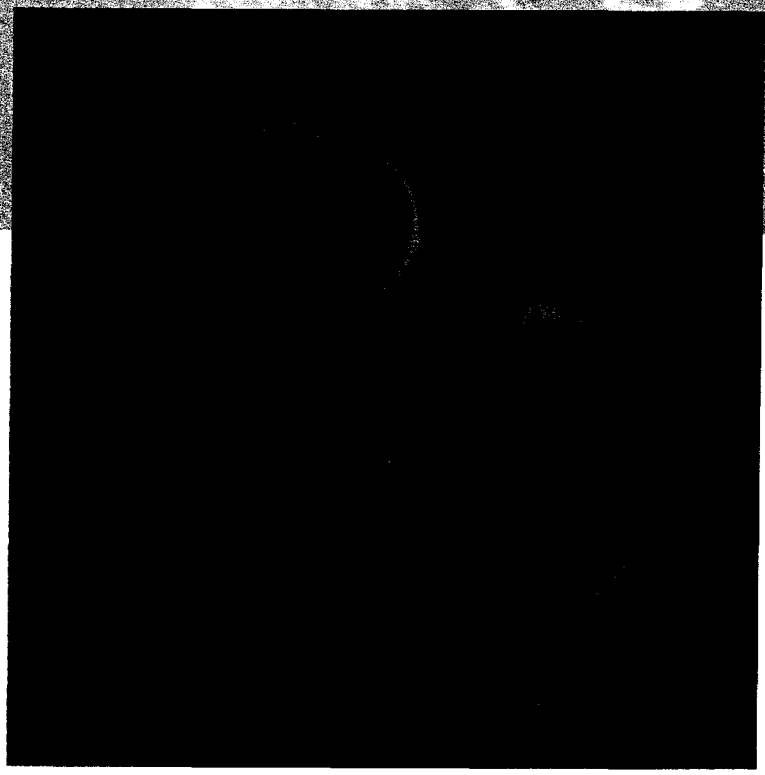
7 Recombinant DNA Technology



▲ Detection of human immunodeficiency virus (HIV-1) in two human lymphocytes by in situ polymerase chain reaction.

Once the structure of DNA and the genetic code were unraveled, it became clear that many deep biological secrets were locked up in the sequence of bases in DNA. But identifying the sequences of long regions of DNA, much less altering them at will, seemed a distant dream. An avalanche of technical discoveries in the 1970s drastically changed this perspective and has led to astounding advances in molecular cell biology in the past few years based on the analysis and manipulation of macromolecules, particularly DNA.

The discovery of two types of enzymes provided the impetus for these recent developments and permit *DNA cloning*. One type, called restriction enzymes, cut the DNA from any organism at specific sequences of a few nucleotides, generating a reproducible set of fragments. The other type, called DNA ligases, can insert DNA restriction fragments into replicating DNA molecules producing *recombinant DNA*. The recombinant DNA molecules then can be introduced into appropriate cells, most often bacterial cells; all the descendants from a single such cell, called a *clone*, carry the same recombinant DNA molecule. Once a clone of cells bearing a desired segment of DNA is isolated, unlimited quantities of this DNA can be prepared. In addition, DNA sequences up to ≈60 bases long can now be chemically synthesized by entirely automated procedures. Recombinant DNAs thus can be produced containing either natural DNA fragments resulting from restriction-enzyme cleavage or any desired chemically synthesized mutant sequences.

The availability of restriction enzymes also facilitated development of techniques for rapid *DNA sequencing* in the late 1970s. A long DNA molecule is first cleaved with restriction enzymes into a reproducible array of fragments, whose order in the original molecule is determined. Procedures also were developed for determining the sequence of bases in fragments up to 500 nucleotides long. Thus there

was no longer any obstacle to obtaining the sequence of a DNA containing 10,000 or more nucleotides. Suddenly, any DNA could be isolated and sequenced. With the aid of computer-automated procedures for sequencing and for storing, comparing, and analyzing data, scientists will sequence the entire human genome in the next few years.

Any cloned DNA segment, whether natural, modified, or completely synthetic, can be reinserted into cells and tested for biological activity. Almost overnight, this group of techniques, often collectively referred to as *recombinant DNA technology,* became the dominant approach for studying many basic biological processes, as examples in the following chapters will illustrate. In this chapter, we describe the various techniques that compose recombinant DNA technology. The power and success of this new technology promise to bring many practical benefits, particularly in medicine and agriculture.

➤ *DNA Cloning with Plasmid Vectors*

The essence of cell chemistry is to isolate a particular cellular component and then analyze its chemical structure and activity. In the case of DNA, this is feasible for relatively short molecules such as the genomes of small viruses. But genomes of even the simplest cells are much too large to directly analyze in detail at the molecular level. The problem is compounded for complex organisms. The human genome, for example, contains about 6×10^9 base pairs (bp) in the 23 pairs of chromosomes. Cleavage of human DNA with restriction enzymes, which produce about one cut for every 3000 base pairs, yields some 2 million fragments, far too many to separate from each other directly. This obstacle to obtaining pure DNA samples from large genomes has been overcome by recombinant DNA technology. With these methods virtually any gene can be purified, its sequence determined, and the functional regions of the sequence explored by altering it in planned ways and reintroducing the DNA into cells and into whole organisms.

The essence of recombinant DNA technology is the preparation of large numbers of identical DNA molecules. A DNA fragment of interest is linked through standard $3' \rightarrow 5'$ phosphodiester bonds to a *vector* DNA molecule, which can replicate when introduced into a host cell. When a single *recombinant* DNA molecule composed of a vector plus an inserted DNA fragment is introduced into a host cell, the inserted DNA is reproduced along with the vector, producing large numbers of recombinant DNA molecules that include the fragment of DNA originally linked to the vector. In this section, the general procedure for cloning DNA fragments in *E. coli* plasmids is described. Further details and techniques are presented in later sections.

Plasmids Are Extrachromosomal Self-Replicating DNA Molecules

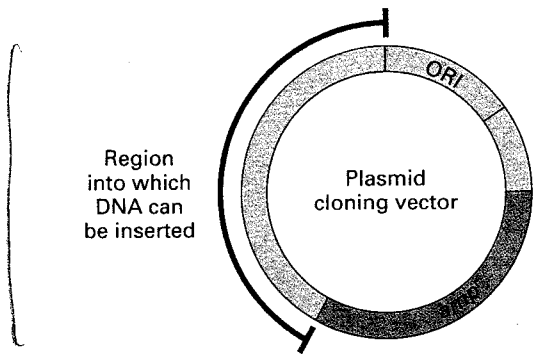
Plasmids are circular, double-stranded DNA (dsDNA) molecules that are separate from a cell's chromosomal DNA. These extrachromosomal DNAs occur naturally in bacteria and in the nuclei of yeast and some higher eukaryotic cells, existing in a parasitic or symbiotic relationship with their host cell. Plasmids range in size from a few thousand base pairs to more than 100 kilobases (kb). Like the host-cell chromosomal DNA, plasmid DNA is duplicated before every cell division. During cell division, at least one copy of the plasmid DNA is segregated to each daughter cell, assuring continued propagation of the plasmid through successive generations of the host cell.

Many naturally occurring plasmids contain genes that provide some benefit to the host cell, fulfilling the plasmid's portion of a symbiotic relationship. For example, some bacterial plasmids encode enzymes that inactivate antibiotics. Therefore, a bacterial cell containing such a plasmid is resistant to the antibiotic and can replicate in an environment containing the antibiotic, whereas the same type of bacterium lacking the drug-resistance plasmid is killed.

Such drug-resistance plasmids have become a major problem in the treatment of a number of common bacterial pathogens. As antibiotic use became widespread, plasmids containing several drug-resistance genes evolved, making their host cells resistant to a variety of different antibiotics simultaneously. Many of these plasmids also contain "transfer genes" encoding proteins that transfer a copy of the plasmid to other host cells of the same or related bacterial species. Plasmid transfer occurs through a complex macromolecular tube called a *pilus,* which is constructed from proteins encoded by some of the transfer genes. Such transfer can result in the rapid spread of drug-resistance plasmids, expanding the number of antibiotic-resistant bacteria in an environment such as a hospital. Coping with the spread of drug-resistance plasmids is an important challenge for modern medicine.

E. Coli Plasmids Can Be Engineered for Use as Cloning Vectors

The plasmids most commonly used in recombinant DNA technology replicate in *E. coli.* Generally, these plasmids have been engineered to optimize their use as vectors in DNA cloning. For instance, to simplify working with plasmids, their length is reduced; many plasmid vectors are only ≈3 kb in length, which is much less than that of naturally occurring *E. coli* plasmids. (The circumference of plasmids usually is referred to as their "length," even though plasmids are almost always circular DNA molecules.) Most plasmid vectors contain little more than the essential nucleotide sequences required for their use in



▲ FIGURE 7-1 Diagram of a simple cloning vector derived from a plasmid, a circular, double-stranded DNA molecule that can replicate within an *E. coli* cell. Plasmid vectors are ≈1.2–3 kb in length and contain a replication origin (ORI) sequence and a gene that permits selection usually by conferring resistance to a particular drug. Here the selective gene is *amp*^{r}; it encodes the enzyme β-lactamase, which inactivates ampicillin. Exogenous DNA can be inserted into the bracketed region without disturbing the ability of the plasmid to replicate or express the *amp*^{r} gene.

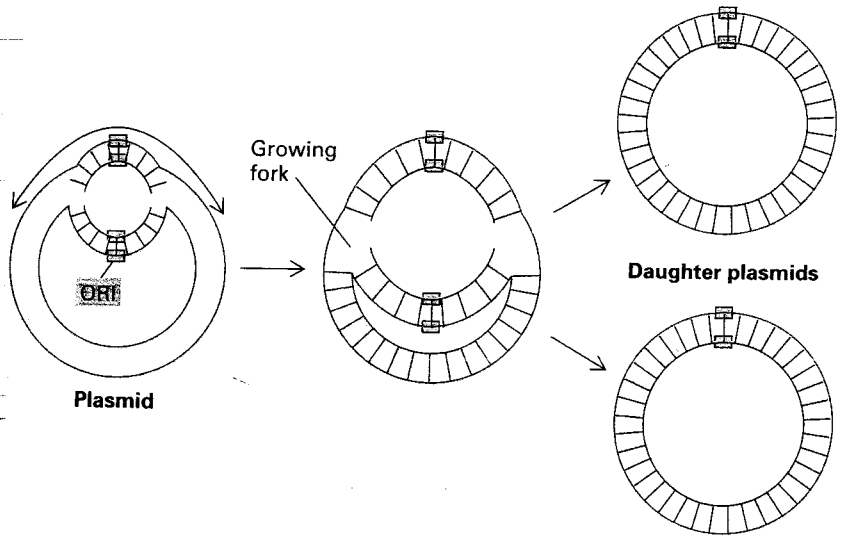
DNA cloning: a replication origin, a drug-resistance gene, and a region in which exogenous DNA fragments can be inserted (Figure 7-1).

Plasmid DNA Replication The *replication origin* (ORI) is a specific DNA sequence of 50–100 base pairs that must be present in a plasmid for it to replicate. Host-cell enzymes bind to ORI, initiating replication of the circular plasmid. Once DNA replication is initiated at ORI, it continues around the circular plasmid regardless of its nucleotide sequence (Figure 7-2). Thus any DNA sequence inserted into such a plasmid is replicated along with the rest of the plasmid DNA; this property is the basis of molecular DNA cloning.

Selection of Transformed Cells In 1944, O. T. Avery, C. M. Macleod, and M. McCarty first demonstrated gene transfer with isolated DNA obtained from *Streptococcus pneumoniae*. This process involved the genetic alteration of a bacterial cell by the uptake of DNA isolated from a genetically different bacterium and its recombination with the host-cell genome. Their experiments provided the first evidence that DNA is the genetic material. Later studies showed that such genetic alteration of a recipient cell can result from the uptake of exogenous extrachromosomal DNA (e.g., plasmids) that does not integrate into the host-cell chromosome. The term *transformation* is used to denote the genetic alteration of a cell caused by the uptake and expression of foreign DNA regardless of the mechanism involved. (Note that *transformation* has a second meaning defined in Chapter 6, namely the process by which normal cells with a finite life span in culture are converted into continuously growing cells similar to cancer cells.)

The phenomenon of transformation permits plasmid vectors to be introduced into and expressed by *E. coli* cells. In order to be useful in DNA cloning, however, a plasmid vector must contain a *drug-resistance gene* encoding an enzyme that inactivates a specific antibiotic. For example, the ampicillin-resistance gene (*amp*^{r}) encodes β-lactamase, which inactivates the antibiotic ampicillin. After plasmid vectors are incubated with *E. coli*, those cells that take up the plasmid can be easily selected from the larger number of cells that do not by growing them in an ampicillin-containing medium. The ability to select transformed cells is critical to DNA cloning by plasmid-vector technology because the transformation of *E. coli* with isolated plasmid DNA is inefficient.

Normal *E. coli* cells cannot take up plasmid DNA from the medium. Exposure of cells to high concentrations of certain divalent cations, however, makes a small fraction of cells permeable to foreign DNA by a mechanism



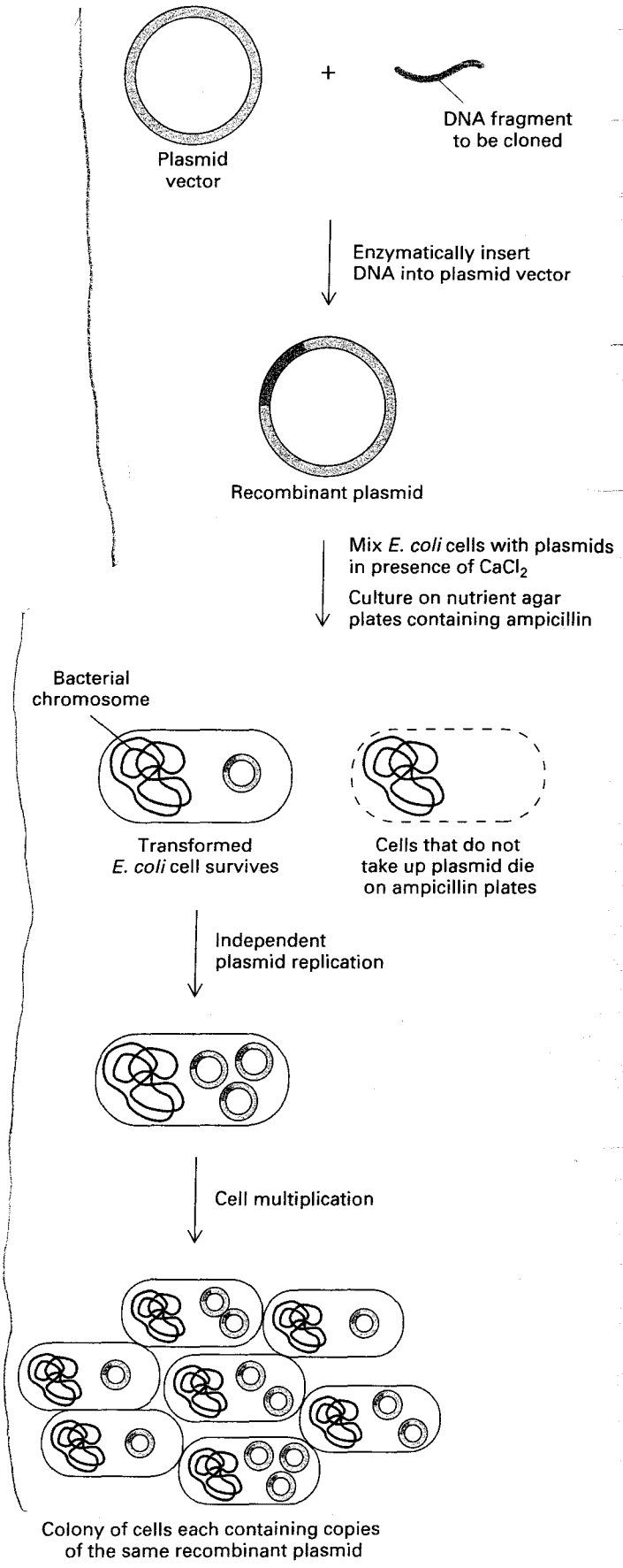
◀ FIGURE 7-2 Plasmid DNA replication. Newly synthesized daughter strands are shown in red. Once DNA replication is initiated at the origin (yellow), it continues in both directions around the circular molecule until the advancing growing forks merge and two daughter molecules are produced. The origin (ORI) is the only specific nucleotide sequence required for replication of the entire circular DNA molecule.

that is not understood. In a typical procedure, *E. coli* cells are treated with $CaCl_2$ and mixed with plasmid vectors; commonly, only 1 cell in about 10,000 or more cells becomes *competent* to take up the foreign DNA. Each competent cell incorporates a *single* plasmid DNA molecule, which carries an antibiotic-resistance gene. When the treated cells are plated on a petri dish of nutrient agar containing the antibiotic, only the rare transformed cells containing the antibiotic-resistance gene on the plasmid vector will survive. All the plasmids in such a colony of selected transformed cells are descended from the single plasmid taken up by the cell that established the colony.

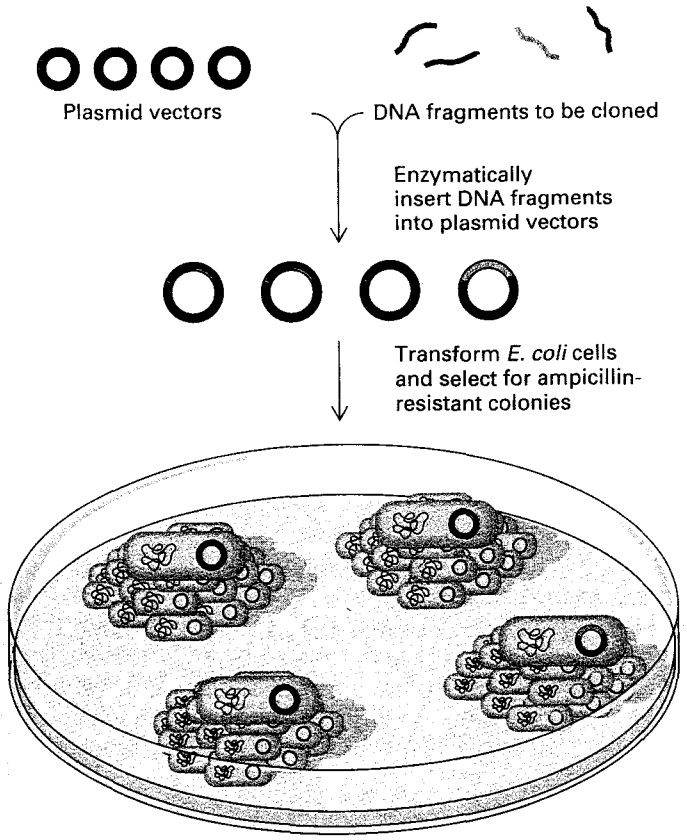
Plasmid Cloning Permits Isolation of DNA Fragments from Complex Mixtures

As discussed in the next section, a DNA fragment of a few base pairs up to ≈20 kilobases (kb) can be inserted into a plasmid vector. When such a recombinant plasmid transforms an *E. coli* cell, all the antibiotic-resistant progeny cells that arise from the initial transformed cell will contain plasmids with the same inserted sequence of DNA (Figure 7-3). The inserted DNA is replicated along with the rest of the plasmid DNA and segregates to daughter cells as the colony grows. In this way, the initial fragment of DNA is replicated in the colony of cells into a large number of identical copies. Since all the cells in a colony arise from a single transformed parental cell, they constitute a *clone* of cells. The initial fragment of DNA inserted into the parental plasmid is referred to as *cloned DNA*, since it can be isolated from the clone of cells.

DNA cloning allows fragments of DNA with a particular nucleotide sequence to be isolated from a complex mixture of fragments with many different sequences. As a simple example, assume you have a solution containing four different types of DNA fragments, each with a unique sequence (Figure 7-4). Each fragment type is individually in-



➤ FIGURE 7-3 General procedure for cloning a DNA fragment in a plasmid vector. Double-stranded DNA is represented by single lines here. Although not indicated by color, the plasmid contains a replication origin and ampicillin-resistance gene. Uptake of plasmids by *E. coli* cells is stimulated by high concentrations of $CaCl_2$. Even in the presence of $CaCl_2$, transformation occurs with a quite low frequency, and only a few cells are transformed by incorporation of a single plasmid molecule. Cells that are not transformed die on ampicillin-containing medium. Once incorporated into a host cell, a plasmid can replicate independently of the host-cell chromosome. As a transformed cell multiplies into a colony, at least one plasmid segregates to each daughter cell.



▲ FIGURE 7-4 Isolation of DNA fragments from a mixture by cloning in a plasmid vector. Four distinct DNA fragments, represented in red, green, dark purple, and light purple, are inserted into plasmid cloning vectors, yielding a mixture of recombinant plasmids each containing a single DNA fragment. *E. coli* cells treated with $CaCl_2$ are incubated with the mixture of recombinant plasmids and then plated on nutrient agar containing ampicillin. Each colony of transformed, antibiotic-resistant cells that grows (represented by a group of cells) arises from a single cell that took up one or another of the recombinant plasmids; all the cells in a given colony thus carry the same DNA fragment. Overnight incubation of *E. coli* at 37°C produces visible colonies containing about a million cells. Since the colonies are isolated from each other on the culture plate, copies of the DNA fragments in the original mixture are separated in the individual colonies. Although not indicated visually, the transformed cells contain multiple copies of a given plasmid.

serted into a plasmid vector. The resulting mixture of recombinant plasmids is incubated with *E. coli* cells under conditions that facilitate transformation; the cells then are cultured on antibiotic selective plates. Since each colony that develops arose from a single cell that took up a *single* plasmid, all of the cells in a colony harbor the identical type of plasmid characterized by the DNA fragment inserted into it. As a result, copies of the DNA fragments in the initial mixture are isolated from each other in the separate bacterial colonies. DNA cloning thus is a powerful, yet simple method for purifying a particular DNA fragment from a complex mixture of fragments and producing large numbers of the fragment of interest.

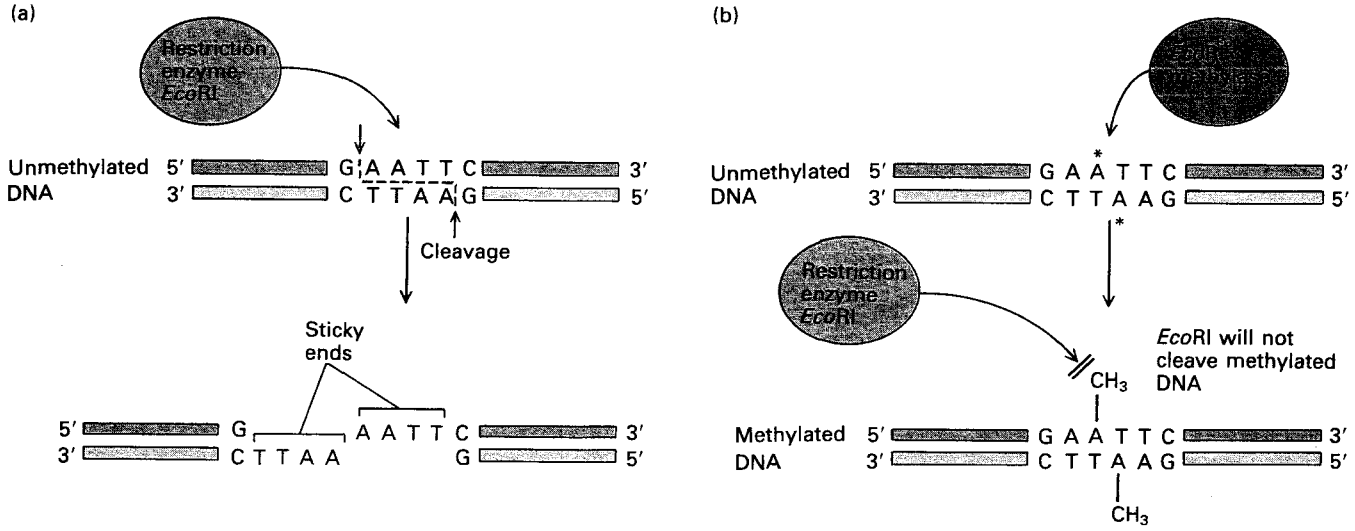
➤ *Production of Recombinant Plasmids*

To clone specific DNA fragments in a plasmid vector, as just described, or in other vectors discussed in later sections, the fragments must be produced and then inserted into the vector DNA. As noted in the introduction, restriction enzymes and DNA ligases are utilized to produce such recombinant cloning vectors.

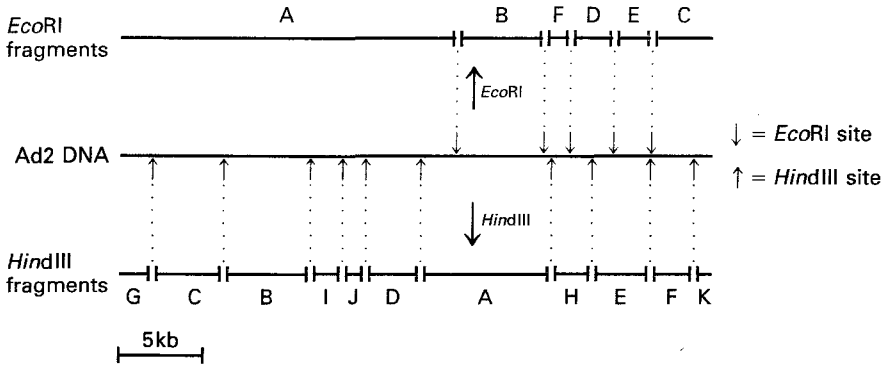
Restriction Enzymes Cut DNA Molecules at Specific Sequences

Restriction enzymes are bacterial enzymes that recognize specific 4- to 8-bp sequences, called *restriction sites*, and then cleave both DNA strands at this site. Since these enzymes cleave DNA within the molecule, they are also called *restriction endonucleases* to distinguish them from *exonucleases*, which digest nucleic acids from an end. Many restriction sites, like the *Eco*RI site shown in Figure 7-5a, are short inverted repeat sequences; that is, the restriction-site sequence is the same on each DNA strand when read in the 5′ → 3′ direction. Because the DNA isolated from an individual organism has a specific sequence, restriction enzymes cut the DNA into a reproducible set of fragments called *restriction fragments* (Figure 7-6).

