Лекция 9.
Геном, плазмиды, вирусы

CHAPTER

Genes and Chromosomes

Every cell of a multicellular organism generally contains the same genetic material. One has only to look at a human being to marvel at the wealth of information contained in each human cell. It should come as no surprise that the DNA molecules containing the cellular genes are by far the largest macromolecules in cells. They are commonly packaged into structures called **chromosomes.** Most bacteria and viruses have a single chromosome; eukaryotes usually have many. A single chromosome typically contains thousands of individual genes. The sum of all the genes and intergenic DNA on all the different chromosomes of a cell is referred to as the cellular **genome.**

Measurements carried out in the 1950s indicated that the largest DNAs had molecular weights of 10^6 or less, equivalent to about 15,000 base pairs. But with improved methods for isolation of native DNAs, their molecular weights were found to be much higher. Today we know that native DNA molecules, such as those from *E. coli* cells, are so large that they are easily broken by mechanical shear forces, and therefore are not readily isolated in intact form.

The size of DNA molecules represents an interesting biological problem in itself. Chromosomal DNAs are often many orders of magnitude longer than the biological packages (cells or viruses) that contain them (Fig. 23–1). In this chapter we move from the secondary structure of DNA considered in Chapter 12 to the extraordinary degree of organization required for the tertiary packaging of DNA into chromosomes. First we examine the size of viral DNAs and cellular chromosomes and the organization of genes and other sequences within them. We then turn to the discipline of DNA topology to give formal definition to the twisting and coiling of DNA molecules. Finally, we consider the protein–DNA interactions that organize chromosomes into compact structures.

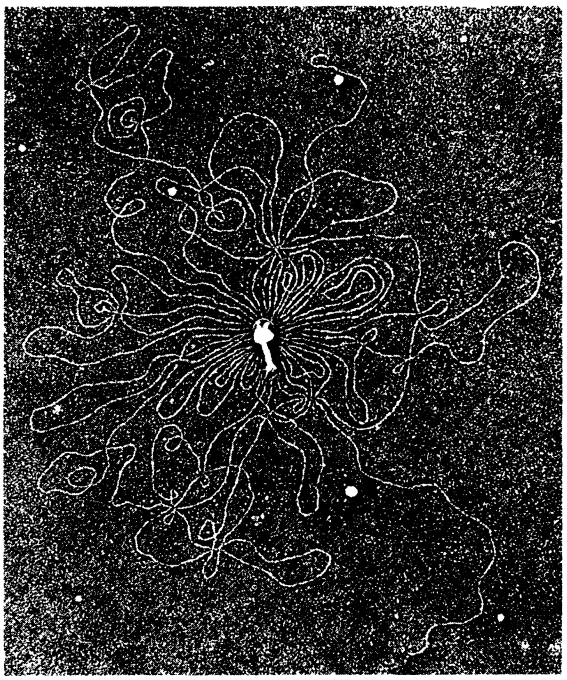
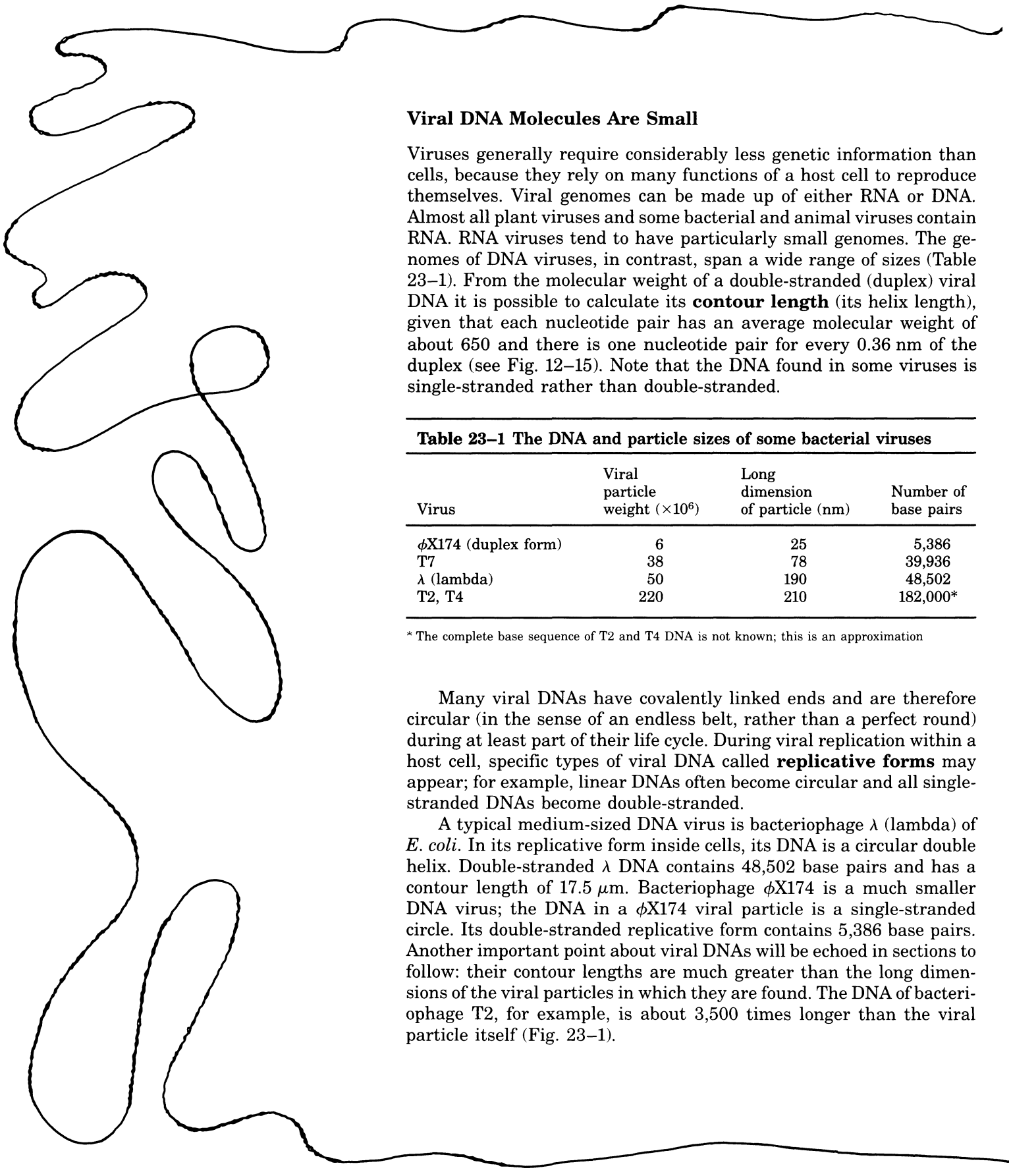


Figure 23–1 Electron micrograph of bacteriophage T2 surrounded by its single, linear molecule of DNA. The DNA was released by lysing the bacteriophage in distilled water and allowing the DNA to spread on the water surface.

The Size and Sequence Structure of DNA Molecules

We begin with a survey of the DNA molecules of viruses and of cells, both prokaryotic and eukaryotic. Chromosomes contain, in addition to genes, special-function sequences that aid in the packaging and segregation of chromosomes to daughter cells at cell division. The structure of chromosomes will be examined, with a focus on the various types of DNA sequences found within them.



Viral DNA Molecules Are Small

Viruses generally require considerably less genetic information than cells, because they rely on many functions of a host cell to reproduce themselves. Viral genomes can be made up of either RNA or DNA. Almost all plant viruses and some bacterial and animal viruses contain RNA. RNA viruses tend to have particularly small genomes. The genomes of DNA viruses, in contrast, span a wide range of sizes (Table 23–1). From the molecular weight of a double-stranded (duplex) viral DNA it is possible to calculate its **contour length** (its helix length), given that each nucleotide pair has an average molecular weight of about 650 and there is one nucleotide pair for every 0.36 nm of the duplex (see Fig. 12–15). Note that the DNA found in some viruses is single-stranded rather than double-stranded.

Table 23–1 The DNA and particle sizes of some bacterial viruses

Virus	Viral particle weight ($\times 10^6$)	Long dimension of particle (nm)	Number of base pairs
ϕX174 (duplex form)	6	25	5,386
T7	38	78	39,936
λ (lambda)	50	190	48,502
T2, T4	220	210	182,000*

* The complete base sequence of T2 and T4 DNA is not known; this is an approximation

Many viral DNAs have covalently linked ends and are therefore circular (in the sense of an endless belt, rather than a perfect round) during at least part of their life cycle. During viral replication within a host cell, specific types of viral DNA called **replicative forms** may appear; for example, linear DNAs often become circular and all single-stranded DNAs become double-stranded.

A typical medium-sized DNA virus is bacteriophage λ (lambda) of *E. coli*. In its replicative form inside cells, its DNA is a circular double helix. Double-stranded λ DNA contains 48,502 base pairs and has a contour length of 17.5 μm. Bacteriophage ϕX174 is a much smaller DNA virus; the DNA in a ϕX174 viral particle is a single-stranded circle. Its double-stranded replicative form contains 5,386 base pairs. Another important point about viral DNAs will be echoed in sections to follow: their contour lengths are much greater than the long dimensions of the viral particles in which they are found. The DNA of bacteriophage T2, for example, is about 3,500 times longer than the viral particle itself (Fig. 23–1).

Bacteria Contain Chromosomes and Extrachromosomal DNA

Bacteria contain much more DNA than the DNA viruses. For example, a single *E. coli* cell contains almost 200 times as much DNA as a bacteriophage λ particle. The DNA in an *E. coli* cell is a single, covalently closed double-stranded circular molecule. It contains about 4.7×10^6 base pairs and has a contour length of about 1.7 mm, some 850 times the length of an *E. coli* cell (Fig. 23–2). Again, the DNA molecule must have a tightly compacted tertiary structure.

In addition to the very large, circular DNA chromosome found in the nucleoid, many species of bacteria contain one or more small, circular DNA molecules that are free in the cytosol. These extrachromosomal elements are called **plasmids** (Fig. 23–3). Many plasmids are only a few thousand base pairs long, but some contain over 10^5 base pairs. Plasmids carry genetic information and undergo replication to yield daughter plasmids, which pass into the daughter cells at cell division. Ordinarily, plasmids exist separately, detached from the chromosomal DNA. A few classes of plasmid DNAs are sometimes inserted into the chromosomal DNA and later excised in a precise manner by means of specialized recombination processes.

Figure 23–3 Electron micrograph of DNA from a lysed *E. coli* cell. Several small, circular plasmid DNAs are indicated by arrows. The black spots and white specks are artifacts of the preparation.

Figure 23–2 The length of the *E. coli* chromosome (1.7 mm) is depicted relative to the length of a typical *E. coli* cell (2 μm).

Plasmids have been found in yeast and other fungi as well as in bacteria. In many cases plasmids confer no obvious advantage on their host, and their sole function appears to be self-propagation. However, some plasmids carry genes that make a host bacterium resistant to antibacterial agents. For example, plasmids carrying the gene for the enzyme β-lactamase confer resistance to β-lactam antibiotics such as penicillin and amoxicillin. Plasmids also may pass from an antibiotic-resistant cell to an antibiotic-sensitive cell of the same or another bacterial species, thus rendering the latter resistant. The extensive use of antibiotics has served as a strong selective force for the spread of these plasmids in disease-causing bacteria, creating multiply resistant bacterial strains, particularly in hospital settings. Physicians are becoming reluctant to prescribe antibiotics unless a bacterial infection is confirmed. For similar reasons, the widespread use of antibiotics in animal feeds is being curbed.

Plasmids are useful models for the study of many processes in DNA metabolism. They are relatively small DNA molecules and hence can quite easily be isolated intact from bacterial and yeast cells. Plasmids have also become a central component of the modern technologies associated with the isolation and cloning of genes. Genes from a variety of species can be inserted into isolated plasmids, and the modified plasmid can then be reintroduced into its normal host cell. Such a plasmid will be replicated and transcribed, and may also cause the host cell to make the proteins coded by the foreign gene, even though it is not part of the normal genome of the cell. Chapter 28 describes how such **recombinant DNAs** are made.

Eukaryotic Cells Contain More DNA than Prokaryotes

An individual cell of a yeast, one of the simplest eukaryotes, has four times more DNA than an *E. coli* cell. Cells of *Drosophila,* the fruit fly used in classical genetic studies, have more than 25 times as much DNA as *E. coli* cells. Each cell of human beings and many other mammals has about 600 times as much DNA as *E. coli,* and the cells of many plants and amphibians have an even greater amount. Note that the nuclear DNA molecules of eukaryotic cells are linear, not circular.

The total contour length of all the DNA in a *single* human cell is about 2 m, compared with 1.7 mm for *E. coli* DNA. In the approximately 10^{14} cells of the adult human body, the total length of all the DNA would be about 2×10^{14} m or 2×10^{11} km. Compare this with the circumference of the earth (4×10^4 km) or the distance between the earth and the sun (1.5×10^8 km). Once again it becomes clear that DNA packaging in cells must involve an extraordinary degree of organization and compaction.

Microscopic observation of nuclei in dividing eukaryotic cells has shown that the genetic material is subdivided into chromosomes, their diploid number depending upon the species of organism (Table 23–2). Human cells, for example, have 46 chromosomes. Each chromosome of a eukaryotic cell, such as that shown in Figure 23–4a, can contain a single, very large, duplex DNA molecule, which may be from 4 to 100 times larger than that of an *E. coli* cell. For example, the DNA of one of the smaller human chromosomes has a contour length of about 30 mm, almost 15 times longer than the DNA of *E. coli.* The DNA molecules in the 24 different types of chromosomes of human cells (22 + X + Y) vary in length over a 25-fold range. Each different chromosome in eukaryotes carries a characteristic set of genes.

