences are called **enhancers**. Enhancers are sequence elements that can act at a distance (often several kb) from a gene to stimulate transcription. Unlike promoters, enhancers are both position- and orientation-independent and can be located either 5′ or 3′ of the transcription start site. Enhancer elements function only in certain cell types and thus appear to be involved in establishing the tissue specificity and/or level of expression of many genes, in concert with one or more transcription factors. In the case of the β-globin gene, several tissue-specific enhancers are present within both the gene itself and in its flanking regions. The interaction of enhancers with particular proteins leads to increased levels of transcription.

Normal expression of the β-globin gene during development also requires more distant sequences called the **locus control region (LCR)**, located upstream of the ε-globin gene (see Fig. 3–7), which is required for appropriate high-level expression. As expected, mutations that disrupt or delete either enhancer or LCR sequences interfere with or prevent β-globin gene expression (see Chapter 11).

RNA Splicing

The primary RNA transcript of the β-globin gene contains two exons, approximately 100 and 850 base pairs in length, that need to be spliced together. The process is exact and highly efficient; 95 percent of β-globin transcripts are thought to be accurately spliced to yield functional globin mRNA. The splicing reactions are guided by specific DNA sequences at both the 5′ and the 3′ ends of introns. The 5′ sequence consists of nine nucleotides, of which two (the dinucleotide GT located in the intron immediately adjacent to the splice site) are virtually invariant among splice sites in different genes (see Fig. 3–10). The 3′ sequence consists of about a dozen nucleotides, of which, again, two, the AG located immediately 5′ to the intron/exon boundary, are obligatory for normal splicing. The splice sites themselves are unrelated to the reading frame of the particular mRNA. In some instances, as in the case of intron 1 of the β-globin gene, the intron actually splits a specific codon (see Fig. 3–10).

The medical significance of RNA splicing is illustrated by the fact that mutations within the conserved sequences at the intron/exon boundaries commonly impair RNA splicing, with a concomitant reduction in the amount of normal, mature β-globin mRNA; mutations in the GT or AG dinucleotides mentioned earlier invariably eliminate normal splicing of the intron containing the mutation. A number of splice site mutations, identified in patients with β-thalassemia, are discussed in detail in Chapter 11.

Polyadenylation

The mature β-globin mRNA contains approximately 130 base pairs of 3′ untranslated material (the 3′ UTR) between the stop codon and the location of the polyA tail (see Fig. 3–10). As in other genes, cleavage of the 3′ end of the mRNA and addition of the polyA tail is controlled, at least in part, by an AAUAAA sequence approximately 20 base pairs before the polyadenylation site. Mutations in this polyadenylation signal in patients with β-thalassemia (as well as mutations in the corresponding polyadenylation signal in the α-globin gene in patients with α-thalassemia) document the importance of this signal for proper 3′ cleavage and polyadenylation (see Chapter 11). The 3′ untranslated region of some genes can be quite long, up to several kb. Other genes have a number of alternative polyadenylation sites, selection among which may influence the stability of the resulting mRNA and thus the steady-state level of each mRNA.

STRUCTURE OF HUMAN CHROMOSOMES

The composition of genes in the human genome, as well as the determinants of their expression, is specified in the DNA of the 46 human chromosomes. As we saw in an earlier section, *each human chromosome is believed to consist of a single, continuous DNA double helix*; that is, each chromosome in the nucleus is a long, linear double-stranded DNA molecule. Chromosomes are not naked DNA double helices, however. The DNA molecule of a chromosome exists as a complex with a family of basic chromosomal proteins called histones and with a heterogeneous group of acidic, nonhistone proteins that are much

less well characterized, but that appear to be critical for establishing a proper environment to ensure normal chromosome behavior and appropriate gene expression. Together, this complex of DNA and protein is called **chromatin**.

There are five major types of histones that play a critical role in the proper packaging of the chromatin fiber. Two copies each of the four core histones H2A, H2B, H3, and H4 constitute an octamer, around which a segment of DNA double helix winds, like thread around a spool (Fig. 3–11). Approximately 140 base pairs of DNA are associated with each histone core, making just under two turns around the octamer. After a short (20 to 60 base-pair) "spacer" segment of DNA, the next core DNA complex forms, and so on, giving chromatin the appearance of beads on a string. Each complex of DNA with core histones is called a **nucleosome**, which is the basic structural unit of chromatin. The fifth histone, H1, appears to bind to DNA at the edge of each nucleosome, in the internucleosomal spacer region. The amount of DNA associated with a core nucleosome, together with the spacer region, is about 200 base pairs.