The word *restriction* in the name of these enzymes refers to their function in the bacteria from which they are isolated: a restriction endonuclease destroys (restricts) incoming foreign DNA (e.g., bacteriophage DNA or DNA taken up during transformation) by cleaving it at all the restriction sites in the DNA. Another enzyme, called a *modification enzyme*, protects a bacterium's own DNA from cleavage by modifying it at or near each potential cleavage site. The modification enzyme adds a methyl group to one or two bases, usually within the restriction site. When a methyl group is present there, the restriction endonuclease is prevented from cutting the DNA (Figure 7-5b). Together with the restriction endonuclease, the methylating enzyme forms a *restriction-modification system* that protects the host DNA while it destroys foreign DNA. Restriction enzymes have been purified from several hundred different species of bacteria, allowing DNA molecules to be cut at a large number of different sequences corresponding to the recognition sites of these enzymes (Table 7-1).



▲ FIGURE 7-5 Restriction-recognition sites are short DNA sequences recognized and cleaved by various restriction endonucleases. (a) *Eco*RI, a restriction enzyme from *E. coli,* makes staggered cuts at the specific 6-bp inverted repeat sequence shown. This cleavage yields fragments with single-stranded, complementary sticky ends. Many other restriction enzymes also produce fragments with sticky ends. The sticky ends on two fragments derived by cleavage of different DNAs with the same restriction enzyme readily base-pair, and the fragments can be enzymatically joined (ligated) to produce recombinant DNAs. (b) Bacterial cells with restriction endonucleases also contain corresponding modification enzymes that methylate bases in the restriction-recognition site. For example, *E. coli* cells containing the *Eco*RI restriction enzyme also contain the *Eco*RI methylase, a modification enzyme that catalyzes addition of a methyl group to two adenines in the *Eco*RI recognition sequence. The methylated restriction site is not cleaved by *Eco*RI, assuring that a cell making this restriction enzyme does not destroy its own DNA.



▲ FIGURE 7-6 Fragments produced by cleavage of the ≈36-kb DNA genome from adenovirus 2 (Ad2) by *Eco*RI and another restriction enzyme, *Hin*dIII from *Haemophilus influenzae.* Double-stranded DNA is represented by single black lines in this figure. Digestion of Ad2 DNA (*center*) with *Eco*RI generates six *Eco*RI fragments (*top*); these result from cleavage at each *Eco*RI restriction site (GAATTC) in the Ad2 sequence. Digestion with *Hin*dIII cleaves the Ad2 DNA at each *Hin*dIII site (AAGCTT), generating eleven specific fragments (*bottom*), all different from the *Eco*RI fragments. By convention, restriction fragments are labeled A–Z in order of decreasing size. By techniques described later, the order of fragments in the original DNA can be determined, thus mapping the restriction sites on the DNA (indicated by short arrows). Such a "restriction-site map" for various restriction enzymes is a unique characteristic of each DNA.

Many Restriction Enzymes Generate DNA Fragments with "Sticky" Ends

As illustrated in Figure 7-5a, *Eco*RI makes staggered cuts in the two DNA strands. Many other restriction enzymes make similar cuts, generating fragments that have a single-stranded "tail" at both ends. The tails on the fragments generated at a given restriction site are complementary to those on all other fragments generated by the same restriction enzyme. At room temperature, these single-stranded regions, often called "sticky ends," can transiently base-pair with those on other DNA fragments generated with the same restriction enzyme, regardless of the source of the DNA. This base pairing of sticky ends permits DNA from widely differing species to be ligated, forming chimeric molecules.

TABLE 7-1 Examples of the Actions of Restriction Endonucleases

Source Microorganism	Enzyme*	Recognition Site (↓)†	Number of Cuts (kb)‡ λ (50)	Ad2 (36)	SV40 (5.2)	pBR322 (4.3)
Arthrobacter luteus	*Alu*I	AG↓CT	143	158	34	14
Thermus aquaticus	*Taq*I	T↓CGA	121	50	1	13
Haemophilus parahaemolyticus	*Hph*I	GGTGA + 5	168	99	4	18
Haemophilus gallinarum	*Hga*I	GACGC + 8	102	87	0	12
Escherichia coli	*Eco*RI	G↓AATTC	5	5	1	1
Haemophilus influenzae	*Hin*dIII	A↓AGCTT	6	12	6	1
Nocardia otitiscaviarums	*Not*I	GC↓GGCCGC	0	7	0	0
Streptomyces fimbriatus	*Sfi*I	GGCCN$_4$↓NGGCC	0	3	1	0

*Enzymes are named with abbreviations of the bacterial strains from which they are isolated; the Roman numeral indicates the enzyme's priority of discovery in that strain (for example, *Alu*I was the first restriction enzyme to be isolated from *Arthrobacter luteus*).

†Recognition sequences are written 5′ → 3′ (only one strand is given) with the cleavage site indicated by an arrow. For example, G↓GATCC is an abbreviation for

```
      ↓
(5′)GGATCC(3′)
(3′)CCTAGG(5′)
      ↑
```

The cleavage site for *Hph*I and *Hga*I occurs five or eight bases away from the recognition sequence; N indicates any base.

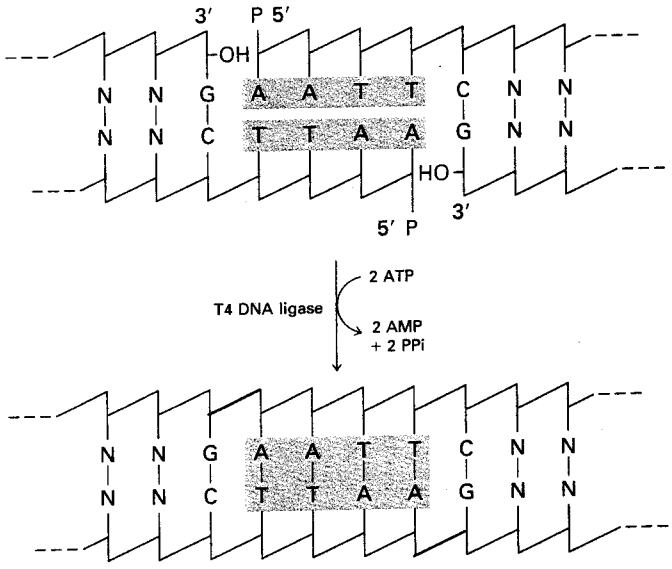
‡These columns list the number of cleavage sites recognized by specific endonucleases on the DNA of bacteriophage λ (λ), adenovirus type 2 (Ad2), simian virus 40 (SV40), and an *E. coli* plasmid (pBR322). The sizes of the DNAs in kilobases (kb) are in parentheses. Note that the actual number of cuts in these sequences deviates from the expected number in random sequences, which would be given by $L/4^n$, where n is the length of the site recognized by an enzyme and L is the length of the sequence.

SOURCE: R. J. Roberts, 1988, *Nuc. Acids Res.* 16(supp):r271.

DNA Ligase Covalently Links Restriction Fragments

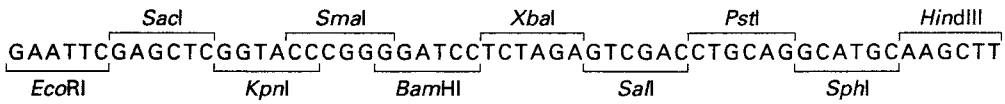
During in vivo DNA replication, DNA ligase catalyzes formation of 3′ → 5′ phosphodiester bonds between the short fragments of the discontinuously synthesized DNA strand at a replication fork (see Figure 4-22). In recombinant DNA technology, purified DNA ligase is used to covalently join the ends of restriction fragments in vitro. This enzyme can catalyze the formation of a 3′ → 5′ phosphodiester bond between the 3′-hydroxyl end of one restriction-fragment strand and the 5′-phosphate end of another restriction-fragment strand during the time that the sticky ends are transiently base-paired (Figure 7-7). When DNA ligase and ATP are added to a solution containing restriction fragments with sticky ends, the restriction fragments are covalently ligated together through the standard 3′ → 5′ phosphodiester bonds of DNA.

Some restriction enzymes cleave both DNA strands exactly in the middle of the recognition site (e.g., *Alu*I in Table 7-1). These restriction enzymes generate DNA restriction fragments with "blunt" ("flush") ends in which all of the nucleotides at the fragment ends are base-paired to nucleotides in the complementary strand. In addition to

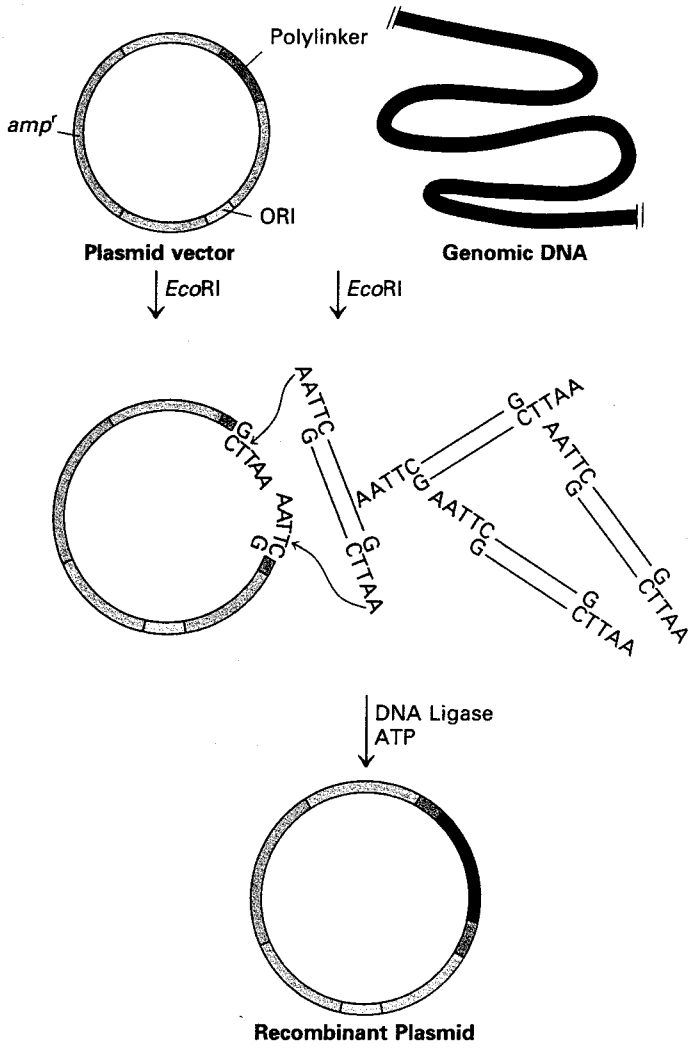


▲ FIGURE 7-7 Ligation of *Eco*RI sticky ends by DNA ligase from bacteriophage T4. After the complementary sticky ends (orange) on two fragments transiently base-pair, the adjacent 3′-hydroxyl and 5′-phosphate groups (red) are covalently joined (ligated). One ATP is consumed for each phosphodiester bond (red) formed.

(a) Sequence of polylinker



(b) Insertion of *Eco*RI restriction fragments



▲ FIGURE 7-8 Plasmid vectors containing a polylinker, or multiple cloning-site sequence, commonly are used to produce recombinant plasmids carrying exogenous DNA fragments. (a) Sequence of a polylinker that includes one copy of the recognition site, indicated by red brackets, for each of the 10 restriction enzymes indicated. Polylinkers are chemically synthesized and then are inserted into a plasmid vector. Only one strand is shown. (b) Insertion of genomic restriction fragments into the pUC19 plasmid vector, which contains the polylinker shown in (a). (The length of the polylinker in relation to the rest of the plasmid is greatly exaggerated here.) One of the restriction enzymes whose recognition site is in the polylinker is used to cut both the plasmid molecules and genomic DNA, generating singly-cut plasmids and restriction fragments with complementary sticky ends (letters at ends of green fragments). By use of appropriate reaction conditions, insertion of a single restriction fragment per plasmid can be maximized. Curved lines indicate base pairing between sticky ends prior to ligation. Note that the restriction sites are reconstituted in the recombinant plasmid. [See C. Yanisch-Perron, J. Vieira, and J. Messing, 1985, *Gene* **33**:103.]

ligating complementary sticky ends, the DNA ligase from bacteriophage T4 can ligate any two blunt DNA ends. However, blunt-end ligation requires a higher DNA concentration than ligation of sticky ends.

Restriction Fragments Are Readily Inserted into Plasmid Vectors

By use of restriction enzymes to create fragments with sticky ends and DNA ligase to covalently link them, foreign DNA can be inserted into plasmid vectors in vitro in a straightforward procedure. *E. coli* plasmid vectors can be constructed with a *polylinker*, or *multiple cloning-site sequence*, a synthetic sequence that contains one copy of several different restriction sites (Figure 7-8a). When such a vector is treated with a restriction enzyme that recognizes a recognition sequence in the polylinker, it is cut at that sequence, generating sticky ends. In the presence of DNA ligase, DNA fragments produced with the same restriction enzyme will be inserted into the plasmid (Figure 7-8b). The ratio of DNA fragments to be inserted to cut vectors and other reaction conditions are chosen to maximize the insertion of one restriction fragment per plasmid vector. The recombinant plasmids produced in in vitro ligation reactions then can be used to transform antibiotic-sensitive *E. coli* cells as shown in Figure 7-4. All of the cells in each antibiotic-resistant clone that remains after selection contain plasmids with the same inserted DNA fragment, but different clones carry different fragments.

➤ *Formation and Uses of Synthetic DNA*

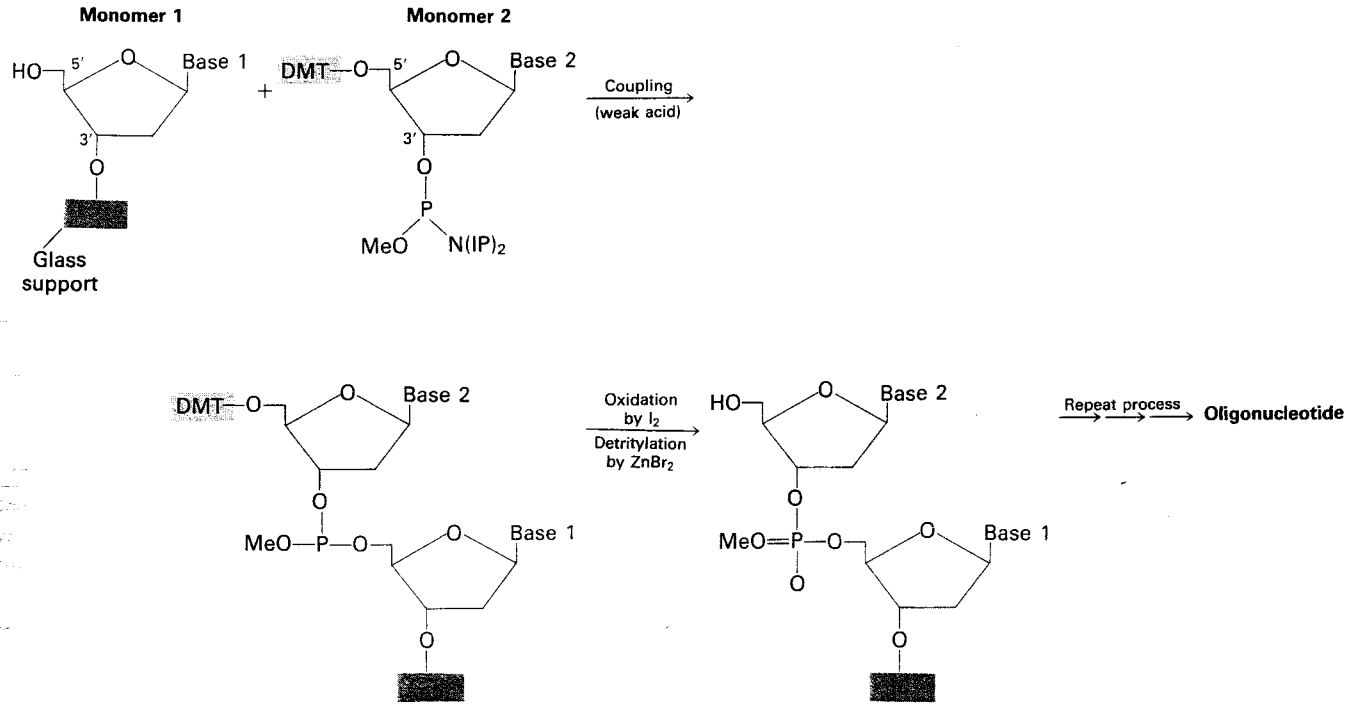
Advances in synthetic chemistry now permit the chemical synthesis of single-stranded DNA (ssDNA) molecules of any sequence up to about 100 nucleotides in length. Syn-

thetic DNA has a number of applications in recombinant DNA technology. Complementary ssDNAs can be synthesized and hybridized to each other to form a dsDNA with sticky ends. Such completely synthetic dsDNAs can be cloned into plasmid vectors just like DNA restriction fragments prepared from living organisms. For example, the 57-bp polylinker sequence shown in Figure 7-8 was chemically synthesized and then inserted into plasmid vectors to facilitate the cloning of fragments generated by different restriction enzymes. This example illustrates the use of synthetic DNAs to add convenient restriction sites where they otherwise do not occur. As described later in the chapter, synthetic DNAs are used in sequencing DNA and as probes to identify clones of interest. Synthetic DNAs also can be substituted for natural DNA sequences in cloned DNA to study the effects of specific mutations; this topic is examined in the next chapter.

The technique for chemical synthesis of DNA oligonucleotides is outlined in Figure 7-9. Note that chains grow in the 3′ → 5′ direction, opposite to the direction of DNA chain growth catalyzed by DNA polymerases. Once the chemistry for producing synthetic DNA was standardized, automated instruments were developed that allow researchers to program the synthesis of oligonucleotides of specific sequences up to about 100 nucleotides long.

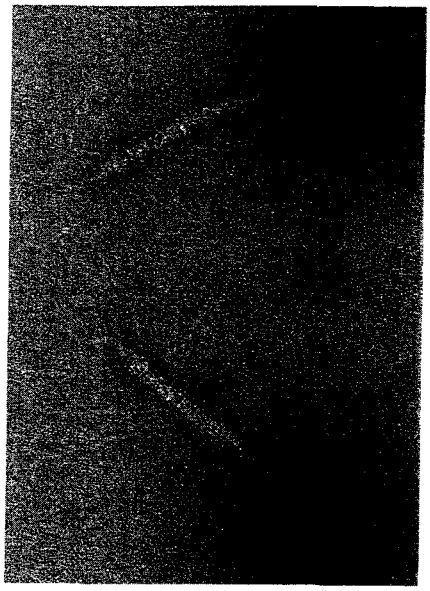
➤ λ-Phage Cloning Vectors and Construction of a Genomic Library

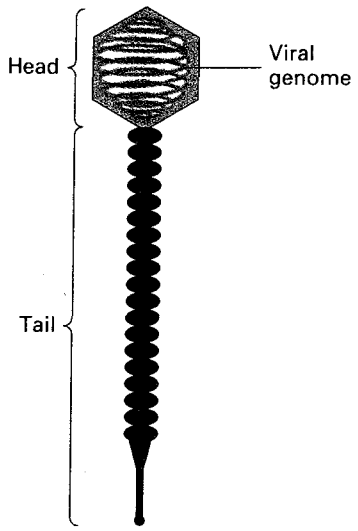
Most DNA cloning is done with *E. coli* plasmid vectors because of the relative simplicity of this procedure. However, the number of individual clones that can be obtained by plasmid cloning is limited by the relatively low efficiency of *E. coli* transformation and the small number (only a few hundred) of individual transformed colonies



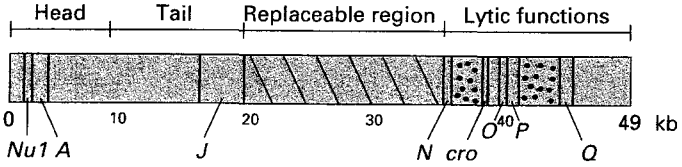
▲ FIGURE 7-9 Chemical synthesis of oligonucleotides by sequential addition of reactive nucleotide derivatives in the 3′ → 5′ direction. The first nucleotide (monomer 1) is bound to a glass support by its 3′ hydroxyl; its 5′ hydroxyl is available. The next nucleotide in the sequence (monomer 2) is derivatized by addition of 4′,4′-dimethoxytrityl (DMT) to its 5′ hydroxyl, thus blocking this hydroxyl from reacting; in addition, a highly reactive methylated diisopropyl phosphoramidite group (red letters) is attached to the 3′ hydroxyl. When the two monomers are mixed in the presence of a weak acid, they form a 5′ → 3′ phosphodiester bond with the phosphorus in the trivalent state. Oxidation with iodine (I_2) increases the phosphorus valency to five; removal of the DMT group by detritylation with zinc bromide ($ZnBr_2$) frees the 5′ hydroxyl, and the process is repeated. When synthesis is complete, all the methyl groups on the phosphates are removed at the same time at alkaline pH, and the bond linking monomer 1 to the glass support is cleaved. [See S. L. Beaucage and M. H. Caruthers, 1981, *Tetrahedron Lett.* **22**:1859.]

(a)





(b) λ-Phage genome



◀ FIGURE 7-10 (a) Electron micrograph of bacteriophage λ virions and schematic diagram of one virion. (b) Simplified map of the λ-phage genome. Genes encoding proteins required for assembly of the head and tail map at the left end; those encoding additional proteins required for the lytic cycle map at the right end. Some regions of the genome can be replaced by exogenous DNA (diagonal lines) or deleted (dotted area) without affecting the ability of λ phage to infect host cells and assemble new virions, permitting insertion of up to ≈25 kb of DNA between the *J* and *N* genes. The λ genome has been mapped with some 60 different genes. Only a few individual genes are shown in this diagram. Small numbers indicate positions in kilobases (kb). [Photograph courtesy of R. Duda and R. Hendrix.]

that can be detected on a typical petri dish. These limitations make plasmid cloning of all the genomic DNA of higher organisms impractical. For example, $\approx 1.5 \times 10^5$ clones carrying 20-kb DNA fragments are required to represent the total human haploid genome, which contains $\approx 3 \times 10^9$ base pairs. Fortunately, cloning vectors derived from bacteriophage λ have proved to be a practical means for obtaining the required number of clones to represent large genomes. Such a collection of λ clones that includes *all* the DNA sequences of a given species is called a *genomic library*. Once a genomic library is prepared, it can be screened for λ clones containing a sequence of interest.

Bacteriophage λ Can Be Modified for Use as a Cloning Vector and Assembled In Vitro

Bacteriophage λ is probably the most extensively studied bacterial virus, and a great deal is known about its molecular biology and genetics. A λ-phage virion has a head region, which contains the viral DNA genome, and a tail region, which functions in infecting *E. coli* host cells (Figure 7-10a). When a λ virion infects a host cell, only the λ DNA enters the cell. As discussed in Chapter 6, the viral DNA then undergoes either lytic or lysogenic growth. In lytic growth, the viral DNA is replicated and assembled into more than 100 progeny virions in each infected cell, killing the cell in the process and releasing the replicated virions (see Figure 6-17). In lysogenic growth, the viral DNA inserts in the bacterial chromosome where it is passively replicated along with the host-cell chromosome as the cell grows and divides (see Figure 6-19).

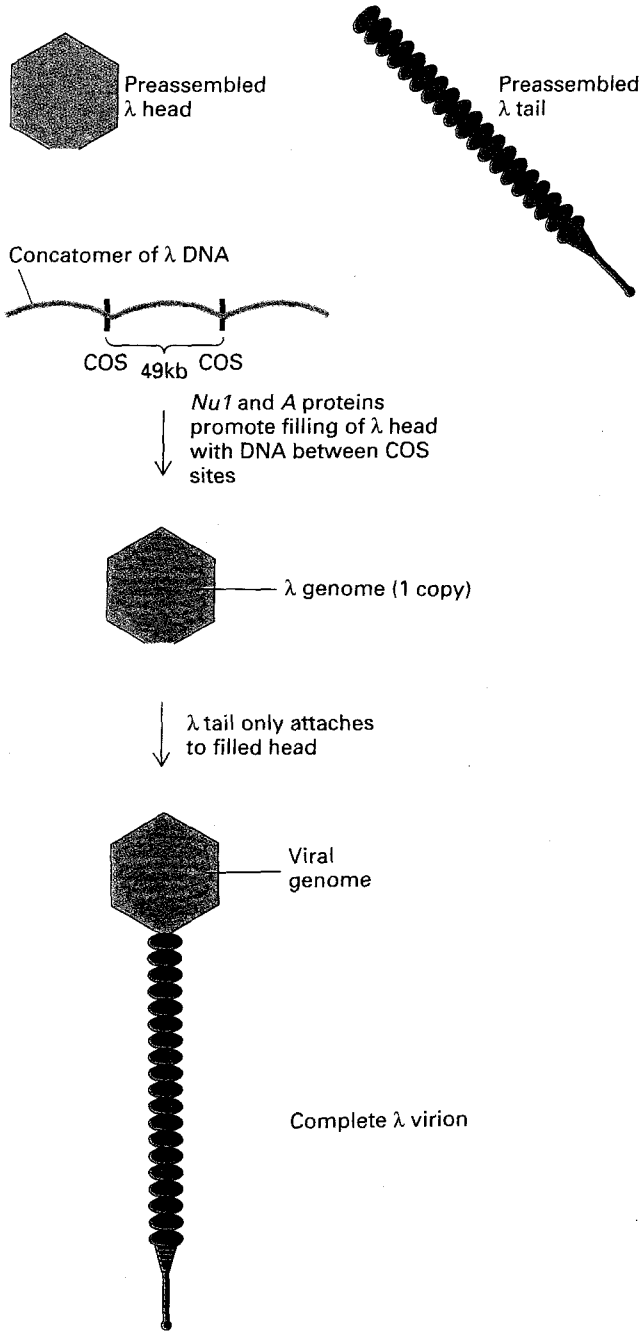
As diagrammed in Figure 7-10b, the λ genes encoding the head and tail proteins as well as various proteins involved in the lytic and lysogenic growth pathways are clustered in discrete regions of the ≈50-kb viral genome. The genes involved in the lysogenic pathway and other viral genes not essential for the lytic pathway are irrelevant for use of bacteriophage λ as a cloning vector. These genes are removed from the viral DNA and replaced with other DNA sequences of interest. Up to ≈25 kb of foreign DNA

can be inserted into the λ genome, resulting in a recombinant DNA that can be packaged to form virions capable of replicating and forming plaques on *E. coli* host cells.