Table 23–2 Normal chromosome number in different organisms*

Bacteria	1
Fruit fly	8
Red clover	14
Garden pea	14
Yeast	16
Honeybee	16
Corn	20
Frog	26
Hydra	30
Fox	34
Cat	38
Mouse	40
Rat	42
Rabbit	44
Human	46
Chicken	78

* For all eukaryotic organisms listed, the diploid chromosome number is shown

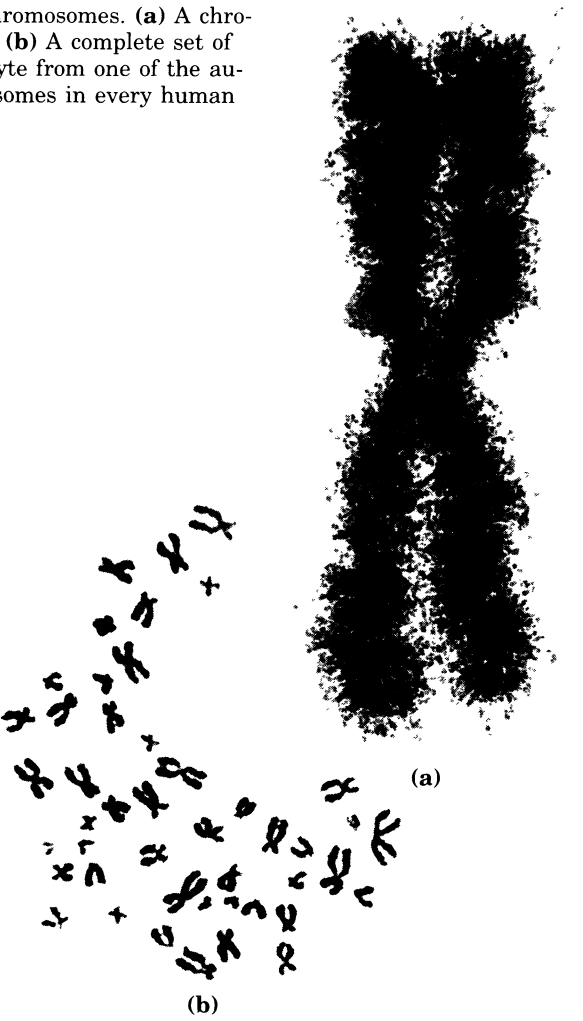


Figure 23–4 Eukaryotic chromosomes. **(a)** A chromosome from a human cell. **(b)** A complete set of chromosomes from a leukocyte from one of the authors. There are 46 chromosomes in every human somatic cell.

Organelles of Eukaryotic Cells Also Contain DNA

In addition to the DNA in the nucleus of eukaryotic cells, very small amounts of DNA, differing in base sequence from nuclear DNA, are present within the mitochondria. Chloroplasts of photosynthetic cells also contain DNA. Usually less than 0.1% of all the cell DNA is present in the mitochondria in typical somatic cells, but in fertilized and dividing egg cells, where the mitochondria are much more numerous, the total amount of mitochondrial DNA is correspondingly larger. Mitochondrial DNA (mDNA) is a very small molecule compared with the nuclear chromosomes. In animal cells it contains less than 20,000 base pairs (16,569 base pairs in human mDNA) and occurs as a circular duplex. Chloroplast DNA molecules also exist as circular duplexes and are considerably larger than those of mitochondria.

The evolutionary origin of mitochondrial and chloroplast DNAs has been the subject of much speculation. A widely accepted view is that they are vestiges of the chromosomes of ancient bacteria that gained access to the cytoplasm of host cells and became the precursors of these organelles (see Fig. 2–17). Mitochondrial DNA codes for the mitochondrial tRNAs and rRNAs and for a few mitochondrial proteins. More than 95% of mitochondrial proteins are encoded by nuclear DNA. Mitochondria and chloroplasts divide when the cell divides (Fig. 23–5). Before and during division of these organelles their DNA is replicated and the daughter DNA molecules pass into the daughter organelles.

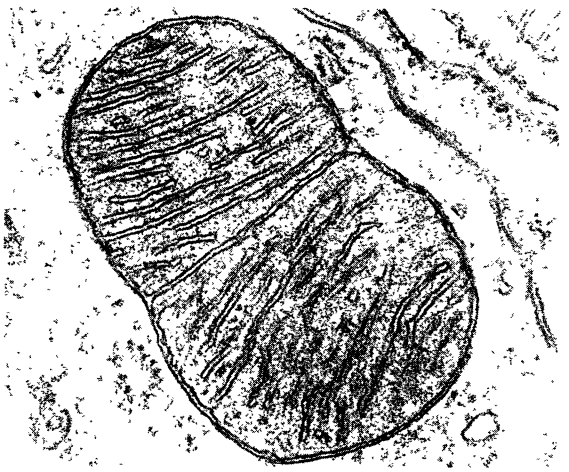


Figure 23–5 A dividing mitochondrion. Many mitochondrial proteins and RNAs are encoded by the mitochondrial DNA (not visible here), which is replicated each time the mitochondrion divides.

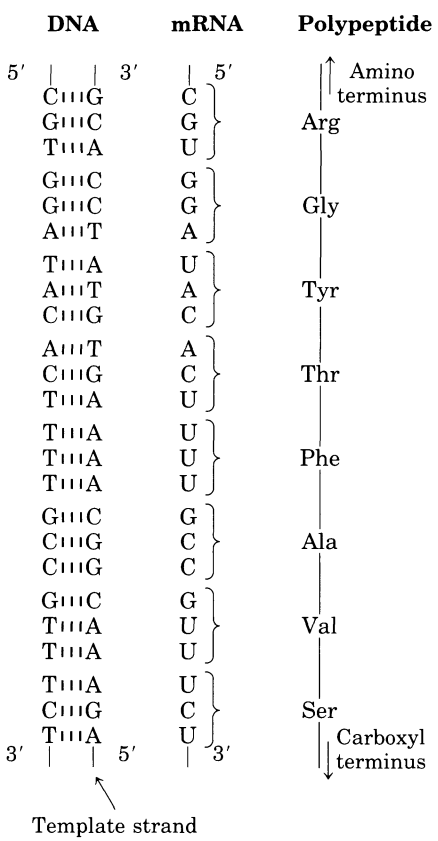


Figure 23–6 Colinearity of the nucleotide sequences of DNA, mRNA, and the amino acid sequence of polypeptide chains. The triplets of nucleotide units in DNA determine the sequence of amino acids in proteins through the intermediary formation of mRNA, which has nucleotide triplets (codons) complementary to those of the DNA. Only one of the DNA strands, the template strand, serves as a template for mRNA synthesis.

Genes Are Segments of DNA That Code for Polypeptide Chains and RNAs

Our present understanding of the gene has evolved considerably over the last century. A gene is defined in the classical biological sense as a portion of a chromosome that determines or affects a single character or **phenotype** (visible property), for example, eye color. But there is also a molecular definition, first proposed by George Beadle and Edward Tatum in 1940. They exposed spores of the mold *Neurospora crassa* to x rays and other agents that damage DNA and sometimes cause alterations in the DNA sequence **(mutations).** Some mutants were found to be deficient in one or another specific enzyme, resulting in the failure of a metabolic pathway. This observation led Beadle and Tatum to conclude that a gene is a segment of the genetic material that determines or codes for one enzyme: the **one gene–one enzyme** hypothesis. Later this concept was broadened to **one gene–one protein,** because many genes code for proteins that are not enzymes.

The present biochemical definition of a gene is somewhat more precise. Recall that many proteins have multiple polypeptide chains (Chapter 6). In some multichain proteins, all the polypeptide chains are identical, in which case they can all be encoded by the same gene. Others have two or more different kinds of polypeptide chains, each with a distinctive amino acid sequence. Hemoglobin A, the major adult hemoglobin of humans, for example, has two kinds of polypeptide chains, α and β chains, which differ in amino acid sequence and are encoded by two different genes. Thus the gene–protein relationship is more accurately described by the phrase "one gene–one polypeptide."

However, not all genes are ultimately expressed in the form of polypeptide chains. Some genes code for the different kinds of RNAs such as tRNAs and rRNAs (Chapters 12 and 25). Genes that code for either polypeptides or RNAs are known as **structural genes:** they encode the primary sequence of some final gene product, such as an enzyme or a stable RNA. DNA also contains other segments or sequences that have a purely regulatory function. **Regulatory sequences** provide signals that may denote the beginning and end of structural genes, or participate in turning on or off the transcription of structural genes, or function as initiation points for replication or recombination (Chapter 27).

The minimum overall size of genes can be estimated directly. As will be described in detail in Chapter 26, each amino acid of a polypeptide chain is coded by a sequence of three consecutive nucleotides in a single strand of DNA (Fig. 23–6). Because there are no signals for "commas" in the genetic code, the coding triplets of DNA are generally arranged sequentially, corresponding to the sequence of amino acids in the polypeptide for which it codes. Figure 23–6 shows the principle of the coding relationships between DNA, RNA, and proteins. A single polypeptide chain may have anywhere from about fifty to several thousand amino acid residues in a specific sequence, thus a gene coding for the biosynthesis of a polypeptide chain must have, correspondingly, at least 150 to 6,000 or more base pairs. For an average polypeptide chain of 350 amino acid residues, this would correspond to 1,050 base pairs. We will see later that many genes in eukaryotes and a few in prokaryotes are interrupted by noncoding DNA segments called introns, and can therefore be considerably longer than the simple calculations outlined above would suggest.

There Are Many Genes in a Single Chromosome

How many genes are in a single chromosome? We can give an approximate answer to this question in the case of *E. coli*. If the average gene is 1,050 base pairs long, the 4.7 million base pairs in the *E. coli* chromosome could accommodate about 4,400 genes. The products of over 1,000 *E. coli* genes have already been characterized, and the number is increasing. A growing fraction of the *E. coli* chromosome has been sequenced, and the number of genes it contains will be known with some precision when this effort is completed.

Eukaryotic Chromosomes Are Very Complex

Bacteria usually have only one chromosome per cell, and in nearly all cases each chromosome contains only one copy of any given gene. A very few genes, such as those for rRNAs, are repeated several times. Regulatory and structural gene sequences account for much of the DNA in prokaryotes. Moreover, almost every gene is precisely colinear with the amino acid sequence (or RNA sequence) for which it codes (Fig. 23–6).

The organization of genes in eukaryotic DNA is structurally and functionally much more complex, and the study of eukaryotic chromosome structure has yielded many surprises. Tests made of the extent to which segments of mouse DNA occur in multiple copies had an unexpected outcome. About 10% of mouse DNA consists of short lengths of less than 10 base pairs that are repeated millions of times per cell. These are called **highly repetitive** segments. Another 20% of mouse DNA was found to occur in lengths up to a few hundred base pairs that are repeated at least 1,000 times, designated **moderately repetitive.** The remainder, some 70% of the DNA, consists of unique segments and segments that are repeated only a few times.

Some of the repetitive DNA may simply be "junk DNA," vestiges of evolutionary sidetracks. At least some of it has functional significance, however. The most highly repeated sequences are called **satellite DNA** because their base compositions are generally unusual, permitting their separation from the rest of the DNA when fragmented cellular DNA samples are centrifuged in cesium chloride density gradients. Satellite DNA is not believed to encode proteins or RNAs. Much of the highly repetitive DNA is associated with two important structures in eukaryotic chromosomes—centromeres and telomeres.

Each chromosome has a single **centromere,** which functions as an attachment point for proteins that link the chromosome to the microtubules of the mitotic spindle (see Fig. 2–14). This attachment is essential for the ordered segregation of chromosomes to daughter cells during cell division. The centromeres of yeast chromosomes have been isolated and studied (Fig. 23–7). The sequences essential to centromere function are about 130 base pairs long and are very rich in A=T pairs. The centromeres of higher eukaryotes are much larger. In higher eukaryotes (but not in yeast), satellite DNA is generally found in the centromeric region and consists of thousands of tandem (side-by-side and in the same orientation) copies of one or a few short sequences. Characterized satellite sequences are generally 5 to 10 base pairs long. The precise role of satellite DNA in centromere function is not yet understood.

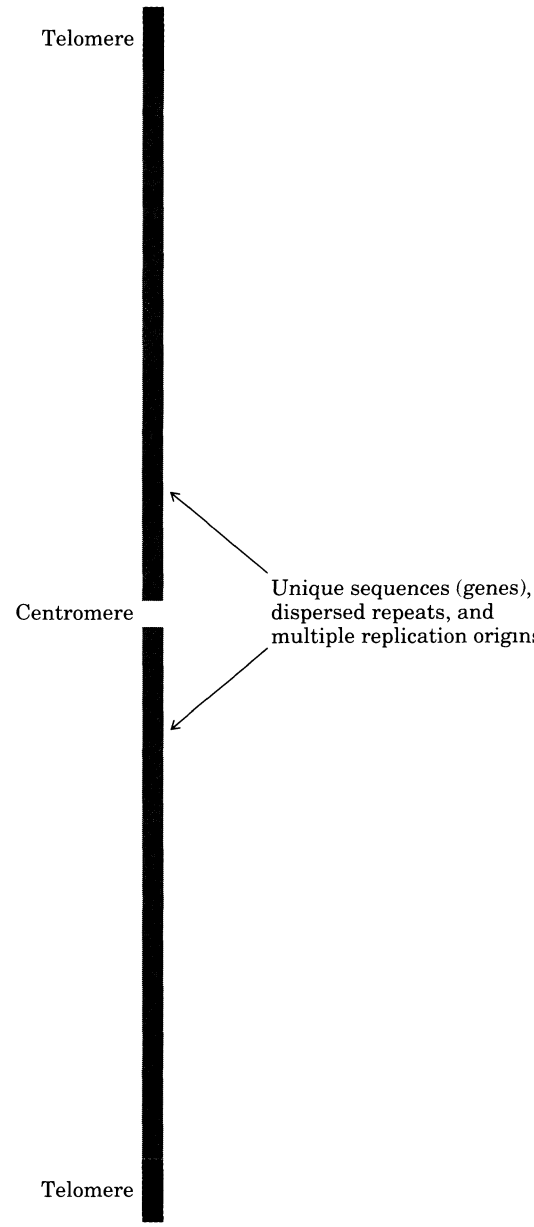


Figure 23–7 Important structural features of a yeast chromosome.

Telomeres are sequences located at the ends of the linear eukaryotic chromosomes, which help stabilize them. The best-characterized telomeres are those of simpler eukaryotes. Yeast telomeres end with about 100 base pairs of imprecisely repeated sequences of the form

$$(5')(T_xG_y)_n$$
$$(3')(A_xC_y)_n$$

where x and y generally fall in the range of 1 to 4. The ends of a linear DNA molecule cannot be replicated by the cellular replication machinery (which may be one reason why bacterial DNA molecules are circular). The repeated sequences in telomeres are added to chromosome ends by special enzymes, one of which is telomerase, which will be discussed in more detail in Chapter 25. What controls the number of repeats in a telomere is not known. The telomere repeats are a very unusual DNA structure.

Efforts have begun to construct artificial chromosomes as a means of better understanding the functional significance of many structural features of eukaryotic chromosomes. A reasonably stable, artificial, linear chromosome requires only three components: a centromere, telomeres at the ends, and sequences that direct the initiation of DNA replication.

Most moderately repetitive DNA consists of 150 to 300 base-pair repeats scattered throughout the genome of higher eukaryotes. Some of these repeats have been characterized. A number of them have some of the structural properties of transposable elements, sequences that move about the genome at very low frequency (Chapter 24). In humans, one class of these repeats (about 300 base pairs long) is called the *Alu* family, so named because their sequence generally includes one copy of the recognition sequence for the restriction endonuclease *Alu*I. (Restriction endonucleases are described in Chapter 28.) Hundreds of thousands of *Alu* repeats occur in the human genome, comprising 1 to 3% of the total DNA. They apparently were derived from a gene for 7SL RNA, a component of a complex called the signal-recognition particle (SRP, Chapter 26) that functions in protein synthesis. The *Alu* repeats, however, lack parts of the 7SL RNA gene sequence and do not produce functional 7SL RNAs. When *Alu* repeats are grouped with other classes of repeats with similar sizes and sequence structures, they make up 5 to 10% of the DNA in the human genome. No function for this DNA is known.