During the cell cycle, as we saw in Chapter 2, chromosomes pass through orderly stages of condensation and decondensation (see Fig. 2–5). In the interphase nucleus, chromosomes and chromatin are quite decondensed in relation to the highly condensed state of chromatin in metaphase. Nonetheless, even in interphase chromosomes, DNA in chromatin is substantially more condensed than it would be as a native, protein-free double helix.

The long strings of nucleosomes are themselves further compacted into a secondary helical chromatin structure that appears under the electron microscope as a thick, 30-nm-diameter fiber (about three times thicker than the nucleosomal fiber) (see Fig. 3–11). This cylindrical "solenoid" fiber (from the Greek *solenoeides*, "pipe-shaped") appears to be the fundamental unit of chromatin organization. The solenoids are themselves packed into **loops** or domains attached at intervals of about 100 kb or so to a nonhistone protein **scaffold** or matrix. It has been speculated that loops are, in fact, functional units of DNA replication or gene transcription, or both, and that the attachment points of each loop are fixed along the chromosomal DNA. Thus, one level of control of gene expression may depend on how DNA and genes are packaged into chromosomes and on their association with chromatin proteins in the packaging process.

The various hierarchical levels of packaging seen in an interphase chromosome are illustrated schematically in Figure 3–11. The enormous amount of DNA packaged into a chromosome can be appreciated when chromosomes are treated to remove most of the chromatin proteins in order to observe the protein scaffold (Fig. 3–12). When DNA is released from chromosomes treated this way, long loops of DNA can be visualized, and the residual scaffolding can be seen to reproduce the outline of a typical metaphase chromosome.

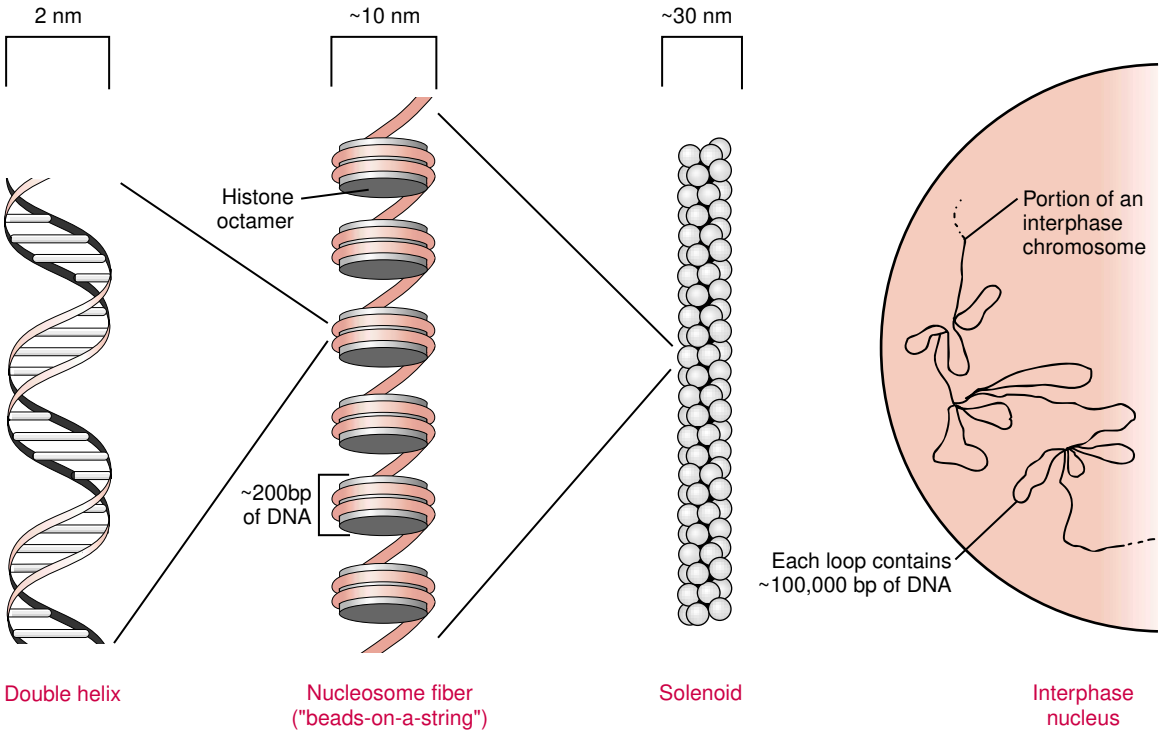


Figure 3–11. Hierarchical levels of chromatin packaging in a human chromosome.