The key to the high efficiency of λ-phage cloning is the ability to assemble λ virions in vitro. During the in vivo assembly of λ virions within infected host cells, viral heads and tails initially are assembled separately from multiple copies of the various proteins that compose these complex structures. Replication of λ DNA in a host cell generates long multimeric DNA molecules, call *concatomers,* that consist of multiple copies of the viral genome linked end to end and separated by specific nucleotide sequences called *COS sites.* Two λ proteins, designated Nu1 and A, bind to COS sites and direct insertion of the DNA between two adjacent COS sites into a preassembled head. This process results in the packaging of a single ≈50-kb λ genome from the multimeric concatomer into each preassembled head. Host-cell chromosomal DNA is not inserted into the λ heads because it does not contain any copies of the COS sequence. Once λ DNA is inserted into a preassembled λ head, the preassembled tail is attached producing a complete virion (Figure 7-11).

To prepare infectious λ virions carrying recombinant DNA, the phage-assembly process is carried out in vitro. In one method, *E. coli* cells are infected with a λ mutant defective in A protein, one of the two proteins required for packaging λ DNA into preassembled phage heads. These cells accumulate preassembled "empty" heads; since tails only attach to heads "filled" with DNA, preassembled tails also accumulate in these cells. After these cells are lysed experimentally, an extract containing high concentrations of heads and tails is prepared. When this extract is mixed with A protein and recombinant λ DNA, which contains a COS site, the DNA is packaged into the empty heads. The tails in the extract then combine with the filled heads, yielding complete virions carrying the recombinant λ DNA.

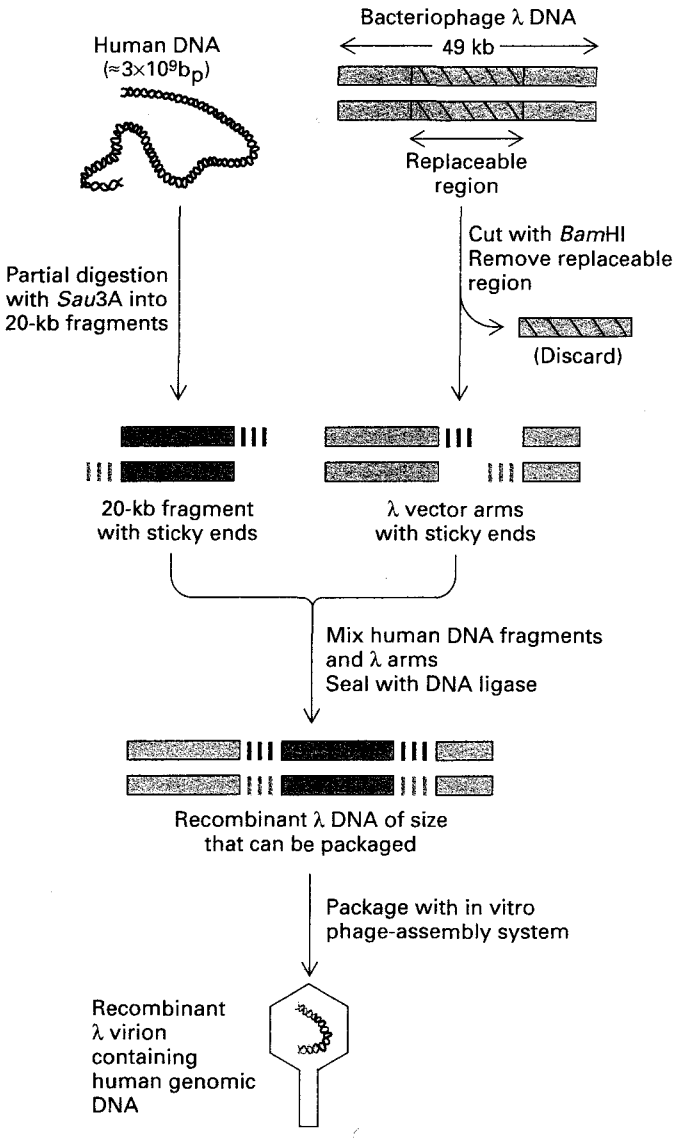
The recombinant virions produced by this method are fully infectious and can efficiently infect *E. coli* cells. Each virion particle binds to receptors on the surface of a host cell and injects its packaged recombinant DNA into the cell. This infection process is about a thousand times more efficient than transformation with plasmid vectors. For instance, $\approx 10^6$ transformed colonies per microgram of recombinant plasmid DNA can be obtained routinely, whereas $\approx 10^9$ plaques representing λ clones can be obtained per microgram of recombinant λ DNA.



▲ FIGURE 7-11 Assembly of bacteriophage λ virions. Empty heads and tails are assembled from multiple copies of several different λ proteins. During the late stage of λ infection, long DNA molecules called concatomers are formed; these multimeric molecules consist of copies of the λ genome linked end to end and separated by COS sites (red), a protein-binding nucleotide sequence that occurs once in each copy of the λ genome. Binding of the λ proteins Nu1 and A to COS sites promotes insertion of the DNA between two adjacent COS sites into an empty head. After the heads are filled with DNA, preassembled λ tails are attached, producing complete λ virions capable of infecting *E. coli* cells.

Nearly Complete Genomic Libraries of Higher Organisms Can Be Prepared by λ Cloning

With the availability of λ-phage cloning vectors, preparation of genomic libraries for higher organisms, including



▲ FIGURE 7-12 Construction of a genomic library of human DNA in a bacteriophage λ vector. The λ DNA first is treated to remove the replaceable region (see Figure 7-10b); the non-essential regions (dotted areas in Figure 7-10b) have been deleted from λ vectors to maximize the size of the exogenous DNA fragment that can be inserted. In this example, the replaceable region of the λ DNA is cut out with *Bam*HI, and the total DNA from human cells is partially digested with *Sau*3A. These two restriction enzymes produce fragments with complementary sticky ends (red lines). The λ vector arms and ≈20-kb genomic fragments are mixed, ligated, and packaged in vitro to produce recombinant λ-phage virions, which are plated on a lawn of *E. coli* cells. Multiple λ vectors have been constructed containing different restriction sites, so that restriction fragments generated by a variety of restriction enzymes can be cloned in λ vectors. A particular 20-kb region of the human genome would occur approximately once in every 1.5×10^5 recombinant λ virions. However, about 10^6 recombinant phages are required to assure that fragments encompassing the entire human genome have a 90–95 percent chance of being included in the library.

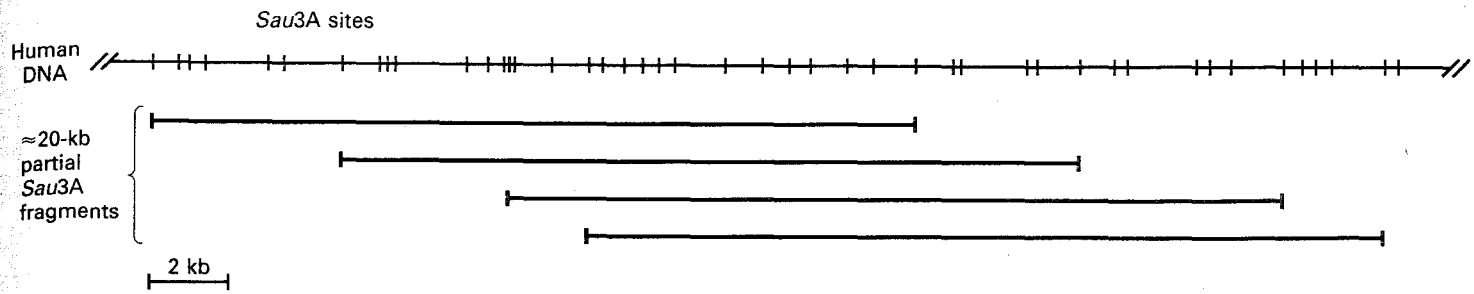
humans, now is feasible. A genomic library is a set of λ (or plasmid) clones that collectively contain every DNA sequence in the genome of a particular organism. Figure 7-12 summarizes the general procedure for constructing a λ genomic library. The λ DNA first is treated with a restriction endonuclease, producing fragments with sticky ends, called λ *vector arms*, that contain all the genes necessary for lytic growth. This step frees the nonessential region in the middle of the λ genome; this region is separated from the λ arms and discarded. Genomic DNA then is extracted from a cell type that contains all the genetic information of the organism under study. Sperm cells or cells of an early embryo often are used as sources of mammalian DNA. The extracted DNA then is cleaved by a restriction enzyme to produce ≈20-kb fragments with sticky ends complementary to the sticky ends on the λ vector arms being used.

The λ arms and the collection of genomic DNA fragments are mixed in about equal amounts. The complementary sticky ends on the fragments and λ arms hybridize and then are joined covalently by DNA ligase. Each of the resulting recombinant DNA molecules contains a foreign DNA fragment located between the two arms of the λ vector DNA. The ligated recombinant DNAs then are packaged into λ virions in vitro as described above. Only DNA molecules of the correct size can be packaged to produce fully infectious recombinant λ virions.

Finally, the recombinant λ virions are plated on a lawn of *E. coli* cells to generate a large number of recombinant λ plaques. Since each plaque arises from a single recombinant virion, all the progeny λ phages that develop are genetically identical and constitute a clone carrying a particular genomic DNA insert. The different plaques correspond to distinct phage clones, each carrying a different DNA insert, and collectively they constitute a λ genomic library.

In constructing a library of genomic DNA, the DNA commonly is cleaved with a restriction enzyme that recognizes a 4-bp restriction site (e.g., *Sau*3A as shown in Figure 7-12). A specific 4-bp sequence will occur on average once every $4^4 = 256$ base pairs. Complete digestion of the human haploid genome, which contains $\approx 3 \times 10^9$ base pairs, would yield somewhat more than 10^7 nonoverlapping different fragments. However, to increase the probability that all regions of the genome are successfully cloned and represented in the λ genomic library, the genomic DNA usually is only partially digested to yield overlapping restriction fragments of ≈20 kb (Figure 7-13). In a large λ library constructed from such overlapping restriction fragments, a specific sequence of genomic DNA may be contained in several "overlapping" clones.

The size of a genomic library for a given organism depends on the amount of DNA in that organism's haploid genome. If the human genome of about 3×10^9 base pairs is cleaved into 20-kb fragments for insertion into a λ vector, then roughly 1.5×10^5 different recombinant λ-phage virions would be required to constitute a complete library. Because the restriction fragments of human DNA are in-



▲ FIGURE 7-13 Production of overlapping restriction fragments by partial digestion of human genomic DNA with *Sau*3A. This restriction endonuclease recognizes the 4-bp sequence GATC and produces fragments with single-stranded sticky ends of the same sequence on the 5′ end of each strand. A hypothetical region of human genomic DNA showing the *Sau*3A recognition sites (red) is shown at the top. Partial digestion of this region of DNA would yield a variety of overlapping fragments (blue) ≈20 kb long. Use of such overlapping fragments increases the probability that all sequences in the genomic DNA will be represented in a λ library. In a large λ library, a specific sequence of genomic DNA (e.g., the gene encoding β-globin) often is contained in or extends over several "overlapping" clones, generated from overlapping restriction fragments.

corporated into phages randomly, about 10^6 recombinant phages are necessary to assure that each region of human DNA has a 90–95 percent chance of being included.

Each plaque produced by a recombinant bacteriophage λ contains large numbers of recombinant virions and, consequently, large numbers of cloned DNA fragments. Hybridization methods for identifying recombinant λ clones of interest are described in a later section. Because these methods allow specific detection of very small plaques, as many as 5×10^4 plaques can be screened on a typical petri dish. Thus only 20–30 petri dishes, each containing about 5×10^4 λ plaques, are sufficient to represent the entire human genome. In contrast, to screen 10^6 recombinant plasmids carrying the entire human genome would require about 5000 petri dishes, because only 200 or so transformed *E. coli* colonies can be detected on a typical petri dish.

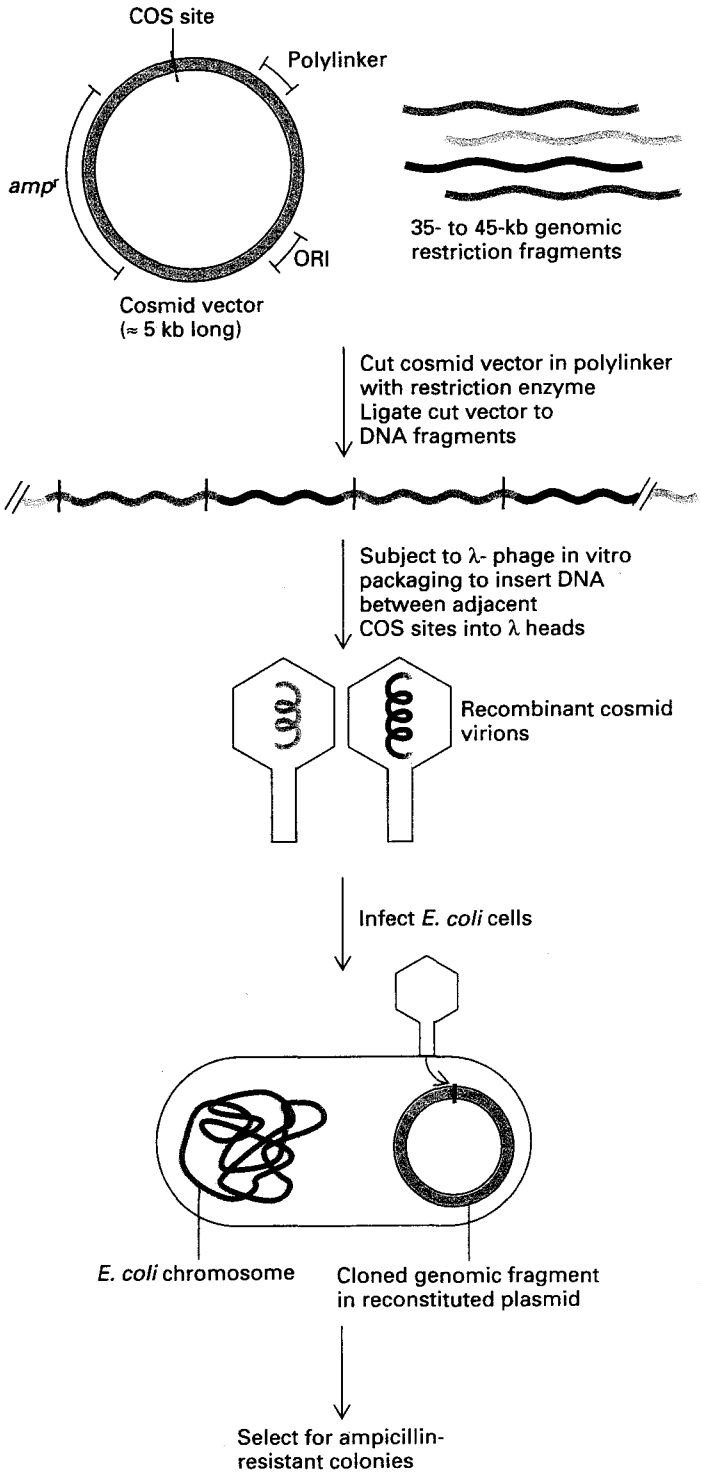
Larger DNA Fragments Can Be Cloned in Cosmids and Other Vectors

Both λ-phage and *E. coli* plasmid vectors are useful for cloning only relatively small DNA fragments. The largest DNA fragment that can be cloned in a λ-phage vector is about 25 kb long; insertion of larger fragments produces recombinant λ DNA that is too large to be packaged into viral heads. Plasmid cloning is practical only with fragments up to about 20 kb long because the efficiency of *E. coli* transformation with plasmid DNA falls off rapidly for larger plasmids. Several other vectors, however, have been developed for cloning larger fragments of DNA (Table 7-2).

One common method for cloning larger fragments makes use of elements of both plasmid and λ-phage cloning. In this method, called *cosmid cloning,* recombinant plasmids containing inserted fragments up to 45 kb long can be efficiently introduced into *E. coli* cells. A cosmid vector is produced by inserting the COS sequence from λ-phage DNA into a small *E. coli* plasmid vector about 5 kb long. Like other plasmid vectors discussed earlier, cosmid vectors contain a replication origin (ORI), an antibiotic-resistance gene (e.g., *amp*^{r}), and a polylinker sequence containing numerous restriction-enzyme recognition sites (Figure 7-14). Next, the cosmid vector is cut with a restriction enzyme and then ligated to 35- to 45-kb restriction fragments of foreign DNA with complementary sticky ends. If the concentration of foreign DNA is high enough, the ligation reaction generates long DNA molecules containing multiple restriction fragments of the foreign DNA separated by the 5-kb cosmid DNA. These ligated DNA molecules, which resemble the concatomers that form during replication of λ phage in a host cell, can be packaged in vitro as described earlier.

TABLE 7-2 Maximum Size of DNA That Can Be Cloned in Vectors

Vector Type	Length of Cloned DNA (kb)
Plasmid	20
Bacteriophage λ	25
Cosmid	45
P1 vector	100
YAC (yeast artificial chromosome)	1000



▲ FIGURE 7-14 General procedure for cloning DNA fragments in cosmid vectors. This procedure has the high efficiency associated with λ-phage cloning and permits cloning of restriction fragments up to ≈45 kb long. In this example, four different types of recombinant cosmid virions could be generated, each carrying one of the genomic fragments (green and orange). Plating of the recombinant virions on *E. coli* cells would yield four different types of colonies. Note that the lengths of vector DNA and genomic fragments are not to scale. See text for further discussion.

In the packaging reaction, the λ Nu1 and A proteins bind to COS sites in the ligated DNA and direct insertion of the DNA between two adjacent COS sites into empty phage heads. Packaging will occur as long as the distance between adjacent COS sites does not exceed about 50 kb (the approximate size of the λ genome). Phage tails then are attached to the filled heads, producing viral particles that contain a recombinant cosmid DNA molecule rather than the λ genome. When these virions are plated on a lawn of *E. coli* cells, they bind to phage receptors on the cell surface and inject the packaged DNA into the cells.

Since the injected DNA does not encode any λ proteins, no viral particles form in infected cells and no plaques develop on the plate. Rather, the injected DNA circularizes, forming in each host cell a large plasmid containing the cosmid vector and an inserted DNA fragment. This plasmid replicates and is segregated to daughter cells like other *E. coli* plasmids (see Figure 7-3), and the colonies that arise from transformed cells can be selected on antibiotic plates. The high efficiency of λ-phage infection of *E. coli* cells makes cosmid cloning a practical method of generating plasmid clones carrying DNA fragments up to 45 kb long. Since many genes of higher eukaryotes are on the order of 30–40 kb in length, cosmid cloning increases the chances of obtaining DNA clones containing the entire sequences of genes.

A recently developed approach similar to cosmid cloning makes use of larger *E. coli* viruses such as bacteriophage P1. Recombinant plasmids containing DNA fragments up to ≈100 kb long can be packaged in vitro with the P1 system. Still larger fragments of DNA can be cloned in yeast artificial chromosomes (YACs). To understand how YACs function requires an explanation of the basic elements of eukaryotic chromosomes; this topic is covered in Chapter 9.

➤ *Construction of a cDNA Library*

In higher eukaryotes, many genes are transcribed into mRNA only in specialized cell types. For example, mRNAs encoding globin proteins are found only in erythrocyte precursor cells, called reticulocytes. Likewise, the mRNA encoding albumin, the major protein in serum, is produced only in liver cells where albumin is synthesized. The specific DNA sequences expressed as mRNAs in a particular cell type can be cloned by synthesizing DNA copies of the mRNAs isolated from that type of cell, and then cloning the DNA copies in plasmid or bacteriophage λ vectors. DNA copies of mRNAs are called *complementary DNAs* (cDNAs); clones of such DNA copies of mRNAs are called *cDNA clones*. In addition to representing only the sequences expressed as mRNAs in a particular cell type, cDNA clones have the advantage that introns present in

genomic DNA clones are absent from cDNA clones. Just as a large collection of clones containing fragments of genomic DNA representing the entire genome of a species is called a genomic library, a large collection of cDNA copies of all the mRNAs in a cell type is called a *cDNA library*.

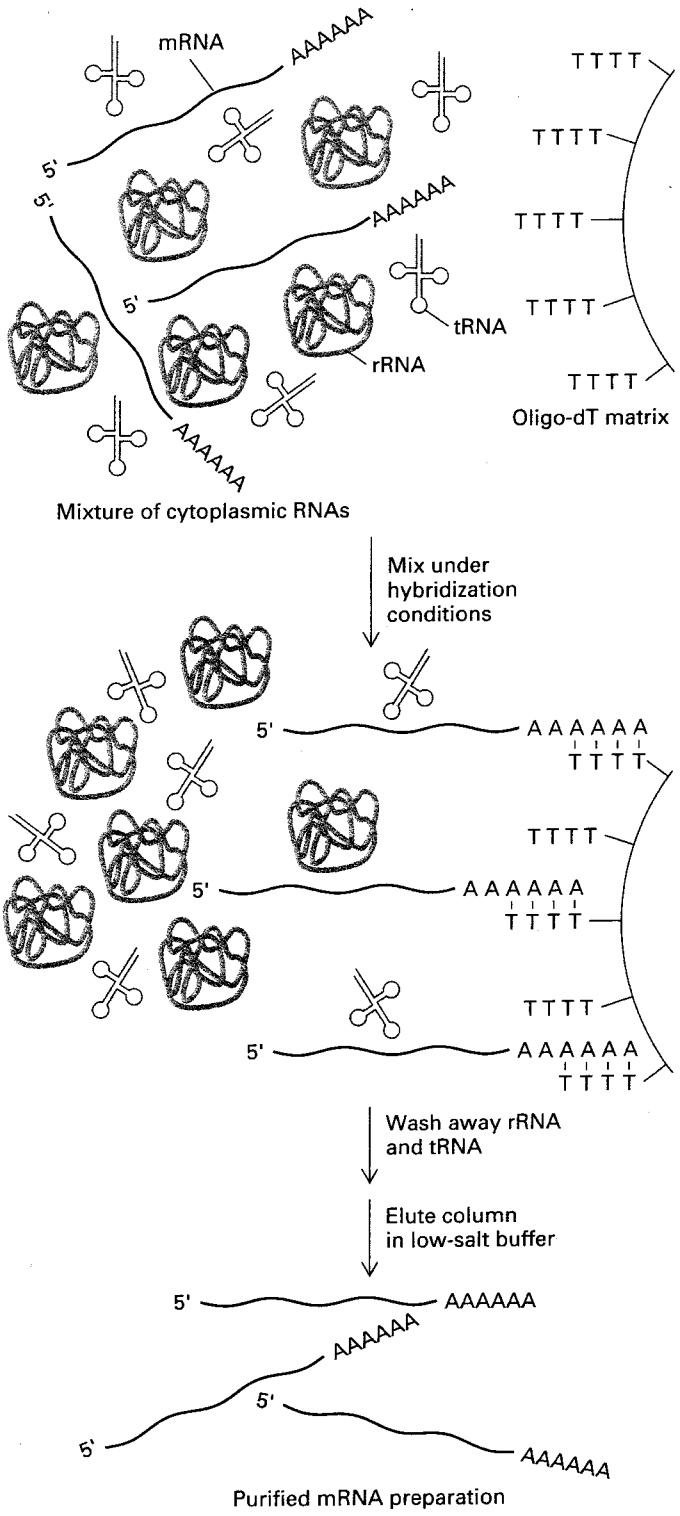
cDNAs Are Produced by Copying Isolated mRNAs with Reverse Transcriptase

The first step in preparing a cDNA library is to isolate the total mRNA from the cell type or tissue of interest. Nature has greatly simplified the isolation of eukaryotic mRNAs: the 3′ end of nearly all eukaryotic mRNAs consists of a string of 50–250 adenylate residues, called a *poly-A tail*. Because of their poly-A tail, mRNAs can be easily separated from the much more prevalent rRNAs and tRNAs present in a cell extract by use of a column to which short strings of thymidylate (oligo-dTs) are linked to the matrix (Figure 7-15). When a cell extract is passed through an oligo-dT column, the mRNA poly-A tails base-pair with the oligo-dTs, binding the mRNAs to the column. Since rRNAs, tRNAs, and other molecules do not bind to the column, they can be washed away. The bound mRNAs are recovered by elution with a low-salt buffer.

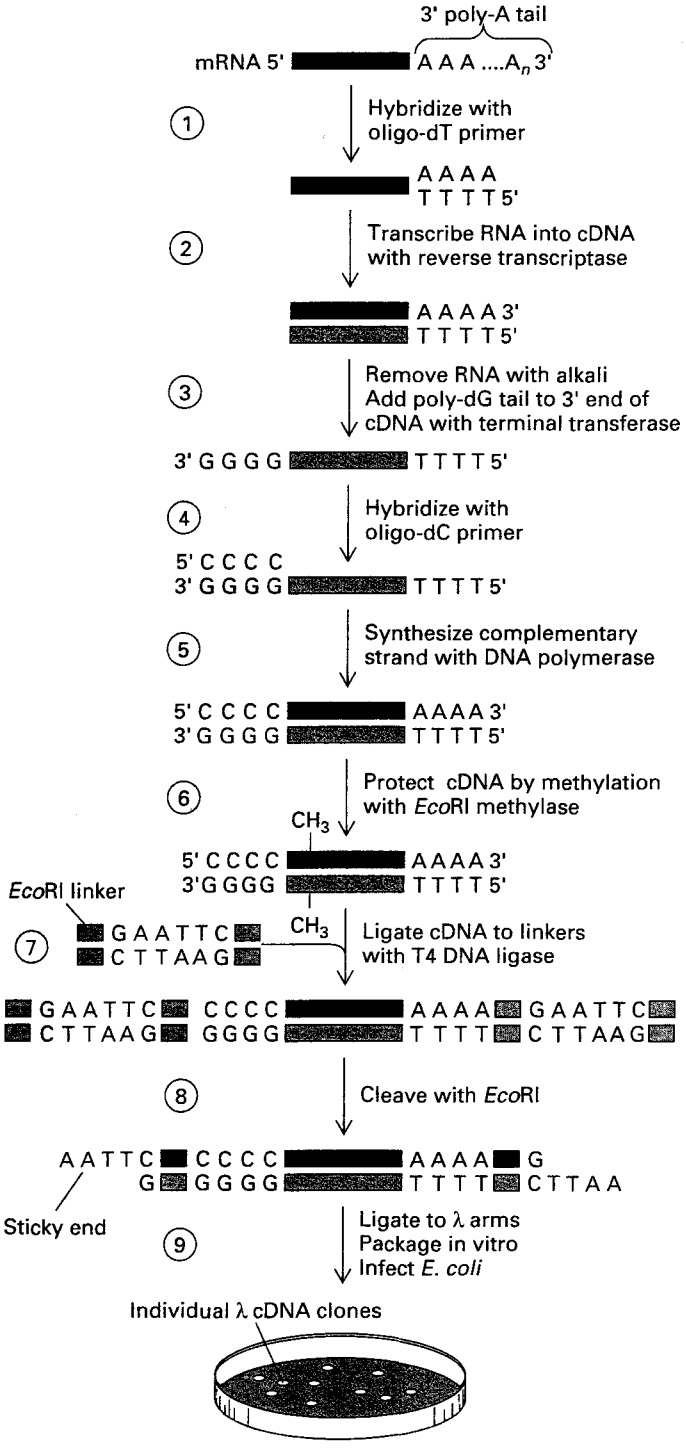
The enzyme *reverse transcriptase*, which is found in retroviruses (see Figure 6-23) is then used to synthesize a strand of DNA complementary to each mRNA molecule (Figure 7-16). This enzyme can polymerize deoxynucleoside triphosphates into a complementary DNA strand using an RNA molecule as template. Like other DNA polymerases, this enzyme can only add nucleotides to the 3′ end of a preexisting primer base-paired to the template. Added free oligo-dT serves this function by hybridizing to the 3′ poly-A tail of each mRNA template.

cDNAs Can Be Enzymatically Converted to Double-Stranded DNA and Cloned

After cDNA copies of isolated mRNAs are synthesized, the mRNAs are removed by treatment with alkali, which hydrolyzes RNA but not DNA. The single-stranded cDNAs then are converted to double-stranded DNA molecules. To do this, the 3′ end of each cDNA strand is elongated by adding several residues of a single nucleotide (e.g., dG) through the action of *terminal transferase*, a unique DNA polymerase that does not require a template, but simply adds deoxynucleotides to free 3′ ends. A chemically synthesized oligo-dC primer then is hybridized to this 3′ oligo-dG. A DNA strand complementary to the original cDNA strand then is synthesized by a DNA polymerase, which uses the oligo-dC as a primer. These reactions produce a complete double-stranded DNA molecule corresponding to each of the mRNA molecules in the original preparation. Each double-stranded-DNA contains an oligo-dC–oligo-dG double-stranded region at one end and an oligo-dT–oligo-dA double-stranded region at the other end.