The unique sequences in eukaryotic chromosomes include most of the genes. There are an estimated 100,000 different genes in the human genome.

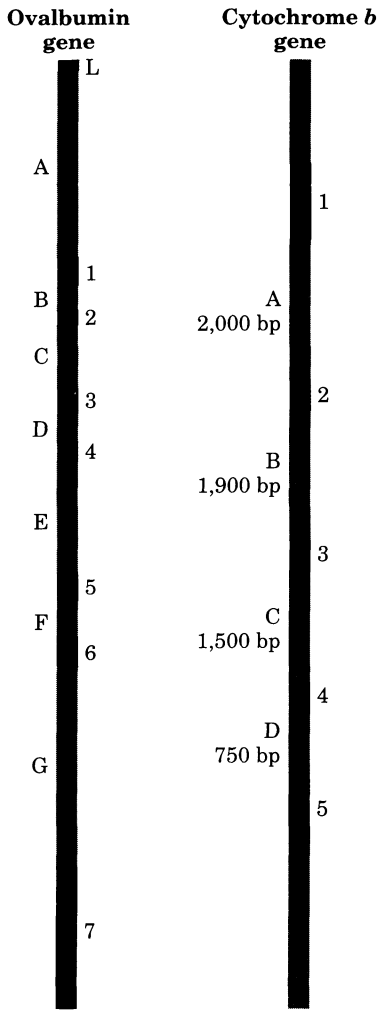


Figure 23–8 Intervening sequences, or introns, in two eukaryotic genes. The gene for ovalbumin has seven introns (A to G), splitting the coding sequences into eight exons (L, 1 to 7). The gene for cytochrome *b* has four introns and five exons. In both cases, more DNA is devoted to introns than to exons. The number of base pairs (bp) in the introns of the cytochrome *b* gene is shown.

Many Eukaryotic Genes Contain Intervening Nontranscribed Sequences (Introns)

Many, if not most, eukaryotic genes have a distinctive and puzzling structural feature: their nucleotide sequences contain one or more intervening segments of DNA that do not code for the amino acid sequence of the polypeptide product. These nontranslated inserts interrupt the otherwise precisely colinear relationship between the nucleotide sequence of the gene and the amino acid sequence of the polypeptide it encodes (Fig. 23–8). Such nontranslated DNA segments in genes are called **intervening sequences,** or **introns,** and the coding segments are called **exons.** A well-known example is the gene coding for the single polypeptide chain of the avian egg protein ovalbumin.

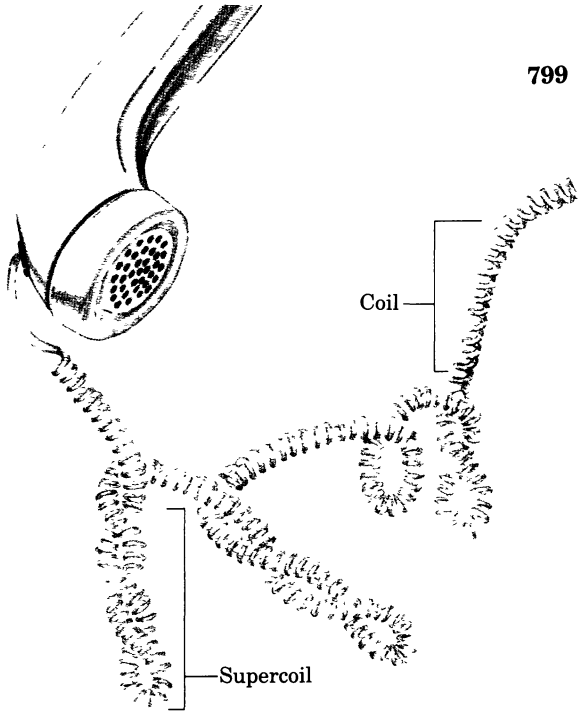


Figure 23–9 Supercoils. A typical phone cord is a coil. A phone cord twisted as shown is a supercoil. The illustration is especially appropriate, because an examination of the twisting of phone cords helped lead Jerome Vinograd and colleagues to the insight that many properties of small, circular DNAs could be explained by supercoiling. They first detected DNA supercoiling in small, circular viral DNAs in 1965.

As can be seen in Figure 23–8, the introns of this particular gene are much longer than the exons; altogether the introns make up 85% of the DNA of this gene. Most eukaryotic genes examined thus far appear to contain introns that vary in number, position, and the fraction of the total length of the gene they occupy. For example, the serum albumin gene contains 6 introns, the gene for the protein conalbumin of the chicken egg contains 17 introns, and a collagen gene has been found to have over 50 introns. Genes for histones provide an example of a family of genes that appear to have no introns. Only a few prokaryotic genes contain introns. In most cases the function of introns is not clear.

DNA Supercoiling

From the examples given above, it is clear that cellular DNA must be very tightly compacted just to fit into the cell. This implies a high degree of structural organization. It is not enough just to fold the DNA into a small space, however. The packaging must permit access to the information in the DNA for processes such as replication and transcription. Before considering how this is accomplished, we must examine an important property of DNA structure that we have not yet considered—DNA supercoiling.

The term "supercoiling" means literally the coiling of a coil. A telephone cord for example, is typically a coiled wire. The twisted path often taken by that wire as it goes from the base of the phone to the receiver generally describes a supercoil (Fig. 23–9). DNA is coiled in the form of a double helix. Let us define an axis about which both strands of the DNA coil. A bending or twisting of that axis upon itself (Fig. 23–10) is referred to as **DNA supercoiling.** As detailed below, DNA supercoiling is generally a manifestation of structural strain. Conversely, if there is no net bending of the DNA axis upon itself, the DNA is said to be in a **relaxed** state.

It is probably apparent that DNA compaction must involve some form of supercoiling. Perhaps less apparent is the fact that replicating or transcribing DNA also must induce some degree of supercoiling.

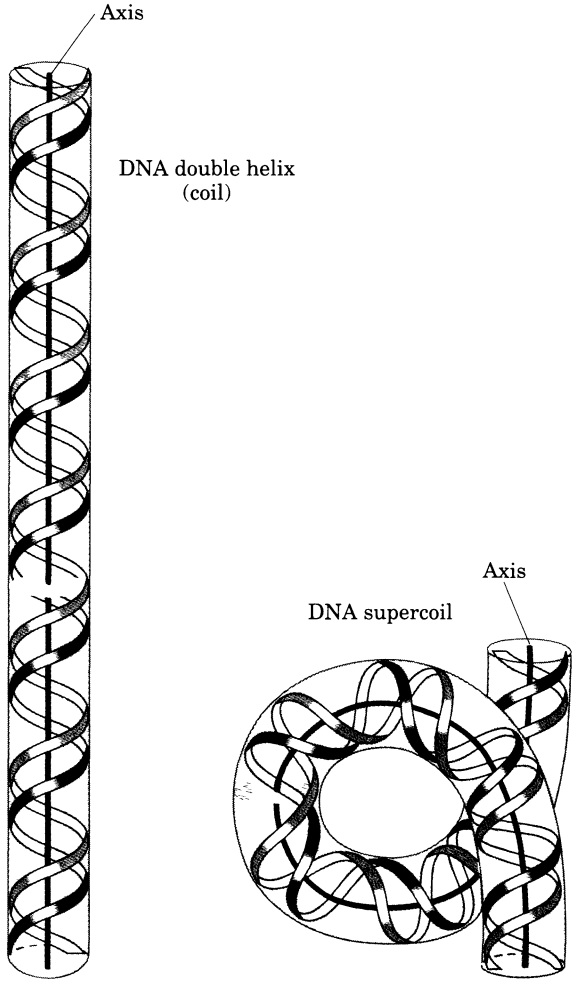


Figure 23–10 Supercoiling of DNA. Supercoiling is the twisting of the DNA axis upon itself.

The DNA molecules in chromosomes are the largest macromolecules in cells. Many smaller DNAs also occur in cells, in the form of viral DNAs, plasmids, and (in eukaryotes) mitochondrial or chloroplast DNAs. Many DNAs, especially those in bacteria, mitochondria, and chloroplasts, are circular. Viral and chromosomal DNAs have one major feature in common: they are generally much longer than the viral particles or cells in which they are packaged. The total DNA content of a eukaryotic cell is much greater than that of a bacterial cell.

Genes are segments of a chromosome that contain the information for a functional polypeptide or RNA molecule. In addition to these structural genes, chromosomes contain a variety of regulatory sequences involved in replication, transcription, and other processes. In eukaryotic chromosomes, there are two important special-function repetitive DNA sequences: centromeres, which are attachment points for the mitotic spindle, and telomeres, which occur at the ends of the linear chromosomes. Many genes in eukaryotic cells, and occasionally in bacteria, are interrupted by noncoding sequences called introns. The coding segments separated by introns are called exons.

Most cellular DNAs are supercoiled. Supercoiling is a manifestation of structural strain imparted by the underwinding of the DNA molecule. Underwinding is a decrease in the total number of helical turns in the DNA relative to the relaxed or B form. To maintain an underwound state, DNA must be a closed circle or be bound with protein. Supercoils resulting from underwinding are defined as negative supercoils. Underwinding is quantified by a topological parameter called linking number, Lk. The linking number of a relaxed, closed-circular DNA is used as a reference (Lk_0) and is equal to the number of base pairs divided by 10.5. Underwinding is measured in terms of the specific linking difference or σ, which equals $(Lk - Lk_0)/Lk_0$. For cellular DNAs, σ typically equals -0.05 to -0.07, which means that approximately 5 to 7% of the helical turns in the DNA have been removed. DNA underwinding facilitates strand separation for processes such as transcription or replication. The plectonemic supercoils in negatively supercoiled DNA in solution are right-handed, and the overall structure is narrow and extended. An alternative form called solenoidal supercoiling provides a much greater degree of compaction, and this form predominates in the cell.

DNAs that differ only in their linking number are called topoisomers. The enzymes that underwind and/or relax DNA are called topoisomerases, and they act by catalyzing changes in linking number. There are two classes, type 1 and type 2, which change Lk in increments of 1 or 2, respectively. In a bacterial cell, the superhelical density of the DNA represents a regulated balance between the activities of topoisomerases that increase and decrease linking number.

In the chromatin of eukaryotic cells, the fundamental unit of organization is the nucleosome, which consists of DNA and a protein particle containing eight histones, two copies each of histones H2A, H2B, H3, and H4. The segment of DNA (about 146 base pairs) wrapped around the protein core is in the form of a left-handed solenoidal supercoil. Nucleosomes are organized into 30 nm fibers, and the fibers themselves are extensively folded to provide the 10,000-fold compaction required to fit a typical eukaryotic chromosome into a cell nucleus. The higher-order folding involves attachment to a nuclear scaffold that contains large amounts of histone H1 and topoisomerase II. Bacterial chromosomes are also extensively compacted into a structure called a nucleoid, but the chromosome appears to be much more dynamic and irregular in structure than eukaryotic chromatin, reflecting the shorter cell cycle and very active metabolism of a bacterial cell.

Further Reading

General

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J.D. (1989) *Molecular Biology of the Cell,* 2nd edn, Garland Publishing, Inc., New York.
An excellent general reference.

Kornberg, A. & Baker, T.A. (1991) *DNA Replication,* 2nd edn, W.H. Freeman and Company, New York.
A good place to start for further information on the structure and function of DNA.

Singer, M. & Berg, P. (1991) *Genes and Genomes: A Changing Perspective,* University Science Books, Mill Valley, CA.
An up-to-date discussion of genes, chromosome structure, and many other topics.

Genes and Chromosomes

Blackburn, E.H. (1990) Telomeres: structure and synthesis. *J. Biol. Chem.* **265,** 5919–5921.

Jelinek, W.R. & Schmid, C.W. (1982) Repetitive sequences in eukaryotic DNA and their expression. *Annu. Rev. Biochem.* **51,** 813–844.

Murray, A.W. & Szostak, J.W. (1987) Artificial chromosomes. *Sci. Am.* **257** (November), 62–68.

Novick, R.P. (1980) Plasmids. *Sci. Am.* **243** (December), 102–127.

Sharp, P.A. (1985) On the origin of RNA splicing and introns. *Cell* **42,** 397–400.

Ullu, E. & Tschudi, C. (1984) *Alu* sequences are processed 7SL RNA genes. *Nature* **312,** 171–172.

Supercoiling and Topoisomerases

Bauer, W.R., Crick, F.H.C., & White, J.H. (1980) Supercoiled DNA. *Sci. Am.* **243** (July), 118–133.

Boles, T.C., White, J.H., & Cozzarelli, N.R. (1990) Structure of plectonemically supercoiled DNA. *J. Mol. Biol.* **213,** 931–951.
A study that defines several fundamental features of supercoiled DNA.

Cozzarelli, N.R., Boles, T.C., & White, J.H. (1990) Primer on the topology and geometry of DNA supercoiling. In *DNA Topology and Its Biological Effects* (Cozzarelli, N.R. & Wang, J.C., eds), pp. 139–184, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
This provides a more advanced and thorough discussion.

Lebowitz, J. (1990) Through the looking glass: the discovery of supercoiled DNA. *Trends Biochem. Sci.* **15,** 202–207.
A short and interesting historical note.

Liu, L.F. (1989) DNA topoisomerase poisons as antitumor drugs. *Annu. Rev. Biochem.* **58,** 351–375.
A review of eukaryotic topoisomerases and the use of topoisomerase inhibitors in cancer chemotherapy.

Wang, J.C. (1985) DNA topoisomerases. *Annu. Rev. Biochem.* **54,** 665–697.

Wang, J.C. (1991) DNA topoisomerases: Why so many? *J. Biol. Chem.* **266,** 6659–6662.
A good short summary of topoisomerase functions.

Chromatin and Nucleosomes

Filipski, J., Leblanc, J., Youdale, T., Sikorska, M., & Walker, P.R. (1990) Periodicity of DNA folding in higher order chromatin structures. *EMBO J.* **9,** 1319–1327.

Kornberg, R.D. (1974) Chromatin structure: a repeating unit of histones and DNA. *Science* **184,** 868–871.
The classic paper that introduced the subunit model for chromatin.

Richmond, T.J., Finch, J.T., Rushton, B., Rhodes, D., & Klug, A. (1984) Structure of the nucleosome core particle at 7Å resolution. *Nature* **311,** 532–537.

van Holde, K.E. (1989) *Chromatin,* Springer-Verlag, New York.

Problems

1. *How Long Is the Ribonuclease Gene?* What is the minimum number of nucleotide pairs in the gene for pancreatic ribonuclease (124 amino acids long)? Suggest a reason why the number of nucleotide pairs in the gene might be much larger than your answer.