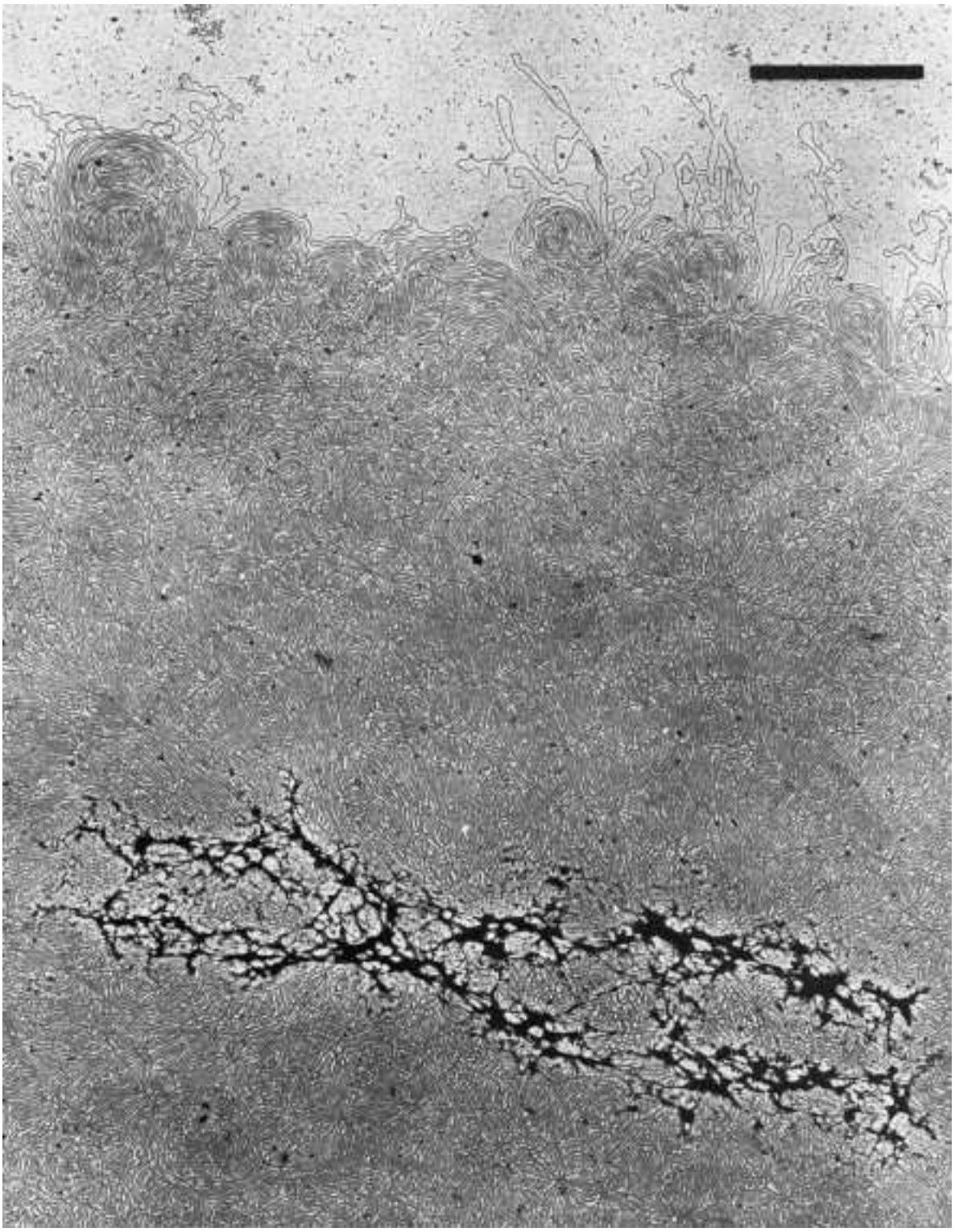


Figure 3–12. Electron micrograph of a protein-depleted human metaphase chromosome, showing the residual chromosome scaffold and loops of DNA. Individual DNA fibers can be best seen at the edge of the DNA loops. Bar = 2 μ. (From Paulson JR, Laemmli UK [1977] The structure of histone-depleted metaphase chromosomes. Cell 12:817–828. Reprinted by permission of the authors and Cell Press.)