▲ FIGURE 7-15 Isolation of eukaryotic mRNA by oligo-dT column affinity chromatography. Isolated cytoplasmic RNA consists mostly of ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) shown in orange and blue. The much less abundant mRNAs (red) have 3′ poly-A tails, which hybridize to oligo-dT covalently coupled to the column matrix. After hybridization, the rRNAs and tRNAs are washed out of the column; then the mRNAs are eluted with a low-salt buffer. The resulting purified mRNA preparation contains many different mRNA molecules encoding different proteins.



◀ FIGURE 7-16 Preparation of a bacteriophage λ cDNA library. A mixture of mRNAs, isolated as shown in Figure 7-15, is used to produce cDNAs corresponding to all the cellular mRNAs (steps 1–3). The single-stranded cDNAs (light green) are then converted into double-stranded cDNAs, which are treated with *Eco*RI methylase to prevent subsequent digestion by *Eco*RI (steps 4–6). The protected double-stranded cDNAs are ligated to a synthetic double-stranded *Eco*RI-site linker at both ends and then cleaved with the corresponding restriction enzyme, yielding cDNAs with sticky ends (red letters); these are incorporated into λ-phage cloning vectors, and the resulting recombinate λ virions are plated on a lawn of *E. coli* cells (steps 7–9). See text for further discussion.

Next, short double-stranded DNA segments (usually ≈10–12 base pairs long) containing the recognition site for a particular restriction enzyme are ligated onto both ends of the double-stranded cDNAs. These short fragments, called restriction-site *linkers*, are prepared by hybridizing chemically synthesized complementary oligonucleotides. As noted earlier, the DNA ligase from bacteriophage T4 can ligate "blunt-ended" double-stranded DNA molecules that do not have sticky ends. Such blunt-end ligations are less efficient than the ligation of DNA fragments with compatible sticky ends. Nonetheless, the ligation reaction can be driven to completion by using high concentrations of linkers. The resulting double-stranded cDNAs, which contain a restriction-site linker at each end, are treated with the restriction enzyme specific for the linker; this generates cDNA molecules with sticky ends at each end. To prevent digestion of any cDNAs that by chance have a recognition sequence for this restriction enzyme within the cDNA sequence, the mixture of double-stranded cDNAs is treated with the appropriate modification enzyme before addition of the linkers. This enzyme methylates specific bases within the restriction-site sequence, preventing the restriction enzyme from digesting the methylated sites (see Figure 7-5b).

The final step in construction of a cDNA library is ligation of the restriction-cleaved double-stranded cDNAs, which now have sticky ends, to plasmid or λ-phage vectors that have been cut to generate complementary sticky ends. The recombinant vectors then are plated on a lawn of *E. coli* cells, producing a library of plasmid or λ clones (see Figure 7-16). Each clone carries a cDNA derived from a single mRNA.

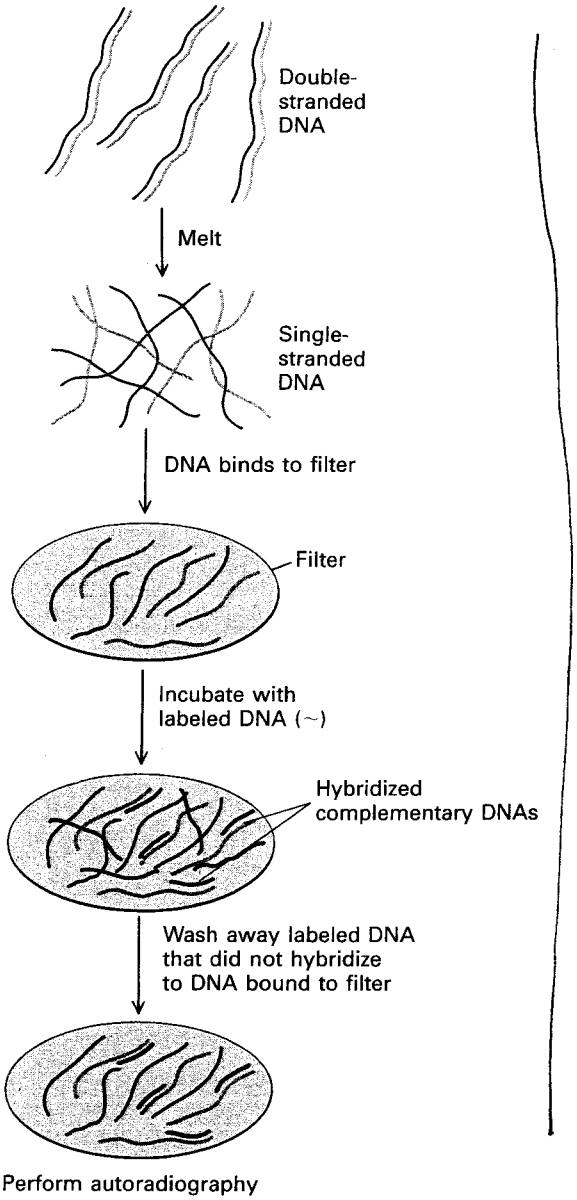
➤ *Identification of Specific Clones in a Genomic or cDNA Library*

Suppose you have isolated a particular protein and want to isolate the gene that encodes it. A complete genomic λ library from mammals contains at least a million different clones; a cDNA library must contain as many clones to include the sequences of scarce mRNAs. How are specific clones of interest identified in such large collections? The most common method involves screening a library by *hybridization* with radioactively labeled DNA or RNA *probes*. In an alternative method, called *expression cloning*, a specific clone in a library of cloned DNA is identified based on some property of its encoded protein. In this section, we first discuss the hybridization method and then describe expression cloning.

Membrane Hybridization Can Be Used to Screen a Library

As discussed in Chapter 4, under the conditions of temperature and ion concentration found in cells, DNA is maintained as a duplex (two-stranded) structure by the hydrogen bonds that form between the A and T bases and G and C bases in each strand (see Figure 4-7). DNA duplexes can be denatured (melted) into single strands by heating them, usually in a dilute salt solution (e.g., 0.01 M NaCl), or by raising the pH above 11. If the temperature is lowered and the ion concentration in the solution is raised, or if the pH is lowered to neutrality, the A-T and G-C base pairs reform between complementary single strands (see Figure 4-11). This process goes by many names: renaturation, reassociation, hybridization, annealing. In a mixture of nucleic acids, only complementary single strands (or strands containing complementary regions) will reassociate; the extent of their reassociation is virtually unaffected by the presence of noncomplementary strands. Such *molecular hybridization* can take place between two complementary strands of either DNA or RNA, or between an RNA strand and a DNA strand.

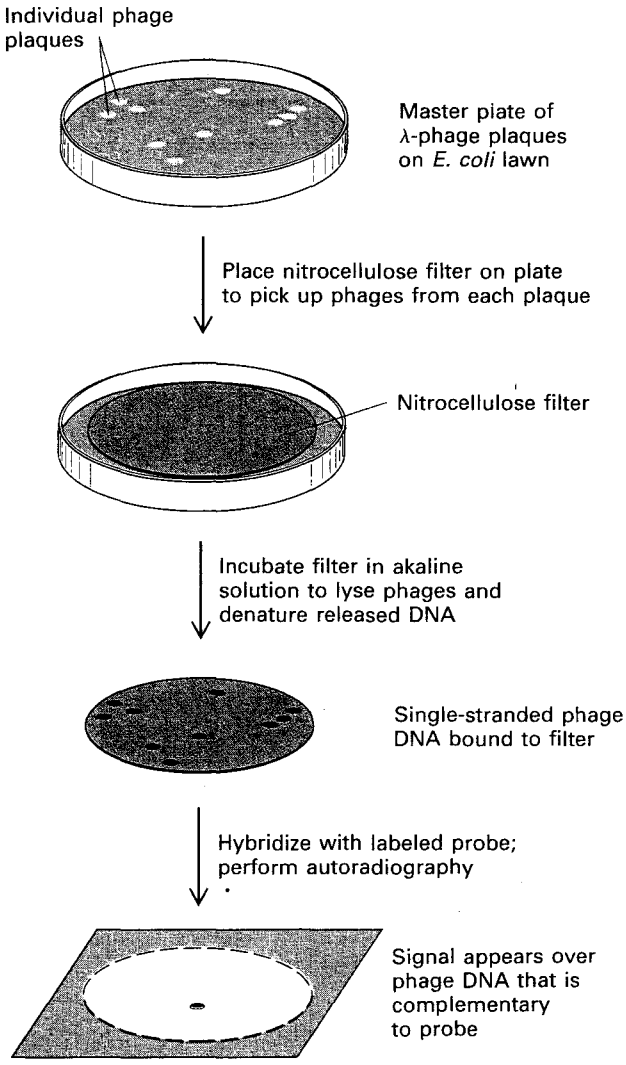
To utilize molecular hybridization in the detection of specific DNA clones, single-stranded DNA from recombinant DNA molecules is attached to a solid support, commonly a nitrocellulose filter or treated nylon membrane (Figure 7-17). When a solution containing single-stranded nucleic acids is dried on such a membrane, the single strands become irreversibly bound to the solid support in a manner that leaves most of the bases available for hybridizing to a complementary strand. Although the chemistry of this irreversible binding is not well understood, the procedure is very useful. The membrane is then incubated in a solution containing a radioactively labeled single-stranded DNA (or RNA) that is complementary to some of the nucleic acid bound to the membrane. Under hybridization conditions (near neutral pH, 40–65°C, 0.3–0.6 M NaCl), this labeled probe hybridizes to the complementary nucleic acid bound to the membrane. Any excess probe that does not hybridize is washed away, and the labeled hybrids are detected by autoradiography of the filter.



▲ FIGURE 7-17 Membrane-hybridization assay for detecting complementary regions of DNA (and RNA). This general assay, which is widely used in recombinant DNA technology, permits as little as 1 part in 10^6 of a particular DNA or RNA to be detected.

The procedure for screening a λ library with this membrane-hybridization technique is outlined in Figure 7-18. The recombinant λ virions present in plaques on a lawn of *E. coli* are transferred to a nylon membrane by placing the membrane on the surface of the petri dish. Many of the viral particles in each plaque absorb to the surface of the membrane, but many virions remain in the plaques on the surface of the nutrient agar in the petri dish. In this way a *replica* of the petri dish containing a large number of individual λ clones is reproduced on the surface of the membrane. The original petri dish is refrigerated to store the collection of λ clones. The membrane is then incubated in an alkaline solution, which disrupts the virions, releasing and denaturing the encapsulated DNA. The membrane is then dried, fixing the recombinant λ DNA to the membrane's surface. Next, the membrane is incubated with a radiolabeled probe under hybridization conditions. Unhybridized probe is washed away, and the filter is subjected to autoradiography.

The appearance of a spot on the autoradiogram indicates the presence of a recombinant λ clone containing DNA complementary to the probe. The position of the spot on the autoradiogram corresponds to the position on the original petri dish where that particular clone formed a



▲ FIGURE 7-18 Identification of a specific clone from a λ-phage library by membrane hybridization to a radiolabeled probe. The position of the signal on the autoradiogram identifies the desired plaque on the plate. In practice, in the initial plating of a library the plaques are not allowed to develop to a visible size so that up to 50,000 recombinants can be analyzed on a single plate. Phage particles from the identified region of the plate are isolated and replated at low density so that the plaques are well separated. Then pure isolates can be obtained by repeating the plaque hybridization as shown in the figure.

plaque. Since the original petri dish still contains many infectious virions in each plaque, viral particles from the identified clone can be recovered for replating by aligning the autoradiogram and the petri dish and removing viral particles from the clone corresponding to the spot. A similar technique can be applied for screening a plasmid library in *E. coli* cells.

Certain cDNAs and Synthetic Oligonucleotides Are Used as Probes

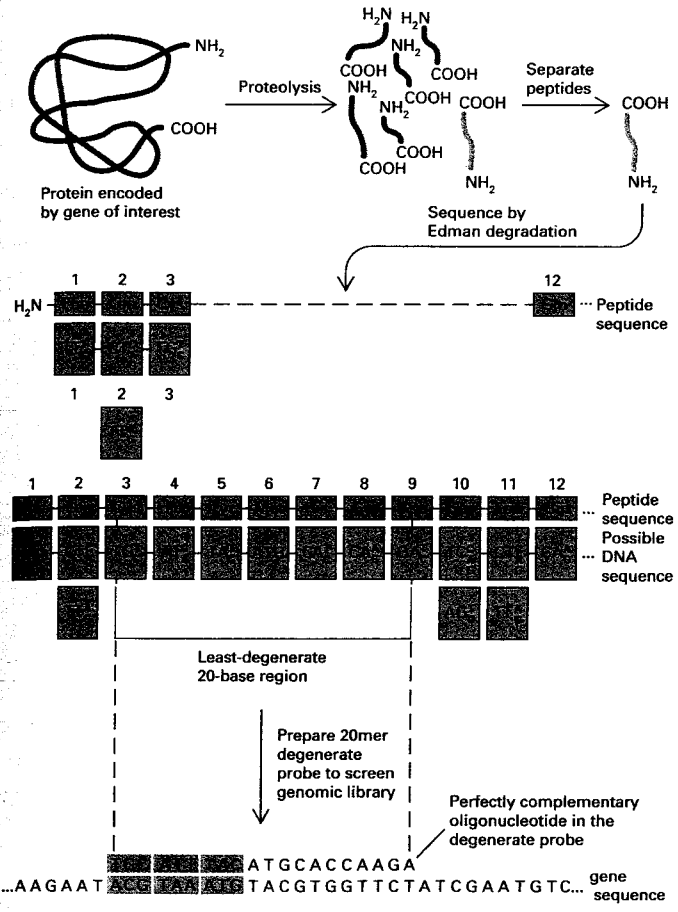
Identification of specific clones by the membrane-hybridization technique depends on the availability of labeled probes that will hybridize with specific DNA sequences. The method used to prepare a particular probe depends on whether the gene of interest encodes a protein that is expressed at a high or low level.

cDNA Probes Prepared from Abundant mRNAs In some cases, a particular mRNA is sufficiently abundant in a particular tissue or cell type that it can be readily isolated and used directly to prepare a probe. For example, over 90 percent of the mRNA isolated from reticulocytes encodes α or β globin. Specific probes for the α- and β-globin genes can be prepared by isolating the globin mRNAs from reticulocytes on an oligo-dT column (see Figure 7-15) and then synthesizing the cDNA with reverse transcriptase in the presence of radiolabeled nucleoside triphosphates. Because of the ease with which globin probes can be prepared, the genes encoding the α and β globins were among the first to be cloned. This same approach has been used to prepare probes directly from several other mRNAs expressed at high levels in certain tissues or cell types.

Synthetic Oligonucleotide Probes In many cases, however, a gene of interest encodes a protein that is not expressed at a high level. For example, the enzyme β-*N*-hexosaminidase A, which is defective in the inherited disorder Tay-Sachs disease (see Figure 5-42) is present at a very low concentration in human cells. Likewise, the mRNA encoding this enzyme is so rare that it cannot be isolated directly and used to prepare a cDNA probe. Nonetheless, specific oligonucleotide probes for the genes encoding this and other low-abundance proteins can be chemically synthesized based on their amino acid sequence.

Because an oligonucleotide probe containing only ≈20 nucleotides is sufficient to screen a library, only a small portion (consisting of six or seven amino acids) of the total protein needs to be sequenced. To prepare a specific probe by this method, the protein of interest usually is purified by sequential column chromatography and SDS-polyacrylamide gel electrophoresis (Chapter 3). The purified protein is digested with one or more proteases (e.g., trypsin) into specific peptides (Figure 7-19). The N-terminal amino acid sequences of a few of these peptides is determined by sequential Edman degradation (see Figure 3-6). Based on the genetic code, oligonucleotide probes encoding the determined peptide sequences can be synthesized and radiolabeled (see Figure 7-9).

In choosing which peptide sequences to use in the design of oligonucleotide probes, the degeneracy of the genetic code must be considered (see Table 4-4). As a result of this degeneracy, many amino acids are encoded by multiple



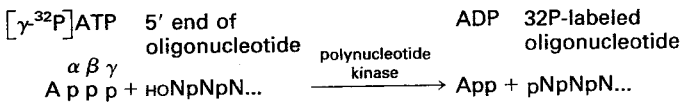
▲ FIGURE 7-19 Designing oligonucleotide probes based on protein sequence. An isolated protein is digested with a selective protease such as trypsin, which specifically cleaves peptide bonds on the carboxyl-terminal side of lysine and arginine residues. The resulting peptides are separated, and several are partially sequenced from their N-terminus by sequential Edman degradation (see Figure 3-6b). The determined sequences then are analyzed to identify the 6- or 7-amino acid region that can be encoded by the smallest number of possible DNA sequences. Because of the degeneracy of the genetic code, the 12-aa sequence (light green) shown here theoretically could be encoded by any of the DNA triplets below it, with the possible alternative bases at the same position indicated. For example, Phe-1 is encoded by TTT or TTC; Leu-2 is encoded by one of six possible triplets (CTT, CTC, CTA, CTG, TTA, or TTG). The region with the least degeneracy for a sequence of 20 bases (20mer) is indicated by the dark red bracket. There are 48 possible DNA sequences in this 20-base region that could encode the peptide sequence Cys-Ile-Tyr-Met-His-Gln-Asp. Since the actual sequence of the gene is unknown, a degenerate 20mer probe consisting of a mixture of all the possible 20-base oligonucleotides is prepared. If a genomic library is screened with this degenerate probe, the one oligonucleotide that is perfectly complementary to the actual gene sequence (blue) will hybridize to it.

codons. Since the specific codons used to encode particular amino acids in the protein of interest are unknown, oligonucleotides containing all possible combinations of codons must be synthesized to assure a perfect match with the gene. For this reason, peptides containing arginine, leucine, or serine are to be avoided if possible, since six different codons encode each of these amino acids. The best amino acids for making such probes are tryptophan and methionine (one codon each); the next best are phenylalanine, tyrosine, histidine, aspartic acid, glutamic acid, asparagine, cysteine, lysine, and glutamine (two codons each).

Once several peptides have been sequenced, the 6- or 7-aa stretch that can be encoded by the smallest number of possible DNA sequences is determined. This approach is illustrated in Figure 7-19 for a partial amino acid sequence. For example, the amino acids in the sequence extending from position 3 through 8 (Cys-Ile-Tyr-Met-His-Gln) can be encoded by 2, 3, 2, 1, 2, and 2 possible codons, respectively. Consequently, 48 ($= 2 \times 3 \times 2 \times 1 \times 2 \times 2$) different 18-base DNA sequences could encode this one sequence of amino acids. The GA added at the 3′ end of these 18-base sequences must be complementary to the gene since the next amino acid in this peptide, Asp-9, is encoded by two codons that both start with GA. To be certain of obtaining a probe based on this amino acid sequence that hybridizes perfectly to the unique sequence present in the gene, all 48 of the 20mer probes must be synthesized. A mixture of 20mer probes based on any other portion of this peptide sequence would have to contain considerably more than 48 oligonucleotides, because of the presence of leucine and/or serine residues, each encoded by six different codons.

In practice, rather than separately synthesizing all possible oligonucleotides that can encode the selected portion of a peptide sequence, researchers usually synthesize a mixture of probes at one time. This can be done by adding more than one nucleotide precursor to the synthesis reaction at those points in the sequence that can be encoded by alternative bases. Such a mixture of oligonucleotides is often called a "degenerate" probe, because it is based on the degeneracy of the genetic code. The final step in preparing this type of probe is to radiolabel the oligonucleotides, usually by transferring a ^{32}P-labeled phosphate group from ATP to the 5′ end of each oligonucleotide using *polynucleotide kinase* (Figure 7-20).

In the most common approach, a radiolabeled degenerate probe is used to screen a λ cDNA library using the membrane-hybridization technique. This screening will identify clones that hybridize to the perfectly complementary oligonucleotide present in the probe mixture. Under the usual experimental conditions, oligonucleotides that differ from the cDNA sequence at one or two bases also will hybridize. The nucleotide sequence of the entire cDNA clone can then be determined by methods described in a later section. From the nucleotide sequence the complete



▲ FIGURE 7-20 Radiolabeling of an oligonucleotide at the 5′ end with phosphorus-32. The three phosphate groups in ATP are designated the α, β, and γ phosphates in order of their position away from the ribose ring. ATP containing the radioactive isotope ^{32}P in the γ-phosphate position is called [γ-^{32}P]ATP. Kinase is the general term for enzymes that transfer the γ-phosphate of ATP to specific substrates. Polynucleotide kinase can transfer the ^{32}P-labeled γ-phosphate of [γ-^{32}P]ATP to the 5′ end of a polynucleotide chain (either DNA or RNA). This reaction is commonly used to radiolabel synthetic oligonucleotides.

amino acid sequence of the encoded protein can be predicted. The nucleotide sequence can be compared with the other peptide sequences initially determined from the peptide fragments to verify that the correct cDNA clone has been isolated.

Once a cDNA clone is obtained, radiolabeled cDNA can be prepared and used to probe a genomic library. This identifies cloned fragments of genomic DNA that contain the gene encoding the protein. Because the genes of higher eukaryotes, often contain large introns, the total length of the gene may be considerably greater than the corresponding cDNA. The genomic DNA also includes sequences outside of the protein-coding region that regulate the transcription of the gene and sequences of the primary RNA transcript required for RNA splicing and polyadenylation.

Expression Cloning Identifies Specific Clones Based on Properties of the Encoded Proteins

Genomic and cDNA libraries can also be screened for the properties of a specific protein encoded in the cloned DNA. This approach uses special cloning vectors, called *λ expression vectors*, in which the cloned DNA is transcribed into the complementary mRNA, which in turn is translated into the encoded protein. For example, λ-phage vectors have been constructed so that the junction of inserted DNA lies in a region of the vector that is transcribed and translated at a high rate. Cloned DNA inserted at this position is transcribed into mRNA in every cell infected by this type of vector. If the cloned DNA contains a protein-coding sequence inserted in the same reading frame as the vector protein, a *fusion protein* is synthesized with an amino terminus encoded by the vector and the remainder of the fusion protein encoded by the cloned DNA (Figure 7-21).

When replica nitrocellulose filters are prepared from a recombinant library constructed in a λ expression vector, fusion proteins expressed from each individual clone are bound to the nitrocellulose filter. The replica filter can be screened by procedures capable of detecting specific fusion proteins. For example, a monoclonal antibody specific for a protein of interest can be incubated with replica filters of a λ cDNA expression library. If one of the λ clones expresses a fusion protein that includes the region of the protein bound by that monoclonal antibody, antibody molecules will bind to the filter at the position of that specific clone. After washing the filter to remove unbound antibody, the position of the specific clone is detected by incubation with a second radioactively labeled antibody that recognizes the first antibody, followed by autoradiography of the filter.