2. *Packaging of DNA in a Virus* The DNA of bacteriophage T2 has a molecular weight of 120×10^6. The head of the T2 phage is about 210 nm long. Assuming the molecular weight of a nucleotide pair is 650, calculate the length of T2 DNA and compare it with the length of the T2 head. Your answer will show the necessity of very compact packaging of DNA in viruses (see Fig. 23–1).

3. *The DNA of Phage M13* Bacteriophage M13 DNA has the following base composition: A, 23%; T, 36%; G, 21%; C, 20%. What does this information tell us about the DNA of this phage?

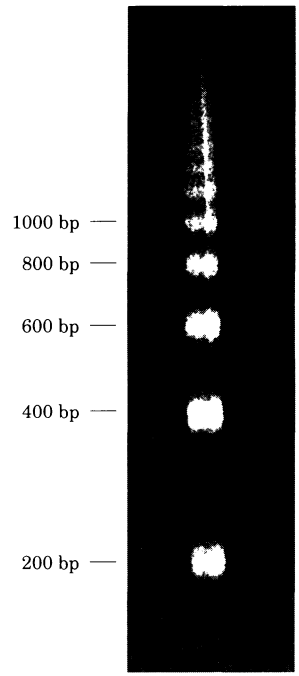
4. *Base Composition of ϕX174 DNA* Bacteriophage ϕX174 DNA occurs in two forms, single-stranded in the isolated virion and double-stranded during viral replication in the host cell. Would you expect them to have the same base composition? Give your reasons.

5. *Size of Eukaryotic Genes* An enzyme present in rat liver has a polypeptide chain of 192 amino acid residues. It is coded for by a gene having 1,440 base pairs. Explain the relationship between the number of amino acid residues in this enzyme and the number of nucleotide pairs in its gene.

6. *DNA Supercoiling* A covalently closed circular DNA molecule has an *Lk* of 500 when it is relaxed. Approximately how many base pairs are in this DNA? How will the linking number be altered (increase, decrease, no change, become undefined) if (a) a protein complex is bound to form a nucleosome, (b) one DNA strand is broken, (c) DNA gyrase is added with ATP, or (d) the double helix is denatured (base pairs are separated) by heat?

7. *DNA Structure* Explain how the underwinding of a B-DNA helix might facilitate or stabilize the formation of Z-DNA.

8. *Chromatin* One of the important early pieces of evidence that helped define the structure of the nucleosome is illustrated by the agarose gel shown below, in which the thick bands represent DNA. It was generated by treating chromatin briefly with an enzyme that degrades DNA, then removing all protein and subjecting the purified DNA to electrophoresis. Numbers at the side of the gel denote the position to which a linear DNA of the indicated size (in base pairs) would migrate. What does this gel tell you about chromatin structure? Why are the DNA bands thick and spread out rather than sharp?



Translesion replication requires DNA polymerase III, as well as the activities of the UmuC, UmuD, and RecA proteins. The mechanism by which the latter three proteins permit DNA polymerase III to replicate past DNA lesions is not understood. DNA polymerase II may also play a role in error-prone DNA repair. This polymerase is induced as part of the SOS response and, unlike DNA polymerase III, is capable of limited replication past lesions such as AP sites. This enzyme has some of the same subunits as DNA polymerase III.

The RecA protein merits some additional discussion because it has several distinct functions (besides mutagenesis) in the bacterial cell. RecA protein is involved in recombination and in the regulation of the SOS response, and in these cases its molecular function is well characterized. The regulation of the SOS response is described in Chapter 27. We now turn to a discussion of genetic recombination.

DNA Recombination

The rearrangement of genetic information in and among DNA molecules encompasses a variety of processes that are collectively placed under the heading of genetic recombination. An understanding of how DNA rearrangements occur is finding practical application as scientists explore new methods for altering the genomes of a variety of organisms (Chapter 28).

Barbara McClintock
1902–1992

Genetic recombination events fall into at least three general classes. **Homologous genetic recombination** involves genetic exchanges between any two DNA molecules (or segments of the same molecule) that share an extended region with homologous sequences. The actual sequence of bases in the DNA is irrelevant as long as the sequences in the two DNAs are similar. **Site-specific recombination** differs in that these exchanges occur only at a defined DNA sequence. **DNA transposition** is distinct in that it usually involves a short segment of DNA with the remarkable capacity to move from one location in a chromosome to another. These "hopping genes" were first observed in maize in the 1950s by Barbara McClintock. In addition to these well-characterized classes, there is a wide range of unusual rearrangements for which no mechanism or purpose has been proposed. We will focus only on the first three classes noted above.

Any discussion of the mechanics of recombination must always include unusual DNA structures. In homologous genetic recombination, the two DNA molecules interact and align their similar sequences at some stage in the reaction. This alignment process may involve the formation of novel DNA intermediates in which three or possibly even four strands are interwound. (Recall the three-stranded structure of H-DNA; see Fig. 12–22.) Branched DNA structures are also found as recombination intermediates. The exchange of information between two large, helical macromolecules often involves a complex interweaving of strands.

The functions of genetic recombination systems are as varied as their mechanisms. The maintenance of genetic diversity, specialized DNA repair systems, the regulation of expression of certain genes, and programmed genetic rearrangements during development represent some of the recognized roles for genetic recombination events. To illustrate these functions, we must first describe the recombination reactions themselves.

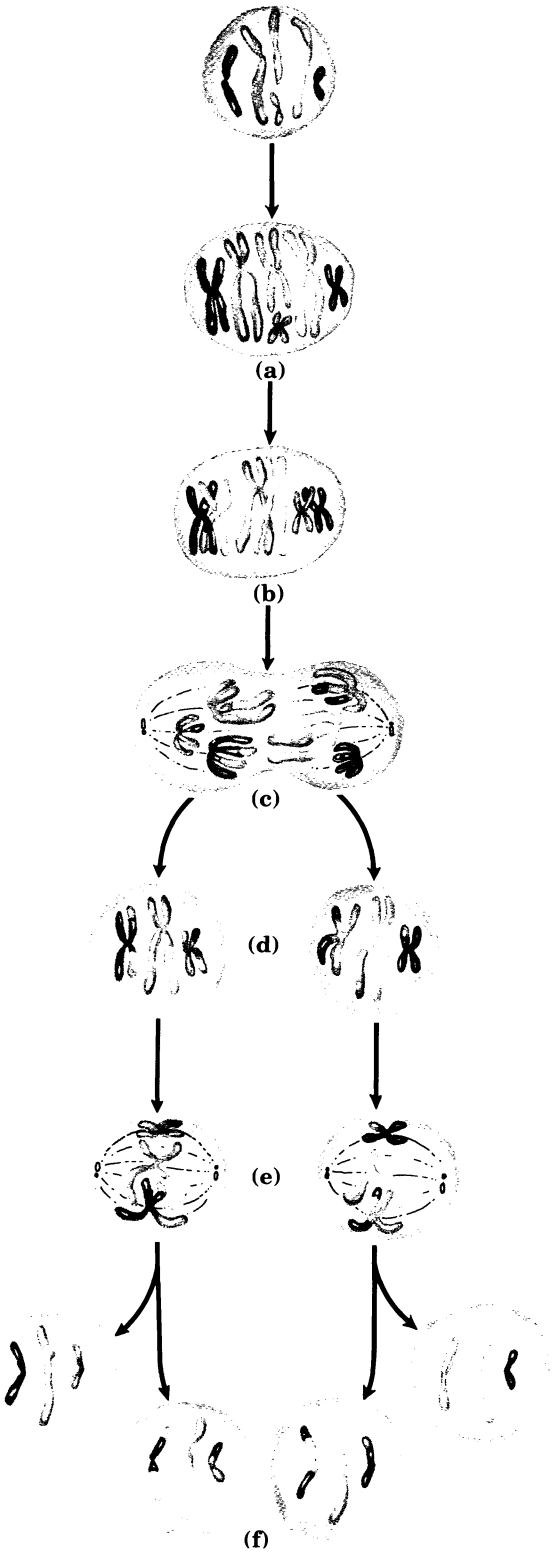
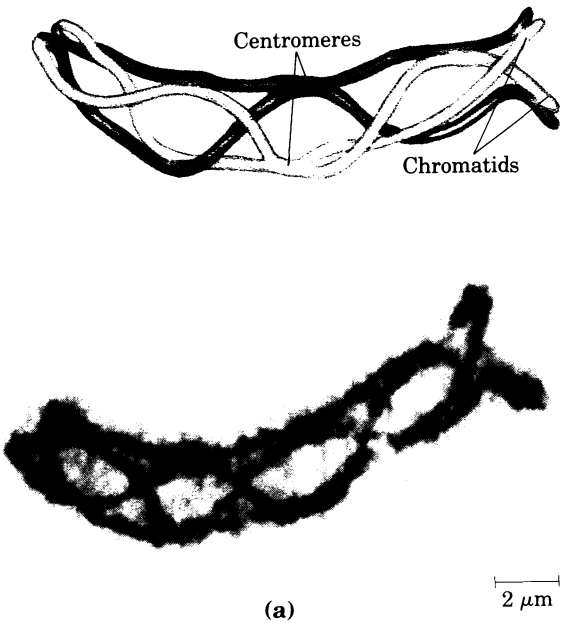


Figure 24–25 Meiosis in eukaryotic germ-line cells. **(a)** The chromosomes of a germ-line cell (six chromosomes; three homologous pairs) are replicated, except for centromeres. While the product DNA molecules remain attached at their centromeres, they are called chromatids (sometimes, "sister chromatids"). **(b)** In prophase I, just prior to the first meiotic division, the three homologous sets of chromatids are aligned to form tetrads, held together by covalent links at homologous junctions (chiasmata). Crossovers occur within the chiasmata (see Fig. 24–26). **(c)** Homologous pairs separate toward opposite poles of the cell. **(d)** The first meiotic division produces two daughter cells, each with three pairs of chromatids. **(e)** The homologous pairs align in the center of the cell in preparation for separation of the chromatids (now chromosomes). **(f)** The second meiotic division produces daughter cells with three chromosomes, half the number of the germ-line cell. The chromosomes have resorted and recombined.

Homologous Genetic Recombination Has Multiple Functions

Homologous genetic recombination (also called general recombination) is tightly linked to cell division in eukaryotes. The process occurs with the highest frequency during **meiosis,** the process in which a germ-line cell with two matching sets of chromosomes (a diploid cell) divides to produce a set of gametes—sperm cells or ova in higher eukaryotes—each gamete having only one member of each chromosome pair (haploid cells). The process of meiosis is illustrated in Figure 24–25. In outline, meiosis begins with replication of the DNA in the germ-line cell so that each DNA molecule is present in four copies. The cell then goes through two meiotic cell divisions that reduce the DNA content to the haploid level in each of four daughter cells.

After the DNA is replicated during prophase I (prophase of the first meiotic division), the resulting DNA copies remain associated at their centromeres and are referred to as sister chromatids. Each set of four homologous DNA molecules is therefore arranged as two pairs of chromatids. Genetic information is exchanged between the closely associated homologous chromatids at this stage of meiosis by means of homologous genetic recombination. This process involves a breakage and rejoining of DNA. The exchange is also called crossing over, and can be observed cytologically (Fig. 24–26). Crossing over links the two pairs of sister chromatids together at points called chiasmata (singular, chiasma). This effectively links together all four homologous chromatids, and this linkage is essential to the proper segregation of chromosomes in the subsequent meiotic cell divisions. To a first approximation, recombination, or crossing over, can occur with equal probability at almost any point along the length of two homologous chromosomes. The frequency of recombination in a region separating two points on a chromosome is therefore proportional to the distance between the points.



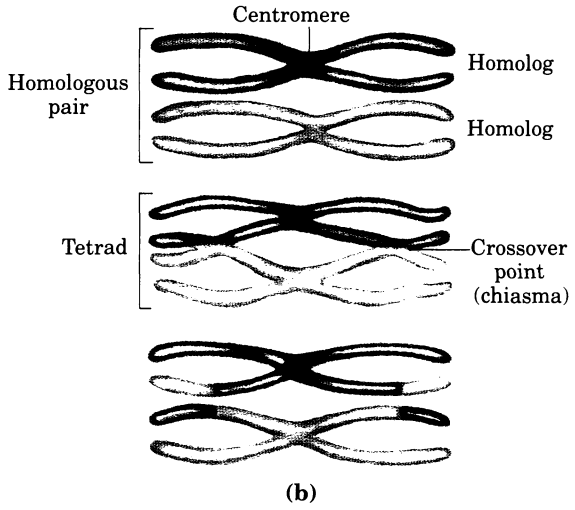


Figure 24–26 Crossing over. **(a)** The homologous chromosomes of a grasshopper are shown during prophase I of meiosis. Multiple points of joining (chiasmata) are evident between the two homologous pairs of chromatids. These chiasmata are the physical manifestation of prior homologous recombination (crossing over) events. **(b)** Crossing over often results in an exchange of genetic material.

This fact has been used by geneticists for many decades to map the relative positions and distances between genes; homologous recombination is therefore the molecular process that underpins much of the classical application of the science of genetics.

In bacteria, which do not of course undergo meiosis, homologous genetic recombination occurs in processes such as conjugation, a mating in which chromosomal DNA is transferred between two closely linked bacterial cells, or it can occur within a single cell between the two homologous chromosomes present during or immediately after replication.

This type of recombination serves at least three identifiable functions: (1) it contributes to genetic diversity in a population; (2) it provides in eukaryotes a transient physical link between chromatids that is apparently critical to the orderly segregation of chromosomes to the daughter cells in the first meiotic cell division; and (3) it contributes to the repair of several types of DNA damage.

The first and second functions are often of most interest to scientists studying genes, and homologous recombination is often described as a source of genetic diversity. However, the DNA repair function is almost certainly the most important role in the cell. DNA repair as described thus far is predicated on the fact that a DNA lesion in one strand can be accurately repaired because the genetic information is preserved in an undamaged complementary strand. In certain types of lesions, such as double-strand breaks, double-strand cross-links, or lesions left behind in single strands during replication (Fig. 24–27), the complementary strand is itself damaged or absent. When this occurs, the information required for accurate DNA repair must come from a separate, homologous chromosome, and the repair involves homologous recombination. These kinds of lesions commonly result from ionizing radiation and oxidative reactions, and their repair is critical to the production of viable gametes in eukaryotes and to the everyday existence of bacteria. Repair that is mediated by homologous genetic recombination is simply called **recombinational repair;** it is discussed in detail later in this chapter.