The Mitochondrial Chromosome

A small but important subset of genes encoded in the human genome resides in the cytoplasm in the mitochondria. Mitochondrial genes exhibit exclusively maternal inheritance (see Chapter 5). Human cells have hundreds of mitochondria, each containing a number of copies of a small circular molecule, the mitochondrial chromosome. The mitochondrial DNA molecule is only 16 kb in length (less than 0.03

percent of the length of the smallest nuclear chromosome!) and encodes only a few dozen genes. Although the products of these genes function in mitochondria, it should be emphasized that the vast majority of proteins found in mitochondria are, in fact, the products of nuclear genes. Mutations in mitochondrial genes have been demonstrated in several maternally inherited, as well as sporadic, disorders (see Chapter 12).

ORGANIZATION OF THE HUMAN GENOME

Regions of the genome with similar characteristics or organization, replication, and expression are not arranged randomly but, rather, tend to be clustered together. This functional organization of the genome correlates remarkably well with its structural organization as revealed by metaphase chromosome banding (introduced in Chapter 2 and discussed in detail in Chapter 9). The overall significance of this functional organization is that chromosomes are not just a random collection of different types of genes and other DNA sequences. Some chromosome regions, or even whole chromosomes, are quite high in gene content ("gene-rich"), whereas others are low ("gene-poor"). Certain types of sequence are characteristic of the different physical hallmarks of human chromosomes. The clinical consequences of abnormalities of genome structure reflect the specific nature of the genes and sequences involved. Thus, abnormalities of gene-rich chromosomes or chromosomal regions tend to be much more severe clinically than similar-sized defects involving gene-poor parts of the genome.

As the Human Genome Project nears completion, it is apparent that the organization of DNA in the human genome is far more complex than was anticipated, as illustrated in Figure 3–13 for the fully characterized region of chromosome 17 in the vicinity of the *BRCA1* gene, mutations in which are responsible for some forms of familial breast cancer (see Chapter 16). Of the DNA in the genome, less than 10 percent actually encodes genes. Only about one half to three quarters of the total linear length of the genome consists of so-called **single-copy** or **unique DNA**—that is, DNA whose nucleotide sequence is represented only once (or at most a few times) per haploid genome. The rest of the genome consists of several classes of **repetitive DNA** and includes DNA whose nucleotide sequence is repeated, either perfectly or with some variation, hundreds to millions of times in the genome. Whereas most (but not all) of the estimated 50,000 genes in the genome are represented in single-copy DNA, sequences in the repetitive DNA fraction contribute to maintaining chromosome structure.

Single-Copy DNA Sequences

Although single-copy DNA makes up most of the DNA in the genome, much of its function remains a mystery because, as mentioned, sequences actually encoding proteins (i.e., the coding portion of genes) constitute only a small proportion of all the single-