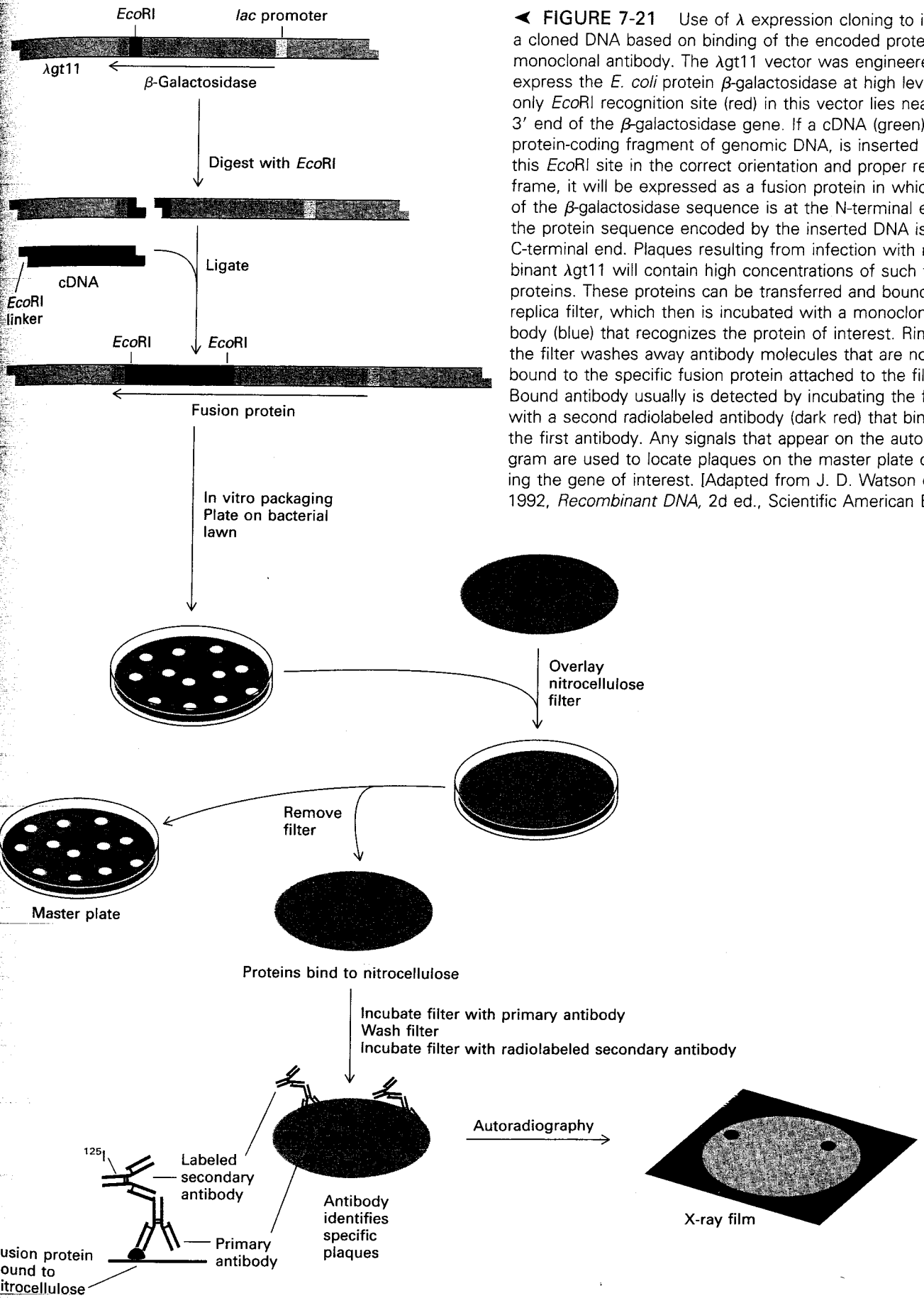
In this method, termed *expression cloning*, any molecule that binds to a protein of interest with high affinity and specificity can be labeled and used as a probe to identify clones expressing the interacting protein. For instance, expression cloning has been useful in identifying cDNA clones encoding proteins that bind to specific DNA sequences; many such proteins are involved in controlling transcription. In this case, a labeled synthetic double-stranded DNA probe is incubated with replica filters prepared from a cDNA library cloned into a λ expression vector. Binding of the labeled DNA by fusion proteins locates the positions of desired clones on the original filter. As described in a later section, other types of expression vectors can be used to produce large amounts of a protein from a cloned gene.

➤ *Analyzing and Sequencing Cloned DNA*

Once a particular genomic DNA fragment or cDNA has been cloned, it can be separated from the vector DNA and analyzed. The most complete characterization of a cloned DNA requires determination of its nucleotide sequence; from this sequence, the amino acid sequence of the encoded protein can be deduced. The sequence of genomic DNA includes introns as well as exons and regions that control gene expression by determining the type of cell in which the encoded protein is expressed as well as the time in development and the amount of protein produced. Genomic DNAs also include replication origins and sequences important in determining how the DNA associates with proteins in chromosomes. In subsequent chapters, we consider how cells use DNA sequences for these functions. In this section, techniques for characterizing and finally sequencing cloned DNA are outlined.

Cleavage with an Appropriate Restriction Enzyme Separates a Cloned DNA from Its Vector

The first step in analyzing a cloned DNA is to separate it from the plasmid or λ vector carrying it. This can be done

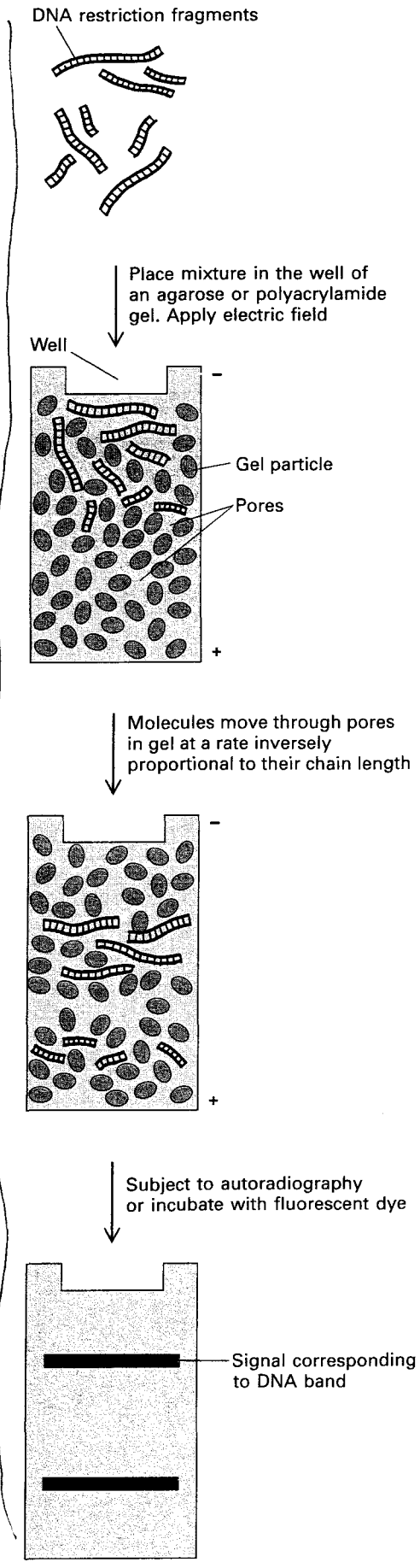


◀ **FIGURE 7-21** Use of λ expression cloning to identify a cloned DNA based on binding of the encoded protein to a monoclonal antibody. The λgt11 vector was engineered to express the *E. coli* protein β-galactosidase at high level. The only *Eco*RI recognition site (red) in this vector lies near the 3′ end of the β-galactosidase gene. If a cDNA (green), or protein-coding fragment of genomic DNA, is inserted into this *Eco*RI site in the correct orientation and proper reading frame, it will be expressed as a fusion protein in which most of the β-galactosidase sequence is at the N-terminal end and the protein sequence encoded by the inserted DNA is at the C-terminal end. Plaques resulting from infection with recombinant λgt11 will contain high concentrations of such fusion proteins. These proteins can be transferred and bound to a replica filter, which then is incubated with a monoclonal antibody (blue) that recognizes the protein of interest. Rinsing the filter washes away antibody molecules that are not bound to the specific fusion protein attached to the filter. Bound antibody usually is detected by incubating the filter with a second radiolabeled antibody (dark red) that binds to the first antibody. Any signals that appear on the autoradiogram are used to locate plaques on the master plate containing the gene of interest. [Adapted from J. D. Watson et al., 1992, *Recombinant DNA*, 2d ed., Scientific American Books.]

by cleaving the recombinant vector with the same restriction enzyme used to insert the DNA fragment initially. During ligation of a cut vector and DNA fragments generated with the same restriction enzyme, the restriction recognition sequence is regenerated between the DNA fragments and vector (see Figure 7-8). Subsequent treatment with the same restriction enzyme will cut the recombinant vector at the same sites, releasing the vector and cloned DNA, which then can be separated by gel electrophoresis.

Gel Electrophoresis Resolves DNA Fragments of Different Size

As described in Chapter 3, gel electrophoresis is a powerful method for separating proteins according to size. Gel electrophoresis also is used to separate DNA and RNA molecules by size. DNA and RNA molecules are highly charged near neutral pH because the phosphate group in each nucleotide contributes one negative charge. As a result, DNA and RNA molecules move toward the positive electrode during gel electrophoresis. Smaller molecules move through the gel matrix more readily than larger molecules, so that molecules of different length, such as restriction fragments, separate (Figure 7-22). Because the gel matrix restricts random diffusion of the molecules, molecules of different length separate into "bands" whose width equals that of the well into which the original DNA mixture was placed. The resolving power of gel electrophoresis is so great that single-stranded DNA molecules up to about 500 nucleotides long can be separated if they differ in length by only one nucleotide. This high resolution is a critical aspect of the procedures for sequencing DNA described later.



➤ **FIGURE 7-22** Separation of DNA fragments of different lengths by gel electrophoresis. A gel is prepared by pouring a liquid containing either melted agarose or unpolymerized acrylamide between two glass plates a few millimeters apart. As the agarose solidifies or the acrylamide polymerizes into polyacrylamide, a gel matrix forms consisting of particles and interconnecting channels, or pores, whose size depends on the concentration of the gel material. Because the pores are larger in agarose gels than in polyacrylamide gels, large DNA fragments are separated on agarose gels and small DNA fragments on polyacrylamide gels. The mixture of DNA fragments to be separated is layered in a well at the top of the gel and an electric current is passed through the gel. DNA fragments move toward the positive pole at a rate inversely proportional to their length, forming bands that can be visualized by autoradiography (if the fragments are radiolabeled) or by treatment with a fluorescent dye. Agarose gels can be run in a horizontal direction; in this case, the melted agarose is allowed to harden on a single horizontal glass or plastic plate. This is not easily done with polyacrylamide gels because oxygen in the atmosphere inhibits polymerization of acrylamide.

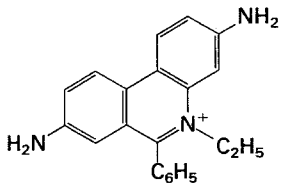


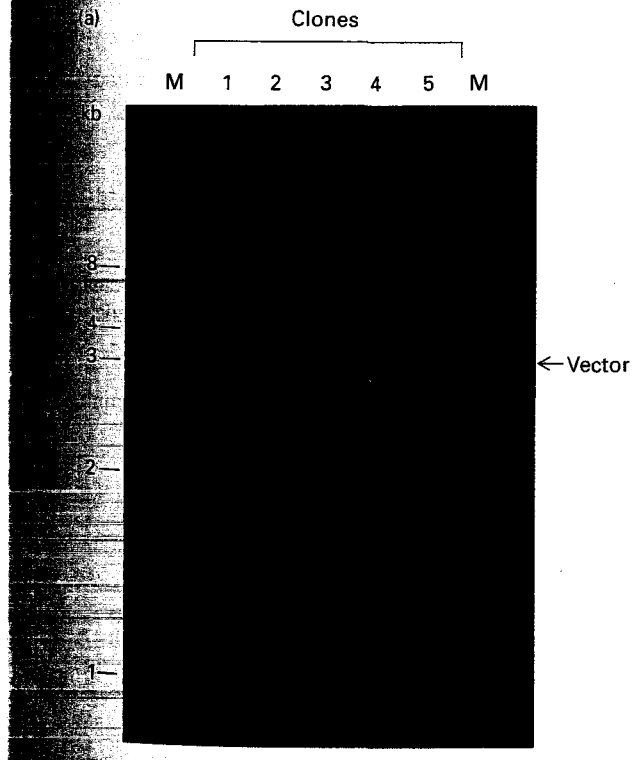
FIGURE 7-23 Structural formula of ethidium, a fluorescent dye commonly used to detect DNA and RNA in agarose and polyacrylamide gels. The bromide salt of ethidium generally is used.

Two methods are common for visualizing separated DNA bands on a gel. If the DNA is not radiolabeled, the gel is incubated in a solution containing the fluorescent dye ethidium, a planar molecule that binds to DNA by intercalating between the base pairs (Figure 7-23). This binding concentrates the ethidium in the DNA and also increases its intrinsic fluorescence. As a result, when the gel is illuminated with ultraviolet light, the regions of the gel containing DNA fluoresce much more brightly than the regions of the gel without DNA (Figure 7-24a). Radioactively labeled DNA can be visualized by autoradiography of the gel. In this case, the gel is laid against a sheet of photographic film in the dark, exposing the film at the positions where labeled DNA is present. When the film is developed, a photographic image of the DNA is observed (Figure 7-24b).

Multiple Restriction Sites Can Be Mapped on a Cloned DNA Fragment

In addition to separating restriction fragments of different lengths, gel electrophoresis provides a means for estimating the length of fragments. The distance that a restriction fragment migrates in a gel is inversely proportional to the logarithm of its length. Thus the length of a restriction fragment can be determined fairly accurately by comparison with restriction fragments of known length subjected to electrophoresis on the same gel (see Figure 7-24a).

The ability to determine the length of restriction fragments makes it possible to locate the positions of restriction sites relative to each other on a DNA molecule (e.g., a newly cloned DNA fragment). A diagram showing the positions of restriction sites on a DNA molecule is called a *restriction map*, and the process of determining these positions is called *restriction-site mapping*. Figure 7-25 illustrates the procedure for mapping two restriction sites relative to each other when only one copy of each site is present in a fragment. In this simple case, three fragment



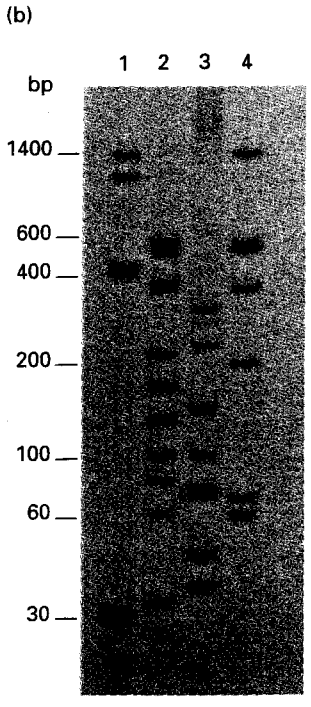
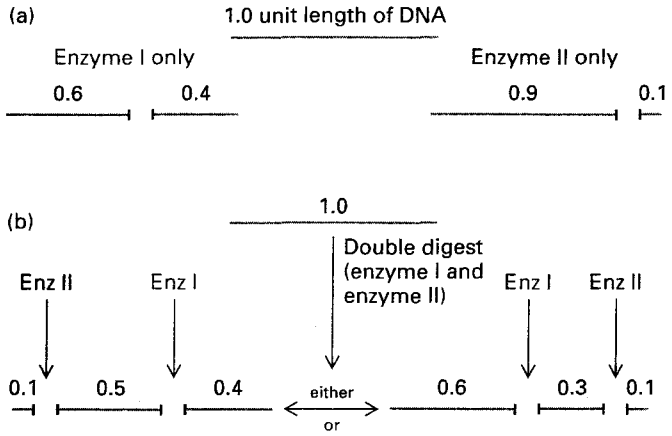
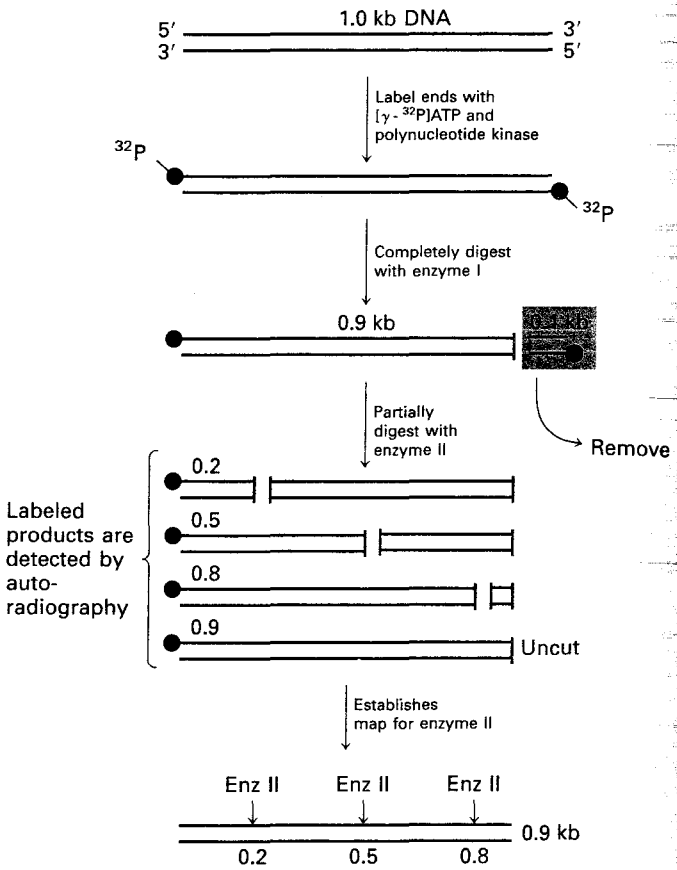


FIGURE 7-24 Separation of restriction fragments by gel electrophoresis. (a) Several different plasmid clones were digested with *Eco*RI and the digested DNA was subjected to agarose gel electrophoresis to separate the cloned fragments from the plasmid vector DNA. Each *Eco*RI-cut plasmid was layered on a separate well. As a "marker" for estimating the sizes of the cloned fragments, adenovirus 2 (Ad2) DNA was digested with *Hin*dIII (see Figure 7-6) and layered in the left and right wells of the gel. The lengths of the Ad2 *Hin*dIII fragments were calculated from the Ad2 DNA sequence. (b) Autoradiogram of ^{32}P-labeled fragments separated by polyacrylamide gel electrophoresis. Lengths of fragments are indicated at the left in base pairs. Parts (a) and (b) courtesy of Carol Eng.



▲ FIGURE 7-25 Mapping the recognition sites for two restriction enzymes relative to each other in a DNA fragment containing one copy of each site. For simplicity, double-stranded DNA is represented by a single line. (a) The fragment is exposed separately to two restriction enzymes (I and II). Each enzyme cuts the fragment once, generating two subfragments, whose lengths are determined by gel electrophoresis. (b) The fragment also is digested with both enzymes simultaneously. Since the lengths of the resulting fragments will depend on the relative position of the two restriction sites, the sites can be mapped based on the lengths observed. By continuing this process with different pairs of enzymes, the investigator can construct a detailed map of restriction sites.



▲ FIGURE 7-26 Mapping the multiple copies of the recognition site for a given restriction enzyme in a DNA fragment. The fragment first is doubly end-labeled (red circles) and then is digested completely with an appropriate restriction enzyme (I) to produce fragments labeled at *one* end only. The larger fragment is partially digested with a second enzyme (II), so that no more than one cut is made in any fragment molecule. Complete digestion would generate only one labeled product (here, the 0.2-kb subfragment), whereas partial digestion generates a labeled product for each restriction site. From the lengths of the labeled pieces, the positions of the multiple recognition sites for enzyme II (red arrows) in the original DNA can be inferred. [See H. O. Smith and M. Birnstiel, 1976, *Nucl. Acids Res.* **3**:2387.]

samples are digested: one with enzyme I, one with enzyme II, and one with both enzymes.

When a DNA fragment contains multiple copies of the recognition site for one or more restriction enzymes, the mapping procedure is more complicated. In this case, the sites for each enzyme must be mapped before the sites for different enzymes can be mapped relative to each other. The first step in this procedure is radiolabeling both ends of the fragment with [γ-^{32}P]ATP and polynucleotide kinase (see Figure 7-20). As shown in Figure 7-26, the doubly end-labeled fragment is treated with a restriction enzyme that cuts the fragment just once, and the resulting singly end-labeled fragments are separated. These fragments then are partially digested with the enzyme whose multiple recognition sites are being mapped.

Multiple restriction sites in a cloned DNA fragment can be mapped by use of these two methods with multiple restriction enzymes. Each distinct DNA sequence has a characteristic restriction-site map (see Figure 7-6). Such maps can be used to align partially overlapping cloned DNA fragments. Also, specific small regions within a large cloned DNA fragment can be prepared by digesting the cloned fragment with various combinations of restriction enzymes; the smaller subfragments then can be isolated by gel electrophoresis.

Pulsed-Field Gel Electrophoresis Separates Large DNA Molecules

The gel electrophoretic techniques described so far can resolve DNA fragments up to ≈20 kb in length. Larger DNAs, ranging from 2×10^4 to 10^7 base pairs (20 kb to 10 megabases [Mb]) in length, can be separated by size with *pulsed-field gel electrophoresis*. This technique depends on the unique behavior of large DNAs in an electric field that is turned on and off (pulsed) at short intervals.

When an electric field is applied to large DNA molecules in a gel, the molecules migrate in the direction of the

field and also stretch out lengthwise. If the current then is stopped, the molecules begin to "relax" into random coils. The time required for relaxation is directly proportional to the length of a molecule. The electric field then is reapplied at 90° or 180° to the first direction. Longer molecules relax less than shorter ones during the time the current is turned off; as a result, longer molecules start moving in the direction imposed by the new field more slowly than shorter ones. Repeated alternation of the field direction gradually forces large DNA molecules of different size farther and farther apart.

Pulsed-field gel electrophoresis is very important for purifying long DNA molecules up to $\approx 10^7$ base pairs in length (Figure 7-27). The technique is required for analyzing cellular chromosomes, which range in size from about 5×10^5 base pairs (smallest yeast chromosomes) to $2–3 \times 10^8$ base pairs (animal and plant chromosomes). Very large chromosomes must be digested into fragments of 10^7 base pairs or less before they can be analyzed. Such large restriction fragments can be generated with restriction enzymes that cut at rarely occurring 8-bp restriction sites.

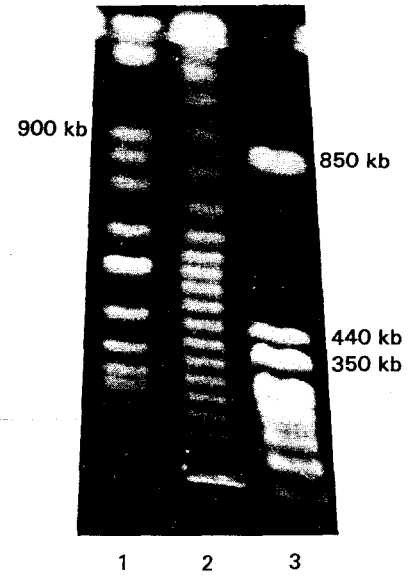


FIGURE 7-27 Pulsed-field gel electrophoretic separation of large DNA molecules. Lane 1 shows individual DNA molecules; each band represents one chromosome from the yeast *S. cerevisiae*. Lane 3 shows *Not*I restriction fragments of the *E. coli* chromosome, which were used in mapping this genome. Lane 2 shows a "ladder" of concatomers of λ phage DNA in which each unit is ≈48.5 kb long (see Figure 7-11). The band at the bottom is a single unit; the other bands are successively ≈48.5 kb longer. This ladder can be used to estimate the length of long DNA fragments electrophoresed in parallel. [See C. L. Smith et al., 1987, *Nucl. Acids Res.* **15**:4481. Photograph courtesy of C. L. Smith.]

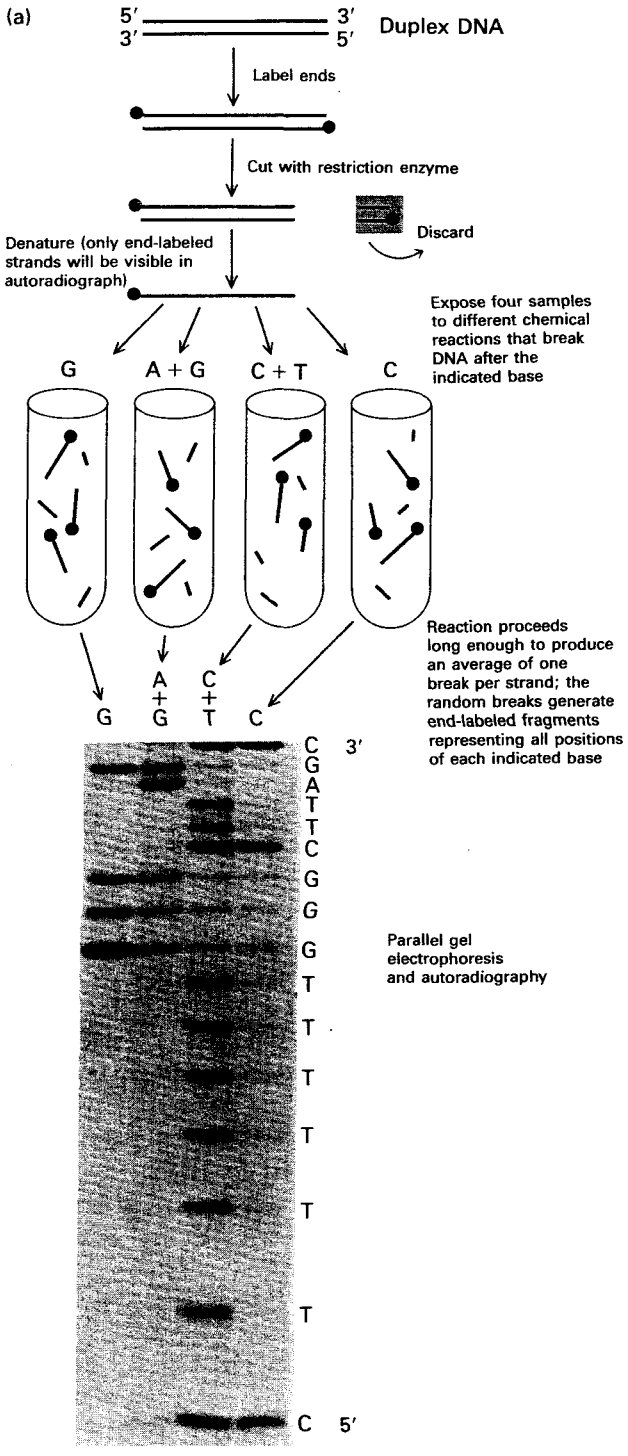
Nucleotide Sequencing of Cloned DNA Fragments Paves the Way for Sequencing Entire Genomes

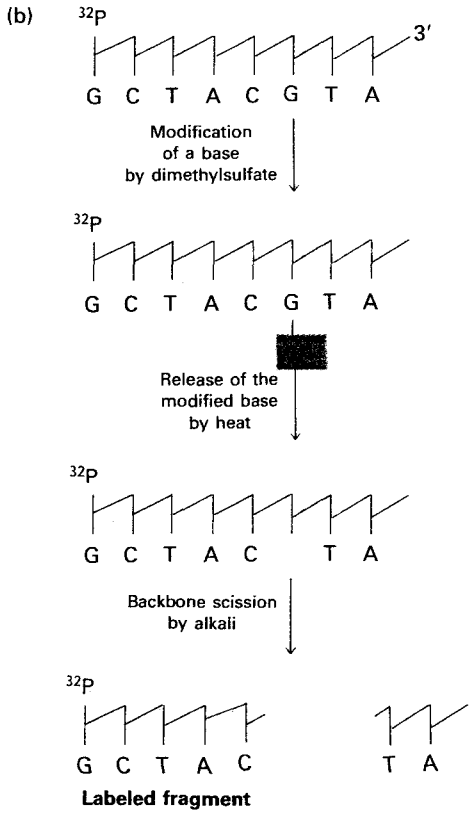
Virtually all the information required for the growth and development of an organism is encoded in the DNA of its genome. The availability of techniques to produce and separate DNA restriction fragments a few hundred nucleotides long led to development of two procedures for determining the exact nucleotide sequence of stretches of DNA up to ≈500 nucleotides long. These DNA sequencing methods together with the technology for constructing a library representing the entire genome of an organism make it possible to determine the exact sequence of the entire DNA of that organism.