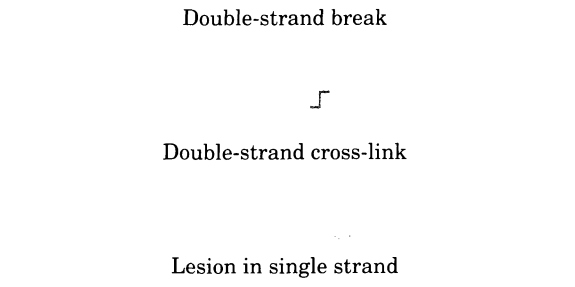
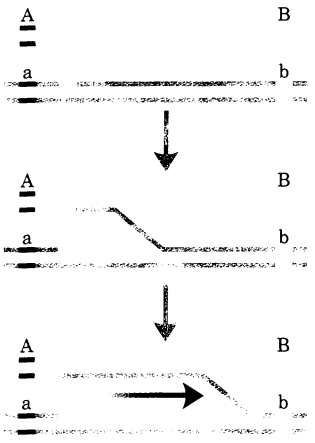
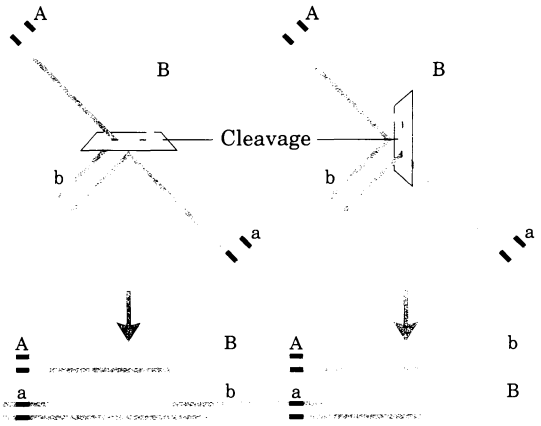


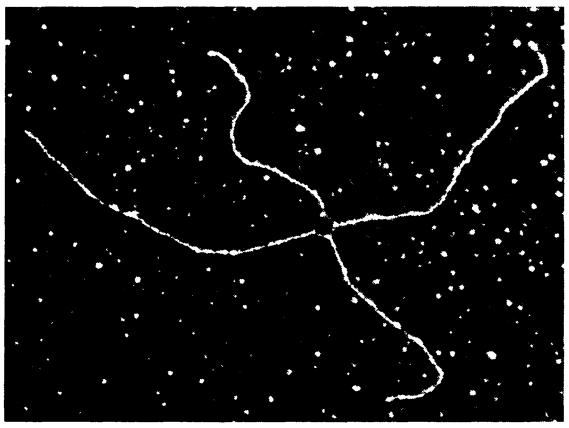
Figure 24–27 Types of DNA damage that require recombinational repair. In each case the damage to one strand cannot be repaired by mechanisms described earlier in this chapter because the complementary strand required to direct accurate repair is damaged or absent.



The Holliday intermediate can be cleaved (or resolved) in two ways, producing two possible sets of products. Below, the orientation of the Holliday intermediate is changed to clarify differences in the two cleavage patterns:



(a)



(b)

Figure 24–28 (a) The Holliday model for homologous genetic recombination. Two genes on the homologous chromosomes are indicated by the regions in red and yellow. Each chromosome has different alleles of these genes, as indicated by uppercase and lowercase letters. Note which alleles are linked in the four final products. **(b)** A Holliday intermediate formed between two bacterial plasmids in vivo, as seen with the electron microscope.

An important contribution to understanding homologous recombination is a model proposed by Robin Holliday in 1964, a version of which is presented in Figure 24–28. There are four key features of this model: (1) homologous DNAs are aligned by an unspecified mechanism; (2) one strand of each DNA is broken and joined to the other to form a crossover structure called a **Holliday intermediate;** (3) the region in which strands from different DNA molecules are paired, called **heteroduplex DNA,** is extended by **branch migration** (Fig. 24–29); and (4) two strands of the Holliday intermediate are cleaved and the breaks are repaired to form recombinant products. Homologous recombination can vary in many details from one species to another, but most of these steps are generally present in some form. Holliday intermediates have been observed in vivo in bacteria and in bacteriophage DNA (Fig. 24–28b). Note that there are two ways to cleave or "resolve" the Holliday intermediate so that the process is conservative, that is, so that the two products contain the same genes linked in the same linear order as in the substrates. If cleaved one way, the DNA flanking the heteroduplex region is recombined; if cleaved the other way, the flanking DNA is not recombined (Fig. 24–28a). Both outcomes are observed in vivo in both eukaryotes and prokaryotes.

Homologous recombination as illustrated in Figure 24–28 is a very elaborate process with subtle molecular consequences. To understand how this process affects genetic diversity, it is important to note that *homologous* does not necessarily mean *identical.* The two homologous chromosomes that are recombined may contain the same linear array

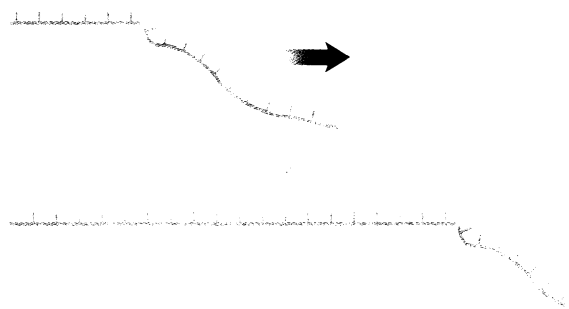


Figure 24–29 Branch migration occurs within a branched DNA structure in which at least one strand is partially paired with each of two complementary strands. The branch "migrates" when a base pair to one of the two complementary strands is broken and replaced with a base pair to the second strand. In the absence of an enzyme to direct it, this process can move the branch spontaneously in either direction.

of genes, but each chromosome may have slightly different base sequences in some of these genes. In a human, for example, one chromosome may contain the normal gene for hemoglobin while the other contains a hemoglobin gene with the sickle-cell mutation. The differences may represent no more than a change in a base pair or two among millions of identical base pairs. Although homologous recombination does not change the linear array of genes, it can determine which of the different versions (or alleles) of the genes are linked together on a single chromosome (Fig. 24–28).

Recombination Requires Specific Enzymes

Enzymes have been isolated from both prokaryotes and eukaryotes that promote one or more steps of homologous recombination. Again, progress in both identifying and understanding these enzymes has been greatest in *E. coli*. Important recombination enzymes are encoded by the *rec*A, B, C, and D genes, and by the *ruv*C gene. The *rec*B, C, and D genes encode the RecBCD enzyme, which can initiate recombination by unwinding DNA and occasionally cleaving one strand. The RecA protein promotes all the central steps in the process: the pairing of two DNAs, formation of Holliday intermediates, and branch migration as described below. A novel class of nucleases that specifically cleave Holliday intermediates have also been isolated from bacteria and yeast. These nucleases are often called resolvases; the *E. coli* resolvase is the RuvC protein.

The RecBCD enzyme binds to linear DNA at one end and uses the energy of ATP to travel along the helix, unwinding the DNA ahead and rewinding it behind (Fig. 24–30). Rewinding is slower than unwinding so that a single-stranded bubble is gradually formed and enlarged. The single strands in the bubble are cut when the enzyme encounters a certain sequence called *chi*, (5′)GCTGGTGG(3′). There are about 1,000 of these sequences in the *E. coli* genome, and they have the effect of increasing the frequency of recombination in the regions where they occur. Sequences that enhance recombination frequency have also been identified in several other organisms.

The RecA protein is unusual among proteins involved in DNA metabolism in that its active form is an ordered, helical filament that assembles cooperatively on DNA and can involve thousands of RecA monomers (Fig. 24–31). Formation of this filament normally occurs on single-stranded DNA such as that produced by the RecBCD enzyme. The filament will also form on a duplex DNA with a single-stranded gap, in which case the first RecA monomers bind to the single-stranded DNA in the gap and then filament assembly rapidly envelops the neighboring duplex.

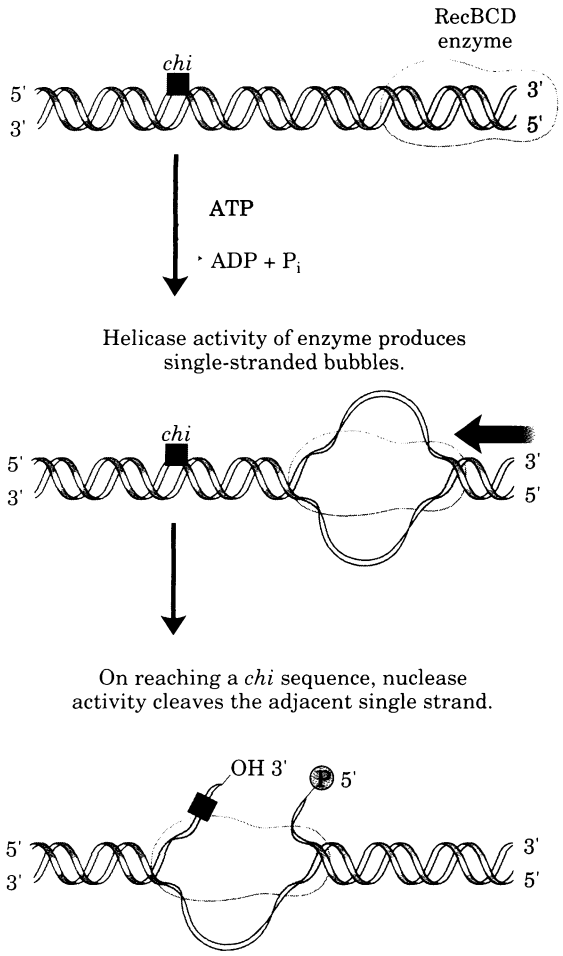


Figure 24–30 Helicase and nuclease activities of the RecBCD enzyme. Unwinding of DNA ahead of the moving enzyme and slower rewinding behind create single-stranded bubbles. One strand is cleaved when the enzyme encounters a *chi* sequence. Movement of the enzyme requires ATP hydrolysis. This enzyme is believed to help initiate homologous genetic recombination in *E. coli*.

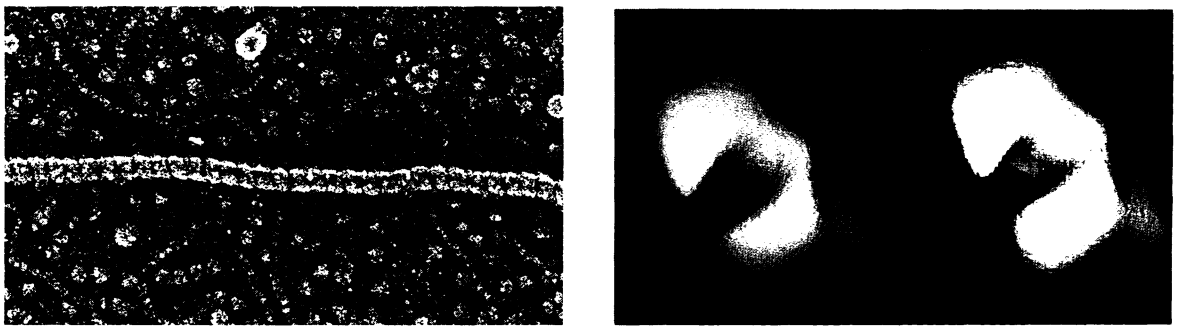


Figure 24–31 **(a)** Nucleoprotein filament of RecA protein on single-stranded DNA, as seen with the electron microscope. The striations make evident the right-handed helical structure of the filament. **(b)** A computer enhancement of the structure seen with the electron microscope.

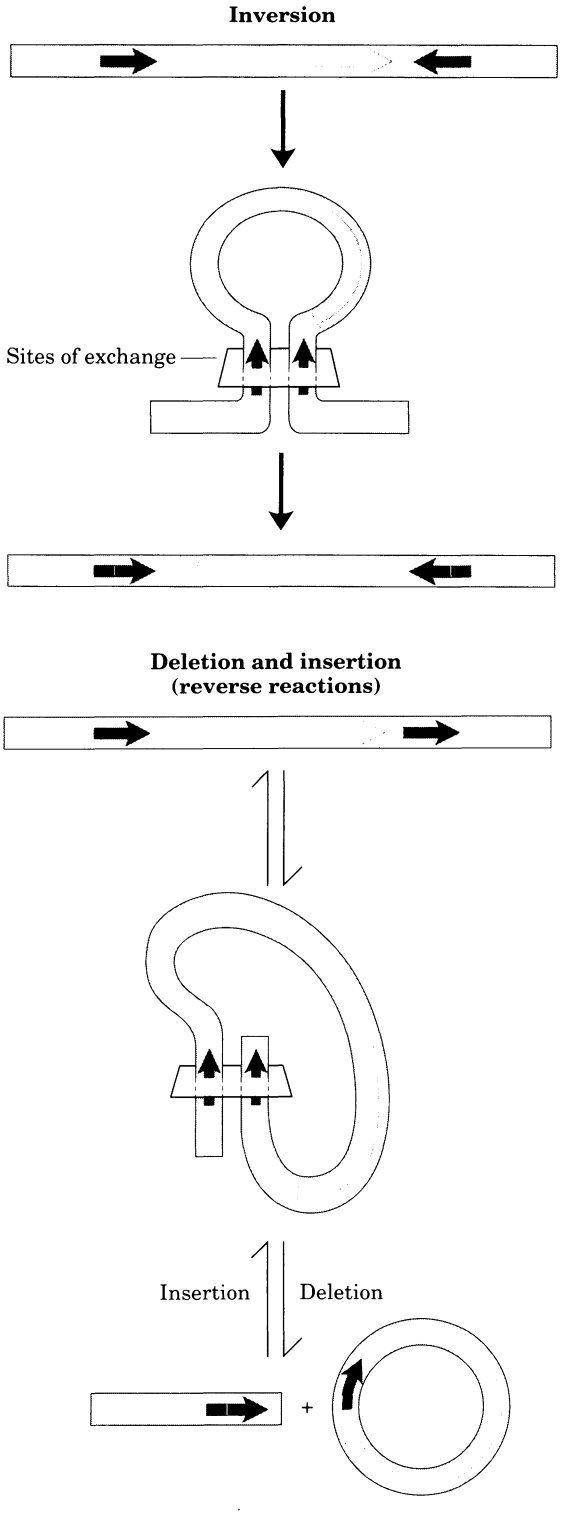


Figure 24–36 Possible outcome of site-specific recombination, depending on location and orientation of recombination sites (red and blue) in a double-stranded DNA molecule. Inversion and deletion and insertion are illustrated. Orientation here refers to the order of nucleotides in the recombination site, not the 5′→3′ direction.

The sequences of recombination sites recognized by these recombinases are partially asymmetric (nonpalindromic), and the two recombining sites are aligned in the same orientation for reaction by the recombinase. The reaction can have several outcomes, depending on the relative location and orientation of the recombination sites (Fig. 24–36). If the two sites are on the same DNA molecule the reaction will result in either inversion or deletion of the DNA between them, depending on whether the sites have the opposite or the same orientation, respectively. If the sites are on different DNAs the recombination is intermolecular, and an insertion reaction occurs if one or both of these DNAs is circular. Some systems are highly specific for one of these reactions (e.g., inversions) and will not act on sites in the wrong relative orientation.