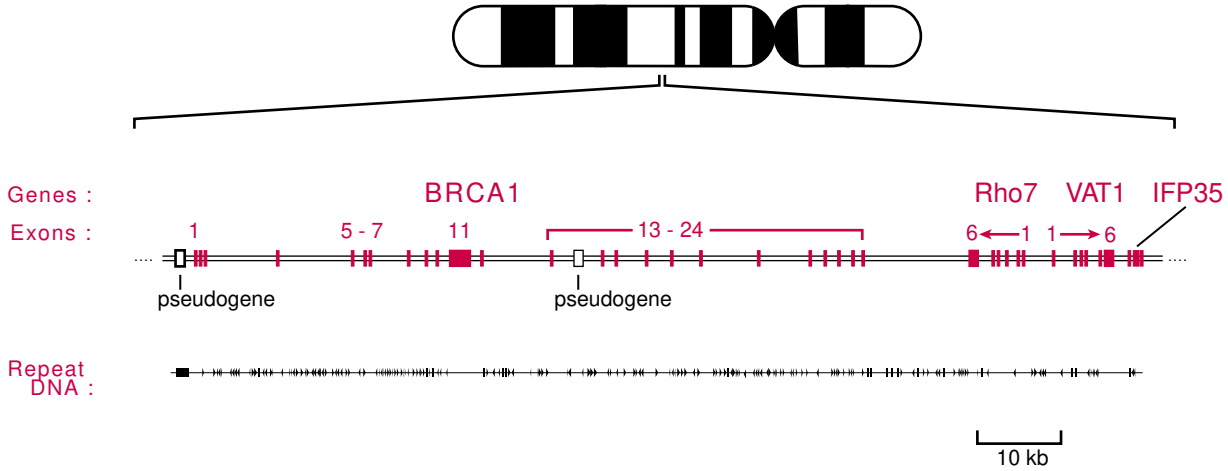


Figure 3–13. Sequence organization of the region of the human genome near the *BRCA1* gene that is responsible for some cases of familial, early-onset breast cancer. Four genes and two pseudogenes are found in this 117-kb sequence from chromosome 17, and their locations and structures are indicated above the chromosome. Nearly half of the sequence consists of various interspersed repetitive elements, including 170 copies of *Alu* repeats and 8 copies of L1 repeats that are indicated below the chromosome. *BRCA1* contains a GC-rich promoter, with a number of specific transcription-factor binding sites. The *RHO7* gene encodes a member of a GTP-binding protein family. *VAT1* encodes a membrane protein found in synaptic vesicles. Both *RHO7* and *VAT1* also have GC-rich promoters. *IFP35* encodes a protein whose expression can be induced by interferon. (Based on Smith TM, Lee MK, Szabo CI et al [1996] Complete genomic sequence and analysis of 117 kb human DNA containing the gene *BRCA1*. Genome Research 6:1029–1049.)

copy DNA. Long stretches of unique DNA sequences (> 25 kb) are quite rare in the genome. Most single-copy DNA is found in short stretches (several kb or less), interspersed with members of various repetitive DNA families (see Fig. 3–13).

Repetitive DNA Families

Several different categories of repetitive DNA are recognized. A useful distinguishing feature is whether the repeated sequences ("repeats") are clustered in one or a few locations or whether they are dispersed throughout the genome, interspersed with single-copy sequences along the chromosome. Clustered repeated sequences constitute an estimated 10 to 15 percent of the genome and consist of arrays of various short repeats organized tandemly in a head-to-tail fashion. The different types of such tandem repeats are collectively called **satellite DNAs**, so named because many of the original tandem repeat families could be purified by density centrifugation from the rest of the genome as "satellite" fractions of DNA.

Satellite DNA families vary with regard to their location in the genome, the total length of the tandem array, and the length of the constituent repeat units that make up the array. In general, satellite arrays can stretch several million base pairs or more in length and constitute up to several percent of the DNA content of an individual human chromosome. Many satellite sequences are important as molecular tools that have revolutionized clinical cytogenetic analysis because of their relative ease of detection (see Chapter 9). Some human satellite sequences are based on repetitions (with some variation) of a short sequence such as a pentanucleotide. Long arrays of such repeats are found in heterochromatic regions on the proximal long arms of chromosomes 1, 9, and 16 and on nearly the entire long arm of the Y chromosome (see Chapter 9). Other satellite DNAs are based on somewhat longer basic repeats. For example, the α-satellite family of DNA is composed of tandem arrays of different copies of an approximately 171 base-pair unit, found at the centromeric region of each human chromosome. This repeat family is believed to play a role in centromere function, ensuring proper chromosome segregation in mitosis and meiosis.

In addition to satellite DNAs, another major class of repetitive DNA in the genome consists of related sequences that are dispersed throughout the genome rather than localized (see Fig. 3–13). Although many small DNA families meet this general description, two in particular warrant discussion because together they make up a significant proportion of the genome and because they have been implicated in genetic diseases. The best-studied dispersed repetitive elements belong to the so-called ***Alu* family**. The members of this family are about 300 base pairs in length and are recognizably related to each other although not identical in sequence. In total, there are about 500,000 *Alu* family members in the genome, making up at least several percent of human DNA. In some regions of the genome, however, including near the *BRCA1* gene as seen in Figure 3–13, they make up a much higher percentage of the DNA. A second major dispersed, repetitive DNA family is called the **L1 family**. L1 elements are long, repetitive sequences (up to 6 kb in length) that are found in about 100,000 copies per genome. They are plentiful in some regions of the genome but relatively sparse in others.

Families of repeats dispersed throughout the genome are clearly of medical importance. Both *Alu* and L1 sequences have been implicated as the cause of mutations in hereditary disease through the process of retrotransposition, introduced in an earlier section. At least a few copies of the L1 and *Alu* families are still transpositionally active and generate copies of themselves that can integrate elsewhere in the genome, occasionally causing insertional inactivation of a medically important gene. The frequency of retrotransposition events causing genetic disease in humans is unknown currently, but they may account for as many as 1 in 500 mutations. In addition, aberrant recombination events between different copies of dispersed repeats can also be a cause of mutation in some genetic diseases (see Chapter 6).