The total genomes of many viruses and much of the *E. coli* genome have already been sequenced. Automation of the techniques for sequencing DNA and computerized storage of the sequence data are facilitating the current effort to determine the sequence of the entire human genome. Within the next decade, if not sooner, researchers also are likely to complete sequencing the entire genomes of several important experimental organisms including the yeast *Saccharomyces cerevisiae*, the round worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and the mouse. Knowledge of these DNA sequences will undoubtedly revolutionize our understanding of how cells and organisms function.

Maxam-Gilbert Method In the late 1970s, A. M. Maxam and W. Gilbert devised the first method for sequencing DNA fragments containing up to ≈500 nucleotides. In this method, four samples of an end-labeled DNA restriction fragment are chemically cleaved at different specific nucleotides. The resulting subfragments are separated by gel electrophoresis, and the labeled fragments are detected by autoradiography. As illustrated in Figure 7-28, the sequence of the original end-labeled restriction fragment can be determined directly from parallel electrophoretograms of the four samples.

Sanger (Dideoxy) Method A few years later, F. Sanger and his colleagues developed a second method of DNA sequencing, which now is used much more frequently than the Maxam-Gilbert method. The Sanger method is also called *dideoxy sequencing* because it involves use of 2′,3′-dideoxynucleoside triphosphates (ddNTPs), which lack a 3′-hydroxyl group (Figure 7-29). In this method, the single-stranded DNA to be sequenced serves as the template strand for in vitro DNA synthesis; a synthetic 5′-end–labeled oligodeoxynucleotide is used as the primer. As shown in Figure 7-30, four separate polymerization reactions are performed, each with a low concentration of one of the four ddNTPs in addition to higher concentrations of the normal deoxynucleoside triphosphates (dNTPs). In each reaction, the ddNTP is randomly incorporated at the positions of the corresponding dNTP; such addition of a

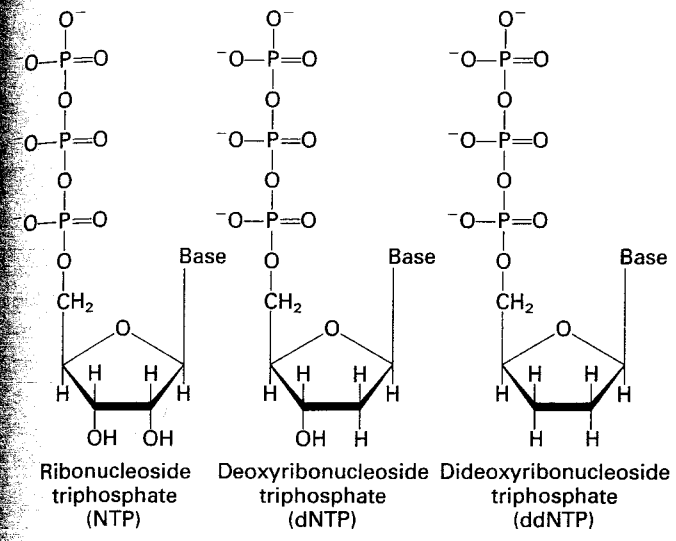




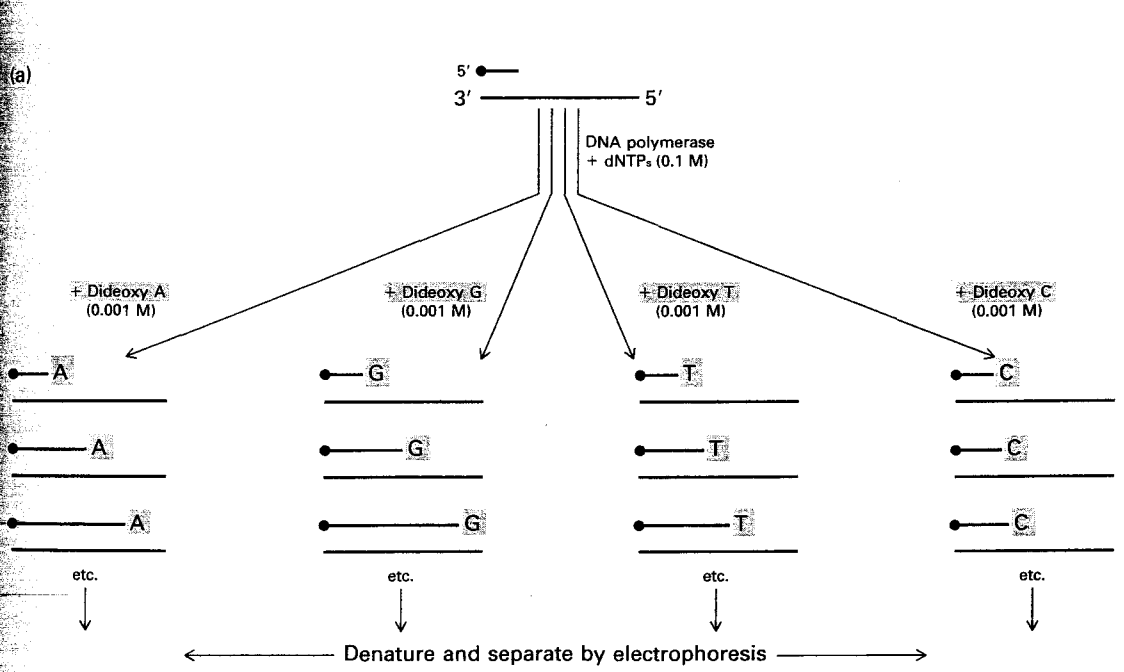
▲ FIGURE 7-28 Maxam-Gilbert method for sequencing DNA fragments up to ≈500 nucleotides in length. (a) The double-stranded fragment to be sequenced is labeled at the 5′ ends with ^{32}P (see Figure 7-20). The label (red circle) is *removed from one end, and the fragment then is denatured.* Four identical samples of the prepared fragment are subjected to four different chemical reactions that selectively cut the DNA backbone at G, G + A, C + T, or C residues. The reactions are controlled so that each labeled chain is likely to be broken only once. An example of the reaction that cleaves at a G is shown in (b). The labeled subfragments created by all four reactions have the label at one end and the chemical cleavage point at the other. Gel electrophoresis and autoradiography of each separate mixture yield one radioactive band for each nucleotide in the original fragment. Bands appearing in the G and C lanes can be read directly. Bands in the A + G lane that are not duplicated in the G lane are read as A. Bands in the T + C lane that are not duplicated in the C lane are read as T. The sequence is read from the bottom of the gel up. [See A. Maxam and W. Gilbert, 1977, *Proc. Nat'l. Acad. Sci. USA* **74**:560. Photograph from L. Stryer, 1988, *Biochemistry,* 3d ed., W. H. Freeman and Company, p. 120; courtesy of Dr. David Dressler.]

ddNTP terminates polymerization because the absence of a 3′ hydroxyl prevents addition of the next nucleotide. The mixture of terminated fragments from each of the four reactions is subjected to gel electrophoresis in parallel; the separated fragments then are detected by autoradiography. The sequence of the original DNA template strand can be read directly from the resulting autoradiogram (see Figure 7-30c). Once the sequence for a particular cloned DNA fragment is determined, primers for overlapping fragments can be chemically synthesized based on that sequence. The sequence of a long continuous stretch of DNA thus can be determined by individually sequencing the overlapping cloned DNA fragments that compose it.

Informatics: The Storage, Distribution and Analysis of DNA Sequence Data Vast amounts of DNA sequence have already been determined, and the pace at which new sequences are characterized is continuously accelerating. Computers are necessary to store and distribute this enormous volume of data. *Informatics* is the rapidly



◀ FIGURE 7-29 Structures of ribonucleoside triphosphate (NTP), deoxyribonucleoside triphosphate (dNTP), and dideoxyribonucleoside triphosphate (ddNTP).



▲ FIGURE 7-30 Sanger (dideoxy) method for sequencing DNA fragments. (a) A single strand of the DNA to be sequenced (blue line) is hybridized to a 5′-end–labeled synthetic deoxynucleotide primer. The primer is elongated in four separate reaction mixtures containing the four normal deoxynucleoside triphosphates (dNTPs) plus one of the four dideoxynucleoside triphosphates (ddNTPs) in a ratio of 100 to 1. A ddNTP molecule can add at the position of the corresponding normal dNTP, but when this occurs, chain elongation stops because the ddNTP lacks a 3′ hydroxyl. In time, each reaction mixture will contain a mixture of prematurely terminated chains ending at every occurrence of the ddNTP (yellow). (b) Three of the labeled chains that would be generated from the specific DNA sequence shown in the presence of ddGTP. (c) An actual audioradiogram of a polyacrylamide gel in which more than 300 bases can be read. Each reaction was carried out in duplicate using Sequenase™, a commercial preparation of the DNA polymerase from bacteriophage T7. [Part (c) courtesy of United States Biochemical Corporation.]

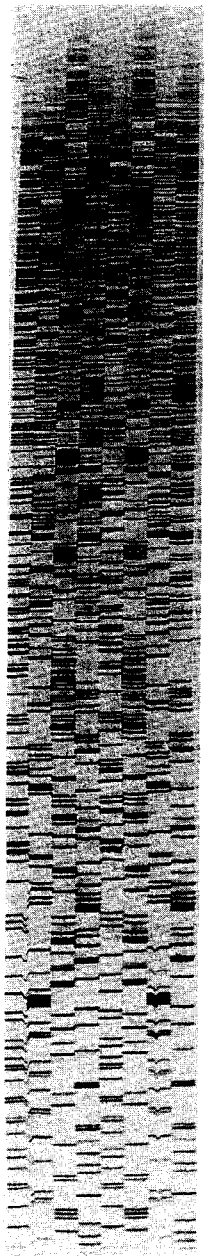
(b)

5′ ^{32}P-TAGCTGACTC 3′
3′ ATCGACTGAGTCAAGAGCTATTGGGCTTAA . . .

DNA polymerase
+dATP, dGTP, dCTP, dTTP
+ddGTP in low concentration

5′ ^{32}P-TAGCTGACTCAG 3′
3′ ATCGACTGAGTCAAGAGCTATTGGGCTTAA . . .
+
5′ ^{32}P-TAGCTGACTCAGTTCTCG 3′
3′ ATCGACTGAGTCAAGAGCTATTGGGCTTAA . . .
+
5′ ^{32}P-TAGCTGACTCAGTTCTCGATAACCCG 3′
3′ ATCGACTGAGTCAAGAGCTATTGGGCTTAA . . .

(c)



C A G T C G A T

developing area of computer science devoted to collecting, organizing, and analyzing DNA and protein sequences. The principal data banks where such sequences are stored are the GenBank at the Los Alamos National Laboratory, and the EMBL Data Library at the European Molecular Biology Laboratory in Heidelberg. These data bases continuously exchange newly reported sequences and make them available to molecular cell biologists throughout the world through an international computer network. Newly derived sequences can be compared with previously determined sequences to search for similarities, called *homologous sequences*. Protein-coding regions can be translated into amino acid sequences, which also can be compared. Because of degeneracy in the genetic code, related proteins often exhibit more homology than the genes encoding them.

As discussed in Chapter 3, proteins with similar functions often contain homologous sequences that correspond to important functional domains in the three-dimensional structure of the proteins. Discovery that a protein encoded by a newly cloned gene exhibits such homologies with proteins of known function can provide revealing insights into the function of the cloned gene. For example, the human gene *NF1* was identified and cloned by methods described in Chapter 8. This gene is associated with the inherited disease neurofibromatosis 1, which results in the development of multiple tumors in the peripheral nervous system. Before the *NF1* gene was cloned, there was little understanding of the molecular basis of the disease. After a cDNA clone of *NF1* was isolated and sequenced, the deduced sequence of the NF1 protein was checked against all other protein sequences in GenBank. A region of homology was identified with a protein called GAP, which interacts with the Ras protein. This latter protein is encoded by the *ras* gene, which is mutated in many human tumors. As we examine in detail in Chapters 20 and 26, the GAP and Ras proteins normally function to control cell growth and differentiation in response to signals from neighboring cells. Because of its sequence homology with GAP, researchers hypothesized that NF1 also would interact with Ras. The NF1 protein subsequently was expressed from the cloned gene by methods described later and shown to interact with Ras. This finding suggests that in individuals with neurofibromatosis, who have a defective *NF1* gene, a mutant NF1 protein, expressed in cells of the peripheral nervous system, interacts abnormally with Ras to signal inappropriate cell division.

This example illustrates how insight into the molecular basis of an inherited disease can be gained by identifying and cloning the associated mutant gene and then comparing the sequence of the encoded protein with the sequences of other proteins stored in data banks. This general approach for revealing the molecular role of various proteins will undoubtedly increase as the sequences of more proteins with known functions are determined and as new computer methods are devised for identifying potentially significant relationships among sequences.

➤ *Analysis of Specific Nucleic Acids in Complex Mixtures*

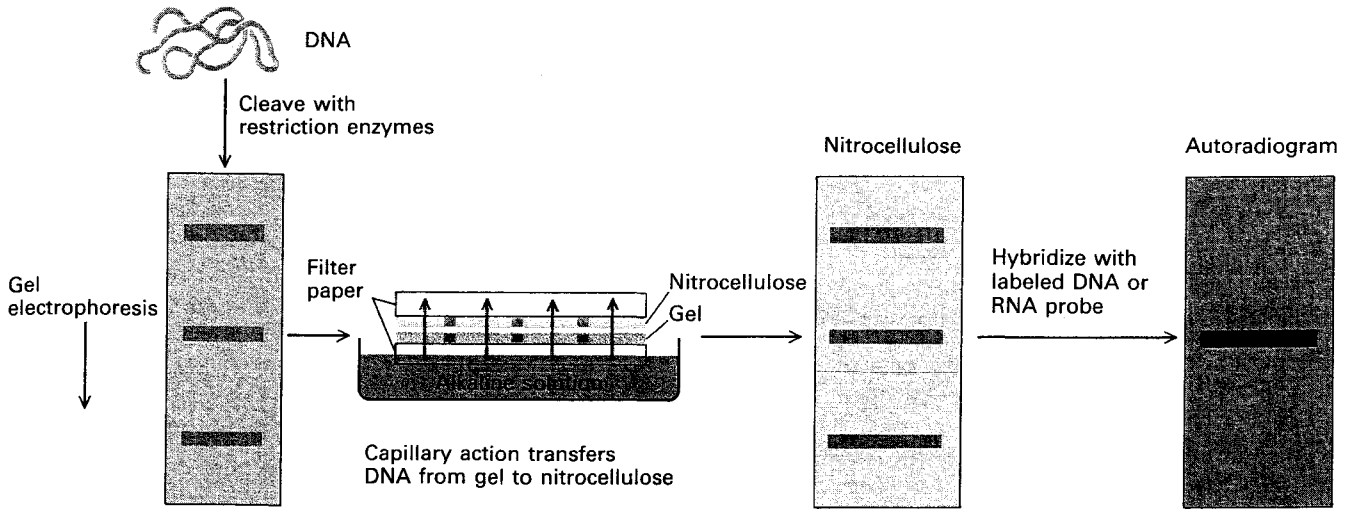
Once a specific DNA sequence has been isolated by cloning, the cloned DNA can be used as a probe to detect the presence, and in some cases the amounts, of complementary nucleic acids in complex mixtures such as total cellular DNA or RNA. These procedures depend on the exquisite specificity of nucleic acid hybridization. Related methods are used to locate DNA regions encoding specific mRNAs and transcription start sites.

Southern Blotting Detects Specific DNA Fragments

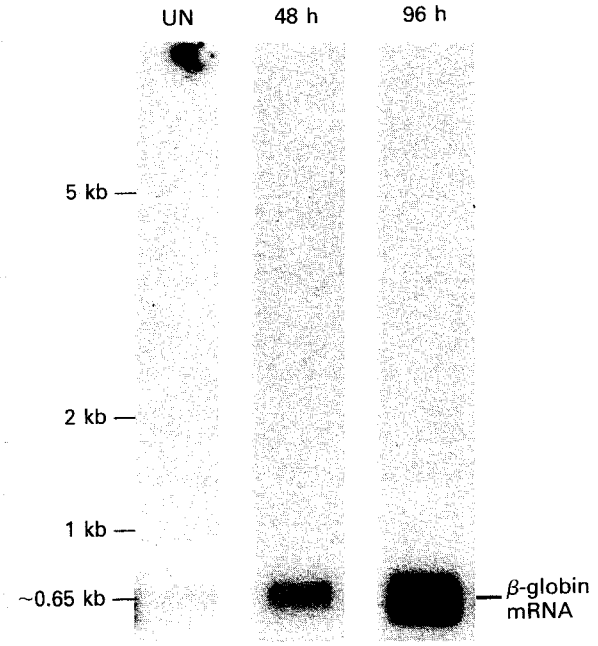
The technique of Southern blotting, named after its originator Edward Southern, can identify specific restriction fragments in a complex mixture of restriction fragments. The DNA to be analyzed, such as the total DNA of an organism, is digested to completion with a restriction enzyme. For an organism with a complex genome, this digestion may generate millions of specific restriction fragments. The complex mixture of fragments is subjected to gel electrophoresis to separate the fragments according to size. However, many different fragments are of exactly the same length, and these do not separate from each other.

Even though all the fragments are not resolved by gel electrophoresis, an individual fragment that is complementary to a specific DNA clone can be detected. The restriction fragments present in the gel are denatured with alkali and transferred onto a nitrocellulose filter or nylon membrane by blotting (Figure 7-31). This procedure preserves the distribution of the fragments in the gel, creating a replica of the gel on the filter, much like the replica filter produced from plaques of a λ library. The filter then is incubated under hybridization conditions with a specific radiolabeled DNA probe usually generated from a cloned restriction fragment. The DNA restriction fragment that is complementary to the probe hybridizes, and its location on the filter can be revealed by autoradiography.

Southern blotting permits a comparison between the restriction map of DNA isolated directly from an organism and the restriction map of cloned DNA. This is necessary to be certain that no rearrangements have occurred during the cloning procedure such as might happen if two restriction fragments that do not normally lie next to each other were inadvertently ligated together before ligation into a cloning vector. Southern blotting also is used to map restriction sites in genomic DNA next to the sequence of a cloned DNA fragment. This provides a rapid method of



▲ FIGURE 7-31 The Southern-blot technique for detecting the presence of specific DNA sequences following gel electrophoresis of a complex mixture of restriction fragments. The diagram depicts three restriction fragments in the gel, but the procedure can be applied to a mixture of millions of DNA fragments. [See E. M. Southern, 1975, *J. Mol. Biol.* **98**:508.]



▲ FIGURE 7-32 Northern blots of β-globin mRNA in extracts of erythroleukemia cells that are growing but uninduced (UN lane) and in cells that are induced to stop growing and allowed to differentiate for 48 h or 96 h. The density of a band is proportional to the amount of mRNA present. The β-globin mRNA is barely detectable in uninduced cells but increases more than 1000-fold after 96 h of differentiation. [Courtesy of L. Kole.]

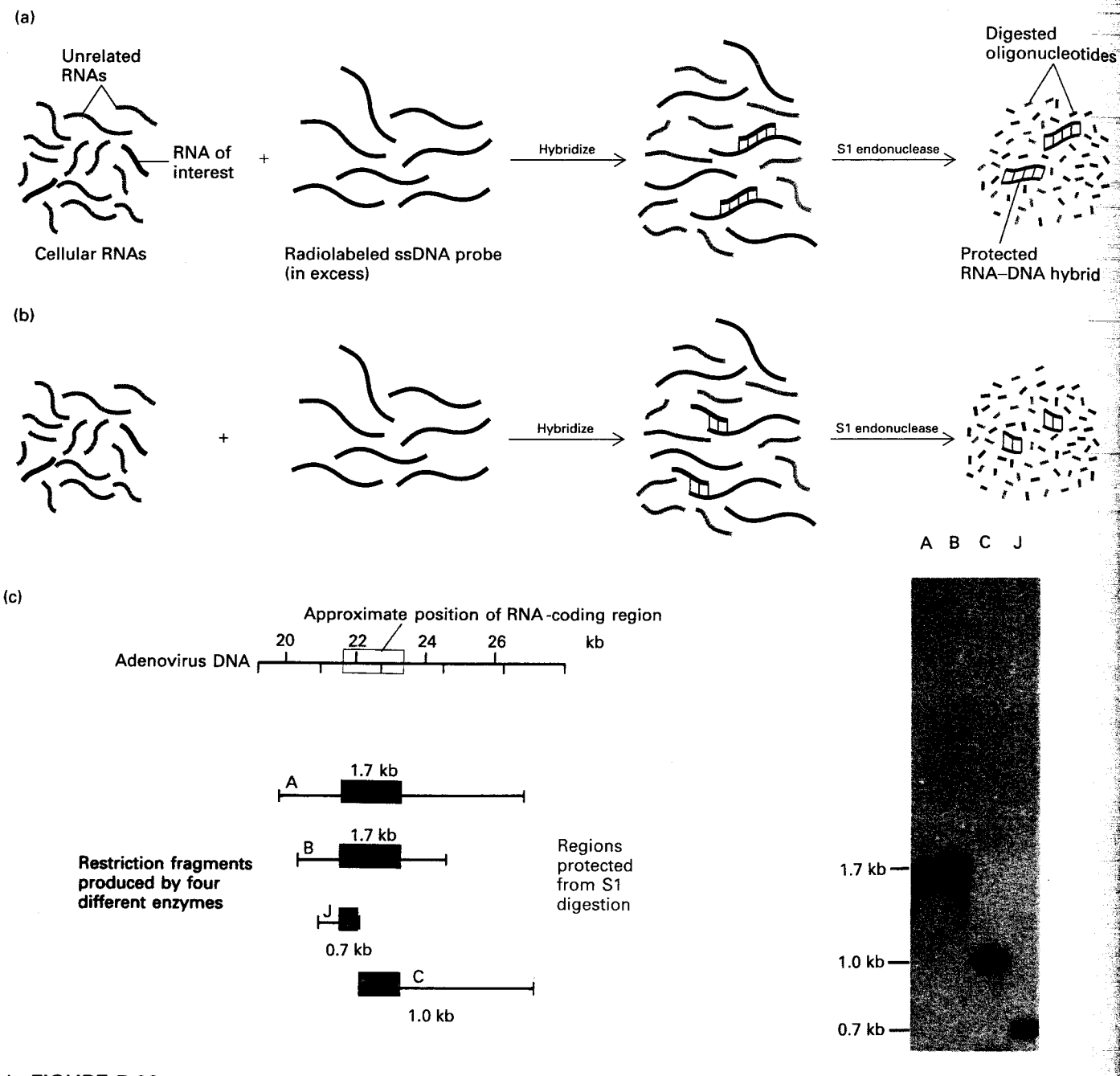
comparing the restriction maps of different individual organisms in the region surrounding a cloned fragment. Deletion and insertion mutations are readily detected, as well as sequence differences in specific restriction sites.

Northern Blotting Detects Specific RNAs

Northern blotting, somewhat humorously named because it is patterned after Southern blotting, is used to detect a particular RNA in a mixture of RNAs. A RNA sample, often the total cellular RNA, is denatured by treatment with an agent (e.g., formaldehyde) that prevents hydrogen bonding between base pairs, ensuring that all the RNA molecules have an unfolded, linear conformation. The individual RNAs then are separated according to size by gel electrophoresis and transferred to a nitrocellulose filter to which the extended denatured RNAs adhere. The filter then is exposed to a labeled DNA probe and subjected to autoradiography. Because the amount of a specific RNA in a sample can be estimated from a Northern blot, the procedure is widely used to compare the amounts of a particular mRNA in cells under different conditions (Figure 7-32).

Nuclease Protection Is Used to Quantitate Specific RNAs and Map the DNA Regions Encoding Them

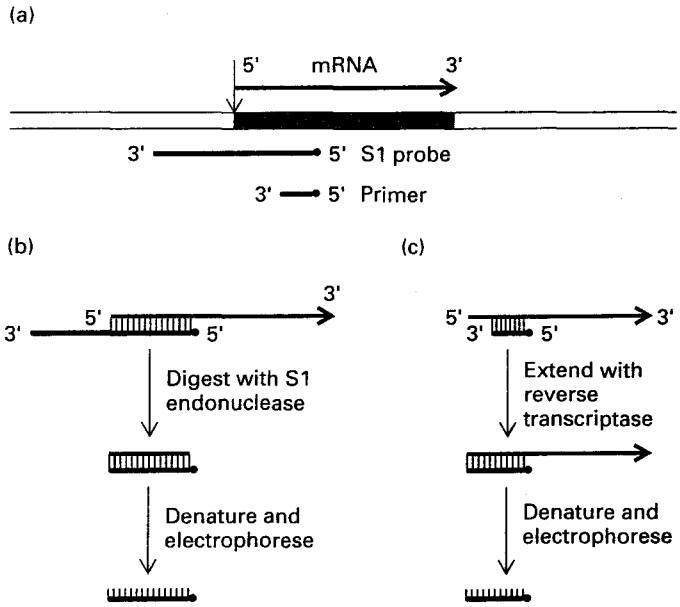
Another important method for detecting and quantitating specific RNA molecules employs endonucleases that digest single-stranded but not double-stranded nucleic acids. The method was originally designed using endonuclease S1, an enzyme from the mold *Aspergillus oryzae* that digests



▲ **FIGURE 7-33** Nuclease-protection method for quantitating specific RNAs in a mixture and mapping them. (a) A radiolabeled, single-stranded DNA probe (red) is mixed with a mixture of cellular RNAs; the probe hybridizes only to the complementary RNA (blue), which is a small fraction of the total RNA sample. In this example, the probe contains a sequence complementary to the entire RNA of interest. Digestion with S1 endonuclease degrades all the unprotected (unhybridized) RNA and DNA sequences, leaving a double-stranded RNA-DNA hybrid equal in length to the RNA. The protected hybrid is detected by gel electrophoresis followed by autoradiography. The density of the resulting band is proportional to the amount of the hybridized RNA in the original mixture. (b) With a "partial" DNA probe, containing only a portion of the DNA sequence complementary to the RNA, the protected S1-digestion product is shorter than the RNA and equal in length to the complementary region of the probe. (c) An example of mapping an RNA on the genome of adenovirus. A 1.7-kb RNA was approximately mapped to the region between 21 and 24 from the left end of the 36-kb viral genome. Four radiolabeled restriction-fragment probes (A, B, C, and J) from this region of the viral DNA were prepared, hybridized with RNA from virus-infected cells, and then treated with S1 endonuclease. An autoradiogram of the S1-digestion products is shown at the right. Probes A and B produced S1-digestion products of 1.7 kb, indicating the RNA sequence maps entirely within these restriction fragments. The results with the partial probes C and J map the RNA sequence relative to the restriction site separating fragments C and J. [Photograph in part (c) from A. J. Berk and P. A. Sharp, 1977, *Cell* **12**:721; copyright M.I.T.]

single-stranded RNA and DNA but not double-stranded molecules. A labeled DNA strand, or probe, complementary to an RNA of interest is prepared from a cloned DNA. A source of RNA, such as the total polyadenylated RNA isolated from a particular tissue or type of cultured cell, is incubated with a high concentration of the labeled DNA probe under conditions where all the RNA complementary to the probe hybridizes to it (Figure 7-33a). The preparation then is treated with endonuclease S1, which digests all the unhybridized RNA and probe molecules. However, the double-stranded region in the RNA-DNA hybrids is protected from nuclease digestion, but the noncomplementary regions of the probe are trimmed to the length of the complementary RNA. The digested preparation is subjected to gel electrophoresis followed by autoradiography of the gel to detect the protected probe. The amount of radioactivity in the resulting band is a measure of the amount of RNA complementary to the probe in the initial sample of RNA. Nuclease protection also can be performed with a complementary labeled RNA and ribonuclease A, a single-strand–specific pancreatic ribonuclease.