The first site-specific recombination system identified and studied in vitro was that encoded by the bacteriophage λ. When λ phage DNA enters an *E. coli* cell, a complex series of regulatory events ensues that commits the DNA to one of two fates: either it is replicated and used to produce more bacteriophages (in which case the host cell is destroyed), or it is integrated into the host chromosome where it can be replicated passively along with the host chromosome for many cell generations. Integration is accomplished by a phage-encoded recombinase called the λ integrase, acting at recombination sites (attachment sites in the bacteriophage λ system) on the phage and bacterial DNAs called *att*P and *att*B, respectively (Fig. 24–37). Several auxiliary proteins also are used in this reaction, some encoded by the bacteriophage and others by the bacterial host cell. Note that a site-specific recombination reaction (Fig. 24–35) is chemically symmetric in terms of the chemical bonds present before and after, and it should have an equilibrium constant of 1.0. A major function of the auxiliary proteins in λ integration is to alter this equilibrium by permitting integration and/or preventing the reverse reaction (excision). The mechanism by which this is accomplished is not understood in detail. When the bacteriophage DNA must eventually be excised from the chromosome (which occurs when the cell is subjected to a variety of environmental stresses), the site-specific excision reaction uses a different set of auxiliary proteins (Fig. 24–37).

The use of site-specific recombination to regulate gene expression will be considered in Chapter 27.

Immunoglobulin Genes Are Assembled by Recombination

An important example of a programmed recombination event that occurs during development is the generation of immunoglobulin genes from gene segments that are separate in the genome. Immunoglobulins (or antibodies), produced by B lymphocytes, are the foot soldiers of the vertebrate immune system—the molecules that bind to infectious agents and all substances foreign to the organism. A mammal such as a human is capable of producing many millions of different antibodies with distinct binding specificities. However, the human genome contains only about 100,000 genes. Recombination allows an organism to produce an extraordinary diversity of antibodies from a relatively small amount of DNA-coding capacity.

Vertebrates generally produce multiple classes of immunoglobulins. To illustrate how antibody diversity is generated, we will focus on the immunoglobulin G (IgG) class from humans. Immunoglobulins consist of two heavy and two light polypeptide chains (Fig. 24–38a).

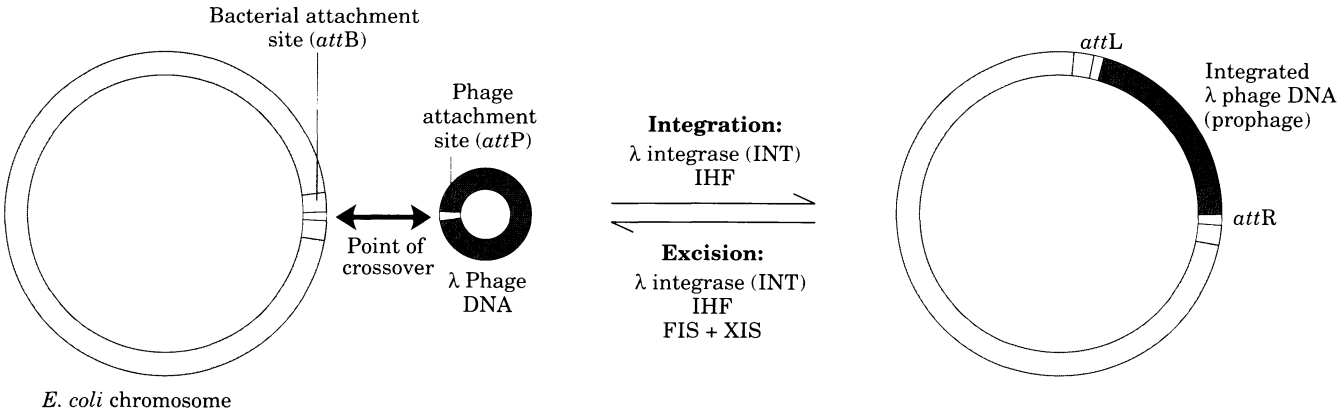


Figure 24–37 The integration and excision of bacteriophage λ DNA at the chromosomal target site. The attachment site on the λ phage DNA (*att*P) shares only 15 base pairs of complete homology with the bacterial site (*att*B) in the region of the crossover. The reaction generates two new attachment sites (*att*R and *att*L) flanking the integrated phage DNA. The recombinase is the λ integrase or INT protein. Integration and excision use different attachment sites and different auxiliary proteins. Excision uses the proteins XIS, encoded by the bacteriophage, and FIS, encoded by the bacterium. Both reactions require the protein IHF (*i*ntegration *h*ost *f*actor), encoded by the bacterium.

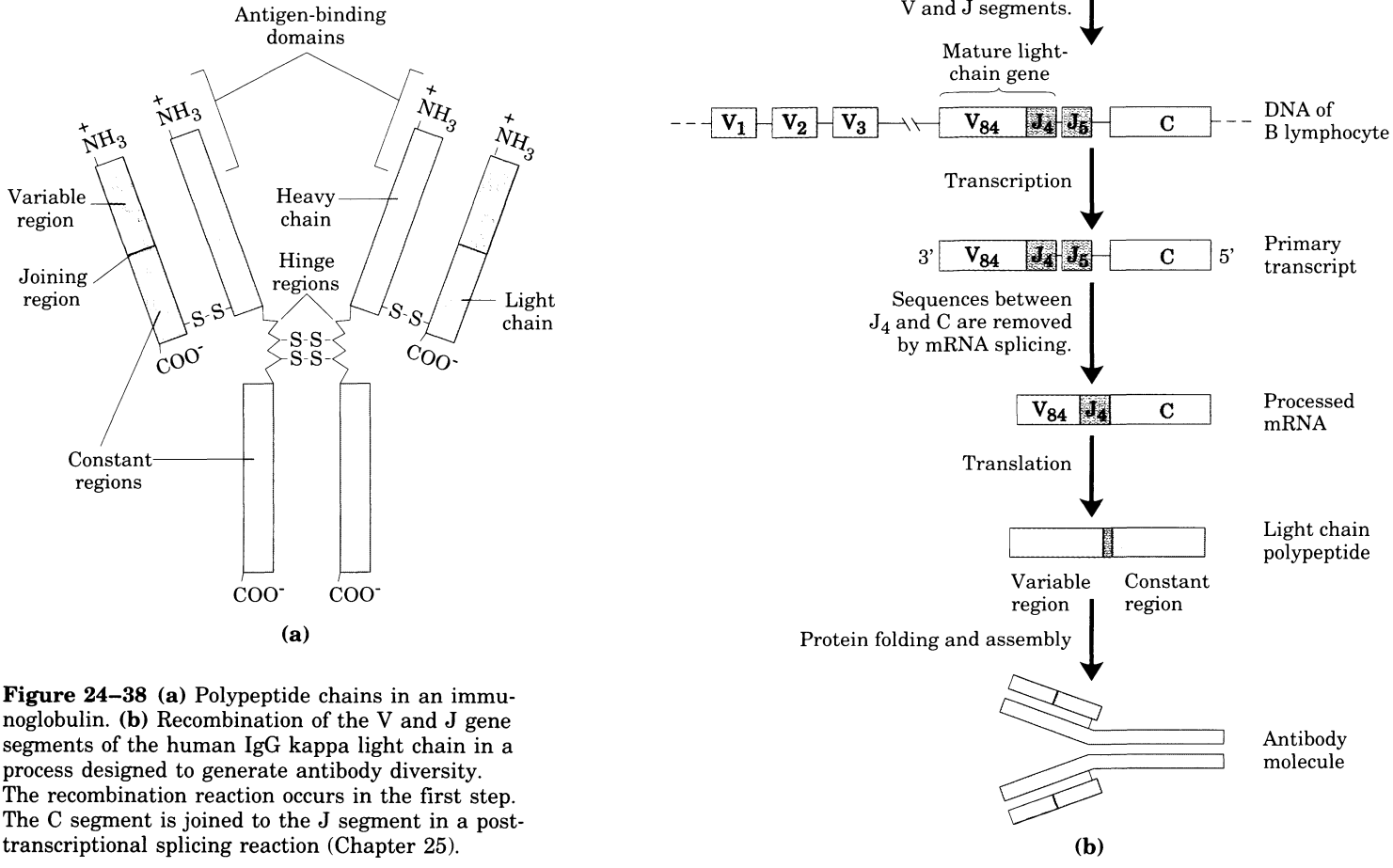


Figure 24–38 **(a)** Polypeptide chains in an immunoglobulin. **(b)** Recombination of the V and J gene segments of the human IgG kappa light chain in a process designed to generate antibody diversity. The recombination reaction occurs in the first step. The C segment is joined to the J segment in a post-transcriptional splicing reaction (Chapter 25).

Each chain has a variable region with a sequence that differs greatly from one immunoglobulin to the next, and another region that is virtually constant within a class of immunoglobulins. There are also two distinct families of light chains, called kappa and lambda, which differ somewhat in the sequences of their constant regions. For each of the three types of polypeptide chain (heavy chain, and kappa or lambda light chain), diversity in the variable regions is generated by a similar mechanism. The genes for these polypeptides are divided into segments, and clusters containing multiple versions of each segment exist in the genome. One version of each segment is joined to create a complete gene.

The organization of the DNA encoding the kappa light chains of human IgG and the process by which a mature kappa light chain is generated are shown in Figure 24–38b. In undifferentiated cells, the coding information for this polypeptide chain is separated into three segments. The V (*v*ariable) segment encodes the first 95 amino acid residues of the variable region, the J (*j*oining) segment encodes the remaining 12 amino acid residues of the variable region, and the C segment encodes the *c*onstant region. There are about 300 different V segments, 4 different J segments, and 1 C segment. As a stem cell in the bone marrow differentiates to form a mature B lymphocyte, one V and one J are brought together by site-specific recombination. This is effectively a programmed DNA deletion event, and the intervening DNA is discarded. There are $300 \times 4 = 1{,}200$ possible combinations. The recombination process is not as precise as the site-specific recombination described earlier, and some additional variation occurs in the sequence at the V–J junction that adds a factor of at least 2.5 to the total variation possible, so that about $2.5 \times 1{,}200 = 3{,}000$ different V–J combinations can be generated. The final joining of this V–J combination to the C region is accomplished by an RNA-splicing reaction after transcription (Fig. 24–38b). RNA splicing will be described in the next chapter. The genes for the heavy chains and lambda light chains are formed similarly. For heavy chains, there are more gene segments and more than 5,000 possible combinations. Because any heavy chain can combine with any light chain to generate an immunoglobulin, there are at least $3{,}000 \times 5{,}000$ or 1.5×10^7 possible IgGs. Additional diversity is generated because the V sequences are subject to high mutation rates (of unknown mechanism) during B-lymphocyte differentiation. Each mature B lymphocyte produces only one type of antibody, but the range of antibodies produced by different cells is clearly enormous. The enzymes that catalyze these gene rearrangements have not been isolated, but sequences critical to the V–J joining process that are presumably recognized by these enzymes have been identified.

This recombination process helps to illustrate the principle that recombination does not destroy the integrity of the genetic material that the replication and repair processes attempt to maintain. Here we see a precisely orchestrated process that occurs only in specialized cells (germ-line DNA is not affected) and enables the organism to make much more efficient use of its genetic information resource.

Transposable Genetic Elements Move from One Location to Another

Finally, we consider the recombination of transposable elements or **transposons.** Transposons are segments of DNA, found in virtually

RNA Processing

Many of the RNA molecules in bacteria and virtually all of the RNA molecules in eukaryotes are processed to some degree after they are synthesized. Many of the most interesting molecular events in RNA metabolism are to be found among these postsynthetic reactions. The study of these processes has revealed that some of them are catalyzed by enzymes made up of RNA rather than protein. The discovery of catalytic RNAs has brought on a revolution in thinking about RNA function and about the origin of life.

A newly synthesized RNA molecule is called a **primary transcript.** Perhaps the most extensive processing of primary transcripts occurs in eukaryotic mRNAs and in tRNAs of both bacteria and eukaryotes. A primary transcript for a eukaryotic mRNA typically contains sequences encompassing one gene. The sequences encoding the polypeptide, however, usually are not contiguous. Instead, in the majority of cases, the coding sequence is interrupted by noncoding tracts called introns; the coding segments are called exons (see the discussion of introns and exons in DNA, p. 798). In a process called **splicing,** the introns are removed from the primary transcript and the exons joined to form a contiguous sequence specifying a functional polypeptide. Eukaryotic mRNAs are also modified at each end. A structure called a cap is added at the 5′ end, and a polymer containing 20 to 250 adenylate residues, poly(A), is added to the 3′ end. These processes are outlined in Figure 25–10 and described in more detail below.

The primary transcripts of most tRNAs (in all organisms) are also processed by the removal of sequences from each end (called cleavage) and sometimes by the removal of introns (splicing). Many bases in tRNAs are also modified; mature tRNAs are replete with unusual bases not found in other nucleic acids.

The ultimate postsynthetic modification reaction is the complete degradation of the RNA. All RNAs eventually meet this fate and are replaced with newly synthesized RNAs. The rate of turnover of RNAs is critical to determining their steady-state level and the rate at which cells can shut down expression of a gene whose product is no longer needed.

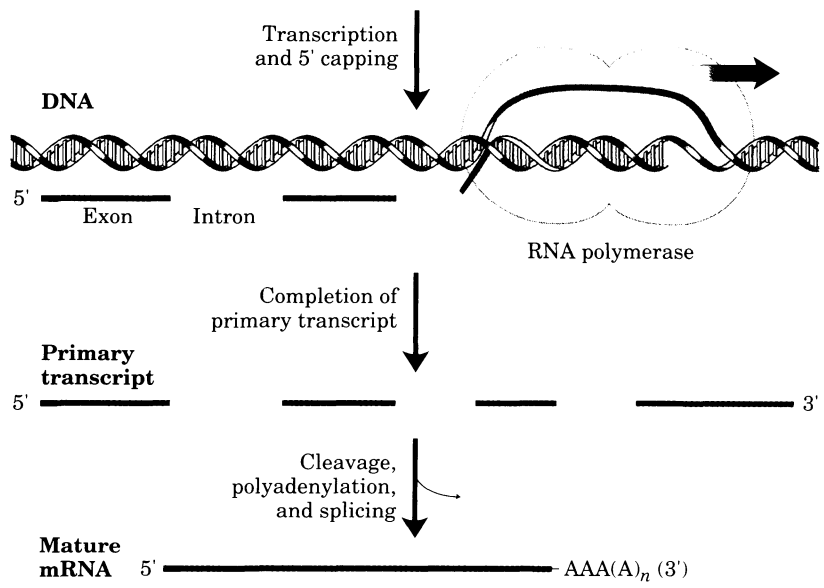


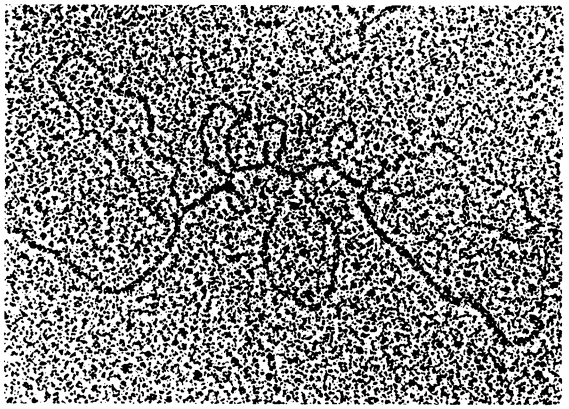
Figure 25–10 Formation of the primary transcript and its processing during maturation of the mRNA in a eukaryotic cell. The 5′ cap (in red) is added before synthesis of the primary transcript is complete. Noncoding sequences following the last exon are shown in orange. Splicing may occur either before or after the cleavage and polyadenylation steps. All of the processes represented here take place within the nucleus.