VARIATION IN GENE EXPRESSION AND ITS RELEVANCE TO MEDICINE

The regulated expression of the estimated 50,000 genes encoded in human chromosomes involves a set of complex interrelationships among different levels of control, including proper gene dosage (controlled by mechanisms of chromosome replication and segregation), gene structure, and, finally, transcription, mRNA stability, translation, protein processing, and protein degradation. For some genes, fluctuations in the level of functional gene product, due either to inherited variation in the structure of a particular gene or to changes induced by nongenetic factors such as diet or the environment, are of relatively little importance. For other genes, changes in the level of expression can have dire clinical consequences, reflecting the importance of those gene products in particular biological pathways. The nature of inherited variation in the structure and function of chromosomes and genes, and the influence of this variation on the

expression of specific traits, is the very essence of medical and molecular genetics and is dealt with in subsequent chapters.

General References

Abel T, Maniatis T (1994) Mechanisms of eukaryotic gene regulation. *In* Stamatoyannopoulos G, Nienhuis AW, Majerus PW, Varmus H, (eds) The Molecular Basis of Blood Diseases. WB Saunders, Philadelphia. pp. 33–70.

Alberts B, Bray D, Lewis J, et al. (1994) Molecular Biology of the Cell, 3rd ed. Garland Publishing, New York.

Bernardi G (1995) The human genome, organization, and evolutionary history. Ann Rev Genet 29:445–476.

Lewin B (2000) Genes VII, 7th ed. Oxford University Press, Oxford, England.

Semenza G (1999) Transcription Factors and Human Disease. Oxford University Press, New York.

Singer M, Berg P (1997) Exploring Genetic Mechanisms. University Science Books, Sausalito, California.

Wolffe A (1998) Chromatin Structure and Function, 3rd ed. Academic Press, San Diego.

References Specific to Particular Topics

Berg P (1981) Dissections and reconstructions of genes and chromosomes (Nobel Prize lecture). Science 213:296–303.

Kazazian HH, Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nature Genetics 19:19–24.

Lawn RM, Efstratiadis A, O'Connell C, et al (1980) The nucleotide sequence of the human β-globin gene. Cell 21:647–651.

Smith TM, Lee MK, Szabo CI, et al (1996) Complete genomic sequence and analysis of 117 kb of human DNA containing the gene *BRCA1*. Genome Research 6:1029–1049.

Wallace DC (1999) Mitochondrial diseases in man and mouse. Science 283:1482–1488.

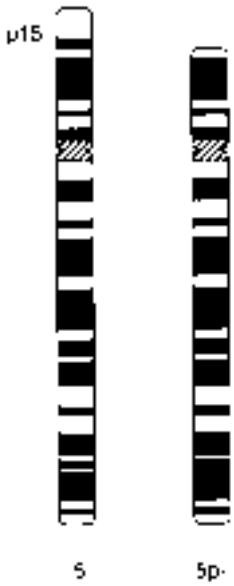
Problems

1. The following amino acid sequence represents part of a protein. The normal sequence and four mutant forms are shown. By consulting Table 3–1, determine the double-stranded sequence of the corresponding section of the normal gene. Which strand is the strand that RNA polymerase "reads"? What would the sequence of the resulting mRNA be? What kind of mutation is each mutant protein most likely to represent?

   Normal -lys-arg-his-his-tyr-leu-
   Mutant 1 -lys-arg-his-his-cys-leu-
   Mutant 2 -lys-arg-ile-ile-ile-
   Mutant 3 -lys-glu-thr-ser-leu-ser-
   Mutant 4 -asn-tyr-leu-

2. The following items are related to each other in a hierarchical fashion. What are these relationships? Chromosome, base pair, nucleosome, G-band, kb pair, intron, gene, exon, chromatin, codon, nucleotide.

3. The following schematic drawing illustrates a chromosome 5 in which the most distal band (band p15) on the short arm is deleted. This deleted chromosome is associated with the cri du chat syndrome (see Chapters 9 and 10). Given what you know about the organization of chromosomes and the genome, approximately how much DNA is deleted? How many genes?



4. Most of the human genome consists of sequences that are not transcribed and do not directly encode gene products. For each of the following, consider ways in which these genome elements can contribute to human disease: introns, *Alu* or L1 repetitive sequences, locus control regions, pseudogenes.