The DNA region that encodes a particular RNA can be mapped with the nuclease-protection technique by use of restriction-fragment probes in which one end is complementary to only a portion of the RNA of interest. In this case, the RNA-DNA hybrid protected from S1 digestion is shorter than the RNA being probed; its length corresponds to that of the DNA region extending from one end of the coding region to a restriction site within it (Figure 7-33b). Comparison of the protected doubled-stranded fragments obtained with two or more such "partial" probes can map the RNA sequence relative to restriction sites in the complementary DNA (Figure 7-33c).



▲ **FIGURE 7-34** Two methods for mapping the start site for transcription of a particular gene in a region of DNA of known sequence. (a) Diagram of the DNA fragment containing the gene of interest (light blue) and the corresponding mRNA (red). The end-labeled (red dot) single-stranded DNA fragment used as a probe in the S1 mapping technique and the end-labeled oligonucleotide primer used in the primer-extension technique are shown below the position of their sequence in the DNA. (b) In the S1 mapping technique, the probe is hybridized with the mRNA, and unpaired nucleic acid is then digested with S1 endonuclease (see Figure 7-33). Denaturation leaves a labeled DNA fragment whose length accurately marks the distance of the starting nucleotide of the mRNA from the nucleotide that hybridized with the labeled DNA end. (c) In the primer-extension technique, a short (approximately 20 nucleotides) oligodeoxynucleotide is synthesized and end-labeled. After the primer is hybridized to the mRNA, it is extended by reverse transcriptase until it reaches the first nucleotide of the mRNA. The length of the primer-extension product, determined by gel electrophoresis, measures the distance from the 5′ end of the primer to the 5′ end of the mRNA.

Transcription Start Sites Can Be Mapped by S1 Protection and Primer Extension

As discussed in Chapter 11, some of the DNA regulatory elements that control transcription of a protein-coding sequence (gene) into mRNA are located near the transcription *start site*. Mapping the start site for synthesis of a particular mRNA often helps in identifying the DNA regulatory sequences that control transcription of the corresponding gene. Two methods are used to map the 5′ end of a particular mRNA on a complementary DNA: *S1 protection* and *primer extension*. The first step in both methods is to identify the general region of a DNA that includes the start site of interest (Figure 7-34a); this can be done by Northern-blot analysis or nuclease protection using various cloned restriction fragments.

In the S1-protection method, the identified DNA region is treated with appropriate restriction enzymes to produce a single-stranded DNA fragment that will hybridize with the 5′ portion of the mRNA. This fragment is radiolabeled at the 5′ end, hybridized with the mRNA, and then trimmed with S1 endonuclease (Figure 7-34b). From the length of the labeled probe segment protected from digestion, the position of the start site in the original DNA can be located.

The primer-extension method uses a synthetic oligonucleotide that is complementary to an approximately 20-nucleotide stretch of the mRNA located 50–100 nucleotides from its 5′ end. This synthetic oligonucleotide is end-labeled at the 5′ end and then used to prime DNA synthesis by reverse transcriptase with the mRNA as the template (Figure 7-34c). The position of the start site can be mapped from the length of the resulting extension product.

➤ *Designing Expression Systems That Produce Abundant Amounts of Specific Proteins*

Many proteins with interesting or useful functions are normally expressed at very low concentrations. A case in point is *granulocyte colony-stimulating factor* (G-CSF). This human protein hormone stimulates the development and replication of granulocytes, the phagocytic white blood cells critical to defense against bacterial infections. Both granulocytes and the cells that produce G-CSF are very sensitive to chemotherapeutic agents used in the treatment of cancer. As a result, one of the most serious side effects of chemotherapy is a fall in the concentration of granulocytes, making patients prone to life-threatening infections. G-CSF is normally made in very low concentrations in the bone marrow where granulocytes differentiate and by some cultured human cell lines. Isolation of G-CSF from these sources is a tedious process yielding minuscule amounts of purified protein. By the techniques described in this section, G-CSF can now be produced in large enough amounts for therapeutic use in cancer patients to diminish the impact of chemotherapy on granulocyte production.

The first step in obtaining large amounts of low-abundance proteins such as G-CSF is to purify a sufficient quantity of the protein to allow determination of a portion of its amino acid sequence. (As discussed in Chapter 3, only very small amounts of a protein are needed for amino acid sequencing.) A synthetic oligonucleotide probe is designed based on the determined amino acid sequence; this probe is used to identify a cDNA encoding the protein of interest, as described previously. Once the desired cDNA is cloned, large amounts of the encoded protein often can be synthesized in engineered *E. coli* cells.

Previously we discussed a λ-phage expression vector that produces fusion proteins consisting of β-galactosidase joined to a protein fragment encoded by a DNA inserted into the vector; this type of expression system can be used to identify DNA clones encoding a specific protein (see Figure 7-21). Here, in contrast, we discuss expression vectors designed to produce full-length proteins at high levels.

Full-Length Proteins Encoded by Cloned Genes Can be Produced in *E. coli* Expression Systems

G-CSF and many other low-abundance proteins can be expressed at high levels in *E. coli* through use of specially designed expression vectors. Although many different expression vectors have been constructed, they all take advantage of the molecular mechanisms that control transcription and translation in *E. coli*.

Plasmid Expression Vectors Carrying a Strong, Regulated Promoter The most commonly used *E. coli* expression vectors are assembled by ligation of a basic plasmid vector, containing a replication origin (ORI) and selectable antibiotic-resistance gene, to a DNA sequence that functions as a strong, regulated *promoter*. A promoter is a DNA sequence where RNA polymerase initiates transcription. At a strong, regulated promoter, transcription is initiated many times per minute under specific environmental conditions.

One example of this type of expression vector contains a cloned fragment of the *E. coli* chromosome that includes the *lac* promoter and the *lacZ* gene encoding β-galactosidase. Transcription from the *lac* promoter only occurs when lactose, or a lactose analog such as isopropylthiogalactoside (IPTG), is added to the culture medium. IPTG generally is used because it cannot be metabolized, and therefore its concentration does not change as the cells grow. After addition of IPTG, the *lacZ* gene is transcribed into mRNA, which then is translated to yield many copies of the β-galactosidase protein (Figure 7-35a).

To modify this plasmid for production of G-CSF, the *lacZ* gene is replaced with a cDNA encoding G-CSF using restriction enzymes and DNA ligase. In this process, the *lac* promoter, which is required for efficient transcription, must be maintained just before the start site of the inserted cDNA. In *E. coli* cells transformed by the resulting plasmid, transcription of the G-CSF cDNA and expression of G-CSF protein occurs in the presence of IPTG (Figure 7-35b).

Plasmid Expression Vectors Carrying the T7 Late Promoter A more complicated expression system, involving two levels of amplification, can produce larger amounts of a desired protein than the system just described. This second-generation system depends on the regulated expression of *T7 RNA polymerase*, an extremely active enzyme that is encoded in the DNA of bacteriophage T7. The T7 RNA polymerase transcribes DNA beginning within a specific 23-bp promoter sequence called the *T7 late promoter*. Copies of the T7 late promoter are located at several sites on the T7 genome, but none is present in *E. coli* chromosomal DNA. As a result, in T7-infected cells, T7 RNA polymerase catalyzes transcription of viral genes but not of *E. coli* genes.

In this expression system, recombinant *E. coli* cells first are engineered to carry the gene encoding T7 RNA polymerase next to the *lac* promoter. In the presence of IPTG, these cells transcribe the T7 gene at a high rate and synthesize abundant amounts of T7 RNA polymerase. These cells then are transformed with plasmid vectors that carry a copy of the T7 late promoter and, adjacent to it, the cDNA encoding the desired protein (Figure 7-36). When IPTG is added to the culture medium containing these transformed, recombinant *E. coli* cells, large amounts of

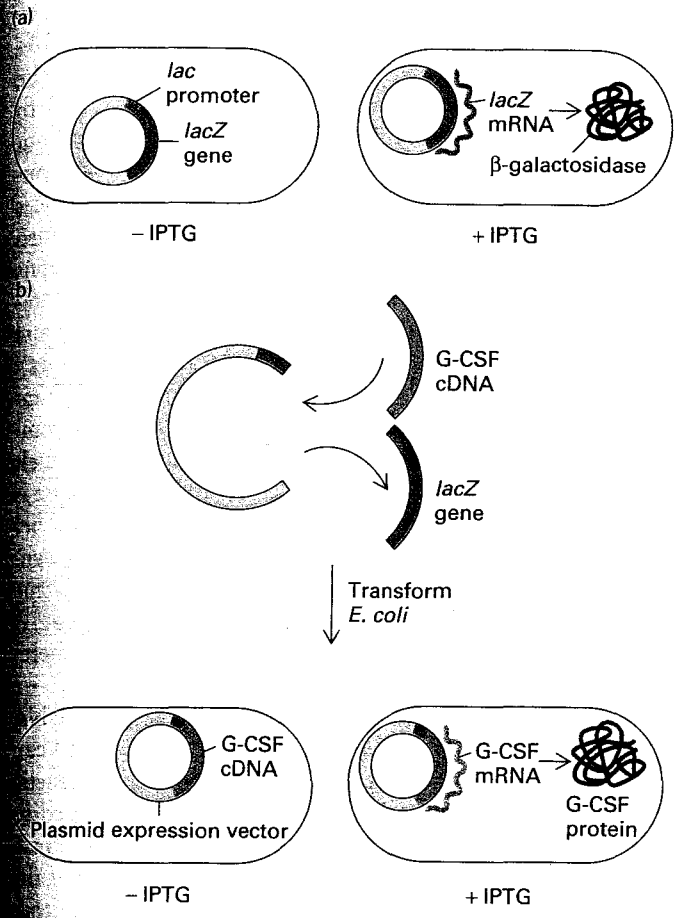
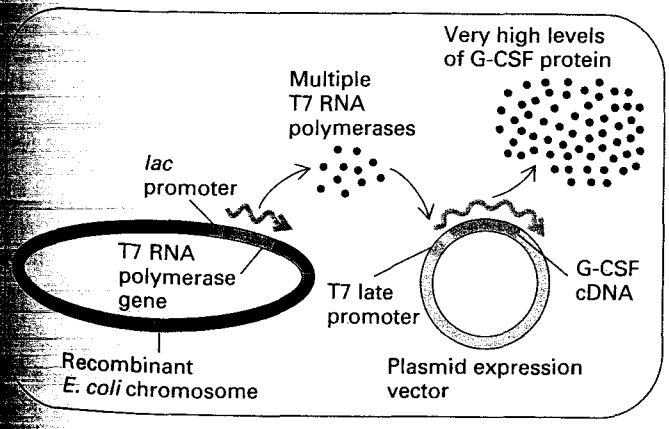


FIGURE 7-35 A simple *E. coli* expression vector utilizing the *lac* promoter. (a) The expression vector plasmid contains a fragment of the *E. coli* chromosome containing the *lac* promoter and the neighboring *lacZ* gene. In the presence of the lactose analog IPTG, RNA polymerase normally transcribes the *lacZ* gene producing *lacZ* mRNA, which is translated into the encoded protein, β-galactosidase. (b) The *lacZ* gene can be cut out of the expression vector with restriction enzymes and replaced by the G-CSF cDNA. When the resulting plasmid is transformed into *E. coli* cells, addition of IPTG and subsequent transcription from the *lac* promoter produces G-CSF mRNA, which is translated into G-CSF protein.



◀ FIGURE 7-36 Two-step expression vector system based on bacteriophage T7 RNA polymerase and T7 late promoter. The chromosome of a specially engineered *E. coli* cell contains a copy of the T7 RNA polymerase gene under the transcriptional control of the *lac* promoter. When transcription from the *lac* promoter is induced by addition of IPTG, the T7 RNA polymerase gene is transcribed, and the mRNA is translated into the enzyme. The T7 RNA polymerase molecules produced then initiate transcription at a very high rate from the T7 late promoter on the expression vector. Multiple copies of the expression vector are present in such cells, although only one copy is diagrammed here. The large quantity of mRNA transcribed from the cDNA cloned next to the T7 late promoter is translated into abundant protein product.

T7 RNA polymerase are produced. The polymerase then binds to the T7 late promoter on the plasmid expression vectors, catalyzing transcription of the inserted cDNA at a high rate. Since each *E. coli* cell contains many copies of the expression vector, prodigious amounts of mRNA corresponding to the cloned cDNA can be produced in this system.

Proteins with Post-Translational Modifications Can Be Produced in Eukaryotic Expression Systems

Most of the enzymes used in recombinant DNA technology (e.g., restriction enzymes, DNA polymerases, DNA ligases, polynucleotide kinase, and reverse transcriptase) are now produced commercially in *E. coli* expression systems. Large quantities of many eukaryotic proteins of interest, such as G-CSF, also can be produced in these systems. However, some eukaryotic proteins cannot be produced in active form in *E. coli* cells. These include proteins that are extensively modified during or following their synthesis, such as glycoproteins to which carbohydrate groups are added. *E. coli* lacks the enzymes that catalyze many of the post-translational modifications found on eukaryotic proteins.

To overcome this limitation of *E. coli* expression systems, researchers have developed *eukaryotic expression vectors* that permit addition of appropriate post-translation modifications to expressed proteins. Such vectors can be used in various types of eukaryotic cells to direct abundant synthesis of eukaryotic proteins from cloned genes.

Proteins Encoded by Cloned Genes and cDNAs Can Be Translated In Vitro

In the *E. coli* and eukaryotic expression systems just described, proteins are synthesized in vivo within living cells.

In vitro expression of proteins encoded by cloned DNA is possible with the T7 late-promoter expression vector diagrammed in Figure 7-36. In this system, the mRNA encoded by the cloned DNA is synthesized in vitro using the purified expression vector and purified T7 RNA polymerase. The mRNA then is translated in vitro to yield the desired protein. Although an in vitro expression system yields less protein than an in vivo system, it allows the desired protein to be radioactively labeled. In addition to the T7 system, plasmid vectors containing late promoters and the corresponding RNA polymerases from related bacteriophages (e.g., T3, T5, and SP6) also can be used for in vitro production of proteins from cloned DNA.

Cell extracts are used to translate the in vitro synthesized mRNA into protein. These extracts, which usually are prepared from rabbit reticulocytes or wheat germ, are rich in the components required for translation: ribosome subunits, tRNAs, amino acyl-tRNA synthases, and initiation, elongation, and termination factors (Chapter 4). Extracts first are treated with low concentrations of micrococcal nuclease to eliminate any endogenous mRNAs. At low levels, this enzyme, which is active only in the presence of Ca^{2+}, digests mRNA but not tRNA or rRNA. After digestion of mRNA is complete, the enzyme is inactivated by addition of EGTA, a chelating-agent that binds Ca^{2+} with a much higher affinity than the nuclease does. When an in vitro synthesized mRNA is added to a treated extract, only the corresponding protein is produced, since the endogenous mRNAs have been destroyed. If a radiolabeled amino acid such as [^{35}S]methionine is included in the translation reaction, the protein product will be labeled. The labeled protein can be used in binding experiments to test its interaction with other proteins and in various other experimental systems.

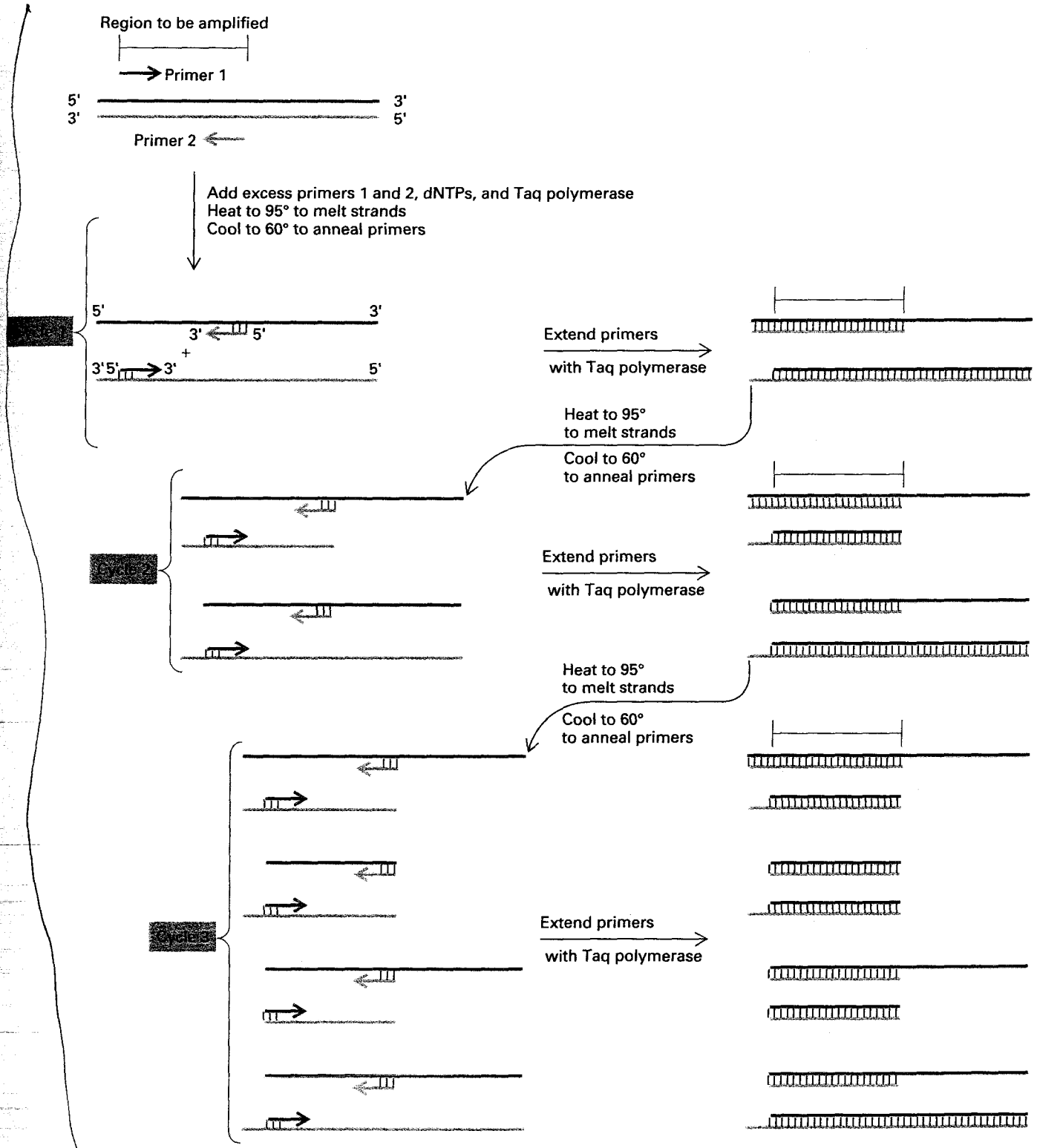
➤ The Polymerase Chain Reaction: An Alternative to Cloning

An alternative to cloning, called the *polymerase chain reaction* (PCR), can be used to directly amplify rare specific DNA sequences in a complex mixture when the ends of the sequence are known. This method of amplifying rare sequences from a mixture has vastly increased the sensitivity of procedures used in human genetics testing. For example, the β-globin gene in a small sample of DNA isolated from an individual can be specifically amplified by the PCR to determine if the person is a carrier of the mutant sickle-cell allele. Quantities of amplified DNA sufficient for sequencing can be prepared rapidly; subsequent sequencing reveals if the mutant allele is present in the sample.

A typical PCR is outlined in Figure 7-37. Genomic DNA is digested into large fragments using a restriction enzyme and then is heat-denatured into single strands. Two synthetic oligonucleotides complementary to the 3′ ends of the DNA segment of interest are added in great excess to the denatured DNA, and the temperature is lowered to 50–60°C. The genomic DNA remains denatured, because the complementary strands are at too low a concentration to encounter each other during the period of incubation, but the specific oligonucleotides, which are at a very high concentration, hybridize with their complementary sequences in the genomic DNA. The hybridized oligonucleotides then serve as primers for DNA chain synthesis, which begins upon addition of a supply of deoxynucleotides and a temperature-resistant DNA polymerase such as that from *Thermus aquaticus* (a bacterium that lives in hot springs). This enzyme, called *Taq polymerase,* can extend the primers at temperatures up to 72°C. When synthesis is complete, the whole mixture is heated further (to 95°C) to melt the newly formed DNA duplexes. When the temperature is lowered again, another round of synthesis takes place because excess primer is still present. Repeated cycles of synthesis (cooling) and melting (heating) quickly amplify the sequence of interest. At each round, the number of copies of the sequence between the primer sites is doubled; therefore the desired sequence increases exponentially.

The PCR is so effective at amplifying specific DNA sequences that DNA isolated from a single human cell can be analyzed for mutations associated with various genetic diseases. In one reported case, this approach was used to screen in vitro fertilized human embryos prepared from sperm and ova from a couple who both are carriers of the genetic disorder cystic fibrosis. This disease results from mutation in the *CFTR* gene, which is located on chromosome 7. The DNA isolated from a single embryonic cell

➤ FIGURE 7-37 The polymerase chain reaction (PCR). The top left of the figure indicates a region of DNA to be amplified. One strand is shown in dark blue and the other in light blue. Oligonucleotide primers are synthesized corresponding to ≈20 nucleotides from each end of the region to be amplified. Arrowheads indicate the 3′ ends of primers and extended strands. An excess of the two primers together with the heat-resistant DNA polyerase Taq polymerase and dNTPs are added to a preparation of DNA that includes the sequence to be amplified. The solution is heated to 95°C to melt the DNA into single strands and then is cooled to 60°C; at the lower temperature, the primers anneal (hybridize) to the complementary strands. Continued incubation at 60°C allows Taq polymerase to extend the primers, resulting in the synthesis of two partial doubled-stranded DNA molecules that include the sequence between the primers. After a second cycle of heating and cooling, four partial double-stranded DNA molecules are synthesized; each includes the sequence between the primers. After the third cycle, two of the eight DNA molecules formed are completely double stranded and extend from the 5′ end of primer 1 to the 5′ end of primer 2. Subsequent cycles of heating and cooling amplify these DNA molecules by a factor of two each cycle. In twenty cycles, the DNA sequence between the primers is amplified more than a millionfold.


Region to be amplified
Primer 1
5'
3'
3'
5'
Primer 2
Add excess primers 1 and 2, dNTPs, and Taq polymerase
Heat to 95° to melt strands
Cool to 60° to anneal primers
5'
3'
3'
5'
+
3'5'
3'
5'
Extend primers
with Taq polymerase
Heat to 95°
to melt strands
Cool to 60°
to anneal primers
Cycle 2
Extend primers
with Taq polymerase
Heat to 95°
to melt strands
Cool to 60°
to anneal primers
Cycle 3
Extend primers
with Taq polymerase

was subjected to PCR amplification and then analyzed for mutations identified in one of the two copies of chromosome 7 in each parent. In this way embryos that had inherited the wild-type chromosome from at least one parent were identified and then transferred to the mother's uterus. (Removal of a single cell from an in vitro fertilized human embryo has no apparent effect on subsequent development of the embryo after it is implanted in a receptive uterus.) By use of this procedure, carrier couples can be assured of having children that will not be at risk for cystic fibrosis. Another medical application of the PCR is early detection of infection with HIV, the virus that causes acquired immunodeficiency syndrome (AIDS). The PCR is so sensitive that it can detect HIV at very early stages in the disease (before symptoms appear) when only a few thousand blood cells in a patient are infected with the virus.

In basic research, the PCR also has numerous applications. For example, this procedure allows the recovery and rapid amplification of entire DNA sequences between any two ends whose sequences are known; the amplified sequences then can be ligated into standard cloning vectors. Only fragments of ≈2 kb or less can be amplified efficiently; however, further development of this technique will likely extend this upper limit.