The Introns Transcribed into RNA Are Removed by Splicing

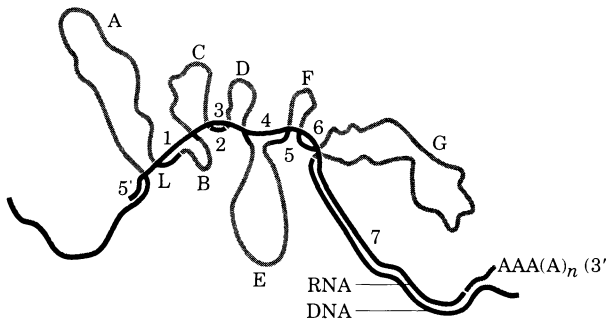
In bacteria, a polypeptide chain is generally encoded by a DNA sequence that is colinear with the amino acid sequence, continuing along the DNA template without interruption until the information needed to specify the polypeptide is complete. The notion that all genes are continuous was unexpectedly disproven in 1977 with the discovery that the genes for polypeptides in eukaryotes are often interrupted by the noncoding sequences now called introns. Introns are present in the vast majority of genes in vertebrates; among the few exceptions are the genes that encode certain histones. The occurrence of introns in other eukaryotes is variable. Most genes in the yeast *Saccharomyces cerevisiae* lack introns, although introns are more prevalent in the genes of some other yeast species. Introns are also found in a few prokaryotic genes.

Introns are spliced from the primary transcript, and exons are joined to form a mature, functional RNA. Introns were discovered when mRNA and the DNA from which it was derived were compared using methods such as that illustrated in Figure 25–11. If the DNA containing a gene is completely denatured and then renatured in the presence of the mature RNA derived from the gene, an RNA–DNA hybrid is formed. This kind of experiment revealed DNA sequences that were not present in the RNA and therefore were looped out as in Figure 25–11. Experiments using this and other methods have shown the presence of multiple introns in many genes, with some genes interrupted by introns more than 40 times. In eukaryotic mRNAs most exons are less than 1,000 nucleotides long, with many clustered in the 100 to 200 nucleotide size range. Most exons therefore encode polypeptide chains that are 30 to 50 amino acids long. Introns are much more variable in size (50 to 20,000 nucleotides). Genes of higher eukaryotes, including humans, typically have much more DNA devoted to introns than to exons; it is not uncommon to find genes that are 50,000 to 200,000 nucleotides long and that contain numerous introns.

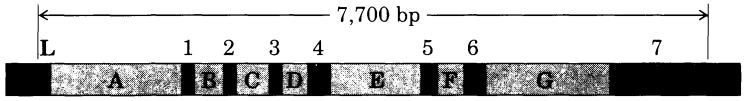
There are four classes of introns. The first two, called group I and group II, share some key characteristics but differ in the details of their splicing mechanisms. Group I introns are found in some nuclear, mitochondrial, and chloroplast genes coding for rRNAs; group II introns are generally found in the primary transcripts of mitochondrial or chloroplast mRNAs. Both groups share the property that no high-energy cofactors (such as ATP) are required for splicing. Both splicing



(a)



(b)



(c)

Figure 25–11 Defining the structure of the chicken ovalbumin gene by hybridization. Mature mRNA was hybridized to denatured DNA containing the ovalbumin gene, and the resulting molecules were visualized with the electron microscope. Some regions of the DNA have no complement in the mRNA because of splicing of the primary transcript. The resulting single-stranded DNA loops are evident in the electron micrograph **(a).** The loops define the locations and sizes of introns. The introns are labeled A to G and the seven exons are numbered in the interpretive drawing **(b).** The poly(A) tail defines the 3′ end of the mRNA. The L sequence encodes a signal sequence that targets the protein for export from the cell. **(c)** A linear representation of the ovalbumin gene showing introns and exons.

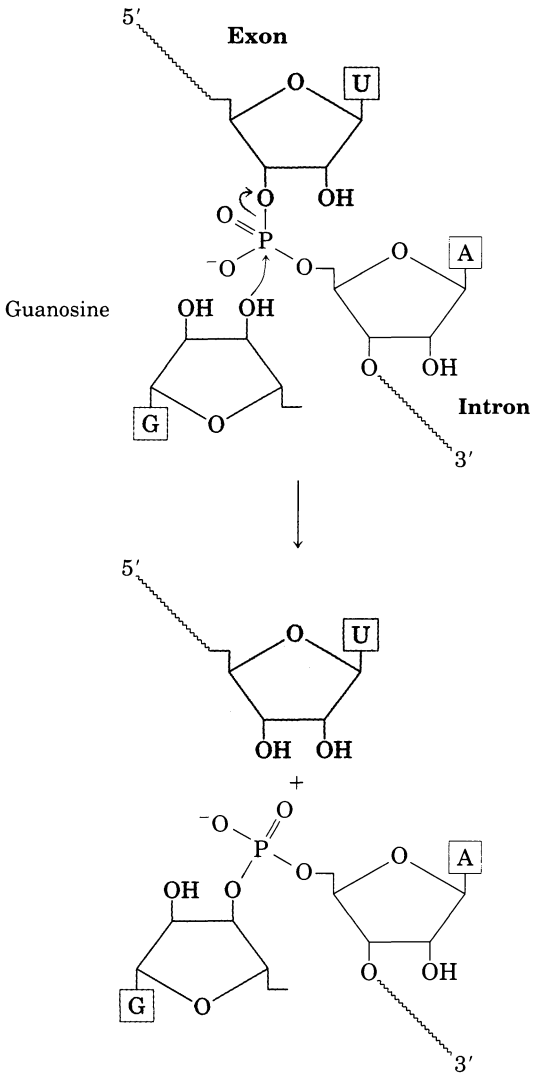


Figure 25–12 A transesterification reaction. This is the first step in the splicing of group I introns. Here, the 3′ OH of a guanosine molecule acts as nucleophile.

mechanisms involve two transesterification reaction steps (Fig. 25–12). A 2′- or 3′-hydroxyl group of a ribose makes a nucleophilic attack on a phosphorus, and in each step a new phosphodiester bond is formed at the expense of the old, maintaining an energy balance. Note that these reactions are very similar to the DNA breaking and rejoining reactions promoted by topoisomerases (Chapter 23) and site-specific recombinases (Chapter 24).

The group I splicing reaction requires a guanine nucleoside or nucleotide cofactor. This cofactor is not used as a source of energy; instead, the 3′-hydroxyl group of guanosine is used as a nucleophile in the first step of the splicing pathway. The guanosine 3′-hydroxyl forms a normal 3′,5′-phosphodiester bond with the 5′ end of the intron (Fig. 25–13). The 3′-hydroxyl of the exon that is displaced in this step then acts as a nucleophile in a similar reaction at the 3′ end of the intron. The result is precise excision of the intron and ligation of the exons.

In group II introns the pattern is similar except for the nucleophile in the first step. Instead of an external cofactor, the nucleophile is the 2′-hydroxyl group of an adenylate residue within the intron (Fig. 25–14). An unusual branched lariat structure is formed as an intermediate.

Attempts to identify the enzymes that promote splicing of group I and group II introns produced a major surprise; many of these introns are *self-splicing*—no protein enzymes are involved. This was first revealed in studies of the splicing mechanism of the group I rRNA intron from the ciliated protozoan *Tetrahymena thermophila* by Thomas Cech and colleagues in 1982. These workers proved that no proteins were involved by transcribing *Tetrahymena* DNA (including the intron) in

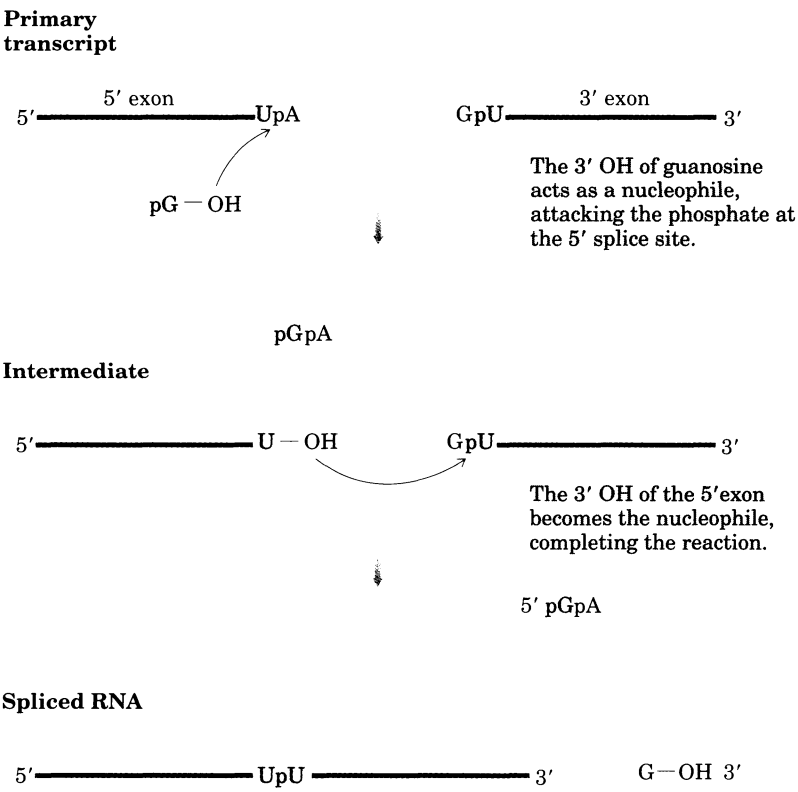


Figure 25–13 Splicing mechanism of group I introns. The nucleophile in the first step may be guanosine, GMP, GDP, or GTP.

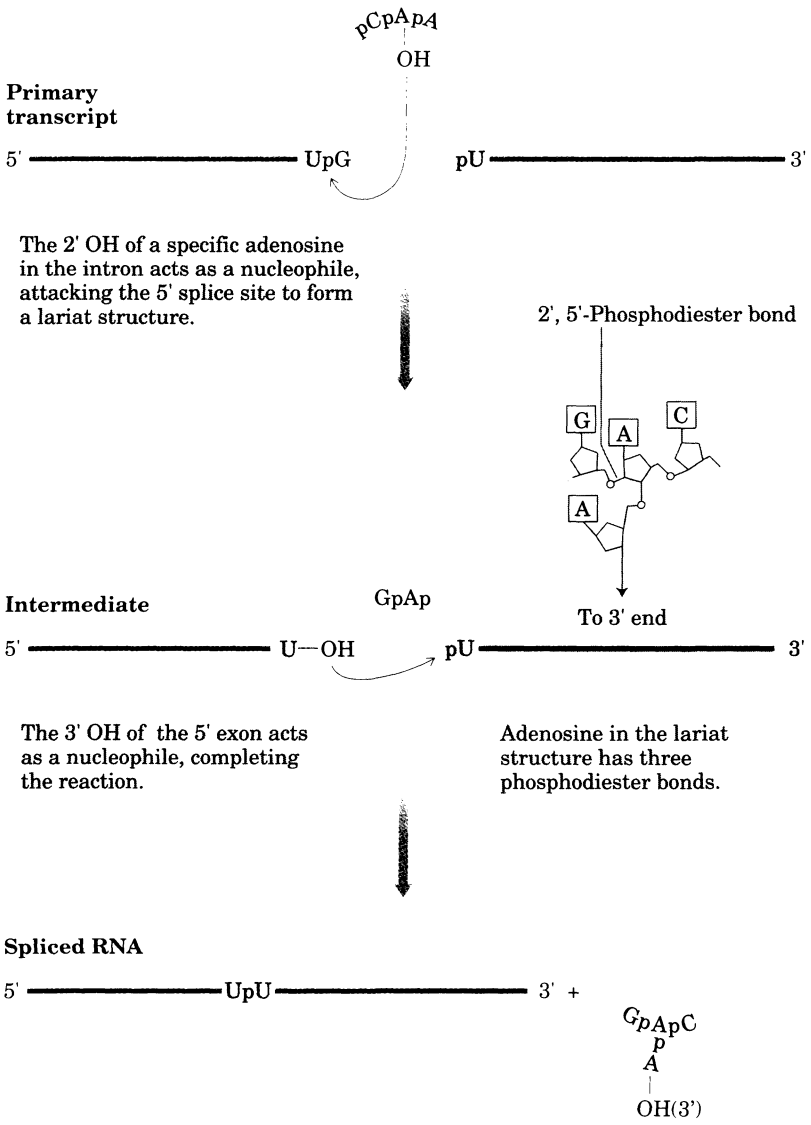


Figure 25–14 Splicing mechanism of group II introns. The chemistry is similar to that of group I intron splicing, except for the nucleophile in the first step and the novel lariatlike intermediate with one branch having a 2′,5′-phosphodiester bond.

vitro using bacterial RNA polymerase. The resulting RNA spliced itself accurately even though it had never been in contact with any enzymes from *Tetrahymena*. The realization that RNAs, as well as proteins, could have catalytic functions was a milestone in thinking about biological systems. RNA catalysts are discussed in more detail later in this chapter.

The third and largest group of introns, found in nuclear mRNA primary transcripts, undergo splicing by the same lariat-formation mechanism as the group II introns. However, they are not self-splicing. Splicing requires the action of specialized RNA–protein complexes containing a class of eukaryotic RNAs called **small nuclear RNAs (snRNAs).** Five snRNAs, U1, U2, U4, U5, and U6, are involved in splicing reactions. They are found in abundance in the nuclei of many eukaryotes, range in size from 106 (U6) to 189 (U2) nucleotides, and are complexed with proteins to form particles called *s*mall *n*uclear *r*ibo*n*ucleo*p*roteins (snRNPs, often referred to as "snurps"). The RNAs and proteins in snRNPs are highly conserved among vertebrates and insects. Small nuclear RNAs similar to these are also found in yeast and slime molds.