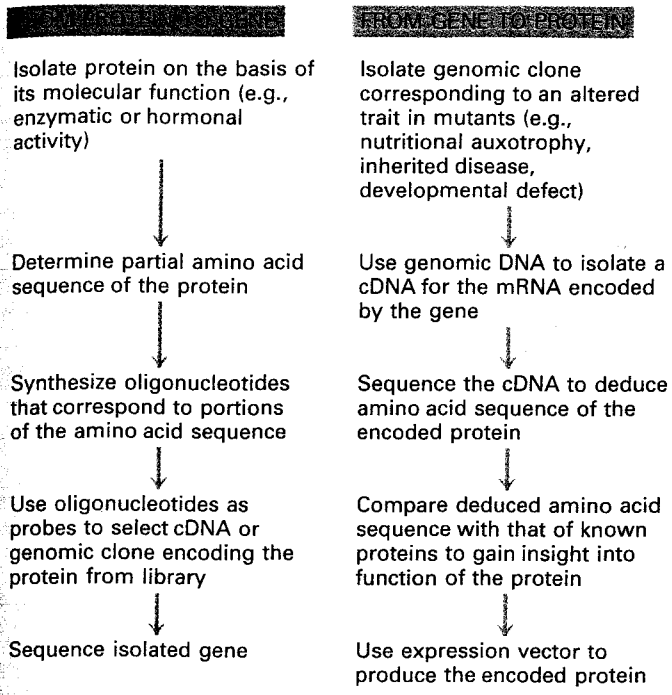
The PCR also provides a powerful approach to cloning a cDNA based on the partial amino acid sequence of a purified protein. The amino acid sequence of two peptides isolated from the protein are used to design two degenerate oligonucleotide mixtures containing all possible DNA sequences encoding the two peptides. Rather than using these oligonucleotides as probes for direct screening of a cDNA library, as described previously, they are used as primers in a PCR. First, cDNA is synthesized from total cellular mRNA using reverse transcriptase. The cDNA is then used as the template for a PCR performed with the two degenerate oligonucleotide primers. This reaction amplifies the region of the cDNA between the sequences encoding the peptides used to design the degenerate primers. The PCR procedure effectively selects the correct oligonucleotides for priming DNA synthesis from the degenerate oligonucleotide mixtures, because only DNA synthesized from the correct cDNA template will hybridize to oligonucleotides present in both degenerate primer mixtures. For exponential amplification to take place, priming must occur from both ends of a fragment. Even if an oligonucleotide in one of the degenerate mixtures hybridizes to an incorrect cDNA and primes DNA synthesis, the DNA strand that is synthesized will not be amplified, because it will not contain a sequence complementary to one of the oligonucleotides in the second degenerate primer mixture. The cDNA sequence amplified by this procedure contains the unique sequence of the naturally occurring mRNA encoding the region between the two peptides originally sequenced. This unique DNA sequence can then be radioactively labeled and used as a probe for screening a cDNA or genomic library.

➤ *From Protein to Gene and From Gene to Protein with Recombinant DNA Technology*

In the past, researchers have had two basic approaches for unraveling the molecular mechanisms underlying various cellular processes. One approach involves the biochemical purification and analysis of a protein based on its functional characteristics by techniques described in Chapter 3. The other approach involves the characterization and mapping of genes defined by mutations using the classical genetic analyses described in Chapter 8. Recombinant DNA technology and methods for chemical synthesis of DNA provide the link between protein and gene. With the methods discussed in this chapter, today's molecular cell biologists can begin with an isolated protein and clone the gene that encodes it. By adding the techniques described in Chapter 8, they can begin with the concept of a gene identified by the characteristics of a mutant organism and isolate a cDNA clone containing the gene. Ultimately, the encoded protein can be produced in sufficient quantities for detailed study. The marriage of biochemical and genetic approaches by recombinant DNA technology provides an enormously powerful strategy for studying the role of particular proteins in complex processes.

For example, suppose that a small amount of a protein with an interesting function is isolated and purified (Figure 7-38 *left*). Portions of its amino acid sequence can be determined and used to design oligonucleotide probes to isolate a cDNA encoding the protein. The cDNA can then be incorporated into an expression vector to produce much larger amounts of the protein, greatly simplifying further studies. The cDNA also can be used as a probe to isolate the gene encoding the protein from a genomic library.

Once the gene encoding a particular protein has been isolated and cloned, it can be studied in various ways. For example, techniques described in Chapter 11 permit direct analysis of the genomic DNA sequences that regulate activity of a gene, and expression of the encoded protein, during development or in response to changes in the cellular environment. Moreover, as discussed in Chapter 8, a normal gene can be replaced by a mutant form of the gene in some organisms. Such gene replacement allows researchers to directly test the function of the gene in the physiology and development of the organism. Portions of the natural gene sequence also can be replaced with synthetic DNA, thus introducing specific mutations in the encoded protein. With this technique, investigators can analyze the effect of specific changes in a protein's structure on its function. These are powerful approaches for studying how a particular protein functions in complex cellular processes. They depend on isolation of the DNA encoding a specific protein through the use of data on its amino acid sequence or generation of specific antibody against the protein.



◀ FIGURE 7-38 Overview of alternative strategies for studying the roles of specific proteins in complex cellular processes. This chapter highlights the protein-to-gene strategy; Chapter 8 describes the gene-to-protein strategy. In both cases, recombinant DNA technology can link a gene and its encoded protein, so that traditional biochemical and genetic approaches are combined in powerful ways.

Alternatively, suppose that a gene encoding an unknown protein is identified by classical genetic analyses of mutants exhibiting an altered trait (Figure 7-38 *right*). Genomic clones carrying the gene can be isolated, as described in Chapter 8; the cloned gene then can be used to probe a cDNA library to identify a cDNA clone complementary to the mRNA corresponding to the gene. The complete amino acid sequence of the encoded protein can be deduced by sequencing the cDNA (even though the protein has never actually been isolated). Comparison of this amino acid sequence with the sequences of previously studied proteins reveals if it is related to other known proteins and may provide clues about the function of the protein. By use of an expression vector, sufficient amounts of the protein can be produced to test hypotheses about its function.

As we will see in later chapters, this approach has been crucial in identifying and isolating proteins involved in the development of *Drosophila* and mice starting with mutants exhibiting defects in development. For example, genes that are critical early in the development of *Drosophila* and mice have been shown to encode proteins closely related to a particular class of transcription factors, which are proteins that control the transcription of multiple genes (Chapters 11 and 13). Other genes with important functions later in *Drosophila* development have been shown to encode proteins homologous to polypeptide-hormone receptors characterized from mammals. These proteins span the plasma membrane of the cell, communicating information about the extracellular environment to the interior of the cell (Chapters 13 and 20). It is easy to imagine how these two classes of proteins could have a vital role in the development of a multicellular organism and to understand why mutations in the genes encoding them have disastrous consequences.

SUMMARY

Recombinant DNA technology has reshaped the way biological research is carried out today. The technology depends on the use of restriction enzymes, which cut DNA at specific 4- to 8-bp sequences, generating a reproducible set of restriction fragments from the genome of any organism. The sequences left at the termini of many restriction fragments can be covalently joined with the enzyme DNA ligase, allowing insertion of restriction fragments into plasmid vectors, circular DNA molecules that replicate in *E. coli* cells. If such recombinant plasmids carry an antibiotic-resistance gene, they can transform *E. coli* cells to antibiotic resistance. Selection of the rare cells that take up the plasmid is done on plates containing the appropriate antibiotic; on such plates only these transformed cells grow into colonies. Since each colony arises from a single transformed cell, it constitutes a clone of cells all carrying the same recombinant plasmid. In this way, a complex mixture of restriction fragments carried in plasmid vectors can be separated into discrete *E. coli* clones each composed of cells harboring a particular restriction fragment.

Restriction fragments also can be ligated to the arms of the bacteriophage λ genome. Such recombinant λ DNA containing inserted fragments of up to ≈25 kb can be efficiently packaged in vitro into infectious λ virions. When the resulting mixture of recombinant λ virions is used to generate plaques on a lawn of *E. coli,* each plaque is a recombinant λ clone containing a distinct inserted restriction fragment. Cloning in λ vectors is so efficient that nearly all of the genomic DNA of complex multicellular organisms can be represented in a λ genomic library of a

few million independent λ clones. Cosmids are plasmid vectors that contain the λ-phage COS sequence, which allows them to be packaged in vitro into λ virions. By use of cosmid vectors, DNA fragments up to ≈45 kb in length can be introduced into *E. coli* cells. A similar strategy using bacteriophage P1 permits cloning of DNA fragments up to ≈100 kb long.

The ability to manipulate recombinant DNA was vastly extended by the development of techniques for chemically synthesizing single-stranded DNA molecules of any sequence containing up to about 60 nucleotides. Such synthetic oligonucleotides have many uses.

Messenger RNAs can be converted into complementary *cDNAs* with reverse transcriptase. Since almost all eukaryotic mRNAs have a 3′ poly-A tail, a mixture of cDNAs representing nearly all the mRNAs extracted from a particular tissue or cell type can be synthesized by priming reverse transcription with oligo-dT. This mixture of cDNAs can be converted into double-stranded DNA and inserted into plasmid or λ vectors to generate a cDNA library.

Two general methods are used to identify specific clones encoding known proteins within a genomic or cDNA library. One method involves membrane hybridization with radiolabeled, single-stranded cDNA or synthetic oligonucleotide probes. The other method utilizes λ expression vectors to synthesize the protein encoded by the cloned DNA. In this case, a specific clone is identified based on the ability of the encoded protein to specifically bind a monoclonal antibody or other ligand.

Cloned genomic restriction fragments and cDNAs can be separated from the vectors carrying them by cleavage with the restriction enzyme used to prepare the recombinant cloning vector. The cloned DNA then is isolated from the vector DNA by polyacrylamide or agarose gel electrophoresis, which separates single-stranded DNA molecules according to chain length. Restriction sites for multiple different restriction enzymes can be mapped on a cloned fragment, generating a restriction map that is characteristic for each distinct DNA sequence.

The Maxam-Gilbert and Sanger (dideoxy) methods are used to sequence cloned genomic restriction fragments and cDNAs. Both methods depend on the ability of polyacrylamide gel electrophoresis to separate single-stranded DNA molecules differing in length by only one nucleotide. From the nucleotide sequence of a cloned gene, the complete amino acid sequence of the encoded protein is easily deduced. If this sequence exhibits homology with sequences of other previously studied proteins, the protein is likely to have similar biochemical activities. A protein identified in this way can be produced and isolated by use of an expression vector carrying the cDNA encoding it. As discussed in the next chapter, cloning of genes identified by genetic analysis of mutants and sequencing of these genes can provide valuable information about cellular processes disrupted by specific mutations.

The availability of cloned restriction fragments led to development of two very useful methods to detect a specific DNA or RNA in a complex mixture: Southern blotting, which detects specific DNA fragments, and Northern blotting, which detects specific RNAs. Both methods involve gel electrophoresis of a nucleic acid mixture, transfer (blotting) of the separated DNA or RNA molecules to a filter, and exposure of the filter to a specific, radiolabeled DNA probe, which hybridizes to complementary molecules.

The polymerase chain reaction (PCR), which is used to amplify a particular DNA sequence, has numerous applications in basic research and medicine. With this technique, the DNA from a single human cell can be amplified enough to allow sequencing of a particular region, providing an exquisitely sensitive way to detect mutations associated with inherited human diseases.

This chapter has focussed on the experimental strategy of starting with a protein identified and purified by biochemical techniques and eventually isolating the gene encoding it. In the next chapter, we examine the opposite strategy of starting with a gene identified by genetic analysis and eventually isolating the protein it encodes. Both strategies employ many of the same recombinant DNA techniques and are potent tools for analyzing how macromolecules and cells function.

REVIEW QUESTIONS

1. Review the action of restriction endonucleases, paying particular attention to the properties of the sites in DNA which they cleave. Also review the concept of a polylinker as shown in Figure 7-8. What is the advantage of having such a region in a plasmid used for cloning?

For the following question, use the information in the table and in the polylinker sequence presented in Box 1a.

Restriction Endonuclease	Sequence of Cleaved DNA*
*Age*I	A ↓ CCGGT
*Bam*HI	G ↓ GATCC
*Cfr*10I	Pu ↓ CCGGPy
*Hpa*II	C ↓ CGG
*Hin*dIII	A ↓ AGCTT
*Kpn*I	GGTAC ↓ C
*Nae*I	GCC ↓ GGC
*Pst*I	CTGCA ↓ G
*Sac*I	GAGCT ↓ C
*Sal*I	G ↓ TCGAC
*Sma*I	CCC ↓ GGG
*Sph*I	GCATG ↓ C
*Xba*I	T ↓ CTAGA
*Xma*I	C ↓ CCGGG

*An arrow indicates the cleavage site.

Suppose that you were going to use a plasmid to clone a region of DNA. See Box 1a for the polylinker region present in your vector; each of the six base pair restriction sites found in this region is unique in that it is not found elsewhere in the plasmid.

a. You cut the plasmid with *Age*I. Where will it cut the polylinker region? Will it produce sticky or blunt ends?
b. You could potentially use this same enzyme to digest the DNA you wish to insert into this plasmid, but you notice that one other enzyme cleaves DNA at the same sequence as *Age*I. Which is it? Would you be able to use it to digest the DNA you wish to insert into the vector? Could any of the other restriction enzymes be used to produce DNA fragments that can be inserted into the plasmid vector cut with *Age*I?
c. A colleague of yours suggests that you digest the sample of DNA you wish to insert by using *Xma*I. Do you think this is feasible? If recombinants were formed when you used *Xma*I, which of the listed restriction enzymes would cut the new plasmids at the junction between the insert and the vector polylinker?

2. Review the techniques for preparing cDNA and genomic DNA libraries, as well as the means of selecting clones of interest. There are several mutagenic chemical agents that induce changes in DNA by forming methyl groups that react with guanine to form O^6-methylguanine (mG). This particular lesion can instruct DNA polymerase to place a T opposite the modified G as opposed to a C, thus producing base changes in DNA. During evolution, cells have devised a way of removing mG from DNA by a reaction involving a protein that is used only once in the reaction and is not regenerated; it is called O^6-alkylguanine–DNA alkyltransferase (AGT), and the reaction is described as follows:

$$O^6\text{-methylguanine} + \text{R–S–H} \longrightarrow \text{guanine} + \text{R–S–CH}_3,$$

where R–S–H is the enzyme. In this reaction, the methyl group is transferred to a cysteine residue, leaving guanine intact in the DNA and resulting in an inactive enzyme. AGT from humans has been characterized extensively. In this enzyme, the amino acid sequence surrounding the cysteine that accepts the methyl group is

-Pro-Ile-Leu-Ile-Pro-Cys-His-Arg-Val-

The underlined Cys represents the methyl-accepting group.

a. You wish to clone the gene for AGT from humans. Would you prepare a cDNA library or a genomic library to do this? What would be the advantage or disadvantage of each? Would you need to consider the conservation of the peptide sequence referred to above? What is the relevance of expression of the gene encoding AGT to your decision? Refer to your answer to these questions to defend your choice. If you chose to prepare both types of libraries, describe how you would obtain them.
b. Assume that you have prepared a cDNA library from the cells you are using to clone AGT, and you employed bacteriophage λ as the vector. How would you select plaques harboring the cDNA of interest? If you used the amino acid sequence given above, how many degenerate probes could exist, and how would you prepare them?

3. Compare and contrast the methods of Southern blotting and Northern blotting. What is each used for? Pay particular attention to the different conditions required for nucleic acid transfer in each case. Can RNA be transferred in an alkaline solution?

O^6-Alkylguanine–DNA alkyltransferase (AGT) activity (see question 2) is not found in all cells. You have a series of human cell lines in your laboratory, some of which are actually derived from tumors.

a. You are interested in using the cloned AGT DNA from the previous problem to detect mRNA corresponding to AGT in each of the cell lines. How would you do this? The results you obtain are shown in the following table, along with those showing the actual protein activity as measured by a specific assay for AGT protein activity.

Cell Line	Relative AGT mRNA	Relative AGT Activity
A	++++	++++
B	++++	−
C	−	−
D	−	−
E	+	+

Box 1a
5′-GAATTCGAGCTCGGTACCCGGGGATCCACCGGTGTCGACCTGCAGGCATGCAAGCTT-3′

Relate these results to expression of the *AGT* gene and translation of the message.

b. Suppose that you wanted to prepare large quantities of the AGT protein. Devise a method for doing this.

c. You wish to determine if the *AGT* gene is present in the cell lines lacking AGT mRNA by using the technique of Southern blotting. What restriction enzyme would you choose for this purpose, or would you choose several? Suppose that you choose the restriction enzymes *Eco*RI and *Hin*dIII to digest the DNA from each of your cell lines. You then resolve the DNA by electrophoresis, transfer it to a membrane, and use your clone to probe the filters. The results are in the following table.

	Sizes of Bands (base pairs)		
Cell Line	***Eco*RI**	***Hin*dIII**	***Eco*RI/*Hin*dIII**
A	25,000	18,000 12,000	14,000 11,000
B	25,000	30,000	25,000
C	25,000	18,000 12,000	14,000 11,000
D	—*	—	—
E	25,000	18,000 12,000	14,000 11,000

*A dash (—) denotes that no bands were detected.

Construct a map showing the relative positions of the restriction sites for *Hin*dIII and *Eco*RI at the genetic locus that encodes AGT. Is this map unique, or is there more than one possible solution? If you wanted to distinguish among several possible restriction maps for these results, what additional data would you need?

d. Based on data from the Southern, Northern, and AGT activity measurements, explain further the pattern of AGT activity found in each cell line.

References

General Methods of Recombinant DNA Technology

AUSUBEL, F. M., et al. 1987. *Current Protocols in Molecular Biology.* Wiley.

SAMBROOK, J., E. F. FRITSCH, and T. MANIATIS. 1989. *Molecular Cloning.* Cold Spring Harbor Laboratory.

DNA Cloning with Plasmid Vectors

BERG, P. 1981. Dissections and reconstructions of genes and chromosomes. *Science* 213:296–303.

BOLIVAR, F., et al. 1977. Construction and characterization of new cloning vehicles. II: A multipurpose cloning system. *Gene* 2:95–107.

COHEN, S. N., et al. 1973. Construction of biologically functional bacterial plasmids in vitro. *Proc. Nat'l. Acad. Sci. USA* 70:3240–3244.

OLD, R. W., and S. B. PRIMROSE. 1985. *Principles of Gene Manipulation: An Introduction to Genetic Engineering.* Blackwell Scientific Publications.

WATSON, J. D., et al., 1992, *Recombinant DNA,* 2d ed. Scientific American Books.

Bacteriophage λ as a Cloning Vehicle

COLLINS, J., and B. HOHN. 1978. Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage λ heads. *Proc. Nat'l. Acad. Sci. USA* 75:4242–4246.

DUNN, I. S., and F. R. BLATTNER. 1987. Charons 36 to 40: Multi enzyme, high-capacity, recombination-deficient replacement vectors with polylinkers and polystuffers. *Nucl. Acids Res.* 15:2677–2701.

HOHN, B. 1979. In vitro packaging of λ and cosmid DNA. *Methods Enzymol.* 68:299–308.

FRISCHAUF, A.-M., et al. 1983. Lambda replacement vectors carrying polylinker sequences. *J. Mol. Biol.* 170:827–848.

MURRAY, N. E., and K. MURRAY. 1974. Manipulation of restriction targets in phage λ to form receptor chromosomes for DNA fragments. *Nature* 251:476–483.

RAMBACH, A., and P. TIOLLAIS. 1974. Bacteriophage λ having *Eco*RI endonuclease sites only in the nonessential region of the genome. *Proc. Nat'l. Acad. Sci. USA* 71:3927–3931.

THOMAS, M., J. R. CAMERON, and R. W. DAVIS. 1974. Viable molecular hybrids of bacteriophage lambda and eukaryotic DNA. *Proc. Nat'l. Acad. Sci. USA* 71:4579–4583.

Chemical Synthesis of DNA

CARUTHERS, M. H. 1985. Gene synthesis machines: DNA chemistry and its uses. *Science* 230:281–285.

ITAKURA, K., and A. D. RIGGS. 1980. Chemical DNA synthesis and recombinant DNA studies. *Science* 209:1401–1405.

ITAKURA, K., J. J. ROSSI, and R. B. WALLACE. 1984. Synthesis and use of synthetic oligonucleotides. *Ann. Rev. Biochem.* 53:323–356.

cDNA Cloning

GUBLER, U., and B. J. HOFFMAN. 1983. A simple and very efficient method for generating cDNA libraries. *Gene* 25:263–289.

HAN, J. H., C. STRATOWA, and W. J. RUTTER. 1987. Isolation of full-length putative rat lysophospholipase cDNA using improved methods for mRNA isolation and cDNA cloning. *Biochemistry* 26:1617–1632.

JENDRISAK, J., R. A. YOUNG, and J. D. ENGEL. 1987. Cloning cDNA into λgt10 and λgt11. *Methods Enzymol.* 152:359–385.

OKAYAMA, H., and P. BERG. 1982. High-efficiency cloning of full-length cDNA. *Mol. Cell. Biol.* 2:161–170.

Identifying Clones of Interest

BENTON, W. D., and R. W. DAVIS. 1977. Screening λgt recombinant clones by hybridization to single plaques in situ. *Science* 196:180–183.

GRUNSTEIN, M., and D. S. HOGNESS. 1975. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proc. Nat'l. Acad. Sci. USA* 72:3961–3965.

HALL, B. D., and S. SPIEGELMAN. 1961. Sequence complementarity of T2 DNA and T2-specific RNA. *Proc. Nat'l. Acad. Sci. USA* 47:137–146.

SINGH, H., et al. 1988. Molecular cloning of an enhancer binding protein: isolation by screening of an expression library with a recognition site DNA. *Cell* 52:415–423.

WALLACE, R. B., et al. 1981. The use of synthetic oligonucleotides as hybridization probes. II: Hybridization of oligonucleotides of mixed sequence to rabbit β-globin DNA. *Nucl. Acids Res.* 9:879–887.

WETMUR, J. G., and N. DAVIDSON. 1968. Kinetics of renaturation of DNA. *J. Mol. Biol.* 31:349–370.

YOUNG, R. A., and R. W. DAVIS. 1983. Efficient isolation of genes by using antibody probes. *Proc. Nat'l. Acad. Sci. USA* 80:1194–1198.

Restriction Enzymes

DAVIES, K. E., ed. 1988. *Genome Analysis: A Practical Approach.* Oxford: IRL Press.

MERTZ, J. E., and R. W. DAVIS. 1972. Cleavage of DNA by RI restriction endonuclease generates cohesive ends. *Proc. Nat'l. Acad. Sci. USA* **69**:3370–3374.

NATHANS, D., and H. O. SMITH. 1975. Restriction endonucleases in the analysis and restructuring of DNA molecules. *Ann. Rev. Biochem.* **44**:273–293.

NELSON, M., and M. MCCLELLAND. 1992. The effect of site-specific methylation on restriction-modification enzymes. *Nucl. Acids Res.* **20** (Suppl.):2145–2157.

ROBERTS, R. J. 1992. Restriction enzymes and their isoschizomers. *Nucl. Acids Res.* **20** (Suppl.):2167–2180.

SMITH, H. O. 1970. Nucleotide sequence specificity of restriction endonucleases. *Science* **205**:455–462.

Gel Electrophoresis of DNA Fragments

AIJ, C., and P. BORST. 1972. The gel electrophoresis of DNA. *Biochem. Biophys. Acta* **269**:192–200.

ANDREWS, A. T. 1986. *Electrophoresis,* 2d ed. Oxford University Press.

CANTOR, C. R., C. L. SMITH, and M. K. MATTHEW. 1988. Pulsed-field gel electrophoresis of very large DNA molecules. *Ann. Rev. Biophys. Biophys. Chem.* **17**:41–72.

CARLE, G. F., M. FRANK, and M. V. OLSON. 1986. Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. *Science* **232**:65–70.

CRICKWOOD, D., and B. D. HAMES, eds. 1982. *Gel Electrophoresis of Nucleic Acids.* London: IRL Press Ltd.

SHARP, P. A., B. SUGDEN, and J. SAMBROOK. 1973. Detection of two restriction endonuclease activities in *Hemophilus parainfluenzae* using analytical agarose-ethidium bromide electrophoresis. *Biochemistry* **12**:3055–3062.

SMITH, C. L., et al. 1987. A physical map of the *Escherichia coli* K12 genome. *Science* **236**:1448–1453.

SMITH, H. O., and M. L. BERNSTIEL. 1976. A simple method for DNA restriction-site mapping. *Nucl. Acids Res.* **3**:2387–2398.

DNA Sequencing

MAXAM, A. M., and W. GILBERT. 1977. A new method for sequencing DNA. *Proc. Nat'l. Acad. Sci. USA* **74**:560–564.

MAXAM, A. M., and W. GILBERT. 1980. Sequencing end-labeled DNA with base-specific chemical-cleavages. *Methods Enzymol.* **65**:499–560.

SANGER, F. 1981. Determination of nucleotide sequences in DNA. *Science* **214**:1205–1210.

Analysis of Specific Nucleic Acids in Complex Mixtures

ALWINE, J. C., D. J. KEMP, and G. R. STARK. 1977. A method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Nat'l. Acad. Sci. USA* **74**:5350–5354.

BERK, A. J., and P. A. SHARP. 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease digested hybrids. *Cell* **12**:721–732.

SOUTHERN, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503–515.

THOMAS, P. S. 1980. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc. Nat'l. Acad. Sci. USA* **77**:5201–5205.

Expression Vectors

ELROY-STEIN, O., T. R. FUERST, and B. MOSS. 1989. Cap-independent translation of mRNA conferred by encephelomyocarditis virus 5′ sequence improves the performance of the vaccinia virus/bacteriophage T7 hybrid expression system. *Proc. Nat'l. Acad. Sci. USA* **86**:6126–6130.

LUCKOW, V. A., and M. D. MILLER. 1989. High level expression of nonfused foreign proteins with *Autographa california* nuclear polyhedrosis virus expression vectors. *Virology* **170**:31–39.

NAGATA, S., et al. 1986. Molecular cloning and expression of cDNA for human granulocyte colony-stimulating factor. *Nature* **319**:415–418.

PELHAM, H. R. B., and R. J. JACKSON. 1976. An efficient mRNA dependent translation system from reticulocyte lysates. *Eur. J. Biochem.* **67**:247–256.

TABOR, S., and C. C. RICHARDSON. 1985. A bacteriophage T7 RNA polymerase promoter system for controlled exclusive expression of specific genes. *Proc. Nat'l. Acad. Sci. USA* **82**:1074–1078.

SOUZA, L. M., et al. 1986. Recombinant human granulocyte-colony stimulating factor: effects on normal and leukemic myeloid cells. *Science* **232**:61–65.

STUDIER, F. W., and B. A. MOFFATT. 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**:113–134.

Polymerase Chain Reaction

ERLICH, H., ed. 1992. *PCR Technology: Principle and Applications for DNA Amplification.* W. H. Freeman and Company.

ERLICH, H. E., R. A. GIBBS, and H. H. KAZAZIAN JR., eds. 1989. *Polymerase Chain Reaction.* Cold Spring Harbor Laboratory.

HANDYSIDE, A. H., et al. 1992. Birth of a normal girl after in vitro fertilization and preimplantation diagnostic testing for cystic fibrosis. *New Engl. J. Med.* **327**:905–909.

INNIS, M. A., et al. 1990. *PCR Protocols: A Guide to Methods and Applications.* Academic Press.

SAIKI, R. K., et al. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**:487–491.