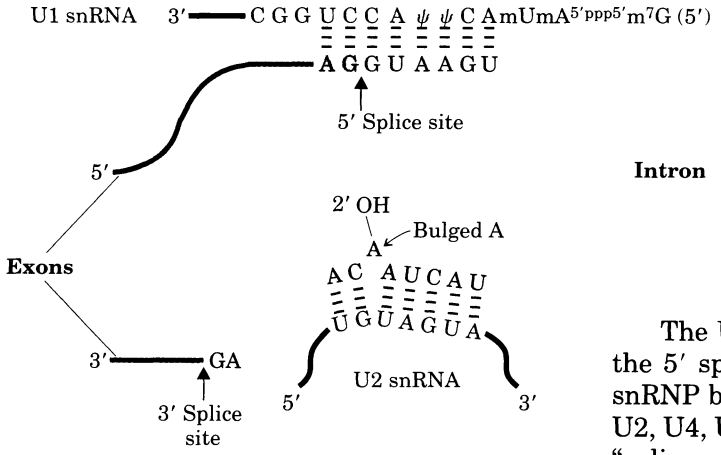


Figure 25–15 Splicing mechanism in mRNA primary transcripts. The splice sites that mark the intron–exon boundaries of many eukaryotic mRNAs have some conserved sequences. The U1 snRNA has a sequence near its 5′ end complementary to the splice site at the 5′ end of the intron. Base pairing of U1 to this region of the primary transcript helps define the 5′ splice site. ψ represents pseudouridine (see Fig. 25–25), and "m" indicates methylated residues. Base pairing of U2 snRNA to the branch site displaces (bulges) and perhaps activates the adenosine, whose 2′ OH forms the lariat structure through a 2′,5′-phosphodiester bond.

The U1 snRNA has a sequence complementary to sequences near the 5′ splice site of nuclear mRNA introns (Fig. 25–15), and the U1 snRNP binds to this region in the primary transcript. Addition of the U2, U4, U5, and U6 snRNPs leads to formation of a complex called the "spliceosome" within which the actual splicing reaction occurs. ATP is required for assembly of the spliceosome, but there is no reason to believe that the splicing reactions require ATP.

The fourth class of intron, found in certain tRNAs, is distinguished from the group I and II introns in that its splicing requires ATP. In this case, a splicing endonuclease cleaves the phosphodiester bonds at both ends of the intron, and the two exons are joined as shown in Figure 25–16. The joining reaction is similar to the DNA ligase mechanism (see Fig. 24–15).

Introns are not limited to eukaryotes. Although very rare, several genes with introns have now been found in bacteria and bacterial viruses. Bacteriophage T4, for example, has several genes with group I introns. Introns appear to be more common in archaebacteria (p. 25) than in *E. coli*.

Figure 25–16 (Facing page) The splicing of yeast tRNA. This splicing pathway requires a high-energy cofactor (ATP) for the ligation step. **(a)** The intron is first removed by endonuclease-catalyzed cleavage at both ends. **(b)** The 2′,3′-cyclic phosphate on the 5′ exon is cleaved by a cyclic nucleotide phosphodiesterase, leaving a 2′ phosphate. **(c)** The 5′ OH left on the 3′ exon is then activated in two steps. **(d)** The free 3′ hydroxyl of the 5′ exon acts as a nucleophile to displace AMP, joining the two exons with a 3′,5′-phosphodiester bond. **(e)** The 2′ phosphate is removed to yield the final product.

Eukaryotic mRNAs Undergo Additional Processing

In eukaryotes, mature mRNAs have distinctive structural features at *both ends. Most have a* ***5′ cap,*** *a residue of 7-methylguanosine linked* to the 5′-terminal residue of the mRNA through an unusual 5′,5′-triphosphate linkage (Fig. 25–17). At the 3′ end, most eukaryotic mRNAs have a "tail" of 20 to 250 adenylate residues, called the **poly(A) tail.** The functions of the 5′ cap and the 3′ poly(A) tail are only partially known. The 5′ cap binds to a protein and may participate in the binding of the mRNA to the ribosome to initiate translation (Chapter 26). The poly(A) tail also is bound by a specific protein. It is likely that the 5′ cap and poly(A) tail and their associated proteins help protect the mRNA from enzymatic destruction.

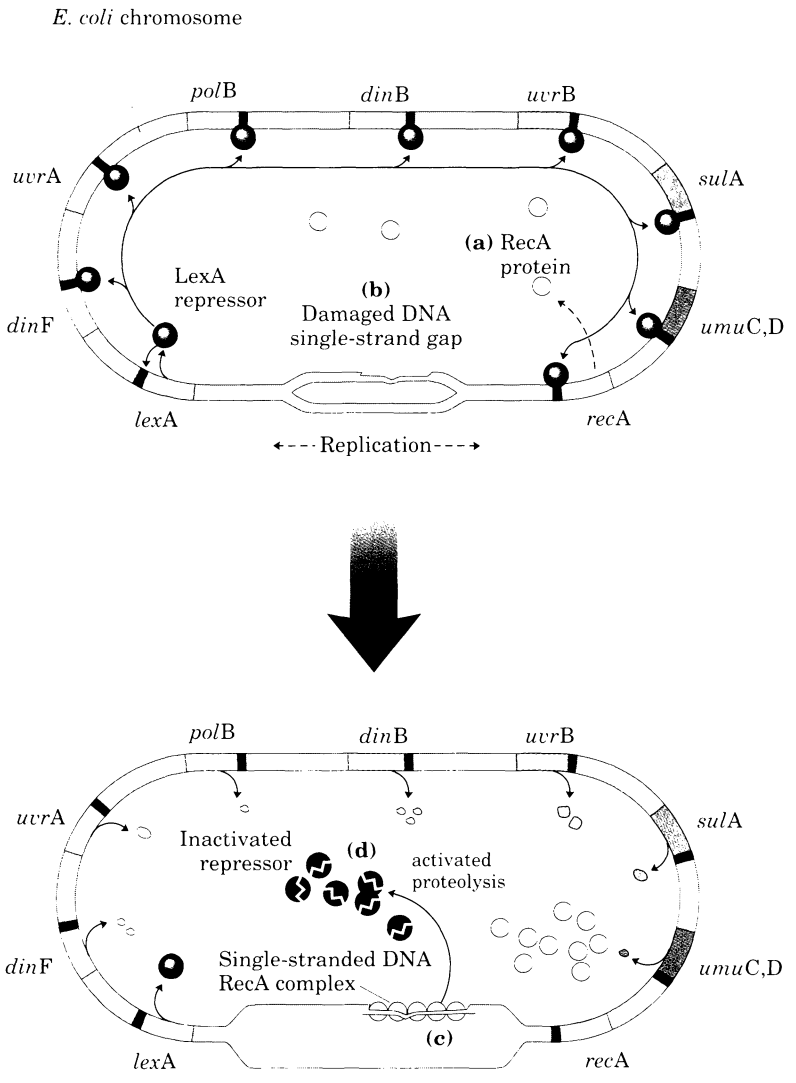


Figure 27–24 The SOS response in *E. coli*. See Table 24–6 for a description of the functions of these genes. The LexA protein is the repressor in this system, with an operator site (indicated in red) near each gene. **(a)** The *rec*A gene is not entirely repressed by the LexA repressor, and about 1,000 RecA protein monomers are normally found in the cell. **(b)** When DNA is extensively damaged (e.g., by UV light), DNA replication is halted and the number of single-strand gaps in the DNA increases. **(c)** RecA protein binds to this single-stranded DNA, activating the protein's coprotease activity. **(d)** While bound to DNA the RecA protein facilitates the cleavage and inactivation of the LexA repressor. When the repressor is inactivated, the SOS genes, including *rec*A, are induced. RecA protein levels increase 50- to 100-fold.

During induction of the SOS response, RecA protein also cleaves and thus inactivates the repressors that allow bacteriophage λ and related bacterial viruses to be propagated in a dormant (lysogenic) state within a bacterial host. These repressors have apparently evolved to mimic the LexA repressor, and they are also cleaved at a specific Ala–Gly peptide bond. This leads to replication of the virus and lysis of the cell to release the new virus particles, permitting the bacteriophage to make a hasty exit from a bacterial cell that is in distress, as described below.

Bacteriophage λ Provides an Example of a Regulated Developmental Switch

The objective of regulation of bacterial virus (bacteriophage) genes is usually the orderly assembly of new phage particles without destroying the host cell too soon. The well-studied bacteriophage λ provides an example of a complex and elegant regulatory circuit that determines the developmental fate of the virus in a given bacterial cell. This provides a paradigm for the complex problem of development in multicellular organisms.

Lysis or Lysogeny: Two Possible Fates Bacteriophage λ (Fig. 27–25) is a medium-sized DNA phage with a chromosome containing 48,502 base pairs. It is a **temperate phage,** meaning that phage infection does not always result in the destruction of the host cell. The DNA of the invading phage has two possible fates (Fig. 27–25b): reproduction leading to generation of new phage particles and lysis of the cell **(lytic**

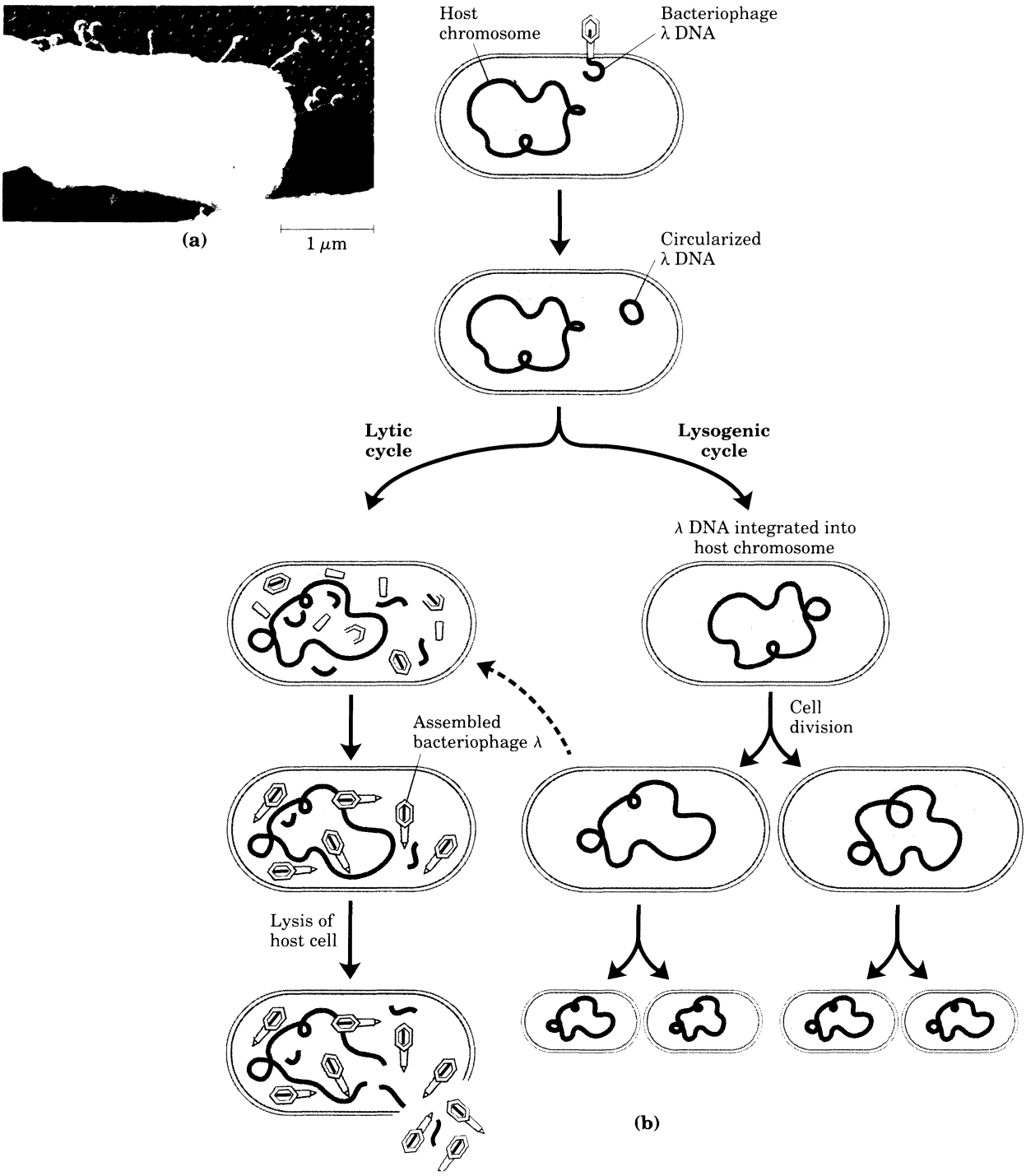


Figure 27–25 Bacteriophage λ. **(a)** Electron micrograph of bacteriophage λ virus particles attached to an *E. coli* cell. **(b)** Two alternative fates for a bacteriophage λ infection: lysis or lysogeny. Under certain conditions (e.g., when the SOS response is induced), the lysogenic state is interrupted and the cell undergoes a lytic cycle (dashed arrow).

cycle) or a relatively benign integration into the host chromosome where it can be replicated passively with the host DNA for many generations **(lysogeny).** The choice between lysis and lysogeny is governed largely by the interactions of five regulatory proteins called CI, CII, Cro, N, and Q. These proteins regulate transcription from a number of promoters in a regulatory region of the phage DNA described below. The functions of the proteins and promoters are summarized in Table 27–1.

Table 27–1 Regulatory elements of bacteriophage λ

Regulatory element	Function
Proteins	
CI	At low concentrations a repressor of P_R and P_L and an activator of P_{RM}; at high concentrations also represses P_{RM}
CII	An activator of P_{RE} and P_{int}
Cro	At low concentrations a repressor of P_{RM}; at high concentrations also represses P_L and P_R
N	An antiterminator at t_{L1}, t_{R1}, t_{R2}, and other terminators
Q	An antiterminator for late gene transcription
Promoters	
P_R	Major rightward transcription
P_L	Major leftward transcription
P_{RM}	Transcription for *r*epressor *m*aintenance
P_{RE}	Transcription for *r*epressor *e*stablishment
P_{int}	Transcription of genes for *int*egration and excision

The CI, CII, and Cro proteins are approximately analogous to the types of regulatory proteins already described. The CI and Cro proteins are repressors, and the CII protein is an activator. The N and Q proteins interact directly with the *E. coli* RNA polymerase to permit readthrough (i.e., transcription) of certain transcription termination sequences built into the phage DNA genome. This activity of the N and Q proteins is referred to as **antitermination,** and it is distinct from the regulatory mechanisms described to this point.

Lysis In the lytic pathway (Fig. 27–26), phage genes are arranged in three sets according to their time of expression. Some products of genes expressed in the first stage (immediate-early) are required to permit transcription of second-stage genes (delayed-early). The products of some second-stage genes, in turn, are required to permit transcription in the last stage (late genes). In general, genes required for replication and recombination are expressed early, and genes required for assembling phage peptides and lysing the host cell are expressed late. The N and Q proteins are the key to temporal regulation of the expression of bacteriophage genes.